# Information, Entropy and Their Geometric Structures

Edited by

Frédéric Barbaresco and

Ali Mohammad-Djafari

Printed Edition of the Special Issue Published in *Entropy*


www.mdpi.com/journal/entropy

Frédéric Barbaresco and Ali Mohammad-Djafari (Eds.)

# Information, Entropy and Their Geometric Structures

MDPI

This book is a reprint of the Special Issue that appeared in the online, open access journal, *Entropy* (ISSN 1099-4300) from 2014–2015 (available at: http://www.mdpi.com/journal/entropy/special_issues/entropy-Geome).

*Guest Editors*
Frédéric Barbaresco
Advanced Radar Concepts Business Unit
Thales Air Systems S.A., Voie Pierre-Gilles de Gennes F91470 Limours
France

Ali Mohammad-Djafari
Laboratoire des Signaux et Systmes
UMR 8506 CNRS-SUPELEC-UNIV PARIS SUD
Gif-sur-Yvette
France

*Editorial Office*
MDPI AG
Klybeckstrasse 64
Basel, Switzerland

*Publisher*
Shu-Kun Lin

*Managing Editor*
Jely He

**1. Edition 2015**

MDPI • Basel • Beijing • Wuhan • Barcelona

ISBN 978-3-03842-103-0 (Hbk)
ISBN 978-3-03842-104-7 (PDF)

# Table of Contents

## Chapter 3: Applications of Information/Entropy Geometric Structures

# List of Contributors

**Roger Balian:** Institut de Physique Théorique, CEA/Saclay, F-91191 Gif-sur-Yvette Cedex, France
**Frédéric Barbaresco:** Thales Air Systems, Advanced Radar Concepts Business Unit, Voie Pierre-Gilles de Gennes, Limours F-91470, France
**Pierre Baudot:** Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, 04103 Leipzig, Germany
**Anass Bellachehab:** Telecom SudParis, Institut Mines-Télécom, UMR CNRS 5157 SAMOVAR, 9 Rue Charles Fourier, 91000 Évry, France
**Youssef Bennani:** CNRS, Laboratoire I3S-UMR 7271, Université de Nice-Sophia Antipolis/CNRS, 06900 Sophia Antipolis, France
**Daniel Bennequin:** Universite Paris Diderot-Paris 7, UFR de Mathematiques, Equipe Geometrie et Dynamique, Batiment Sophie Germain, 5 rue Thomas Mann, 75205 Paris Cedex 13, France
**Jérémy Bensadon:** Laboratoire de Recherche en Informatique, Université Paris-Sud, 91400 Orsay, France
**Stefano Bordoni:** Department of Pharmacy and Biotechnology, University of Bologna—Rimini Campus, Via Dei Mille 39—47921 Rimini, Italy
**Jaehyung Choi:** Department of Applied Mathematics and Statistics, State University of New York (SUNY), StonyBrook, NY 11794, USA
**Shinto Eguchi:** The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan
**Misha Gromov:** Institut Hautes Études Scientifiques, 35, Route de Chartres, F-91440 Bures-sur-Yvette, France
**Mitsuhiro Itoh:** Institute of Mathematics, University of Tsukuba, 1-1-1, Ten-noudai, Tsukuba, 305-8571, Japan
**Osamu Komori:** The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan
**José Carlos Magossi:** School of Technology (FT), University of Campinas (Unicamp), Rua Paschoal Marmo 1.888, 13484-370 Limeira, São Paulo, Brazil
**Luigi Malagò:** Department of Electrical and Electronic Engineering, Shinshu University, Nagano, Japan; Inria Saclay, Île-de-France, Orsay Cedex, France
**Nina Miolane:** INRIA, Asclepios project-team, 2004 Route des Lucioles, BP93, Sophia Antipolis Cedex F-06902, France
**Ali Mohammad-Djafari:** Laboratoire des Signaux et Systèmes, UMR 8506 CNRS-SUPELEC-UNIV PARIS SUD, SUPELEC, Plateau de Moulon, 3 rue Juliot-Curie, 91192 Gif-sur-Yvette, France
**Andrew P. Mullhaupt:** Department of Applied Mathematics and Statistics, State University of New York (SUNY), StonyBrook, NY 11794, USA
**Xavier Pennec:** INRIA, Asclepios project-team, 2004 Route des Lucioles, BP93, Sophia Antipolis Cedex F-06902, France

**Giovanni Pistone:** De Castro Statistics, Collegio Carlo Alberto, Moncalieri, Italy
**Luc Pronzato:** CNRS, Laboratoire I3S-UMR 7271, Université de Nice-Sophia Antipolis/CNRS, 06900 Sophia Antipolis, France
**Maria João Rendas:** CNRS, Laboratoire I3S-UMR 7271, Université de Nice-Sophia Antipolis/CNRS, 06900 Sophia Antipolis, France
**Olivier Rioul:** Télécom ParisTech, Institut Mines-Télécom, CNRS LTCI, 46 Rue Barrault, 75013, Paris, France
**Hiroyasu Satoh:** Nippon Institute of Technology, Saitama, 345-8501, Japan
**Takashi Takenouchi:** Future University Hakodate, 116-2 Kamedanakano, Hakodate Hokkaido 041-8655, Japan
**Geert Verdoolaege:** Department of Applied Physics, Ghent University, Sint-Pietersnieuwstraat 41, B-9000 Ghent, Belgium; Laboratory for Plasma Physics—Royal Military Academy (LPP-ERM/KMS), Avenue de la Renaissancelaan 30, B-1000 Brussels, Belgium
**Udo von Toussaint:** Max-Planck-Institute for Plasmaphysics, Boltzmannstrasse 2, 85748 Garching, Germany
**Jun Zhang:** Department of Psychology and Department of Mathematics, University of Michigan, 530 Church Street, Ann Arbor, MI 48109, USA

# About the Guest Editors

**Frédéric Barbaresco** received his State Engineering degree from the French Grand Ecole SUPELEC, Paris, France, in 1991. Since then, he has worked for the THALES Group where he is now Senior Scientist and Advanced Studies Manager in the Advanced Radar Concept Business Unit of THALES AIR SYSTEMS.

He has been an Emeritus Member of SEE since 2011 and he was awarded the Aymé Poirson Prize (for application of sciences to industry) by the French Academy of Sciences in 2014, the SEE Ampere Medal in 2007, the Thévenin Prize in 2014 and the NATO SET Lecture Award in 2012. He is President of SEE Technical Club ISIC "Engineering of Information and Communications Systems" (https://www.see.asso.fr/ct-isic) and a member of the SEE administrative board.

In 2009, he founded the Leon Brillouin Seminar on "Geometric Sciences of Information" (http://repmus.ircam.fr/brillouin/home) hosted by IRCAM in Paris, France. He has organized a tutorial on "Modern Radar Processing based on Geometry of Structured Matrices and Information Geometry" (http://www.radar2014.org/tutorials/98#TUTORIAL1) for the Radar'14 Conference, and a Short Course on "Geometric Radar Processing Breakthrough based on Fisher Information Geometry" (http://www.eumweek.com/docs/workshops/SCW01.pdf) for the EuRAD'15 Conference. He was co-organizer of the French-Indian MIG'11 Workshop at Ecole Polytechnique and Thales Research and Technology (https://www.sonycsl.co.jp/person/nielsen/infogeo/MIG/) and a coordinator of Springer Lecture Notes on "Matrix Information Geometry" published in 2012 (http://www.springer.com/us/book/9783642302312). He was an invited speaker at the SMC-NEGAA "Structured Matrix Computations in Non Euclidean Geometries" workshop (http://www-lmpa.univ-littoral.fr/SMC-NEGAA2012/participants.html) at CIRM Luminy in 2012. In 2012 he organized a Symposium at Institut Henri Poincaré on "Optimal Transport and Information Geometry" (https://www.ceremade.dauphine.fr/~peyre/mspc/mspc-thales-12/). He was an invited lecturer for UNESCO on "Advanced School and Workshop on Matrix Geometries and Applications" in Trieste (http://indico.ictp.it/event/a12193) at the ITCP (International Center for Theoretical Physics) in June 2013. He is the General Co-chairman of the new international conference GSI "Geometric Sciences of Information", first edition GSI'13 at Ecole des Mines in 2013 (https://www.see.asso.fr/gsi2013) and 2nd edition GSI'15 at Ecole Polytechnique in 2015 (https://www.see.asso.fr/gsi2015). He is the co-editor of "Geometric Science of Information" (http://www.springer.com/us/book/9783642400193), a book published by Springer in 2013, and one of the contributors to the Springer book "Geometric Theory of Information" (http://www.springer.com/us/book/9783319053165) published in 2014. He co-organized MaxEnt'14 conference in Amboise in 2014 (https://www.see.asso.fr/maxent14) and is the editor of Proceedings published by the AIP "American Institute of Physics" (http://scitation.aip.org/content/aip/proceeding/aipcp/1641). He was invited to be a keynote speaker for the GIO'14 workshop in Bordeaux on "Geometry of Information and Optimization" (https://sites.google.com/site/bordeauxgio2014/).

**Ali Mohammad-Djafari** received his B.Sc. degree in electrical engineering from the Polytechnic of Teheran, in 1975, a diploma degree (M.Sc.) from Ecole Supérieure d'Electricit (SUPELEC), Gif sur Yvette, France, in 1977, a “Docteur-Ingénieur” (Ph.D.) degree and “Doctorat d'Etat” in Physics, from the University of Paris Sud 11 (UPS), Orsay, France, respectively in 1981 and 1987.

He was Assistant Professor at UPS for two years (1981–1983). Since 1984, he has held a permanent position at “Centre National de la Recherche Scientifique (CNRS)” and works at “Laboratoire des signaux et systèmes (L2S)” at SUPELEC. He was a visiting Associate Professor at the University of Notre Dame, Indiana, USA, from 1997–1998. From 1998 to 2002, he has been the Head of Signal and Image Processing Division at this laboratory.

Presently, he is the “Directeur de recherche” and his main scientific interests are in developing new probabilistic methods based on Bayesian inference, information theory and maximum entropy approaches for inverse problems in general, in all aspects of data processing, and, more specifically, in imaging and vision systems: image reconstruction; signal and image deconvolution; blind source separation; sources localization; data fusion; multi and hyper spectral image segmentation. The main application domains of his interests are medical imaging, computed tomography (X rays, PET, SPECT, MRI, microwave, ultrasound and eddy current imaging) either for medical imaging or for non-destructive testing (NDT) in industry, multivariate and multi-dimensional data, signal and image processing, data mining, clustering, classification and machine learning methods for biological or medical applications.

He has supervised over 20 Ph.D. thesis’, 20 post-doc research activities and 50 M.Sc. student research projects. He has published over 50 full journal papers, 10 book and proceedings editions and more than 200 papers in national and international conferences. He has organized or co-organized more than 10 international workshops and conferences. He has been further been the expert and consultant for a great number of French national and international projects. Since 1988 he has held many teaching activities as Professor in M.Sc. and Ph.D. Level in SUPELEC, University of Paris Sud (UPS), ENSTA and Ecole Centrale de Paris (ECP).

He also participated and managed industrial contracts with many French national industries such as EDF, RENAUL, PEUGEOT, THALES, SAFRAN, CARESTREAM and great research institutions such as CEA, INSERM, INRIA, ONERA, as well as regional (such as Digiteo), national (such as ANR) and European projects (such as ERASYSBIO).

For an overview and access to more details of his activities, please see his web page: http://djafari.free.fr for general information; http://djafari.free.fr/news.htm for news and activities; and http://publicationslist.org/djafarie for the list of publications.

# Preface
## From Information Theory to Geometric Science of Information

Venus at the Forge of Vulcan, Le Nain Brothers, Musée Saint-Denis, Reims (Vulcan is the god of fire and god of metalworking and the forge, often depicted with a blacksmith's hammer)

"Intelligence is the faculty of manufacturing artificial objects, especially tools to make tools, and of indefinitely varying the manufacture."—Henri Bergson

Information theory was founded in the 1950s based on the work of Claude Shannon and Jacques Laplume in communication and Léon Brillouin in statistical physics, among other main contributors. These foundations have conventionally been built on linear algebra theory and probability models in conventional spaces (vector space, normed spaces, ...).

What's new since 1950s
INFORMATION THEORY

**Jacques Laplume**
**Radar Dept. Thomson-Houston**

**Claude Shannon**
**MIT**

**Léon Brillouin**
**Collège de France**

At the turn of the century, new and fruitful interactions were found between several branches of science: ***Information Science*** (information theory, digital communications, statistical signal processing, …), ***Mathematics*** (group theory, geometry and topology, probability, statistics, ...) and ***Physical Sciences*** (geometric mechanics, thermodynamics, statistical physics, quantum mechanics, ...).

**From Probability to Geometry**

The probability theory was conceived by Blaise Pascal and Jacob Bernoulli. Pierre de Fermat also helped in his exchange of correspondence with Blaise Pascal to develop the foundations of probability theory, a mathematical accident that caused the study of Chevalier de Méré's game (Antoine Gombaud, Chevalier de Méré, a French nobleman with an interest in gaming and gambling questions, called Pascal's attention to an apparent contradiction concerning a popular dice game). Then, probability theory was consolidated by many contributors, such as Pierre Simon Laplace, Abraham de Moivre and Carl Friedrich Gauss during the XVIII century and by Emile Borel, Andreï Kolmogorov and Paul Levy last century. Probability is again the subject of a new foundation to apprehend new structures and generalize the theory to more abstract spaces (metric spaces, homogeneous manifolds, graphs ....). A first attempt at probability generalization in metric spaces was developed by Maurice Fréchet in the middle of last century, in the framework of abstract spaces topologically affine and "distance space" ("espace distancié") with triangular inequality constraint.

# LES ESPACES ABSTRAITS TOPOLOGIQUEMENT AFFINES.

PAR

MAURICE FRÉCHET

à STRASBOURG.

Un grand nombre des propriétés topologiques de l'espace euclidien s'étendent immédiatement à tous les espaces où une définition de la limite étant donnée (qui est en général imposée par la nature des éléments ou points de l'espace et les applications qu'on a en vue), cette définition peut s'exprimer par l'intermédiaire d'une *distance*.[1] Nous entendons par là qu'à tout couple $A, B$ d'éléments ou points de l'espace considéré correspond un nombre $(A, B)=(B, A)\geq 0$, qui n'est nul que si $A$ et $B$ ne sont pas distincts et qui satisfait aux deux conditions suivantes:

I. Pour trois points $A, B, C$ arbitraires, on a toujours

$$(A, B)\leq(A, C)+(C, B).$$

II. La condition nécessaire et suffisante pour qu'une suite de points $A_1, A_2, \ldots$ de cet espace tende vers le point $A$ de cet espace est que la distance $(A, A_n)$ tende vers zéro.

Un tel espace sera appelé un espace $(D)$ (initiale de distance).[1] Dans le cas où l'on n'impose pas la condition I, $(A, B)$ sera un écart[1] et l'espace sera un espace $(E)$[1].

## From Statistics to Geometry

In the middle of last century, another branch of geometric approaches of statistical problems has been initiated by Calyampudi Radhakrishna Rao that introduced a metric space in the parameters space of probability densities. The metric tensor was proved to be equal to the Fisher Information matrix. This result was axiomatized by Nikolai Nikolaevich Chentsov in the framework of category theory. Having been introduced in 1939, the lower bound in statistics, six years before C.R. Rao, this idea was latent in the work of Maurice Fréchet, who had noticed that the "distinguished densities" that reach this lower bound are defined by a function that is given by a solution of Legendre-Clairaut equation. Nowadays, this Legendre-Clairaut equation is the cornerstone of "Information Geometry" theory linking two dual potential functions in dual spaces. In parallel, Jean-Louis Koszul had constructed a Hessian geometry on convex cones, through the concept of Koszul-Vinberg characteristic function and Koszul forms. Koszul Information Geometry is a generalization of information geometry theory, where invariance with respect to densities parameters is replaced by invariance with respect to automorphisms of these convex cones where these parameters lie. In 1957, the framework was consolidated by the principle of Maximum Entropy, expounded by E. T. Jaynes in two papers where he emphasized a natural correspondence between statistical mechanics and information theory. In particular, Jaynes offered a link to statistical physics and a rationale as to why the Gibbsian method of statistical mechanics works. He argued that the entropy of statistical mechanics and the information entropy of information theory are principally the same thing.

Consequently, statistical mechanics should be seen just as a particular application of a general tool of logical inference and information theory.

**From Thermodynamics to Geometry**

On the side of statistical physics and thermodynamics—which were based on the seminal works of Sadi Carnot, Rudolf Clausius, Ludwig Boltzmann, François Massieu and Williard Gibbs—several geometric attempts were developed later as a "general equation of thermodynamics" by Pierre Duhem unifying in the same equations all changes of systems' positions and states. In his 1891 Paper, « Sur les équations générales de la Thermodynamique », Pierre Duhem wrote "*We made a special case, the dynamics of thermodynamics, a science that embraces common principles in all the changes of state of the bodies, both changes of places and changes in physical qualities.*" Four scientists were credited by Duhem with having carried out the most important researches on that subject: François Massieu to derive thermodynamics from a characteristic function and its partial derivatives; J. W. Gibbs to show that Massieu's functions could play the role of potentials in the determination of the states of equilibrium in a given system; H. von Helmholtz to put forward similar ideas (and analogy between thermodynamic and mechanics); and A. von Oettingen to give an exposition of thermodynamics of remarkable generality based on the general duality concept. More recently, we can make references to the "Lie Group thermodynamics" theory created by Jean-Marie Souriau in the framework of geometric mechanics and symplectic geometry, or the concept of thermodynamics contact manifolds that was conceptualized by Vladimir Arnold. This geometrization of thermodynamics and mechanics was also extended to quantum mechanics by Roger Balian, providing also a bridge with information geometry. Roger Balian, in 1986, introduced a geometric structure through extension of the Fisher metric in statistical physics and quantum mechanics, compatible with gauge theory of thermodynamics.

**From Mechanics to Geometry**

The last branch of geometric structure elaboration for information is emerging through the inter-relations between "geometric mechanics" and the "geometric science of information", that will be largely debated at the GSI'15 conference (www.gsi2015.org). We can imagine that other links could be discovered between mechanics and geometry, for instance based on the elastic theory of the Cosserat brothers that should enlighten new seminal works as discovered by Jean-François Pommaret. In 1926, Louis-Maurice Roy, published in Annals, "a thermodynamic theory of elastic line […]", directly inspired by the relatively recent work of Duhem and M. M. Cosserat. This idea was also developed in Louis de Broglie's book on thermodynamics.

Regarding geometry and mechanics, for the anecdote, we can observe that the master of geometry during the last century, Elie Cartan, was the son of Joseph Cartan who was the village blacksmith, and Elie recalled that his childhood had passed under "blows of the anvil, which started every morning from dawn". We can imagine easily that the child, Elie Cartan, watching his father Joseph "coding curvature" on metal between the hammer and the anvil, insidiously influencing Elie's mind with germinal intuition of fundamental geometric concepts. The alliance of geometry

and mechanics is beautifully given by this image of Forge, as illustrated in this painting of Velasquez about Vulcan God. This concordance of meaning is confirmed by the etymology of the word "Forge", that comes from the late XIV century, "a smithy", from Old French forge "forge, smithy" (XII century), earlier faverge, from Latin fabrica "workshop, smith's shop", from faber (genitive fabri) "workman in hard materials, smith". One can imagine the hammer blows given by Joseph on the anvil, giving shape and curvature to the metal, inspired the curious mind of Elie that surely inspired later his intuition of "moving frame" and "nonholonomic space" in geometry. Elie Cartan was motivated by the objective to build new foundations of geometry. He said "*distinguished service that has rendered and will make even the absolute differential calculus of Ricci and Levi-Civita should not prevent us to avoid too exclusively formal calculations, where debauchery indices often mask a very simple geometric fact. It is this reality that I have sought to put in evidence everywhere.*" (« *Les services éminents qu'a rendus et que rendra encore le Calcul différentiel absolu de Ricci et Levi-Civita ne doivent pas nous empêcher d'éviter les calculs trop exclusivement formels, où les débauches d'indices masquent une réalité géométrique souvent très simple. C'est cette réalité que j'ai cherché à mettre partout en évidence.*» in É. Cartan, Leçons sur la théorie des espaces de Riemann, Paris: Gauthier-Villars, 2e éd., 1946, p. VII).

Into the Flaming Forge of Vulcan, into the Ninth Sphere, Mars descends in order to retemper his flaming sword and conquer the heart of Venus (Diego Velázquez, Museo Nacional del Prado)

**Groups Everywhere and Metrics Everywhere**

Geometric structure can also be considered through group theory. As observed by Gaston Bachelard, mathematical physics, incorporating at its core the concept of group, brand supremacy. All rational geometries, and without doubt more generally all mathematical organizations of experience, are characterized by a special group of transformations. The group provides evidence of mathematics closed on itself. Its discovery closes the era of conventions, more or less independent, more or less coherent. Henri Poincaré said that if we strip the mathematical theory of which appears to be an accident, that is to say its material, there will remain only the essential, that is to say, the form; and this form, which is as it were the solid skeleton of the theory, will be the group's structure. Concerning Elie Cartan's work, Henri Poincaré said that "*the problems addressed by Elie Cartan are among the most important, most abstract and most general dealing with mathematics; group theory is, so to speak, the whole mathematics, stripped of its material and reduced to pure form. This extreme level of abstraction has probably made my presentation a little dry; to assess each of the results, I would have had virtually render him the material which he had been stripped; but this refund can be made in a thousand different ways; and this is the only form that can be found as well as a host of various garments, which is the common link between mathematical theories that are often surprised to find so near*". "Groups everywhere" and "metrics everywhere" are then the new leitmotiv in mathematics and physics. In particular, a central role could be attributed to Misha Gromov and his contribution to metric spaces. The analysis of the invariants and the transformations preserving them is at the core of Gromov's work on "geometrical group theory".

## RAPPORT SUR LES TRAVAUX DE M. CARTAN

fait à la Faculté des Sciences de l'Université de Paris.

PAR

H. POINCARÉ.

. . . . . Le rôle prépondérant de la théorie des groupes en mathématiques a été longtemps insoupçonné; il y a quatre-vingts ans, le nom même de groupe était ignoré. C'est GALOIS qui, le premier, en a eu une notion claire, mais c'est seulement depuis les travaux de KLEIN et surtout de LIE que l'on a commencé à voir qu'il n'y a presque aucune théorie mathématique où cette notion ne tienne une place importante.

**Entropy Everywhere**

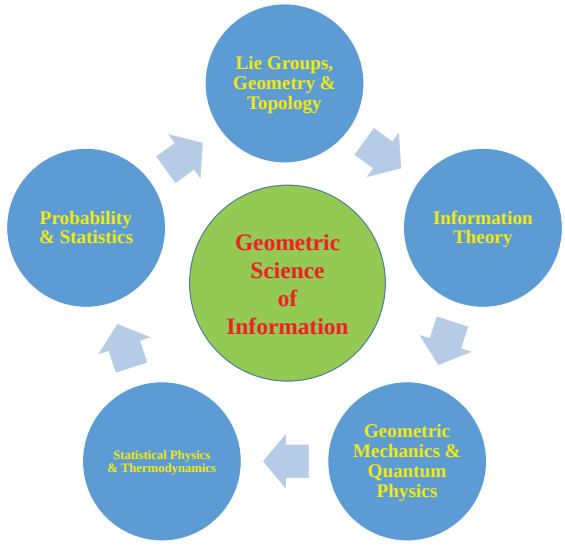
From its beginnings, the theory of information has also been linked to statistical physics through the concept of entropy. This intimate relationship between information and entropy was studied by Léon Brillouin and Claude Shannon. The latter writes: "*My biggest concern was what to call it. I thought to call information, but the term was used too, so I decided to call it uncertainty. When I was talking with John von Neumann, he had a better idea. He said to me, you should call it entropy, for two reasons. First, your uncertainty function was used in statistical mechanics under that name, so it already has a name. Second, and most important, no one really knows what the entropy, so in a debate you would always have the advantage*". René Thom also tried to show in which direction a real information theory could go, being halfway between semantics and semiotics, thermodynamics of real forms, would attempt to return a proper analysis of morphological forms of messages.

**Linguistics Everywhere**

To conclude this preface, if we go back further in history, let's look at the etymological origins of the word "information". First written as "enformer" the word "inform" appears in French in 1286, the Latin word "informare", literally "shape". The word "information" appears in the XIII century. From the Greek etymology, μορφή, morphs ("shape"), we reached the sense of morphology, the science of forms. For Plato, the concept of "form" is designated "morph", "Eidos" and "idea"; Henri Bergson gave his definition of the Greek concept of "Eidos" in the book "Creative Evolution": "*The word, eidos, which we translate here by "Idea", has, in fact, this threefold meaning. It denotes (1) the quality, (2) the form or essence, (3) the end or design (in the sense of intention) of the act being performed, that is to say, at bottom, the design (in the sense of drawing) of the act supposed accomplished*." These three aspects are those of the adjective, substantive and verb, and correspond to the three essential categories of language, proving, as Jean-Marie Souriau did, that we have to apprehend "the grammar of nature".

**Geometric Science of Information as a Federative Structure and Grammar**

Henri Poincaré said that "Mathematics is the art of giving the same name to different things" (« La mathématique est l'art de donner le même nom à des choses différentes» in «Science et méthode", 1908). By paraphrasing Henri Poincaré, we could claim that « Geometric Science of Information » is the art of giving the same name to different sciences. The rules, the structure and architecture of this new "manufacture" is a kind of new Grammar for Sciences.




**Book Chapters Survey**

The aim of this book is to provide an overview of current work addressing this topic of research that explores the geometric structures of information and entropy. These papers are an extended version of the paper published in Proceedings (http://printorders.aip.org/proceedings/1641) of the 34th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MaxEnt 2014), Amboise, France, 21–26 September 2014 (https://www.see.asso.fr/maxent14).

Chapter 1 of the book is a historical review of the origins of thermodynamics and information theory:

- Stefano Bordoni analyses and puts in perspective crosswise the work of J.J. Thomson and P. Duhem in thermodynamics in recent decades of the nineteenth century, with two abstract and phenomenological approaches to thermodynamics. After the analysis of intermediate solutions by Helmholtz, Planck and Oettingen, he describes J.J. Thomson's general theory for physical and chemical processes, and P. Duhem's design of energetics as unification between physics and chemistry.

- Olivier Rioul and José Carlos Magossi then detail the history of the discovery of the Shannon's formula for Gaussian channel, with seminal Hartley's rule twenty years before Shannon for uniform channel, and the first published work in April 1948 of the French engineer, Jacques Laplume from Radar/Hyper Department of Thomson-Houston, among others.

Chapter 2 discusses the mathematical and physical foundations of geometric structures related to information and entropy:

- Misha Gromov, IHES (Institute of Advanced Scientific Studies), Abel Prize 2009, indicates possibilities for (homological and non-homological) linearization of basic notions of the probability theory and also the replacement of the real numbers as values of probabilities by objects of suitable combinatorial categories.
- Pierre Baudot from Max-Planck Institute, and Daniel Bennequin from Institut mathématique de Jussieu, observe that entropy is a universal co-homological class in a theory associated to a family of observable quantities and a family of probability distributions. This gives rise to a new kind of topology for information processes that accounts for the main information function.
- Nina Miolane and Xavier Pennec compute bi-Invariant pseudo-Metrics on Lie Groups for consistent statistics to define a Riemannian metric compatible with the group structure, to perform statistics on Lie groups for computational anatomy.
- Frédéric Barbaresco introduces the Symplectic Structure of Information Geometry based on Souriau's "Lie Group Thermodynamics model", with a covariant definition of Gibbs equilibrium via invariances through co-adjoint action of a group on its momentum space. The Fisher metric is identified as a Souriau Geometric Heat Capacity. This model is compared with hessian Geometry of Jean-Louis Koszul, which is the main pillar of Information Geometry theory.
- Roger Balian introduces in the space of quantum density matrices, a Riemann metric as hessian of the von Neumann entropy, which is physically founded and which characterizes the amount of quantum information lost, underlying the canonical mapping between the spaces of states and of observables, which involves the Legendre transform. Roger Balian provides then its general expression and its explicit form for q-bits.
- Mitsuhiro Itoh and Hiroyasu Satoh study the geometry of Fisher metrics and geodesics on a space of probability measures defined on a compact manifold and its application to geometry of a barycenter map associated with Busemann function on a Hadamard manifold X. They describe a fibre space structure of barycenter map.

Lastly, Chapter 3 is dedicated to applications with numerical schemes for geometric structures of information and entropy:

- Ali Mohammad-Djafari proposes to review the main inference tools using the Bayes rule, maximum entropy principle (MEP), information theory, relative entropy and the Kullback–Leibler (KL) divergence, Fisher information and its corresponding geometries. The second part of the paper is focused on the ways these tools have been used in data, signal and image processing and in the inverse problems, which arise in different physical sciences and engineering applications.
- Jérémy Bensadon, a PhD student of Yann Ollivier, extends the IGO (information geometric optimization) method, a general framework for stochastic optimization problems aiming at limiting the influence of arbitrary parametrization choices. He defines the geodesic IGO, a fully parametrization-invariant algorithm, named GIGO, using the Riemannian structure, and illustrates it for the manifold of Gaussians, thanks to Noether's theorem.
- Luigi Malagò and Giovanni Pistone develop Amari's natural gradient flows of real functions defined on the densities belonging to an exponential family on a finite sample space, that converges to densities with reduced support that belong to the border of the exponential family. They provides an extension based on the algebraic concept of an exponential variety.
- Anass Bellachehab, a PhD student of Jérémie Jakubowicz, presents an application of distributed consensus algorithms to metamorphic systems (a set of identical units that can self-assemble to form a rigid structure). He proposes a distributed algorithm that synchronizes all of the systems in the network, by casting the problem as a consensus problem on a metric space, and using recently distributed consensus algorithms that only makes use of metrical notions.
- Jaehyung Choi and Andrew P. Mullhaupt propose two papers. In the first paper, they construct geometric shrinkage priors for Kählerian signal filters and introduce an algorithm for finding superharmonic priors which outperform the Jeffreys prior, with implication of the algorithm to time series models. In the second paper, they prove the correspondence between the information geometry of a signal filter and a Kähler manifold, and several time series models are studied in the Kählerian information geometry.
- Youssef Bennani, Luc Pronzato and Maria João Rendas propose a new non-parametric density estimator from region-censored observations with application in the context of population studies, with a maximum entropy estimator that satisfies a set of constraints imposing a close fit to the empirical distributions associated with the set of censoring regions.
- Udo von Toussaint derives an explicit prior distribution for the parameters of multivariate linear regression problems in the absence of further prior information, based on geometric invariance properties. The derived prior distribution generalizes the already known special cases, e.g., 2D plane in three dimensions.

- Geert Verdoolaege discusses a new general regression method, called geodesic least squares regression (GLS), based on minimization of the Rao geodesic distance on a probabilistic manifold. He demonstrates the robustness of the method on synthetic data in the presence of significant uncertainty on both the data and the regression model, with application to a scaling law in magnetic confinement fusion.
- Jun Zhang presents an extension to α-geometry. It is further shown here that the resulting metric and α-connections obtained through arbitrary monotone embeddings is a unique extension of the α-geometric structure.
- Takashi Takenouchi, Osamu Komori and Shinto Eguchi investigate the basic properties of binary classification with a pseudo model based on the Itakura–Saito distance and propose a novelmulti-task learning algorithm based on the pseudo model in the framework of the ensemble learning method.

We hope that this vast survey on the geometric structure of information and entropy will motivate readers to go further and explore the emerging domain of "Science of Information".

> "*As regards human intelligence, there is not enough noticed that mechanical invention was first its essential approach ... If we could rid ourselves of all pride, if, to define our species, we kept strictly to what the historic and the prehistoric periods show us to be the constant characteristic of man and of intelligence, we should say perhaps not Homo sapiens, but Homo faber. In short, intelligence, considered in what seems to be its original feature, is the faculty of manufacturing artificial objects, especially tools to make tools, and of indefinitely varying the manufacture.*"
>
> Henri Bergson, The Creative Evolution, 1907

Frédéric Barbaresco
Advanced Radar Concepts Business Unit, Thales Air Systems S.A.,
Voie Pierre-Gilles de Gennes F91470 Limours, France

Prof. Dr. Ali Mohammad-Djafari
Laboratoire des Signaux et Systmes,
UMR 8506 CNRS-SUPELEC-UNIV PARIS SUD,
Gif-sur-Yvette, France

*Guest Editors*

**References**


1. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423, 623–656.
2. Laplume, J. Sur le nombre de signaux discernables en présence du bruit erratique dans un système de transmission à bande passante limitée. *Comptes rendus de l'Académie des Sciences de Paris* **1948**, *226*, 1348–1349. (In French)
3. Brillouin, L. *Science and Information Theory*; Academic Press: New York, NY, USA, 1956. (traduction (fr) La science et la théorie de l'information, Masson, 1959)
4. Pascal, B. Oeuvres complètes, Bibliothèque de la Pléiade, published by Gallimard, Paris, France
5. De Fermat, P. *Œuvres de Fermat*; Ministère de l'instruction publique, Gauthier-Villars et cie: Paris, France, 1891–1922, 5 Volumes.
6. Bernoulli, J. Some Questions about Interest, with a Solution of a Problem about Games of Chance. *Acta eruditorum* **1685**, 219–223.
7. De Moivre, A. *The doctrine of chances: or, A method of calculating the probabilities of events in play*; W. Pearson: London, UK, 1718.
8. Laplace, P.S. *Mémoire sur les Probabilités*; Mémoires de l'Académie royale des sciences de Paris: Paris, France, 1781
9. Gauss, C.F. *Abhandlungen zur Methode der kleinsten Quadrate*; Boersch, A., Simon, P., Eds.; Reprint of the edition of 1887. Contains translations from Latin/reprints of all Gauss's works on the theory of errors (mainly 1809; 1811; 1816; 1823; 1828, as well as their abstracts compiled by Gauss himself); Physica-Verlag: Würzburg, Germany, 1964.
10. Kolmogorov, A. On Analytical Methods in Probability Theory. *Math. Ann.* **1931**, *104*, 415–458. (in German); also in *Selected Works of A. N. Kolmogorov: Volume II Probability Theory and Mathematical Statistics*; Shiryayew, A.N., Ed.; Kluwer: Dordrecht, The Netherlands, 1992.
11. Borel, E. *Traité du calcul des probabilités et ses applications*; Tome III, Les Applications de la Théorie des Probabilités aux Sciences Economiques et Biologiques. Fascicule I. Assurances sur la Vie. Calcul des Primes, by Henri Galbrun; Gauthier-Villars: Paris, France, 1924.
12. Levy, P. Calcul des probabilités; Gauthier-Villars: Paris, France, 1925; reprinted in 2004 by Jacques Gabay.
13. Fréchet, M.R. Les espaces abstraits topologiquement affines. *Acta Mathematica* **1925**, *47*, 25–52.
14. Fréchet, M.R. Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l'institut Henri Poincaré* **1948**, *10*, 215–310.
15. Fréchet, M.R., Généralisations de la loi de probabilité de Laplace. *Annales de l'institut Henri Poincaré* **1951**, *12*, 1–29.
16. Carnot, S. réflexions sur la puissance motrice du feu; published in French in 1824; translation edited by E Mendoza; Dover: New York, NY, USA, 1960.
17. Clausius, R. *On the Mechanical Theory of Heat*; translated by WR Browne 1879; Macmillan: London, UK, 1879.
18. Boltzmann, L. *Vorlesungen über die Gastheorie*; Volume 1; Leipzig: Barth, Germany, 1896.
19. Boltzmann, L. *Wissenschaftliche Abhandlungen*; Volume 3; Leipzig: Barth, Germany, 1909.

20. Massieu, F. Sur les Fonctions caractéristiques des divers fluides. *Comptes Rendus de l'Académie des Sciences* **1869**, *69*, 858–862. (In French)
21. Massieu, F. Addition au précédent Mémoire sur les Fonctions caractéristiques. *Comptes Rendus de l'Académie des Sciences* **1869**, *69*, 1057–1061. (In French)
22. Massieu, F. *Thermodynamique: Mémoire sur les Fonctions Caractéristiques des Divers Fluides et sur la Théorie des Vapeurs*; Académie des Sciences: Paris, France, 1876; p. 92. (In French)
23. Gibbs, J.W. On the equilibrium of heterogeneous substances. *Am. J. Sci.* **1878**, *96*, 441–458.
24. Poincaré, H. Sur les tentatives d'explication mécanique des principes de la thermodynamique. *Comptes rendus de l'Académie des sciences* **1889**, *108*, 550–553.
25. Poincaré, H. Réflexions sur la théorie cinétique des gaz. *J. Phys. Theor. Appl*. **1906**, *5*, 369–403.
26. Poincaré, H. *Thermodynamique, Cours de Physique Mathématique*; G. Carré: Paris, France, 1892.
27. Cosserat, F. *Théorie des corps déformables*; Hermann: Paris, France, 1909.
28. Pommaret, J.F. François Cosserat et le secret de la théorie mathématique de l'élasticité. *Annales de l'Ecole des Ponts et Chaussées* **1997**, *82*, 59–66.
29. Pommaret, J.F. Group interpretation of coupling phenomena. *Acta Mech.* **2001**, *149*, 23–39.
29. Roy, L. Théorie thermodynamique de la ligne élastique et la propagation des ondes. *Annales de la faculté des sciences de Toulouse* **1926**, *18*, 117–195.
30. Duhem, P. Sur les équations générales de la Thermodynamique. *Annales Scientifiques de l'Ecole Normale Supérieure* **1891**, *8*, 231–266. (In French)
31. De Broglie, L. *La Thermodynamique de la particule isolée (ou Thermodynamique cachée des particules)*; Gauthier-Villars: Paris, France, 1964.
32. Souriau, J.M. Structure des systèmes dynamiques; Dunod: Paris, France, 1970.
33. Souriau, J.M. Thermodynamique et géométrie. In *Differential Geometrical Methods in Mathematical Physics II*; Springer: Berlin, Germany, 1978; pp. 369–397.
34. Balian, R.; Alhassid, Y.; Reinhardt, H. Dissipation in many-body systems: A geometric approach based on information theory. *Phys. Rep*. **1986**, *131*, 1–146.
35. Balian, R. *From Microphysics to Macrophysics: Methods and Applications of Statistical Physics*; Springer: Berlin, Germany, 2006.
36. Balian, R. The entropy-based quantum metric. *Entropy* **2014**, *16*, 3878–3888.
37. Balian, R. François Massieu et les potentiels thermodynamiques; Académie des Sciences in Histoire des sciences /Évolution des disciplines et histoire des découvertes: Paris, France, April 2015. Available online: http://www.academie-sciences.fr/en/Evolution-des-disciplines-et-histoire-des-decouvertes/francois-massieu-et-les-potentiels-thermodynamiques.html (accessed on 10 August 2015).
38. Fréchet, M.R. Sur l'extension de certaines évaluations statistiques au cas de petits échantillons. *Rev. Int. Stat. Inst*. **1943**, *11*, 182–205. (Published in IHP Lecture in 1939)
39. Rao, C.R. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc*. **1945**, *37*, 81–89.

40. Chentsov, N.N. *Statistical Decision Rules and Optimal Inferences*; Transactions of Mathematics Monograph, Volume 53; American Mathematical Society: Providence, RI, USA, 1982. (Published in Russian in 1972)
41. Jaynes, E.T. Information Theory and Statistical Mechanics. *Phys. Rev*. **1957**, *106*, 620–630.
42. Jaynes, E.T. Information Theory and Statistical Mechanics II. *Phys. Rev*. **1957**, *108*, 171–190.
43. Jaynes, E.T. Prior Probabilities. *IEEE Trans. Syst. Sci. Cybern*. **1968**, *4*, 227–241.
44. Koszul, J.L. Ouverts convexes homogènes des espaces affines. *Math. Z*. **1962**, *79*, 254–259.
45. Koszul, J.L. Variétés localement plates et convexité. *Osaka J. Math*. **1965**, *2*, 285–290.
46. Cartan, E. La structure des groupes de transformations continus et la théorie du trièdre mobile. *Bull. Sci. Math.* **1910**, *34*, 250–284
47. Cartan, E. Les récentes généralisations de la notion d'espace. *Bull. Sci. Math.* **1924**, *48*, 294–320.
48. Cartan, E. Le rôle de la théorie des groupes de Lie dans l'évolution de la géométrie moderne. *C.R. Congrès Int. Oslo* **1936,** *1*, 92–103. Available online: www.mathunion.org/ICM/ICM1936.1/Main/icm1936.1.0092.0103.ocr.pdf (accessed on 10 August 2015).
49. Gromov, M. In a Search for a Structure, Part 1: On Entropy. July 6, 2012, Available online: http://www.ihes.fr/~gromov/PDF/structre-serch-entropy-july5–2012.pdf (accessed on 6 August 2015).
50. Gromov, M. Six Lectures on Probability, Symmetry, Linearity. Available online: http://www.ihes.fr/~gromov/PDF/probability-huge-Lecture-Nov-2014.pdf (accessed on 6 August 2015).
51. Ollivier, Y. Aspects de l'entropie en mathématiques et en physique (théorie de l'information, systèmes dynamiques, grandes déviations, irréversibilité). Available online: http://www.yann-ollivier.org/entropie/entropie.pdf (accessed on 7 August 2015).
52. Villani, C. (Ir)rréversibilité et entropie. Presented in Séminaire Poincaré XV, December 2010; pp. 17–75. Available online: http://www.bourbaphy.fr/villani.pdf (accessed on 5 August 2015).
53. Darrigol, O. The origins of the entropy concept. Presented in Séminaire Poincaré December 2003; pp.1–12. Available online: http://www.bourbaphy.fr/darrigol.pdf (accessed on 5 August 2015).
54. Deza, M.M.; Deza, E. *Encyclopedia of Distances*, 3rd ed.; Springer: Berlin/Heidelberg, Germany, 2014.

# Chapter 1:
# Origins of Entropy and Information Theory

# J.J. Thomson and Duhem's Lagrangian Approaches to Thermodynamics

**Stefano Bordoni**

**Abstract:** In the last decades of the nineteenth century, different attitudes towards mechanics led to two main theoretical approaches to thermodynamics: an abstract and phenomenological approach, and a very different approach in terms of microscopic models. In reality some intermediate solutions were also put forward. Helmholtz and Planck relied on a mere complementarity between mechanical and thermal variables in the expressions of state functions, and Oettingen explored the possibility of a more demanding symmetry between mechanical and thermal capacities. Planck refused microscopic interpretations of heat, whereas Helmholtz made also recourse to a Lagrangian approach involving fast hidden motions. J.J. Thomson incorporated the two mechanical attitudes in his theoretical framework, and put forward a very general theory for physical and chemical processes. He made use of two sets of Lagrangian coordinates that corresponded to two components of kinetic energy: alongside macroscopic energy, there was a microscopic energy, which was associated with the absolute temperature. Duhem put forward a bold design of unification between physics and chemistry, which was based on the two principles of thermodynamics. From the mathematical point of view, his thermodynamics or *energetics* consisted of a Lagrangian generalization of mechanics that could potentially describe every kind of irreversible process, explosive chemical reactions included.

Reprinted from *Entropy*. Cite as: Bordoni, S. J.J. Thomson and Duhem's Lagrangian Approaches to Thermodynamics. *Entropy* **2014**, *16*, 5876ï 5890.

## 1. Introduction

In the second half of the nineteenth century, the recently emerged thermodynamics underwent a process of mathematisation, and new theoretical frameworks were put forward. Moreover a widespread philosophical and cosmological debate on the second law also emerged. On the specific physical side, two main traditions of research were at stake: the refinement of the kinetic theory of gases, and a questionable alliance between mechanical models and statistical procedures, on the one hand, and the attempt at recasting thermodynamics in accordance with the mathematical structures of Analytical mechanics, on the other. Both research traditions attempted to bridge the gap between the mechanical and thermal domains (Some conceptual aspects of the theoretical pathway leading from Clausius to Duhem are developed in [1,2]. A detailed mathematical account of the emergence of abstract thermodynamics can be found in [3]. For the methodological and philosophical debate that stemmed from the second principle of thermodynamics, see [4]).

James Clerk Maxwell and Ludwig Boltzmann pursued the integration of thermodynamics with the kinetic theory of gases and statistics. At the turn of the twentieth century, the alliance between microscopic mechanical models and probabilistic laws was successfully applied to the field of

electromagnetic radiation [5]. Other scientists relied on a macroscopic and abstract approach in term of continuous variables, setting aside specific mechanical models. The second research tradition was based on the mathematical and physical concept of *potential*, and had its roots in Rudolf Clausius and William Macquorn Rankine's researches in the mid-nineteenth century. Nevertheless the simplified picture of two traditions of research in thermodynamics overshadows the existence of many nuances and different theoretical streams. Different "mechanical theories of heat", and different meanings of the adjective *mechanical* were on stage. In the abstract approach we can find at least three conceptual streams, which corresponded to different attitudes toward mechanics:

(1). a macroscopic and phenomenological approach,
(2). a macroscopic approach based on a structural analogy with abstract mechanics,
(3). a combination of macroscopic and microscopic approaches.

The third stream represented an attempt to bridge the gulf between the two main traditions. It is worth remarking that even Clausius had followed a twofold pathway: a very general mathematical approach to thermodynamics in some memoirs, and an attempt at devising kinetic models of gases in other memoirs. Some scientists contributed to different streams: Max Planck and Arthur von Oettingen contributed to the first and second, Hermann von Helmholtz developed the second and third, and Joseph John Thomson was also at ease along the second and third. Pierre Duhem developed the second stream in an original way: at first he recast thermochemistry, where the second principle of thermodynamics and the concept of free energy were in prominence. Subsequently he attempted to set up a mathematical theory for hysteresis and other irreversible processes. In the meantime he had developed a generalized Lagrangian theory where geometrical, thermal, and other kinds of generalised coordinates were at stake. After some reference to the early developments of the abstract pathway, I will focus on J.J. Thomson and Duhem's Lagrangian approaches. They had different attitudes towards a Lagrangian approach to thermodynamics. J.J. Thomson looked upon Lagrange's equations as a powerful language that could unify microphysics and macrophysics, whereas Duhem refused any reference to microscopic structures. The latter looked upon Lagrange's equations as a model for a more general mathematical framework that could account for a wide set of physical and chemical processes.

## 2. The Second Research Tradition

The first tradition was pursued and refined by Ludwig Boltzmann. He tried to go far beyond Maxwell's microscopic interpretation of equilibrium in rarefied gases: he aimed at clarifying the processes leading to equilibrium. In a long paper he published in 1872 he assumed that molecules were continuously in motion, and those microscopic undetectable motions gave rise to "well-defined laws" at the macroscopic level, which involved the observed average values. A thermodynamic theory required therefore two different levels: a microscopic invisible, and a macroscopic visible one. Statistics and probability could bridge the gap between the two levels. According to Boltzmann, probability did not mean uncertainty: probabilistic laws were ordinary mathematical laws as certain as the other mathematical laws. In 1877, in an even longer paper, he stressed the structural similarity

between his function Ω, representing the probability of a given state, and the entropy $dQ/T$ in any "reversible change of state" [6,7] (Dugas reminded us that Boltzmann's theoretical representation of atoms and molecules evolved over time. In the first volume of his *Vorlesungen über Gastheorie* (1895–1898), we find molecules as "elastic spheres", and then molecules as "centers of force", whereas in the second volume molecules are represented as "mechanical systems characterized by generalized coordinates" [8]).

With regard to the second tradition and its theoretical roots, it is worth remarking that in 1854 Clausius had looked upon the second law of thermodynamics as a law of equivalence between "transformations," in order to maintain a sort of symmetry in the axiomatic structure of thermodynamics. This formulation of the second law, pivoted on the concept of "equivalence value" $dQ/T$, where $T$ was a function of temperature. From the linguistic and conceptual points of view, the two laws of thermodynamics were two principles of the same kind: while the first stated the equivalence between heat and work, the second stated the equivalence between mathematically well-defined "transformation values". In the case of "reversible cyclic processes", the sum or the integral vanished, namely $\oint dQ/T = 0$. A formal analogy between mechanics and thermodynamics was thus established. The sum of the "transformation content" [*Verwandlungsinhalt*] had to vanish in pure, "reversible" thermodynamic processes, as well as the sum of mechanical works along a closed path had to vanish in non-dissipative mechanics. When the processes were irreversible, there was a loss of the transformation content, and the above integral became positive: the initial conditions could not be restored, and the transformation was "uncompensated" [9].

Another formal development was put forward by the Scottish engineer Rankine in 1855. The concept of "*Actual energy*" became a generalization of the mechanical *living force*: it included "heat, light, electric current", and so on. The concept of "*Potential energy*" was extended far beyond gravitation, elasticity, electricity and magnetism. It included "chemical affinity of uncombined elements", and "mutual actions of bodies, and parts of bodies". In general, work was the result of a sum of different terms, where every "variation" of a generalized variable was multiplied "by the corresponding effect" [10]:

$$W = Xdx + Ydy + Zdz + \ldots\ldots \tag{1}$$

In 1869, the mining engineer François Massieu took the path of a mathematical generalization of thermodynamics. After having chosen the volume *v* and the temperature *t* as independent variables, and after some computations, he arrived at a function $\psi$ whose differential was an exact differential of the same variables. Massieu labeled "*characteristic function of the body*" the function $\psi$. The most important mathematical and physical step consisted in deriving "all properties dealing with thermodynamics" from $\psi$ and its derivatives. More specifically, the internal energy $U$ and the entropy $S$ could be expressed in terms of the function $\psi$:

$$U = T^2 \frac{\partial \psi}{\partial t} \text{ and } S = \psi + T\frac{\partial \psi}{\partial t}, \text{ or } S = \frac{\partial}{\partial t}(T\psi) \text{ and } \psi = S - \frac{U}{T} \tag{2}$$

He also introduced a second *characteristic function* $\psi$ n in terms of the two variables *t* and pressure *p*. Besides $U$, $p$, $v$, $Q$ and $S$, even the specific heats at constant pressure or volume, and the

coefficient of dilatation at constant pressure or volume could be derived from $\psi$ and $\psi$ n . According to Massieu, this "mechanical theory of heat" allowed mathematicians and engineer to "settle a link between similar properties of different bodies". Thermodynamics could rely on a consistent set of general and specific laws, and his "characteristic functions" could be looked upon as the mathematical and conceptual link between general and specific laws. In Massieu's theoretical and meta-theoretical context, the adjective "mechanical" did not mean microscopic mechanical models in the sense of Maxwell and Boltzmann, but a mathematical approach on the track of abstract mechanics [11–13].

An abstract approach and wide-scope generalizations were also the hallmarks of Josiah Willard Gibbs's researches on thermodynamics, which he published in the years 1875–1878. The American scientist put forward three "fundamental" thermodynamic functions:

$$\psi = \varepsilon - t\eta\,,\ \chi = \varepsilon + pv\,,\ \zeta = \varepsilon - t\eta + pv \tag{3}$$

The adjective "fundamental" meant that all "thermal, mechanical, and chemical properties" of a physical-chemical system could be derived from them. Under specific conditions, the functions $\psi$ , $\chi$ , and $\zeta$ led to specific conditions of equilibrium [14] (The modern names and symbols for Gibbs's functions $\psi, \chi, \zeta$ are *free energy* $F = U - TS$ , *enthalpy* $H = U + pV$ , and *free enthalpy* or *Gibbs free energy* $G = U - TS + pV$ [15,16]).

In 1880, the young German physicist Max Planck remarked that the theory of elasticity had been put forward without any connection with the thermal properties of bodies, and the thermal actions on them. He aimed at filling the gap between thermodynamics and the theory of elasticity, and outlined a mathematical theory where the mechanics of continuous media merged with thermodynamics. Both mechanical work and heat flow could act on the body: under those actions, both the reciprocal of density [*spezifische Volumen*] and temperature could change from $(v; T)$ to $(v\,\mathrm{n}; T\,\mathrm{n})$ . In particular the geometrical co-ordinates of a point inside the body, and its temperature, underwent a transformation in accordance with the equations

$$x = x_0 + \xi; y = y_0 + \eta; z = z_0 + \zeta \ \text{ and } \ T\,\mathrm{n} = T + \tau \tag{4}$$

where $x_0, y_0, z_0$ and $T$ were the initial values and $\zeta, \eta, \zeta$ , and $\tau$ the infinitesimal variations. Energy depended on $\tau$ and Cauchy's six strain components. Planck showed that energy, entropy, and elastic stresses depended on a combination of mechanical and thermal variables, which were multiplied by a combination of mechanical and thermal coefficients. The two elastic constants could be expressed in terms of those coefficients [17].

After two years, the physicist and physiologist Helmholtz put forward a mathematical theory of heat pivoted on the concept of "free energy". Helmholtz labeled $\vartheta$ the absolute temperature, and $p_\alpha$ the parameters defining the state of the body: they depended neither on each other nor on temperature. If $P_\alpha$ was the external force corresponding to the parameter $p_\alpha$ , and $P_\alpha \cdot dp_\alpha$ the corresponding work, then the total external work was $dW = \sum_\alpha P_\alpha \cdot dp_\alpha$ . Provided that $U$ was the internal energy of the physical system, $S$ its entropy, and $J$ the mechanical equivalent of heat, the function $F = U - J \cdot \vartheta \cdot S$ played the role of a generalized potential for the forces $P_\alpha$ :

$$P_\alpha = -\frac{\partial F}{\partial p_\alpha}. \tag{5}$$

According to Helmholtz, the function $F$ represented the potential energy in the thermodynamic context. The functions $U$ and $S$ could be derived from $F$ by simple derivation. The function $F$ also represented "the free energy", namely the component of the internal energy that could be transformed into every kind of work. If $U$ represented the total internal energy, the difference between $U$ and $F$, namely $J \cdot \vartheta \cdot S$, represented "the bound energy", namely the energy stored in the system as a sort of *entropic* heat [18] (Helmholtz did not seem aware of Massieu's result, which had probably not crossed the France borderlines).

In 1884 Helmholtz attempted to give a microscopic representation of heat, but without any recourse to specific mechanical models. He introduced a global microscopic Lagrangian coordinate, corresponding to a fast, hidden motion, and a set of macroscopic coordinates, corresponding to slow, visible motions. The energy associated with the first coordinate corresponded to thermal energy, whereas the energy associated with the others corresponded to external thermodynamic work [19].

In 1885 Oettingen undertook an even more ambitious design: a formal theory, where mechanical work and heat flows represented the starting point of a dual mathematical structure. The whole body of knowledge of thermodynamics could be based on four "main variables" and two kinds of energy. Temperature and entropy corresponded to "the actual energy [*actuelle Energie*]" $Q$, or in other words the exchanged heat. Volume and pressure corresponded to "the potential energy S", namely the mechanical energy that actually appeared under the form of mechanical work. In brief

$$dQ = t \cdot du\ ,\ dS = -p \cdot dv \tag{6}$$

where $t$ was "the absolute temperature", $u$ "the entropy or Adiabate", $p$ the pressure, and $v$ "the specific volume". He insisted on the physical and linguistic symmetry between thermal and mechanical variables and functions. He put forward a list of "energy coefficients" or "capacities": both "heat capacities [*Wärmecapacitäten*]" and "work capacities [*Arbeitscapacitäten*]" were at stake. In particular, "thermal heat capacities" and "thermal work capacities" [20] corresponded to

$$\left(\frac{dQ}{dt}\right)_v = C_v,\ \left(\frac{dQ}{dt}\right)_p = C_p;\ \left(\frac{dS}{dt}\right)_u = \Phi_u,\ \left(\frac{dS}{dt}\right)_p = \Phi_p. \tag{7}$$

### 3. J.J. Thomson's "Applications of Dynamics"

In 1888 Joseph John Thomson published a book, *Applications of Dynamics to physics and Chemistry*, where he put forward a very general approach to physical and chemical problems. From the outset he remarked that physicists had at their disposal two different methods of establishing "the connection between two different phenomena": a detailed mechanical description of the physical system, or a more general description, "which does not require a detailed knowledge of the mechanism required to produce the phenomena". The second method depended on "the properties of a single function of quantities fixing the state of the system", and had already been "enunciated by M. Massieu and Prof. Willard Gibbs for thermodynamic phenomena". The structure of Lagrange's

equations was suitable for dealing with a set of generalized coordinates $q_i$, and generalized forces $Q_i$; $L = T - V$ was the difference between kinetic and potential energy. Temperature or a distribution of electricity could be interpreted as "coordinates" in a very general sense. Thomson insisted on this opportunity: "any variable quantities" could be considered as coordinates if the corresponding Lagrangian function could be expressed "in terms of them and their first differential coefficients" [21].

He applied the method to those cases "in which we have to consider the effects of temperature upon the properties of bodies": temperature was a measure of "the mean energy due to the translatory motion of the molecules of the gas". In the general structure

$$\frac{d}{dt}\frac{dL}{d\dot{q}_i} - \frac{dL}{dq_i} = Q_i \qquad i = 1, ....., n \tag{8}$$

he introduced kinetic terms of the kind $(1/2)K\,\dot{u}^2$, where $u$ was a Lagrangian coordinate "helping to fix the position or configuration of a molecule". There was "an essential difference" between this kind of coordinates and those "which fix the geometrical, strain, electric, and magnetic configuration of the system". If the latter could be labelled "controllable coordinates" because they were "entirely under our control", the former were much more elusive and "individually" unattainable. Only "the average value of certain functions of a large number of these coordinates" was actually observable or measurable: he labeled them "unconstrainable" coordinates. He could not exclude that the above kinetic terms depended on some "controllable coordinate $\phi$", namely

$$\frac{1}{2}K\,\dot{u}^2 + ..... = \frac{1}{2} f(\phi)\left[(uu)'\,\dot{u}^2 + .....\right] \tag{9}$$

where "the coefficients $(uu)$n do not involve $\phi$". On the contrary, the temperature $\theta$, which was proportional to those kinetic expressions, did not involve "controllable coordinates" [21].

Thomson found convenient to "divide the kinetic energy of a system into two parts": the first part $T_u$ depended on "the motion of unconstrainable coordinates", and was proportional to the absolute temperature $\theta$, whereas the second part $T_c$ depended on the motion of "controllable coordinates". He stressed that $T_c$ corresponded to what Helmholtz had called "die freie Energie [free energy]". He also assumed that the generalized velocities $\dot{u}$ and $\dot{\varphi}$ could not mix, and in particular

$$\frac{dT_u}{d\dot{\phi}} = 0 \tag{10}$$

As already pointed pout, $T_u$ might contain $\phi$, and Lagrange's equations for the coordinates $\phi$ was

$$\Phi = \frac{d}{dt}\frac{dL}{d\dot{\phi}} - \frac{dL}{d\phi} = \frac{d}{dt}\frac{d(T_c + T_u - V)}{d\dot{\phi}} - \frac{d(T_c + T_u - V)}{d\phi} = \frac{d}{dt}\frac{dT_c}{d\dot{\phi}} + \frac{d}{dt}\frac{dT_u}{d\dot{\phi}} - \frac{dT_c}{d\phi} - \frac{dT_u}{d\phi} + \frac{dV}{d\phi} \tag{11}$$

where $\Phi$ was "the external force of this type acting on the system". Taking into account the above mentioned assumptions, the equation could be written [21] as

$$\Phi = \frac{d}{dt}\frac{dT_c}{d\dot{\phi}} - \frac{dT_c}{d\phi} - \frac{dT_u}{d\phi} + \frac{dV}{d\phi} \tag{12}$$

The last equation was the starting point of a mathematical derivation which led to a differential relationship between the invisible kinetic energy $T_u$ and the applied forces $\Phi$, and then between heat fluxes and $\Phi$. In the end, simple relationships between thermal and mechanical effects in elastic bodies could be derived. The first step consisted in computing

$$\begin{aligned}\frac{dT_u}{d\phi} &= \frac{d}{d\phi}\left\{\frac{1}{2} f(\phi)\left[(uu)\,\mathrm{n}\,\dot{u}^2 + .....\right]\right\} = \frac{1}{2} f\,\mathrm{n}(\phi)\left[(uu)\,\mathrm{n}\,\dot{u}^2 + .....\right] = \\ &\frac{1}{2}\frac{f\,\mathrm{n}(\phi)}{f(\phi)} f(\phi)\left[(uu)\,\mathrm{n}\,\dot{u}^2 + .....\right] = \frac{f\,\mathrm{n}(\phi)}{f(\phi)} T_u\end{aligned} \tag{13}$$

As a consequence, Equation (l) became

$$\Phi = \frac{d}{dt}\frac{dT_c}{d\dot{\phi}} - \frac{dT_c}{d\phi} - \frac{f\,\mathrm{n}(\phi)}{f(\phi)} T_u + \frac{dV}{d\phi} \tag{14}$$

When no purely mechanical transformation took place, and only the energy depending on "uncontrollable" coordinates could change, the last equation yielded

$$\frac{d\Phi}{dT_u} = -\frac{f\,\mathrm{n}(\phi)}{f(\phi)} \tag{15}$$

This was the second equation involving the ratio $f\,\mathrm{n}(\phi)/f(\phi)$: the comparison between the two equations gives [21]

$$-\frac{d\Phi}{dT_u} = \frac{1}{T_u}\frac{dT_u}{d\phi} \text{ or } \frac{dT_u}{d\phi} = -T_u\frac{d\Phi}{dT_u} \tag{16}$$

Now a flux of heat $\delta Q$ was called into play, and the conservation of energy required that

$$\delta Q + \sum \Phi \cdot \delta\phi = \delta T_c + \delta T_u + \delta V \tag{17}$$

The term $\delta V$ depended only on $\delta\phi$, and therefore

$$\delta V = \sum \frac{dV}{d\phi}\delta\phi \tag{18}$$

whereas the term $\delta T_c$ required some computations, which led to

$$\delta T_c = \sum\left(\frac{d}{dt}\frac{dT_c}{d\dot{\phi}} - \frac{dT_c}{d\phi}\right)\delta\phi \tag{19}$$

The expression corresponding to the conservation of energy thus became

$$\delta Q = \sum\left(\frac{d}{dt}\frac{dT_c}{d\dot{\phi}} - \frac{dT_c}{d\phi}\right)\delta\phi - \sum \Phi \cdot \delta\phi + \delta T_u + \sum\frac{dV}{d\phi}\delta\phi \tag{20}$$

Equation (l) offered an expression for the generalized forces $\Phi$, which allowed Thomson [21] to simplify the expression for $\delta Q$:

$$\delta Q = \sum\left(\frac{d}{dt}\frac{dT_c}{d\dot{\phi}} - \frac{dT_c}{d\phi}\right)\delta\phi - \sum\left(\frac{d}{dt}\frac{dT_c}{d\dot{\phi}} - \frac{dT_c}{d\phi} - \frac{dT_u}{d\phi} + \frac{dV}{d\phi}\right)\cdot\delta\phi + \delta T_u + \sum\frac{dV}{d\phi}\delta\phi = \\ = \sum\left(\frac{dT_u}{d\phi}\right)\cdot\delta\phi + \delta T_u \tag{21}$$

Now Equation (p) was called into play, and therefore

$$\delta Q = \sum\left(-T_u\frac{d\Phi}{dT_u}\right)_{\phi=const}\cdot\delta\phi + \delta T_u \tag{22}$$

When he took into account isothermal transformations, he assumed that "the quantity of work communicated to the system" was "just sufficient to prevent $T_u$ from changing", where $T_u$ was "proportional to the absolute temperature $\theta$. As a consequence,

$$\delta Q = \sum\left(-T_u\frac{d\Phi}{dT_u}\right)_{\phi=const}\cdot\delta\phi \\ \left(\frac{dQ}{d\phi}\right)_{\theta=const} = \left(-T_u\frac{d\Phi}{dT_u}\right)_{\phi=const} \quad \text{or} \quad \left(\frac{dQ}{d\phi}\right)_{\theta=const} = -\theta\left(\frac{d\Phi}{d\theta}\right)_{\phi=const} \tag{23}$$

Thomson stressed the importance of the last equation, which linked the dependence of heat fluxes on mechanical coordinates to the dependence of external forces on temperature. A deep connection between thermal and mechanical effects was at stake. He made use of this equation in order to tackle "the relations between heat and strain", and in particular the "effects produced by the variation of the coefficients of elasticity $m$ and $n$ with temperature" [21] (In 1845 George Gabriel Stokes had introduced two distinct kinds of elasticity, "one for restoration of volume and one for restoration of shape" [22,23]).

The Greek letters $\alpha, \beta, \gamma$ corresponded to "the components parallel to the axes x, y, z of the displacements of any small portion of the body". Six Latin letters corresponded to longitudinal and transverse strains:

$$e = \frac{d\alpha}{dx}, f = \frac{d\beta}{dy}, g = \frac{d\gamma}{dz},\ a = \frac{d\gamma}{dy} + \frac{d\beta}{dz}, b = \frac{d\alpha}{dz} + \frac{d\gamma}{dx}, c = \frac{d\beta}{dx} + \frac{d\alpha}{dy} \tag{24}$$

He assumed that $\Phi$ corresponded to "a stress of type $e$", and therefore

$$\Phi = m(e+f+g) + n(e-f-g)\ \frac{d\Phi}{d\theta} = \frac{dm}{d\theta}(e+f+g) + \frac{dn}{d\theta}(e-f-g) \tag{25}$$

What had been labeled $\phi$ in Equation (w) corresponded now to the coordinate $e$, and $\delta Q$ corresponded to the amount of heat which had to be supplied to the unit volume of a bar "to keep its temperature from changing when $e$ is increased by $\delta e$" [21]:

$$\frac{dQ}{de} = -\theta\frac{d\Phi}{d\theta} = -\theta\left[\frac{dm}{d\theta}(e+f+g) + \frac{dn}{d\theta}(e-f-g)\right] \text{ or}$$
$$\delta Q = -\left[\frac{dm}{d\theta}(e+f+g) + \frac{dn}{d\theta}(e-f-g)\right]\theta\delta e \tag{26}$$

If the coefficients of elasticity decreased as the temperature increased ( $dm/d\theta < 0$ and $dn/d\theta < 0$) then the equation showed that $\delta Q > 0$: a given amount of heat had to be supplied in order "to keep the temperature of a bar constant when it is lengthened". In other words, "a bar will cool when it is extended", if no heat is supplied from outside.

In the case of twist, $\Phi$ represented "a couple tending to twist the bar about the axis of x", and $a$ was the corresponding twist:

$$\Phi = na\,,\ \frac{d\Phi}{d\theta} = \frac{dn}{d\theta}a \tag{27}$$

The amount of heat that assured the temperature to be preserved was

$$\delta Q = -\frac{dn}{d\theta}\theta\,\delta a \tag{28}$$

The physical interpretation was not different from the previous one: when a rod is twisted, "it will cool if left to itself", provided that "the coefficient of rigidity diminishes as the temperature increases", which is what usually happens (Thomson reminded the readers that William Thomson had first obtained those results "by means of the Second Law of thermodynamics" [21]).

**4. Duhem's "General Equations"**

In 1891, Pierre Duhem began to outline a systematic design of mathematisation and generalization of thermodynamics. He took into account a system whose elements had the same temperature: the state of the system could be completely specified by its temperature $\vartheta$ and $n$ independent coordinates α, β, …, λ. He then introduced some "external forces", which depended on α, β, …, λ and $\vartheta$, and held the system in equilibrium. At the thermodynamic equilibrium, a series of equations of the kind

$$\frac{\partial A}{\partial \beta} - \frac{\partial B}{\partial \alpha} = 0 \tag{29}$$

could be derived. The equations suggested that "a uniform, finite, and continuous function $F(\alpha,\beta,...,\lambda,\vartheta)$ of $n$ + 1 coordinates α, β, …, λ, and $\vartheta$ does exist". In other words, apart from $\Theta$, which was "independent of the function $F$", generalized forces could be written as the components of $F$ gradient:

$$A = \frac{\partial}{\partial\alpha}F(\alpha,\beta,...,\lambda,\vartheta),\ B = \frac{\partial}{\partial\beta}F(\alpha,\beta,...,\lambda,\vartheta),\ \ldots\ L = \frac{\partial}{\partial\lambda}F(\alpha,\beta,...,\lambda,\vartheta) \tag{30}$$

The function $F$ was nothing else but Helmholtz's free energy of Gibbs' first potential [24].

In 1892 Duhem put forward Lagrange's equations for a physical system at the thermodynamic equilibrium. When $dQ = 0$,

$$\frac{d}{dt}\frac{\partial T}{\partial \alpha'}-\frac{\partial T}{\partial \alpha}+E\frac{\partial U}{\partial \alpha}=A,\quad ...,\quad ...,\quad \frac{d}{dt}\frac{\partial T}{\partial \lambda'}-\frac{\partial T}{\partial \lambda}+E\frac{\partial U}{\partial \lambda}=L \tag{31}$$

where $T$ was the kinetic energy, $U$ the internal energy, and $E$ the mechanical equivalent of heat. In 1894 he generalized the equations, and introduced a perturbation, which represented a source of irreversibility for the physical system:

$$\frac{d}{dt}\frac{\partial T}{\partial \alpha'}-\frac{\partial T}{\partial \alpha}+\frac{\partial F}{\partial \alpha}=A'+f_\alpha,\quad ...,\quad ...,\quad \frac{d}{dt}\frac{\partial T}{\partial \lambda'}-\frac{\partial T}{\partial \lambda}+\frac{\partial F}{\partial \lambda}=L'+f_\lambda \tag{32}$$

The new functions $f_\alpha, f_\beta, ..., f_\lambda$ represented "passive resistances to be overcome by the system", and depended on the coordinates α, β, …, λ, $\vartheta$, their time derivatives $\alpha', \beta', ..., \lambda'$, and time *t*. Equilibrium was *perturbed* by physical or chemical actions that represented the generalization of mechanical *viscosity* [25,26].

In the meantime Duhem was committed to updating thermochemistry. In 1893 he focused on experiments performed at high temperatures, and in particular the phenomenon of "false equilibrium". Thermodynamics forbade some transformations, and they did not really happen, but sometimes even permitted transformations did not take place. Duhem qualified the first case as "true equilibrium", and the latter as "false equilibrium". The concept of "false" equilibrium allowed Duhem to interpret chemical reactions that were associated with "a powerful release of heat" or explosions. When mixtures of hydrogen and oxygen, or hydrogen and chlorine, reached their "true" equilibrium, namely water and muriatic acid, they released such a great amount of heat as to trigger off an explosion. In Duhem's theoretical framework, an explosion was therefore a passage "from a state of false equilibrium to a state of true equilibrium", where "a remarkable amount of heat" was released [27].

From 1894 onwards he published a series of papers dealing with mechanical and magnetic hysteresis, and other kinds of physical and chemical irreversible transformations. He started from a simplified physical system defined by a temperature *T* and a single "normal variable *x*", and applied to it "the classic propositions of thermodynamics". The condition of equilibrium under an external force *X* was $X = \partial \mathcal{F}(x,T)/\partial x$. If the differentiation of the external force required in general that

$$dX=\frac{\partial^2 F(x,T)}{\partial x^2}dx+\frac{\partial^2 F(x,T)}{\partial x\,\partial T}dT \tag{33}$$

a more general expression

$$dX=\frac{\partial^2 F(x,T)}{\partial x^2}dx+\frac{\partial^2 F(x,T)}{\partial x\,\partial T}dT+f(x,T,X)\cdot|\,dx\,| \tag{34}$$

was required in order to describe the presence of permanent deformations. The function $f(x,T,X)$ was an unspecified "*uniform* and *continuous* function of the three variables $x, T, X$". It was the existence of a term depending on $|\,dx\,|$ that assured that "*a continuous series of states of equilibrium of the system is not, in general, a reversible transformation*". The mathematical model became sensitive to the direction of transformations. At that stage, Duhem confined himself to isothermal

transformations, for he was interested mainly in mechanical deformations. The simplified equation yielded [28]

$$dX = \frac{\partial^2 F(x,T)}{\partial x^2} dx + f(x,T,X) \cdot |\, dx \,| \tag{35}$$

He assumed the existence of a new kind of closed cycle, a cycle of hysteresis, which was the fundamental entity of the new thermodynamics of permanent, irreversible transformations. When a force *dX* was applied to the physical system, and then applied in the opposite direction, the sum of forces vanished, but the sum of the corresponding strains $dx_1$ and $dx_2$ did not. According to the simplified equation,

$$0 = dX - dX = \frac{\partial^2 F(x,T)}{\partial x^2} \sum_{k=1}^{2} dx_k + f(x,T,X) \sum_{k=1}^{2} |\, dx_k \,| \text{ or } \sum_{k=1}^{2} dx_k = -\frac{f(x,T,X)}{\dfrac{\partial^2 F(x,T)}{\partial x^2}} \sum_{k=1}^{2} |\, dx_k \,| \tag{36}$$

The physical system did not return to its initial conditions: it experienced an irreversible strain. Duhem made use of the non-simplified equation in order to describe simple mechanical systems: "a homogeneous cylinder submitted to a traction", or "torsion", or "flexion". Other kinds of permanent deformations corresponded to processes like quenching. If traction, torsion and flexion represented the mechanical side, quenching represented the thermal side of Duhem's theory of permanent deformations [28].

In 1896, he put forward a further generalization of his Lagrangian equations, which relied on the structural analogy between chemical "false" equilibrium and mechanical "friction". From the mathematical point of view, the condition of unstable equilibrium that preceded an explosive chemical reaction was not so different from the equilibrium experienced by a body at rest on a rough inclined plane when the tilt angle was slowly increased. Only after having crossed a critical value of the inclination, the body suddenly slid down. The new equations involved a set of functions $g_a, g_b, \ldots\ldots, g_l$, and terms of the kind $g_a \cdot a' / |a'|$ that represented the generalization of static friction:

$$\frac{d}{dt}\frac{\partial T}{\partial \alpha'} - \frac{\partial T}{\partial \alpha} + \frac{\partial F}{\partial \alpha} = A' + f_\alpha + g_\alpha \frac{\alpha'}{|\alpha'|}, \quad \ldots, \quad \ldots, \quad \frac{d}{dt}\frac{\partial T}{\partial \lambda'} - \frac{\partial T}{\partial \lambda} + \frac{\partial F}{\partial \lambda} = L' + f_\lambda + g_\lambda \frac{\lambda'}{|\lambda'|} \tag{37}$$

The generalized frictional terms depended on generalized coordinates, velocities, and forces. Differently from the "viscous" forces, the new terms did not vanish when the velocities vanished: on the contrary, they tended to the limiting functions $\gamma_\alpha, \gamma_\beta, \ldots\ldots, \gamma_\lambda$, which depended only on coordinates and forces. In this case, every equation gave rise to two different sets of forces that corresponded to two thresholds for the physical-chemical system [29]:

$$A' \pm \gamma_\alpha, \quad \ldots, \quad \ldots, \quad L' \pm \gamma_\lambda \tag{38}$$

Duhem set up a general and pliable mathematical structure that could be further widened in order to account for phenomena of increasing complexity. When he took into account chemical false equilibrium and explosions, he dropped the traditional "inertial" Lagrangian terms. After having widened the scope and the mathematical structure of traditional mechanics, he disregarded the original component of that structure, and focused on the complementary terms, which corresponded

to a sort of complementary mechanics. It was a chemical mechanics or a new kind of mechanics suitable for chemical reactions. The thermodynamic potential $H = F + PV$ (Duhem's potential $H$ corresponded to Massieu's potential $\varphi'$ and Gibb's potential $\zeta$ ) was the suitable potential for physical-chemical processes taking place at constant pressure, and the general equations were reduced to a mathematical structure [29] of the kind

$$\frac{\partial H(P,\alpha,T)}{\partial \alpha} - f(P,\alpha,T,\alpha') - g(P,\alpha,T,\alpha')\frac{\alpha'}{|\alpha'|} = 0 \tag{39}$$

Duhem had added dissipative terms to Lagrange's equations in order to generalize analytical mechanics. In the new mathematical structure, no inertial terms appeared, while dissipative terms were in prominence: traditional Analytical mechanics and Chemistry represented two opposite poles in the new formal framework.

The equation described a chemical mixture: the three coordinates represented the degree of combination α, "a uniform and constant pressure $P$ ", and "a variable temperature $T$ ". The time derivative α represented "the velocity of transformation of the system", or in other words, the velocity of the chemical reaction. Some approximations allowed Duhem to derive that velocity, which was in some way the solution of the mathematical procedure. He assumed that $g(P,\alpha,T,\alpha')$ did not depend on α,

$$g(P,\alpha,T,\alpha \mathrm{n}) \approx \gamma(P,\alpha,T) \tag{40}$$

and $f(P,\alpha,T,\alpha')$ was a linear function of $\alpha'$:

$$f(P,\alpha,T,\alpha') \approx \varphi(P,\alpha,T)\cdot\alpha' \tag{41}$$

The simplified *equation of motion*

$$\frac{\partial H(P,\alpha,T)}{\partial \alpha} - \varphi(P,\alpha,T)\cdot\alpha' \pm \gamma(P,\alpha,T) = 0 \tag{42}$$

yielded the "velocity" of reaction [29]

$$\alpha' = \frac{\dfrac{\partial H(P,\alpha,T)}{\partial \alpha} \pm \gamma(P,\alpha,T)}{\varphi(P,\alpha,T)} \tag{43}$$

Duhem's complementary or chemical mechanics led to results that were paradoxical from the point of view of traditional mechanics but consistent with explosive chemical reactions. When the viscous term vanished, the velocity of reaction became infinite. Pure mechanics and chemical reactions represented the opposite poles in Duhem's generalized mechanics, which could encompass physics and chemistry in a very general mathematical structure.

## 5. Concluding Remarks

In the context of an abstract approach to thermodynamics, late nineteenth-century Lagrangian theories represented one of the most interesting theoretical streams. J.J. Thomson put forward a bold mathematical framework that could host microscopic motions, macroscopic stresses, and

macroscopic heat fluxes. Duhem put forward an even bolder mathematical framework where traditional Lagrangian terms stood alongside dissipative terms that could account for irreversible processes. The concept of motion underwent a deep transformation: it corresponded to any variation of a Lagrangian coordinate. It does not seem that the two authors were influenced by one another. Duhem put forward the first historical reconstruction of the emergence of an abstract approach to thermodynamics. In general he acknowledged the scientific contributions of other scholars: he explicitly mentioned Massieu, Gibbs, Helmholtz, and Oettingen, but not J.J. Thomson. This is a weak clue about the non-influence of Thomson on Duhem, but stronger evidence is given by the fact that Duhem sharply opposed any microscopic approach. It is definitely more evident that Duhem could not influence Thomson because Duhem's systematic research programme was put forward after 1888.

Today we know that J.J. Thomson's approach did not leave disciples whereas Duhem is acknowledged as the creator of modern phenomenological thermodynamics or the theory of continuous media based on thermodynamics (For the role played by Duhem in the emergence of twentieth-century thermodynamics of nolinear irreversible processes, see [30]. He was the first scholar to put forward a general thermodynamic framework for widespread dissipative processes such as hysteresis and explosions.

**Acknowledgments**

I would like to thank Frédéric Barbaresco, Ali Mohammad-Djafari, Olivier Darrigol, the scholars of *Max-Planck-Institut für Wissenschaftsgeschichte* (Berlin), and the scholars of *Department of Basic Sciences and Foundations* (Urbino, Italy) for the opportunity of discussing some parts of the present paper in formal and informal talks. I also thank the anonymous referees for helpful criticism and remarks.

**Conflicts of Interest**

The author declares no conflict of interest.

**References**

1. Bordoni, S. Unearthing a Buried Memory: Duhem's Third Way to thermodynamics. Part 1. *Centaurus* **2012**, *54*, 124–147.
2. Bordoni, S. Unearthing a Buried Memory: Duhem's Third Way to thermodynamics. Part 2. *Centaurus* **2012**, *54*, 232–249.
3. Bordoni, S. Routes towards an Abstract Thermodynamics in the late nineteenth century. *Eur. Phys. J. H* **2013**, *38*, 617–660.
4. Kragh, H. *Entropic Creation—Religious Context of Thermodynamics and Cosmology*; Ashgate: Farnham, UK, 2008.
5. Darrigol, O.; Renn, J. *La Nascita Della Meccanica Statistica, in Storia Della Scienza*; Istituto della Enciclopedia Italiana: Roma, Italy, 2003; Chapter XLV, pp. 496–507.

6. Boltzmann, L. Weiteren Studien über das Wärmegleichgewicht unter Gasmolekülen; in Boltzmann, L., *Wissenschaftlichen Abhandlungen*: Barth: Leipzig, Germany, 1909; Volume I, pp. 317–402.
7. Boltzmann, L. Über die Beziehung zwischen dem zweiten Hauptsatze der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung respektive den Sätzen über das Wärmegleichgewicht. In Boltzmann, L. *Wissenschaftlichen Abhandlungen*; Barth: Leipzig, Germany, 1909; Volume II, pp. 164–223.
8. Dugas, R. *La théorie Physique au Sens de Boltzmann*; Éditions du Griffon: Neuchatel, Switzerland, 1959.
9. Clausius, R. Ueber eine veränderte Form des zweiten Hauptsatzes der mechanischen Wärmetheorie. In *Abhandlungen Über Die Mechanische Wärmelehre*; Clausius, R.; Friedrich Vieweg und Sohn: Braunschweig, Germany, 1864.
10. Rankine, W.J.M. Outlines of the Science of Energetics. Rankine, W.J.M. In *Miscellaneous Scientific Papers*; Charles Griffin and Company: London, UK, 1881; pp. 209–228.
11. Massieu, F. Sur les Fonctions caractéristiques des divers fluides. *Comptes Rendus Acad. Sci.* **1869**, *LXIX*, 858–862.
12. Massieu, F. Addition au précédent Mémoire sur les Fonctions caractéristiques. *Comptes Rendus Acad. Sci.* **1869**, *LXIX*, 1057–1061.
13. Massieu, F. Mémoire sur les Fonctions Caractéristiques des Divers Fluides et sur la Théorie des Vapeurs. *Mémoires présentés par divers savants à l'Académie des Sciences de l'Institut National de France* **1876**, *XXII*, 1–92.
14. Gibbs, J.W. On the Equilibrium of Heterogeneous Substances. Gibbs, J.W.; In *The Scientific Papers of J. Willard Gibbs*; Longmans, Green, and Co.: London, New York, and Bombay, 1906; pp. 55–349.
15. Kragh, H.; Weininger, S.J. Sooner silence than confusion: The tortuous entry of entropy into chemistry. *Hist. Stud. Phys. Biol. Sci.* **1996**, *27*, 91–130.
16. Müller, I. A History of Thermo-Dynamics. In *The Doctrine of Energy and Entropy*; Dover: New York, NY, USA, 2007.
17. Planck, M. *Gleichgewichtszustände Isotroper Körper in Verschiedenen Temperaturen*; Theodore Ackermann: München, Germany, 1880.
18. Helmholtz, H. Die Thermodynamik Chemischer Vorgänge. Helmholtz, H.; In *Wissenschaftliche Abhandlungen*; Barth: Leipzig, Germany, 1883; Volume II, pp. 958–978.
19. Bierhalter, G. Helmholtz's Mechanical Foundation of thermodynamics. In *Hermann von Helmholtz and the Foundations of Nineteenth-Century*; Cahan, D., Ed.; University of California Press: Berkeley, CA, USA, 1993; pp. 291–333.
20. Oettingen, A. Die thermodynamischen Beziehungen antithetisch entwickelt. *Mémoires de l'Académie impériale des Sciences de Saint-Pétersbourg* **1885**, *XXXII*, 1–70.
21. Thomson, J.J. *Applications of Dynamics to Physics and Chemistry*; Macmillan and Co.: London, UK; New York, NY, USA, 1888.

22. Darrigol, O. Between Hydrodynamics and Elasticity Theory: The First Five Births of the Navier-Stokes Equation. *Arch. History Exact Sci.* **2002**, *56*, 95–150.
23. Stokes, G.G. *Mathematical and Physical Papers Volume II*; Cambridge University Press: Cambridge, UK, 1883.
24. Duhem, P. Sur les équations générales de la Thermodynamique. *Annal. Sci. Ecole Normale Supérieure* **1891**, *3e série, Tome VIII*, 231–266.
25. Duhem, P. Commentaire aux principes de la Thermodynamique—Première partie. *J. Math. Pures Appl.* **1892**, *4e série, Tome VIII*, 269–330.
26. Duhem, P. Commentaire aux principes de la Thermodynamique—Troisième partie. *J. Math. Pures Appl.* **1894**, *4e série, Tome X*, 207–286.
27. Duhem, P. *Introduction à la Mécanique Chimique*; Carré: Paris, France, 1893.
28. Duhem, P. Sur les Déformations Permanentes et l'hystérésis. In Duhem, P. *Sur les Déformations Permanentes et l'hystérésis*; Hayez, Imprimeur de l'Académie Royale de Belgique: Bruxelles, Belgium, 1896.
29. Duhem, P. *Théorie thermodynamique de la viscosité, du frottement et des faux équilibres chimiques*; Hermann: Paris, France, 1896.
30. Maugin, G.A. *The Thermomechanics of Nonlinear Irreversible Behaviors: An Introduction*; World Scientific: Singapore, Singapore, 1999.

# On Shannon's Formula and Hartley's Rule: Beyond the Mathematical Coincidence

**Olivier Rioul and José Carlos Magossi**

**Abstract:** In the information theory community, the following "historical" statements are generally well accepted: (1) Hartley did put forth his rule twenty years before Shannon; (2) Shannon's formula as a fundamental tradeoff between transmission rate, bandwidth, and signal-to-noise ratio came out unexpected in 1948; (3) Hartley's rule is inexact while Shannon's formula is characteristic of the additive white Gaussian noise channel; (4) Hartley's rule is an imprecise relation that is not an appropriate formula for the capacity of a communication channel. We show that all these four statements are somewhat wrong. In fact, a careful calculation shows that "Hartley's rule" in fact coincides with Shannon's formula. We explain this mathematical coincidence by deriving the necessary and sufficient conditions on an additive noise channel such that its capacity is given by Shannon's formula and construct a sequence of such channels that makes the link between the uniform (Hartley) and Gaussian (Shannon) channels.

Reprinted from *Entropy*. Cite as: Rioul, O.; Magossi, J.C. On Shannon's Formula and Hartley's Rule: Beyond the Mathematical Coincidence. *Entropy* **2014**, *16*, 4892–4910.

## 1. Introduction

As researchers in information theory, we all know that the milestone event that founded our field is Shannon's publication of his seminal 1948 paper [1] that created a completely new branch of applied mathematics and called it to immediate worldwide attention. What has rapidly become the emblematic classical expression of the theory is *Shannon's formula* [1,2]

$$C = \tfrac{1}{2}\log_2\Big(1+\frac{P}{N}\Big) \tag{1}$$

for the information capacity of a communication channel with signal-to-noise ratio $P/N$.

Hereafter we shall always express information capacity in binary units (bits) per *sample*. Shannon's well-known original formulation was in bits per second:

$$C = W\log_2\Big(1+\frac{P}{N}\Big) \qquad \text{bits/s.}$$

The difference between this formula and (1) is essentially the content of the *sampling theorem*, often referred to as Shannon's theorem, that the number of independent samples that can be put through a channel of bandwidth $W$ hertz is $2W$ samples per second. We shall not discuss here whether the sampling theorem should be attributed to Shannon or to other authors that predate him in this discovery; see e.g., [3] for a recent account and extensive study on this subject.

The classical derivation of (1) was done in [1] as an application of Shannon's coding theorem for a memoryless channel, which states that the best coding procedure for reliable transmission

achieves a maximal rate of $C = \max_X I(X;Y)$ bits per sample, where $X$ is the channel input with average power $P = \mathbb{E}(X^2)$ and $Y = X + Z$ is the channel output. Here $Z$ denotes the additive Gaussian random variable (independent of $X$) that models the communication noise with power $N = \mathbb{E}(Z^2)$. By expanding mutual information $I(X;Y) = h(Y) - h(Y|X)$ as a difference of differential entropies, noting that $h(Y|X) = h(Z) = \log_2 \sqrt{2\pi eN}$ is constant, and choosing $X$ Gaussian so as to maximize $h(Y)$, Shannon arrived at his formula $C = \max_X h(Y) - h(Z) = \log_2 \sqrt{2\pi e(P+N)} - \log_2 \sqrt{2\pi eN} = \frac{1}{2}\log_2(1 + P/N)$.

Formula (1) is also known as the *Shannon–Hartley formula*, and the channel coding theorem stating that (1) is the maximum rate at which information can be transmitted reliably over a noisy communication channel is often referred to as the *Shannon–Hartley theorem* (see, e.g., [4]). The reason for which Hartley's name is associated to the theorem is commonly justified by the so-called *Hartley's law*, which is described as follows:

> *During 1928, Hartley formulated a way to quantify information and its line rate (also known as data signalling rate R bits per second) [5]. This method, later known as Hartley's law, became an important precursor for Shannon's more sophisticated notion of channel capacity. (...)*
>
> *Hartley argued that the maximum number of distinguishable pulse levels that can be transmitted and received reliably over a communications channel is limited by the dynamic range of the signal amplitude and the precision with which the receiver can distinguish amplitude levels. Specifically, if the amplitude of the transmitted signal is restricted to the range of* $[-A, +A]$ *volts, and the precision of the receiver is* $\pm\Delta$ *volts, then the maximum number of distinct pulses* $M$ *is given by* $M = 1 + \frac{A}{\Delta}$*. By taking information per pulse in bit/pulse to be the base-2-logarithm of the number of distinct messages* $M$ *that could be sent, Hartley [5] constructed a measure of the line rate* $R$ *as* $R = \log_2(M)$ *[bits per symbol].*
>
> —Wikipedia [4]

In other words, within a noise amplitude limited by $\Delta$, by taking regularly spaced input symbol values in the range $[-A, A]$ with step $2\Delta$:

$$-A, -A + 2\Delta, \ldots, A - 2\Delta, A,$$

one can achieve a maximum total number of $M = A/\Delta + 1$ possible distinguishable values. This holds in the most favorable case where $A/\Delta$ is an integer, where the "+1" is due to the sample values at the boundaries—otherwise, $M$ would be the integer part of $A/\Delta + 1$. Therefore, error-free communication is achieved with at most

$$C' = \log_2\Big(1 + \frac{A}{\Delta}\Big) \tag{2}$$

bits per sample. This equation strikingly resembles (1). Of course, the "signal-to-noise ratio" $A/\Delta$ is a ratio of amplitudes, not of powers, hence should not be confused with the usual definition $P/N$;

accordingly, the factor $1/2$ in Formula (1) is missing in (2). Also, (2) is only considered as an approximation of (1):

> *Hartley's rate result can be viewed as the capacity of an errorless $M$-ary channel (...). But such an errorless channel is an idealization, and if M is chosen small enough to make the noisy channel nearly errorless, the result is necessarily less than the Shannon capacity of the noisy channel (...), which is the Hartley–Shannon result that followed later [in 1948].*
>
> —Wikipedia [4]

In the information theory community, the following "historical" statements are generally well accepted:

(1) *Hartley did put forth his rule* (2) *twenty years before Shannon.*

(2) *The fundamental tradeoff* (1) *between transmission rate, bandwidth, and signal-to-noise ratio came out unexpected in 1948: the time was not even ripe for this breakthrough.*

(3) *Hartley's rule is inexact while Shannon's formula is characteristic of the additive white Gaussian noise (AWGN) channel ($C' \neq C$).*

(4) *Hartley's rule is an imprecise relation between signal magnitude, receiver accuracy and transmission rate that is not an appropriate formula for the capacity of a communication channel.*

In this article, we show that all these four statements are somewhat wrong. The organisation is as follows. Sections 2–5 will each defend the opposite view of statements (1)–(4) correspondingly. Section 6 concludes through a detailed mathematical analysis.

## 2. Hartley's Rule is not Hartley's

Hartley [5] was the first researcher to try to formulate a theory of the transmission of information. Apart from stating explicitly that the amount of transmitted information is proportional to the transmission bandwidth, he showed that the number $M$ of possible alternatives from a message source over given a time interval grows exponentially with the duration, suggesting a definition of information as the logarithm $\log M$. However, as Shannon recalled in 1984:

> *I started with information theory, inspired by Hartley's paper, which was a good paper, but it did not take account of things like noise and best encoding and probabilistic aspects.*
>
> —Claude Elwood Shannon [6]

Indeed, no mention of signal vs. noise, or of amplitude limitation $A$ or $\Delta$ was ever made in Hartley's paper [5]. One may then wonder how (2) was coined as Hartley's law.

The oldest reference that attributes (2) to Hartley—and incidentally cited in the Wikipedia page [4]—seems to be the classical 1965 textbook of Wozencraft and Jacobs, most notably its introduction chapter:

*(...) in 1928, Hartley [5] reasoned that Nyquist's result, when coupled with a limitation on the accuracy of signal reception, implied a restriction on the amount of data that can be communicated reliably over a physical channel. Hartley's argument may be summarized as follows. If we assume that (1) the amplitude of a transmitted pulse is confined to the voltage range* $[-A, A]$ *and (2) the receiver can estimate a transmitted amplitude reliably only to an accuracy of* $\pm\Delta$ *volts, then, as illustrated in [the] Figure (...), the maximum number of pulse amplitudes distinguishable at the receiver is* $(1 + A/\Delta)$*. (...)*

*[in* the Figure's legend:*] Hartley considered received pulse amplitudes to be distinguishable only if they lie in different zones of width* $2\Delta$ *(...)*

*Hartley's formulation exhibits a simple but somewhat inexact interrelation among (...) the maximum signal magnitude* $A$*, the receiver accuracy* $\Delta$*, and the allowable number of message alternatives. Communication theory is intimately concerned with the determination of more precise interrelations of this sort.*

—John M. Wozencraft; Irwin Mark Jacobs [7]

The textbook was highly regarded and still widely used today. Its introductive text has become famous to many researchers in the field of communication theory and has had a tremendous impact. This would explain why (2) is now widely known as Hartley's capacity law.

One may then wonder whether Wozencraft and Jacobs have found such a result themselves while attributing it to Hartley or whether it was inspired from other researchers. We found that the answer is probably in very first tutorial article in information theory that was ever published by E. Colin Cherry in 1951:

*Although not explicitly stated in this form in his paper, Hartley [5] has implied that the quantity of information which can be transmitted in a frequency band of width* $B$ *and time* $T$ *is proportional to the product:* $2BT \log M$*, where* $M$ *is the number of "distinguishable amplitude levels." [...] He approximates the waveform by a series of steps, each one representing a selection of an amplitude level. [...] For example, consider a waveform to be traced out on a rectangular grid [...], the horizontal mesh-width representing units of time (equal to* $1/2B$ *in order to give the necessary* $2BT$ *data in a time* $T$*), and the vertical the "smallest distinguishable" amplitude change; in practice this smallest step may be taken to equal the noise level* $n$*. Then the quantity of information transmitted may be shown to be proportional to* $BT \log(1 + a/n)$ *where* $a$ *is the maximum signal amplitude, an expression given by Tuller [8], being based upon Hartley's definition of information.*

—E. Colin Cherry [9]

Cherry attributes (2) to an *implicit* derivation of Hartley but cites the explicit derivation of Tuller [8]. The next section investigates the contribution of Tuller and others.

## 3. Independent 1948 Derivations of Shannon's Formula

In the introduction to his classic textbook, Robert McEliece wrote:

> *With many profound scientific discoveries (for example Einstein's discovery in 1905 of the special theory of relativity) it is possible with the aid of hindsight to see that the times were ripe for a breakthrough. Not so with information theory. While of course Shannon was not working in the vacuum in the 1940's, his results were so breathtakingly original that even the communication specialists of the day were at a loss to understand their significance.*
>
> —Robert McEliece [10]

One can hardly disagree with this statement when one sees the power and generality of Shannon's results. Just to mention a few examples: the introduction of the formal architecture of communication systems (Shannon's paradigm) with explicit distinction between source, channel and destination; the emphasis on digital representation to make the chance of error as small as desired; the consideration of codes in high dimensions; and the use of probabilistic models for the signal as well as for the noise, *via* information theoretic tools like entropy and mutual information. Shannon's ideas were revolutionary, in keeping with J.R. Pierce's famous quote:

> *In the end, [1] and the book based on it came as a bomb, and something of a delayed-action bomb.*
>
> —John R. Pierce [11]

Indeed, [1] being so deep and profound, did not have an immediate impact. As Robert Gallager recalls:

> *(...) two important papers (...) were almost concurrent to [1].*
>
> *The first subsequent paper was [12], whose coauthors were B. R. Oliver and J. R. Pierce. This is a very simple paper compared to [1], but it had a tremendous impact by clarifying a major advantage of digital communication. (...) It is probable that this paper had a greater impact on actual communication practice at the time than [1].*
>
> *The second major paper written at about the same time as [1] is [2]. This is a more tutorial amplification of the AWGN channel results of [1]. (...) This was the paper that introduced many communication researchers to the ideas of information theory.*
>
> —Robert Gallager [13]

In [12], Shannon's Formula (1) was used without explicit reference to the Gaussian nature of the added white noise, as the capacity of an "ideal system". On the other hand, [2] was devoted to a geometric proof of Formula (1).

It appears, therefore, that Shannon's Formula (1) was *the* emblematic result that impacted communication specialists at the time, as expressing the correct tradeoff between transmission

rate, bandwidth, and signal-to-noise ratio. It is one Shannon's result that is the best known and understood among communications engineers. As Verdú has noticed in [14], only a few months after the publication of [2], M. Golay [15] referred to (1) as "the now classical expression for the information reception capacity of a channel." In the following years, finding "codes to reach the promised land (1)" [16] became the "holy grail of information theory" [14].

As far as (1) is concerned, Shannon, after the completion of [1], acknowledged other works:

> *Formulas similar to* (1) *for the white noise case have been developed independently by several other writers, although with somewhat different interpretations. We may mention the work of N. Wiener [17], W. G. Tuller [8], and H. Sullivan in this connection.*
>
> —Claude Elwood Shannon [1]

Unfortunately, Shannon gave no specific reference to H. Sullivan. S. Verdú cited many more contributions during the same year of 1948:

> *By 1948 the need for a theory of communication encompassing the fundamental tradeoffs of transmission rate, reliability, bandwidth, and signal-to-noise ratio was recognized by various researchers. Several theories and principles were put forth in the space of a few months by A. Clavier [18], C. Earp [19], S. Goldman [20], J. Laplume [21], C. Shannon [1], W. Tuller [8], and N. Wiener [17]. One of those theories would prove to be everlasting.*
>
> —Sergio Verdú [14]

Lundheim reviewed some of these independent discoveries and concludes:

> *(...) this result [Shannon's formula] was discovered independently by several researchers, and serves as an illustration of a scientific concept whose time had come.*
>
> —Lars Lundheim [22]

This can be contrasted to the above citation of R. McEliece.

Wiener's independent derivation [17] of Shannon's formula is certainly the one that is closest to Shannon's. He also used probabilistic arguments, logarithmic measures (in base 2) and differential entropy, the latter choice being done "mak[ing] use of a personal communication of J. von Neumann". Wiener considers "the information gained by fixing one or more variables in a problem", e.g., fixing $Y = X + Z$ where $X$ and $Z$ are independent Gaussian. By computing the difference $h(X) - h(X|Y)$, he concludes that "the excess of information concerning $X$ when we know $Y$ is" (1). Unlike Shannon, however, his definition of information is not based on any precise communication problem. There is also no relation to Hartley's argument leading to (2).

Concerning the idea of information theory, Wiener wrote in his book Cybernetics:

> *This idea occurred at about the same time to several writers, among them the statistician R. A. Fisher, Dr. Shannon of the Bell Telephone Laboratories, and the author. Fisher's motive in studying this subject is to be found in classical statistical theory; that of Shannon in the problem of coding information; and that of the author in the problem of noise and message in electrical filters. Let it be remarked parenthetically that some of my speculations in this direction attach themselves to the earlier work of Kolmogoroff in Russia, although a considerable part of my work was done before my attention was called to the work of the Russian school.*
>
> —Norbert Wiener[17]

It is likely that it is the importance of Shannon's formula for which he has made an independent derivation that lead him to declare:

> *Information theory has been identified in the public mind to denote the theory of information by bits, as developed by C. E. Shannon and myself.*
>
> —Norbert Wiener[23]

J.R. Pierce comments:

> *Wiener's head was full of his own work and an independent derivation of* (1) *(...) Competent people have told me that Wiener, under the misapprehension that he already knew what Shannon had done, never actually found out.*
>
> —John R. Pierce [11]

All other independent discoveries in the year of 1948 were in fact essentially what is now referred to Hartley's rule leading to (2). Among these, the first published work in April 1948 was by the French engineer Jacques Laplume [21] from Thompson-Houston. He essentially gives the usual derivation that gives (2) for a signal amplitude range $[0, A]$. C. Earp's publication [19] in June 1948 also makes a similar derivation of (2) where the signal-to-noise amplitude ratio is expressed as a "root-mean-square ratio" for the "step modulation", which is essentially pulse-code modulation. In a footnote, Earp claims that his paper "was written in original form in October, 1946". In an another footnote at the first page, he mentions that

> *A symposium on "Recent Advances in the Theory of Communication" was presented at the November 12, 1947, meeting of the New York section of the Institute of Radio Engineers. Four papers were presented by A. G. Clavier (...); B.D. Loughlin (...); and J. R. Pierce and C. E. Shannon, both of Bell Telephone Laboratories.*
>
> —C.W. Earp [19]

André Clavier is another French engineer from LMT laboratories (subsidiary of ITT Corporation), who published "Evaluation of transmission efficiency according to Hartley's expression of information content" [18] in December 1948. He again makes a similar derivation of (2) as Earp's, expressed with root-mean-square values. As Lundheim notes [22],"it is, perhaps, strange that neither Shannon nor

Clavier have mutual references in their works, since both [2] and [18] were orally presented at the same meeting (...) and printed more than a year afterwards."

In May 1948, Stanford Goldman again re-derived (2), acknowledging that the equation "has been derived independently by many people, among them W. G. Tuller, from whom the writer first learned about it" [20]. William G. Tuller's thesis was defended in June 1948 and printed as an article in May 1949 [8]. His derivation uses again root-mean-square (rms) ratios.

> *Let $S$ be the rms amplitude of the maximum signal that may delivered by the communication system. Let us assume, a fact very close to the truth, that a signal amplitude change less than noise amplitude cannot be recognized, but a signal amplitude change equal to noise is instantly recognizable. Then, if $N$ is the rms amplitude of the noise mixed with the signal, there are $1 + S/N$ significant values of signal that may be determined. (...) the quantity of information available at the output of the system [is $= \log(1 + S/N)$].*
>
> —William G. Tuller [8]

In the 1949 article [8] he explains that

> *The existence of [Shannon's] work was learned by the author in the spring of 1946, when the basic work underlying this paper had just been completed. Details were not known by the author until the summer of 1948, at which time the work reported here had been complete for about eight months.*
>
> —William G. Tuller [8]

In view of this note it is perhaps not completely fair so say, following J.R. Pierce [11] (Shannon's co-author of [12]), that

> *(...) much of the early reaction to Shannon's work was either uninformed or a diversion from his aim and accomplishment. (...) In 1949, William G. Tuller published a paper giving his justification of* (1) *[8].*
>
> —John R. Pierce [11]

Considering that Tuller's work is—apart from Wiener's—the only work referenced by Shannon in [1], and that the oldest reference known (1946) is Tuller's, it should be certainly appropriate to refer to (2) as *Tuller's formula* or to (1) as the *Tuller–Shannon formula*.

There is perhaps no better conclusion for this section than to cite Shannon's 1949 article [2] where he explicitly mentioned (and criticized) Hartley's Law as the property that the maximum amount of information per second is proportional to the bandwidth (without reference to noise limitation), and where he proposed his own interpretation of (2) making the link with his formula (1):

*How many different signals can be distinguished at the receiving point in spite of the perturbations due to noise? A crude estimate can be obtained as follows. If the signal has a power $P$, then the perturbed signal will have a power $P+N$. The number of amplitudes that can be reasonably well distinguished is $K\sqrt{\frac{P+N}{N}}$ where $K$ is a small constant in the neighborhood of unity depending on how the phrase "reasonably well" is interpreted. (...) The number of bits that can be sent in this time is $\log_2 M$ [$= \frac{1}{2}\log_2 K^2\Big(1+\frac{P}{N}\Big)$].*

—Claude Elwood Shannon [2]

It may be puzzling to notice, as Hodges did in his historical book on A. Turing [24], that Shannon's article [2] mentioned a manuscript with a received date of 23 July, 1940! But this was later corrected by Shannon himself in 1984 (cited in [6], Reference 10):

*(...) Hodges cites a Shannon manuscript date 1940, which is, in fact, a typographical error.*
*(...) First submission of this work for formal publication occurred soon after World War ll.*

—Claude Elwood Shannon [6]

This would mean in particular that Shannon's work leading to his formula was completed in 1946, at about the same time as Tuller's.

**4. Hartley's Rule yields Shannon's Formula:** $C' = C$

Let us consider again the argument leading to (2). The channel input $X$ is taking $M = 1 + A/\Delta$ values in the set $\{-A, -A+2\Delta, \ldots, A-2\Delta, A\}$, which is the set of values $(M-1-2k)\Delta$ for $k = 0, \ldots, M-1$. A maximum amount of information will be conveyed through the channel if the input values are equiprobable. Then, using the well-known formula for the sum of squares of consecutive integers, one finds:

$$P = \mathbb{E}(X^2) = \frac{1}{M}\sum_{k=0}^{n}(M-1-2k)^2 = \Delta^2\frac{M^2-1}{3}$$

Interestingly, this is the classical formula for the average power of a $M$-state pulse-code modulation or pulse-amplitude modulation signal, as was derived by Oliver, Pierce and Shannon in [12].

The input is mixed with additive noise $Z$ with accuracy $\pm\Delta$. The least favorable case would be that $Z$ follows a uniform distribution in $[-\Delta, \Delta]$. Then its average power is

$$N = \mathbb{E}(Z^2) = \frac{1}{2\Delta}\int_{-\Delta}^{\Delta} z^2\,\mathrm{d}z = \frac{\Delta^2}{3}$$

It follows that (2) takes the form of a striking identity!

$$C' = \log_2 M = \tfrac{1}{2}\log_2(1+M^2-1) = \tfrac{1}{2}\log_2\Big(1+\frac{3P}{\Delta^2}\Big) = \tfrac{1}{2}\log_2\Big(1+\frac{P}{N}\Big) = C.$$

A mathematical coincidence?

One may perhaps argue that if Tuller or others knew about such a coincidence, they would probably have followed Wiener's attitude in claiming paternity of information theory. In any case, such an identification of (1) and (2) calls for verification that Hartley's rule would in fact be "mathematically correct" as a capacity formula.

## 5. Hartley's Rule as a Capacity Formula

Consider the *uniform channel*, a memoryless channel with additive white noise $Z$ with uniform density in the interval $[-\Delta, \Delta]$. If $X$ is the channel input, the output will be $Y = X + Z$, where $X$ and $Z$ are independent. We assume that the input has the amplitude constraint $|X| \leqslant A$. The following calculation was proposed as a homework exercise in the excellent textbook by Cover and Thomas [25].

**Theorem 1.** *Assuming $A/\Delta$ is integral, the uniform channel has capacity $C'$ given by* (2).

(If $A/\Delta$ is not integral, then the proof of the theorem shows that $C' \leqslant \log_2(1 + A/\Delta)$, yet $C'$ cannot be obtained in closed form.)

**Proof.** From Shannon's coding theorem, the channel's capacity is $C = \max_X I(X;Y)$ bits per sample, where the maximum is taken over all distributions of $X$ such that $|X| \leqslant A$, *i.e.*, with support $[-A, A]$. By expanding mutual information $I(X;Y) = h(Y) - h(Y|X)$ as a difference of differential entropies, and noting that $h(Y|X) = h(Z) = \log_2(2\Delta)$ is constant, the required capacity $C'$ is obtained by maximizing $h(Y)$.

Now since $|X| \leqslant A$, by the triangular inequality, the output amplitude is limited by $|Y| \leqslant |X| + |Z| \leqslant A + \Delta$. Choosing $X = X^*$ to be discrete uniform taking $M = 1 + A/\Delta$ equiprobable values in the set $\{-A, -A + 2\Delta, \ldots, A - 2\Delta, A\}$, it is immediate to see that $Y = X^* + Z$ will have the uniform density over the interval $[-A - \Delta, A + \Delta]$, which is known to maximize $h(Y)$ under the constraint $|Y| \leqslant A + \Delta$. Therefore such an $X^*$ achieve the capacity and we have $C' = \max_X h(Y) - h(Z) = \log_2(2(A + \Delta)) - \log_2(2\Delta) = \log_2(1 + A/\Delta)$. $\square$

Thus there is a sense in which the "Tuller–Shannon Formula" (2) is indeed *correct* as the capacity of a communication channel, except that the communication noise is *not* Gaussian, but uniform, and that the signal limitation is *not* on the power, but on the amplitude (as a side remark, it is interesting to mention that $C'$ is in fact a zero-error capacity and that no coding is actually necessary to achieve it).

The analogy between the Gaussian and uniform channels can be pushed further. Both channels are memoryless and additive, with $Y = X + Z$ where $X$ and $Z$ are independent. Both have "additive" constraints on their inputs of the form $\Phi(X) \leqslant c$, where additivity means that $\Phi(X) \leqslant c$ and $\Phi(Z) \leqslant c'$ imply $\Phi(X + Z) \leqslant c + c'$. Specifically, in the Gaussian case, $\Phi(X) = \mathbb{E}(X^2)$ and additivity results from the fact that $X$ and $Z$ are uncorrelated; and in the uniform case, $\Phi(X) = |X|$ and additivity is simply a consequence of the inequality $|X + Z| \leqslant |X| + |Z|$. Also in both cases, the noise $Z = Z^*$ maximizes the differential entropy $h(Z)$ under the constraint $\Phi(Z) \leqslant c'$, and the input $X = X^*$ that maximizes mutual information $I(X;Y) = I(X; X + Z^*)$ is such that the

corresponding output $Y^* = X^* + Z^*$ also maximizes the differential entropy $h(Y)$ under the constraint $\Phi(Y) \leqslant c + c'$. When $\Phi(X) = \mathbb{E}(X^2)$ (power limitation), both $Y^*$ and $Z^*$ are Gaussian while for $\Phi(X) = |X|$ (amplitude limitation), both $Y^*$ and $Z^*$ have a uniform distribution.

Shannon used these properties for $\Phi(X) = \mathbb{E}(X^2)$ to show that under limited *power*, Gaussian noise is the *worst* possible noise that one can inflict in the channel (in terms of its capacity). To show this, he considered an arbitrary additive noise $Z$ and defined $\tilde{Z}$ as a random variable of the same distribution type as $Z^*$ but with the same differential entropy as $Z$. Thus for $\Phi(X) = \mathbb{E}(X^2)$, $\tilde{Z}$ is a zero-mean Gaussian variable of average power $\tilde{N} = 2^{2h(Z)}/2\pi e$, which is referred to as the *entropy power* [1] of $Z$. He then established that the capacity associated with the noise $Z$ satisfies [1]

$$\tfrac{1}{2}\log_2\Big(1 + \alpha\frac{P}{N}\Big) \leqslant C \leqslant \tfrac{1}{2}\log_2\Big(1 + \frac{P}{N}\Big) + \tfrac{1}{2}\log_2\alpha, \tag{3}$$

where we have noted $\alpha = N/\tilde{N}$. The first inequality was in fact derived by Shannon as a consequence of the *entropy power inequality* (see, e.g., [26] for more details on this inequality). Since $h(\tilde{Z}) = h(Z) \leqslant h(Z^*)$, one has $\tilde{N} \leqslant N$ so that $\alpha \geqslant 1$ (with equality $\alpha = 1$ only in the case of Gaussian noise). It follows from the above inequality that the capacity has the lowest value for Gaussian noise.

The uniform channel enjoys a similar property: under limited *amplitude*, *uniform* noise is the worst possible noise that one can inflict in the channel. To show this, consider the following

**Definition 2** (*Entropic Amplitude*)**.** *Given an arbitrary additive noise $Z$, let $\tilde{Z}$ be a random variable of the same distribution type as $Z^*$ but with the same differential entropy as $Z$. Thus for $\Phi(X) = |X|$, $\tilde{Z}$ is a zero-mean uniformly distributed variable with amplitude $\tilde{\Delta}$. The entropic amplitude of $Z$ is*

$$\tilde{\Delta} = 2^{h(Z)-1}.$$

The squared entropic amplitude is related to the entropy power by the relation $\tilde{\Delta}^2 = \tilde{N}\pi e/2$.

**Theorem 3.** *When $\Phi(X) = |X|$ (amplitude limitation) under the same conditions as Theorem 1, the capacity $C'$ associated with an arbitrary additive noise $Z$ satisfies*

$$\log_2\Big(1 + \frac{A}{\Delta}\Big) \leqslant C' \leqslant \log_2\Big(1 + \frac{A}{\Delta}\Big) + \log_2\alpha, \tag{4}$$

*where $\alpha = \Delta/\tilde{\Delta} \geqslant 1$ (with equality $\alpha = 1$ only for uniform noise).*

It follows as announced that the capacity has the lowest value for uniform noise.

**Proof.** One has $I(X; X+Z) = h(X+Z) - h(Z)$ where $h(Z) = \log_2(2\tilde{\Delta})$ ; since $|Y| \leqslant A + \Delta$, $h(Y) \leqslant \log_2(2(A+\Delta))$. Therefore, $I(X; X+Z) \leqslant \log_2(2(A+\Delta)) - \log_2(2\tilde{\Delta}) = \log_2\Big(1 + \frac{A}{\Delta}\Big) + \log_2\alpha$. Maximizing $I(X; X+Z)$ over the distribution de $X$ in this inequality gives the second inequality in (4).

To prove the first inequality, notice that $C = \max_X I(X; X+Z) \geqslant I(X^*; X^*+Z) = h(X^*+Z) - h(Z)$ where, as above, $X^*$ is discrete uniform in the $M$-ary set $\mathcal{X} = \{-A, -A+2\Delta, \ldots, A-2\Delta, A\}$

with $M = 1 + A/\Delta$. Now $Y = X^* + Z$ follows the density $p_Y(y) = \frac{1}{M}\sum_{x\in\mathcal{X}} p_Z(y-x)$ where $p_Z(z)$ is the density of $Z$. Since $|Z| \leqslant \Delta$ all terms in this sum have disjoint supports. Therefore,

$$h(X^*+Z) = -\sum_{x\in\mathcal{X}}\int_{-\Delta}^{\Delta}\Big(\frac{1}{M}p_Z(y-x)\Big)\log_2\Big(\frac{1}{M}p_Z(y-x)\Big)\,\mathrm{d}y = \log_2 M - \int p_Z(z)\log_2 p_Z(z)\,\mathrm{d}z$$

which reduces to the simple formula $h(X^* + Z) = \log_2 M + h(Z)$. Therefore, $C \geqslant h(X^* + Z) - h(Z) = \log_2 M = \log_2\Big(1 + \frac{A}{\Delta}\Big)$, which proves the first inequality in (4). □

## 6. A Mathematical Analysis

### *6.1. Conditions for Shannon's Formula to Hold*

In this section, we consider a memoryless additive noise channel with zero-mean input $X$ and output $Y = X + Z$. Such a channel is defined by:

- the probability density function (pdf) $p_Z$ of the zero-mean noise $Z$, which is assumed independent of $X$;
- a constraint set $\mathcal{C}$ on the possible distributions of $X$. The channel capacity is computed under this constraint as

$$C = \max_{X\in\mathcal{C}} I(X;Y) = \max_{X\in\mathcal{C}} h(Y) - h(Z) = \big(\max_{X\in\mathcal{C}} h(X+Z)\big) - h(Z).$$

We let $X^*$ be the input that attains this maximum and let $Y^* = X^* + Z$ be the corresponding output. Thus $C = h(Y^*) - h(Z) = h(X^* + Z) - h(Z)$. We also let $P = \mathbb{E}(X^{*2})$ and $N = \mathbb{E}(Z^2)$ so that $P/N$ denotes the signal-to-noise ratio at the optimum.

**Lemma 4.** *If there exists a number $\alpha > 1$ such that $\alpha Z$ and $Y^*$ share the same distribution, then the channel capacity $C$ is given by Shannon's Formula* (1)*.*

**Proof.** One has $C = h(Y^*) - h(Z) = h(\alpha Z) - h(Z) = \log_2|\alpha| = \frac{1}{2}\log_2\alpha^2$. However, $P + N = \mathbb{E}(X^2) + \mathbb{E}(Z^2) = \mathbb{E}(Y^2) = \alpha^2\,\mathbb{E}(Z^2) = \alpha^2 N$ and so $\alpha^2 = 1 + P/N$. This gives (1). □

**Example 1** (Gaussian channel)**.** *Here both $Z$ and $Y^* = X^* + Z$ are zero-mean Gaussian so that the condition of the lemma is satisfied. We recover* (1) *as the classical expression for the channel capacity.*

**Example 2** (uniform channel)**.** *Here both $Z$ and $Y^* = X^* + Z$ are uniformly distributed over a centered interval so the condition of the lemma is also satisfied. This explains anew the coincidence found in the calculation of Section 4.*

In the following we note $\phi_X(\omega) = \mathbb{E}(e^{i\omega X})$, the characteristic function of any random variable $X$.

**Lemma 5.** *The condition of Lemma 4 is satisfied if and only if there exists $\alpha > 1$ such that*

$$\frac{\phi_Z(\alpha\omega)}{\phi_Z(\omega)} = \phi_{X^*}(\omega)$$

**Proof.** $\alpha Z$ and $Y^* = X^* + Z$ have the same distribution if and only if they share the same characteristic function, which is equal to $\phi_{\alpha Z}(\omega) = \phi_Z(\alpha\omega)$ and to $\phi_{Y^*}(\omega) = \phi_{X^*}(\omega)\phi_Z(\omega)$. □

In particular the above quotient must be a characteristic function of some random variable. This shows that the distribution of $Z$ should be *divisible*.

**Example 3** (Gaussian channel (continued))**.** *Here* $\alpha^2 = \frac{P+N}{N}$ *and*

$$\frac{\phi_Z(\alpha\omega)}{\phi_Z(\omega)} = \frac{e^{-\alpha^2\omega^2 N/2}}{e^{-\omega^2 N/2}} = e^{-\omega^2 P/2}$$

*which the characteristic function of* $X^* \sim \mathcal{N}(0, P)$.

**Example 4** (uniform channel (continued))**.** *Here* $\alpha = \frac{A+\Delta}{\Delta} = M$ *is assumed integral and*

$$\frac{\phi_Z(M\omega)}{\phi_Z(\omega)} = \frac{\operatorname{sinc}(M\Delta\cdot\omega)}{\operatorname{sinc}(\Delta\cdot\omega)} = \frac{\sin(M\Delta\cdot\omega)}{M\sin(\Delta\cdot\omega)} = \frac{1}{M}(e^{-i(M-1)\omega\Delta} + e^{-i(M-3)\omega\Delta} + \cdots + e^{i(M-1)\omega\Delta})$$

*where* $\operatorname{sinc} x = \frac{\sin x}{x}$ *is the sine cardinal function and where the last equality is the well-known Dirichlet kernel expression. The result is the characteristic function of* $X^*$ *which take* $M$ *equiprobable values in the set* $\{-(M-1)\Delta, -(M-3)\Delta, \ldots, (M-3)\Delta, (M-1)\Delta\}$.

**Example 5** (Cauchian channel)**.** *Let* $Z$ *be Cauchy distributed with* $p_Z(z) = \frac{1}{\pi}\frac{a}{a^2+z^2}$*, where* $a > 0$*. Then for any* $\alpha > 0$*,*

$$\frac{\phi_Z(\alpha\omega)}{\phi_Z(\omega)} = \frac{e^{-a\alpha|\omega|}}{e^{-a|\omega|}} = e^{-a(\alpha-1)|\omega|}$$

*is the characteristic function of* $X^*$*, which is Cauchy distributed with parameter* $(\alpha-1)a$*. However, in this particular case,* $P = \mathbb{E}(X^{*2}) = +\infty$ *and* $N = \mathbb{E}(Z^2) = +\infty$ *so that the signal-to-noise ratio is not defined.*

**Lemma 6.** *Let* $p_Z$ *and* $p_{Y^*}$ *be the pdf's of* $Z$ *and* $Y^*$*, respectively. Then* $X^*$ *attains capacity subject to an average cost per channel use of the form* $\mathbb{E}(b(X)) \leqslant C$*, where*

$$b(x) = \mathbb{E}\Big(\log_2 \frac{p_Z(Z)}{p_{Y^*}(x+Z)}\Big). \tag{5}$$

Thus given the pdf of $Y^*$, (5) defines an adequate constraint set $\mathcal{C}$ so that $C = h(Y^*) - h(Z)$.

**Proof.** Let $p_Y$ be the pdf of $Y = X + Z$. By the information inequality $D(p_Y \| p_{Y^*}) \geqslant 0$, we obtain

$$h(Y) \leqslant \mathbb{E}\log_2 \frac{1}{p_{Y^*}(Y)} = \mathbb{E}_X\Big(\mathbb{E}_Z \log \frac{1}{p_{Y^*}(X+Z)}\Big).$$

Therefore,

$$I(X;Y) = h(Y) - h(Z) \leqslant \mathbb{E}_X\Big(\mathbb{E}_Z \log \frac{p_Z(Z)}{p_{Y^*}(X+Z)}\Big) = \mathbb{E}(b(X))$$

Equality holds if and only if $p_Y = p_{Y^*}$, that is, when the channel capacity is attained. In this case $\max I(X;Y) = \mathbb{E}(b(X))$ should be equal to the capacity $C$. The assertion follows. □

**Example 6** (Gaussian channel (continued))**.** *Here $Z \sim \mathcal{N}(0, N)$ and $Y^* \sim \mathcal{N}(0, P+N)$. Therefore,*

$$b(x) = \log_2 \sqrt{\frac{P+N}{N}} + \mathbb{E}\log_2 \exp\Big(\frac{(x+Z)^2}{2(P+N)} - \frac{Z^2}{2N}\Big) = C + \frac{\log_2 e}{2}\Big(\frac{x^2+N}{P+N} - 1\Big).$$

*The constraint $\mathbb{E}(b(X)) \leqslant C$ is now equivalent to $\mathbb{E}(X^2) \leqslant P$ as expected.*

**Example 7** (uniform channel (continued))**.** *Here $Z$ is uniformly distributed on the interval $[-\Delta, \Delta]$ and $Y^*$ is uniformly distributed on $[-A-\Delta, A+\Delta]$ where $A = (\alpha - 1)\Delta > 0$. Therefore,*

$$b(x) = \log_2 \frac{A+\Delta}{\Delta} + \mathbb{E}\log \frac{1}{\mathbf{1}_{|x+Z| \leqslant A+\Delta}}$$

*where $\mathbf{1}$ denotes the indicator function. The first term in the r.h.s. is equal to $C$. If $|x| \leqslant A$ then $|x+Z| \leqslant A+\Delta$ a.e. so that the second term equals $\log 1 = 0$. Otherwise, $\mathbf{1}_{|x+z| \leqslant A+\Delta}$ vanishes for $z$ in some subinterval of $[-\Delta, \Delta]$ of positive length and the second term is infinite. Hence*

$$b(x) = \begin{cases} C & \text{if } |x| \leqslant A \\ +\infty & \text{otherwise.} \end{cases}$$

*The constraint $\mathbb{E}(b(X)) \leqslant C$ is equivalent to $|X| \leqslant A$ a.e. as expected.*

**Theorem 7.** *Assume that there exists $\alpha > 1$ such that $\dfrac{\phi_Z(\alpha\omega)}{\phi_Z(\omega)}$ is a characteristic distribution and let $\mathcal{C}$ be defined by the condition $\mathbb{E}(b(X)) \leqslant C$ where*

$$b(x) = \mathbb{E}\log_2 \frac{\alpha p_Z(Z)}{p_Z((x+Z)/\alpha)}. \tag{6}$$

*Then the channel capacity $C = \log_2 \alpha$ of the corresponding additive noise channel is given by Shannon's Formula* (1)*.*

**Proof.** Apply the preceding lemmas, noting that $p_{Y^*}(y) = \frac{1}{\alpha} p_Z(\frac{y}{\alpha})$. □

*6.2. B-Spline Channels of Any Degree*

Equipped with Theorem 7 we can construct many additive noise channels whose capacities are given by Shannon's Formula (1).

**Definition 8** (*B-spline Channel*)**.** *Let $U_\Delta$ be uniformly distributed over the interval $[-\Delta, \Delta]$ and let $d \in \mathbb{N}$. Define*

$$Z_d = U_{\Delta,0} + U_{\Delta,1} + \ldots + U_{\Delta,d}$$

*where the $U_{\Delta,i}$ are independent copies of $U_\Delta$. The (uniform)* B-spline channel *of degree $d$ is the associated additive noise channel $Y = X + Z_d$ with capacity $C_d$.*

For $d = 0$ one recovers the uniform channel. It is easily seen and well-known that the pdf of $Z_d$ is the uniform $B$-spline function:

$$p_{Z_d}(z) = \frac{1}{2\Delta} \cdot \beta_d\Big(\frac{z}{2\Delta}\Big)$$

where $\beta_d$ is the standard central B-spline [27] of order $d$, the $(d+1)$th convolution power of the indicator function of the interval $[-1/2, 1/2]$.

**Theorem 9.** *For all $d \in \mathbb{N}$ and any choice of a positive integer $M$, the capacity $C_d$ of the B-spline channel of degree $d$ under the input constraint $\mathbb{E}(b_d(X)) \leqslant C_d$ where*

$$b_d(x) = \mathbb{E}\log_2 \frac{M\beta_d\Big(\frac{Z}{2\Delta}\Big)}{\beta_d\Big(\frac{x+Z}{2M\Delta}\Big)}. \tag{7}$$

*is given by Shannon's Formula* (1).

**Proof.** Since $p_{Z_d}(z) = \frac{1}{2\Delta} \cdot \beta_d\big(\frac{z}{2\Delta}\big)$ is the $(d+1)$th convolution power of the rectangle function of the interval $[-\Delta, \Delta]$, the corresponding characteristic function is a $(d+1)$th power of a cardinal sine:

$$\phi_{Z_d}(\omega) = \mathrm{sinc}^{d+1}(\Delta \cdot \omega).$$

Let $M > 0$ be an integer. From Example 4, we have

$$\begin{aligned}\frac{\phi_{Z_d}(M\omega)}{\phi_{Z_d}(\omega)} &= \frac{\mathrm{sinc}^{d+1}(M\Delta \cdot \omega)}{\mathrm{sinc}^{d+1}(\Delta \cdot \omega)} = \Big(\frac{\sin(M\Delta \cdot \omega)}{M\sin(\Delta \cdot \omega)}\Big)^{d+1} \\ &= \Big(\frac{1}{M}\big(e^{-i(M-1)\omega\Delta} + e^{-i(M-3)\omega\Delta} + \cdots + e^{i(M-1)\omega\Delta}\big)\Big)^{d+1}.\end{aligned}$$

This is the characteristic function of the random variable

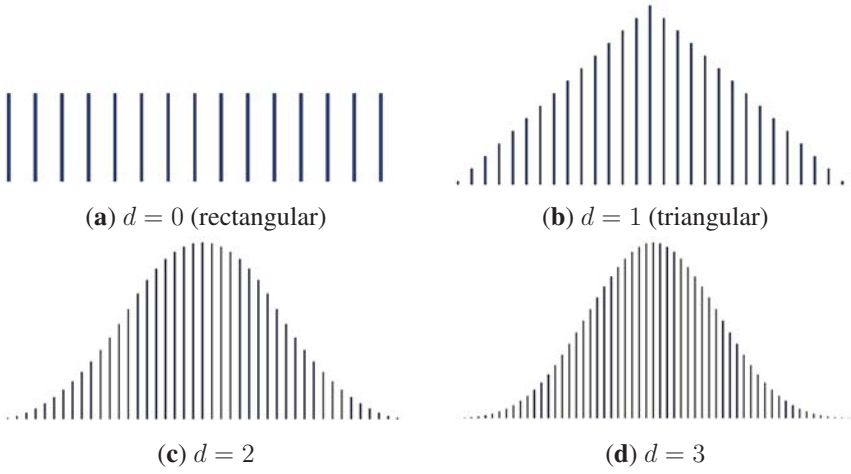
$$X_d = X_{M,0} + \cdots + X_{M,d},$$

where the $X_{M,i}$ are i.i.d. and take $M$ equiprobable values in the set $\{-(M-1)\Delta, -(M-3)\Delta, \ldots, (M-3)\Delta, (M-1)\Delta\}$. Hence, Theorem 7 applies with $\alpha = M$ and cost function (7). □

Again for $d = 0$ one recovers the case of the uniform channel with input $X_0 = X_{M,0}$ taking $M$ equiprobable values in the set $\{-(M-1)\Delta, -(M-3)\Delta, \ldots, (M-3)\Delta, (M-1)\Delta\}$ (Figure 1a). In general, the probability distribution of $X_d$ is the $(d+1)$th discrete convolution power of the uniform distribution. For $d = 1$, the pdf of the noise has a triangular shape and the distribution of $X_d$ is also triangular (Figure 1b). As $d$ increases, it becomes closer to a Gaussian shape (Figure 1c,d).

### *6.3. Convergence as $d \to +\infty$*

To determine the limit behavior as $d \to +\infty$, we need to apply some normalization on the probability distributions. Since the pdf of $Z_d$ is obtained by successive convolutions of rectangles of length $2\Delta$, its support $[-(d+1)\Delta, (d+1)\Delta]$ as well as its average power (or variance) $N =$



(**a**) $d = 0$ (rectangular) (**b**) $d = 1$ (triangular)

(**c**) $d = 2$ (**d**) $d = 3$

**Figure 1.** Discrete plots of input probability distributions (of $X_d$) that attain capacity for $M = 15$ and different values of $d$.

$(d+1)\Delta^2/3$ increase linearly with $(d+1)$. Similarly, the support and average power $P$ of $X^*$ also increase linearly with $(d+1)$. Although this does not affect the ratio $P/N$, in order for average powers $P$ and $N$ to converge as $d \to +\infty$ we need to divide $Z_d$ and $X^*$, hence their sum $Y$, by $\sqrt{d+1}$. The capacity will remain unaltered because

$$\begin{aligned} I\Big(\frac{X}{\sqrt{d+1}}; \frac{Y}{\sqrt{d+1}}\Big) &= h\Big(\frac{Y}{\sqrt{d+1}}\Big) - h\Big(\frac{Z}{\sqrt{d+1}}\Big) \\ &= h(Y) - \tfrac{1}{2}\log(d+1) - h(Z) + \tfrac{1}{2}\log(d+1) \\ &= h(Y) - h(Z) \\ &= I(X;Y). \end{aligned}$$

Therefore, in what follows, we assume that all random variables $X, Y, Z$ have been normalized by the factor $\sqrt{d+1}$. We then say that the additive channel with input $X_d$, output $Y_d$, noise $Z_d$, and cost function $b_d(x)$ converges as $d \to +\infty$ to the additive channel with input $X$, output $Y$, noise $Z$, and cost function $b(x)$ if $X_d \to X$, $Y_d \to Y$, $Z_d \to Z$ in distribution, and $b_d(x) \to b(x)$.

**Theorem 10.** *The B-spline channel of degree $d$ converges to the Gaussian channel as $d \to +\infty$.*

**Proof.** By the central limit theorem,

$$\frac{Z_d}{\sqrt{d+1}} = \frac{U_{\Delta,0} + U_{\Delta,1} + \ldots + U_{\Delta,d}}{\sqrt{d+1}}$$

converges in distribution to the Gaussian $Z \sim \mathcal{N}(0, N)$ (in fact, the B-spline pdf converges uniformly to the Gaussian pdf) [27]. Since $Y_d$ has the same distribution as $M \cdot Z_d$, it also converges in distribution to the Gaussian $Y \sim \mathcal{N}(0, P+N)$. Again by the central limit theorem,

$$\frac{X^*}{\sqrt{d+1}} = \frac{X_0^* + \cdots + X_d^*}{\sqrt{d+1}}$$

converges in distribution to the Gaussian $\mathcal{N}(0, P)$. Finally, we can write

$$b_d(x) = \mathbb{E}\Big(\log_2 \frac{M p_{Z_d}(Z_d)}{p_{Z_d}(\frac{x+Z_d}{M})}\Big) = \mathbb{E}\Big(\log_2 \frac{p_{Z_d}(Z_d)}{p_Z(Z_d)}\Big) - \mathbb{E}\Big(\log_2 \frac{p_{Z_d}(\frac{x+Z_d}{M})}{p_Z(\frac{x+Z_d}{M})}\Big) + \mathbb{E}\Big(\log_2 \frac{M p_Z(Z_d)}{p_Z(\frac{x+Z_d}{M})}\Big)$$

The first term in the r.h.s. tends to zero by the strengthened central limit theorem of Barron [28] in relative entropy. The second term also tends to zero by a similar argument and change of variable. By a calculation identical to that of Example 6, the third term is equal to

$$\log_2 M + \mathbb{E}\log_2 \exp\Big(\frac{(x+Z_d)^2}{2(P+N)} - \frac{Z_d^2}{2N}\Big) = C + \frac{\log_2 e}{2}\Big(\frac{x^2+N}{P+N} - 1\Big) = b(x)$$

which shows that $b_d(x) \to b(x)$ as $d \to +\infty$. $\square$

Figure 2 shows the graphs of the cost functions $b_d(x)$ for different values of degree $d$. As the degree increases, the curves converge to the parabola that represents the quadratic cost function $b(x)$ for the Gaussian channel.

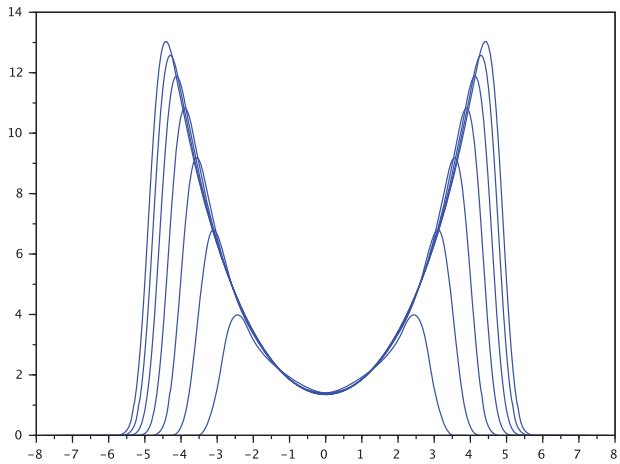



**Figure 2.** Cost functions $b_d(x)$ for $d = 1$ to 7 (with $M = 4$ and $\Delta = 1$). Convergence holds to the quadratic cost function $b(x)$.

Thus we have built a sequence of additive noise "B-spline" channels indexed by $d \in \mathbb{N}$ that makes the transition from the uniform ($d = 0$) to the Gaussian channel ($d \to \infty$). Shannon's Formula (1) holds for all these channels.

**Acknowledgments**

The authors wish to thank Max H. M. Costa for valuable discussions and suggestions. This work was partially supported by São Paulo Research Foundation (FAPESP) Grant # 2014/13835-6, under

the FAPESP thematic project *Segurança e Confiabilidade da Informação: Teoria e Prática*, Grant # 2013/25977-7.

**Author Contributions**

Both authors performed the historical research. Olivier Rioul wrote the paper and carried out the mathematical analysis. Both authors have read and approved the final manuscript.

**Conflicts of Interest**

The authors declare no conflict of interest.

**References**

1. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423, 623–656. Reprinted in C.E. Shannon and W. Weaver *The Mathematical Theory of Communication*; University Illinois Press: Champaign, IL, USA, 1949.
2. Shannon, C.E. Communication in the presence of noise. *Proc. Inst. Radio Eng.* **1949**, *37*, 10–21.
3. Butzer, P.; Dodson, M.; Ferreira, P.; Higgins, J.; Lange, O.; Seidler, P.; Stens, R. Multiplex signal transmission and the development of sampling techniques: The work of Herbert Raabe in contrast to that of Claude Shannon. *Appl. Anal.* **2011**, *90*, 643–688.
4. Wikipedia: Shannon–Hartley theorem. Available online: http://en.wikipedia.org/wiki/Shannon-Hartley_theorem (accessed on 28 August 2014).
5. Hartley, R.V.L. Transmission of information. *Bell Syst. Tech. J.* **1928**, *7*, 535–563.
6. Ellersick, F.W. A conversation with Claude Shannon. *IEEE Commun. Mag.* **1984**, *22*, 123–126.
7. Wozencraft, J.M.; Jacobs, I.M. *Principles of Communication Engineering*; John Wiley & Sons: New York, NY, USA, 1965; pp. 2–5.
8. Tuller, W.G. Theoretical limitations on the rate of transmission of information. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 1948; Reprinted in *Proc. Inst. Radio Eng.* **1949**, *37*, 468–478.
9. Cherry, E.C. A history of information theory. *Proc. Inst. Elect. Eng.* **1951**, *98*, 383–393.
10. McEliece, R.J. *The Theory of Information and Coding*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2002.
11. Pierce, J.R. The early days of information theory. *IEEE Trans. Inf. Theory* **1973**, *19*, 3–8.
12. Oliver, B.; Pierce, J.; Shannon, C.E. The Philosophy of PCM. *Proc. Inst. Radio Eng.* **1948**, *36*, 1324–1331.
13. Gallager, R. Claude E. Shannon: A retrospective on his life, work, and impact. *IEEE Trans. Inf. Theory* **2001**, *47*, 2681–2695.
14. Verdú, S. Fifty years of Shannon theory. *IEEE Trans. Inf. Theory* **1998**, *44*, 2057–2078.
15. Golay, M.J.E. Note on the theoretical efficiency of information reception with PPM. *Proc. Inst. Radio Eng.* **1949**, *37*, 1031.

16. Slepian, D. Information theory in the fifties. *IEEE Trans. Inf. Theory* **1973**, *19*, 145–148.
17. Wiener, N. Time series, Information and Communication. In *Cybernetics*; John Wiley & Sons: New York, NY, USA, 1948; Chapter III, pp. 10–11.
18. Clavier, A.G. Evaluation of transmission efficiency according to Hartley's expression of information content. *Electron. Commun. ITT Tech. J.* **1948**, *25*, 414–420.
19. Earp, C.W. Relationship between rate of transmission of information, frequency bandwidth, and signal-to-noise ratio. *Electron. Commun. ITT Tech. J.* **1948**, *25*, 178–195.
20. Goldman, S. Some fundamental considerations concerning noise reduction and range in radar and communication. *Proc. Inst. Radio Eng.* **1948**, *36*, 584–594.
21. Laplume, J. Sur le nombre de signaux discernables en présence du bruit erratique dans un système de transmission à bande passante limitée. *Comptes rendus de l'Académie des Sciences de Paris* **1948**, *226*, 1348–1349. (In French)
22. Lundheim, L. On Shannon and "Shannon's Formula". *Telektronikk* **2002**, *98*, 20–29.
23. Wiener, N. What is information theory? *IRE Trans. Inf. Theory* **1956**, *2*, 48.
24. Hodges, A. *Alan Turing: The Enigma*; Simon and Schuster: New York, NY, USA, 1983; p. 552.
25. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; John Wiley & Sons: New York, NY, USA, 2006.
26. Rioul, O. Information theoretic proofs of entropy power inequalities. *IEEE Trans. Inf. Theory* **2011**, *57*, 33–55.
27. Unser, M.; Aldroubi, A.; Eden, M. On the asymptotic convergence of B-spline wavelets to Gabor functions. *IEEE Trans. Inf. Theory* **1992**, *38*, 864–872.
28. Barron, A.R. Entropy and the central limit theorem. *Ann. Probab.* **1986**, *14*, 336–342.

# Chapter 2:
# Mathematical and Physical Foundations of Information and Entropy Geometric Structures

# Symmetry, Probabiliy, Entropy: Synopsis of the Lecture at MAXENT 2014

**Misha Gromov**

**Abstract:** In this discussion, we indicate possibilities for (homological and non-homological) linearization of basic notions of the probability theory and also for replacing the real numbers as values of probabilities by objects of suitable combinatorial categories.

Reprinted from *Entropy*. Cite as: Gromov, M. Symmetry, Probabiliy, Entropy: Synopsis of the Lecture at MAXENT 2014. *Entropy* **2015**, *17*, 1273–1277.

The success of the probability theory decisively, albeit often invisibly, depends on symmetries of systems this theory applies to. For instance:

- The symmetry group of *a single round of gambling with three dice* has order $288 = 6 \times 6 \times 8$: it is a semidirect product of the permutation group $S_3$ of order 6 and the symmetry group of the $3d$ cube, that is, in turn, is a semidirect product of $S_3$ and $\{\pm 1\}^3$.

- *The Bernoulli spaces* $(\blacksquare_p,\ \blacklozenge_{1-p})^{\mathbb{Z}}$, $0 < p < 1$, of $(\blacksquare,\ \blacklozenge)$-sequences indexed by integers $z \in \mathbb{Z} = \{\cdots, -2, -1, 0, 1, 2, \cdots\}$ are acted upon by a semidirect product of the infinite permutation group

$$S_{\infty=\mathbb{Z}} \supset \mathbb{Z} = \{\cdots, -2, -1, 0, 1, 2, \cdots\}$$

  and the (compact) group $\{\pm 1\}^{\mathbb{Z}} = \{\ \blacksquare \leftrightarrow \blacklozenge\}^{\mathbb{Z}}$, with the role of the latter being essential even for $p \neq \frac{1}{2}$ where the probability measure is not preserved.

- The system of *identical point-particles* $\bullet_i$ *in the Euclidean* 3*-space* $\mathbb{R}^3$, that are indexed by a countable set $I \ni i$, is acted upon by the isometry group of $\mathbb{R}^3$ times the infinite permutation group $S_{\infty=I}$.

- *Buffon's probabilistic needle formula for* $\pi = 3.141592653589793\cdots$ relies on the invariance of the Haar measure on the circle.

**I.** What happens if the symmetry is enhanced, e.g., from the permutation group $S_{\infty=I}$ to the group $GL_{\mathbb{F}}(\infty)$ of liner transformations of the vector space $\mathbb{F}^I$ (formally) spanned by symbols $[i]$, $i \in I$, regarded as (linearly independent) vectors over a filed $\mathbb{F}$?

**II.** What could you do if your system is *inherently heterogeneous*, such as *a folding polypeptide chain* or a *natural language*, for instance?

*Hilbertisation/unitarisation/quantization* of set categories brought along a development of several magnificent *non-commutative probability theories*, e.g., of those under the headings of *von-Neumann algebras, von Neumann entropy* [1,2], *free probabilities* [3].

By comparison, the achievements of the *non-unitary linearisation* of probability theory are modest—just a few amusing observations.

**Example 1.** *Linearized Loomis-Whitney-Shannon-Shearer Submultiplicativity Inequality* [4,5].

*Let $\Phi = \Phi(x_1, x_2, x_3, x_4)$ be a 4-linear function (form) over some field (where the variables $x_i$ run over some vector spaces $X_i$). Then the ranks of the following four* ***bilinear*** *forms $\Phi(x_1, x_2 \otimes x_3 \otimes x_4), \Phi(x_1 \otimes x_2, x_3 \otimes x_4), \Phi(x_1 \otimes x_3, x_2 \otimes x_4)$ and $\Phi(x_1 \otimes x_4, x_2 \otimes x_3)$ satisfy*

$$(rank[1, 234])^2 \leq rank[12, 34] \cdot rank[13, 24] \cdot rank[14, 23].$$

**Example 2.** *Homology Measures* [6].

*Homologies $H_*(X) = \oplus_i H_i(X)$ of topological spaces $X$ and natural subgroups in $H_*$ are graded Abelian groups: their ranks are properly represented not by individual numbers $r_i$, but by Poincaré polynomials $P_X(t) = \sum_i r_i \cdot t^i$.*

*The polynomial valued set function $U \mapsto P_U$, $U \subset X$, has some measure/entropy-like properties that become more pronounced for the ideal valued function that assigns the kernels*

$$Ker_{X \smallsetminus U} \subset H^*(X; A)$$

*of the inclusion/restriction cohomology homomorphisms for the complements $X \smallsetminus U \subset X$ for subsets $U \subset X$,*

$$U \mapsto \mu^*(U) =_{def} Ker_{X \smallsetminus U} =_{def} Ker\big[H^*(X; A) \to H^*(X \smallsetminus U; A)\big],$$

*for some Abelian (cohomology coefficient) group $A$.*

The basic properties of this $\mu^*$ (stated slightly differently in topology textbooks) have an attractive measure theoretic flavour. Namely,

*$\mu^*(U)$ is* **additive** *for the sum-of-subsets in the group $H^*(X; A)$ and, if $A$ is a commutative ring, then $\mu^*$ is* **super-multiplicative** *for the the $\smile$-product of ideals:*

$$\mu^*(U_1 \cup U_2) = \mu^*(U_i) \dotplus \mu^*(U_2)$$

for *disjoint* open subsets $U_1$ and $U_2$ in $A$, and

$$\mu^*(U_1 \cap U_2) \supset \mu^*(U_1) \smile \mu^*(U_2)$$

for all open $U_1, U_2 \subset A$.

Next, given a linear subspace $\Theta \subset H^*(X; A)$, let

$$\mu_\Theta(U) = \Theta \cap Ker_{X \smallsetminus U}$$

and, assuming $A$ is (the additive group of) a field, denote the rank of $\mu_\Theta(U)$ over this field by $|\mu_\Theta(U)| = |\mu_\Theta(U)|_A$.

*Linearized Matsumoto-Tokushige Separation Inequality in the $N$-torus.*

Let $U_1, U_2 \subset \mathbb{T}^N$ be *non-intersecting* (closed or open) subsets and let

$$\Theta_1 = H^{n_1}(\mathbb{T}^N; A), \text{and } \Theta_2 = H^{n_2}(\mathbb{T}^N; A)$$

for $n_i \leq N/2$, $i = 1, 2$, and some field $A$. Then

$$|\mu_{\Theta_1}(U_1)| \cdot |\mu_{\Theta_2}(U_2)| \leq c \cdot |\Theta_1| \cdot |\Theta_2|$$

*for* $c = n_1 n_2 / N^2$ *and where, observe,* $|\Theta_i = \wedge^{n_i} A| = \binom{N}{n_i}$.

If we think of the torus $\mathbb{T}^N$ as a physical system of $N$ uncoupled linear oscillators then the "measures" $\mu^*(U)$ and/or $\mu_\Theta(U)$ may be interpreted as

*"the numbers of persistent degrees of freedom" of this system that are observable from* $U$.

*Probabilistic/entropic interpretation of homology*, which is kind of "dual" to "homological interpretation of entropy-like invariants" by Bennequin [7], and also by Drummond-Cole *et al.* [8,9], is also possible for "coupled systems" [10] where particularly attractive ones are systems of moving *disjoint* balls in space where the configuration spaces of these systems support rich homology structures that are induced from the classifying spaces of (subgroups of) infinite symmetric groups $S_{\infty=I}$ [11], that is expanded/corrected in [12].

A mathematical study of "loose structures" such as what you find in biology and linguistics needs generalisations that would allow a use of relaxed, rather than enhanced, symmetries.

For instance, just to warm up, one may start by elaborating on the category theoretic definition of the entropy suggested "In a Search for a Structure, Part 1: On Entropy" [13], where the entropy of a finite probability space $P = \{p_i\}$, $p_i > 0$, $\sum_i p_i = 1$, comes as the class $[P]_{Gro}$ of $P$ in the *Grothendieck group* $Gro(\mathcal{P})$ of the topological category $\mathcal{P}$ of finite probability spaces $P$ and probability/measure preserving maps $P \to Q$ with a properly defined topological structure in $\mathcal{P}$.

Since *the group* $Gro(\mathcal{P})$ *is isomorphic to the multiplicative group of positive real numbers* [13]—this is a reformulation of the Bernoulli law of large numbers – the Grothendieck class $[P]_{Gro}$ can be identified with $\exp ent(P)$.

In general, such a Grothendieck-style entropy would be not a *number valued* function of any kind, but (not quite) a functor from an elaborate combinatorial (not quite) category, e.g., comprised of fragments of a natural language with some (not always composable) "morphisms/arrows" between them, to some "simple category" e.g., the category of weighted trees.

The so modified probability/entropy theory is badly needed for designing algorithms that would model what we call *(ego)learning* described in " Ergostructures, Ergodic and the Universal Learning Problem" [14] and in " Understanding Languages and Making Dictionaries" [15], (in preparation) but I have not progressed much in pursuing this direction yet.

**Acknowledgments**

I want to thank Frederic Barbaresco for his interest in the subject matter of this paper and for inviting me to the MaxEnt'14 meeting in Amboise, France, and the anonymous referees for their friendly comments and suggestions.

**Conflicts of Interest**

The author declares no conflict of interest.

## References


1. Parthasarathy, K.R. *An Introduction to Quantum Stochastic Calculus*; Modern Birkhäuser Classics; Springer: Basel, Switzerland, 1992.
2. Meyer, P.-A. *Quantum Probability for Probabilists*, 2nd ed.; Lecture Notes in Mathematics; Springer: Berlin/Heidelberg, Germany, 1995.
3. Nica, A.; Speicher, R.; Voiculescu, D. Free Probability Theory. Available online: https://www.birs.ca/workshops/2004/04w5028/report04w5028.pdf (accessed on 10 March 2015).
4. Gromov, M. Entropy and isoperimetry for linear and non-linear group actions. *Groups Geom. Dyn.* **2008**, *2*, 499–593. Available online: http://www.ihes.fr/~gromov/topics/grig-final-june11-08.pdf (accessed on 10 March 2015).
5. Gromov, M. Six Lectures on Probabiliy, Symmetry, Linearity. Available online: http://www.ihes.fr/~gromov/PDF/probability-paris-Oct-2014.pdf (accessed on 10 March 2015).
6. Gromov, M. Singularities, expanders and topology of maps. Part 2: From combinatorics to topology via algebraic isoperimetry. *Geom. Funct. Anal.* **2010**, *20*, 416–526. Available online: http://www.ihes.fr/~gromov/PDF/morse2_gafa.pdf (accessed on 10 March 2015).
7. Bennequin, D. Homological interpretation of entropy-like invariants. In Proceedings of 34th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, 21–26 September 2014, Amboise, France.
8. Drummond-Cole, G.C.; Park, J.-S.; Terilla, J. Homotopy Probability Theory I. *J. Homotopy Relat. Struct.* **2013**, doi: 10.1007/s40062-013-0067-y.
9. Drummond-Cole, G.C.; Park, J.-S.; Terilla, J. Homotopy Probability Theory II. *J. Homotopy Relat. Struct.* **2013**, doi: 10.1007/s40062-014-0078-3.
10. Bertelson, M.; Gromov, M. *Dynamical Morse Entropy*; Modern dynamical systems and applications; Springer: Berlin, Germany, 2010. Available online: www.ihes.fr/~gromov/PDF/dynamical-entropy-final-bis.pdf (accessed on 10 March 2015).
11. Gromov, M. Number of Questions. Available online: http://www.ihes.fr/~gromov/PDF/sept2014-copy.pdf (accessed on 10 March 2015); Section 4, pp. 82–102.
12. Gromov, M. Morse Spectra, Homology Measures and Parametric Packing Problems. **2015**, in preparation.
13. Gromov, M. In a Search for a Structure, Part 1: On Entropy. Available online: http://www.ihes.fr/~gromov/PDF/structre-serch-entropy-july5-2012.pdf (accessed on 10 March 2015).
14. Gromov, M. Ergostuctures, Ergologic and the Universal Learning Problem: Chapters 1,2. Available online: http://www.ihes.fr/~gromov/PDF/ergologic3(1).pdf (accessed on 10 March 2015).
15. Gromov, M. Understanding Languages and Making Dictionaries. **2015**, in preparation.

# The Homological Nature of Entropy

**Pierre Baudot and Daniel Bennequin**

**Abstract:** We propose that entropy is a universal co-homological class in a theory associated to a family of observable quantities and a family of probability distributions. Three cases are presented: (1) classical probabilities and random variables; (2) quantum probabilities and observable operators; (3) dynamic probabilities and observation trees. This gives rise to a new kind of topology for information processes, that accounts for the main information functions: entropy, mutual-informations at all orders, and Kullback–Leibler divergence and generalizes them in several ways. The article is divided into two parts, that can be read independently. In the first part, the introduction, we provide an overview of the results, some open questions, future results and lines of research, and discuss briefly the application to complex data. In the second part we give the complete definitions and proofs of the theorems A, C and E in the introduction, which show why entropy is the first homological invariant of a structure of information in four contexts: static classical or quantum probability, dynamics of classical or quantum strategies of observation of a finite system.

Reprinted from *Entropy*. Cite as: Baudot, P.; Bennequin, D. The Homological Nature of Entropy. *Entropy* **2015**, *17*, 3253–3318.

## Contents

## 1. Introduction

### *1.1. What is Information ?*

"What is information ?" is a question that has received several answers according to the different problems investigated. The best known definition was given by Shannon [1], using random variables and a probability law, for the problem of optimal message compression. However, the first definition was given by Fisher, as a metric associated to a smooth family of probability distributions, for optimal discrimination by statistical tests; it is a limit of the Kullback–Leibler divergence, which was introduced to estimate the accuracy of a statistical model of empirical data, and which can be also viewed as a quantity of information. More generally Kolmogorov considered that the concept of information must precede probability theory (*cf.* [2]). However, Evariste Galois saw the application of group theory for discriminating solutions of an algebraic equation as a first step toward a general theory of ambiguity, that was developed further by Riemann, Picard, Vessiot, Lie, Poincare and Cartan, for systems of differential equations; it is also a theory of information. In another direction Rene Thom claimed that information must have a topological content (see [3]); he gave the example of the unfolding of the coupling of two dynamical systems, but he had in mind the whole domain of algebraic or differential topology.

All these approaches have in common the definition of secondary objects, either functions, groups or homology cycles, for measuring in what sense a pair of objects departs from independency. For instance, in the case of Shannon, the mutual information is $I(X;Y) = H(X) + H(Y) - H(X,Y)$, where $H$ denotes the usual Gibbs entropy ($H(X) = -\sum_x P(X = x) \ln_2 P(X = x)$), and for Galois it is the quotient set $IGal(L_1; L_2|K) = (Gal(L_1|K) \times Gal(L_2|K))/Gal(L|K)$, where $L_1, L_2$ are two fields containing a field $K$ in an algebraic closure $\Omega$ of $K$, where $L$ is the field generated by $L_1$ and $L_2$ in $\Omega$, and where $Gal(L_i|K)$ (for $i = \emptyset, 1, 2$) denotes the group introduced by Galois, made by the field automorphisms of $L_i$ fixing the elements of $K$.

We suggest that all information quantities are of co-homological nature, in a setting which depends on a pair of categories (*cf.* [4,5]); one for the data on a system, like random variables or functions of solutions of an equation, and one for the parameters of this system, like probability laws or coefficients of equations; the first category generates an algebraic structure like a monoid, or more generally a monad (*cf.* [4]), and the second category generates a representation of this structure,

as do for instance conditioning, or adding new numbers; then information quantities are co-cycles associated with this module.

We will see that, given a set of random variables on a finite set $\Omega$ and a simplicial subset of probabilities on $\Omega$, the *entropy* appears as the only one universal co-homology class of degree one. The higher mutual information functions that were defined by Shannon are co-cycles (or twisted co-cycles for even orders), and they correspond to higher homotopical constructions. In fact this description is equivalent to the theorem of Hu Kuo Ting [6], that gave a set theoretical interpretation of the mutual information decomposition of the total entropy of a system. Then we can use information co-cycles to describe forms of the information distribution between a set of random data; figures like ordinary links, or chains or Borromean links appear in this context, giving rise to a new kind of topology.

### *1.2. Information Homology*

Here we call random variables (r.v) on a finite set $\Omega$ *congruent* when they define the same partition (remind that a partition of $\Omega$ is a family of disjoint non-empty subsets covering $\Omega$ and that the partition associated to a r.v $X$ is the family of subsets $\Omega_x$ of $\Omega$ defined by the equations $X(\omega) = x$); the join r.v $YZ$, also denoted by $(Y, Z)$, corresponds to the less fine partition that is finer than $Y$ and $Z$. This defines a monoid structure on the set $\Pi(\Omega)$ of partitions of $\Omega$, with $1$ as a unit, and where each element is idempotent, *i.e.*, $\forall X, XX = X$. An *information category* is a set $\mathcal{S}$ of r.v such that, for any $Y, Z \in \mathcal{S}$ less fine than $U \in \mathcal{S}$, the join $YZ$ belongs to $\mathcal{S}$, *cf.* [7]. An ordering on $\mathcal{S}$ is given by $Y \leq Z$ when $Z$ refines $Y$, which also defines the morphisms $Z \to Y$ in the category $\mathcal{S}$. In what follows we always assume that $1$ belongs to $\mathcal{S}$. The simplex $\Delta(\Omega)$ is defined as the set of families of numbers $\{p_\omega; \omega \in \Omega\}$, such that $\forall \omega, 0 \leq p_\omega \leq 1$ and $\sum_\omega p_\omega = 1$; it parameterizes all probability laws on $\Omega$. We choose a simplicial sub-complex $\mathcal{P}$ in $\Delta(\Omega)$, which is stable by all the conditioning operations by elements of $\mathcal{S}$. By definition, for $N \in \mathbb{N}$, an information $N$-cochain is a family of measurable functions of $P \in \mathcal{P}$, with values in $\mathbb{R}$ or $\mathbb{C}$, indexed by the sequences $(S_1; ...; S_N)$ in $\mathcal{S}$ majored by an element of $\mathcal{S}$, whose values depend only of the image law $(S_1, ..., S_N)_* P$. This condition is natural from a topos point of view, *cf.* [4]; we interpret it as a "locality" condition. Note that we write $(S_1; ...; S_N)$ for a sequence, because $(S_1, ..., S_N)$ designates the joint variable. For $N = 0$ this gives only the constants. We denote by $\mathcal{C}^N$ the vector space of $N$-cochains of information. The following formula corresponds to the *averaged conditioning* of Shannon [1]:

$$S_0.F(S_1; ...; S_N; \mathbb{P}) = \sum \mathbb{P}(S_0 = v_j) F(S_1; ...; S_N; \mathbb{P}|S_0 = v_j), \tag{1}$$

where the sum is taken over all values of $S_0$, and the vertical bar is ordinary conditioning. It satisfies the associativity condition $(S'_0 S_0).F = S'_0.(S_0.F)$.

The coboundary operator $\delta$ is defined by

$$\begin{aligned} & \delta F(S_0; ...; S_N; \mathbb{P}) \qquad (2) \\ = \;& S_0.F(S_1; ...; S_N; \mathbb{P}) + \sum_0^{N-1} (-1)^{i+1} F(...; (S_i, S_{i+1}); ...; S_N; \mathbb{P}) + (-1)^{N+1} F(S_0; ...; S_{N-1}; \mathbb{P}), \end{aligned}$$

It corresponds to a standard non-homogeneous bar complex (*cf.* [5]). Another co-boundary operator on $\mathcal{C}^N$ is $\delta_t$ ($t$ for twisted or trivial action or topological complex), that is defined by the above formula with the first term $S_0.F(S_1;...;S_N;\mathbb{P})$ replaced by $F(S_1;...;S_N;\mathbb{P})$. The corresponding co-cycles are defined by the equations $\delta F = 0$ or $\delta_t F = 0$, respectively. We easily verify that $\delta \circ \delta = 0$ and $\delta_t \circ \delta_t = 0$; then co-homology $H^*(\mathcal{S};\mathcal{P})$ resp. $H_t^*(\mathcal{S};\mathcal{P})$ is defined by taking co-cycles modulo the elements of the image of $\delta$ resp. $\delta_t$, called co-boundaries. The fact that classical entropy $H(X;\mathbb{P}) = -\sum_i p_i \log_2 p_i$ is a 1-co-cycle is the fundamental equation $H(X,Y) = H(X) + X.H(Y)$.

**Theorem A.** (*cf.* Theorem 1 section 2.3, [7]): For the full simplex $\Delta(\Omega)$, and if $\mathcal{S}$ is the monoid generated by a set of at least two variables, such that each pair takes at least four values, then the information co-homology space of degree one is one-dimensional and generated by the classical entropy.

**Problem 1.** Compute the homology of higher degrees.

We conjecture that for binary variables it is zero, but that in general non-trivial classes appear, deduced from polylogarithms. This could require us to connect with the works of Dupont, Bloch, Goncharov, Elbaz-Vincent, Gangl *et al.* on motives (*cf.* [8]), which started from the discovery of Cathelineau (1988) that entropy appears in the computation of the degree one homology of the discrete group $SL_2$ over $\mathbb{C}$ with coefficients in the adjoint action (*cf.* [9]).

Suppose $\mathcal{S}$ is the monoid generated by a finite family of partitions. The higher mutual informations were defined by Shannon as alternating sums:

$$I_N(S_1;...;S_N;\mathbb{P}) = \sum_{k=1}^{k=N} (-1)^{k-1} \sum_{I\subset[N];card(I)=k} H(S_I;\mathbb{P}), \tag{3}$$

where $S_I$ denotes the join of the $S_i$ such that $i \in I$. We have $I_1 = H$ and $I_2 = I$ is the usual mutual information: $I(S;T) = H(S) + H(T) - H(S,T)$ .

**Theorem B.** (*cf.* section 3, [7]): $I_{2m} = \delta_t\delta\delta_t...\delta\delta_t H$, $I_{2m+1} = -\delta\delta_t\delta\delta_t...\delta\delta_t H$, where there are $m-1$ $\delta$ and $m$ $\delta_t$ factors for $I_{2m}$ and $m$ $\delta$ and $m$ $\delta_t$ factors for $I_{2m+1}$.

Thus odd information quantities are information co-cycles, because they are in the image of $\delta$, and even information quantities are twisted (or topological) co-cycles, because they are in the image of $\delta_t$.

In [7] we show that this description is equivalent to the theorem of Hu Kuo Ting (1962) [6], giving a set theoretical interpretation of the mutual information decomposition of the total entropy of a system: mutual information, join and averaged conditioning correspond respectively to intersection, union and difference $A\backslash B = A \cap B^c$ . In special cases we can interpret $I_N$ as homotopical algebraic invariants. For instance for $N = 3$, suppose that $I(X;Y) = I(Y;Z) = I(Z;X) = 0$, then $I_3(X;Y;Z) = -I((X,Y);Z)$ can be defined as a Milnor invariant for links, generalized by Massey, as they are presented in [10] (*cf.* page 284), through the 3-ary obstruction to associativity of products

in a subcomplex of a differential algebra, *cf.* [7]. The absolute minima of $I_3$ correspond to Borromean links, interpreted as synergy, *cf.* [11,12].

*1.3. Extension to Quantum Information*

Positive hermitian $n \times n$-matrices $\rho$, normalized by $Tr(\rho) = 1$, are called density of states (or density operators) and are considered as quantum probabilities on $E = \mathbb{C}^n$. Real quantum observables are $n \times n$ hermitian matrices, and, by definition, the amplitude, or expectation, of the observable $Z$ in the state $\rho$ is given by the formula $\mathbb{E}(Z) = Tr(Z\rho)$ (see e.g., [13]). Two real observables $Y, Z$ are said *congruent* if their eigenspaces are the same, thus orthogonal decomposition of $E$ are the quantum analogs of partitions. The join is well defined for commuting observables. An information structure **S** is given by a subset of observables, such that, if $Y, Z$ have common refined eigenspaces decomposition in **S**, their join $(Y, Z)$ belongs to **S**. We assume that $\{E\}$ belongs to **S**. What plays the role of a probability functor is a map **Q** from **S** to sets of positive hermitian forms on $E$, which behaves naturally with respect to the quantum direct image, thus **Q** is a covariant functor. We define information $N$-cochains as for the classical case, starting with the numerical functions on the sets $\mathbf{Q}_X; X \in \mathbf{S}$, which behave naturally under direct images.

The restriction of a density $\rho$ by an observable $Y$ is $\rho_Y = \sum_A E_A^* \rho E_A$, where the $E_A$'s are the spectral projectors of the observable $Y$. The functor **Q** is said to match **S** (or to be complete and minimal with respect to **S**) if, for each $X \in \mathbf{S}$, the set $\mathbf{Q}_X$ is the set of all possible densities of the form $\rho_X$.

The action of a variable on the cochains space $\mathcal{C}_Q^*$ is given by the *quantum averaged conditioning*:

$$Y.F(Y_0; ...; Y_m; \rho) = \sum_A Tr(E_A^* \rho E_A) F(Y_0; ...; Y_m; E_A^* \rho E_A) \tag{4}$$

>From here we define coboundary operators $\delta_Q$ and $\delta_{Qt}$ by the formula (22), then the notions of co-cycles, co-boundaries and co-homology classes follow. We have $\delta_Q \circ \delta_Q = 0$ and $\delta_{Qt} \circ \delta_{Qt} = 0$; *cf.* [7].

When the unitary group $U_n$ acts transitively on **S** and **Q**, there is a notion of invariant cochains, forming a subcomplex of information cochains, and giving a more computable co-homology than the brut information co-homology. We call it the invariant information co-homology and denote it by $H_U^*(\mathbf{S}; \mathbf{Q})$.

The *Von-Neumann entropy* of $\rho$ is $S(\rho) = \mathbb{E}_\rho(-\log_2(\rho)) = -Tr(\rho \log_2(\rho))$; it defines a 0-cochain $S_Y$ by restricting $S$ to the sets $\mathbf{Q}_X$. The classical entropy is $H(Y; \rho) = -\sum_A Tr(E_A^* \rho E_A) \log_2(Tr(E_A^* \rho E_A))$. Both these co-chains are invariant. It is well known that $S_{(X,Y)}(\rho) = H(X; \rho) + X.S_Y(\rho)$ when $X, Y$ commute, *cf.* [13]. In particular, by taking $Y = 1_E$ we see that classical entropy measures the default of equivariance of the quantum entropy, *i.e.*, $H(X; \rho) = S_X(\rho) - (X.S)(\rho)$. But using the case where $X$ refines $Y$, we obtain that the entropy of Shannon is the co-boundary of (minus) the Von Neumann entropy.

**Theorem C.** (*cf.* Theorem 3 section 4.3): For $n \geq 4$ and when **S** is generated by at least two decompositions such that each pair has at least four subspaces, and when **Q** is matching **S**, the

invariant co-homology $H^1_U$ of $\delta_Q$ in degree one is zero, and the space $H^0_U$ is of dimension one. In particular, the only invariant $0$-cochain such that $\delta S = -H$ is the Von Neumann entropy.

(This statement, which will be proved below, corrects a similar statement which was made in the announcement [14].)

*1.4. Concavity and Convexity Properties of Information Quantities*

The simplest classical information structure $\mathcal{S}$ is the monoid generated by a family of "elementary" *binary* variables $S_1, ..., S_n$. It is remarkable that in this case, the information functions $I_{N,J} = I_N(S_{j_1}; ...S_{j_N})$ over all the subsets $J = \{j_1, ..., j_N\}$ of $[n] = \{1, ..., n\}$, different from $[n]$ itself, give algebraically independent functions on the probability simplex $\Delta(\Omega)$ of dimension $2^n - 1$. They form coordinates on the quotient of $\Delta(\Omega)$ by a finite group.

Let $\mathcal{L}_d$ denotes the Lie derivative with respect to $d = (1, ..., 1)$ in the vector space $\mathbb{R}^{2^n}$, and $\triangle$ the Euclidian Laplace operator on $\mathbb{R}^{2^n}$, then $\Delta = \triangle - 2^{-n}\mathcal{L}_d \circ \mathcal{L}_d$ is the Laplace operator on the simplex $\Delta(\Omega)$ defined by equating the sum of coordinates to 1.

**Theorem D.** (*cf.* [15]): On the affine simplex $\Delta(\Omega)$ the functions $I_{N,J}$ with $N$ odd (resp. even) satisfies the inequality $\Delta I_N \geq 0$ (resp. $\Delta I_N \leq 0$).

In other terms, for $N$ odd the $I_{N,J}$ are super-harmonic which is a kind of weak concavity and for $N$ even they are sub-harmonic which is a kind of weak convexity. In particular, when $N$ is even (resp. odd) $I_{N,J}$ has no local maximum (resp. minimum) in the interior of $\Delta(\Omega)$.

**Problem 2.** What can be said of the other critical points of $I_{N,J}$? What can be said of the restriction of one information function on the intersection of levels of other information functions? Information topology depends on the shape of these intersections and on the Morse theory for them.

*1.5. Monadic Cohomology of Information*

Now we consider the category $\mathcal{S}^*$ of *generalized ordered partitions* of $\Omega$ over $\mathcal{S}$: they are sequences $S = (E_1, ..., E_m)$ of subsets of $\Omega$ such that $\cup_j E_j = \Omega$ and $E_i \cap E_j = \emptyset$ as soon as $i \neq j$. The number $m$ is named the degree of $S$. Note the important technical point that some of the sets $E_j$ can be the empty set. In the same spirit we introduce *generalized ordered orthogonal decompositions* of $E$ for the quantum case; but in this summary, for simplicity we restrict ourselves to the classical case. Also we forget to add generalized to ordered up to now in this summary. A *rooted tree decorated* by $\mathcal{S}^*$ is an oriented finite planar tree $\Gamma$, with a marked initial vertex $s_0$, named the *root* of $\Gamma$, where each vertex $s$ is equipped with an element $F_s$ of $\mathcal{S}^*$, such that edges issued from $s$ correspond to the values of $F_s$. When we want to mention that we restrict to partitions less fine than a partition $X$ we put an index $X$, like in $\mathcal{S}^*_X$.

The notation $\mu(m; n_1, ..., n_m)$ denotes the operation which associates to an ordered partition $S$ of degree $m$ and to $m$ ordered partitions $S_i$ of respective degrees $n_i$, the ordered partition that is obtained

by cutting the pieces of $S$ using the pieces of $S_i$ and respecting the order. An evident unit element for this operation is the unique partition $\pi_0$ of degree 1. The symbol $\mu_m$ denotes the collection of those operations for $m$ fixed. The introduction of empty subsets in ordered partitions insures that the result of $\mu(m; n_1, ..., n_m)$ is a partition of length $n_1 + ... + n_m$, thus the $\mu_m$ do define what is named an *operad*; *cf.* [10,16]. The axioms of unity, associativity and covariance for permutations are satisfied. See [10,16–18] for the definition of operads.

The most important algebraic object which is associated to an operad is a *monad* (*cf.* [4,16]), *i.e.*, a functor $\mathcal{V}$ from a category $\mathcal{A}$ to itself, equipped with two natural transformations $\mu : \mathcal{V} \circ \mathcal{V} \to \mathcal{V}$ and $\eta : \mathbb{R} \to \mathcal{V}$, which satisfy to the following axioms:

$$\mu \circ (\mathcal{V}\mu) = \mu \circ (\mu\mathcal{V}), \quad \mu \circ (\mathcal{V}\eta) = Id = \mu \circ (\eta\mathcal{V}) \tag{5}$$

In our situation, we can apply the Schur construction (*cf.* [16]) to the $\mu_m$ to get a monad: take for $V$ the real vector space freely generated by $\mathcal{S}^*$; it is naturally graded, so it is the direct sum of spaces $V(m); m \geq 1$ where the symmetric group $\mathfrak{S}_m$ acts naturally to the right, then introduce, for any real vector space $W$ the real vector space $\mathcal{V}(W) = \bigoplus_{m\geq 0} V(m) \otimes_{\mathfrak{S}_m} W^{\otimes m}$; the Schur composition is defined by $\mathcal{V}\circ\mathcal{V} = \bigoplus_{m\geq 0} V(m)\otimes_{\mathfrak{S}_m} \mathcal{V}^{\otimes m}$. It is easy to verify that the collection $(\mu_m; m \in \mathbb{N})$ defines a natural transformation $\mu : \mathcal{V} \circ \mathcal{V} \to \mathcal{V}$, and the trivial partition $\pi_0$ defines a natural transformation $\eta : \mathbb{R} \to \mathcal{V}$, that satisfied to the axioms of a monad.

Also we fix a functor of probability laws $Q_X$ over the category $\mathcal{S}$. Let $\mathcal{M}_X(m)$ be the vector space freely generated over $\mathbb{R}$ by the symbols $(P, i, m)$ where $P$ belongs to $Q_X$, and $1 \leq i \leq m$. In the last section of the second part we show how this space arises from the consideration of divided probabilities. This is apparent on the following definition of the right action of the operad $\mathcal{V}$ on the family $\mathcal{M}_X(m); m \in \mathbb{N}^*$: a sequence $S_1, ..., S_m$ or ordered partitions in $\mathcal{S}_X^*$ acts to a generator $(P, i, m)$ by giving the vector $\sum_j p_j(P_j, (i, j), n)$ where $p_j$ is the probability $P(S_i = j)$ and $P_j$ is the conditioned probability $P|(S_i = j)$. We denote by $\theta_m((P, i, m), (S_1, ..., S_m))$ this vector.

Now we consider the Schur functor $\mathcal{M}_X(W) = \bigoplus_m \mathcal{M}_X(m) \otimes_{\mathfrak{S}_m} W^{\otimes m}$; the operations $\theta_m$ define a natural transformation $\theta : \mathcal{M} \circ \mathcal{V} \to \mathcal{M}$, which is an action to the right in the sense of monads, *i.e.*, $\theta \circ (\mathcal{F}\mu) = \theta \circ (\theta\mathcal{V})$; $\quad \theta \circ (\mathcal{F}\eta) = Id$. (We forgot the index $X$ for simplicity.)

Now we consider the bar resolution of $\mathcal{M}$: .... $\to \mathcal{M} \circ \mathcal{V}^{\circ(k+1)} \to \mathcal{M} \circ \mathcal{V}^{\circ k} \to$ ..., as in Beck (triples, ...) [19], and Fresse [16], with its simplicial structure deduced from $\theta$ and $\mu$, and the complex of natural transformations of $\mathcal{V}$-right modules $C^*(\mathcal{M}) = Hom_V(\mathcal{M} \circ \mathcal{V}^{\circ *}, \mathcal{R})$, where $\mathcal{R}$ is the trivial right module given by $\mathcal{R}(m) = \mathbb{R}$. As in the classical case, we restrict us to co-chains that are measurable in the probability $(P, i, m)$.

The co-boundary is defined by the Hochschild formula, extended by MacLane and Beck to monads (see Beck [19]):

$$\delta F = F \circ (\theta\mathcal{V}^{\circ k}) - \sum_{i=0,...,k-1} (-1)^i F \circ \mathcal{M}\mathcal{V}^{\circ i}\mu\mathcal{V}^{\circ k-i-1} - (-1)^k F \circ \mathcal{M}\mathcal{V}^{\circ k}\epsilon. \tag{6}$$

The cochains are described by families of scalar measurable functions $F_X(S_1; ..., S_k; (P, i, m)$, where $S_1; ...; S_k$ is a forest of $m$ trees of level $k$ labelled by $\mathcal{S}_X^*$, and where the value on $(P, i, m)$ depends only on the tree $S_1^i; S_2^i; ...; S_k^i$.

We impose now the condition, named *regularity*, that $F_X(S_1;...,S_k;(P,i,m)) = F_X(S_1^i;S_2^i;...;S_k^i;P)$. The regular co-chains form a sub-complex $C_r^*(\mathcal{M})$; by definition, its homology is the *arborescent information co-homology*.

The regular cochains of degree $k$ are determined by their values for $m=1$ and decorated trees of level $k$, where the co-boundary takes the form:

$$\begin{aligned}&\delta F(S;S_1;...;S_k;\mathbb{P})\\ =\ &\sum_i \mathbb{P}(S=i)F(S_1^i;...;S_k^i;\mathbb{P}|(S=i)) + \sum_{i=1}^{i=k}(-1)^i F(S;...;\mu(S_{i-1}\circ S_i);S_{i+1};...;S_k;\mathbb{P}) \quad (7)\\ &+(-1)^{k+1}F(S;...;S_{k-1};\mathbb{P})\end{aligned}$$

This gives co-homology groups $H_\tau^*(\mathcal{S},\mathcal{P})$, $\tau$ for tree. The fact that entropy $H(S_*\mathbb{P}) = H(S;\mathbb{P})$ defines a 1-cocycle is a result of an equation of Fadeev, generalized by Baez, Fritz and Leinster [20], who gave another interpretation, based on the operad structure of the set of all finite probability laws. See also Marcolli and Thorngren [21].

**Theorem E.** (*cf.* Theorem 4 section 6.3, [22]): If $\Omega$ has more than four points, $H_\tau^1(\Pi(\Omega),\Delta(\Omega))$ is the one dimensional vector space generated by the entropy.

Another co-boundary $\delta_t$ on $\mathcal{C}_r^*(\mathcal{M})$ corresponds to another right action of the monad $\mathcal{V}_X$, which is deduced from the maps $\theta_t$ that send $(P,i,m)\otimes S_1\otimes...\otimes S_m)$ to the sum of the vectors $(P,(i,j),n)$ for $j=1,...,n_i$ that are associated to the end branches of $S_i$. It gives a twisted version of information co-homology as we have done in the first paragraph. This allows us to define higher information quantities for strategies: for $N=2M+1$ odd, $I_{\tau,N} = -(\delta\delta_t)^M H$, and for $N=2M+2$ even, $I_{\tau,N} = \delta_t(\delta\delta_t)^M H$.

This gives for $N=2$, a notion of mutual information between a variable $S$ of length $m$ and a collection $T$ of $m$ variables $T_1,...,T_m$:

$$I_\tau(S;T;\mathbb{P}) = \sum_{i=1}^{i=m}(H(T_i;\mathbb{P}) - \mathbb{P}(S=i)H(T_i;\mathbb{P}|S=i)). \quad (8)$$

When all the $T_i$ are equals we recover the ordinary mutual information of Shannon plus a multiple of the entropy of $T_i$.

### *1.6. The Forms of Information Strategies*

A rooted tree $\Gamma$ decorated by $\mathcal{S}_*$ can be seen as a strategy to discriminate between points in $\Omega$. For each vertex $s$ there is a minimal set of chained edges $\alpha_1,...,\alpha_k$ connecting $s_0$ to $s$; the cardinal $k$ is named the *level* of $s$; this chain defines a sequence $(F_0,v_0;F_1,v_1;...;F_{k-1},v_{k-1})$ of observables and values of them; then we can associate to $s$ the subset $\Omega_s$ of $\Omega$ where each $F_j$ takes the value $v_j$. At a given level $k$ the sets $\Omega_s$ form a partition $\pi_k$ of $\Omega$; the first one $\pi_0$ is the unit partition of length 1, and $\pi_l$ is finer than $\pi_{l-1}$ for any $l$. By recurrence over $k$ it is easy to deduce from the orderings

of the values of $F_s$ an embedding in the Euclidian plane of the subtrees $\Gamma(k)$ at level $k$ such that the values of the variables issued from each vertex are oriented in the direct trigonometric sense, thus $\pi_k$ has a canonical ordering $\omega_k$. Remark that many branches of the tree gives the empty set for $\Omega_s$ after some level; we name them dead branches. It is easy to prove that the set $\Pi(\mathcal{S})_*$ of ordered partitions that can be obtained as a $(\pi_k, \omega_k)$ for some tree $\Gamma$ and some level $k$ is closed by the natural ordered join operation, and, as $\Pi(\mathcal{S})_*$ contains $\pi_0$, it forms a monoid, which contains the monoid $M(\mathcal{S}_*)$ generated by $\mathcal{S}_*$.

Complete discrimination of $\Omega$ by $\mathcal{S}_*$ exists when the final partition of $\Omega$ by singletons is attainable as a $\pi_k$; optimal discrimination correspond to minimal level $k$. When the set $\Omega$ is a subset of the set of words $x_1, ..., x_N$ with letters $x_i$ belonging to given sets $M_i$ of respective cardinalities $m_i$, the problem of optimal discrimination by observation strategies $\Gamma$ decorated by $\mathcal{S}_*$ is equivalent to a problem of minimal rewriting by words of type $(F_0, v_0), (F_1, v_1), ..., (F_k, v_k)$; it is a variant of optimal coding, where the alphabet is given. The topology of the poset of discriminating strategies can be computed in terms of the free Lie algebra on $\Omega$, *cf.* [16].

Probabilities $\mathbb{P}$ in $\mathcal{P}$ correspond to *a priori* knowledge on $\Omega$. In many problems $\mathcal{P}$ is reduced to one element, that is the uniform law. Let $s$ be a vertex in a strategic tree $\Gamma$, and let $\mathcal{P}_s$ be the set of probability laws that are obtained by conditioning through the equations $F_i = v_i; i = 0, ..., k-1$ for a minimal chain leading from $s_0$ to $s$. We can consider that the sets $\mathcal{P}_s$ for different $s$ along a branch measure the evolution of knowledge when applying the strategy. The entropy $H(F; \mathbb{P}_s)$ for $F$ in $\mathcal{S}_*$ and $\mathbb{P}_s$ in $\mathcal{P}_s$ gives a measure of information we hope to obtain when applying $F$ at $s$ in the state $\mathbb{P}_s$. The maximum entropy algorithm consists in choosing at each vertex $s$ a variable that has the maximal conditioned entropy $H(F; \mathbb{P}_s)$.

**Theorem F.** (*cf.* [22]): To find one false piece of different weight among $N$ pieces for $N \geq 3$, when knowing the false piece is unique, by the minimal numbers of weighing, one can use the maximal entropy algorithm.

However we have another measure of information of the resting ambiguity at $s$, by taking for the *Galois group* $G_s$ the set of permutations of $\Omega_s$ which respects globally the set $\mathcal{P}_s$ and the set of restrictions of elements of $\mathcal{S}_*$ to $\Omega_s$, and which preserves one by one the equations $F_i = v_i$. Along branches of $\Gamma$ this gives a decreasing sequence of groups, whose successive quotients measure the evolution of acquired information in an algebraic sense.

**Problem 3.** Generalize Theorem F. Can we use algorithms based on the Galoisian measure of information? Can we use higher information quantities associated to trees for optimal discrimination?

### *1.7. Conclusion and Perspective*

Concepts of Algebraic topology were recently applied to Information theory by several researchers. In particular notions coming from category theory, homological algebra and differential geometry were used for revisiting the nature and scope of entropy, *cf.* for instance Baez *et al*.

[20], Marcolli and Thorngren [21] and Gromov [23]. In the present note we interpreted entropy and Shannon information functions as co-cycles in a natural co-homology theory of information, based on categories of observable and complexes of probability. This allowed us to associate topological figures, like Borromean links, with particular configuration of mutual dependency of several observable quantities. Moreover we extended these results to a dynamical setting of system observation, and we connected probability evolutions with the measures of ambiguity given by Galois groups. All those results provide only the first steps toward a developed Information Topology. However, even at this preliminary stage, this theory can be applied to the study of distribution and evolution of Information in concrete physical and biological systems. This kind of approach already proved its efficiency for detecting collective synergic dynamic in neural coding [12], in genetic expression [24], in cancer signature [25], or in signaling pathways [26]. In particular, information topology could provide the principles accounting for the structure of information flows in biological systems and notably in the central nervous system of animals.

## 2. Classical Information Topos. Theorem One

### *2.1. Information Structures and Probability Families*

Let $\Omega$ be a finite set, the set $\Pi(\Omega)$ of all partitions of $\Omega$ constitutes a category with one arrow $Y \to Z$ from $Y$ to $Z$ when $Y$ is more fine than $Z$, we also say in this case that $Y$ divides $Z$. In $\Pi(\Omega)$ we have an initial element, which is the partition by points, denoted $\omega$ and a final element, which is $\Omega$ itself and is denoted by $1$. The joint partition $YZ$ or $(Y, Z)$, of two partitions $Y, Z$ of $\Omega$ is the less fine partition that divides $Y$ and $Z$, *i.e.*, their gcd. For any $X$ we get $XX = X$, $\omega X = \omega$ and $1.X = X$.

By definition an *information structure* $\mathcal{S}$ on $\Omega$ is a subset of $\Pi(\Omega)$, such that for any element $X$ of $\mathcal{S}$, and any pair of elements $Y, Z$ in $\mathcal{S}$ that $X$ refines, the joint partition $YZ$ also belongs to $\mathcal{S}$.
In addition we will always assume that the final partition $1$ belongs to $\mathcal{S}$. In terms of observations, it means that at least something is a certitude.

Examples: start with a set $\Sigma = \{S_i; 1 \leq i \leq n\}$ of partitions of $\Omega$. For any subset $I = \{i_1, ..., i_k\}$ of $[n] = \{1, ..., n\}$, the joint $(S_{i_1}, ..., S_{i_k})$, also denoted $S_I$, divides each $S_{i_j}$. The set $W = W(\Sigma)$ of all the $S_I$, when $I$ describes the subsets of $[n]$ is an information struture. It is even a commutative monoid, because any product of elements of $W$ belongs to $W$, and the partition associated to $\Omega$ itself gives the identity element of $W$. The product $S_{[n]}$ of all the $S_i$ is maximal; it divides all the other elements. As $\Pi(\Omega)$ the monoid $W(\Sigma)$ is idempotent, *i.e.*, for any $X$ we have $XX = X$.

By definition, the faces of the abstract simplex $\Delta([n])$ are the subsets of $[n]$; its vertices are the singletons. Thus the monoid $W(\Sigma)$ can be identified with the first barycentric subdivision of the simplex $\Delta([n])$.

Remind that a simplicial subcomplex of $\Delta([n])$ is a subset of faces that contains all faces of any of its elements. Then any simplicial sub-complex $K$ of $\Delta([n])$ gives a simplicial information structure $\mathcal{S}(K)$, embedded in $W(\Sigma)$. In fact, if $Y$ and $Z$ are faces of a simplex $X$ belonging to $K$, $YZ$ is also a face in $X$, thus it belongs to $K$. The maximal faces $\Sigma_a; a \in A$ of $K$ correspond to the finest

elements in $\mathcal{S}(K)$; the vertices of a face $\Sigma_a$ gives a family of partitions, which generates a sub-monoid $W_a = W(\Sigma_a)$ of $W$; it is a sub-information structures (full sub-category) of $\mathcal{S}(K)$, having the same unit, but having its own initial element $\omega_a$. These examples arise naturally when formalizing measurements if some obstructions or *a priori* decisions forbid a set of joint measurements.

This kind of examples were considered by Han [27] see also McGill [28].

**Example 1.** $\Omega$ has four elements $(00), (01), (10), (11)$; the variable $S_1$ (resp. $S_2$) is the projection $pr_1$ (resp. $pr_2$), on $E_1 = E_2 = \{0, 1\}$; $\Sigma$ is the set $\{S_1, S_2\}$. The monoid $W(\Sigma)$ has four elements $1, S_1, S_2, S_1S_2$. The partition $S_1S_2 = S_2S_1$ corresponds to the variable $Id : \Omega \to \Omega$.

**Example 2.** Same $\Omega$ as before, with the same names for the elements, but we take all the partitions of $\Omega$ in $\mathcal{S}$. In addition to $1, S_1, S_2$ and $S = S_1S_2$, there is $S_3$, the last partition in two subsets of cardinal two, which can be represented by the sum of the indices: $S_3(00) = 0, S_3(11) = 0, S_3(01) = 1, S_3(10) = 1$, the four partitions $Y_\omega$, for $\omega \in \Omega$, formed by a singleton $\{\omega\}$ and its complementary, and finally the six partitions $X_{\mu,\nu} = Y_\mu Y_\nu$, indexed by pairs of points in $\Omega$ satisfying $\mu < \nu$ in the lexical order. The product of two distinct $Y$ is a $X$, the product of two distinct $X$ or two distinct $S_i$ is $S$, the product of one $Y$ and a $S_i$ is a $X$, of one $Y$ and a $X$ is this $X$ or $S$, of one $S$ and a $X$ is this $X$ or $S$. In particular the monoid $W$ is also generated by the three $S_i$ and the four $Y_\omega$; it is called the monoid of partitions of $\Omega$, and the associative algebra $\Lambda(\mathcal{S})$ of this monoid is called the partition algebra of $\Omega$.

**Example 3.** Same $\Omega$ as before, that is $\Omega = \Delta(4)$, with the notations of example 2 for the partitions; but we choose as generating family the set $\Upsilon$ of the four partitions $Y_\mu$; $\mu \in \Omega$; the joint product of two such partitions is either a $Y_\mu$ (when they coincide) or a $X_{\mu\nu}$ (when they are different). The monoid $W(\Upsilon)$ has twelve elements.

**Example 4.** $\Omega$ has $8$ elements, noted $(000), ..., (111)$, and we consider the family $\Sigma$ of the three binary variables $S_1, S_2, S_3$ given by the three projections. If we take all the joints, we have a monoid of eight elements. However, if we forbid the maximal face $(S_1, S_2, S_3)$, we have a structure $\mathcal{S}$ which is not a monoid; it is the set formed by $1$, $S_1, S_2, S_3$ and the three joint pairs $(S_1, S_2), (S_1, S_3), (S_2, S_3)$.

On the side of probabilities, we choose a Boolean algebra $\mathcal{B}$ of sets in $\Omega$, *i.e.*, a subset $\mathcal{B}$ of the set $\mathcal{P}(\Omega)$ of subsets of $\Omega$ that contains the empty set $\emptyset$ and the full set $\Omega$, and is closed by union and intersection. In this finite context, it is easy to prove that $\mathcal{B}$ is constituted by all the unions of its minimal elements (called atoms). Associated to this case, we will consider only information structures that are made by partitions whose each element belongs to $\mathcal{B}$. Consequently we could replace everywhere $\Omega$ by the finite set $\Omega_\mathcal{B}$ of the atoms of $\mathcal{B}$, but we will see that several Boolean sub-algebras appear naturally in the process of observation, thus we prefer to mention the choice of $\mathcal{B}$ at the beginning of observations.
Then we consider the set $\Delta(\Omega_\mathcal{B})$, or $\Delta(\mathcal{B})$, of all probability laws on $(\Omega, \mathcal{B})$, *i.e.*, all real functions

$p_x$ of the atoms $x$ of $\mathcal{B}$ (the points of $\Omega_{\mathcal{B}}$), satisfying $p_x \geq 0$ and $\sum_x p_x = 1$. We see that this set of probabilities is also a simplex $\Delta([N])$, where $N$ is the cardinality of $\Omega_{\mathcal{B}}$.

As on the side of partitions, we will consider more generally any simplicial sub-complex $\mathcal{Q}$ of $\Delta(\mathcal{B})$, and call it a probability complex. In the appendix, we show that this kind of examples correspond to natural forbidding rules, that can express physical constraints on the observed system.

A partition $Y$ which is measurable with respect to $\mathcal{B}$ is made by elements $Y_i$ for $i = 1, ..., m$, belonging to $\mathcal{B}$. Let $P$ be an element of $\Delta(\mathcal{B})$; the conditioning of $P$ by the element $Y_i$ is defined only if $P(Y_i) \neq 0$, and given by the formula $P(B|Y = y_i) = P(B \cap Y_i)/P(Y_i)$. We will consider it as a probability on $\Omega$ equipped with $\mathcal{B}$, not as a probability on $Y_i$. Remark that if $P$ belongs to a simplicial family $\mathcal{Q}$, the probability $P(B|Y = y_i)$ is also contained in $\mathcal{Q}$. In fact, if the smallest face of $\mathcal{Q}$ which contains $P$ is the simplex $\sigma$ on the vertices $x_1, ..., x_k$, then the conditioning of $P$ by $Y_i$, being equal to 0 for the other atoms $x$, belongs to a face of $\sigma$, which is in $\mathcal{Q}$, because $\mathcal{Q}$ is a complex.

For a probability family $\mathcal{Q}$, *i.e.*, a set of probabilities on $\Omega$, and a set of partitions $\mathcal{S}$, we say that $\mathcal{Q}$ and $\mathcal{S}$ are adapted one to each other if the conditioning of every element of $\mathcal{Q}$ by every element of $S$ belongs to $\mathcal{Q}$.

By definition, the algebra $\mathcal{B}_Y$ is the set of unions of elements of the partition $Y$. We can consider it as a Boolean algebra on $\Omega$ contained in $\mathcal{B}$ or as Boolean algebra on the quotient set $\Omega/Y$.
The image $Y_*Q$ of a probability $Q$ for $\mathcal{B}$ by the partition $Y$ is the probability on $\Omega$ for the sub-algebra $\mathcal{B}_Y$, that is given by $Y * Q(t) = Q(t)$ for $t \in \mathcal{B}_Y$. It is the forgetting operation, also frequently named marginalization by $Y$.

By definition, the set $\mathcal{Q}_Y$ is the image of $Y_*$. Let us prove that it is a simplicial sub-complex of $\Delta(\mathcal{B}_Y)$: take a simplex $\sigma$ of $\mathcal{Q}$, denote its vertices by $x_1, ..., x_k$, note $\delta_j$ the Dirac mass of $x_j$, and look at the partition $\sigma_i = Y_i \cap \sigma$ of $\sigma$ induced by $Y$, then for all the $x_j \in \sigma_i$ the images $Y_*\delta_j$ coincide. Let us denote this image by $\delta(Y, \sigma_i)$; it is an element of $\mathcal{Q}_Y$. For every law $Q$ in $\sigma$, the image $Y_*Q$ belongs to the simplex on the laws $\delta(Y, \sigma_i)$, and any point in this simplex belongs to $\mathcal{Q}_Y$. Q.E.D.

If $X \to Y$ is an arrow in $\Pi(\Omega_{\mathcal{B}})$, the above argument shows that the map $\mathcal{Q}_X \to \mathcal{Q}_Y$ is a simplicial mapping.

Conditioning by $Y$ and marginalization by $Y_*$ are related by the barycentric law (or theorem of total probability, Kolmogorov 1933 [29]): for any measurable set $A$ in $\mathcal{B}$ we have

$$P(A) = P(Y = y_1)P|(Y = y_1)(A) + ... + P(Y = y_m)P|(Y = y_m)(A). \tag{9}$$

Remark that the notions of information structures and probability complexes extend to infinite sets; this is developed in paper [7].

In this context, we have a formula for any integrable function $\varphi$ on $\Omega$ with respect to $P$:

$$\int_\Omega \varphi(\omega) dP(\omega) = \int_{\Omega/Y} d(Y_*P)(\omega') \int_\Omega \varphi(\omega) d(P|(Y = \omega'))(\omega). \tag{10}$$

Consider a finite set $\Omega$, equipped with a Boolean algebra $\mathcal{B}$, a probability family $\mathcal{Q}$ for it and an information structure $\mathcal{S}$ adapted to $\mathcal{B}$.

For each object $X$ in $\mathcal{S}$, the set $\mathcal{S}_X$ made by the partitions $Y$ that are divided by $X$ is a closed sub-category, possessing an internal law of monoid. The object $X$ is initial. To any arrow $X \to Y$

is associated the inclusion $\mathcal{S}_Y \to \mathcal{S}_X$, thus we get a contra-variant functor from $\mathcal{S}$ to the category of monoids.

On the other side we have a natural co-variant functor of $\mathcal{S}$ to the category of sets, which associates to each partition $X \in \mathcal{S}$ the set $\mathcal{Q}_X$ of probability laws in the image of $\mathcal{Q}$ on the quotient set $\Omega/X$, and which associates to each arrow $X \to Y$ the surjection $\mathcal{Q}_X \to \mathcal{Q}_Y$ which is given by direct image $P_X \mapsto Y_* P_X$. If $\mathcal{Q}$ is simplicial the functor goes to the category of simplicial complexes.

**Definition 1.** For $X \in \mathcal{S}$, the *functional module* $\mathcal{F}_X(\mathcal{Q})$ is the real vector space of measurable functions on the space $\mathcal{Q}_X$; for each arrow of divisibility $X \to Y$, we have an injective linear map $f \mapsto f^{Y|X}$ from $\mathcal{F}_Y$ to $\mathcal{F}_X$, given by

$$f^{Y|X}(P_X) = f(Y_* P_X). \tag{11}$$

In this manner, we obtain a contra-variant functor $\mathcal{F}$ from the category $\mathcal{S}$ to the category of real vector spaces.

If $\mathcal{Q}$ and $\mathcal{S}$ are adapted one to each other, the functor $\mathcal{F}$ admits a canonical action of the monoid functor $X \mapsto \mathcal{S}_X$, given by the average formula

$$(Y.f)(P) = \int dY_* P(y) f(P|(Y = y)). \tag{12}$$

To verify this is an action of monoid, we must verify that for any $Z$ which divides $Y$, and any $f \in \mathcal{F}_Y$, we have, in $\mathcal{F}_X$ the identity

$$(Z.f)^{Y|X} = Z.(f^{Y|X}); \tag{13}$$

that means, for any $P \in \mathcal{Q}_X$:

$$\int_{E_Z} dZ_* P(z) f^{Y|X}(P|(Z = z)) = \int_{E_Z} dZ_* P(z) f((Y_* P)|(Z = z)). \tag{14}$$

But this results from the identity $Y_*(P|(Z = z)) = (Y_* P)|(Z = z)$ due to $Y_* P(Z = z) = P(Z = z)$.

The arrows of direct images and the action of averaged conditioning satisfy the axiom of distributivity: if $Y$ and $Z$ divide $X$, but not necessarily $Z$ divides $Y$, we have

$$Z.(f^Y)(P, X) = (Z.f)((Z, Y)_* P, (Y, Z)) = (Z.f)^{(Z,Y)}(P, X). \tag{15}$$

**Proof.** The first identity comes from the fact that $(Z, Y)_*(P|(Z = z)) = Y_*(P|(Z = z))$; the second one follows from the fact that we have an action of the monoid $\mathcal{S}_X$.

As the formula (12) is central in our work, we insist a bit on it, and comment its meaning, at least in this finite setting:

Let $P \mapsto f(P)$ be an element of $\mathcal{F}_X$, and $Y$ be the goal of an arrow $X \to Y$, we have

$$Y.f(P) = \sum_j \mathbb{P}(Y = y_j) f(\mathbb{P}|Y = j). \tag{16}$$

where $j$ describes the indices of the partition $Y$.

We will see when discussing functions of several partitions that this formula is due to Shannon and correspond to conditional information.

**Lemma 1.** for any pair $(Y, Z)$ of variables in $\mathcal{S}_X$, and any $F$ for which the integrals converge, we have $(Y, Z).F = Y.(Z.F)$.

**Proof.** We note $p_i$ the probability that $Y = y_i$, $\pi_{ij}$ the joint probability of $(Y = y_i, Z = z_j)$, and $q_{ij}$ the conditional probability of $Z = z_j$ knowing that $Y = y_i$, then

$$
\begin{aligned}
&(Y, Z).F(\mathbb{P}) = \sum_i \sum_j \pi_{ij} F(\mathbb{P}|(Y = y_i, Z = z_j)) \\
&= \sum_i p_i (\sum_j q_{ij} F(\mathbb{P}|(Y = y_i, Z = z_j)) \\
&= \sum_i p_i (\sum_j q_{ij} F(\mathbb{P}|(Y = y_i))|(Z = z_j)) \\
&= \sum_i p_i (Z.F)(\mathbb{P}|(Y = y_i)) \\
&= Y.(Z.F)(\mathbb{P}).
\end{aligned}
$$

**Remark 1.** In the general case, where $\Omega$ is not necessarily finite and $\mathcal{B}$ is any sigma-algebra, the Lemma 1 is a version of the Fubini theorem.

Let us consider the category $\mathcal{S}$ equipped with the discrete topology, to get a site (*cf.* SGA [30]). Over a discrete site every presheaf is a sheaf. The contravariant functor $X \mapsto \mathcal{S}_X$ gives a structural sheaf of monoids, and by passing to the algebras $\mathcal{A}_X$ over $\mathbb{R}$ which are generated by the (finite) monoids, we get a sheaf in rings, thus $S$ becomes a ringed site. Moreover, by considering all contra-variant functors $X \mapsto \mathcal{N}_X$ from $\mathcal{S}$ to modules over the algebra functor $\mathcal{A}$, we obtain a ringed topos, that we name the *information topos* associated to $\Omega, \mathcal{B}, \mathcal{S}$. This ringed topos concerns only the observables given by partitioning.

Take now in account a probability family $\mathcal{Q}$ which is adapted to $\mathcal{S}$, for instance a simplicial family; we obtain a functor $X \mapsto Q_X$ translating the marginalization by the partitions, considered as observable quantities, and the conditioning by observables is translated by a special element $X \mapsto \mathcal{F}_X$ of the information topos.

In this way it is natural to expect that topos co-homology, as introduced by Grothendieck, Verdier and their collaborators (see SGA 4 [30]), captures the invariant structure of observation, and defines in this context what information is. This is the main outcome of our work.

As a consequence of Grothendieck's article (Tohoku, 1957 [31]), a ringed topos possesses enough injective objects, *i.e.*, any object is the sub-object of an injective object, moreover, up to isomorphism, there is a unique minimal injective object containing a given object, called its injective envelope (*cf.* Gabriel, seminaire Dubreil, exp. 17 [32]). Thus each object in the category $\mathcal{D}_\mathcal{S}$ of modules over a ringed site $\mathcal{S}$ possesses a canonical injective resolution $I_*(N)$; then the group $Ext^n_\mathcal{D}(M, N)$

can be defined as the homology of the complex $Hom_{\mathcal{D}}(M, I_n(N))$. Those groups are denoted by $H^n(M; N)$.

The "comparison theorem" (*cf.* Bourbaki, Alg.X Th1, p.100 [33], or MacLane 1975, p. 261 [5]) asserts that, for any projective (resp. injective) resolution of $M$ (resp. $N$) there exists a natural map of complexes between the resulting complex of homomorphisms and the above canonical complex, and that this map induces an isomorphism in co-homology.

In our context, we take for $M$ the trivial constant module $\mathbb{R}_{\mathcal{S}}$ over $\mathcal{S}$, and we take for $N$ the functional module $\mathcal{F}(\mathcal{Q})$.

The existence of free resolutions of $\mathbb{R}_{\mathcal{S}}$ makes things easier to handle.

Hence we propose that the natural information quantities are classes in the co-homology groups $H^*(\mathbb{R}_{\mathcal{S}}, \mathcal{F}(\mathcal{Q}))$.

This is reminiscent of Galois co-homology see SGA [30], where $M$ is also taken as the constant sheaf over the category of $G$-objects seen as a site.

In [7] we develop further this more geometric approach, by considering several resolutions. But in this paper, in order to be concrete, we will only focus on a more elementary approach, associated to a special resolution, called the non-homogeneous bar-resolution, which also leads to the general result. This is the object of the next section.

### *2.2. Non-Homogeneous Information Co-Homology*

For each relative integer $m \geq 0$, and each object $X \in \mathcal{S}$, we consider the real vector space $S_m(X)$, freely generated by the $m$-uples of elements of the monoid $\mathcal{S}_X$, and we define $C^m(X)$ as the real vector space of linear functions from $S_m(X)$ to the space $\mathcal{F}_X$ of measurable functions from $\mathcal{Q}_X$ to $\mathbb{R}$.

Then we define the set $\mathcal{C}^m$ of $m$-cochains as the set of collections $F_X \in C^m(X)$ satisfying the following condition, named *joint locality*:

For each $Y$ divided by $X$, when each variable $X_j$ is divided by $Y$, we must have

$$F_Y(X_1; ...; X_m; Y_*\mathbb{P}) = F_X(X_1; ...; X_m; \mathbb{P}). \tag{17}$$

Thus a co-chain $F$ is a natural transformation from the functor $S_m(X)$ from $\mathcal{S}$ to the category of real vector spaces to the functor $\mathcal{F}$ of measurable functions on $\mathcal{Q}_X$. Hence, $F$ is not an ordinary numerical function of probability laws $\mathbb{P}$ and a set $(X_1, ..., X_m)$ of $m$ random variables, but we can speak of its value $F_X(X_1; ...; X_m; \mathbb{P})$ for each $X$ in $\mathcal{S}$. For $X$ given the co-chains form a sub-vector space $\mathcal{C}^m(X)$ of $C^m(X)$.

If we apply the condition to $Y = (X_1, ..., X_m)$ we find that $F(X_1; ...; X_m; \mathbb{P})$ depends only on the direct image of $\mathbb{P}$ by the joint variable of the $X_i$'s. This implies that, if $F$ belongs to $\mathcal{C}^m(X)$, we have

$$F(X_1; ...; X_m; \mathbb{P}) = F(X_1; ...; X_m; (X_1...X_n)_*\mathbb{P}), \tag{18}$$

Conversely, suppose that $F$ satisfies the conditions (18) and consider $X, Y$ two variables such that $X$ divides $Y$, and that $Y$ divides each $X_j$, and let $P$ be a probability in $\mathcal{Q}_X$; then the joint variable $Z = (X_1, ..., X_m)$ divides $Y$ and $X$, thus we have $Z_*P = Z_*(X_*P) = Z_*(Y_*P)$, and

$$F(X_1; ...; X_m; Y_*P) = F(X_1; ...; X_m; Z_*P) = F(X_1; ...; X_m; X_*P). \tag{19}$$

Which proves that $F$ belongs to $\mathcal{C}^m(X)$.

Let $F$ be an element of $\mathcal{C}^m(X)$, and $Y$ an element of $\mathcal{S}_X$; then we define

$$Y.F(X_1; ...; X_m; \mathbb{P}) = \sum \mathbb{P}(Y = y_j)F(X_1; ...; X_m; \mathbb{P}|Y = y_j). \tag{20}$$

It follows from the equivalent condition (18) that $Y.F$ also belongs to $\mathcal{C}^m(X)$.

Moreover, the proof of Lemma 1 applies and give that, for any pair $(Y, Z)$ of variables in $\mathcal{S}_X$, and any $F$ in $\mathcal{C}^m(X)$, we have $(Y, Z).F = Y.(Z.F)$.

Thus (1) defines an action of the semigroup $\mathcal{S}_X$ on the vector spaces $\mathcal{C}^m(X)$.

**Remark 2.** The operation of $\mathcal{S}_X$ can be rewritten more compactly by using integrals:

$$Y.F(X_1; ...; X_m; \mathbb{P}) = \int_\Omega F(X_1; ...; X_m; \mathbb{P}|(Y = Y(\omega)))dP(\omega). \tag{21}$$

The differential $\delta$ for computing co-homology is given by the Eilenberg-MacLane formula (1943):

$$\begin{aligned}&\delta^m F(Y_1; ...; Y_{m+1}; P)\\ =&Y_1.F(Y_2; ...; Y_{m+1}; P) + \sum_1^m (-1)^i F(...; (Y_i, Y_{i+1}); ...; Y_{m+1}; P) + (-1)^{m+1} F(Y_1; ...; Y_m; P).\end{aligned} \tag{22}$$

Since this formula corresponds to the standard inhomogeneous bar-resolution in the case of semi-groups and algebras (Cf. MacLane p.115 [4] and Cartan-Eilenberg pp.174–175. [34]), we name $\delta$ the Hochschild co-boundary, as in the case of semi-groups, and algebras.

Remark that a function $F$ satisfying the joint locality condition, (*i.e.*, the hypothesis that $F(Y_1; ...; Y_m; P)$ depends only on $(Y_1, ..., Y_m)_*P$), has a co-boundary which is also jointly local, because the variables appearing in the definition are all joint variables of the $Y_j$. (This this would not have been true for the stronger locality hypothesis asking that $F$ depends only on the collection $(Y_j)_*P; j = 1, ..., m.$)

It is easy to verify that $\delta^m \circ \delta^{m-1} = 0$. We denote by $Z^m$ the kernel of $\delta^m$ and by $B^m$ the image of $\delta^{m-1}$. The elements of $Z^m$ are named $m$-cocycles, we consider them as *information quantities*, and the elements of $B^m$ are $m$-coboundaries.

**Definition 2.** For $m \geq 0$, the quotient

$$H^m(\mathcal{C}^*) = Z^m/B^m \tag{23}$$

is the $m$-th *cohomology group of information* of the information structure $\mathcal{S}$ on the simplicial family of probabilities $\mathcal{Q}$. We denote it by $H^m(\mathcal{S}; \mathcal{Q})$.

The information co-homology satisfies functoriality properties:

Consider two pairs of information structures and probability families, $(\mathcal{S}, \mathcal{Q})$ and $(\mathcal{S}', \mathcal{Q}')$ on two sets $\Omega, \Omega'$ equipped with the $\sigma$-algebras $\mathcal{B}, \mathcal{B}'$ respectively, and $\varphi$ a *surjective* measurable map from $(\Omega, \mathcal{B})$ to $(\Omega', \mathcal{B}')$, such that $\varphi_*(\mathcal{Q}) \subseteq \mathcal{Q}'$ (*i.e.*, $\varphi_*(Q) \in \mathcal{Q}'$ for every $Q \in \mathcal{Q}$), and such that $\mathcal{S} \subseteq \varphi^*\mathcal{S}'$ (*i.e.*, $\forall X \in \mathcal{S}, \exists X' \in \mathcal{S}', X = X' \circ \varphi$); then we have the following construction:

**Proposition 1.** For each integer $m \geq 0$, a natural linear map

$$\varphi^* : H^m(\mathcal{Q}'; \mathcal{S}') \to H^m(\mathcal{Q}; \mathcal{S}), \tag{24}$$

is defined by the following application at the level of local co-chains:

$$\varphi^*(F')(X_1; ...; X_m; P) = F'(X'_1; ...; X'_m; \varphi_*(P)), \tag{25}$$

for a collection of variables $X'_j; j = 1, ..., m$ satisfying $X_j = X'_j \circ \varphi$ for each $j$.

**Proof.** First, remark that $X_j = X"_j \circ \varphi$ implies $X'_j = X"_j$ because $\varphi$ is surjective. As $F'$ is (jointly) local, the co-chain $F = \varphi^*(F')$ is also (jointly) local. Finally, it is evident that the map $F' \mapsto F$ commutes with the co-boundary operator. Therefore the proposition follows.

Another co-homological construction works in the reversed direction:

Consider two information structures $(\mathcal{S}, \mathcal{Q})$ and $(\mathcal{S}', \mathcal{Q}')$ on two sets $\Omega, \Omega'$ equipped with $\sigma$-algebras $\mathcal{B}, \mathcal{B}'$ respectively, and $\varphi$ a measurable map from $(\Omega, \mathcal{B})$ to $(\Omega', \mathcal{B}')$, such that $\mathcal{Q}' \subseteq \varphi_*(\mathcal{Q})$ (*i.e.*, $\forall Q' \in \mathcal{Q}', \exists Q \in \mathcal{Q}, Q' = \varphi_*(Q)$), and such that $\varphi^*\mathcal{S}' \subseteq \mathcal{S}$ (*i.e.*, $\forall X' \in \mathcal{S}', X = X' \circ \varphi \in \mathcal{S}$); then the following result is true:

**Proposition 2.** For each integer $m \geq 0$, a natural linear map

$$\varphi_* : H^m(\mathcal{Q}'; \mathcal{S}') \to H^m(\mathcal{Q}; \mathcal{S}), \tag{26}$$

is defined by the following application at the level of co-chains:

$$\varphi_*(F)(X'_1; ...; X'_m; P') = F(X'_1 \circ \varphi; ...; X'_m \circ \varphi; P), \tag{27}$$

for a probability law $P \in \mathcal{Q}$ and its image $P' = \varphi_*(P)$.

**Proof.** First, remark that, if $Q$ also satisfies $P' = \varphi_*(Q)$, we have $F(X'_1 \circ \varphi; ...; X'_m \circ \varphi; P) = F(X'_1 \circ \varphi; ...; X'_m \circ \varphi; Q)$. To establish that point, let us denote $X_j = X'_j \circ \varphi; j = 0, ..., m$, and $X' = (X'_1, ..., X'_m)$, $X = (X_1, ..., X_m)$ the joint variables; the quantity $F(X'_1 \circ \varphi; ...; X'_m \circ \varphi; P)$ depends only on $X_*P$, but this law can be rewritten $X'_*P'$, which is also equal to $X_*Q$. In particular, if $F$ is local, then $F' = \varphi_*F$ is local.

As it is evident that the map $F \mapsto F'$ commutes with the co-boundary operator, the proposition follows.

Remark this way of functoriality uses the locality of co-cycles.

**Corollary 1.** In the case where $\mathcal{Q}' = \varphi_*(\mathcal{Q})$ and $\mathcal{S} = \varphi^*\mathcal{S}'$, the maps $\varphi^*$ and $\varphi_*$ in information co-homology are inverse one of each other.

This is our formulation of the invariance of the information co-homology for equivalent information structures.

When $m = 0$, co-cochains are functions $f$ of $P_X$ in $\mathcal{Q}_X$ such that $f(Y_*P_X) = f(P_X)$ for any $Y$ multiple of $X$ (*i.e.*, coarser than $X$). As we assume $1$ belongs to $\mathcal{S}$, and the set $Q_1$ has only one element, $f$ must be a constant. And every constant is a co-cycle, because

$$\delta.f(X_0; P) = X_0.f(P) - f(P) = \sum_j P(X_0 = x_j) f(P|X_0 = x_j) - f(P) = f(1)(1-1) = 0. \quad (28)$$

Consequently $H^0$ is $\mathbb{R}$. This corresponds to the hypothesis $1 \in \mathcal{S}$, meaning connexity of the category. If $m$ components exist, we recover them in the same way and $H^0$ is isomorphic to $\mathbb{R}^m$.

We now consider the case $m = 1$. From what precedes we know that there is no non-trivial co-boundary.

Non-homogeneous 1-cocycles of information are families of functions $f_X(Y; P_X)$, measurable in the variable $P$ in $\mathcal{Q}$, labelled by elements $Y \in \mathcal{S}_X$, which satisfies the locality condition, stating that each time we have $Z \to X \to Y$ in $\mathcal{S}$, we have

$$f_X(Y; X_*P_Z) = f_Z(Y; P_Z) \quad (29)$$

and the co-cycle equation, stating that for two elements $Y, Y'$ of $\mathcal{S}_X$, we have

$$f((Y, Y'); P) = f(Y; P) + Y.f(Y'; P). \quad (30)$$

Remark that locality implies that it is sufficient to know the $f_Y(Y; Y_*P)$ to recover $f_X(Y; P)$ for all partition $X$ in $\mathcal{S}$ that divides $Y$.

It is in this sense that we frequently omit the index $X$ in $f_X$.
Remark also that for any 1-cocycle $f$ we have $f(1; P) = 0$.

In fact, the co-cycle equation tells that

$$f((1, 1); P) = f(1; P) + 1.f(1; P). \quad (31)$$

but

$$1.f(1; P) = f(1; P|1 = 1) = f(1; P), \quad (32)$$

and $(1, 1) = 1$, thus $f(1; P) = 0$.

More generally, for any $X$, and any value $x_i$ of $X$, we have

$$f(X; P|(X = x_i)) = 0, \quad (33)$$

In fact a special case of Equation (30) is

$$f((X,X);P) = f(X;P) + X.f(X;P). \tag{34}$$

which implies $X.f(X;P) = 0$; however, by definition,

$$X.f(X;P) = \sum_i P(X = x_i) f(X;P|(X = x_i)), \tag{35}$$

thus for every $i$ we must have $f(X;P|(X = x_i)) = 0$, due to $P \geq 0$. This generalizes $f(1;P) = 0$ for any $P$, because, for a probability conditioned by $X = x_i$, the partition $X$ appears the same as 1, that is a certitude.

Remark also that for each pair of variables $(X,Y)$, a 1-cocycle must satisfy the following symmetric relation:

$$f(Y;\mathbb{P}) - Z.f(Y;\mathbb{P}) = f(Z;\mathbb{P}) - Y.f(Z;\mathbb{P}). \tag{36}$$

*2.3. Entropy*

Any multiple of the Shannon entropy is a non-homogeneous information co-cycle. Remind that entropy $H$ is defined for one partition $X$ by the formula

$$H(X;\mathbb{P}) = -\sum_i p_i \log p_i, \tag{37}$$

where the $p_i$ denotes the values of $\mathbb{P}$ on the elements of the partition $X$. In particular the function $H$ depends only on $X_*(\mathbb{P})$, which is locality. The co-cycle equation expresses the fundamental property for an information quantity, writen by Shannon:

$$H(X,Y) = H(X) + H_X(Y) \tag{38}$$

Thus every constant multiple $f = \lambda H$ of $H$ defines a co-cycle.
Remark that the corresponding "homogeneous 1-cocycle" is the *entropy variation*:

$$F(X;Y;\mathbb{P}) = H(X;\mathbb{P}) - H(Y;\mathbb{P}). \tag{39}$$

This means that it satisfies the "invariance property":

$$\begin{aligned} F((Z,X);(Z,Y)) &= H(Z,X) - H(Z,Y) \\ &= H(Z) + H_Z(X) - H(Z) - H_Z(Y) \\ &= Z.F(X;Y), \end{aligned}$$

and the "simplicial equation":

$$F(Y;Z) - F(X;Z) + F(X;Y) = 0 \tag{40}$$

Note that the entropy variation $H(X;P) - H(Y;P)$ exists in a wider range of condition, *i.e.*, when $\Omega$ is infinite, if the laws of $X$ and $Y$ are absolutely continuous with respect to a same probability law

$\mathbb{P}_0$: we only have to replace the finite sum by the integral of the function $-\varphi \log \varphi$ where $\varphi$ denotes the density with respect to $\mathbb{P}_0$. Changing the reference law $\mathbb{P}_0$ changes the quantities $H(X)$ and $H(Y)$ by the same constant, thus does not change the variation $H(X;P) - H(Y;P)$.

We will prove now that, for many *simplicial* structures $\mathcal{S}$, and sufficiently large adapted probability complexes $\mathcal{Q}$, any information co-homology class of degree one is a multiple of the entropy class.

In particular this would be true for $\mathcal{S} = W(\Sigma)$ and $\mathcal{Q} = \Delta(\Omega)$, when $\Sigma$ has more than two elements and $\Omega$ more than four elements, but this is also true in more refined situation, as we will see.

We assume that the functor of probabilities $\mathcal{Q}_X$ contains all the laws on $\Omega/X$, when $X$ belongs to $\mathcal{S}$. In such a case, by definition, we say that $\mathcal{Q}$ is *complete* with respect to $\mathcal{S}$.

Let us consider a probability law $P$ in $\mathcal{Q}$ and two partitions $X, Y$ in the structure $\mathcal{S}$, such that the joint $XY$ belongs to $\mathcal{S}$. We denote by Greek letters $\alpha, \beta, ...$ the indices labelling the partition $Y$ and by Latin letters $k, l, ...$ the indices of the partition $X$; the probability that $X = \xi_k, Y = \eta_\alpha$ is noted $p_{k,\alpha}$, then the probability of $X = \xi_k$ is equal to $p_k = \sum_\alpha p_{k,\alpha}$ and the probability of $Y = \eta_\alpha$ is equal to $q_\alpha = \sum_k p_{k,\alpha}$.

To simplify the notations, let us write $F = f(X;\mathbb{P})$,$G = f((Y,X);\mathbb{P})$,$H = f(Y;\mathbb{P})$, $F_\alpha = f(X;\mathbb{P}|(Y = \eta_\alpha))$,$H_k = f(Y;\mathbb{P}|(X = \xi_k))$.

The Hochschild co-cycle equation gives

$$\sum_\alpha q_\alpha F_\alpha(\frac{p_{k_1,\alpha}}{q_\alpha}, ..., \frac{p_{k_m,\alpha}}{q_\alpha}) = G((p_{k,\alpha})) - H(q_{\alpha_1}, ..., q_{\alpha_n}) \tag{41}$$

But we also have the relation obtained by exchanging $X$ and $Y$, which gives

$$\sum_k p_k H_k(\frac{p_{k,\alpha_1}}{p_k}, ..., \frac{p_{k,\alpha_n}}{p_k}) = G((p_{k,\alpha})) - F(p_{k_1}, ..., p_{k_m}). \tag{42}$$

Suppose that $p_{k,\alpha} = 0$ except when $\alpha = \alpha_1$ and $k = k_2, k_3, ..., k_m$ or $\alpha = \alpha_2$ and $k = k_1$; we put $p_{k_i,\alpha_1} = x_i; \quad i = 2, ..., m$ and $p_{k_1,\alpha_2} = x_1$, which implies that we have $x_1 + x_2 + ... + x_m = 1$. Then Equation (33) implies that each term $H$ in Equation (42) is zero, because only one value of the image law is non-zero, thus we can replace the only term $G$ by $F(p_{k_1}, ..., p_{k_m})$, and we get from Equation (41):

$$H(1 - x_1, x_1, 0, ..., 0) = F(x_1, x_2, ..., x_m) - (1 - x_1)F_{\alpha_1}(0, \frac{x_2}{1 - x_1}, ..., \frac{x_m}{1 - x_1}). \tag{43}$$

Only the term $F$ for $\alpha_1$ subsists because, the possible other one, for $\alpha_2$, concerns a certitude.

Consequently, by imposing $x_2 = 1 - x_1 = a, x_3 = ... = x_m = 0$, we deduce the identity $H(a, 1 - a, 0, ..., 0) = F(1 - a, a, 0, ..., 0)$. This gives a recurrence equation to calculate $F$ from the binomial case:

$$F(x_1, x_2, ..., x_m) = F(x_1, 1 - x_1, 0, ..., 0) + (1 - x_1)F(0, \frac{x_2}{1 - x_1}, ..., \frac{x_m}{1 - x_1}). \tag{44}$$

That is due to the fact that $F_{\alpha_1}$ is a special case of $F$, thus independent from $Y$ and $\alpha_1$.

Then coming back to the co-cycle equation, we obtain in particular a functional equation for the binomial variables.

**Lemma 2.** With the notations of the example 1 (*cf.* example 1), $\Omega = \{(00),(01),(10),(11)\}$, $S_1$ (resp. $S_2$) the projection $pr_1$ (resp. $pr_2$), on $E_1 = E_2 = \{0,1\}$, $S = \{S_1, S_2\}$; then the (measurable) information co-homology of degree one is generated by the entropy, i.e., there exists a constant $C$ such that, for any $X$ in $W(\Sigma)$, $P \in \mathcal{P}$, $f(X;P) = CH(X;P)$.

**Proof.** We consider a 1-cocycle $f$. We have $f(1;P) = 0$. Let us note $f_i(P) = f(S_i;P)$, and $f_{ijk}(u)$ the function $f(S_i;P|(S_j = k))$, the variable $u$ representing the probability of the first point in the fiber $S_j = k$ in the lexicographic order. For each tableau $2\times 2$, $P = (p_{00},p_{01},p_{10},p_{11})$, the symmetry formula (36) gives

$$\begin{aligned} & (p_{00}+p_{10})f_{120}(\frac{p_{00}}{p_{00}+p_{10}}) + (p_{01}+p_{11})f_{121}(\frac{p_{01}}{p_{01}+p_{11}}) - f_1(P) \\ = \;& (p_{00}+p_{01})f_{210}(\frac{p_{00}}{p_{00}+p_{01}}) + (p_{10}+p_{11})f_{211}(\frac{p_{10}}{p_{10}+p_{11}}) - f_2(P) \end{aligned} \tag{45}$$

imposing $p_{10} = 0$,$p_{00} = u$,$p_{11} = v$,$p_{01} = 1-u-v$ in this relation, we obtain the equation:

$$\begin{aligned} & (1-u)f_1(0,\frac{1-u-v}{1-u},0,\frac{v}{1-u}) - f_1(u,1-u-v,0,v) \\ = \;& (1-v)f_2(\frac{u}{1-v},\frac{1-u-v}{1-v},0,0) - f_2(u,1-u-v,0,v). \end{aligned} \tag{46}$$

By hypothesis, $f_1, f_2$ depend only on the image law by $S_1, S_2$ respectively, thus, again by noting a binomial probability from the value of the first element in lexicographic order, we get

$$(1-u)f_1(\frac{1-u-v}{1-u}) - f_1(1-v) = (1-v)f_2(\frac{u}{1-v}) - f_2(u). \tag{47}$$

By equating $u$ to $1-v$, we find that $f_1(u) = f_2(u)$; then we arrive to the following functional equation for $h = f_1 = f_2$:

$$h(u) - h(v) = (1-v)h(\frac{u}{1-v}) - (1-u)h(\frac{v}{1-u}) \tag{48}$$

This is the functional equation which was considered by Tverberg in 1958 [35]. As a result of the works of Tverberg [35], Kendall [36] and Lee (1964, [37]), (see also Kontsevich, 1995 [38]), it is known that every measurable solution of this equation is a multiple of the entropy function:

$$h(x) = C(x\log(x) + (1-x)\log(1-x)). \tag{49}$$

>From here it follows that, for any $m$-uple $(x_1,...,x_m)$ of real numbers such that $x_1+...+x_m = 1$,

$$F(x_1,x_2,...,x_m) = C\sum_i x_i\log(x_i). \tag{50}$$

The same is true for $H$ and $G$ with the appropriate number of variables.

A pair of variables $X, Y$, such that $X, Y, (XY)$ belong to $\mathcal{S}$, is called an edge of $\mathcal{S}$; we says this edge is *rich* if $X$ and $Y$ contain at least two elements and $(X,Y)$ at least four elements which cross the elements of $X$ and $Y$, in such a manner that the Lemma 2 applies if $\mathcal{Q}$ is complete. We say that $\mathcal{S}$ is *connected*, if every pair of elements $X, X'$ in $\mathcal{S}$ can be joined by a sequence of edges. We say that

$\mathcal{S}$ is *sufficiently rich* if each vertex belongs to at least one rich edge. By the the recurrence Equation (100), these two conditions guaranty that the constant $C$ which appears in the Lemma 2 is the same for all rich edges. Then the same recurrence Equation (100) implies that the whole co-cycle is equal to $CH$. If $\mathcal{S}$ has $m$ connected components, we get necessarily $m$ independent constants.

Thus we have established the following result:

**Theorem 1.** For every connected structure of information $\mathcal{S}$, which is sufficiently rich, and every set of probability $\mathcal{Q}$, which is complete with respect to $\mathcal{S}$, the information co-homology group of degree one is one-dimensional and generated by the classical entropy.

The theorem applies to rich simplicial complexes, in particular to the full simplex $\mathcal{S} = W(\Sigma)$ , which is generated by a family $\Sigma$ of partitions $S_1, ..., S_n$, when $n \geq 2$, such that, for every $i$ at least of the pairs $(S_i, S_j)$ is rich.

Note that most of the axiomatic characterizations of entropy have used convexity, and recurrence over the dimension, see Khintchin [39], Baez *et al.* [20].

In our characterization, we assumed no symmetry hypothesis, this was a consequence of co-homology. Moreover, we do not assume any stability property relating to a higher dimensional simplex, this was also a consequence of the homological definition.

There exists a notion of symmetric information co-homology:

The group of permutations $\mathfrak{S}(\Omega, \mathcal{B})$, made by the permutations of $\Omega$ that respect the algebra $\mathcal{B}$, acts naturally on the set of partitions $\Pi(\Omega)$; in fact, if $X \in \Pi(\Omega)$ is made by the subsets $\Omega_1, ..., \Omega_k$, the partition $\sigma^* X$ is made by the subsets $\sigma^{-1}(\Omega_1), ..., \sigma^{-1}(\Omega_1)$, in such a manner that, if $\sigma, \tau$ are two permutations of $\Omega$, we have $\tau^*(\sigma^* X) = (\sigma \circ \tau)^* X$.

We say that a classical information structure $\mathcal{S}$ on $(\Omega, \mathcal{B})$ is *symmetric* if it is closed by the action of the group of permutations $\mathfrak{S}(\Omega, \mathcal{B})$, *i.e.*, if $X \in \mathcal{S}$, and $\sigma \in \mathfrak{S}(\Omega)$, the partition $\sigma^* X$ also belongs to $\mathcal{S}$.

In the same way, we say that a probability functor $\mathcal{Q}$ is *symmetric*, if it is stable under local permutations, *i.e.*, if $X \in \mathcal{S}$ and $P \in \mathcal{Q}_X$, and if $\sigma \in \mathfrak{S}(\Omega/X)$, then the probability law $\sigma^* P = P \circ \sigma$ on $\Omega/X$ also belongs to $\mathcal{Q}_X$.

Remark that we also have $\tau^* \sigma^* P = (\sigma \circ \tau)^* P)$. Thus the actions of symmetric groups are defined here on the right. However, we have actions to the left by taking $\sigma_* = (\sigma^{-1})^*$. For the essential role of symmetries in information theory, see the article of Gromov in this volume.

A $m$-cochain $F_X : \mathcal{S}^m \times \mathcal{Q}_X \to \mathbb{R}$ is said symmetric, when, for every $X \in \mathcal{S}$, every probability $P \in \mathcal{Q}_X$, every collection of partitions $Y_1, ..., Y_m$ in $\mathcal{S}_X$, we have

$$F_{\sigma_* X}(\sigma_* Y_1; ...; \sigma_* Y_m; \sigma_* P) = F_X(Y_1; ...; Y_m; P). \tag{51}$$

It is evident that symmetric cochains form a subcomplex of the information cochains complex; *i.e.*, the coboundary of a symmetric cochain being a symmetric cochain. Consequently we get a symmetric information co-homology, that we name $H^*_{\mathfrak{S}}(\mathcal{S}; \mathcal{Q})$.

In particular the entropy is a symmetric 1-cocycle.

The above proof of Theorem 1 applies to symmetric cocycle as well, thus, under the convenient hypothesis of connexity, richness, and completeness for $\mathcal{S}$ and $\mathcal{Q}$ we have $H^1_{\mathfrak{S}}(\mathcal{S};\mathcal{Q}) = \mathbb{R}H$.

Remark that an equivalent way to look at symmetric information cochains, consists in enlarging the category $\mathcal{S}$ in a "symmetric category" $\mathcal{S}^{\mathfrak{S}}$, by putting an arrow associated to each element $\sigma_X \in \mathfrak{S}(\Omega/X)$ from $X$ to $\sigma_* X$, and completing the category by composing the two kind of arrows, division and permutation. In this case, the probability functor $\mathcal{Q}$ must behave naturally with respect to permutation, which implies it is symmetric. Moreover, the natural notion of functional sheaf and local cochains are a symmetric sheaf and symmetric cochains.

*2.4. Appendix. Complex of Possible Events*

In each concrete situation, physical constraints produce exclusion rules between possible events, which select a sub-complex $\mathcal{Q}$ in the full probability simplex $\mathcal{P} = \Delta_N$ on $\Omega$. The aim of this appendix is to make this remark more precise.

Let $A^0, A^1, A^2, A^3, ...$ the $N+1$ vertices of the large simplex $\Delta_N$, a point of $\Delta_N$ is interpreted as a probability $\mathbb{P}$ on the set of thee vertices; each vertex can be seen as an elementary event, and we will say that a general event $A$ is *possible* for $\mathbb{P}$ when $\mathbb{P}(A)$ is different from zero. An event $A$ is said *impossible* for $\mathbb{P}$ in the other case, that is when $\mathbb{P}(A) = 0$.

The star $S(A)$ of a vertex $A$ of $\Delta_N$ is the complementary set of the opposite face to $A$, *i.e.*, it is the set of probabilities $P$ in $\Delta_N$ such that $A$ is possible, *i.e.*, has non-zero probability. The relative star $S(A|K)$ of $A$ in subcomplex $K$ is the intersection of the star of $A$ with $K$.

We denote $F = (A, B, C, D, ...)$ the face of $\Delta_N$ whose vertices are $A, B, C, D, ....$ We note $L(F)$ the set of points $p$ in $\Delta_N$ such that at least one of the points $A, B, C, D, ...$ is impossible for $p$. This is also the reunion of the faces which are opposite to the vertices $A, B, C, D, ...$ . Then $L(F)$ is a simplicial complex. The complementary set in $F$ of the interior of $F$, *i.e.*, the boundary of $F$, is the reunion of the intersections of $F$ with all faces opposite to $A, B, C, D, ...$; it is also the set of probabilities $p$ in $F$ such that at least one of the points $A, B, C, D, ...$ is impossible for $p$, thus it is equal to $L(F) \cap F$. If $G$ is a face containing $F$ the complex $L(G)$ contains the complex $L(F)$.

Let $K$ be a simplicial complex contained in a $N$-simplex; then $K$ is obtained by deleting from $\Delta_N$ a set $E = E_K$ of open faces. Let $\dot{F} = F \backslash \partial F$ be an element of $E$, then each faces $G$ of $\Delta_N$ containing $F$ belongs to $E$, because $K$ is a complex.

In this case $K$ is contained in $L(F)$. In fact $L(F)$ is the smallest sub-complex of $\Delta_N$ which does not contain $\dot{F}$. This can be proved as follows: if $p$ in $K$ makes that every vertices of $F$ is possible, it belongs to a face $G$ such that every vertex of $F$ is a vertex of $G$, thus $K$ contains $G$ which contains $F$. So, if $K$ does not contain $\dot{F}$, $K$ is contained in $L(F)$.

Let $L = L_K$ be the intersection of the $L(F)$, where $F$ describe the faces in $E_K$. From what precedes we know that $K$ is contained in $L$. However, every $\dot{F}$ in $E$ is included in the complementary set of $L(F)$, thus it is included in the complementary set of $L$, which is the union of the complementary sets of the $L(F)$. Consequently the complementary set of $K$ is included in the complementary set of $L$. Then $K = L$.

This discussion establishes the following result:

**Theorem 2.** A subset $K$ of the simplex $\Delta_N$ is a simplicial sub-complex if and only if it is defined by a finite number of constraints of the type: "for any $p$ in $K$, the fact that $A, B, C, ...$ are possible for $p$ implies that $D$ is impossible for $p$".

In other terms, more imaged but also more ambiguous, every sub-complex $K$ is defined by constraints of the type: "if $A, B, C, ...$ are simultaneously allowed it is excluded that $D$ can happen".

The statement of the theorem is just a rewriting of the discussion, using elementary propositional calculus: let $K$ be a sub-complex of $\Delta_N$, we have shown that $K$ is the intersection of the $L(F)$ where the open face $\dot{F}$ is not in $K$, but if $A, B, C, D, ...$ denote the vertices of the face $F$, a point $p$ belongs to $L(F)$ if and only if "($A$ is impossible for $p$) or ($B$ is impossible for $p$) or ...", and this sentence is equivalent to "if ($A$ is possible for $p$) and ($B$ is possible for $p$) and ..., then ($D$ is impossible for $p$)". This results from the equivalence between "($P$ implies $Q$) is true" and "(no $P$ or $Q$) is true". Reciprocally any $L(F)$ is a simplicial complex, then every intersection of sets of the form $L(F)$ is a simplicial complex too.

### 3. Higher Mutual Informations. A Sketch

The topological co-boundary operator on $\mathcal{C}^*$, denoted by $\delta_t$, is defined by the same formula as $\delta$, except that the first term $Y_1.F(Y_2; ...; Y_n; \mathbb{P})$ is replaced by the term $F(Y_2; ...; Y_n; \mathbb{P})$ without $Y_1$:

$$\begin{aligned}&\delta_t^m F(Y_1; ...; Y_{m+1}; P_X)\\ =&F(Y_2; ...; Y_{m+1}; P_X) + \sum_1^m (-1)^i F(...; (Y_i, Y_{i+1}); ...; Y_{m+1}; P_X) + (-1)^{m+1} F(Y_1; ...; Y_m; P_X).\end{aligned} \tag{52}$$

It is the coboundary of the bar complex for the trivial module $\mathcal{F}_t$, which is the same as $\mathcal{F}$ except no conditioning appears, *i.e.*, $Y.F = F$. Hence it is the ordinary simplicial co-homology of the complex $\mathcal{S}$ with local coefficients in $\mathcal{F}$.

Remark that this operator also preserves locality, because all the functions of $\mathbb{P}$ which comes in the development depends only on $(Y_2, ..., Y_n) * \mathbb{P}$, $(Y_1, ..., Y_n) * \mathbb{P}$ and $(Y_1, ..., Y_{n-1}) * \mathbb{P}$.

By definition a topological cocycle of information is a cochain $F$ that satisfies $\delta_t F = 0$, and a topological co-boundary is an element in the image of $\delta_t$.

It is easy to show that $\delta_t \circ \delta_t = 0$, which allows to define a co-homology theory that we will name topological co-homology.

Now assume that the information structure $\mathcal{S}$ is a set $W(\Sigma) = \Delta(n)$ generated by a family $\Sigma$ of partitions $S_1, ..., S_n$, when $n \geq 2$.

Higher mutual information quantities were defined by Hu Kuo Ting [6] (see also Yeung [40]), generalizing the Shannon mutual information.

$$I_N(S_1; ...; S_N; \mathbb{P}) = \sum_{k=1}^{k=N} (-1)^{k-1} H_k(S_1; ...; S_N; \mathbb{P}), \tag{53}$$

where

$$H_k(S_1;...;S_N;\mathbb{P}) = \sum_{I\subset[N];card(I)=k} H(S_I;\mathbb{P}), \tag{54}$$

$S_I$ denoting the joint partition of the $S_i$ such that $i \in I$. We also define $I_1 = H$.

The definition of $I_N$ makes evident it is a symmetric function, invariant by all permutation of the partitions $S_1, ..., S_N$.

For instance $I_2(S;T) = H(S) + H(T) - H(S,T)$ is the usual mutual information.

It is easily seen that $I_2 = \delta_t H$. The following formula generalizes this remark to higher mutual informations of even orders:

$$I_{2m} = \delta_t\delta\delta_t...\delta\delta_t H, \tag{55}$$

where the right member contains $2m-1$ terms.

And for odd mutual information we have

$$I_{2m+1} = -\delta\delta_t\delta\delta_t...\delta\delta_t H, \tag{56}$$

where the right member contains $2m$ terms.

We deduce from here that higher mutual informations are co-boundaries for $\delta$ or $\delta_t$ according that their order is odd or even respectively.

The result which proves the two above formulas is the following:

**Lemma 3.** Let $n$ be even or odd we have

$$I_N((S_0,S_1);S_2;...;S_N;\mathbb{P}) = I_N(S_0;S_2;...;S_N;\mathbb{P}) + S_0.I_N(S_1;S_2;...;S_N;\mathbb{P}) \tag{57}$$

This lemma can be proved by comparing the completely developed forms of the quantities. It seems to signify that, with respect to one variable, $I_N$ satisfies the equation of information 1-cocycle, thus $I_N$ seems to be a kind of "partial 1-cocycle"; however this is misleading, because the locality condition is not satisfied. In fact $I_N$ is a $N$-cocycle, either for $\delta$, either for $\delta_t$ depending on the parity of $N$.

For any $N$-cochain $F$ we have

$$(\delta - \delta_t)F(S_0;S_1;...;S_N;\mathbb{P}) = ((S_0-1).F)(S_1;...;S_N;P), \tag{58}$$

where $S_0 - 1$ denotes the sum of the two operators of mean conditioning and minus identity. That implies:

$$(\delta\delta_t - \delta_t\delta)F(S_0;S_1;S_2;...;S_N;\mathbb{P}) = ((1+S_0+S_1-S_0S_1).F)(S_2;...;S_N;\mathbb{P}), \tag{59}$$

**Remark 3.** Reciprocally the functions $I_N$ decompose the entropy of the finest joint partition:

$$H(S_1,S_2,...,S_N;\mathbb{P}) = \sum_{k=1}^{k=N}(-1)^{k-1}\sum_{I\subset[N];card(I)=k} I_k(S_{i_1};S_{i_2};...;S_{i_k};\mathbb{P}) \tag{60}$$

For example, we have $H(S,T) = I_1(S) + I_1(T) - I_2(S;T)$, and

$$H(S,T,U) = H(S) + H(T) + H(U) - I_2(S;T) - I_2(T;U) - I_2(S;U) + I_3(S;T;U). \quad (61)$$

Let us also note the recurrence formula whose proof is left to the reader (*cf.* Cover and Thomas [41]):

$$I_{N+1}(S_0;S_1;...;S_N) = I_N(S_1;...;S_N) - S_0.I(S_1;...;S_N). \quad (62)$$

## 4. Quantum Information and Projective Geometry

### *4.1. Quantum Measure, Geometry of Abelian Conditioning*

In finite dimensional quantum mechanics the role of the finite set $\Omega$ of atomic events is played by a complex vector space $E$ of finite dimension.

In fact, to each set $\Omega$, of cardinal $N$, is naturally associated a vector space of dimension $N$ over $\mathbb{C}$, which is the space freely generated over $\mathbb{C}$ by the elements of $\Omega$. Then we can identify $E$ with $\mathbb{C}^N$, the canonical basis being the points $x$ of $\Omega$. In this case the canonical positive hermitian metric on $E$ corresponds to the quadratic mean: if $f$ and $g$ are elements of $E$, we have

$$h_0(f,g) = \langle f|g\rangle_0 = \int \bar{f}(\omega)g(\omega)d\omega = \frac{1}{N}\sum_j \overline{f_j}g_j \quad (63)$$

Remark that, in the infinite dimensional situation, the space which would play the role of $E$ is the space of $L^2$ functions for a fixed probability $P_0$.

Probability laws $\mathbb{P}$, which are elements of the big simplex $\Delta(N)$, give other hermitian structures, the ones which are expressed by diagonal matrices, with positive coefficients, and trace equal to $1$.

In the general quantum case, described by $E$, a quantum probability law is every positive non-zero hermitian product $h$. If a basis is chosen, $h$ is described by an $N \times N$-matrix $\rho$. In the physical literature, every such $\rho$ is called a *density of states*; and it is considered as a full description of the physical states of the finite quantum system. Usually $\rho$ is normalized by $Tr(\rho) = 1$.

Note that this condition on the trace has no meaning for a positive hermitian form $h$ if no additional structure is given, for instance a non-degenerate form $h_0$ of reference. Why is it so? Because *a priori* a hermitian form $h$ on $E$ is a map from $E$ to $\overline{E}^*$, where $*$ denotes duality and $bar$ denotes conjugation, the conjugate space $\overline{E}$ being the same set $E$, with the same structure of vector space over the real numbers as $E$, but with structure of vector space over the complex numbers changed by changing the sign of the action of the imaginary unit $i$. The complexification of the real vector space $H$ of hermitian forms is $Hom_{\mathbb{C}}(E,\overline{E}^*) \cong E^* \otimes \overline{E}^*$. The space $H$ is the set of fixed points of the $\mathbb{C}$-anti-linear map $u \mapsto^t \bar{u}$. A trace is defined for an endomorphism of the space $E$, as a linear invariant quantity on $E^* \otimes E$. Here we could take the trace over $\mathbb{R}$, because $E$ and $\overline{E}$ are the same over $\mathbb{R}$, but the duality would be an obstacle, because even over the field $\mathbb{R}$, the spaces $E$ and $E^*$ cannot be identified, and there exits no linear invariant in $E^* \otimes E^*$, even over $\mathbb{R}$. In fact, a non-degenerate positive $h_0$ is one of the way to identify $E$ and $\overline{E}^*$. A basis is another way, also defining canonically a form $h_0$. More precisely, when $h_0$ is given, every hermitian form

$h$ diagonalizes in an orthonormal basis for $h_0$, thus all the spectrum of $h$ makes sense not only the trace.

This $h_0$ is tacitly assumed in most presentations. However it is better to understand the consequences of this choice. In non-relativistic quantum mechanics, it is not too grave, however in relativist quantum mechanics, it is; for instance, considering the system of two states as a spinor on the Lorentz space of dimension $4$, the choice of $h_0$ is equivalent to the choice of a coordinate of time. See Penrose and Rindler [42].

A much less violent way to do is to consider hermitian structures $h$ up to multiplication by a strictly positive number. This would have the same effect as fixing the trace equals to one, without introducing any choice. In quantum mechanics only non-zero positive $h$ are considered, not necessarily positive definite, but non-zero. This indicates that a good space of states is not the set $H_+$ of all positive non-zero hermitian products but a convex part $PH_+$ of the real projective space of real lines in the vector space $H$ of hermitian forms. In this space, the complex projective space $\mathbb{P}(E)$ of dimension $N-1$ over $\mathbb{C}$ is naturally embedded, its image consists of the rank one positive hermitian matrices of trace $1$; these matrices correspond to the orthogonal projectors on one dimensional directions in $E$.

When a basis of $E$ is chosen, particular elements of $\mathbb{P}(E)$ are given by the generators of $\mathbb{C}^N$; they correspond to the Dirac distributions on classical states. We see here a point defended in particular by Von Neumann, that quantum states are projective objects not linear objects.

The classical random variables, *i.e.*, the measurable functions on $\Omega$ with values in $\mathbb{C}$, are generalized in Quantum Mechanics by the operators in $E$, they are all the endomorphisms, *i.e.*, any $N \times N$-matrix, and they are named *observables*. Classical observables are recovered by diagonal matrices, their action on $E$ corresponding to the multiplication of functions. Real valued variables are generalized by hermitian operators. Again this supposes that a special probability law $h_0$ is given. If not "to be hermitian" for an operator has no meaning. (What could have a meaning for an operator is to be diagonalizable over $\mathbb{R}$, which is something else.)

Then if $h_0$ is chosen, the only difference between real observable and density of states is the absence of the positivity constraint.

By definition, the *amplitude*, or *expectation*, of the observable $Z$ in the state $\rho$ is the number given by the formula

$$\mathbb{E}_\rho(Z) = Tr(Z\rho). \tag{64}$$

It is important to note that $h_0$ plays a role in this formula. Consequently the definition of expectation requires to fix an $h_0$ not only a $\rho$. This imposes a departure from the relativistic case, which shall not be surprising, since considerations in relativistic statistical physics show that the entropy, for instance, depends on the choice of a coordinate for time. Cf. Landau-Lifschitz, Fluid Mechanics, second edition [43].

The partitions of $\Omega$ associated to random variables are replaced in the quantum context by the spectral decompositions of the hermitian operators $X$. As $h_0$ is given, this decomposition is given by a set of positive hermitian commuting projectors of sum equal to the identity. The additional data for

recovering the operator $X$ is one real eigenvalue for each projector. The underlying fact from linear algebra is that every hermitian matrix is diagonalizable in a unitary basis, which means that

$$Z = \sum_j z_j E_j, \tag{65}$$

where the number $z_j$ are real, two by two different, and where the matrices $E_j$ are hermitian projectors, which satisfy, for any $j$ and $k \neq j$,

$$E_j^2 = E_j; \quad E_j^* = E_j; \quad E_j E_k = E_k E_j = 0; \tag{66}$$

and

$$\sum_j E_j = Id_N \tag{67}$$

When the hermitian operator $Z$ commutes with the canonical projectors on the axis of $\mathbb{C}^N$, its spectral measure gives an ordinary partition of the canonical basis, and we recover the classical situation.

Note that the extension of the notion of partition is given by any decomposition of the vector space $E$ in orthogonal sum, not necessarily compatible with a chosen basis. Again this assumes a given positive definite $h_0$.

To generalize what we presented in the classical setting, quantum information theory must use only the spectral support of the decomposition, not the eigenvalues.

It would have been tempting to consider any decomposition of $E$ in direct sum as a possible observable, however not every linear operator, or projective transformation, corresponds to such a decomposition, due to the existence of non-trivial nilpotent operators. What could be their role in quantum information? Moreover, the presence of $h_0$ fully justifies the limitation to orthogonal decompositions.

In the general case, hermitian but not necessarily diagonal, we define the probability of the elementary events $Z = z_j$ by the following formula

$$\mathbb{P}_\rho(Z = z_j) = Tr(E_j^* \rho E_j) \tag{68}$$

And we define the conditional probability $\rho|(Z = z_j)$ by the formula

$$\rho|(Z = z_j) = E_j^* \rho E_j / Tr(E_j^* \rho E_j). \tag{69}$$

One can notice that this definition can be extended to any projector, not necessarily hermitian. By definition, the conditioning of $\rho$ by a projector $Y$ is the matrix $Y^* \rho Y$, normalized to be of trace 1. However, here, as it is done in most of the texts on Quantum Mechanics, we will mostly restrict ourselves to the case of hermitian projectors, *i.e.*, $Y^* = Y$.

**Remark 4.** What justifies these definitions of probability and conditioning? First they allow to recover the classical notions when we restrict to diagonal densities and diagonal observables, *i.e.*, when $\rho$ is diagonal, real, positive, of trace 1, $Z$ is diagonal, and the $E_j$ are diagonals, in which case

they give a partition of $\Omega$. The mean of $Z$ is its amplitude. The probability of the event $Z = z_j$ is the sum of the probabilities $p(\omega) = \rho_{\omega\omega}$ for $\omega$ in the image of $E_j$; this the trace of $\rho E_j$. Moreover, the conditioning by this event is the probability obtained by projection on this image, as prescribed by the above formula.

Second, pure states are defined as rank one hermitian matrices. In this case $\rho$ is the orthogonal projection on a vector $\psi$ of norm equal to 1 (the finite dimensional version of the Schrodinger wave vector), the exact relation is

$$\rho = |\psi\rangle\langle\psi| \tag{70}$$

or, in coordinates, if $\psi$ has for coordinates the imaginary numbers $\psi(\omega)$, we have

$$\rho_{\omega\omega'} = \overline{\psi(\omega)}\psi(\omega'). \tag{71}$$

Let $Z$ be any hermitian operator, the result of quantum experiments indicate that the probability of the event $Z = z_j$, for the state $\psi$, is equal to

$$P_j = \langle\psi|E_j\psi\rangle. \tag{72}$$

But this quantity can also be written

$$P_j = Tr_{\mathbb{C}}(\langle\psi|E_j\psi\rangle) = Tr_E(|\psi\rangle\langle\psi|E_j) = Tr(\rho E_j). \tag{73}$$

Starting from this formula and the fact any $\rho$ can be written as a classical mixture of commuting quantum pure states,

$$\rho = \sum_a p_a|\psi_a\rangle\langle\psi_a|, \tag{74}$$

we get the general formula of a quantum probability that we recalled.

Moreover, physical experiments indicate that after the measurement of an observable $Z$, giving the quantity $z_j$, the system is reduced to the space $E_j$, and every pure state $\psi$ is reduced to its projection $E_j\psi$, which is compatible with the above definition of conditioning for pure states. Here again, the general formula can be deduced by Equation (74). The division by the probability is achieved to normalize to a trace 1. Thus conditioning in general is given by orthogonal projection in $E$, and it corresponds to the operation of measurement.

However, as claimed in particular by Roger Balian [44], the fact that the decomposition in pure states is non-unique implies that pure states cannot be so pertinent for understanding quantum information.

**Definition 3.** The density of states associated to a given variable $Z$ and a given density $\rho$ is given by the sum:

$$\rho_Z = \sum_j \mathbb{P}_\rho(Z = z_j)\rho|(Z = z_j) = \sum_j E_j^*\rho E_j, \tag{75}$$

where $(E_j)_{j\in J}$ designates the spectral decomposition of $Z$, also named spectral measure of $Z$. Thus $\rho_Z$ is usually seen as representing the density of states after the measurement of the variable $Z$. This formula is usually interpreted by saying that the statistical analysis of the repeated measurements of the observable $Z$ transforms the density $\rho$ into the density $\rho_Z$.

Remark that $\rho_Z$ is better understood as being a collection of conditional probabilities $\rho|(Z = z_j)$, indexed by $j$.

In quantum physics as in classical physics the symmetries, discrete and continuous, have always played a fundamental role. For example, in quantum mechanics, a fundamental principle is the unitarity of the evolution in time, which claims that the states evolve as $\rho_t = U_t\rho$ and that the observables evolve as $Z_t = U_t Z U_t^{-1}$, with $U_t$ respecting the fundamental scalar product $h_0$. In fact, as we already mentioned, a deeper principle associates the choice of a time coordinate $t$ to the choice of $h_0$, which gives birth to a unitary group $U(E; h_0)$, isomorphic to $U_N(\mathbb{C})$. For stationary systems the family $(U_t)_{t\in\mathbb{R}}$ forms a one parameter group, *i.e.*, $U_{t+s} = U_t U_s = U_s U_t$, and there exists a hermitian generator $H$ of $U_t$ in the sense that $U_t = \exp(2\pi i t H/h)$; by definition, this particular observable $H$ is the energy, the most important observable. Even if we have a privileged basis, like $\Omega$ in the relation with classical probability, the consideration of another basis which makes the energy $H$ diagonal is of great importance. In the stationary case, a *symmetry* of the dynamical system is defined as any unitary operator, which commutes with the energy $H$. The set of symmetries forms a Lie group $G$, a closed sub-group in $U_N$. The infinitesimal generators are considered as hermitian observables (obtained by multiplying the elements of the Lie algebra $L(G)$ by $i$); in general they do not commute between themselves.

All these axioms extend to the infinite dimensional situation when $E$ has a structure of an Hilbert space, but the spectral analysis of the un-bounded operators is more delicate and diverse than the analysis in finite dimension. Three kinds of spectrum appear, discrete, absolutely continuous and singular continuous. The symmetries could not form a Lie group in general, and so on.

In our simple case of elementary quantum probability, without fixed dynamics, the classical symmetries of the set of probabilities are given by the permutations of $\Omega$, the vertices of $\Delta(N)$. They correspond to the unitary matrices which have one and only one non-zero element in each line and each column. They do not diagonalize in the same basis because they do not commute, but they form a group $\mathfrak{S}_N$. Another subgroup of $U_N$ is natural for semi-classical study, it is the diagonal torus $\mathbb{T}^N$, its elements are the diagonal matrices with elements of modulus 1, they correspond to sets of angles. The group $\mathfrak{S}_N$ normalizes the torus $\mathbb{T}^N$, *i.e.*, for each permutation $\sigma$ and each diagonal element $Z$, the matrix $\sigma Z \sigma^{-1}$ is also diagonal; its elements are the same as the elements of $Z$ but in a different orders. The subgroup generated by $\mathfrak{S}_N$ and $\mathbb{T}^N$ is the full normalizer of $\mathbb{T}^N$.

One of the strengths of the quantum theory, with respect to the classical theory, is that it gives a similar status to the states, the observables and the symmetries. States are hermitian forms, generalizing points in the sphere (or in the projective space) which are pure states, observables are hermitian operators, or better spectral decompositions, and symmetries are unitary operators, infinitesimal symmetries being anti-hermitian matrices.

All classical groups should appear in this framework. First, by choosing a special structure on $E$ we restrict the linear group $GL_N(\mathbb{C})$ to an algebraic subgroup $G_{\mathbb{C}}$. For instance, by choosing a symmetric invertible bilinear form on $E$ we obtain $O_N(\mathbb{C})$, or, when $N$ is even, by choosing an antisymmetric invertible bilinear form on $E$ we obtain $Sp_N(\mathbb{C})$. In each of these cases there exists a special maximal torus (formed by the complexification of a maximal abelian subgroup $T$ of unitary

operators in $G_{\mathbb{C}}$), and a Weyl group, which is the quotient of the normalizer $N(T)$ by the torus $T$ itself. This Weyl group generalizes the permutation group when more algebraic structures are given in addition to the linear structure. The compact group of symmetries is the intersection $G$ of $G_{\mathbb{C}}$ with $U_N$. In fact, given any compact Lie group $G_c$, and any faithful representation $r_c$ of $G_c$ in $\mathbb{C}^N$, we can restrict real observables to generators of elements in $C_c$, and general observables to complex combinations of these generators, which integrate in a reductive linear group $G$. The spectral decomposition corresponds to the restriction to parabolic sub-groups of $G_{\mathbb{C}}$. The densities of states are restricted to the Satake compactification of the symmetric space $G_{\mathbb{C}}/G_c$ [45].

*4.2. Quantum Information Structures and Density Functors*

To define information quantities in the quantum setting, we have *a priori* to consider families of operators $(Y_1, Y_2, ..., Y_m)$ as joint variables. However, the efforts made in Physics and Mathematics were not sufficient to attribute a clear probability to the joint events $(Y_1 = y_1, Y_2 = y_2, ..., Y_m = y_m)$, when $Y_1, ..., Y_m$ do not commute; we even suspect that this difficulty is revelator of a principle, that information requires a form of commutativity. Thus, in our study, we will adopt the convention that *every time we consider joint observables, they do commute*. Hence we will consider only collections of commuting hermitian observables; their natural amplitudes in a given state are vectors in $\mathbb{R}^m$. However we do not exclude the consideration in our theory of sequences $(Y_1; ...; Y_m)$ such that the $Y_i$ do not commute.

A joint observable $(Y_1, Y_2, ..., Y_m)$ define a linear decomposition of the total space $E$ in direct orthogonal sum

$$E = \bigoplus_{\alpha \in A} E_\alpha, \tag{76}$$

where $E_\alpha; \alpha \in A$ is the collection of joint eigenspaces of the operators $Y_j$. Note that any orthogonal decomposition can be defined by a unique operator.

Another manner to handle the joint variables is to consider linear families of commuting operators

$$Y(\lambda_1, ..., \lambda_m) = \lambda_1 Y_1 + ... + \lambda_m Y_m, \tag{77}$$

or in equivalent terms, linear maps from $\mathbb{R}^m$ to $End(E)$. Then assigning a probability number and perform probability conditioning can be seen as functorial operations.

In what follows we denote indifferently by $E_\alpha$ the subspace of $E$ or the orthogonal projection on this subspace.

>From the point of view of information, two sets of observables are equivalent if they give the same linear decomposition of $E$. We say that a decomposition $E_\alpha; \alpha \in A$ refines a decomposition $E'_\beta; \beta \in B$, when each $E'_\beta$ is a sum of spaces $E_\alpha$ for $\alpha$ in a subset $A_\beta$ of $A$. In such a case, we say that $E_\alpha; \alpha \in A$ *divides* $E'_\beta; \beta \in B$.

For instance, for commuting decompositions $Y, Z$ it is possible to define the joint variable, as the less fine decomposition which is finer than $Y$ and $Z$.

We insist that only decompositions have a role in information study at this moment. We will see that observation trees in the last section imposes to consider a supplementary structure, which consists in an *ordering* of the factors in the decomposition.

An *information structure* on $E$ is a set $\mathbf{S}$ of decompositions $X$ of $E$ in direct sum, such that when $Y$ and $Z$ are elements of $\mathbf{S}$ which refine $X \in \mathbf{S}$, then $Y, Z$ commute and the finer decomposition $(Y, Z)$ they generate belongs to $\mathbf{S}$. In this text, we will only consider orthogonal decompositions.

Remark: in fact, the necessity of this condition in the quantum context was the original motivation to introduce the definition of classical information structure, as exposed in the first section. This can be seen as a comfortable flexibility in the classical context, or as a step from classical to quantum information theory.

As in the classical case, an information structure gives a category, denoted by the letter $\mathbf{S}$, whose objects are the elements of $\mathbf{S}$, and whose arrows $X \to Y$ are given by the divisions $X|Y$ between the decompositions in $\mathbf{S}$.

In what follows we always assume that $1$, which corresponds to the trivial partition $E$, belongs to $\mathbf{S}$, and is a final object. If not we will not get a topos.

Note that we are not the first to use categories and topos to formulate quantum or classical probability. In particular Doring and Isham propose a reformulation of the whole quantum and classical physics by using topos theory, see [46] and references inside. This theory followed remarkable works of Isham, Butterfield and Hamilton, made beween 1998 and 2002, and was further developed by Flori, Heunen, Landsman, Spitters, specially in the direction of a quantum logic. A common point between these works and our work is the consideration of sheaves over the category made by the partial ordering in commutative subalgebras. However, Doring *et al*. consider only the set of maximal algebras, and do not look at decompositions, *i.e*., they consider also the spectral values. In [46], Doring and Isham defined topos associated to quantum and classical probabilities. However, they focused on the definition of truth values in this context. For instance, in the classical setting, the topos they define is the topos of ordinary topological sheaves over the space $(0, 1)_L$ which has for open sets the intervals $]0, r[$ for $0 \leq r \leq 1$, and particular points in their topos are given by arbitrary probabilized spaces, which is far from the objects we consider, because our classical topos are attached to sigma-algebras over a given set. In fact, our aim is more to develop a kind of geometry in this context, by using homological algebra, in the spirit of Artin, Grothendieck, Verdier, when they developed topos for studying the geometry of schemes.

**Example 5.** The most interesting structures $\mathbf{S}$ seem to be provided by the quantum generalization of the simplicial information structure in classical finite probability. A finite family of commuting decompositions $\Sigma = \{S_1, ..., S_n\}$ is given, they diagonalize in a common orthogonal basis, but it can happen that not all diagonal decompositions associated to the maximal torus belongs to the set of joints $W(\Sigma)$. In such a case a subgroup $G_\Sigma$ appears, which corresponds to the stabilizer of the finest decomposition $S_{[n]} = (S_1...S_n)$. This group is in general larger than a maximal torus of $U_N$, it is a product of unitary groups (corresponding to common eigenvalues of observables in $W(\Sigma)$), and it is named a Levy subgroup of the unitary group. In addition we consider a closed subgroup $G$ in the

group $U(E;h_0)$ (which could be identified with $U_N$), and all the conjugates $gYg^{-1}$ of elements of $W(\Sigma)$ by elements of $G$; this gives a manifold of commutative observable families $\Sigma_g; g \in G$. More generally we could consider several families $\Sigma_\gamma; \gamma \in \Gamma$ of commuting observables, where $\Gamma$ is any set. It can happen that an element of $\Sigma_\gamma$ is also an element of $\Sigma_\lambda$ for $\lambda \neq \gamma$. The family $\Gamma * \Sigma$ of the $\Sigma_\gamma$ when $\gamma$ describes the set $\Gamma$ forms a quantum information structure. The elements of this structure are (perhaps ambiguously) parameterized by the product of an abstract simplex $\Delta(n)$ with the set $\Gamma$ (in particular $\Gamma = G$ for conjugated families).

A simplicial information structure is a subset of $\Gamma * \Sigma$ which corresponds to a family $K_\gamma$ of simplicial sub-complexes of $\Delta(n)$. In the invariant case, when $\Gamma = G$, several restrictions could be usefull, for instance using the structure of the manifold of the conjugation classes of $G_\Sigma$ under $G$. The simplest case is given by taking the same complex $K$ for all conjugates $g\Sigma g^{-1}$. By definition this latter case is a simplicial invariant family of quantum observables.

An *event* associated to **S** is a subspace $E_A$, which is an element of one of the decompositions $X \in \mathbf{S}$. For instance, if $Y = (Y_1, ..., Y_m)$, the joint event $A = (Y_1 = y_1, Y_2 = y_2, ..., Y_m = y_m)$ gives the space $E_A$ which is the maximal vector subspace of $E$ where $A$ happens, *i.e.*,

$$(f \in E_A) \Leftrightarrow (Y_1(f) = y_1 f, Y_2(f) = y_2 f, ..., Y_m(f) = y_m f). \tag{78}$$

We say that $A$ is measurable for a decomposition $Y$ whenever it is obtained by unions of elements of $Y$.

The role of the Boolean algebra $\mathcal{B}$ introduced in the first section, could have been accounted here by a given decomposition **B** of $E$ such that any decomposition in **S** is divided by **B**.

However this choice of **B** is too rigid, in particular it forbids invariance by the unitary group $U(h_0)$. Thus we decided that a better analog of the Boolean algebra $\mathcal{B}$ is the set $U\mathbf{B}$ of all decompositions that are deduced from a given **B** by unitary transformations.

On the side of density of states, *i.e.*, quantum probabilities, we can consider a subspace $\mathbf{Q}_1$ of the space $\mathbf{P} = \mathbb{P}\mathbf{H}_+$ of hermitian positive matrices modulo multiplication by a constant. Concretely, we identify the elements of $\mathbf{Q}_1$ with positive hermitian operators $\rho$ such that $Tr\rho = 1$. The space **P** is naturally stratified by the rank of the form; the largest cell $\mathbb{P}\mathbf{H}_{++}$ corresponds to the non-degenerate forms; the smallest cells correspond to the rank one forms, which are called pure states in Quantum Mechanics.

We will only consider subsets $\mathbf{Q}_1$ of **P** which are *adapted* to **S**, *i.e.*, which satisfy that if $\rho$ belongs to $\mathbf{Q}_1$, the conditioning of $\rho$ by elements of **S** also belongs to $\mathbf{Q}_1$. This means that $\mathbf{Q}_1$ is closed by orthogonal projections on all the elements $E_A$ of the orthogonal decompositions $X$ belonging to **S**. Note that a subset of **P** which is closed by all orthogonal projections is automatically adapted to any information category **S**.

Remind that, if $\rho$ is a density of states and $E_A$ is an elementary event (*i.e.*, a subspace of $E$), we define the conditioning of $\rho$ by $A$ by the hermitian matrix

$$\rho|A = E_A^* \rho E_A / Tr(E_A^* \rho E_A). \tag{79}$$

And we define the *probability of the event* $E_A$ for $\rho$ as the trace:

$$\mathbb{P}_\rho(A) = Tr(E_A^* \rho E_A), \tag{80}$$

In the same manner we define the density of a joint observable by

$$\rho_Y = \sum_A \mathbb{P}_\rho(A)\rho|A = \sum_A E_A^* \rho E_A, \tag{81}$$

A nice reference studying important examples is Paul-Andre Meyer, Quantum probability for probabilists [47].

If $X$ is an orthogonal decomposition of $E$, we can associate to it a subset $\mathbf{Q}_X$ of $\mathbf{Q}_1$, which contains at least all the forms $\rho_X$ where $\rho$ belongs to $\mathbf{Q}_1$. The natural axiom that we assume for the function $X \mapsto \mathbf{Q}_X$, is that for each arrow of division $X \to Y$, the set $\mathbf{Q}_Y$ contains the set $\mathbf{Q}_X$; then we note $Y_*$ the injection from $\mathbf{Q}_X$ to $\mathbf{Q}_Y$. The fact that $\mathbf{Q}_X$ is stable by conditioning by every element of a decomposition $Y$ which is less fine than $X$ is automatic; it follows from the fact that $\mathbf{Q}_1$ is adapted to $\mathbf{S}$. We will use conditioning in this way.

In what follows we denote by the letter $\mathbf{Q}$ such a functor $X \mapsto \mathbf{Q}_X$ from the category $\mathbf{S}$ to the category of quantum probabilities, with the arrows given by direct images. The set $\mathbf{Q}_1$ is the value of the functor $\mathbf{Q}$ for the certitude $1$. We must remind that many choices are possible for the functor when $\mathbf{Q}_1$ is given; the two extreme being the functor $\mathbf{Q}^{max}$ where $\mathbf{Q}_X = \mathbf{Q}_1$ for every $X$, and the functor $\mathbf{Q}^{min}$ where $\mathbf{Q}_X$ is restricted to the set of forms $\rho_X$ where $\rho$ describes $\mathbf{Q}_1$; in this last case the elements of $\mathbf{Q}_X$ are positive hermitian forms on $E$, which are decomposed in blocs according to $X$.

From the physical point of view, $\mathbf{Q}^{min}$ appears to have more sense than $\mathbf{Q}^{max}$, but we prefer to consider both of them.

A special probability functor, which will be noted $\mathbf{Q}^{can}(\mathbf{S})$, is canonically associated to a quantum information structure $\mathbf{S}$:

**Definition 4.** The *canonical density functor* $\mathbf{Q}_X^{can}(\mathbf{S})$ is made by all positive hermitian forms matched to $X$, i.e., all the forms $\rho_X$ when $\rho$ describes $P\mathbf{H}_+$.

It is equal to the functor $\mathbf{Q}^{min}$ associated to the full set $\mathbf{Q}_1 = P\mathbf{H}_+$. When the context is clear, we will simply write $\mathbf{Q}^{can}$.

An important difference appears between the quantum and the classical frameworks: if $X$ divides $Y$, there exist more (quantum) probability laws in $\mathbf{Q}_Y$ than in $\mathbf{Q}_X$, but there exist less classical laws at the place $Y$ than at the place $X$, because classical laws are defined on smaller sigma-algebras.

In particular, the trivial partition has only one classical state, which is $Tr(\rho) = 1$, but it has the richest structure in terms of quantum laws, any hermitian positive form.

Let us consider the classical probabilities, *i.e.*, the maps that associate the number $P_\rho(A)$ to an event $A$; then, for an event which is measurable for $Y$, the law $Y_*\rho_X$ gives the same result than the law $\rho_X$.

Remark: This points to a generalized notion of direct image, which is a correspondence $q_X Y_*$ between $\mathbf{Q}_X$ and $\mathbf{Q}_Y$, not a map: we say that the pair $(\rho_X, \rho_Y)$ in $\mathbf{Q}_X \times \mathbf{Q}_Y$ belongs to $q_X Y_*$, if for any event which is measurable for $Y$, we have the equality of probabilities

$$\mathbb{P}_{\rho_X}(A) = \mathbb{P}_{\rho_Y}(A). \tag{82}$$

Let us look at the relation of quantification, between a classical information structure and a quantum one:

Consider a maximal family of commuting observables $\mathcal{S}$ in the quantum information structure $\mathbf{S}$, *i.e.*, the full subcategory associated to an initial object $X_0$. This family is a classical information structure. Conversely, if we start with a classical information structure $\mathcal{S}$, made by partitions of a finite set $\Omega$, we can always consider it as a quantum structure associated to the vector space $E = \mathbb{C}^\Omega$ freely generated over $\mathbb{C}$ by the elements of $\Omega$. Note that $E$ comes with a canonical positive definite form $h_0$, and, to be interesting from the quantum point of view, it is better to extend $\mathcal{S}$ by applying to it all unitary transformations of $E$, generating a quantum structure $\mathbf{S} = U\mathcal{S}$.

**Remark 5.** Suppose that $\mathbf{S}$ is unitary invariant, we can define a larger category $\mathbf{S}^U$ by taking as arrows the isomorphisms of ordered decomposition, and close by all compositions of arrows of $\mathbf{S}$ with them. Such an invariant extended category $\mathbf{S}^U$ is not far to be equivalent to the category $\mathcal{S}^{\mathfrak{S}}$, made by adding arrows for permutations of the sets $\Omega/X$ (cf. above section), from the point of view of category theory: let us work an instant, as we will do in the last part of this paper, with ordered partitions of $\Omega$, being itself equipped with an order, and ordered orthogonal decompositions of $E$. In this case we can associate to any ordered partition $X = (E_1, ..., E_m)$ of $E$, the unique ordered partition $\Omega$ compatible with the sequence of dimensions and the order of $\Omega$. It gives a functor $\tau$ from $\mathbf{S}$ to $\mathcal{S}$ such that $\iota \circ \tau = Id_{\mathcal{S}}$, where $\iota$ denotes the inclusion of $\mathcal{S}$ in $\mathbf{S}$. These two functors are extended, preserving this property, to the categories $\mathbf{S}^U$ and $\mathcal{S}^{\mathfrak{S}}$. In fact, the functor $\iota$ sends a permutation to the unitary map which acts by this permutation on the canonical basis, and the functor $\tau$ sends a unitary transformation $g$ between $X \in \mathbf{S}$ and $gXg^* \in \mathbf{S}$ to the permutation it induces on the orthogonal decompositions. Moreover, consider the map $f$ which associates to any $X \in \mathbf{S}^U$ the unique morphism from the decomposition $\iota \circ \tau(X)$ to $X$; it is a natural transformation from the functor $\iota \circ \tau$ to the functor $Id_{\mathbf{S}^U}$, which is invertible, then it defines an equivalence of category between $\mathcal{S}^{\mathfrak{S}}$ and $\mathbf{S}^U$. However a big difference begins with probability functors.

Let $\mathbf{Q}$ be a quantum density functor adapted to $\mathbf{S}$, and note $\iota^*\mathbf{Q}$ the composite functor on $\mathcal{S}$; we can consider the map $Q$ which associates to $X \in \mathcal{S}$ the set of classical probabilities $\mathbb{P}_\rho$ for $\rho \in \mathbf{Q}_X$. If $X$ divides $Y$, the fact that the direct image $Y_*\mathbb{P}(\rho)$ of $\rho \in \mathbf{Q}_X$ coincides with the law $\mathbb{P}_{Y_*(\rho)}$ gives the following result:

**Lemma 4.** $\rho \mapsto \mathbb{P}_\rho$ is a natural transformation from the functor $\iota^*\mathbf{Q}$ to the functor $Q$.

**Definition 5.** This natural transformation is called the *Trace*, and we denote by $Tr_X$ its value in $X$, *i.e.*, $Tr_X(\rho) = \mathbb{P}_\rho$, seen as a map from $\mathbf{Q}_X$ to $\mathcal{Q}_X$.

In general there is no natural transformation in the other direction, from $\mathcal{Q}_X$ to $\mathbf{Q}_X$.

Remark that the trace sends a unitary invariant functor to a symmetric functor.

*4.3. Quantum Information Homology*

As in the classical case, we can consider the ringed site given by the category **S**, equipped with the sheaf of monoids $\{\mathbf{S}_X; X \in \mathbf{S}\}$. In the ringed topos of sheaves of **S**-modules, the choice of a probability functor **Q** generates remarkable elements in this topos, formed by the functional space **F** of measurable functions on **Q** with values in $\mathbb{R}$. The action of the monoid (or the generated ring) being given by averaged conditioning, and the arrows being given by transposition of direct images. Then, the quantum information co-homology is the topos co-homology:

$$H^m(\mathbf{S}, \mathbf{Q}) = Ext^m_{\mathbf{S}}(\mathbb{R}; \mathbf{F}) \tag{83}$$

However, as in the classical case, we can define directly the co-homology with a bar resolution of the constant sheaf, as follows:

A set of functions $F_X$ of $m$ observables $Y_1, ..., Y_m$ divided by $X$, and one density $\rho$ indexed by $X \in \mathbf{S}$, is said *local*, when for any decomposition $X$ dividing a decomposition $Y$, we have, for each $\rho$ in $\mathbf{Q}_X$,

$$F_X(Y_1; ...; Y_m; \rho) = F_X(Y_1; ...; Y_m; Y_*(\rho)). \tag{84}$$

For $m = 0$ this equation expresses that the family $F_X$ is an element of the topos.

For every $m$, a collection $F_X, X \in \mathbf{S}$ is a natural transform $F$ from a free functor $\mathbf{S}_m$ to the functor **F**.

Be careful that in the quantum context, it is not true in general that locality is equivalent to the condition saying that the value $F_X(Y_1; ...; Y_n; \rho)$ depends only on the family of conditioned densities $E^*_{A_i}\rho E_{A_i}; i = 0, ..., m$, where $A_i$ is one of the possible events defined by $Y_i$.

In fact it depends on the choice of **Q**; for instance it is false for a $\mathbf{Q}^{max}$, but it is true for a $\mathbf{Q}^{min}$.

The counter-example in the case of $\mathbf{Q}^{max}$ is given by a function $F(\rho)$ which is independent of $X$. It is local (in the sense of topos that we adopt) but it is non-local in the apparently more natural sense that it depends only of $\rho_X$. This is important to have this quantum particularity in the mind for understanding the following discussion.

As in the classical case, the action of observables on local functions is given by the average of conditioning, in the manner of Shannon, but using the Von Neumann conditioning:

$$Y.F(Y_0; ...; Y_m; \rho) = \sum_A Tr(E^*_A \rho E_A) F(Y_0; ...; Y_m; \rho|A) \tag{85}$$

where the $E_A$'s are the spectral projectors of the bundle $Y$. In this definition there is no necessity to assume that $Y$ commutes with the $Y_j$'s.

Remind that, when $E^*_A \rho E_A$ is non-zero, $\rho|A$ is equal to $E^*_A \rho E_A / Tr(E^*_A \rho E_A)$, and verifies the normalization condition that the trace equals to one. When $E^*_A \rho E_A$ is equal to zero, the factor $Tr(E^*_A \rho E_A)$ is zero, then by convention the corresponding term $F$ is absent.

The proof of the Lemma 1 applies without significant change to prove that the above formula defines an action of the monoid functor $\mathbf{S}_X$.

Then, the definition of co-homology is given exactly as we have done for the classical case, by introducing the Hochschild operator:

$$\begin{aligned}&\widehat{\delta}^m F(Y_1;...;Y_{m+1};\rho)\\ =&Y_1.F(Y_2;...;Y_{m+1};\rho)+\sum_1^m(-1)^i F(...;(Y_i,Y_{i+1});...;Y_{m+1};\rho)+(-1)^{m+1}F(Y_1;...;Y_m;\rho).\end{aligned}\tag{86}$$

The Von-Neumann entropy is defined by the following formula

$$S(\rho)=\mathbb{E}_\rho(-\log_2(\rho))=-Tr(\rho\log_2(\rho)).\tag{87}$$

For any density functor $\mathbf{Q}$ which is adapted to $\mathbf{S}$, the Von-Neumann entropy defines a local $0$-cochain, that we will call $S_X$, and is simply the restriction of $S$ to the set $\mathbf{Q}_X$. If $\rho$ belongs to $\mathbf{Q}_X$ and if $X$ divides $Y$, the law $Y_*\rho$, which is the same hermitian form as $\rho$ belongs to $\mathbf{Q}_Y$ by functoriality, thus $S(Y_*\rho)=S(\rho)$ is translated by $S_X(\rho)=S_Y(Y_*\rho)$. This $0$-cochain will be simply named the Von Neumann entropy.

In the case of $\mathbf{Q}^{max}$, $S_X$ gives the same value at all places $X$. In the case of $\mathbf{Q}^{min}$ it coincides with $S(\rho_X)$, where $\rho_X$ denotes the restriction to the decomposition $X$.

Be careful: $\rho\mapsto S(\rho_X)$ is not a local $0$-cochain for $\mathbf{Q}^{max}$. In fact in the case of $\mathbf{Q}^{max}$ we have the same set $\mathbf{Q}=\mathbf{Q}_X$ for every place $X$, thus, if we take for $X$ a strict divisor of $Y$ and if we take a density $\rho$ such that, for the restrictions of $\rho$, the spectrum of $\rho_Y$ and $\rho_X$ are different, then, in general, we do not have $S_X(\rho)=S_Y(Y_*\rho)$, even if, as it is the case in the quantum context, $Y_*\rho=\rho$.

Remark that in the case of $\mathbf{Q}^{max}$, where every function of $\rho$ independent of $X$ is a cochain of degree zero, the particular functions which depends only on the spectrum of $\rho$ are invariant under the action of the unitary group, and they are the only $0$-cochains which are invariant by this group.

**Definition 6.** Suppose that $\mathbf{S}$ and $\mathbf{Q}$ are invariant by the unitary group, as is $U\mathbf{B}$, we say that an $m$-cochain $F$ is *invariant*, if for every $X$ in $\mathbf{S}$ dividing $Y_1,...,Y_m$ in $\mathbf{S}$, every $\rho$ in $\mathbf{Q}_X$ and every $g$ in the group $U(h_0)$, we have

$$F_{g.X}(g.Y_1,...,g.Y_m;g.\rho)=F_X(Y_1;...;Y_m;\rho);\tag{88}$$

where $g.X=gXg^*$, $g.Y_i=gY_ig^*;i=1,...,m$ and $g.\rho=g\rho g^*$.

This is compatible with the naturality assumption (functoriality by direct images), because direct image is a covariant operation.

Note that conditioning is also covariant if we change all variables and laws coherently. Thus the action of the monoids $\mathbf{S}_X$ on cochains respects the invariance.

Then the coboundary $\hat{\delta}$ preserves invariance. Thus the co-homology of the invariant co-chains is well defined. We call it the *invariant information co-homology*, and we will denote it by $H^*_U(\mathbf{S};\mathbf{Q})$, $U$ for unitary.

Invariant co-cochains form a subcomplex of ordinary cochains, then we have a well defined map from $H^*_U(\mathbf{S};\mathbf{Q})$ to $H^*(\mathbf{S};\mathbf{Q})$.

The invariant $0$-co-chains depend only on the spectrum of $\rho$ in the sets $\mathbf{Q}_X$.

The invariant co-homology is probably a more natural object from the point of view of Physics. It is also on this co-homology that we were able to obtain constructive results.

The classical entropy of the decomposition $\{E_j\}$ and the quantum law $\rho$ is

$$H(X;\rho) = -\sum_j Tr(E_j^*\rho E_j)\log_2(Tr(E_j^*\rho E_j)) \tag{89}$$

In general it is not true that $H(X;\rho) = H(Y;Y_*\rho)$ when $X$ divides $Y$. Thus the Shannon (or Gibbs) entropy is *not* a local $0$-cochain, but it is a local $1$-cochain, *i.e.*, if $X \to Y \to Z$ we have

$$H_X(Z;\rho_X) = H_Y(Z;Y_*\rho_X), \tag{90}$$

Moreover it is a spectral $1$-cochain for any $\mathbf{Q}^{min}$.

The following result is well known, *cf.* Nielsen and Chuang [13].

**Lemma 5.** Let $X, Y$ be two commuting families of observables; we have

$$S_{(X,Y)}(\rho) = H(Y;\rho) + Y.S_X(\rho) \tag{91}$$

**Proof.** We denote by $\alpha, \beta, ...$ the indices of the different values of $X$, by $k, l, ...$ the indices of the different values of $Y$, and by $i, j, ...$ the indices of a basis $I_{k,\alpha}$ of eigenvectors of the conditioned density $\rho_{k,\alpha} = E^*_{k,\alpha}\rho E_{k,\alpha}$ constrained by the projectors $E_{k,\alpha}$ of the pair $(Y, X)$. The probability $p_k = P_\rho(X = \xi_k)$ is equal to the sum over $i, \alpha$ of the eigenvalues $\lambda_{i,k,\alpha}$ of $\rho_{k,\alpha}$. We have

$$\begin{aligned} Y.S(X;\rho) &= -\sum_k p_k \sum_{i,\alpha} \frac{\lambda_{i,k,\alpha}}{p_k}\log_2(\frac{\lambda_{i,k,\alpha}}{p_k}) \\ &= -\sum_{i,k,\alpha} \lambda_{i,k,\alpha}\log_2(\lambda_{i,k,\alpha}) + \sum_{i,k,\alpha}\lambda_{i,k,\alpha}\log_2(p_k) \\ &= -\sum_{i,k,\alpha} \lambda_{i,k,\alpha}\log_2(\lambda_{i,k,\alpha}) + \sum_k p_k\log_2(p_k). \end{aligned}$$

**Remark 6.** Taking $X = 1$, or any scalar matrix, the preceding Lemma 5 expresses the fact that classical entropy is a derived quantity measuring the default of equivariance of the quantum entropy:

$$H(Y;\rho) = S_Y(\rho) - (Y.S_Y)(\rho). \tag{92}$$

**Lemma 6.** For any $X \in \mathbf{S}$, dividing $Y \in \mathbf{S}$ and $\rho \in \mathbf{Q}_X$,

$$\hat{\delta}(S_X)(Y;\rho) = -H_X(Y;\rho). \tag{93}$$

**Proof.** This is exactly what says the Lemma 5 in this particular case, because in this case $(X, Y) = X$, and, by definition, we have $\hat{\delta}(S_X)(Y;\rho) = Y.S_X(\rho) - S_X(\rho)$.
To insist, we give a direct proof with less indices for this case:

$$\begin{aligned} Y.S_X(\rho) &= -\sum_i p_i \sum_k \frac{\lambda_{ik}}{p_i} \log_2 \frac{\lambda_{ik}}{p_i} \\ &= -\sum_{ik} \lambda_{ik} \log_2 \lambda_{ik} + \sum_{ik} \lambda_{ik} \log_2 p_i \\ &= S_X(\rho) + \sum_i \log_2 p_i \sum_k \lambda_{ik} = S_X(\rho) + \sum_i (\log_2 p_i) p_i \\ &= S_X(\rho) - H_X(Y; \mathbb{P}_\rho) = S_X(\rho) - H_X(Y;\rho). \end{aligned}$$

The Lemma 6 says that (up to the sign) the Shannon entropy is the co-boundary of the Von-Neumann entropy. This implies that the Shannon entropy is a 1-co-cycle, as in the classical case, but now it gives zero in co-homology.

Note that the result is true for any $\mathbf{Q}$, thus for $\mathbf{Q}^{min}$ and for $\mathbf{Q}^{max}$ as well.

Consider a maximal observable $X_0$ in $\mathbf{S}$, *i.e.*, a maximal set of commuting observables in $\mathbf{S}$, the elements of this maximal partition form a finite set $\Omega_0$. If $\mathbf{S}$ is invariant by the group $U(E; h_0)$, all the maximal observables are deduced from $X_0$ by applying a unitary base change. Suppose that the functor $\mathbf{Q}$ is invariant also; then we get automatically a symmetric classical structure of information $\mathcal{S}$ on $\Omega_0$, given by the elements of $\mathbf{S}$ divided by $X_0$. And $\mathcal{S}$ is equipped with a symmetric classical functor of probability, given by the probability laws associated to the elements of $\mathcal{S}$.

Remind that we defined the *trace* from quantum probabilities to classical probabilities, by taking the classical $\mathbb{P}_\rho$ for each $\rho$, and we noticed that the trace is compatible with invariance and symmetry by permutations.

**Definition 7.** To each classical co-chain $F^0$ we can associate a quantum co-chain $F = tr^* F^0$ by putting

$$tr^*(F)_X(Y_1; ...; Y_m; \rho) = F^0_X(Y_1; ...; Y_m; tr_X(\rho)). \tag{94}$$

The following result is straightforward:

**Proposition 3.** $(i)$ The trace of co-chains defines a map of the classical information Hochschild complex to the quantum one, which commutes with the co-boundaries, i.e., the map $tr^*$ defines a map from the classical information Hochschild complex to the quantum Hochschild complex; $(ii)$ this map sends symmetric cochains to invaraint cochains; it induces a natural map from the symmetric classical information co-homology $H^*_{\mathfrak{S}}(\mathcal{S}; \mathcal{Q})$ to the invariant quantum information co-homology $H^*_U(\mathbf{S}; \mathbf{Q})$.

The Lemma 6 says that the entropy class goes to zero.

**Remark 7.** In a preliminary version of these notes, we considered the expression $s(X;\rho) = S(\rho_X) - S(\rho)$ and showed it satisfies formally the 1-cocycle equation. But we suppress this consideration now,

because $s$ is not local, thus it plays no interesting role in homology. For instance in $\mathbf{Q}^{min}$, $S(\rho_X)$ is local but $S(\rho)$ is not and in $\mathbf{Q}^{max}$, $S(\rho)$ is local but $S(\rho_X)$ is not.

**Definition 8.** In an information structure $\mathbf{S}$ we call *edge* a pair of decompositions $(X, Y)$ such that $X, Y$ and $XY$ belong to $\mathbf{S}$; we say that an edge is *rich* when both $X$ and $Y$ have at least two elements and $XY$ cuts those two in four distinct subspaces of $E$. The structure $\mathbf{S}$ is connected if every two points are joined by a sequence of edges, and it is *sufficiently rich* when every point belongs to a rich edge. We assume a maximal set of subspaces $U\mathbf{B}$ is given in the Grassmannian of $E$, in such a way that the maximal elements $X_0$ of $\mathbf{S}$ (i.e., initial in the category) are made by pieces in $U\mathbf{B}$. The density functor $\mathbf{Q}$ is said *complete* with respect to $\mathbf{S}$ (or $U\mathbf{B}$) if for every $X$, the set $\mathbf{Q}_X$ contains the positive hermitian forms on the blocs of $X$, that give scalar blocs $\rho_{\alpha\beta}$ for two elements $E_\alpha, E_\beta$ of a maximal decomposition. (All that is simplified when we choose a basis, and take maximal commutative subalgebras of operators, but we want to be free to consider simplicial complexes.)

**Theorem 3.** $(i)$ for any unitary invariant quantum information structure $\mathbf{S}$, which is connected and sufficiently rich, and for the canonical invariant density functor $\mathbf{Q}^{can}(\mathbf{S})$, (i.e., the density functor which is minimal and complete with respect to $\mathbf{S}$), the invariant information co-homology of degree one $H^1_U(\mathbf{S}; \mathbf{Q})$ is zero. $(ii)$ Under the same hypothesis, the invariant co-homology of degree zero has dimension one, and is generated by the constants. Then, up to an additive constant, the only invariant $0$-cochain which has the Shannon entropy as co-boundary is (minus) the Von-Neumann entropy.

**Proof.** (I) Let $X, Y$ be two orthogonal decompositions of $E$ belonging to $\mathbf{S}$ such that $(X, Y)$ belongs to $\mathbf{S}$, and $\rho$ an element of $\mathbf{Q}$. We name $A_{k_i}; i = 1, ..., m$ the summands of $X$, and $B_{\alpha_j}; j = 1, ..., l$ the summands of $Y$; the projections $E_{k_i}\rho E_{k_i}; i = 1, ..., m$ resp. $E_{\alpha_j}\rho E_{\alpha_j}; j = 1, ..., l$ of $\rho$ on the summands of $X$, resp. $Y$ are denoted by $\rho_{k_i}; i = 1, ..., m$ and $\rho_{\alpha_j}; j = 1, .., l$ respectively. The projections by the commutative products $E_{k_i}E_{\alpha_j}$ are denoted by $\rho_{k_i,\alpha_j}; i = 1, ..., m, j = 1, .., l$.

Let $f$ be a 1-cocycle, we write $f(X; \rho) = F(\rho)$, $f(Y; \rho) = H(\rho)$ and $G(\rho) = f(X, Y; \rho)$. Note that in $\mathbf{Q}^{min}$, $F$ is a function of the $\rho_{k_i}$, $H$ a function of the $\rho_{\alpha_j}$ and $G$ a function of the $\rho_{k_i,\alpha_j}$, but there is no necessity too assume this property; we can always consider these functions restricted to diagonal blocs, which are arbitrary due to the completeness hypothesis.

For any positive hermitian $\rho'$, we write $\rho'|\alpha$, resp. $\rho'|i$ the form conditioned by the event $B_\alpha$ resp. $A_i$.

The co-cycle equation gives the two following equations, that are exchanged by permuting $X$ and $Y$:

$$\sum_{\alpha_j} Tr(\rho_{\alpha_j}) F((\rho_{k_i}|\alpha_j); i = 1, ..., m) = G((\rho_{k_i,\alpha_j}); i, j) - H((\rho_{\alpha_j}); j), \tag{95}$$

$$\sum_{i} Tr(\rho_{k_i}) H((\rho_{\alpha_j}|k_i); j) = G((\rho_{k_i,\alpha_j}); i, j) - F((\rho_{k_i}); i). \tag{96}$$

Now we consider a particular case, where the small blocs $\rho_{k,\alpha}$ are zero except for $(k_1, \alpha_2)$ and $(k_j, \alpha_1)$ for $j = 2, ..., m$. We denote by $h_1$ the forme $\rho_{k_1,\alpha_2}$ and by $h_i$ the form $\rho_{k_i,\alpha_1}$, for $i = 2, ..., m$. Remark that $Tr(h_1 + h_2 + ... + h_m) = 1$.

(II) As in the classical case, it is a general fact for a 1-cocycle $f$ and any variable $Z$ the value $f(Z;\rho)$ is zero if $\rho$ is zero outside one of the orthogonal summand $C_a$ of $Z$; because the equation $f_X(Z,Z;\rho) = f_X(Z;\rho) + Z.f_X(Z;\rho)$ implies $Z.f_X(Z,\rho) = 0$, and if $\rho$ has only one non-zero factor $\rho_a$, we have

$$Z.f(Z;\rho) \;=\; \sum_b Tr(\rho_b) f(Z;\rho_b/Tr(\rho_b)) \;=\; Tr(\rho_a) f(Z;\rho_a/Tr(\rho_a)) \;=\; 1.f(Z;\rho_a). \quad (97)$$

Therefore in the particular case that we consider, we get for any $i$ that $H((\rho_{\alpha_j}|k_i);j) = 0$. Consequently the Equation (96) equals the term in $G$ to the term in $F$, and we can report this equality in the first equation. By denoting $1 - x_1 = Tr(\rho_{\alpha_1})$, this gives

$$H((\rho_{\alpha_j}); j = 1,2) = F((\rho_{k_i}); i = 1,...,m) - (1-x_1) F((0, \frac{h_2}{1-x_2}, ..., \frac{h_m}{1-x_m})). \quad (98)$$

Now if we add the condition $h_3 = ... = h_m = 0$ we have $F(0, h_2/(1-x_1), 0, ..., 0) = 0$ for the reason which eliminated the $H((\rho_{\alpha_j}|k_i);j)$; thus we obtain

$$H(\rho_{\alpha_1}); j = 1,2) = F((\rho_{k_1}); i = 1,2). \quad (99)$$

This is a sufficiently strong constraints for implying that both terms are functions of $h_1, h_2$ only, and that of course they coincide as functions of these small blocs.
First this gives a recurrence equation, which, as in the classical case is able to reconstruct $F((\rho_{k_i}); i = 1,...,m)$ from the case of two blocs:

$$F(X;(\rho_{k_i}); i = 1,...,m) = F(X;(\rho_{k_1}, \rho_{k_2}, 0, ..., 0) - (1-x_1) F(X;(0, \frac{h_2}{1-x_2}, ..., \frac{h_m}{1-x_m})). \quad (100)$$

(III) We are left with the study of two binary variables $Y, Z$, forming a rich edge.

The blocs of $\rho$ adapted to the joint $ZY$ are denoted by $\rho_{00}, \rho_{01}, \rho_{10}, \rho_{11}$, where the first index refers to $Y$ and the second index refers to $Z$, but the blocs that are allowed for $Y$ and $Z$ are more numerous than four; there exist out of diagonal blocs, and their role will be important in our analysis. For $Y$ we have matrices $\rho_0^0$ and $\rho_1^0$, and for $Z$ we have matrices $\rho_0^1$ and $\rho_1^1$;

$$\rho_0^0 = \begin{pmatrix} \rho_{00} & \rho_{001}^0 \\ \rho_{010}^0 & \rho_{01} \end{pmatrix} \quad \rho_1^0 = \begin{pmatrix} \rho_{10} & \rho_{101}^0 \\ \rho_{111}^0 & \rho_{11} \end{pmatrix} \quad (101)$$

$$\rho_0^1 = \begin{pmatrix} \rho_{00} & \rho_{001}^1 \\ \rho_{010}^1 & \rho_{01} \end{pmatrix} \quad \rho_1^1 = \begin{pmatrix} \rho_{10} & \rho_{101}^1 \\ \rho_{111}^1 & \rho_{11} \end{pmatrix} \quad (102)$$

They are disposed in sixteen blocs for $\rho$, but certain of them, noted with stars, cannot be seen from $\rho_Y$ or $\rho_Z$:

$$\rho = \begin{pmatrix} \rho_{00} & \rho_{001}^0 & \rho_{001}^1 & \rho_{001}^* \\ \rho_{010}^0 & \rho_{01} & \rho_{101}^* & \rho_{001}^1 \\ \rho_{010}^1 & \rho_{010}^* & \rho_{10} & \rho_{101}^0 \\ \rho_{111}^* & \rho_{111}^1 & \rho_{111}^0 & \rho_{11} \end{pmatrix} \quad (103)$$

Now the co-cycle equations are

$$F(Y,Z;\rho) = Y.F(Z;\rho) + F(Y;\rho) = Z.F(Y;\rho) + F(Z;\rho), \tag{104}$$

giving the symmetrical relation:

$$Y.F(Z;\rho) - F(Z;\rho) = Z.F(Y;\rho) - F(Y;\rho). \tag{105}$$

The conditioning makes many blocs disappear. Then, by denoting with latin letters the corresponding traces, and taking in account explicitly the blocs that must count, the symmetrical identity gives, for any $\rho$, the following developed equation:

$$\begin{aligned} & (p_{00}+p_{01})F_Z(\frac{\rho_{00}}{p_{00}+p_{01}}, \frac{\rho_{01}}{p_{00}+p_{01}}, 0, 0) + (p_{10}+p_{11})F_Z(0, 0, \frac{\rho_{10}}{p_{10}+p_{11}}, \frac{\rho_{01}}{p_{10}+p_{11}}) \\ & -F_Z(\rho_{00}, \rho^1_{001}, \rho_{01}, \rho^1_{001}, \rho^1_{010}, \rho_{10}, \rho^1_{111}, \rho_{11}) \\ = \; & (p_{00}+p_{10})F_Y(\frac{\rho_{00}}{p_{00}+p_{10}}, 0, \frac{\rho_{10}}{p_{00}+p_{11}}, 0) + (p_{01}+p_{11})F_Y(0, \frac{\rho_{10}}{p_{01}+p_{11}}, 0, \frac{\rho_{11}}{p_{01}+p_{11}}) \\ & -F_Y(\rho_{00}, \rho^0_{001}, \rho_{01}, \rho^0_{001}, \rho^0_{010}, \rho_{10}, \rho^0_{111}, \rho_{11}). \end{aligned} \tag{106}$$

(IV) Now we make appeal to the invariance hypothesis: let us apply a unitary transformation $g$ which respects the two summands of $Y$ but does not necessarily respect the summands of $Z$ we replace $Z$ by $gZg^*$, and $\rho$ by $g\rho g^*$, the value of $F_Y(\rho^0_0, \rho^0_1)$ does not change. Our claim is that the only function $F_Y$ which is compatible with the Equation (106) for every $\rho$ are functions of the traces of the blocs.

For the proof, we assume that all the blocs are zero except the eight blocs concerning $Y$. In this case, we see that the last function $-F_Y$ of the right member, involves the eight blocs, but all the other functions involve only the four diagonal blocs. Thus our claim follows from the following result:

**Lemma 7.** A measurable function $f$ on the set $H$ of hermitian matrices which is invariant under conjugation by the unitary group $U_n$ and invariant by the change of the coefficient $a_{1n}$, the farthest from the diagonal, is a function of the trace.

**Proof.** An invariant function for the adjoint representation is a function of the traces of the exterior powers $\Lambda^k(\rho)$, but these traces are coefficients in the basis $e_{i_1} \wedge e_{i_1} \wedge ... \wedge e_{i_k}$, and the elements divisible by $e_1 \wedge e_n$ cannot be neglected, as soon as $k \geq 2$.

Therefore the co-cycle $F_Y, F_Z$ comes from the image of $tr^*$ in proposition 3. Then the recurrence relation (100) implies that the same is true for the whole co-cycle $F$.

(V) For concluding the proof of $(i)$, we appeal to the Theorem 1, that the only non-zero cocycles in this context, connected and sufficiently rich, are multiples of the classical entropy. However, the Lemma 5 says that the entropy is a co-boundary.

(VI) To prove $(ii)$, we have to show that every 0-cocycle $X \mapsto f_X(\rho)$, which depends only on the spectrum of $\rho$, is a constant. We know that a spectral function is a measurable function $\varphi(\sigma_1, \sigma_2, ...)$ of the elementary symmetric functions $\sigma_1 = \sum_i \lambda_i$,$\sigma_2 = \Sigma_{i<j}\lambda_i\lambda_j$,....

And, to be a 0-cocycle, $f$ must verify, for every pair of decompositions, $X \to Y$, the equation

$$f_X(\rho) = \sum_i P_\rho(Y = i) f_X(\rho|(Y = i)). \tag{107}$$

Explicitly, if $f_X(\rho) = \varphi_X(\sigma_1, \sigma_2, ...)$,

$$\varphi_X(\sigma_1, \sigma_2, ...) = \sum_i \sigma_1(\lambda_{k,i}) \varphi_X(\sigma_1(\lambda_{ki}), ...) \tag{108}$$

where each bloc $\rho|i$ has the spectrum $\{\lambda_{k,i}; k \in J_i\}$. For a sufficiently rich edge $X = YZ$, we have with four eigenvalues repeated as it must be to fulfill the dimensions:

$$\begin{aligned} & f(\lambda_{00}^{(n_{00})}, n_{00}^{(n_{00})}, \lambda_{01}^{(n_{01})}, \lambda_{10}^{(n_{10})}, \lambda_{11}^{(n_{11})}) \\ = \; & (n_{00}\lambda_{00} + n_{01}\lambda_{01}) f(\frac{\lambda_{00}^{(n_{00})}}{n_{00}\lambda_{00} + n_{01}\lambda_{01}}, \frac{\lambda_{01}^{(n_{01})}}{n_{00}\lambda_{00} + n_{01}\lambda_{01}}) \\ & + (n_{10}\lambda_{10} + n_{11}\lambda_{11}) f(\frac{\lambda_{10}^{(n_{10})}}{n_{10}\lambda_{10} + n_{11}\lambda_{11}}, \frac{\lambda_{11}^{(n_{11})}}{n_{10}\lambda_{10} + n_{11}\lambda_{11}}), \end{aligned} \tag{109}$$

and

$$\begin{aligned} & f(\lambda_{00}^{(n_{00})}, n_{00}^{(n_{00})}, \lambda_{01}^{(n_{01})}, \lambda_{10}^{(n_{10})}, \lambda_{11}^{(n_{11})}) \\ = \; & (n_{00}\lambda_{00} + n_{10}\lambda_{10}) f(\frac{\lambda_{00}^{(n_{00})}}{n_{00}\lambda_{00} + n_{10}\lambda_{10}}, \frac{\lambda_{10}^{(n_{10})}}{n_{00}\lambda_{00} + n_{10}\lambda_{10}}) \\ & + (n_{01}\lambda_{01} + n_{11}\lambda_{11}) f(\frac{\lambda_{01}^{(n_{01})}}{n_{01}\lambda_{01} + n_{11}\lambda_{11}}, \frac{\lambda_{11}^{(n_{11})}}{n_{01}\lambda_{01} + n_{11}\lambda_{11}}), \end{aligned} \tag{110}$$

By equating the two second members, taking $\lambda_{01} = \lambda_{00} = 0$, and varying $\lambda_{10}, \lambda_{11}$, we find that $f(x, y)$ is the sum of a constant and a linear function.

At the end, $f_X$ must be the sum of a constant and a linear function for every $X$. However, a linear symmetric function is a multiple of $\sigma_1$. As $\rho$ is normalized by the condition $Tr(\rho) = 1$, only the constant survives.

**Remark 8.** In his book "Structure des Systemes Dynamiques", J-M. Souriau [48] showed that the mass of a mechanical system is a degree one class of co-homology of the relativity group with values in its adjoint representation; this class being non-trivial for classical Mechanics, with the Galileo group, and becoming trivial for Einstein relativistic Mechanics, with the Lorentz-Poincare group. Even if we are conscious of the big difference with our construction, the above result shows the same thing happens for the entropy, but going from classical statistics to quantum statistics.

From the philosophical point of view, it is important to mention that the main difference between classical and quantum information co-homology in degree less than one, is the fact that the certitude, 1, becomes highly non-trivial in the quantum context. This point is discussed in particular by Gabriel Catren [49]. In geometric quantization the first ingredient, discovered by Kirillov, Kostant and Souriau in the sixties, is a circular bundle over the phase space that allows a non-trivial representation

of the constants. The second ingredient also discovered by the same authors, is the necessity to choose a polarization, which correspond to the choice of a maximal commutative Poisson sub-algebra of observable quantities. This second ingredient appears in our framework through the limitations of information categories to collection of commutative Boolean algebras, coming from the impossibility to define manageable joints for arbitrary pair of observables.

## 5. Product Structures, Kullback–Leibler Divergence, Quantum Version

In this short section, we use both the homogeneous bar-complex and the non-homogeneous complex.

A natural extension of the information co-cycles is to look at the measurable functions

$$F(X_0; X_1; ...; X_m; P_0; P_1, P_2, ..., P_n; X), \tag{111}$$

of several probability laws $P_j$ (or density of states respectively) on $\Omega$ (or $E$ respectively) belonging to the space $\mathcal{Q}_X$ that are absolutely continuous with respect to $P_0$, and several decompositions $Y_i$ less fine than $X$. To be homogeneous co-chains these functions have to behave naturally under direct image $Y_*(P_i)$, and to satisfy the equivariance relation:

$$\begin{aligned} & F((Y, X_0); (Y, X_1); ...; (Y, X_m); P_0; P_1, P_2, ..., P_n; X) \\ = \; & Y.F(X_0; X_1; ...; X_m; P_0; P_1, P_2, ..., P_n; X), \end{aligned} \tag{112}$$

for any $Y \in \mathcal{S}_X$ (resp. $\mathbf{S}_X$), where

$$\begin{aligned} & Y.F(X_0; X_1; ...; X_m; P_0; P_1, P_2, ..., P_n; X) \\ = \; & \int_{E_Y} dY_* P_0(y) F(X_0; X_1; ...; X_m; P_0|Y = y; P_1|Y = y, ..., P_n|Y = y; X). \end{aligned} \tag{113}$$

Note that a special role is played by the law $P_0$, which justifies the coma notation.

The proof of the Lemma 1 in Section 2.1 extends without modification to show that this defines an action of semi-group.

Then we define the homogeneous co-boundary operator by

$$\begin{aligned} & \delta F(X_0; X_1; ...; ...; X_m; X_{m+1}; P_0; P_1, P_2, ..., P_n; X) \\ = \; & \sum_i (-1)^i F(X_0; ...; \widehat{X_i}; ...; X_m; X_{m+1}; P_0; P_1, P_2, ..., P_n; X). \end{aligned} \tag{114}$$

The co-cycles are the elements of the kernel of $\delta$ and the co-boundaries the elements of the image of $\delta$ (with a shift of degree). The co-homology groups are the quotients of the spaces of co-cycles by the spaces of co-boundaries.

This co-homology is the topos co-homology $H^*_{\mathcal{S}}(\mathbb{R}, \mathcal{F}_n)$, of the module functor $\mathcal{F}_n$ of measurable functions of $n+1$-uples of probabilities, in the ringed topos $\mathcal{S}$ (resp. $\mathbf{S}$ in the quantum case).

There is also the non-homogeneous version: a $m$-cocycle is a family of functions $F_X(X_1; ...; ...; X_m; P_0; P_1, P_2, ..., P_n)$ which behave naturally under direct images, without equivariance condition.

The co-boundary operator is copied on the Hochschild operator: then we define the homogeneous co-boundary operator by

$$\begin{aligned} & \widehat{\delta}F_X(X_0;X_1;...;...;X_m;P_0;P_1,P_2,...,P_n) \\ = \; & (X_0.F_X)(X_1;...;...;X_m;P_0;P_1,P_2,...,P_n) \\ & +\sum_i (-1)^{i+1}F(X_0;...;\widehat{X_i};...;X_m;P_0;P_1,P_2,...,P_n;X). \end{aligned} \tag{115}$$

Let us recall the definition of the Kullback–Leibler divergence (or relative entropy) between two classical probability laws $P, Q$ on the same space $\Omega$, in the finite case:

$$H(P;Q) = -\sum_i p_i \log \frac{q_i}{p_i}. \tag{116}$$

Over an infinite set, it is required that $Q$ is absolutely continuous with respect to $P$ with a $L^1$-density $dQ/dP$, and the definition is

$$H(P;Q) = -\int_\Omega dP(\omega) \log \frac{dQ(\omega)}{dP(\omega)}. \tag{117}$$

When $dQ(\omega)/dP(\omega) = 0$, the logarithm is $-\infty$ and due to the sign minus, we get a contribution $+\infty$ in $H$, thus, if this happens with probability non-zero for $P$ the divergence is infinite positive. To get a finite number we must suppose also that $P$ is absolutely continuous with respect to $Q$, *i.e.*, $P$ and $Q$ are equivalent.

The analogous formula defines the quantum Kullback–Leibler divergence (or quantum relative entropy), *cf.* Nielsen-Chuang [13], between two density of states $\rho, \sigma$ on the same Hilbert space $E$, in the finite dimensional case:

$$S(\rho;\sigma) = -Tr(\rho(\log\sigma - \log\rho)). \tag{118}$$

In the case of an infinite dimensional Hilbert space, it is required that the trace is well defined.

These quantities are positive or zero, and they are zero only in the case of equality of the measures (resp. the densities of states). It is the reason why it is frequently used as a measure of distance between two laws.

**Proposition 4.** The map which associates to $X$ in $\mathcal{S}$, $Y$ divided by $X$, and two laws $P, Q$ the quantity $H(Y_*P;Y_*Q)$ defines a non-homogeneous 1-cocycle, denoted $H_X(Y;P;Q)$.

**Proof.** As we already know that the classical Shannon entropy is a non-homogeneous 1-cocycle, it is sufficient to prove the Hochschild relation for the new function

$$H_m(Y;P;Q) = -\sum_i p_i \log q_i. \tag{119}$$

Let us denote by $p_{ij}$ (resp. $q_{ij}$) the probability for $P$ (resp. $Q$) of the event $Y = x_i, Z = y_j$, and by $p^j$ (resp. $q^j$) the probability for $P$ (resp. $Q$) of the event $Z = y_j$; then the probability $p_i^j$ (resp. $q_i^j$) of $Y = x_i$ knowing that $Z = y_j$ for $P$ (resp. for $Q$) is equal to $p_{ij}/p^j$ (resp. $q_{ij}/q^j$), and we have

$$H_m((Z,Y);P,Q) = -\sum_i \sum_j p_{ij} \log q_{ij} \tag{120}$$

$$= -\sum_j p^j \sum_i p_i^j \log(q^j q_i^j) \tag{121}$$

$$= -\sum_j p^j \log q^j (\sum_i p_i^j) - \sum_j p^j \sum_i p_i^j \log q_i^j \tag{122}$$

$$= -\sum_j p^j \log q^j - \sum_j p^j \sum_i p_i^j \log q_i^j; \tag{123}$$

the first term on the right is $H_m(Z;P;Q)$ and the second is $(Z.H_m)(Y;P;Q)$,
Q.E.D.

This defines a homogeneous co-cycle for pairs of probability laws $H_X(Y;Z;P;Q) = H_X(Y;P;Q) - H_X(Z;P;Q)$, named *Kullback-divergence variation*.

In the quantum case, for two densities of states $\rho, \sigma$ we define in the same manner a classical Kullback–Leibler divergence $H_X(Y;\rho;\sigma)$ by the formula

$$H_X(Y;\rho;\sigma) = \sum_k (Tr(\rho_k \log(Tr(\rho_k)) - \log(Tr(\sigma_k)))); \tag{124}$$

where the index $k$ parameterizes the orthogonal decomposition $E_k$ associated to $Y$ and where $\rho_k$ (resp. $\sigma_k$) denotes the matrix $E_k^* \rho E_k$ (resp. $E_k^* \sigma E_k$). It is the Kullback–Leibler divergence of the classical laws associated to the direct images $\rho$ and $\sigma$ respectively.

But in the case of quantum information theory, we can also define a quantum divergence, for any pair densities of states $(\rho, \sigma)$ in $\mathbf{Q}_X$,

$$S_X(\rho;\sigma) = -Tr(\rho \log \sigma). \tag{125}$$

**Lemma 8.** For any pair $(X,Y)$ of commuting hermitian operators, such that $Y$ divides $X$, the function $S_X$ satisfies the relation

$$S(X,Y)(\rho;\sigma) = H_Y(X;\rho;\sigma) + X.S_Y(\rho;\sigma); \tag{126}$$

where $H_X$ of two variables denotes the mixed entropy, defined by Equation (119).

**Proof.** As in the proof of the Lemma 4, we denote by $\alpha, \beta, ...$ (resp. $k, l, ...$) the indices of the orthogonal decomposition $Y$ (resp. $X$), and by $i, j, ...$ the indices of a basis $\phi_{i,k,\alpha}$ of the space $E_{k,\alpha}$ made by eigenvectors of the matrix $\varrho_{k,\alpha} = E_{k,\alpha}^* \rho E_{k,\alpha}$ belonging to the joint operator $(X,Y)$. In a general manner if $M$ is an endomorphism of $E_{k,\alpha}$ we denote by $M_{i,k,\alpha}$ the diagonal coefficient of index $(i,k,\alpha)$. The probability $p_k$ (resp. $q_k$) for $\rho$ (resp. $\sigma$) of the event $X = \xi_k$ is equal to the sum over $i, \alpha$ of the eigenvalues $\lambda_{i,k,\alpha}$ of $\rho_{k,\alpha}$ (resp. $\mu_{i,k,\alpha}$ of $\sigma_{k,\alpha}$). And the restricted density $\rho^{Y_k}$ (resp.

$\sigma^{Y_k}$), conditioned by $X = \xi_k$, is the sum over $\alpha$ of $\varrho_{k,\alpha}$ (resp. of $\sigma_{k,\alpha}$) divided by $p_k$ (resp. $q_k$). We have

$$X.S_Y(\rho;\sigma) = -\sum_k p_k Tr(\rho_{Y_k} \log \sigma^{Y_k}) \tag{127}$$

$$= -\sum_k p_k \sum_{i,\alpha} \frac{\lambda_{i,k,\alpha}}{p_k} (\log \frac{\sigma_k}{q_k})_{i,k,\alpha} \tag{128}$$

$$= \sum_{i,k,\alpha} \lambda_{i,k,\alpha} \log q_k - \sum_{i,k,\alpha} \lambda_{i,k,\alpha} (\log \sigma_k)_{i,k,\alpha} \tag{129}$$

$$= \sum_k p_k \log q_k - Tr(\rho_{k,\alpha} \log(\sigma_{k,\alpha}) \tag{130}$$

$$= -H_Y(X;\rho;\sigma) + S_{(X,Y)}(\rho;\sigma). \tag{131}$$

As a corollary, with the argument proving the Lemma 5 from the Lemma 4, we obtain that the classical Kullback divergence is minus the co-boundary of the $0$-cochain defined by the quantum divergence.

This shows that the generating function of all the co-cycles we have considered so far is the quantum $0$-cochain for pairs $S(\rho;\sigma) = -Tr(\rho \log \sigma)$.

## 6. Structure of Observation of a Finite System

Up to now the considered structures and the interventions of entropy can be considered as forming a kind of statics in information theory. The aim of this section is to indicate the elements of dynamics which could correspond. This more dynamical study could be more adapted to the known intervention of entropy in the theory of dynamical systems, as defined by Kolmogorov and Sinai.

### *6.1. Problems of Discrimination*

The problem of *optimal discrimination* consists in separating the various states of a system, by using in the most economical manner, a family of observable quantities. One can also only want to detect a state satisfying a certain chosen property. A possible measure of the cost of discrimination is the number of step before ending the process.

First, let us define more precisely what we mean by a system, a state, an observable quantity and a strategy for using observations. As before, for simplicity, the setting is finite sets.

The symbol $[n]$ denotes the set $\{1, ..., n\}$. We have $n$ finite sets $M_i$ of respective cardinalities $m_i$, and we consider the set $M$ of sequences $x_1, ..., x_n$ where $x_i$ belongs to $M_i$; by definition a *system* is a subset $X$ of $M$ and a *state* of the system is an element of $X$. The set of (classical) observable quantities is a (finite) subset $A$ of the functions from $X$ to $\mathbb{R}$.

A use of observables, named an *observation strategy*, is an oriented tree $\Gamma$, starting at its root, that is the smallest vertex, and such that each vertex is labelled by an element of $A$, and each arrow (naturally oriented edge) is labelled by a possible value of the observable at the initial vertex of the arrow.

For instance, if $F_0$ marks the root $s_0$, it means that we aim to measure $F_0(x)$ for the states; then branches issued from $t_0$ are indexed by the values $v$ of $F_0$, and to each branch $F_0 = v$ corresponds a subset $X_v$ of states, giving a partition of $X$. If $F_{1,v}$ is the observable at the final vertex $\alpha_v$ of the branch $F_0 = v$, the next step in the program is to evaluate $F_{1,v}(x)$ for $x \in X_v$; then branches issued from $\alpha_v$ corresponds to values $w$ of $F_{1,v}$ restricted to $X_v$, and so on.

For each vertex $s$ in $\Gamma$ we note $\nu(s)$ the number of edges that are necessary for joining $s$ to the root $s_0$. The function $\nu$ with values in $\mathbb{N}$ is called the *level* in the tree.

It can happen that a set $X_v$ consists of one element only; in this case we decide to extend the tree to the next levels by a branch without bifurcation, for instance by labelling with the same observable and the same value, but it could be any labelling, and its value on $X_v$. In such a way, each level $k$ gives a well defined partition $\pi_k$ of $X$.

The level $k$ also defines a sub-tree $\Gamma_k$ of $\Gamma$, such that its final branches are bearing $\pi_k$. This gives a sequence $\pi_0, \pi_1, ..., \pi_l$ of finer and finer partitions of $X$, *i.e.*, a growing sequence of partitions (if the ordering on partition is the opposite of the sense of arrows in the information category $\Pi(X)$). The tree is said *fully discriminant* if the last partition $\pi_l$, which is the finest is made by singletons.

The minimal number of steps that are necessary for separating the elements of $X$, or more modestly for detecting a certain part of states, can be seen as a measure of complexity of the system with respect to the observations $A$. A refined measure could take in account the cost of use of a given observable, for instance the difficulty to compute its values.

Standard examples are furnished by weighting problems: in this case the states are mass repartitions in $n$ objects, and allowed observables are weighting, which are functions of the form

$$F_{I,J}(x) = \sum_{i \in I} x_i - \sum_{j \in J} x_j \tag{132}$$

where $I$ et $J$ are disjoint subsets of $[n]$.

We underline that such a function, which requires the choice of two disjoint subsets in $[n]$, makes use of the definition of $M$ as a set of sequences, not as an abstract finite set.

The kind of problems we can ask in this framework were studied for instance in "*Problemes plaisants et delectables qui se font par les nombres*" from Bachet de Meziriac (1612, 1624) [50].

The starting point of our research in this direction was a particular classical problem signaled to us by Guillaume Marrelec: given $n$ objects $\xi_1, ..., \xi_n$, if we know that $m$ have the same mass and $n - m$ have another common mass, how many measures must be performed, to separate the two groups and decide which is the heavier?

Even for $m = 1$ the solution is interesting, and follows a principle of choice by maximum of entropy. In the present text we only want to describe the general structures in relation to this kind of problem without developing a specific study, in particular we want to show that the co-homological nature of the entropy extends to a more dynamical context of discrimination in time.

**Remark 9.** The discrimination problem is connected with the *coding* problem. In fact a finite system $X$ (as we defined it just before) is nothing else than a particular set of words of length $n$, where the letter appearing at place $i$ belongs to an alphabet $M_i$. Distinguishing between different words with

a set $A$ of variables $f$, is nothing else than rewriting the words $x$ of $X$ with symbols $v_f$ (labelling the image $f(X)$). To determine the most economical manner to do that, consists to find the smallest maximal length $l$ of words in the alphabet $(f, v_f); f \in A, v_f \in f(X)$ translating all the words $x$ in $X$. This translation, when it is possible, can be read on the branches of a fully discriminating rooted tree, associated to an optimal strategy, of minimal level $l$. The word that translate $x$ being the sequence $(F_0, v_0), (F_1, v_1), ..., (F_k, v_k)$, $k \leq l$, of the variables put on the vertices along the branch going from $0$ to $x$, and the values of these variables put along the edges of this branch.

*6.2. Observation Trees. Galois Groups and Probability Knowledge*

More generally, we consider as in the first part (resp. in the second part) a finite set $\Omega$, equipped with a Boolean algebra $\mathcal{B}$ (resp. a finite dimensional complex vector space $E$ equipped with a positive definite hermitian form $h_0$ and a family of direct decompositions in linear spaces $U\mathbf{B}$). In each situation we have a natural notion of *observable quantity*: in the case of $\Omega$ it is a partition $Y$ compatible with $\mathcal{B}$ (*i.e.*, less fine than $\mathcal{B}$) with numbering of the parts by the integers $1, .., k$ if $Y$ has $k$ elements; in the case of $E$ it is a decomposition $Y$ compatible with $U\mathbf{B}$ (*i.e.*, each summand is direct sum of elements of one of the decompositions $u\mathbf{B}$; for $u \in U(h_0)$), with a numbering of the summands by the integers $1, .., k$ if $Y$ has $k$ elements. We also have a notion of *probability*: in the case of $(\Omega, Y)$ it is a classical probability law $P_Y$ on the quotient set $\Omega/Y$; in the case of $(E, Y)$ it is a collection of non-negative hermitian forms $h_{Y,i}$ on each summands of $Y$.

We will consider information structures, denoted by the symbol $\mathbf{S}$, for both cases (which could be distinguished by the typography, $\mathcal{S}$ or $\mathbf{S}$, if necessary): they are categories made by objects that are observables and arrows that are divisions, satisfying the condition that if $X \in \mathbf{S}$ divides $Y$ and $Z$ in $\mathbf{S}$, then the joint $(Y, Z)$ belongs to $\mathbf{S}$.

We will also consider probability families adapted to these information structures; they form a covariant functor $X \mapsto Q_X$ (which can be typographically distinguished in the two cases by $\mathcal{Q}_X$ and $\mathbf{Q}_X$) of direct images. When $\mathcal{S}$ is a classical subcategory of the quantum structure $\mathbf{S}$, we suppose that we have a trace transformation from $\iota^*\mathbf{Q}$ to $\mathcal{Q}$, and if $\mathbf{S}$ and $\mathbf{Q}$ are unitary invariant, we remind that, thanks to the ordering, we have an equivalence of category between $\mathbf{S}^U$ and $\mathcal{S}$, and a compatible morphism from the functional module $\mathcal{F}_{\mathcal{Q}}$ to the functional module $\mathcal{F}_{\mathbf{Q}}$.

Except the new ingredient of orderings, they are familiar objects for our reader. The letter $\mathbf{X}$ will denote both cases $\Omega$ and $E$, then the letters $\mathbf{S}$, $\mathbf{B}$, $\mathbf{Q}$ will denote respectively $\mathcal{S}$, $\mathcal{B}$, $\mathcal{Q}$ or $\mathbf{S}$, $U\mathbf{B}$, $\mathbf{Q}$. Be careful that now all observable quantities are ordered, either partitions, either direct decomposition. We will always assume the compatibility condition between $\mathbf{Q}$ and $\mathbf{S}$, meaning that every conditioning of $P \in \mathbf{Q}$ by an event associated to an element of $\mathbf{S}$ belongs to $\mathbf{Q}$.

In addition we choose a subset $A$ of observables in $\mathbf{S}$, which play the role of allowed elementary observations.

We say that a bijection $\sigma$ from $\Omega$ to itself, measurable for $\mathcal{B}$, *respects* a set of observables $\mathcal{A}$ if for any $Y \in \mathcal{A}$, there exists $Z \in \mathcal{A}$ such that $Y \circ \sigma = Z$. It means that $\sigma$ establishes an ordered bijection between the pieces $Y(i)$ and the pieces $Z(i)$, *i.e.*, $x \in Z(i)$ if and only if $\sigma(x) \in Y(i)$. In

other words the permutation $\sigma$ respects $\mathcal{A}$ when the map $\sigma^*$ which associates the partition $Y \circ \sigma$ to any partition $Y$, sends $\mathcal{A}$ into $\mathcal{A}$.

In the same way, we say that $\sigma$ respects a family of probabilities $\mathcal{Q}$ if the associated map $\sigma_*$ sends an element of $\mathcal{Q}$ to an element of $\mathcal{Q}$.

In the quantum case, with $E$, $h_0$ and $U\mathbf{B}$, we do the same by asking in addition that $\sigma$ is a linear unitary automorphism of $E$.

**Definition 9.** If $\mathbf{X}$, $\mathbf{S}$, $\mathbf{Q}$, $\mathbf{B}$ and $A$ are given, the *Galois group* $G_0$ is the set of permutations of $X$ (resp. linear maps) that respect $\mathbf{S}$, $\mathbf{Q}$, $\mathbf{B}$ and $A$.

**Example 6.** Consider the system $\mathbf{X}$ associated to the simple classical weighting problem: states are parameterized by points with coordinates $0, 1$ or $-1$ in the sphere $S^{n-1}$ of radius $1$ in $\mathbb{R}^n$, according to their weights, either normal, heavier or lighter. Thus in this case $\Omega = \mathbf{X}$ possesses $2n$ points. The set $A$ of elementary observables is given by the weighting operations $F_{I,J}$, Equation (132). For $\mathcal{S}$ we take the set $\mathcal{S}(A)$ of all ordered partitions $\pi_k$ obtained by applications of discrimination trees labelled by $A$. And we consider only the uniform probability $P_0$ on $X$; in $\mathcal{Q}$ this gives the images of this law by the elements of $\mathcal{S}$, and the conditioning by all the events associated to $\mathcal{S}$.

Then the Galois group $G_0$ is the subgroup $\mathfrak{S}_n \times C_2$ of $\mathfrak{S}_{2n}$ made by the product of the permutation group of $n$ symbols by the group changing the signs of all the $x_i$ for $i$ in $[n]$.

*Proof*: the elements of $\mathfrak{S}_n$ respect $A$, and the uniform law. Moreover if $\sigma$ changes the sign of all the $x_i$, one can compensate the effect of $\sigma$ on $F_{I,J}$ by taking $G_{I,J} = F_{J,I}$, *i.e.*, by exchanging the two sides of the balance.

To finish we have to show that permutations of $X$ outside $\mathfrak{S}_n \times C_2$ do not respect $A$. First, consider a permutation $\sigma$ that does not respect the indices $i$. In this case there exists an index $i \in [n]$ such that $\sigma(i^+)$ and $\sigma(i^-)$ are states associated to different coins, for instance $\sigma(i^+) = j^+$ and $\sigma(i^-) = k^+$, with $j \neq k$, or $\sigma(i^+) = j^+$ and $\sigma(i^-) = k^-$, with $j \neq k$. Two cases are possible: these states have the same mass, or they have opposite mass. In both cases let us consider a weighting $F_{j,h}(x) = x_j - x_h$, where $h \neq k$; by applying $\sigma^* F_{j,h}$ to $x = \sigma(i^+)$ we find $+1$ (or $-1$), and by applying $\sigma^* F_{j,h}$ to $x = \sigma(i^-)$ we find $0$. However, this cannot happen for a weighting, because for a weighting, either the change of $i^+$ into $i^-$ has no effect, either it exchanges the results $+1$ and $-1$. Finally, consider a permutation $\sigma$ that respects the indices but exchanges the signs of a subset $I = \{i_1, ..., i_k\}$, with $0 < k < n$. In this case let us consider a weighting $F_{i,j}(x) = x_i - x_j$ with $i \in I$ and $j \in [n] \backslash I$, the function $F_{i,j} \circ \sigma$ takes the value $+1$ for the states $i^-, j^-$, the value $-1$ for $i^+, j^+$ and the value $0$ for the other states, which cannot happen for any weighting, because this weighting must involve both $i$ and $j$, but it cannot be $F_{j,i}(x) = x_j - x_i$, which takes the value $-1$ for $j^-$, and it cannot be $F_{i,j}$ which takes the value $+1$ for $i^+$.

The probability laws we are considering express the beliefs in initial knowledge on the system, in this case it is legitimate to consider that they constrain the initial Galois group $G_0$. This corresponds to the *Jaynes principle* [51,52].

We define in this framework the notion of *observation tree* adapted to a given subset $A$ of $\mathbf{S}$: it is a finite oriented rooted tree $\Gamma$ where each vertex $s$ is labelled by an observable $F_s$ belonging to $A$ and each arrow $\alpha$ beginning at $s$ is labelled by an element $F_s(i)$ of $F_s$. *A priori* we introduce as many branches as there exist elements in $F_s$. The disposition of the arrows in the trigonometric circular order makes that the tree $\Gamma$ is imbedded in the Euclidian plane up to homotopy.

A *branch* $\gamma$ in the tree $\Gamma$ is a sequence $\alpha_1, ..., \alpha_k$ of oriented edges, such that, for each $i$ the initial extremity of $\alpha_{i+1}$ is the terminal extremity of $\alpha_i$. Then $\alpha_{i+1}$ starts with the label $F_i$ and ends with the label $F_{i+1}$. We will say that $\gamma$ starts with the root if the initial extremity of $\alpha_1$ is the root $s_0$, with a label $F_0$.

For any edge $\alpha$ in $\Gamma$, there exists a unique branch $\gamma(\alpha)$ starting from the root, and abutting in $\alpha$. Along this branch, the vertices are decorated with the variables $F_i; i = 0, ..., F_k$ and the edges are decorated with values $v_i$ of these functions; we note

$$S(\alpha) = (F_0, v_0; F_1, v_1; ...; F_{k-1}, v_{k-1}; F_k) \tag{133}$$

By definition, the set $X(\alpha)$ of states which are *compatible with* $\alpha$ is the subset of elements of $X$ such that $F_0(x) = v_0, ..., F_{k-1}(x) = v_{k-1}$.

At any level $k$ the sets $X(\alpha)$ form a partition $\pi_k$ de $X$.

**Definition 10.** We say that an observation tree $\Gamma$ labelled by $A$ is *allowed by* $\mathbf{S}$, if all joint observable along each branch belongs to $\mathbf{S}$.
We say simply *allowed* if their is no risk of confusion.

In what follows this restriction is imposed on all considered tree. Of course if we start with the algebra of all ordered partitions this gives no restriction, but this would exclude the quantum case, where the best we can do is to take maximal commutative families.

**Definition 11.** Let $\alpha$ be an edge of $\Gamma$, we note $\mathcal{Q}(\alpha)$ the set of probability laws on $X(\alpha)$ which are obtained by conditioning by the values $v_0, v_1..., v_{k-1}$ of the observables $F_0, F_1, ..., F_{k-1}$ along the branch $\gamma(\alpha)$ starting in the root and ending with $\alpha$.

**Definition 12.** The Galois group $G(\alpha)$ is the set of permutations of elements of $X(\alpha)$ that belongs to $G_0$, preserve all the equations $F_i(x) = v_i$ (resp. all the summands of the orthogonal decomposition $F_i$ labelling the edges) and preserve the sets of probability $Q(\alpha)$ (resp. quantum probabilities).

We consider $G(\alpha)$ as embedded in $G_0$ by fixing point by point all the elements of $X$ outside $X(\alpha)$.

**Remark 10.** Let $P$ be a probability law (either classical or quantum) on $X$, $\Phi = (F_i; i \in I)$ a collection of observables, and $\varphi = (v_i; i \in I)$ a vector of possible values of $\Phi$; the law $P|(\Phi = \varphi)$ obtained by conditioning $P$ by the equations $\Phi(x) = \varphi$, is defined only if the set $X_\varphi$ of all solutions of the system of equations $\Phi(x) = \varphi$ has a non-zero probability $p_\varphi = P(X_\varphi)$. It can be viewed either as a law on $X_\varphi$, or as a law on the whole $X$ by taking the image by the inclusion of $X_\varphi$ in $X$.

**Definition 13.** The edge $\alpha$ is said *Galoisian* if the set of equations and probabilities that are invariant by $G(\alpha)$ coincide respectively with $X(\alpha)$ and $\mathcal{Q}(\alpha)$.
A tree $\Gamma$ is said *Galoisian* when all its edges are Galoisian.

At each level $k$ we define the group $G_k$ which is the product of the groups $G(\alpha)$ for the free edges at level $k$; it is a subgroup of $G_0$ preserving elements by elements the pieces of the partition $\pi_k$.

Along the path $\gamma$ the partition (or decomposition) $\pi_l$, $l \leq k$ of $X$ is increasing (finer and finer) and the sequence of groups $G_l$, $l \leq k$ is decreasing.

Along a branch the sets $X(\alpha)$ are decreasing and the sequence of groups $G_0, G(\alpha_1), ..., G(\alpha_k)$ is decreasing. We propose that the quotient $G(\alpha_{i+1})/G(\alpha_i)$ gives a measure of the *Galoisian information* gained by applying $F_i$ and obtaining the value $v_i$.

On each set $X(\alpha)$ the images of the elements of the probability family $\mathcal{Q}$ form sets $\mathcal{Q}(\alpha)$ of probabilities on $X(\alpha)$.

Thus also imposed in the group $G(\alpha)$ to preserve the set $\mathcal{Q}(\alpha)$.

**Remark 11.** In terms of coding, introducing probabilities on the $X(\alpha)$ permits to formulate the principle, that it is more efficient to choose, after the edge $\alpha$, the observation having the largest conditional entropy in $Q(\alpha)$. In what circumstances it gives the optimal discrimination tree is a difficult problem, even if the folklore admit that as a theorem. It is the problem of optimal coding.

In virtue of a Shannon's theorem, the minimal length is bounded below by entropy of the law on $X$ if this law is unique. We found it works in a simple example of weighting (*cf.* paper 3 [22]).

Note however important differences between our approach and the traditional one for coding: for us $A$ is given and $\mathcal{Q}$ is given; they correspond respectively to an *a priori* limitation of possible codes for use (like a natural language), and to a set of possible *a priori* knowledges, for instance taking in account the Galois ambiguity in the system (Jaynes principle). All that is Bayesian in spirit.

**Definition 14.** We say that an observation tree $\Gamma$ labelled by $A$ is *allowed by* $\mathbf{S}$ *and by* $X \in \mathbf{S}$, if it is allowed by $\mathbf{S}_X$, which means that all joint observable along each branch is divided by $X$.

**Definition 15.** $\mathbf{S}(A)$ is the set of (ordered) observables $\pi_k$ which can be obtained by allowed observation trees. For $X \in \mathbf{S}$ we note $\mathbf{S}_X(A)$ the set of (ordered) observables $\pi_k$ which can be obtained by observation trees that are allowed by $\mathbf{S}$ and $X$.

**Lemma 9.** The joint product defines a structure of monoid on the set $\mathbf{S}_X(A)$.

**Proof.** Let $\Gamma, \Gamma'$ be two observation trees allowed by $A$, $\mathbf{S}$ and $X \in \mathbf{S}$, of respective lengths $k, k'$, giving final decompositions $S, S'$. To establish the lemma we must show that the joint $SS'$ is obtained by a tree associated with $A$, allowed by $\mathbf{S}$ and $X$.

For that we just graft one exemplar of $\Gamma'$ on each free edge of $\Gamma$. This new tree $\Gamma\Gamma'$ is associated with $A$, and its final partition is clearly finer than $S$. It is also finer than $S'$, because at the end of any branch of $\Gamma\Gamma'$ we have an $X(\beta)$ which is contained in the corresponding element of the

final partition $\pi_{k'}(\Gamma')$. To finish the proof we have to show that each element of $\pi_{k+k'}(\Gamma\Gamma')$ is the intersection of element of $\pi_k(\Gamma)$ with one element of $\pi_{k'}(\Gamma')$, because we know these observables are in $\mathcal{S}_X$, which is a monoid, by the definition of information structure. But a complete branch $\gamma.\gamma'$ in $\Gamma\Gamma'$, going from the root to a terminal edge at level $k+k'$, corresponds to a word $(F_0, v_0, F_1, v_1, ..., F_{k-1}, v_{k-1}, F'_0, v'_0, ..., F'_{k'-1}, v'_{k'-1}$, thus the final set of the branch $\gamma.\gamma'$ is defined by the equations $F_i = v_i; i = 0, ..., k-1$ et $F'_j = v'_j; j = 0, ..., k'-1$, and is the intersection of the sets respectively defined by the first and second groups of equations, that belong respectively to $\pi_k(\Gamma)$ and $\pi_{k'}(\Gamma')$.

Then $\mathbf{S}(A)$ form an information structure. In particular there is a unique maximal partition, initial element for each subcategory $\mathbf{S}_X(A)$ in the information structure $\mathbf{S}(A)$.

But on $\mathbf{S}(A)$ the operation of *grafting*, that we will describe now, is much richer than what we used in the above Lemma 9: we can graft an allowed tree on each free edge of an allowed tree, and this introduces to a theory of operads and monads for information theory.

### *6.3. Co-Homology of Observation Strategies*

Remember that the elements of the partitions or decompositions $Y$ we are considering, are now numbered by the ordered set $\{1, ..., L(Y)\}$, where $L(Y)$ is the number of elements in the partition, or the decomposition, also called its length. In particular we consider as different two partitions which are labelled differently by the integers. This was already taken into account in the definition of the Galois groups.

We define the *multi-products* $\mu(m; n_1, ..., n_m)$ on the set of ordered partitions:

They are defined between a partition equipped with an ordering $(\pi, \omega)$ with $m$ pieces and $m$ ordered partitions $(\pi_1, \omega_1), ..., (\pi_m, \omega_m)$ of respective lengths $n_1, ..., n_m$; the results is the ordered partition obtained by cutting each piece $X_i$ of $\pi$ by the corresponding decomposition $\pi_i$ and renumbering the non-empty pieces by integers in the unique way compatible with the orderings $\omega, \omega_1, ..., \omega_m$. Observe the important fact that the result has in general less than $n = n_1 + ... + n_m$ pieces. This introduces a strong departure from usual multi-products (*cf.* P. May [17,53], Loday-Vallette [10]). We do not have an *operad*, when introducing vector spaces $V(m)$ generated by decompositions of length $m$, we get filtered but not graded structures. However a form of associativity and neutral element are preserved, hence we propose to name this structure a *filtered operads*.

There exists an evident unit to the right which is the unique decomposition of length 1.

The action of the symmetric group $\mathfrak{S}_m$ on the products is evident, and does not respect the length of the result. We will designate by $\mu_m$ the collection of products for the same length $m$.

The numbers $m_i$ between 1 and $n_i$ that counts the pieces of the decomposition of the element $X_i$ of $\pi$ are functions $m_i(\pi, \omega, \pi_i, \omega_i)$. There exists a growing injection $\eta_i : [m_i] \to [n_i]$, which depends only on $(\pi, \omega, \pi_i, \omega_i)$ telling what indices of $(\pi_i, \omega_i)$ survive in the product. These injections are integral parts of the structure of filtered operad. In particular, if we apply a permutation $\sigma_i$ to $[n_i]$, *i.e.*, if we replace $\omega_i$ by $\omega_i \circ \sigma_i$, the number can change.

The axioms of operadic unity and associativity, conveniently modified are easy to verify (*cf.* [22]). The reference we follow here is Fresse "Basic concepts of operads" [16]. For unity nothing has to be modified. For associativity (Figure 1.3 in Fresse [16]), we modify by saying that if the $(\pi_i, \omega_i)$ of lengths $n_i$, for $i$ between 1 et $k$, are composed from $\mu(n_i; n_i^1, ..., n_i^{n_i})$ with the $n_i$-uples $(..., (\pi_i^j, \omega_i^j), ...)$ whose respective lengths are $n_i^j$, and if the result $\mu_i$ for each $i$ has length $(m_i^1 + ... + m_i^{n_i})$ where $m_i^j$ is function of $(\pi_i, \omega_i)$ and $(\pi_i^j, \omega_i^j)$, then the product of $(\pi, \omega)$ of length $k$ with the $\mu_i$ is the same as the one we would have obtained by composing $\mu(k; n_1, ..., n_k)((\pi, \omega); (\pi_1, \omega_1), ...))$ with the $m = m_1 + ... + m_k$ ordered decompositions $(\pi_i^j, \omega_i^j)$ for $j$ belonging to the image of $\eta_i : [m_i] \to [n_i]$. This result is more complicate to write than to prove, because it only expresses the associativity of the ordinary join of three partitions; from which ordering follows.

Moreover, the first axiom concerning permutations (Figure 1.1 in Fresse [16]), can be modified, by considering only permutations of $n_i$ letters which preserve the images of the maps $\eta_i$.

The second axiom, which concerns a permutation $\sigma$ of $k$ elements in $\pi$, and the inverse permutation of the partitions $\pi_i$ can be reformulated by telling the effect of $\sigma$ on the multiple product $\mu$ is the same as the effect of $\sigma$ on the indices of the $(\pi_i, \omega_i)$. In other terms, the effect of $\sigma$ on $\omega$ is compensated by the action of $\sigma^{-1}$ on the indices of the $(\pi_i, \omega_i)$. One has to be careful, because the result of $\mu$ applied to $(\pi, \omega \circ \sigma)$ has in general not the same length as $\mu$ applied to $(\pi, \omega)$. However the compensation implies that $\mu_k$ is well defined on the quotient of the set of sequences $((\pi, \omega), (\pi_1, \omega_1), ...)$ by the diagonal action of $\mathfrak{S}_k$, which permutes the $k$ pieces of $\pi$ and which permutes the indices $i$ of the $n_i$ in the other factors.

Geometrically, if the partition $(\pi, \omega)$ in $\mathbf{S}(A)$ is generated by an observation tree $\Gamma$ with $m$ ending edges and the partitions $(\pi_i, \omega_i); i = 1, ..., m$ are generated by a collection of observation trees $\Gamma_i$; then the result of the application of $\mu(m; n_1, ..., n_m)$ to $(\pi, \omega)$ and $(\pi_i, \omega_i); i = 1, ..., m$ is generated by the observation tree that is obtained by grafting each $\Gamma_i$ on the vertex number $i$. Drawing the planar trees associated to three successive sets of decompositions for two successive grafting operations helps to understand the associativity property.

The fact that in general this does not give a tree with $n_1 + ... + n_m$ free edges, where $n_i$ denotes the number of free edges of $\Gamma_i$ comes from the possibility to find an empty set $X(\beta)$ at some moment along a branch of the grafted tree; this we call a *dead branch*. It expresses the fact that the empty set is excluded from the elements of a partition in the classical context, and the zero space excluded from the orthogonal decomposition in the quantum context. When computing conditioned probabilities we encounter the same problem if a set $X(\beta)$ at some place in a branch has measure zero.

The dead branches and the lack of graduation cause a lot of difficulties for studying algebraically the operations $\mu_m$, thus we introduce more flexible objects, which are the *ordered partitions with empty parts* of $\Omega$, resp. *ordered orthogonal decompositions with zero summands* of $E$: such a partition $\pi^*$ (resp. decomposition) is a family $(E_1, ..., E_m)$ of disjoint subsets of $\Omega$ (resp. orthogonal subspaces of $E$), such that their union (resp. sum) is $\Omega$ (resp. $E$). The only difference with respect to ordered partitions, resp. decompositions, is that we accept to repeat $\emptyset$ (resp. $0$) an arbitrary high number of times. For shortening we will name *generalized decompositions* these new objects.

The number $m$ is named the degree of $\pi^*$. These objects are the natural results of applying rooted observation trees embedded in an oriented half plane.

The notions of adaptation to $A$, $\mathbf{S}$ and $X$ in $\mathbf{S}$ concerning the trees, apply to the generated generalized decompositions. The corresponding sets of generalized objets are written $\mathbf{S}^*(A)$ and $\mathbf{S}^*_X(A)$.

The multi-product $\mu(m; n_1, ..., n_m)$ extends naturally to generalized decompositions, and in this case the degrees are respected, *i.e.*, the result of this operation is a generalized decomposition of degree $n_1 + n_2 + ... + n_m$.

Remark that we could write $\mu^*(m; n_1, ..., n_m)$ for the multi-products extended to generalized decompositions, however we prefer to keep the same notation $\mu(m; n_1, ..., n_m)$; this is justified by the following observation: to a generalized decomposition $\pi^*$ is associated a unique ordered decomposition $(\pi, \omega)$, by forgetting the empty sets (resp. zero spaces) in the family, and the multi-product is compatible with this forgetting application. The gain of the extension is the easy construction of a monad we expose now.

The definition of operad was introduced by P. May [17] as the right tool for studying the homology of infinite loop spaces; then it was recognized as a fundamental tool for algebraic topology, and many other topics, see Loday and Valette, Fresse.

We will encounter only "symmetric" operads.

The multiple products $\mu_m$ on generalized decompositions can be assembled in a structure of *monad* by using the standard Schur construction (*cf.* Loday et Valette [10], or Fresse, "on partitions" [16]): For each $X \in \mathbf{S}$, we introduce the real vector space $V_X = V_X(A)$ freely generated by the set $\mathbf{S}^*_X(A)$ of generalized decompositions obtained by observation trees that are allowed by $A$, $\mathbf{S}$ and $X$; the length $m$ define a graduation $V_X(m)$ of $V_X$. We put $V_X(0) = 0$.

The maps $\mu_m$ generate $m$-linear applications from products of these spaces to themselves which respect the graduation; these applications, also denoted by $\mu_m$, are parameterized by the sets $\mathbf{S}^*_X(m)$, whose elements are the generalized decompositions of degree $m$ which are divided by $X$:

$$\mu_m : V_X(m) \otimes_{\mathfrak{S}_m} V_X^{\otimes m} \to V_X \tag{134}$$

The linear *Schur functor* from the category of real vector spaces to itself, is defined by the direct sum of symmetric co-invariants:

$$\mathcal{V}_X(W) = \bigoplus_{m\geq 0} V_X(m) \otimes_{\mathfrak{S}_m} W^{\otimes m} \tag{135}$$

The composition of Schur functors is defined by

$$\mathcal{V}_X \circ \mathcal{V}_X = \bigoplus_{m\geq 0} V_X(m) \otimes_{\mathfrak{S}_m} \mathcal{V}_X^{\otimes m}. \tag{136}$$

*i.e.*, for each real vector space $W$:

$$\mathcal{V}_X \circ \mathcal{V}_X(W) = \bigoplus_{m\geq 0} \bigoplus_{l\geq m} \bigoplus_{n_1,...,n_m; \sum_i n_i = l} V_X(m) \otimes_{\mathfrak{S}_m} \bigotimes_i V_X(n_i) \otimes_{\mathfrak{S}_{n_i}} W^{\otimes n_i} \tag{137}$$

$$= \bigoplus_{l\geq 0} \bigoplus_{m\geq 0} \bigoplus_{n_1,...,n_m; \sum_i n_i = l} V_X(m) \otimes_{\mathfrak{S}_m} \bigotimes_i V_X(n_i) \otimes_{\mathfrak{S}_{n_1,...,n_k}} W^{\otimes l}; \tag{138}$$

where $\mathfrak{S}_{n_1,...,n_m}$ denotes the groups of permutations by blocs.

**Proposition 5.** For each $X$ in $\mathbf{S}$, the collection of operations $\mu_m$ defines a linear natural transformation of functors $\mu_X : \mathcal{V}_X \circ \mathcal{V}_X \to \mathcal{V}_X$; and the trivial partition defines a linear natural transformation of functors $\eta_X : \mathbb{R} \to \mathcal{V}_X$, which satisfy the axioms of a *monad* (cf. MacLane "Categories for Working Mathematician" 2nd ed. [4], and Alain Proute, Introduction a la Logique Categorique, 2013, Prepublications [54]):

$$\mu_X \circ (\mathcal{V}_X \mu_X) = \mu_X \circ (\mu_X \mathcal{V}_X), \quad \mu_X \circ (\mathcal{V}_X \eta_X) = Id = \mu_X \circ (\eta_X \mathcal{V}_X) \tag{139}$$

**Proof.** The argument is the same as the argument given in Fresse (partitions ...). The fact that the natural transformation $\mu_X$ is well defined on the quotient by the diagonal action of the symmetric group $\mathfrak{S}_m$ on $V_X(m) \otimes \bigotimes_i V_X(n_i) \otimes_{\mathfrak{S}_{n_1,...,n_m}} W^{\otimes s}$ comes from the verification of the symmetry axiom and the properties of associativity and neutral element comes from the verification of the corresponding axiom.

Moreover all these operations are natural for the functor of inclusion from the category $\mathbf{S}_Y$ to the category $\mathbf{S}_X$ of observables divided by $Y$ and $X$ respectively when $X$ divides $Y$; therefore we have the following result:

**Proposition 6.** To each arrow $X \to Y$ in the category $\mathbf{S}$ is associated a natural transformation of functors $\rho_{X,Y} : \mathcal{V}_Y \to \mathcal{V}_X$, making a morphism of monads; this defines a contravariant functor $\mathcal{V}$ from the category $\mathbf{S}$ to the category of monads, that we name the arborescent structural sheaf of $\mathbf{S}$ and $A$.

Considering the discrete topology on $\mathbf{S}$, we introduce the topos of sheaves of modules over the functor in monads $\mathcal{V}$, which we call the *arborescent information topos* associated to $\mathbf{S}$ and $A$.

As explained in Proute *loc.cit.* [54] a monad in a category $\mathcal{C}$ becomes a monoid in the category of endo-functors of $\mathcal{C}$, thus the topos we introduce is equivalent to an ordinary ringed topos.

The monad $\mathcal{V}_X$, and the contravariant monadic functor $\mathcal{V}$ on $\mathbf{S}$, are better understood by considering trees, *cf.* Getzler-Jones [55], Ginzburg-Kapranov [56] and Fresse [16]; in our context we consider all observation trees labelled by elements of $\mathbf{S}_X^* A$:
if $\Gamma$ is an oriented rooted tree of level $k$, each vertex $v$ of $\Gamma$ gives birth to $m_v$ edges; we define

$$V_X(\Gamma)(W) = \bigotimes_{v \in \Gamma} V_X(m_v) \otimes_{\mathfrak{S}_{m_v}} W^{\otimes m_v}. \tag{140}$$

The space $V(\Gamma)(W)$ is the direct sum of spaces $V_X(\Gamma_Y)(W)$ associated to trees which are decorated by a subset $Y$ in $\mathbf{S}_X^*(A)$, with one element $Y_v$ of $\mathbf{S}_X(m)$ for each vertex $v$ which gives birth to $m_v$ edges. Then the iterated functors $\mathcal{V}^{\circ k} = \mathcal{V} \circ ... \circ \mathcal{V}$ for $k \geq 1$ are the direct sums of the functors $V(\Gamma)$ of level $k$.
Remark that we could have worked directly with observation trees labelled by elements of $A$ in spite of working with generalized partitions; this would have given a strictly larger monad but equivalent results.

Associated to probability families we define now a right $\mathcal{V}_X$-module (in the terms of Fresse, Partitions, the term $\mathcal{V}_X$-algebra being reserved to a structure of left module on a constant functor).

For that we introduce the notion of *divided probability*.

**Definition 16.** A divided probability law of degree $m$ is a sequence of triplets $(p, P, U) = (p_1, P_1, U_1; ...; p_m, P_m, U_m)$, where $p_i; i = 1, ..., m$ are positive numbers of sum one, *i.e.*, $p_1 + ... + p_m = 1$, where each $P_i; i = 1, ..., m$ is a classical (resp. quantum) probability law when the corresponding $p_i$ is strictly positive, and a probability law or the empty set when the corresponding $p_i$ is equal to $0$, and where each $U_i; i = 1, ..., m$ is the support in $\mathbf{X}$ of $P_i$; moreover the $U_i$ are assumed to be orthogonal (resp. disjoint in the classical case). The letter $P$ will designate the probability $p_1P_1 + ... + p_mP_m$, where $0.\emptyset = 0$ when it happens.

The symbol $\mathcal{D}(m)$ designates the set of divided probabilities of degree $m$ on $\mathbf{X}$, and $\mathcal{D}_X(m)$ denotes the subset made with probability laws in $\mathbf{Q}_X$ adapted to a variable $X$.
The vector space generated by $\mathcal{D}_X(m)$ will be written $\mathcal{L}_X(m)$. We put $\mathcal{L}_X(0) = 0$.

We also introduce the subspace $\mathcal{K}(m)$ of $\mathcal{L}_X(m)$ which is generated by two families of vectors in $\mathcal{L}_X(m)$:

First the vectors

$$L(\lambda, p', p", P, U) = \lambda(p', P, U) + (1 - \lambda)(p", P, U) - (\lambda p' + (1 - \lambda)p", P, U), \tag{141}$$

where $\lambda$ is any real number between $0$ and $1$, and $(p', P, U)$, $(p", P, U)$ two divided probabilities associated to the same sequence of probability laws $(P_1, ..., P_m)$ and the same supports $(U_1, ..., U_m)$;

Second the vectors

$$D(p, P, U, Q, V) = (p, P, U) - (p, P', U'), \tag{142}$$

where for each index $i$ between $1$ and $m$, such that $p_i > 0$ we have $P_i = P'_i$, and consequently $U_i = U'_i$.

The we define the space of classes of divided probabilities as the quotient real vector space $\mathcal{M}_X(m) = \mathcal{L}_X(m)/\mathcal{K}(m)$. In particular $M_X(0) = 0$, $M_X(1)$ is freely generated over $\mathbb{R}$ by the elements of $\mathbf{Q}_X$.

**Lemma 10.** The space $\mathcal{M}_X(m)$ is freely generated over $\mathbb{R}$ by the vectors $(\emptyset, ..., \emptyset, P_i, \emptyset, ..., \emptyset)$ of length $m$, where at the rank $i$, $P_i$ is an element of $\mathbf{Q}_X$.

**Proof.** Let $D = (p_1, P_1, U_1), ..., (p_m, P_m, U_m)$ be a divided probability; we consider for each $i$ between $1$ and $m$ the divided probability

$$D_i = (0, P_1, U_1), ..., (0, P_{i-1}, U_{i-1}), (1, P_i, U_i), (0, P_{i+1}, U_{i+1}), ..., (0, P_m, U_m),$$

then the vector $D - \sum_i p_i D_i$ is a sum of vectors of type $L$ in $\mathcal{K}_X(m)$. However, for each $i$, the vector $D_i - (\emptyset, ..., \emptyset, P_i, \emptyset, ..., \emptyset)$ is of type $D$, thus the particular vectors of the Lemma 10 generate $\mathcal{M}_X(m)$.

Now, we prove that, if a linear combination of $r$ of these vectors belongs to $\mathcal{K}_X$, the coefficients of this combination must all be equal to $0$. We proceed by recurrence on $r$, the result being evident for

$r = 1$. We also can suppose that at least two involved vectors have a non-empty element at the same place, which we can suppose to be $i = 1$. All vectors with $p_1 = 0$ can be replaced by a vector where $P_1 = \emptyset$ using an element of type $D$ in $\mathcal{K}_X(m)$, then we can assume that at least one of the vectors has a $p_1$ strictly positive, *i.e.*, equals to 1. Let us consider all these vectors $D_1, ..., D_s$, for $2 \leq s \leq r$, their other numbers $p_i$ for $i > 1$ are zero. The other vectors $D_j$, for $j > s$ having the coordinate $p_1$ equal to zero. Let $\sum_j \lambda_j D_j$ be the linear combination of length $r$ belonging to $\mathcal{K}_X(m)$; this vector is a linear combination of vectors of type $L$ and $D$. We can suppose that every $\lambda_j$ is non-zero. Let us consider an element $Q$ of $\mathbf{Q}_X$ which appears in at least one of the $D_j$, $j \leq s$; this $Q$ cannot appear in only one $D_j$, because the sum of coefficients $\lambda$ multiplied by the first $p_1$ in front of any given $Q$ in a vector $L$ or $D$ is zero. Thus we have at least two $D_j$ with the same $P_1$. We can replace the sum of them with $\lambda_j$ positive (resp. negative) by only one special vector of the Lemma 10 using a sum of multiples of vectors of type $L$. Then we are left with the case of two vectors, $D_1, D_2$ having $P_1 = Q$ such that $\lambda_1 + \lambda_2 = 0$, which means that $\lambda_1 D_1 + \lambda_2 D_2$ is multiple of a vector of type $D$. Subtracting it we can apply the recurrence hypothesis and conclude that the considered linear relation is trivial.

As a corollary an equivalent definition of the spaces $\mathcal{M}_X(m)$ would be the real vector space freely generated by pairs $(P, i)$ where $P \in \mathbf{Q}_X$ and $i \in [m]$. Such a vector, identified with $(\emptyset, .., P, ..., \emptyset)$ in $\mathcal{L}_X(m)$, where only the place $i$ is non-empty, will be named a *simple vector* of degree $m$.

Let $S = (S_1, ..., S_m)$ be a sequence of generalized decompositions in $\mathbf{S}^*_X(A)$, of respective degrees $n_1, ..., n_m$, with $n = n_1 + ... + n_m$, and let $(p, P, U)$ be an element of $\mathcal{D}_X(m)$, we define $\theta((p, P, U), S)$ as the following divided probability of degree $n$: if, for $i = 1, ..., m$ the decomposition $S_i$ is made of pieces $E_i^{j_i}$ where $j_i$ varies between 1 and $n_i$, we take for $p_i^{j_i}$ is the classical probability $\mathbb{P}(E_i^{j_i} \cap U_i)$; we take for $P_i^{j_i}$ the law $P_i$ conditioned by the event $S_i = j_i$ which corresponds to $E_i^{j_i}$; and we take for $U_i^{j_i}$ the support of $P_i^{j_i}$. Then we order the obtained family of triples $(p_i^{j_i}, P_i^{j_i}, U_i^{j_i})_{i=1,...,m; j_i=1,...,n_i}$ by the lexicographic ordering. It is easy to verify that the resulting sequence is a divided probability.

Extending by linearity we get a linear map,

$$\lambda_m : \mathcal{L}_X(m) \otimes V_X(n_1) \otimes ... \otimes V_X(n_m) \to \mathcal{L}_X(n_1 + ...n_m), \tag{143}$$

By linearity a vector of type $L$ in $\mathcal{L}_X(m)$, tensorized with $S_1 \otimes ... \otimes S_m$ goes to a linear combination of vectors of type $L$ in $\mathcal{L}_X(n)$. Moreover, if $p_i = 0$ for an index $i$ in $[m]$, all the $p_i^{j_i}$ are zero, thus a vector of type $D$ goes to a vector of type $D$. Then the map $\lambda_m$ sends the subspace $\mathcal{K}_X(m) \otimes V_X(n_1) \otimes ... \otimes V_X(n_m)$ into the subspace $\mathcal{K}_X(n_1 + ...n_m)$, thus it defines a linear map

$$\theta_m : \mathcal{M}_X(m) \otimes V_X(n_1) \otimes ... \otimes V_X(n_m) \to \mathcal{M}_X(n_1 + ...n_m), \tag{144}$$

On a simple vector $(P, i)$, the operation $\theta_m$ is independent of the $S_j$ for $i \neq i$.

Now we introduce the Schur functor $\mathcal{M}_X$ of symmetric co-invariant spaces $\mathcal{M}_X(W) = \bigoplus_m \mathcal{M}_X(m) \otimes_{\mathfrak{S}_m} W^{\otimes m}$ from the category of real vector space to itself, associated to the $\mathfrak{S}$-module $\mathcal{M}^*_X$ (*cf.* Loday and Valette [10], Fresse [16]), formed by the graded family $\mathcal{M}_X(m); m \in \mathbb{N}$.

Then the maps $\theta_m$ define a natural transformation of functors:

$$\theta_X : \mathcal{M}_X \circ \mathcal{V} \to \mathcal{M}_X. \tag{145}$$

In addition, this set of transformations behaves naturally with respect to $X$ in the information category $\mathbf{S}$. Note that it defines a co-variant functor, not a presheaf.

For simplicity, we will note in general $\theta, \mu, \mathcal{F}, \mathcal{V}, ...$ and not $\theta_X, \mu_X, \mathcal{F}_X, \mathcal{V}_X, ...$, but we memorize this is an abuse of language.

Then the composite functor $\mathcal{M} \circ \mathcal{V}(W)$ is given by

$$\begin{aligned}\mathcal{M}_X \circ \mathcal{V}_X(W) &= \bigoplus_{m\geq 0} \mathcal{M}_X(m) \otimes_{\mathfrak{S}_m} \bigotimes_i (V_X(n_i) \otimes_{\mathfrak{S}_{n_i}} W^{\otimes n_i}) \\ &= \bigoplus_{n\geq 0} \bigoplus_{m\geq 0} \bigoplus_{n_1,...,n_m;\sum_i n_i = n} \mathcal{M}_X(m) \otimes_{\mathfrak{S}_m} \bigotimes_i V_X(n_i) \otimes_{\mathfrak{S}_{n_1,...,n_k}} W^{\otimes n};\end{aligned}$$

where $\mathfrak{S}_{n_1,...,n_m}$ denotes the groups of permutations by blocs.

**Proposition 7.** The natural transformation $\theta$ defines a right action in the sense of monads, i.e., we have

$$\theta \circ (\mathcal{F}\mu) = \theta \circ (\theta \mathcal{V}); \quad \theta \circ (\mathcal{F}\eta) = Id. \tag{146}$$

**Proof.** The proof is the same as for proposition 5, by using the associativity of conditioning, and the Bayes identity $P(A \cap B) = P(A|B)P(B)$.

Ginzburg and Kapranov [56] gave a construction of the (co)bar complex of an operad based on decorated trees. It is a graded complex of operads, with a differential operator of degree $-1$. The dual construction can be found in Getzler et Jones [55]; it gives a graded complex of co-operads with a differential operator of degree $+1$. The link with quasi-free co-operads and operads (Quillen's construction) is developed by Fresse (in "partitions" [16]); in this article Fresse also shows that these constructions correspond to the simplicial bar construction for the monads (Maclane) and to the natural notions of derived functors in this context.

In our case, with two right modules, the easiest way is to use the bar construction of Beck (1967) [19], further explicited by Fresse with decorated trees in the case of monads coming from operads.

A *morphism* from a right module $\mathcal{M}$ over $\mathcal{V}$ to a right module $\mathcal{R}$ over $\mathcal{V}$ is a natural transformation $f$ of the first functor in the second such that $f \circ \theta_M = \theta_R \circ f\mathcal{V}$.

In what follows we will use the module $\mathcal{R}$ which comes from the functor of symmetric powers:

$$\mathcal{R}(W) = \bigoplus_m S^m(W); \tag{147}$$

it is the Schur functor associated to the trivial $\mathfrak{S}_*$-module, $\mathcal{R}(m) = \mathbb{R}$, *i.e.*, the action of $\mathfrak{S}_m$ on $\mathcal{R}(m)$ is trivial. We put $\mathcal{R}(0) = \mathbb{R}$.

The right action of $\mathcal{V}_X$ is given by the map

$$\rho_m : \mathcal{R}_X(m) \otimes V_X(n_1) \otimes ... \otimes V_X(n_m) \to \mathcal{R}_X(n_1 + ...n_m), \tag{148}$$

which send each generator $(1, S_1, ..., S_m)$ to 1 in $\mathcal{R}(n) = \mathbb{R}$.

The axioms of a right module are easy to verify.

This $\mathcal{V}$-module $\mathcal{R}$ will play the dual role of the trivial module in the case of information structure co-homology.

Following Beck (Triples, Algebras, Cohomology, 1967, 2002 [19]), we consider the simplicial bar complex $\mathcal{M}_X \circ \mathcal{V}_X^*$ extending the right module $\mathcal{M}$ on $\mathcal{V}$ by the sequence of modules .... $\to \mathcal{M}_X \circ \mathcal{V}_X^{\circ(k+1)} \to \mathcal{M}_X \circ \mathcal{V}_X^{\circ k} \to$ .... Then we introduce the growing complex $C^*(\mathcal{M}_X)$ of *measurable* morphisms from $\mathcal{M}_X \circ \mathcal{V}_X^*$ to the symmetric right module $\mathbf{R}$.

For a given $k \geq 0$, a morphism $F$ from $\mathcal{M}_X \circ \mathcal{V}_X^{\circ k}$ to $\mathcal{R}$ is defined by a family of maps $F(N) : \mathcal{M}_X \circ \mathcal{V}_X^{\circ k}(N) \to \mathcal{R}(N) = \mathbb{R}$, for $N \in \mathbb{N}$.

This gives a family of measurable numerical functions of a divided probability law $(p, P, U)$, of degree $m \leq N$, indexed by forests having $m$ components trees of height $k$ and having total number of ending branches $N$.

We denote such a family of functions by the symbol $F_X(S_1; S_2; ...; S_k; (p, P, U))$, indexed by $X$ in $\mathbf{S}$, where $S_1; ...; S_k$ here designates the sets of decompositions present in the trees at each level from 1 to $k$.

First we remark that the compatibility with the action of $\mathcal{V}_X$ to the right imposes that for any allowed set of variables $S_{k+1}$ we must have

$$F_X(S_1; S_2; ...; \mu(S_k, S_{k+1}); (p, P, U)) = F_X(S_1; S_2; ...; S_k; (p, P, U)). \tag{149}$$

By taking for $S_k$ the collection $(\pi_0, ..., \pi_0)$, we deduce that $F_X$ is independent of the last variable.

This has the effect of decreasing the degree in $k$ by one, for respecting the preceding conventions on information cochains; *i.e.*, we pose $\mathcal{C}^k(M_X) = Hom(\mathcal{M}_X \circ \mathcal{V}^{\circ(k+1)}, \mathcal{R})$.

Secondly, as we are working with the quotient of the space generated by divided probabilities $(p, P, U)$ by the space generated by linearity relations on the external law $p$, for $(p, P, U)$ of degree $m$, we have

$$F_X(S_1; S_2; ...; S_k; (p, P, U)) = \sum_{i=1}^{m} p_i F_X(S_1; S_2; ...; S_k; (P_i; i, m)); \tag{150}$$

where $(Q; i, m)$ designates the divided probability of degree $m$ where all the laws in the sequence are empty except for the number $i$ where it is equal to $Q$.

Moreover, from the definition of $\theta$ and the rule of composition of functors, for any $m \geq 1$ and $i \in [m]$, and any simple vector $(Q, i, m)$, the value of $F$ on any forest depends only on the tree component of index $i$; that we can summarize by the following identity:

$$F_X(S_1; S_2; ...; S_k; (Q; i, m)) = F_X(T(S_1^i; S_2^i; ...; S_k^i); (Q; i, m)); \tag{151}$$

where $T(S_1^i; S_2^i; ...; S_k^i)$ designates the tree numbered by $i$, prolonged in any manner at all the places $j \neq i$.

**Definition 17.** An element of $\mathcal{C}^k(M_X)$ is said *regular* when for each degree $m$ and each index $i$ between 1 and $m$, we have, for each ordered forest $S_1; S_2; ...; S_k$ of $m$ trees, and each probability $Q$,

$$F_X(S_1; S_2; ...; S_k; (Q; i, m)) = F_X(S_1^i; S_2^i; ...; S_k^i; Q); \tag{152}$$

where $S_1^i; S_2^i; ...; S_k^i$ designates the tree number $i$.

Due to Equation (150), this makes that regular elements are defined by their values on trees and ordinary, not divided probabilities.

The adjective regular can be better interpreted as "local in the sense of observation trees".

The vector space $\mathcal{C}_X^k(N)$ is generated by families of functions of divided probabilities $F_X(S_1; S_2; ...; S_k; (p, P, U))$, indexed by $X$ in $\mathbf{S}$ and forests $S_1; ...; S_k$ of level $k$. These families are supposed *local* with respect to $X$, which means that it is compatible with direct image of probabilities under observables in $\mathbf{S}^*$.

**Remark 12.** As we showed in the static case, in the classical context, locality is equivalent to the fact that the values of the functions depend on $\mathbb{P}$ through the direct images of $\mathbb{P}$ by the joint of all the ordered observables which decorate the tree (the joint of the joints along branches); but this is not necessarily true in the quantum context, where it depends on $\mathbf{Q}$. However it is true for $\mathbf{Q}^{min}$, in particular $\mathbf{Q}^{can}$ which is the most natural choice.

The spaces $\mathcal{C}^k(\mathcal{M}_X)$ form a natural degree one complex:

The faces $\delta_i^{(k)}; 1 \leq i \geq k$ are given by applying $\mu$ on $\mathcal{V} \circ \mathcal{V}$ at the places $(i, i+1)$; the last face $\delta_{k+1}^{(k)}; 1 \leq i \geq k$ consists in forgetting the last functor, the operation denoted by $\epsilon$; and the zero face is given by the action $\theta$. Then the boundary $\delta^{(k)}$ is the alternate sum of the operators $\delta_i^{(k)}; 0 \leq i \geq k+1$: if $F$ is measurable morphism from $\mathcal{M} \circ \mathcal{V}^{\circ k}$ to $\mathbb{R}$, then

$$\delta F = F \circ (\theta \mathcal{V}^{\circ k}) - \sum_{i=0,...,k-1} (-1)^i F \circ \mathcal{M}\mathcal{V}^{\circ i} \mu \mathcal{V}^{\circ k-i-1} - (-1)^k F \circ \mathcal{M}\mathcal{V}^{\circ k} \epsilon. \tag{153}$$

The zero face in the complex $\mathcal{C}_X^*$ corresponds to the right action of the monad $V_X$ on divided probabilities; on regular cochains it is expressed by a generalization of the formula (20): if $(P, i, m)$ is a simple vector of degree $m$ and $S_0; S_1; ...; S_k$ a forest of level $k+1$, with $m$ component trees, then

$$\begin{aligned} F_{S_0}(S_1; ...; S_k; (P, i, m)) &= F(S_1; ...; S_k; \theta((P, i, m)S_0)) \\ &= \sum_{j_i=1,...,n_i} \mathbb{P}(S_0^i = j_i) F((S_1^{j_i}; S_2^{j_i}; ...; S_2^{j_i}; (P|(S_0^i = j_i)), \end{aligned} \tag{154}$$

where $S_1^{j_i}; S_2^{j_i}; ...; S_k^{j_i}$ designates the tree number $j_i$ grafted on the branch $j_i$ of the variable $S_{0,i}$ at the place $i$ in the collection $S_0$.

The formula (154) is compatible with the existence of dead branches.

Note that natural integers come into the play under two different aspects: $m$ is for the internal monadic degree and counts the number of components, or the length of partitions, $k$ is for the height of the trees in the forest. The number $k$ gives the degree in co-homology.

The *coboundary* $\delta$ of $\mathcal{C}^*$ is of degree $+1$ with respect to $k$ and degree $0$ with respect to $m$. For any $m \in \mathbb{N}$, the operator $\delta$ has the formula of the coboundary given by the simplicial structure associated to $\theta$ and $\mu$:

$$\begin{aligned}\delta F(S_0;S_1;...;S_k;(p,P,U)) &= F_{S_0}(S_1;...;S_k;(p,P,U)) \\ &+\sum_{i=1}^{i=k}(-1)^i F(S_0;...;\mu(S_{i-1}\otimes S_i);S_{i+1};...;S_k;(p,P,U)) \\ &+(-1)^{k+1}F(S_0;...;S_{k-1};(p,P,U))\end{aligned} \tag{155}$$

We constat that locality is preserved by $\delta$.

**Lemma 11.** If the transformation $F$ is regular, then $\delta F$ is regular; in other terms, the regular elements form a sub-complex $\mathcal{C}_r^k(\mathcal{M}_X)$.

**Proof.** Let $(P,i,m)$ be a simple vector and $S_0;...;S_k$ a forest with $m$ components; let us denote by $S_0^j$ the variable number $j$ having degree $n_j$, and $n = n_1 + ... + n_m$; we have

$$\begin{aligned}&\delta F(S_0;...;S_k;(P,i,m)) \\ =\ & F(S_1;...;S_k;\theta((P,i,m)S_0^i)) - F(\mu(S_0,S_1);...;S_k;(P,i,m)) - ... \\ &+(-1)^k F(S_0;...;\mu(S_{k-1},S_k);(P,i,m)) + (-1)^{k+1}F(S_0;...;S_{k-1};(P,i,m)).\end{aligned} \tag{156}$$

The first term on the right is a combination of the image of $F$ for the $n_i$ simple vectors $P.S_0^{i,j_i}$ of degree $n = n_1 + ... + n_m$ which result from the division of $(P,i,m)$ by $S_0^i$. If $F$ is regular, this combination is the same as the combination of the simple vectors of degree $n_i$ constituting the division of $(P,i,m)$ by $S_0^i$, which gives the same result as the first term on the right in the formula

$$\begin{aligned}\delta F(S_0^i;...;S_k^i;(P,1,1)) &= F(S_1^i;...;S_k^i;\theta(P,S_0^i)) - F(\mu(S_0^i,S_1^i);...;S_k^i;P) - ... \\ &+(-1)^k F(S_0^i;...;\mu(S_{k-1}^i,S_k^i);P) + (-1)^{k+1}F(S_0^i;...;S_{k-1}^i;P).\end{aligned} \tag{157}$$

If $F$ is regular the term number $l > 1$ on the right of the equation (156) coincides with the corresponding term on the right of the Equation (157).

Therefore the terms on the left in Equation (156) coincides with the left term in (157); which establishes the lemma.

We define $\mathcal{C}_r^*(\mathcal{M}_X)$ as the sub-complex of regular vectors in $\mathcal{C}^*(\mathcal{M}_X)$. Its elements are named *tree information cochains* or *arborescent information cochains*.

By definition, the *tree information co-homology* is the homology of this regular complex, considered as a sheaf of complexes over the category $\mathbf{S}(A)$, *i.e.*, a contravariant functor. This corresponds to the topos information co-homology in the monadic context.

To recover the case of the ordinary algebra of partitions, and the formulas of the bar construction in the first sections of this article, we have to take the special case where all the decompositions of the same level coincide at every level of the forests. In this case, we can replace the quotient $\mathcal{M}_X$ by the modules of conditioning by a redefinition of the action on functions $\mathcal{F}_X$. However the notion of

divided probabilities for observation trees and the definition of co-homology in the monadic context can be seen as the natural basis of information co-homology.

When $k = 0$, in the classical case, a cochain is a function $f(\mathbb{P})$, the locality condition tells that it is a constant; and in this case it is a cocycle because the sum of probabilities equals one implies $f(\mathbb{P}) = f_S(\mathbb{P})$. Then $H^0_\tau$ has dimension one.

When $k = 0$, in the quantum case, the spectral functions of $\rho$ in the $\mathbf{Q}_X$ gives invariant information co-chains. Among them the Von Neumann entropy is specially relevant because its co-boundary gives the classical entropy. However, only the constant function is an invariant zero degree co-cycle. Thus again $H^0_U$ has dimension one.

For $k = 1$, a cochain is given by a function $F_X(S; P)$, such that, each time we have $X \to Y \to S$ and elements of $Y$ refines $S$, we have $F_X(S; P) = F_Y(S; Y_*P)$. It is a cocycle when for every collection $S_1, ..., S_m$ of $m$ observables, where $m$ is the length of $S$, we have

$$F(\mu_m(S, (S_1, ..., S_m)); P) = F(S; P) + \sum_i \mathbb{P}(S = i)F(S_i; P|S = i). \tag{158}$$

Note that the partition $\mu_m(S, (S_1, ..., S_m))$ is not the joint of $S$ and the $S_i$ for $i \geq 1$, except when all the $S_i$ coincide. Thus it is amazing that the ordinary entropy also satisfies this functional equation, finer than the Shannon's identity:

**Proposition 8.** The usual entropy $H(S_*\mathbb{P}) = H(S; \mathbb{P})$ is an arborescent co-cycle.

**Proof.** By linearity on the module of divided probabilities $\mathcal{M}_X$, we can decompose the probability $\mathbb{P}$ in the conditional probabilities $\mathbb{P}|(S = s)$, thus we can restrict the proof of the lemma to the case where $S = \pi_0$ is the trivial partition, *i.e.*, $m = 1$.

Let $X_i; i = 1, ..., m$ denote the elements of the partition associated to $S_0$ and $X^j_i; j = 1, ..., n_i$ the pieces of the intersection of $X_i$ with the elements of the partition associate to $S_i$; note $p_i$ the probability of the event $X_i$ and $p^j_i$ the probability of the event $X^j_i$; we have

$$H(\mu_m(S_0; (S_1, ..., S_m)); \mathbb{P}) = -\sum_{i=1}^{i=m} \sum_{j=1}^{j=n_i} p^j_i \log_2 p^j_i, \tag{159}$$

and

$$H_{S_0}(S_1; ...; S_m; \mathbb{P}) = -\sum_{i=1}^{i=m} p_i \sum_{j=1}^{j=n_i} \frac{p^j_i}{p_i} \log_2 \frac{p^j_i}{p_i} \tag{160}$$

$$= -\sum_{i=1}^{i=m} \sum_{j=1}^{j=n_i} p^j_i (\log_2 p^j_i - \log_2 p_i) \tag{161}$$

$$= -\sum_{i=1}^{i=m} \sum_{j=1}^{j=n_i} p^j_i \log_2 p^j_i + \sum_{i=1}^{i=m} \log_2 p_i \sum_{j=1}^{j=n_i} p^j_i \tag{162}$$

$$= -\sum_{i=1}^{i=m} \sum_{j=1}^{j=n_i} p^j_i \log_2 p^j_i + \sum_{i=1}^{i=m} p_i \log_2 p_i, \tag{163}$$

then

$$H(\mu_m(S_0;(S_1,...,S_m));\mathbb{P}) - H_{S_0}(S_1;...;S_m;\mathbb{P}) = H(S_0;\mathbb{P}). \tag{164}$$

Q.E.D.

This identity was discovered by Faddeev, Baez, Fritz, Leinster see [20]. However, we propose that information homology explains its significance.

When the category of quantum information **S**, the set $A$ and the probability functor **Q** are invariant under the unitary group, and if we choose a classical full subcategory $\mathcal{S}$, there is trace map from **Q** to $\mathcal{Q}$, induces a morphism from the classical arborescent co-homology of $\mathcal{S}$, $A$ and $\mathcal{Q}$ to the invariant quantum arborescent co-homology of **S**, $A$ and **Q**.
As a corollary of the Lemma 10 and the Theorems 1 and 3, we obtain the following result:

**Theorem 4.** $(i)$ both in the classical and the invariant quantum context, if $\mathbf{S}(A)$ is connected, sufficiently rich, and if **Q** is canonical, every 1-co-cycle is co-homologous to the entropy of Shannon; $(ii)$ in the classical case $H^1(\mathcal{S},A,\mathcal{Q})$ is the vector space of dimension 1 generated by the entropy; $(iii)$ in the quantum case $H^1_U(\mathbf{S},A,\mathbf{Q}) = 0$, and the only invariant 0-cochain which has for co-boundary the Shannon entropy is (minus) the Von-Neumann entropy.

*6.4. Arborescent Mutual Information*

For $k=2$, a cochain is given by a local function of a probability and a rooted decorated tree of level 2. It is a cocycle when the following functional equation is satisfied

$$\begin{aligned} &\sum_i \mathbb{P}(S=i)F(T_i;U_i;P|S=i) - F(S;T;P) \\ = \;& F(\mu_m(S\circ T);U;P) - F(S;(\mu_{n_i}(T_i\circ U_i);i\in[m]);P), \end{aligned} \tag{165}$$

where $S$ denotes a variable of length $m$, $T$ a collection of $m$ variables $T_1,...,T_m$ of respective lengths $n_1,...,n_m$ and $U$ a collection of variables $U^k_{i,j}$ of respective lengths $n_{i,j}$, with $i$ going from 1 to $m$, $j$ going from 1 to $n_i$ and $k$ going from 1 to $n_{i,j}$; the notation $U_i$ denoting the collection of variables $U^k_{i,j}$ of index $i$.

Our aim is to extend in the monadic context the topological action of the ordinary information structure on functions of probability used in the discussion of mutual information.

For that, we define another structure of $\mathcal{V}_X$-right module on the functor $\mathcal{M}_X$ associated to probabilities, by defining the following map $\theta_t(m)$ from $\mathcal{M}_X(m)$ tensorized with $V_X(n_1)\otimes...\otimes V_X(n_m)$ to $\mathcal{M}_X(n)$, for $n=n_1+...+n_m$:

$$\theta_t((P,i,m)\otimes S_1\otimes...\otimes S_m) = \sum_{j=1,...,n_i}(P,(i,j),n). \tag{166}$$

Remark that the generalized decompositions $S_j$ are used only through the orders on their elements.

As for $\mathcal{R}$, it is easy to verify that the collection of maps $\theta_t(m)$ defines a right action of the monad $\mathcal{V}_X$ on the Schur functor $\mathcal{M}_X$.

Then we consider as before, the graded vector space $C^*(\mathcal{M}_X)$ of homomorphisms of $\mathcal{V}$-modules from the functors $\mathcal{M} \circ \mathcal{V}^{\circ k}; k \geq 0$ to the functor $\mathcal{R}$ which are measurable in the probabilities $P$. As before, on $C^*(\mathcal{M}_X)$, we shift the degree by one, because of the independency with respect to the last stage of the forest, which follows from the trivial action on $\mathcal{R}$.

The topological coboundary operator $\delta_t$ is defined in every degree by the formula of the simplicial bar construction, as in Equation (153) for $\delta$, but with $\theta_t$ replacing $\theta$. It corresponds to the usual simplicial complex of the family $\mathcal{V}^{\circ k}$. A cochain is represented by a family of functions of probability laws $F_X(S_1; ...; S_k; (P, i, m))$, where $S_1; ...; S_k$ denotes a forest with $m$ trees of level $k$. The operator $\delta_t$ is given by

$$\begin{aligned}\delta_t F(S_0; ...; S_k; (P, i, m)) &= F(S_1; ...; S_k; \theta_t((P, i, m), S_0)) \\ &- F(\mu(S_0, S_1); ...; S_k; (P, i, m)) - ... + (-1)^k F(S_0; ...; \mu(S_{k-1}, S_k); (P, i, m)) \\ &+ (-1)^{k+1} F(S_0; ...; S_{k-1}; (P, i, m)). \quad (167)\end{aligned}$$

where $n = n_1 + ... + n_m$ is the sum of numbers of branches of the generalized decompositions $S_0^i$ for $i = 1, ..., m$.

As for $\delta$, a value $F(S_1; ...; S_k; (P, j, n)$ depends only on the tree $S_1^j; ...; S_k^j$ rooted at the place numbered by $j$ in the forest $S_1; ...; S_k$.

**Lemma 12.** The coboundary $\delta_t$ sends a regular cochain to a regular cochain.

**Proof.** Consider a simple vector $(P, i, m)$ in $\mathcal{M}_X(m)$ and a forest $S_0; ...; S_k$ with $m$ components; we denote by $S_0^j$ the variable number $j$ having degree $n_j$, and $n = n_1 + ... + n_m$, and we consider the formula (167).

If $F$ is regular the first term on the right is the sum of the images by $F$ for $P$ and the $n_i$ trees $S_1^{i,j_i}$ which result from the forgetting of the first branches $S_0^i$, and the other terms on the right are equal to the value of $F$ for $P$ and the tree rooted at $i$ in $S_0$. On the other side for the tree $S_0^i; ...; S_k^i$, if $F$ is regular, we have

$$\begin{aligned}\delta F(S_0^i; ...; S_k^i; (P, 1)) &= \sum_j F(S_1^{i,j}; ...; S_k^{i,j}; (P, 1)) - F(\mu(S_0^i, S_1^i); ...; S_k^i; (P, 1)) - ... \\ &+ (-1)^k F(S_0^i; ...; \mu(S_{k-1}^i, S_k^i); (P, 1)) + (-1)^{k+1} F(S_0^i; ...; S_{k-1}^i; (P, 1)). \quad (168)\end{aligned}$$

Thus $\delta F$ is topologically regular.

Consequently we can restrict $\delta_t$ to the subcomplex $\mathcal{C}_r^*(N_X)$, and name its homology the *arborescent*, or *tree*, *topological information co-homology*, written $H_{\tau,t}^*(\mathbf{S}^*, A, \mathbf{Q})$.

Now we suggest to extend the notion of mutual information $I(X; Y; \mathbb{P})$ in the way it will be a cocycle for this co-homology as it was the case for the Shannon mutual information in the ordinary topological information complex. We suggest to adopt the formulas using $\delta$ and $\delta_t$, as in the standard case:

**Definition 18.** Let $H(T;(P,i,m))$ denotes the regular extension to forests of the usual entropy; then the *mutual arborescent information* between a partition $S$ of length $m$ and a collection $T$ of $m$ partitions $T_1,...,T_m$ is defined by

$$I_\alpha(S;T;\mathbb{P}) = \delta_t H(S;T;\mathbb{P}). \tag{169}$$

The identity $\delta H = 0$ implies

$$I_\alpha(S;T;\mathbb{P}) = \sum_{i=1}^{i=m} H(T_i;\mathbb{P}) - \mathbb{P}(S=i)H(T_i;\mathbb{P}|S=i)). \tag{170}$$

In the particular case were all the $T_i$ are equal to a variable $T$, it gives

$$\begin{aligned} I_\alpha(S;T;\mathbb{P}) =& \sum_{i=1}^{i=m} \mathbb{P}(S=i)(H(T;\mathbb{P}) - H(T;\mathbb{P}|S=i)) + (m-1)H(T;\mathbb{P}) \\ &= H(T;P) - \sum_{i=1}^{i=m} \mathbb{P}(S=i)H(T;\mathbb{P}|S=i)) + (m-1)H(T;\mathbb{P}) \\ &= H(T;\mathbb{P}) - H_S(T;\mathbb{P}) + (m-1)H(T;\mathbb{P}), \end{aligned}$$

then

$$I_\alpha(S;T;\mathbb{P}) = I(S;T;\mathbb{P}) + (m-1)H(T;\mathbb{P}). \tag{171}$$

For $\mathcal{S}(A)$, the function $I_\alpha$ is an arborescent topological 2-cocycle.

It satisfies the Equation (165) were $\mathbb{P}$ replaces conditional probabilities $\mathbb{P}|(S=i)$ and where the factors $\mathbb{P}(S=i)$ disappear. Remark that, in this manner, maximization of $I_\alpha(S;T;\mathbb{P})$ comports maximization of usual mutual information $I(S;T;\mathbb{P})$ and unconditioned entropies $H(T_i;\mathbb{P})$.

Pursuing the homological interpretation of higher mutual information quantities given by the Formulas (55) and (56), we suggest the following definition:

**Definition 19.** The mutual arborescent informations of higher orders are given by $I_{\alpha,N} = -(\delta\delta_t)^M H$ for $N = 2M+1$ odd and by $I_{\alpha,N} = \delta_t(\delta\delta_t)^M H$ for $N = 2M+2$ even.

**Acknowledgments**

We thank MaxEnt14 for the opportunity to present these researches to the information science community. We thank Guillaume Marrelec for discussions and notably his participation to the research of the last part on optimal discrimination. We thank Frederic Barbaresco, Alain Chenciner, Alain Proute and Juan-Pablo Vigneaux for discussions and comments on the manupscript. We thank the "Institut des Systemes complexes" (ISC-PIF) region Ile-de-France, and Max Planck Institute For Mathematic in the Science for the financial support and hosting of P. Baudot.

## Author Contributions

Both authors contribute equally to the research, the second author wrote the manuscript. Both authors have read and approved the final manuscript.

## Conflicts of Interest

The authors declare no conflict of interest.

## References


1. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.
2. Kolmogorov, A. Combinatorial foundations of information theory and the calculus of probabilities. *Russ. Math. Surv.* **1983**, *38*, doi:10.1070/RM1983v038n04ABEH004203.
3. Thom, R. *Stabilité struturelle et morphogénèse*, deuxième ed.; Dunod: Paris, France, 1977. (in French)
4. Mac Lane, S. *Categories for the Working Mathematician*; Springer: Berlin/Heidelberg, Germany, 1998.
5. Mac Lane, S. *Homology*; Springer: Berlin/Heidelberg, Germany, 1975.
6. Hu, K.T. On the Amount of Information. *Theory Probab. Appl.* **1962**, *7*, 439–447.
7. Baudot, P.; Bennequin, D. *Information Topology I*; in preparation.
8. Elbaz-Vincent, P.; Gangl, H. On poly(ana)logs I. *Compos. Math.* **2002**, *130*, 161–214.
9. Cathelineau, J. Sur l'homologie de sl2 a coefficients dans l'action adjointe. *Math. Scand.* **1988**, *63*, 51–86.
10. Loday, J.L.; Valette, B. *Algebraic Operads*; Springer: Berlin/Heidelberg, Germany, 2012.
11. Matsuda, H. Information theoretic characterization of frustrated systems. *Physica A* **2001**, *294*, 180–190.
12. Brenner, N.; Strong, S.; Koberle, R.; Bialek, W. Synergy in a Neural Code. *Neural Comput.* **2000**, *12*, 1531–1552.
13. Nielsen, M.; Chuang, I. *Quantum Computation and Quantum Information*; Cambridge University Press: Cambridge, UK, 2000.
14. Baudot, P.; Bennequin, D. Topological forms of information. *AIP Conf. Proc.* **2015**, *1641*, 213–221.
15. Baudot, P.; Bennequin, D. *Information Topology II*; in preparation.
16. Fresse, B. Koszul duality of operads and homology of partitionn posets. *Contemp. Math. Am. Math. Soc.* **2004**, *346*, 115–215.
17. May, J.P. *The Geometry of Iterated Loop Spaces*; Springer: Berlin/Heidelberg, Germany, 1972.
18. May, J.P. *Einfinite Ring Spaces and Einfinite Ring Spectra*; Springer: Berlin/Heidelberg, Germany, 1977.

19. Beck, J. Triples, Algebras and Cohomology. Ph.D. Thesis, Columbia University, New York, NY, USA, 1967.
20. Baez, J.; Fritz, T.; Leinster, T. A Characterization of Entropy in Terms of Information Loss. *Entropy* **2011**, *13*, 1945–1957.
21. Marcolli, M.; Thorngren, R. Thermodynamic Semirings. **2011**, arXiv:10.4171/JNCG/159.
22. Baudot, P.; Bennequin, D. *Information Topology III*; in preparation.
23. Gromov, M. In a Search for a Structure, Part 1: On Entropy. 2013. Available online: http://www.ihes.fr/ gromov/PDF/structre-serch-entropy-july5-2012.pdf (accessed on 6 May 2015).
24. Watkinson, J.; Liang, K.; Wang, X.; Zheng, T.; Anastassiou, D. Inference of Regulatory Gene Interactions from Expression Data Using Three-Way Mutual Information. *Chall. Syst. Biol. Ann. N.Y. Acad. Sci.* **2009**, *1158*, 302–313.
25. Kim, H.; Watkinson, J.; Varadan, V.; Anastassiou, D. Multi-cancer computational analysis reveals invasion-associated variant of desmoplastic reaction involving INHBA, THBS2 and COL11A1. *BMC Med. Genomics* **2010**, *3*, doi:10.1186/1755-8794-3-51.
26. Uda, S.; Saito, T.H.; Kudo, T.; Kokaji, T.; Tsuchiya, T.; Kubota, H.; Komori, Y.; ichi Ozaki, Y.; Kuroda, S. Robustness and Compensation of Information Transmission of Signaling Pathways. *Science* **2013**, *341*, 558–561.
27. Han, T.S. Linear dependence structure of the entropy space. *Inf. Control* **1975**, *29*, 337–368.
28. McGill, W. Psychometrika. *Multivar. Inf. Transm.* **1954**, *19*, 97–116.
29. Kolmogorov, A.N. *Grundbegriffe der Wahrscheinlichkeitsrechnung*; Springer, Berlin/Heidelberg, Germany, 1933. (in German)
30. Artin, M.; Grothendieck, A.; Verdier, J. *Théorie des topos et cohomologie étale des schémas—(SGA 4) Tome I,II,III*; Springer: Berlin/Heidelberg, Germany. (in French)
31. Grothendieck, A. Sur quelques points d'algèbre homologique, I. *Tohoku Math. J.* **1957**, *9*, 119–221.
32. Gabriel, P. Objets injectifs dans les catégories ab liennes. *Séminaire Dubreil. Algèbre et théorie des nombres* **1958–1959**, *12*, 1–32.
33. Bourbaki, N. *Algèbre, chapitre 10, Algèbre homologique*; Masson: Paris, France, 1980. (in French)
34. Cartan, H.; Eilenberg, S. *Homological Algebra*; The Princeton University Press: Princeton, NJ, USA, 1956.
35. Tverberg, H. A new derivation of information function. *Math. Scand.* **1958**, *6*, 297–298.
36. Kendall, D. Functional Equations in Information Theory. *Z. Wahrscheinlichkeitstheorie* **1964**, *2*, 225–229.
37. Lee, P. On the Axioms of Information Theory. *Ann. Math. Stat.* **1964**, *35*, 415–418.
38. Kontsevitch, M. The 1+1/2 logarithm. *Unpublished note. Reproduced in Elbaz-Vincent & Gangl, 2002 On poly(ana)logs I. Compositio Mathematica* **1995** e-print math.KT/0008089.

39. Khinchin, A. *Mathematical Foundations of Information Theory*; Dover: New York, NY, USA; Translated by Silverman, R.A., Friedman, M.D., Eds.; From two Russian articles in Uspekhi Matematicheskikh Nauk, 1957; pp. 17–75.
40. Yeung, R. *Information Theory and Network Coding*; Springer: Berlin/Heidelberg, Germany, 2007.
41. Cover, T.M.; Thomas, J. *Elements of Information Theory*; Wiley: Weinheim, Germany, 1991.
42. Rindler, W.; Penrose, R. *Spinors and Spacetime*, 2nd ed.; Cambridge University Press: Cambridge, UK, 1986.
43. Landau, L.D.; Lifshitz, E.M. *Fluid Mechanics*, 2nd ed.; Volume 6 of a Course of Theoretical Physics. Pergamon Press, 1959.
44. Balian, R. Emergences in Quantum Measurement Processes. *KronoScope* **2013**, *13*, 85–95.
45. Borel, A.; Ji, L. Compactifications of Symmetric and Locally Symmetric Spaces. In *Unitary Representations and Compactifications of Symmetric Spaces*; Springer: Berlin/Heidelberg, Germany, 2006.
46. Doering, A.; Isham, C. Classical and quantum probabilities as truth values. *J. Math. Phys.* **2012**, *53*, doi:10.1063/1.3688627.
47. Meyer, P. *Quantum Probability for Probabilists*; Springer: Berlin, Germany, 1993.
48. Souriau, J. *Structure des Systemes Dynamiques*; Jacques Gabay: Paris, France , 1970. (in French)
49. Catren, G. Towards a Group-Theoretical Interpretation of Mechanics. *Philos. Sci. Arch.* **2013**, http://philsci-archive.pitt.edu/10116/.
50. Bachet Claude-Gaspar *Problèmes plaisans et délectables, qui se font par les nombres*; A. Blanchard: Paris, France, 1993; p. 1612. (in French)
51. Jaynes, E.T. Information Theory and Statistical Mechanics. In *Statistical Physics*; Ford, K., Ed.; Benjamin: New York, NY, USA, 1963; p. 181.
52. Jaynes, E.T. Prior Probabilities. *IEEE Trans. Syst. Sci. Cybern.* **1968**, *4*, 227–241.
53. Cohen, F.; Lada, T.; May, J. *The Homology of Iterated Loop Spaces*; Springer: Berlin, Germany, 1976.
54. Prouté, A. Introduction la Logique Catégorique. 2013. Available online: www.logique.jussieu.fr/~alp/ (accessed on 6 May 2015).
55. Getzler, E.; Jones, J.D.S. Operads, homotopy algebra and iterated integrals for double loop spaces. **1994**, arXiv:hep-th/9403055v1.
56. Ginzburg, V.; Kapranov, M.M. Koszul duality for operads. *Duke Math. J.* **1994**, *76*, 203–272.

# Computing Bi-Invariant Pseudo-Metrics on Lie Groups for Consistent Statistics

**Nina Miolane and Xavier Pennec**

**Abstract:** In computational anatomy, organ's shapes are often modeled as deformations of a reference shape, *i.e.*, as elements of a Lie group. To analyze the variability of the human anatomy in this framework, we need to perform statistics on Lie groups. A Lie group is a manifold with a consistent group structure. Statistics on Riemannian manifolds have been well studied, but to use the statistical Riemannian framework on Lie groups, one needs to define a Riemannian metric compatible with the group structure: a bi-invariant metric. However, it is known that Lie groups, which are not a direct product of compact and abelian groups, have no bi-invariant metric. However, what about bi-invariant pseudo-metrics? In other words: could we remove the assumption of the positivity of the metric and obtain consistent statistics on Lie groups through the pseudo-Riemannian framework? Our contribution is two-fold. First, we present an algorithm that constructs bi-invariant pseudo-metrics on a given Lie group, in the case of existence. Then, by running the algorithm on commonly-used Lie groups, we show that most of them do not admit any bi-invariant (pseudo-) metric. We thus conclude that the (pseudo-) Riemannian setting is too limited for the definition of consistent statistics on general Lie groups.

Reprinted from *Entropy*. Cite as: Miolane, N.; Pennec, X. Computing Bi-Invariant Pseudo-Metrics on Lie Groups for Consistent Statistics. *Entropy* **2015**, *17*, 1850–1881.

## 1. Introduction

### *1.1. Modeling with Lie Groups*

Data can be modeled as elements of Lie groups in many different fields: computational anatomy, robotics, paleontology, *etc*. Indeed, Lie groups are continuous groups of transformations and, thus, appear naturally whenever one deals with articulated objects or shapes.

Regarding articulated objects, one can take examples in robotics or in computational anatomy. In robotics, first, a spherical arm is obviously an articulated object. The positions of the arm can be modeled as the elements of the three-dimensional Lie group of rotations $SO(3)$. In computational anatomy, then, the spine can be modeled as an articulated object. In this context, each vertebra is considered as an orthonormal frame that encodes the rigid body transformation from the previous vertebra. Thus, as the human spine has $24$ vertebrae, a configuration of the spine can be modeled as an element of the Lie group $SE(3)^{23}$, where $SE(3)$ is the Lie group of rigid body transformations in 3D, *i.e.*, the Lie group of rotations and translations in $\mathbb{R}^3$, also called the special Euclidean group.

Regarding shapes, the general model of d'Arcy Thompson suggests representing shape data as the diffeomorphic deformations of a reference shape [1], thus as elements of an infinite dimensional Lie group of diffeomorphisms. This framework can be applied as well in paleontology compared to

in computational medicine. In palaeontology, first, a monkey skull or a human skull can be modeled as the diffeomorphic deformation of a reference skull. In computational medicine, then, the shape of a patient's heart can be modeled as the diffeomorphic deformation of a reference shape. Obviously, many more examples could be given, also in other fields.

### *1.2. Statistics on Lie Groups*

Once data are represented as elements of a Lie group, we may want to perform statistical analysis on them for prediction or quantitative modeling. Thus, we want to perform statistics on Lie groups. How can we define an intrinsic statistical framework that is efficient on all Lie groups? How do we compute the mean or the principal modes of variation for a sample of Lie group elements? In order to train our intuition, we consider finite dimensional Lie groups here.

To define a statistical framework, it seems natural to start with the definition of a mean. The definition of mean on a Lie group exemplifies the issues one can encounter while defining the whole statistical framework. We know that the usual definition of the mean is the weighted sum of the data elements of the sample. However, this definition is linear, and Lie groups are not linear in general. Consequently, we cannot use this definition on Lie groups: we could get a mean of Lie group elements that is not a Lie group element. One can consider as an example the half sum of two rotation matrices that is not always a rotation matrix.

In fact, the definition of the mean on a Lie group should be consistent with the group structure. This consistency leads to several requirements of the mean, or properties. First, the mean of Lie group elements should be in the Lie group. Then, it seems natural to require that a left or right translation of the dataset should translate its mean accordingly. Figure 1 illustrates the case when this condition is fulfilled. Finally, the inversion of all data elements should lead to an inverted mean. A mean verifying all of these properties is said to be bi-invariant.

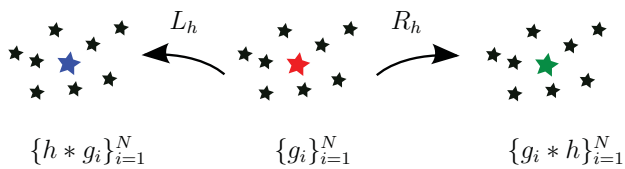



**Figure 1.** Left and right translation of a dataset $\{g_i\}_{i=1}^N$ on the Lie group $G$. The initial dataset $\{g_i\}_{i=1}^N$ has a mean represented in red. The left translated dataset $\{h * g_i\}_{i=1}^N$ has a mean represented in blue. The right translated dataset $\{g_i * h\}_{i=1}^N$ has a mean represented in green. We require that the mean of the (right or left) translated dataset is the translation of the red mean, which is the case in this illustration: the blue mean is the left translation of the red mean, and the green mean is the right translation of the red mean.

A naturally bi-invariant candidate for the mean on Lie groups is the group exponential barycenter [2] defined as follows. A group exponential barycenter $m$ of the dataset $\{g_i\}_{i=1,..,N}$ is a solution, if there are some, of the following group barycenter equation:

$$\sum_{i=1}^{N} \mathrm{Log}(m^{(-1)} * g_i) = 0 \tag{1}$$

where Log is the group logarithm. As the group exponential barycenter is naturally bi-invariant, we call a group exponential barycenter a bi-invariant mean. The local existence and uniqueness of the bi-invariant mean have been proven if the dispersion of the data is small enough. "Local" means that the data are assumed to be in a sufficiently small normal convex neighborhood of some point of the Lie group.

Now, we want to provide a computational framework for the bi-invariant mean that would set the foundations for computations on Lie groups statistics in general. For that, we are interested in characterizing the global domains of existence and uniqueness of the bi-invariant mean. By "global domain", we mean, for example, a ball of maximal radius, such that any probability measure with support included in it would have a unique bi-invariant mean. Note that there is *a priori* no problem having several means, which can be called several "modes", or no mean at all. Our aim is rather to characterize the different situations that may occur: no mean, one unique mean, several means.

### *1.3. Using Riemannian and Pseudo-Riemannian Structures for Statistics on Lie Groups*

To this aim, we are interested in additional geometric structures on Lie groups that could help, by providing computational tools. For example, we are interested in a distance on a Lie group, that could enable one to measure the radii of balls. Such a distance could obviously help with characterizing balls of maximal radius.

However, a Lie group is a group that carries an additional manifold structure, and one can define a pseudo-metric on a manifold, making it a pseudo-Riemannian manifold. Thus, we can add a pseudo-metric on Lie groups, which then induces a pseudo-distance. Could this additional pseudo-Riemannian structure help to define the statistical framework on Lie groups in practice?

We consider first the case of the Riemannian structure, *i.e.*, when the pseudo-metric is in fact a metric (positive definite). Several definitions of the mean on Riemannian manifolds have been proposed in the literature: the Fréchet mean, the Karcher mean or the Riemannian exponential barycenter [3–8]. For example, the Riemannian exponential barycenters are defined as the critical points of the variance of the data, defined as: $\sigma^2(y) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{dist}(x_i, y)^2$, where $\{x_i\}_{i=1}^{N}$ are the data and dist the distance induced by the Riemannian metric. The Riemannian framework provides theorems for the global existence and uniqueness domains of this mean [7–11], ensuring the computability of statistics on Riemannian manifolds. These represent exactly the kind of results that we would like to have for the bi-invariant mean on Lie groups. Thus, one may wonder if we can apply this computational framework for statistics on Lie groups and, more particularly, for the bi-invariant mean, by adding a Riemannian metric on the Lie group.

In fact, the notions of Riemannian mean and group exponential barycenter (or bi-invariant mean) coincide when the Riemannian metric is itself bi-invariant. In this case, the Riemannian geodesics coincide with the geodesics of the Cartan–Schouten connection [12]. Thus, we can use the computational framework for Riemannian means only if we can add a bi-invariant metric on a Lie group.

However, it is known that a Lie group does not have any bi-invariant Riemannian metric in general. The Lie group $ST(n)$ of scalings and translations of $\mathbb{R}^n$, the Heisenberg group $H$, the Lie group $UT(n)$ of upper triangular matrices of size $n \times n$ and the Lie group $SE(n)$ of rotations and translations of $\mathbb{R}^n$ do not have any bi-invariant metric, while they admit a locally unique bi-invariant mean [2]. Therefore, if we want to characterize the bi-invariant mean with an additional geometric structure on Lie groups, we have to consider a structure that is more general than the Riemannian one.

The pseudo-Riemannian framework is a generalization of the Riemannian framework. Thus, it represents a tempting alternative for the characterization of the bi-invariant mean and for the definition of computational statistics on Lie groups in general. The pseudo-metric is not required to be positive definite anymore, only definite: the class of Lie groups that admit a bi-invariant pseudo-metric is larger than the class of those with a bi-invariant metric. Therefore, we could try to generalize the Riemannian statistical framework to a pseudo-Riemannian statistical framework and apply it for Lie groups. For instance, the mean on a pseudo-Riemannian manifold could still be defined as a critical point of the variance $\sigma^2(y) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{dist}(x_i, y)^2$, but dist would now be the pseudo-distance induced by the pseudo-metric. Of course, existence and uniqueness theorems would have to be re-established, but we could get intuition from the Riemannian case.

In order to use the pseudo-Riemannian framework to characterize the bi-invariant mean, the first issue is: how many Lie groups do admit a bi-invariant pseudo-metric? Is it the case for the real Lie groups $ST(n)$, $H$, $UT(n)$ and $SE(n)$, which have a locally unique bi-invariant mean?

### *1.4. Lie Groups and Lie Algebras with Bi-Invariant Pseudo-Metrics*

If $\mathcal{G}$ is a connected Lie group, it admits a bi-invariant non-degenerate symmetric bilinear form if and only if its Lie algebra admits a nondegenerate symmetric bilinear inner product, also called a bi-invariant pseudo-metric. Lie algebras with bi-invariant pseudo-metric were known to exist since the 1910s with the classification of simple Lie algebra [13] and the well-known Cartan–Killing form, which is not degenerate in this case, but their specific study began in the 1950s with the works of [14,15]. Later, [16] started to study the properties of these Lie algebras from their structural point of view and introduced the decomposability or indecomposability of these Lie algebras as a direct sum of ideals. However, the decomposition of [16] was not enough to characterize all Lie algebras with bi-invariant pseudo-metrics, as some authors [17–19] remark that the so-called oscillator algebra arising in quantum mechanics carried a bi-invariant pseudo-metric without being decomposable in the sense of [16]. This leads, Medina and Revoy [20,21] and Keith [17] to build independently a

classification of these Lie algebras, by showing that they all arise through direct sums and a structure, called the double extension in [20,21] and the bi-extension in [17].

These results have been complemented by [22] and then generalized by Bordemann to any non-associative algebras with the bi-invariant form through the $T^*$-extension structure [23]. They have been completely described for certain dimensions in specific cases. The classification of the nilpotent quadratic Lie algebras of dimensions $\leq 7$ is obtained in [24], of the real solvable quadratic Lie algebras of dimensions $\leq 6$ in [25] and the irreducible non-solvable Lie algebras of dimensions $\leq 13$ in [26]. The specific cases of indecomposable quadratic Lie algebras with pseudo-metrics of different indices have been studied: bi-invariant pseudo-metrics of index one are described in [21,27], of index two in [28] and finally of the general index in [29]. The dimension of the space of bi-invariant pseudo-metrics has been studied in [30] where bounds are provided.

Authors from other fields than pure algebra have also contributed to the study of bi-invariant pseudo-metrics. For example in functional analysis, Manin triples are a special type of Lie algebra with the bi-invariant pseudo-metric that allow one to interpret the solutions of the classical Yang–Baxter equation [31]. In this context, the Manin triples have been themselves classified for semi-simple Lie algebras in [32] and for complex reductive Lie algebra in [33].

Simultaneously, people started to gain interest in computational aspects on finite dimensional Lie algebras, implementing the identification of a Lie algebra from its structure constants given in any basis [34,35] or the Levi decomposition [36,37]. The state-of-the-art regarding implementations on finite dimensional Lie algebra is summarized in [38]. However, computations deal with the algebraic aspects of Lie algebras and, to the knowledge of the authors, do not consider metrics or pseudo-metrics.

### *1.5. Contributions and Outline*

Our contribution is an algorithmic reformulation of a classification theorem for Lie algebras [20,21] that answers these questions. More precisely, taking a Lie group $\mathcal{G}$ as input, the algorithm constructs a bi-invariant pseudo-metric on $\mathcal{G}$ in the case of existence. Using this algorithm, we show that most Lie groups that have a locally unique bi-invariant mean do not possess a bi-invariant pseudo-metric. We conclude that, for the purpose of statistics on general real Lie groups and, more precisely, for the computational framework of the bi-invariant mean, generalizing the Riemannian statistical framework to a pseudo-Riemannian framework may not be the optimal program.

The paper is organized as follows. In the first section, we introduce notions on quadratic Lie groups that will be useful for the understanding of the paper. In the second section, we present the (tree-structured) algorithm that constructs bi-invariant pseudo-metrics on a given Lie group, in the case of existence. In the third section, we apply the algorithm on $ST(n)$, $H$, $UT(n)$ and $SE(n)$ and show that most of them do not have any bi-invariant pseudo-metric.

## 2. Introduction to Lie Groups with Bi-Invariant Pseudo-Metrics

Here, we define the algebraic and geometric notions that will be used throughout the paper.

*2.1. Quadratic Lie Groups and Lie Algebras*

In the following, we consider finite dimensionalsimply connected Lie groups over the field $\mathbb{F}$, where $\mathbb{F}$ is $\mathbb{R}$ or $\mathbb{C}$.

2.1.1. Lie Groups

A Lie Group $\mathcal{G}$ is a smooth manifold with a compatible group structure. It is provided with an identity element $e$, a smooth composition law $* : (g, h) \mapsto g * h \in \mathcal{G}$ and a smooth inversion law $Inv : f \mapsto f^{(-1)} \in \mathcal{G}$. Its tangent space at $g$ is written $T_g\mathcal{G}$.

The map $L_h : \mathcal{G} \ni g \mapsto h * g \in \mathcal{G}$ is the left translation by hand is a diffeomorphism of $\mathcal{G}$. Therefore, its differential (at $g$), $DL_h(g) : T_g\mathcal{G} \mapsto T_{L_h g}\mathcal{G}$ is an isomorphism that connects tangent spaces of $\mathcal{G}$. Similarly, one can define $R_h : \mathcal{G} \ni g \mapsto g * h \in \mathcal{G}$, the right translation by $h$.

A vector field $X$ on $\mathcal{G}$ is left invariant if $(dL_h)(X(g)) = X(L_h(g)) = X(h*g)$ for each $g, h \in \mathcal{G}$. Similarly, one could define right invariant vector fields. The left invariant vector fields form a vector space that we denote $\Gamma(T\mathcal{G})^L$ and that is isomorphic to $T_e\mathcal{G}$. The Lie bracket of two left invariant vector fields is a left-invariant vector field [39].

2.1.2. Lie Algebras

As $\Gamma(T\mathcal{G})^L$ is closed under the Lie bracket of vector fields, we can look at $T_e\mathcal{G}$ as a Lie algebra. More precisely, we define $\mathfrak{g}$ the Lie algebra of $\mathcal{G}$ as $T_e\mathcal{G}$ with the Lie bracket induced by its identification with $\Gamma(T\mathcal{G})^L$. The Lie algebra essentially captures the local structure of the group. In the case of Lie algebras of matrices, the Lie bracket corresponds to the commutator. For a more complete presentation of Lie groups and Lie algebras, we refer the reader to [40].

Writing the expression of the Lie bracket $[,]_\mathfrak{g}$ on a given basis $\mathcal{B}_\mathfrak{g} = \{e_i\}_{i=1}^n$ of $\mathfrak{g}$, we define the structure constants $f_{ijk}$ as:

$$[e_i, e_j]_\mathfrak{g} = f_{ijk} e_k \tag{2}$$

The structure constants $f_{ijk}$ depend on the basis $\mathcal{B}_\mathfrak{g}$ chosen. They are always skew-symmetric in the first two indices, but they may have additional symmetry properties if we write them in a well-chosen basis (see below). The structure constants $f_{ijk}$ completely determine the algebraic structure of the Lie algebra. Therefore, the structure constants are often the starting point, or the input, of algorithms on Lie algebras [34–36,38]. It will also be the case for the algorithm we present in this paper.

2.1.3. Pseudo-Metrics

A pseudo-metric $<,>$ on $\mathcal{G}$ is defined as a smooth collection of definite inner products $<,>|_g$ on each tangent space $T_g\mathcal{G}$. Then, $\mathcal{G}$ becomes a pseudo-Riemannian manifold. A metric is defined as a pseudo-metric whose inner products are all positive definite. In this case, $\mathcal{G}$ is called a Riemannian manifold.

The signature $(p, q)$ of a pseudo-metric is the number (counted with multiplicity) of positive and negative eigenvalues of the real symmetric matrix representing the inner product $<,>|_g$ at a point $g$ and with respect to a basis of $T_g\mathcal{G}$. The signature is independent of the choice of the point $g$ and on the basis at $T_g\mathcal{G}$. By definition, a pseudo-metric is definite; thus, there are no null eigenvalues, and we have $p + q = n$, where $n$ is the dimension of $\mathcal{G}$. By definition, a metric is positive definite, and thus, its signature is $(n, 0)$. Again, further details about such differential geometry can be found in [39].

2.1.4. Quadratic Lie Groups and Algebras

A left-invariant pseudo-metric is a pseudo-metric $<,>$, such that for all $X, Y \in T_g\mathcal{G}$ and for all $g, h \in \mathcal{G}$, we have:

$$< DL_h(g)X, DL_h(g)Y > |_{L_hg} =< X, Y > |_g \tag{3}$$

where $L_h$ is the left translation by $h$. In other words, the left translations are isometries for this pseudo-metric. Similarly, we can define right-invariant and bi-invariant pseudo-metrics $<,>$. Note that any Lie group admits a left (or right) invariant pseudo-metric: we can define an inner product on the Lie algebra $\mathfrak{g} = T_e\mathcal{G}$ and propagate it on each tangent space $T_g\mathcal{G}$ through $DL_g(e)$ (or $DR_g(e)$). However, no Lie group admits a bi-invariant pseudo-metric.

The Lie groups that admit a bi-invariant pseudo-metric are called quadratic Lie groups. The corresponding Lie algebras are called quadratic Lie algebras. Note that quadratic Lie groups or algebras are called differently in the literature. We find the appellation metrizable or metrized in [14–16], *metric* in [28,29], quasi-classical in [25] and, finally, quadratic in [24,26].

Figure 2 shows a summary of the structures that we just introduced.

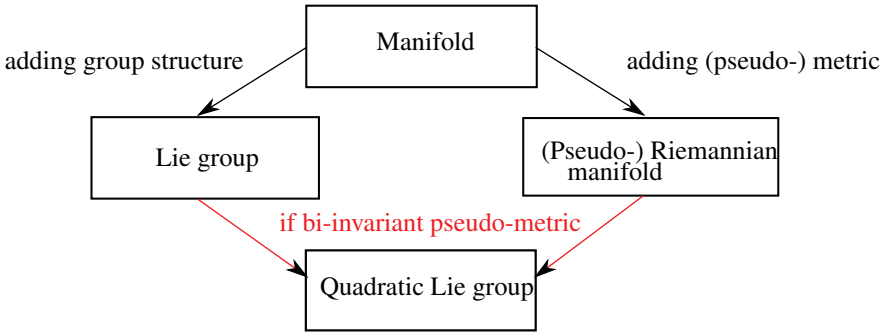



**Figure 2.** Algebraic and geometric structures. If we require compatible algebraic and geometric structures on the manifold, we get a quadratic Lie group: a Lie group with a bi-invariant pseudo-metric.

We now recall that a non-degenerate bi-invariant inner product on a finite dimensional Lie algebra $\mathfrak{g}$ gives rise to a bi-invariant pseudo-metric on every Lie group whose Lie algebra is $\mathfrak{g}$ (see, for example, [41]). Therefore, we focus on Lie algebras from now on. We will still use the terms

"pseudo-metric" or "metric" and the notation "$<,>$" in order to refer to the corresponding inner products on the Lie algebra $\mathfrak{g} = T_e\mathcal{G}$.

### 2.1.5. Characterization of Quadratic Lie Algebras

We give here different formulations of an equation characterizing a pair $(\mathfrak{g}, <,>)$ as a quadratic Lie algebra. A Lie algebra $\mathfrak{g}$ is quadratic if and only if it has a pseudo-metric $<,>$ verifying:

$$\forall x, y, t \in \mathfrak{g}, \qquad < [x, y]_{\mathfrak{g}}, t > + < y, [x, t]_{\mathfrak{g}} >= 0 \tag{4}$$

A proof for this characterization is given in [39] and [12].

First, taking advantage of the linearity in $x, y, t$, we can rewrite Equation (4) on basis vectors. Let $\mathcal{B}_{\mathfrak{g}} = \{e_i\}_{i=1}^{n}$ be a basis of $\mathfrak{g}$; we consider: $x = e_i$, $y = e_j$ and $z = e_k$. Thus, we can express the Lie bracket in terms of the structure constants, and we get:

$$\forall i, j, k \in \{1, .., n\} \qquad f_{ijl} < e_l, e_k > + f_{jkl} < e_l, e_j >= 0 \tag{5}$$

In particular, we observe that the structure constants written in a basis orthonormal with respect to a bi-invariant metric are totally skew-symmetric. The structure constants written in a basis orthonormal with respect to a bi-invariant pseudo-metric will have additional symmetric properties, as well.

Then, as we consider finite dimensional Lie groups, we can also rewrite Equation (4) in terms of matrices:

$$\forall x \in \mathfrak{g}, \quad A(x)^T.Z + Z.A(x) = 0 \tag{6}$$

where $A(x)$ is the matrix of the endomorphism denoted $[x, \bullet]$, defined as $y \mapsto [x, y]$, and $Z$ a symmetric invertible (not necessarily positive) matrix representing $<,>$ on $\mathcal{B}_{\mathfrak{g}}$, the basis of $\mathfrak{g}$. Note that: $x \mapsto A(x)$ is itself linear.

Finally, taking advantage of the linearity again and writing: $A(e_i) = A_i$, we can again reformulate Equation (4), and we get:

$$\forall i \in \{1, .., n\}, \quad A_i^T.Z + Z.A_i = 0 \tag{7}$$

which is now a linear system of $n$ matrix equations. Note that Equation (5) corresponds to Equations (7) written in coordinates.

### 2.1.6. How to Compute Bi-Invariant Pseudo-Metrics?

Given a Lie algebra $\mathfrak{g}$ as input, we see now that the computation of bi-invariant pseudo-metrics on $\mathfrak{g}$ amounts to the resolution of the linear system of Equations (7) for $Z$. The solutions of the linear system Equations (7) form a vector space, which is called the quadratic space $\mathcal{Q}(\mathfrak{g})$ [30]:

$$Q(\mathfrak{g}) = \{Z \in \text{Sym}(n) \quad | \quad \forall i \in \{1, .., n\}, \quad A_i^T.Z + Z.A_i = 0\} \tag{8}$$

Obviously, the vector space $\mathcal{Q}(\mathfrak{g})$ contains invertible and non-invertible solutions. Recalling the definition of a pseudo-metric, we emphasize that we will be interested in invertible solutions only.

In order to solve the system of Equations (7) for $Z$, *i.e.*, to compute the quadratic space $Q(\mathfrak{g})$, we could adopt an analytic point of view. At $i$ fixed, a single equation of the system of Equations (7) is a particular case of a Lyapunov equation that is studied in the context of control theory [42]. Thus, computational methods exist for studying one of our linear matrix equations [43]. For our purpose, however, we want to understand the structure of a quadratic Lie group, in order to get an intuition for the generalization to infinite dimensional Lie groups of diffeomorphisms. Thus, we do not rely on an analytic point of view to solve the system of Equations (7).

We rather consider the whole system of Equations (7) from an algebraic point of view. The pure algebraic point of view enables one to solve the system of Equations (7) completely in most cases, like in the examples provided at the end of the paper. In the other cases, it leads to a smaller system of equations that can be solved analytically or computationally. Thus, the algebraic point of view provides not only a theoretical understanding of quadratic Lie groups, it also either solves the problem or reduces the problem in order for the analytic point of view to solve it.

Therefore, we present in the next subsection the algebraic and geometric notions needed to set up, and later implement, the algebraic point of view.

### *2.2. Lie Algebra Representations*

How can we understand the structure of a Lie algebra? An idea is to represent the Lie algebra elements as matrices acting on vectors. Then, the study of the behavior of these matrices helps to understand the Lie algebra as a whole. This is the purpose of the theory of Lie algebra representations, which we present briefly relying on [13,21,38,40] in all of this subsection.

#### 2.2.1. Lie Algebras Representations

A $\mathfrak{g}$-representation on the vector space $V$ is a Lie algebra homomorphism $\eta : \mathfrak{g} \mapsto \mathfrak{gl}(V)$, which represents the elements of $\mathfrak{g}$ as matrices acting on the vector space $V$. The $\mathfrak{g}$-representations $\theta_1$ and $\theta_2$ are said to be isomorphic if there is an isomorphism of representations between them, *i.e.*, an isomorphism of vector spaces $l : V_1 \mapsto V_2$ that verifies: $\theta_2(x) \circ l = l \circ \theta_1(x)$. We denote $\mathrm{Hom}_{\mathfrak{g}}(V_1, V_2)$ the vector space of isomorphisms of representations between $V_1$ and $V_2$.

In order to understand the representations of a Lie algebra $\mathfrak{g}$ and, thus, the Lie algebra $\mathfrak{g}$ itself, a strategy is to decompose the representations into smaller bricks, and then study those bricks. In this context, a $\mathfrak{g}$-subrepresentation of the $\mathfrak{g}$-representation $V$ is a subspace of $V$ stable by the elements of $\eta(\mathfrak{g})$. An irreducible $\mathfrak{g}$-subrepresentation is a $\mathfrak{g}$-subrepresentation without proper $\mathfrak{g}$-subrepresentation. An indecomposable $\mathfrak{g}$-subrepresentation is a $\mathfrak{g}$-subrepresentation that cannot be decomposed into $\mathfrak{g}$-subrepresentations.

Note that irreducibility implies indecomposability, but the converse is false: a $\mathfrak{g}$-representation can have a $\mathfrak{g}$-subrepresentation that does not have a supplementary that is also a $\mathfrak{g}$-subrepresentation (it would be "only" a vector space). Thus, it is not always possible to decompose a $\mathfrak{g}$-representation into irreducible $\mathfrak{g}$-subrepresentations, but only into indecomposable ones. In this context, a

$\mathfrak{g}$-representation that can be decomposed into irreducible $\mathfrak{g}$-representations is called completely reducible.

2.2.2. Adjoint and Co-adjoint Representation

We can choose the vector space $V$ on which we represent $\mathfrak{g}$. Taking $V = \mathfrak{g}$, thus representing the Lie algebra on itself, we define the so-called adjoint representation of $\mathfrak{g}$, $\text{ad} : \mathfrak{g} \ni x \mapsto \text{ad}(x) = [x, \bullet]_{\mathfrak{g}} \in \mathfrak{gl}(\mathfrak{g})$. In its matricial version, we recognize the matrices $A$ of the previous subsection. We see also that the set of matrices $A_i$ defining the adjoint representation is equivalent to the set of structure constants of $\mathfrak{g}$.

We can rewrite again the Equation (4), but now in terms of the adjoint representation. We get:

$$\forall x, y, t \in \mathfrak{g}, \qquad < \text{ad}(x).y, t > + < y, \text{ad}(x).t >= 0 \tag{9}$$

Thus, the statement that $\mathfrak{g}$ is quadratic with bi-invariant pseudo-metric $<,>$ is equivalent to the requirement that all endomorphisms $\text{ad}(x)$ are skew-symmetric endomorphisms with respect to $<,>$. Recalling the matrix version of Equation (4), that is Equation (6), we see that solving for a bi-invariant $Z$ amounts to finding a symmetric isomorphism of representations $Z$ between the adjoint representation of $\mathfrak{g}$, written in its matricial form as $x \mapsto A(x)$ and the representation written in its matricial form as $x \mapsto -A(x)^T$.

If we choose to represent the Lie algebra $\mathfrak{g}$ on the dual vector space $\mathfrak{g}^*$, *i.e.*, we choose $V = \mathfrak{g}^*$, we can define the co-adjoint representation $\theta$: $\mathfrak{g} \ni x \mapsto \theta(x) \in \mathfrak{gl}(\mathfrak{g}^*)$, where $< \theta(x).f, t >=< f, \text{ad}(x).t >$ for $f \in \mathfrak{g}^*$, $x, y \in \mathfrak{g}$ and $<,>$ the inner product used to define the dual basis. If we write $A(x)$ the matrix of the endomorphism $\text{ad}(x)$, $T(x)$ the matrix of the endomorphism $\theta(x)$ and $Z$ the inner product defining the dual basis, the previous definition states that $Z$ is in fact an isomorphism of representation between the co-adjoint representation $x \mapsto T(x)$ and the representation: $x \mapsto A(x)^T$.

Now, if the inner product $<,>$ used to define the dual basis is bi-invariant, by identifying the vector spaces $\mathfrak{g}$ and $\mathfrak{g}^*$, we can again rewrite Equation (4) to get:

$$\forall x, y, t \in \mathfrak{g}, \qquad < \text{ad}(x).y, t > + < \theta(x).y, t >= 0 \tag{10}$$

We conclude that the bi-invariance of the inner product implies the following relation between the adjoint and co-adjoint representations: $\text{ad} = -\theta$. As $Z$ (that represents $<,>$) is an isomorphism of representations between the co-adjoint and the representation $x \mapsto A(x)^T$, we recover that the statement of $Z$ being a bi-invariant pseudo-metric on $\mathfrak{g}$ is equivalent to $Z$ being a symmetric isomorphism of representations between $x \mapsto A(x)$ and $x \mapsto -A(x)^T$.

2.2.3. Some Vocabulary of Algebra

The adjoint representation is related to the structure constants of $\mathfrak{g}$ and, thus, completely characterizes $\mathfrak{g}$. Thus, it links the language of abstract algebras and the language of representations for $\mathfrak{g}$.

For the special case of the adjoint representation ad, $\mathfrak{g}$-subrepresentations are ideals of $\mathfrak{g}$, irreducible $\mathfrak{g}$-representations are minimal ideals of $\mathfrak{g}$ and indecomposable $\mathfrak{g}$-representations are ideals of $\mathfrak{g}$ that cannot be decomposed into a direct sum of ideals of $\mathfrak{g}$. We will use the two languages of ideals or of representations.

If the adjoint representation is itself irreducible, but not one-dimensional, $\mathfrak{g}$ is said to be simple. If the adjoint representation is completely reducible, $\mathfrak{g}$ is said to be reductive. If the adjoint representation is completely reducible without one-dimensional subrepresentations, $\mathfrak{g}$ is semi-simple. If the adjoint representation is completely reducible with only one-dimensional subrepresentations, $\mathfrak{g}$ is abelian. A reductive Lie algebra is thus the sum (in the sense of subrepresentations) of a semi-simple Lie algebra and an abelian Lie algebra.

#### 2.2.4. Some Vocabulary of Geometry

An ideal $I$ of a Lie algebra $B$ is said to be isotropic with respect to a pseudo-metric given on $B$ if $I \cap I^{\perp} \neq \{0\}$. The ideal $I$ is said to be totally isotropic if $I \subset I^{\perp}$. The intersection between $I$ and $I^{\perp}$ represents the vectors that are orthogonal to themselves and, thus, that have zero norm, even if they are themselves non-zero.

Thus, isotropic ideals appear only in the case of a pseudo-metric that is not a metric. From the intuition provided by theoretical physics, we can interpret the vectors in $I \cap I^{\perp}$ as photons: they have zero mass even if they have non-zero velocity.

### *2.3. Constructions with Lie Algebra Representations*

We have seen that we can study the structure of a given Lie algebra by looking at its representations and more particularly at its adjoint representation. Here, we study decompositions of the adjoint representation that will be pertinent for the characterization of quadratic Lie algebras: the direct sum decomposition and the double extension decomposition. We show how these decompositions can be implemented in a computational framework. In this subsection, we use the notation $(B, [,]_B)$ to denote the Lie algebra, because this is the notation that we will use in the core of our algorithm (see Section 4).

#### 2.3.1. Definition of Direct Sum

$B = B_1 \oplus_B B_2$ is the direct sum of $B_1$, $B_2$ if:

- $B = B_1 \oplus B_2$ in terms of vector spaces,
- $[B, B_1]_B \subset B_1$ and $[B, B_2]_B \subset B_2$, making $B_1$ and $B_2$ subrepresentations of the adjoint representation of $B$, in other words: ideals of $B$.

This decomposition was first studied by [16]. We illustrate it with the matrices $A$ representing the adjoint representation $b \mapsto [b, \bullet]_B$ of $B$, *i.e.*, the matrices denoted: $b \mapsto A(b) = [b, \bullet]_B$. The direct

sum of $B$ is equivalent to the decomposition of the adjoint representation into the $B$-representations $B_1$ and $B_2$ *i.e.*,:

$$A(b) = \begin{pmatrix} A(b_1) & 0 \\ 0 & A(b_2) \end{pmatrix} \tag{11}$$

on a basis respecting $B = B_1 \oplus_B B_2$. Note that we write $\oplus_B$ to emphasize the fact that this direct sum decomposition is more than the direct sum decomposition into vector spaces.

2.3.2. Direct Sum Decomposition and Bi-Invariant Pseudo-Metrics

We have the following property: $B$ being quadratic is equivalent to $B_1$ and $B_2$ being quadratic. Indeed, if $<,>_{B_1}$, $<,>_{B_2}$ are bi-invariant pseudo-metrics on $B_1$, $B_2$ and represented by the matrices $Z_{B_1}$, $Z_{B_2}$, then:

$$Z_{B_1 \oplus_B B_2} = \begin{pmatrix} Z_{B_1} & 0 \\ 0 & Z_{B_2} \end{pmatrix} \tag{12}$$

is bi-invariant on $B$. Conversely, if $<,>_B$ is bi-invariant on $B$, its restrictions $<,>_B|_{B_1}$ and $<,>_B|_{B_2}$ are bi-invariant on $B_1$, $B_2$ [20,21].

2.3.3. Computing the Direct Sum

The direct sum decomposition of a Lie algebra $B$ into indecomposable subrepresentations is unique, up to isomorphisms. In practice, writing $\mathcal{B}_B = \{e_k\}_{k=1}^{\dim(B)}$ a basis of $B$ and $A_k = A(e_k)$, computing the direct sum decomposition of $B$ into indecomposable $B_i$'s amounts to the simultaneous bloc diagonalization of the matrices $A_k$.

2.3.4. Definition of Double Extension

$B = W \oplus S \oplus S^*$ is the double extension of $W$ by a simple $S$ if:

- $B = W \oplus S \oplus S^*$ in terms of vector spaces,
- $(W, [,]_W)$ is a Lie algebra and $[S, W]_B \subset W$ makes $W$ a $S$-representation,
- $(S, [,]_S)$ is a simple Lie subalgebra of $B$: $[s, s']_B = [s, s']_S$,
- $S^*$ is the dual space of $S$ and $[S, S^*]_B \subset S^*$ makes $S^*$ the co-adjoint representation,
- $\forall w, w' \in W: \quad [w, w']_B = [w, w']_W + \beta(w, w')$ where $\beta : \Lambda^2 W \mapsto S^*$ is a (skew-symmetric) $S$-equivariant map, *i.e.*, a map that commutes with the action of $S$.

This definition relies on the framework introduced in [21], or in [17] under the appellation "bi-extension". Here, we can illustrate it with the matrices representing the adjoint representation $b \mapsto [b, \bullet]_B$ of $B$, *i.e.*, the matrices denoted: $b \mapsto A(b)$. The double extension decomposition is equivalent to the following decomposition of the adjoint representation of $B$:

$$A(b) = \begin{pmatrix} [w, \bullet]_W + [s, \bullet]_B & [w, \bullet]_B & 0 \\ 0 & [s, \bullet]_S & 0 \\ \beta(w, \bullet) & [f, \bullet]_B & [s, \bullet]_B \end{pmatrix} \tag{13}$$

on a basis respecting $B = W \oplus S \oplus S^*$ and $b = w + s + f$. Note that, in the blocks of the matrix $A(b)$, we have identified endomorphisms with their corresponding matrices.

The definition of double extension uses a number of different notations. First, we recognize $\text{ad}(s) = [s, \bullet]_S$ and $\text{ad}(w) = [w, \bullet]_W$ to be respectively the adjoint representation of $S$ (on $S$) and the adjoint representation of $W$ (on $W$). However, $[s, \bullet]_B$ is a $S$-representation on $W$ that has nothing to do with the adjoint (the adjoint is a representation of a Lie algebra on itself).

Then, we should be careful with the structures that are manipulated. For example, we can consider the vector space $S^*$ as an abelian Lie subalgebra of $B$. However, we cannot consider $W$ as a subalgebra of $B$. The skew-symmetric map $\beta$ represents precisely the corresponding obstruction.

2.3.5. Double Extension Decomposition and Bi-Invariant Pseudo-Metrics

We have the following property: $B$ being quadratic is equivalent to $W$ being quadratic. Indeed, if $<,>_W$ is bi-invariant on $W$, represented by $Z_W$, then:

$$Z_{W\oplus S\oplus S^*} = \begin{pmatrix} Z_W & 0 & 0 \\ 0 & 0 & \mathbb{I} \\ 0 & \mathbb{I} & 0 \end{pmatrix} \tag{14}$$

is bi-invariant on $B$. Conversely, if $B$ is quadratic and written as a double extension of $W$ with $S$ simple (or one-dimensional), then the restriction $<,>_W = <,>_B |_W$ is bi-invariant [20,21]. Note here that we can write the $\mathbb{I}$-blocks, because the basis of $S$ and $S^*$ are chosen to be duals of each other. If two different basis were chosen, the corresponding bi-invariant pseudo-metric on $B = W \oplus S \oplus S^*$ would have the form:

$$Z_{W\oplus S\oplus S^*} = \begin{pmatrix} Z_W & 0 & 0 \\ 0 & 0 & L \\ 0 & L^T & 0 \end{pmatrix} \tag{15}$$

with $L$ an invertible matrix representing precisely the change of basis. More precisely, by computing Equation (6) on this last $Z_{W\oplus S\oplus S^*}$ while choosing $s \in S$, we show that $L$ is necessarily an isomorphism of $S$-representations on $S$ and $I$, *i.e.*, $L \in \text{Hom}_S(S, S^*)$. This remark will be used in practice in the algorithm (see Section 4).

2.3.6. Computing Double Extensions

Contrary to the direct sum decomposition, the decomposition of a quadratic Lie algebra $B$ as a double extension is not necessary unique. For example, given a quadratic indecomposable non-simple $B$, we can build a double extension decomposition from each minimal ideal of $B$ [21]. It proceeds as follows. We take a minimal ideal $I$ of $B$ and consider $I^\perp$ its orthogonal with respect to a bi-invariant pseudo-metric $<,>_B$. The decomposition:

$$B = W \oplus S \oplus S^* \qquad \text{where:} \qquad W = I^\perp/I, \quad S = B/I^\perp \text{ and } S^* = I$$

is a double extension of $W$ with $S$ simple (or one-dimensional). Moreover, one can show that $I$ and $I^\perp$ verify the following properties:

- $I$ is abelian,
- $I^\perp$ is a maximal ideal,
- $I \subset I^\perp$ (total isotropy),
- $[I, I^\perp] = 0$ (commutativity),
- $\text{codim}(I^\perp) = \dim(I)$.

These necessary conditions are taken from [16,20,21].

In practice, in our algorithm, we will have to build a double extension from a $B$ in order to compute a bi-invariant pseudo-metric on $B$, if it exists (see Section 4). Therefore, even if we know an abelian minimal ideal $I$ of $B$, we will not have its orthogonal $I^\perp$ needed for the construction shown above: we do not know any bi-invariant pseudo-metric, as we want to build one!Thus, given an abelian minimal ideal $I$, we shall test all ideals $J$ that could be an $I^\perp$ for a bi-invariant pseudo-metric, *i.e.*, all ideals $J$ that verify the necessary conditions listed above.

We show here that the only plausible ideals that can play the role of $I^\perp$ are either $J = C_B(I)$ the centralizer of $I$ in $B$ in the case $C_B(I) \neq B$ or the maximal ideals of codimension one containing $I$ in the case $C_B(I) = B$.

We have seen above that the first necessary condition for a $J$ to be an $I^\perp$ is its commutativity with $I$: $[I, J] = 0$. We recall that the centralizer $C_B(I)$ of $I$ in $B$ is defined as the set of elements that commute with $I$. Thus: $J \subset C_B(I)$.

Another necessary condition for a plausible $J$ is to be a maximal ideal. As $I$ is an ideal, $C_B(I)$ is also an ideal. Thus, $J$ is a maximal ideal included in the ideal $C_B(I)$: we have necessarily $J = C_B(I)$ in the case $C_B(I) \neq B$. In this case, the condition $I \subset J$ is fulfilled as $I$ is abelian. The last necessary condition to check is $\text{codim}(C_B(I)) = \dim(I)$.

However, if $C_B(I) = B$, then we shall look for maximal ideals of $B$. However, in this case, $I$ commutes with all elements of $B$, and therefore, $I$ is necessarily of dimension one as a minimal ideal. Therefore, we shall look for maximal ideals $J$ of codimension one. Adding the last necessary condition, we conclude that in the case $C_B(I) = B$, we shall consider only maximal ideals of codimension one containing $I$.

## 3. Structure of Quadratic Lie Groups

Here, we characterize the structure of quadratic Lie algebras, using the constructions defined in the previous section. We first present a reformulation of a classification theorem of quadratic Lie algebras. Then, we emphasize which Lie algebras we add by asking for a bi-invariant pseudo-metric instead of a bi-invariant metric. We finally investigate how we can go from a bi-invariant pseudo-metric to a bi-invariant dual metric on a special class of Lie algebra with bi-invariant pseudo-metrics.

*3.1. A Classification Theorem*

To characterize the structure of a quadratic Lie algebra, we use a reformulation of a classification theorem than can be found in [21] or [17].

**Theorem 1** (Classification of quadratic Lie algebras)**.** *The Lie algebra* $\mathfrak{g}$ *is quadratic if and only if its adjoint representation decomposes into indecomposable subrepresentations* $B$ *that are of the following types:*

- *Type (1):* $B$ *is simple (or one-dimensional),*
- *Type (2):* $B = W \oplus S \oplus S^*$ *is a double extension of a quadratic* $W$ *by* $S$ *simple (or one-dimensional).*

This means that any quadratic Lie algebra writes $g = B_1 \oplus_{\mathfrak{g}} ... \oplus_{\mathfrak{g}} B_N$, where each $B$ is of Type (1) or of Type (2). In particular, we can already conclude that any reductive (*a fortiori*, semi-simple or abelian) Lie algebra $\mathfrak{g}$ is quadratic. Moreover, if $\mathfrak{g}$ is quadratic, but not reductive, then $\mathfrak{g}$ has non-irreducible indecomposable subrepresentations, and these are necessarily double extensions of Type (2).

We recall that the notions of representation decomposition come from a simultaneous diagonalization of matrices. Therefore, they depend on the base field $\mathbb{F}$: a Lie algebra reductive in $\mathbb{R}$ is reductive in $\mathbb{C}$, but the converse is false. Thus, being quadratic also depends on the field that we consider. A Lie algebra quadratic on $\mathbb{R}$ will be quadratic on $\mathbb{C}$, but the converse is false.

3.1.1. Elementary Bi-Invariant Pseudo-Metrics

The previous characterization of quadratic Lie algebras in terms of their structure is useful in practice. It enables one to construct a type of bi-invariant pseudo-metric $<,>_{\mathfrak{g}}$ that exists necessarily on a quadratic $\mathfrak{g}$. We call this type of pseudo-metrics the elementary bi-invariant pseudo-metrics of $\mathfrak{g}$.

The elementary bi-invariant pseudo-metric $<,>_B$ of a one-dimensional Lie algebra $B$ is defined to be the multiplication. The elementary bi-invariant pseudo-metric $<,>_B$ of a simple Lie algebra $B$ is defined to be the Killing form. Now, let us define recursively the elementary bi-invariant pseudo-metrics of a general quadratic $\mathfrak{g}$.

Let us be given a quadratic Lie algebra $\mathfrak{g}$ on which we know an auxiliary bi-invariant pseudo-metric $<,>_{\mathfrak{g}}$ (not necessarily of the elementary type). First, we decompose the adjoint representation of $\mathfrak{g}$ into indecomposable subrepresentations $B$'s: $\mathfrak{g} = B_1 \oplus_{\mathfrak{g}} ... \oplus_{\mathfrak{g}} B_N$. Then, we study separately the two cases: the $B$'s of Type (1) and the $B$'s of Type (2).

On the $B$'s of Type (1), we define the elementary bi-invariant pseudo-metric $<,>_B$ as above: the multiplication if $B$ is one-dimensional or the Killing form if $B$ is simple.

On the $B$'s of Type (2), we build a double extension. To this aim, we consider a minimal ideal $I$, and using the auxiliary bi-invariant pseudo-metric $<,>_{\mathfrak{g}}$ of $\mathfrak{g}$, we compute $I^{\perp}$. We get the double extension $B = W \oplus S \oplus S^*$ with $W = I^{\perp}/I$, $S = B/I^{\perp}$ and $S^* = I$. We construct an elementary

bi-invariant pseudo-metric $<,>_W$ on $W$ recursively. We then define an elementary bi-invariant pseudo-metric $<,>_B$ on the double extension $B = W \oplus S \oplus S^*$ to be of the form of Equation (14).

Finally, we define the elementary bi-invariant pseudo-metric $<,>_{\mathfrak{g}}$ on the direct sum decomposition $\mathfrak{g} = B_1 \oplus_{\mathfrak{g}} ... \oplus_{\mathfrak{g}} B_N$ to be of the form of Equation (12). This construction defines (and proves the existence of) elementary bi-invariant pseudo-metrics on a quadratic $\mathfrak{g}$.

### *3.2. Riemannian and Pseudo-Riemannian Quadratic Lie Groups*

The previous characterization of quadratic Lie algebras can be refined to distinguish between quadratic Lie algebras that admit bi-invariant metrics with respect to quadratic Lie algebras with bi-invariant pseudo-metrics. In other words, it answers the questions: which Lie algebras do we add by removing the positivity of the metric?

#### 3.2.1. Studying the Signature

We recall from Section 2 that a metric on $\mathfrak{g}$ of dimension $n$ has signature $(n, 0)$. Now, we take a quadratic $\mathfrak{g}$ that is decomposed into indecomposable pieces $\mathfrak{g} = B_1 \oplus_{\mathfrak{g}} ... \oplus_{\mathfrak{g}} B_N$, where the $B_i$ are either simple (or one-dimensional) or double extensions. The signature on the direct sum is the sum of the signatures on the $B_i$ [39]:

$$\text{sgn}_{\mathfrak{g}} = \text{sgn}_{B_1} + ... + \text{sgn}_{B_N} \tag{16}$$

Therefore, asking for a positive definite signature on $\mathfrak{g}$ is equivalent to asking for a positive definite signature on each of the $B$'s.

If $B$ is simple, it possesses a bi-invariant metric if and only if it is compact. If $B$ is a double extension, a bi-invariant pseudo-metric has necessary a non-positive definite signature of the form [21]:

$$\text{sgn}_B = \text{sgn}_W + (m, m) \tag{17}$$

where $m$ is the dimension of the minimal ideal $I$ used to build the double extension.

We conclude that $\mathfrak{g}$ admits a bi-invariant metric if and only if its indecomposable parts are simple compact or one-dimensional, *i.e.*, if and only if $\mathfrak{g}$ is reductive with compact simple parts.

#### 3.2.2. Comparison

The trees of Figures 3 and 4 illustrate the comparison between Lie algebras with bi-invariant metrics and Lie algebras with bi-invariant pseudo-metrics.

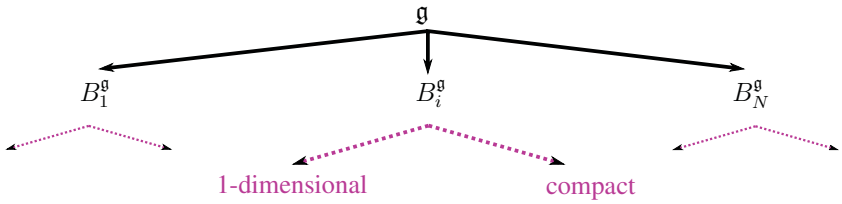



**Figure 3.** Structure of a Lie algebra with bi-invariant metrics.

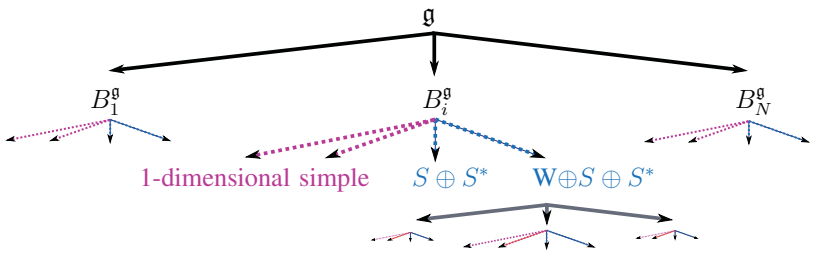



**Figure 4.** Structure of a Lie algebra with bi-invariant pseudo-metrics.

Thus, going from Riemannian to pseudo-Riemannian enables to add the simple algebras that generalize the compact algebras and the double extension structures (in blue) with its recursive construction that is not present in the Riemannian case.

### *3.3. From a Bi-Invariant Pseudo-Metric to a Bi-Invariant Dual Metric?*

We investigate here a special case of Lie algebras that we gain by going from Riemannian to pseudo-Riemannian: the double extension of $W = \{0\}$ by a compact simple Lie algebra $K$, which is an example of a Manin triple (see [31,32]). We will see in this subsection that we can view this case as a Riemannian case by changing the base field $\mathbb{F}$ (which is $\mathbb{R}$ or $\mathbb{C}$ for us) to its dual algebra $\mathbb{D}$. This development is a new contribution, which is a justification and an extension of the dual quaternions for $SE(3)$.

#### 3.3.1. Dual Numbers and Vectors

Given a field $\mathbb{F}$, the algebra $\mathbb{D}$ of dual numbers over this field is defined as $\mathbb{D} = \mathbb{F} + \epsilon\mathbb{F}$, where $\epsilon^2 = 0$ and $\epsilon \neq 0$ defines the multiplication [44]. We can define an $m$-dimensional dual vector space $\mathbb{D}^m = \mathbb{F}^m + \epsilon\mathbb{F}^m$, whose elements are dual vectors. Note here that the term "vector" is abusive in the sense that a vector space is usually defined on a field, not on an algebra. In the following, in order to study the properties of the dual vector space, we will use the dual map:

$$\begin{aligned} \psi : \quad & \mathbb{F}^m \oplus \mathbb{F}^m \mapsto \mathbb{D}^m \\ & x_0 + x_\epsilon \mapsto x_0 + \epsilon x_\epsilon \end{aligned}$$

using the same notation $\psi$ for mapping either to dual numbers or to dual vectors.

3.3.2. From the Double Extension $\mathfrak{g} = K \oplus K^*$ to Its Dual $\overline{\mathfrak{g}} = K + \epsilon K^*$

Now, we consider the double extension $\mathfrak{g} = K \oplus K^*$, where $K$ is compact simple and $\dim(K) = m$, so that $\dim(\mathfrak{g}) = 2m$. We take the following elementary bi-invariant pseudo-metric on $\mathfrak{g}$:

$$Z_{K\oplus K^*} = \begin{pmatrix} \mathbb{I} & \mathbb{I} \\ \mathbb{I} & 0 \end{pmatrix} \tag{18}$$

As $K$ and $K^*$ have same $\mathbb{F}$-dimension $m$, we consider the dual space $\overline{\mathfrak{g}} = K + \epsilon K^*$, of $\mathbb{D}$-dimension $m$. Its dual vectors write $\overline{x} = x_0 + \epsilon.x_\epsilon$, where $x_0 \in K$ and $x_\epsilon \in K^*$.

**Proposition 1.** *The dual map:*

$$\begin{aligned} \psi : \qquad \mathfrak{g} = K \oplus K^* &\mapsto \overline{\mathfrak{g}} \\ x_0 + x_\epsilon &\mapsto x_0 + \epsilon x_\epsilon \end{aligned}$$

*is an isomorphism of Lie algebras that respects the sum $K \oplus K^*$. The canonical inner product on $\overline{\mathfrak{g}}$ is bi-invariant and corresponds to the bi-invariant pseudo-metric $Z_{K\oplus K^*}$ above.*

This can been shown as follows. First, consider the Lie bracket on $\overline{\mathfrak{g}}$ inherited from $\psi$. We have:

$$\begin{aligned} [\psi(x), \psi(x')] &= [x_0 + \epsilon x_\epsilon, x_0' + \epsilon x_\epsilon'] \\ &= [x_0, x_0'] + \epsilon([x_0, x_\epsilon'] + [x_\epsilon, x_0']) \qquad (\text{as } \epsilon^2 = 0) \\ &= \psi([x, x']) \qquad (\text{definition of double extension}) \end{aligned}$$

which proves the isomorphism of Lie algebras.

We now show that the pseudo-metric $Z_{K\oplus K^*}$ on the Lie $\mathbb{F}$-algebra $\mathfrak{g}$ maps to the canonical metric $Z = \mathbb{I}$ on the Lie $\mathbb{D}$-algebra $\overline{\mathfrak{g}}$:

$$\begin{aligned} \psi(x)^T.\psi(x') &= (x_0 + \epsilon.x_\epsilon)^T.(x_0' + \epsilon x_\epsilon') \\ &= x_0^T.x_0' + \epsilon(x_\epsilon^T.x_0' + x_0^T.x_\epsilon') \\ &= \psi(x^T.Z_{K\oplus K^*}.x) \qquad (\text{using } \psi \text{ for dual numbers}) \end{aligned}$$

In others words, the spaces $\mathfrak{g}$ and $\overline{\mathfrak{g}}$ are isometric. However, again, the term "isometric" is abusive, as we recall that $\mathfrak{g}$ and $\overline{\mathfrak{g}}$ are not defined on the same field, the latter being defined on an algebra.

3.3.3. Towards Statistics on Dual Riemannian Manifolds

We have shown that a double extension $\mathfrak{g} = K \oplus K^*$ of $W = \{0\}$ by a compact simple $K$, endowed with a bi-invariant pseudo-metric, is isometrically isomorphic to a dual Lie algebra $\overline{\mathfrak{g}}$ with a bi-invariant metric. Thus, we could think of generalizing the theory of statistics on Riemannian manifolds to a theory of statistics on dual Riemannian manifolds. However, the fact that the space is defined on an algebra may cause some problems.

3.3.4. Generalization?

One could wonder if we can use this construction for any general double extension. However, we should note that this construction takes advantage of the fact that $K^*$ is totally isotropic and abelian. The element $\epsilon$, such that $\epsilon^2 = 0$, enables one to represent the commutativity of $K^*$ (Lie bracket is null) and the self-orthogonality of $K^*$ (the inner product is null) at the same time. A general Lie algebra with the bi-invariant pseudo-metric is not necessarily decomposable into two subspaces of same dimension, such that one of them is abelian and isotropic. For example, take a Lie algebra of an odd dimension.

## 4. An Algorithm to Compute Bi-Invariant Pseudo-Metrics on a Given Lie Group

We go back to the general case of any quadratic Lie algebra over the field $\mathbb{F}$ ($\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$). We present in this section an algorithm that computes bi-invariant pseudo-metrics on a Lie algebra given as input.

Then, we show how one could generalize the algorithm to compute all bi-invariant pseudo-metrics on $\mathfrak{g}$. Finally, we apply the algorithm to some Lie groups known to possess a unique bi-invariant mean: we find that most of them are not quadratic.

### *4.1. The Algorithm: Computation of One Bi-Invariant Pseudo-Metric*

For the computations, we will use matrix representations $Z$ of pseudo-metrics $<,>$, where the basis will be specified. The input is $\mathcal{B}_\mathfrak{g} = \{e_i\}_{i=1}^n$, a basis of $\mathfrak{g}$ and the structure constants $f_{ijk}$ on this basis. The output is a symmetric invertible matrix $Z_\mathfrak{g}$ on the basis $\mathcal{B}_\mathfrak{g}$, representing an elementary bi-invariant pseudo-metric, or a message of error: "the Lie algebra $\mathfrak{g}$ is not quadratic".

4.1.1. Core of the Algorithm

The core of the algorithm tests the structure of the Lie algebra given as input, to determine if it matches the characteristic tree-structure of quadratic Lie algebras described in the Section 3 (see Figure 4). Simultaneously with the progress through the tree, the algorithm tries to construct recursively an elementary bi-invariant pseudo-metric $<,>_\mathfrak{g}$ by testing all possible candidates. If it succeeds, we return the bi-invariant elementary pseudo-metric, proving that $\mathfrak{g}$ is quadratic. If not, we conclude that $\mathfrak{g}$ is not quadratic, and we return the error message. More precisely, the algorithm is divided into four steps as follows.

Step 1, direct sum decomposition: In this step, we decompose the adjoint representation of $\mathfrak{g}$ into indecomposable $B$'s, in other words: we decompose $\mathfrak{g}$ as a direct sum of $B$'s.

$$\mathfrak{g} = B_1 \oplus_\mathfrak{g} ... \oplus_\mathfrak{g} B_N \tag{19}$$

An implementation of this step can be found in [35].

From now on, we work on the basis $\mathcal{B}'_\mathfrak{g}$ that respects the direct sum: $\mathfrak{g} = B_1 \oplus_\mathfrak{g} ... \oplus_\mathfrak{g} B_N$. The $B$'s are indecomposable Lie algebras; thus, we can take advantage of the classification theorem 1

of Section 3. In the following two steps, we test if each $B$ is either of Type (1) (one-dimensional or simple) or of Type (2) (a double extension). testing Type (1): In this step, we test if the indecomposable $B$ is of Type (1), *i.e.*, if $B$ is one-dimensional or simple (see the dichotomy of Theorem 1).

To test if $B$ is one-dimensional, we can obviously count the number of basis vectors of $B$ in the basis $\mathcal{B}'_{\mathfrak{g}}$. If $B$ is found one-dimensional, we return the multiplication, which is an elementary bi-invariant pseudo-metric on $B$.

To test if $B$ is simple, we use a function that computes the radical of the Levi decomposition of $B$ [45]. The indecomposable piece $B$ is simple if and only if the radical is null. Such a function can be found in [36]. If $B$ is found simple, we return the Killing form, which is an elementary bi-invariant pseudo-metric.

If $B$ is neither one-dimensional nor simple, we conclude that $B$ is not of Type (1). We test in the following step if $B$ is of Type (2).

Step 3, testing Type (2): In this step, we test if $B$ is of Type (2), *i.e.*, if $B$ is a double extension of a quadratic $W$ by a simple $S$ (see the dichotomy of Theorem 1). We recall that the double extension structure of $B$ is not necessarily unique. Therefore, it might seem that we need to test all possible candidates for a double extension structure of $B$, in order to answer if $B$ is of Type (2). We proceed slightly differently.

As $B$ is indecomposable and not of Type (1) (see the previous steps), $B$ being of Type (2) is equivalent to $B$ being quadratic. More precisely, at this step of the algorithm, the following assertions are equivalents:

(a) $B$ is of Type (2),
(b) $B$ is quadratic,
(c) $\forall I$ minimal, $I$ abelian, there is a double extension decomposition of $B$,
(d) $\exists I$ minimal, abelian, such that there is a double extension decomposition of $B$.

Thus, we will consider only one minimal ideal $I$ of $B$ and try to construct a double extension out of it, of the form: $B = W \oplus S \oplus I$. Note that this step will need to call the algorithm recursively, to determine if the candidate for $W$ in the double extension structure is quadratic or not. The details of this step are below.

Step 3.a: First, we compute a minimal ideal $I$. More precisely, recalling the necessary conditions of the double extension structure of Section 2, we compute $I$, an abelian minimal ideal, which is also a minimal abelian ideal. A function that finds a minimal abelian ideal of $B$ can be derived from an algorithm of [46] that computes all abelian ideals of $B$: we can choose one of minimal dimension among those.

Step 3.b: Then, we compute $C_B(I)$, the maximal ideals $J$'s and the corresponding candidates for the double extension structure of $B$. The computation of $C_B(I)$ is implemented in [47].

If $C_B(I) \neq B$, we take $J = C_B(I)$ and verify the condition $\mathrm{codim}(J) = \dim(I)$. If the condition is not fulfilled, there is no double extension structure possible for $B$. Therefore, we conclude that $B$ is not of Type (2).

If $C_B(I) = B$, we compute the maximal ideals $J$ of $B$ of codimension one containing $I$ (see Section 2). If no such ideals are found, there is no double extension structure possible for $B$. Again, in this case, we conclude that $B$ is not of Type (2).

If $J$'s are found, we compute the corresponding double extension candidates of $B$, one per $J$, as:

$$B = W \oplus S \oplus S^* \text{ where: } W = J/I, S = B/J \text{ and: } S^* = I. \tag{20}$$

We call the algorithm recursively on $W$, *i.e.*, we determine recursively if $W$ is quadratic. If there is no double extension candidate with a quadratic $W$, we conclude that $B$ is not of Type (2). Otherwise, we keep the double extension candidates that have a quadratic $W$ (with an elementary bi-invariant pseudo-metric $Z_W$).

Step 3.c: Then, we try to compute an elementary pseudo-metric for all double extension candidates of the form: $B = W \oplus S \oplus S^*$, where $W = J/I$ is quadratic with corresponding $Z_W$, $S = B/J$ and $S^* = I$. Given a double extension candidate, we know from Section 2 that an elementary pseudo-metric on $B$ has the form:

$$Z_{B=W\oplus S\oplus I} = \begin{pmatrix} Z_W & 0 & 0 \\ 0 & 0 & L \\ 0 & L^T & 0 \end{pmatrix} \tag{21}$$

where $L \in \mathrm{Hom}_S(S, I)$.

Therefore, we need to compute $\mathrm{Hom}_S(S, I)$. We recall that $S$ is simple; thus, its adjoint representation is irreducible. As we are in the case of a finite dimensional irreducible representation, we can apply Schur's lemma. Its general form states that $\mathrm{Hom}_S(S, S)$ is an associative division algebra over $\mathbb{F}$ ($= \mathbb{R}$ or $\mathbb{C}$), which is of finite degree, because $S$ is finite dimensional [48]. When the base field is $\mathbb{F} = \mathbb{C}$, we use the fact that a finite-dimensional division algebra over an algebraically closed field is necessarily itself. Thus, $\mathrm{Hom}_S(S, S) = \mathbb{C}$ and $\dim_{\mathbb{C}}(\mathrm{Hom}_S(S, S)) = 1$. When the base field is $\mathbb{F} = \mathbb{R}$, we use the Frobenius theorem, which asserts that the only real associative division algebras are $\mathbb{R}$, $\mathbb{C}$ or $\mathbb{H}$, the field of quaternionnumbers [49]. Thus, $\mathrm{Hom}_S(S, S)$ is $\mathbb{R}$, $\mathbb{C}$ or $\mathbb{H}$, and $\dim_{\mathbb{R}}(\mathrm{Hom}_S(S, S))$ is 1, 2 or 4. Now, if $I$ and $S$ are isomorphic, $\mathrm{Hom}_S(S, I)$ is isomorphic to $\mathrm{Hom}_S(S, S)$ and, thus, of maximal dimension four over $\mathbb{F}$. Otherwise, if $I$ and $S$ are not isomorphic, we have $\mathrm{Hom}_S(S, I) = \{0\}$.

The computation of $\mathrm{Hom}_S(S, I)$ is implemented in [50], more generally for any finite-dimensional modules of a finitely generated algebra.

Step 3.d: To conclude Step 3, we determine if one of the possible elementary pseudo-metrics computed above is bi-invariant. To this aim, we plug the expression of $Z_{B=W\oplus S\oplus I}$ into Equations (7) and solve it for $L$. Thus, the initial system of Equations (7) has been reduced to an equation in maximum one (complex case) or in four (real case) parameters.

We run this step for each double extension candidate. If a bi-invariant elementary pseudo-metric $Z_B$ is found on one of the candidates, we return $Z_B$. Otherwise, we conclude that $B$ is not of Type (2).

Step 4, construction of a bi-invariant pseudo-metric on the whole $\mathfrak{g}$: In this step, we construct a bi-invariant (elementary) pseudo-metric on $\mathfrak{g}$, if it exists. If one $B$ of the direct sum decomposition

$\mathfrak{g} = B_1 \oplus_{\mathfrak{g}} ... \oplus_{\mathfrak{g}} B_N$ is neither of Type (1), nor of Type (2), we conclude from Theorem 1 that $\mathfrak{g}$ is not quadratic. We return the error message. Otherwise, we glue together the elementary bi-invariant pseudo-metrics $Z_B$'s that have been returned on the $B$'s.

More precisely, we follow the construction of Section 2 to build the elementary bi-invariant pseudo-metric $Z'_{\mathfrak{g}}$ on the basis $\mathcal{B}'_{\mathfrak{g}}$ of $\mathfrak{g}$ that respects the direct sum decomposition:

$$Z_{\mathfrak{g}=B_1\oplus_{\mathfrak{g}}...\oplus_{\mathfrak{g}}B_N} = \begin{pmatrix} Z_{B_1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & Z_{B_N} \end{pmatrix} \tag{22}$$

Finally, we perform a change of basis from $\mathcal{B}'_{\mathfrak{g}}$ to $\mathcal{B}_{\mathfrak{g}}$ in order to return $Z_{\mathfrak{g}}$, an elementary bi-invariant pseudo-metric on the basis of the Lie algebra given as input.

4.1.2. Tree Structure of the Algorithm

The algorithm has a natural tree structure presented in Figure 5. The bi-invariant pseudo-metric $Z_{\mathfrak{g}}$ is computed in a postfix manner. A tree level corresponds to a reduction of an adjoint representation: reduction of $\mathfrak{g}$ into $B$'s for the first level, reductions of the $W$'s into $B$'s for the others. The arrows in dashes represent the cases that we investigate to test if $\mathfrak{g}$ is quadratic. If $B$ is not in one of such cases, then $B$ is not quadratic, so neither is $\mathfrak{g}$, and we exit the algorithm.

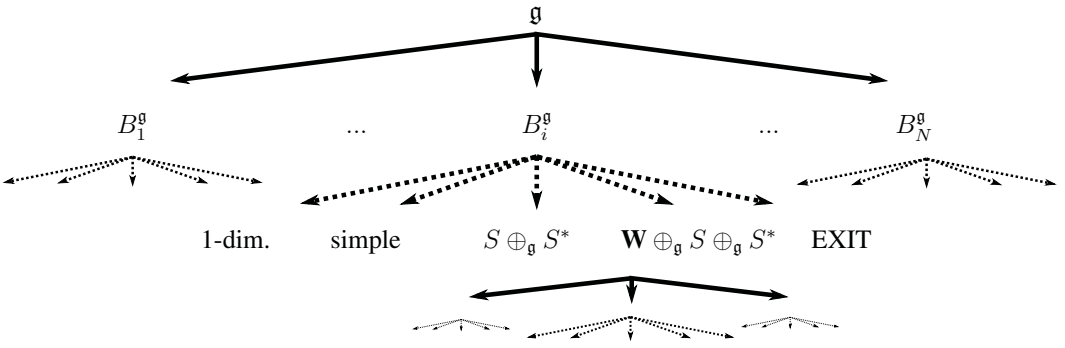



**Figure 5.** Tree structure of the algorithm.

In pseudo-code, the algorithm is written as follows.

**Algorithm 1** Computation of a bi-invariant pseudo-metric on $\mathfrak{g}$.

**Input:** $\mathcal{B}_{\mathfrak{g}} = \{e_i\}_i$ basis of $\mathfrak{g}$, Structure constants $f_{ijk}$ on this basis.
**Initialization:** $B = \mathfrak{g}$.
**Core**:
**switch** $(B)$:

- **case** $B$ is decomposable**:**
  decompose into $B = B_1 \oplus_B ... \oplus_B B_N$;
  call algorithm recursively on the $B_i$'s;
  **return**: $Z_B = \text{Diag}(Z_{B_1}, \ldots, Z_{B_N})$;
- **case** $B$ is 1-dimensional**:**
  **return** $Z_B = \Big(1\Big)$;
- **case** $B$ is simple**:**
  **return** $Z_B = Z_{\text{Killing}}$;
- **default:**
  compute $I$ minimal abelian ideal; **if** no $I$ exists: "EXIT"; **break**;
  compute its centralizer $C_B(I)$; **if** $codim(C_B(I)) \neq \dim(I)$: "EXIT"; **break**;
  compute $S = B/C_B(I)$, $W = C_B(I)/I$;
  call algorithm recursively on $B = W$;
  compute $\text{Hom}_S(S, I)$;
  solve Equation(7) for $L \in \text{Hom}_S(S, I) = \{0\}$ by plugging:

$$Z_{B=W\oplus S\oplus I} = \begin{pmatrix} Z_W & 0 & 0 \\ 0 & 0 & L \\ 0 & L^T & 0 \end{pmatrix},$$

  **if** there is no solution: "EXIT"; **break**;
  **return** $Z_B$.

**end switch**
**Output**:

- **if** "EXIT": **return** the message "The Lie algebra $\mathfrak{g}$ is not quadratic";
- **else: return** the elementary bi-invariant pseudo metric on $\mathfrak{g}$.

This gives a bi-invariant pseudo-metric on the Lie algebra $\mathfrak{g}$. We can then make it a bi-invariant pseudo-metric on the Lie group $\mathcal{G}$ by propagating it through $DL_g(e)$ (or $DR_g(e)$) on all tangent spaces $T_g\mathcal{G}$ (see Section 2).

All in all, the algorithm allows one to compute one bi-invariant pseudo-metric of $\mathfrak{g}$, *i.e.*, one invertible element of the quadratic space $\mathcal{Q}(\mathfrak{g})$. We can generalize the algorithm, in order to compute

all bi-invariant pseudo-metrics of $\mathfrak{g}$, thus the whole quadratic space $\mathcal{Q}(\mathfrak{g})$. This is the purpose of the next subsection.

*4.2. Generalization of the Algorithm: Computation of All Bi-Invariant Pseudo-Metrics*

Here, we present how one should proceed in order to compute all bi-invariant pseudo-metrics of a given Lie algebra $\mathfrak{g}$, *i.e.*, the whole quadratic space $\mathcal{Q}(\mathfrak{g})$. Note that the dimension of $\mathcal{Q}(\mathfrak{g})$ is unknown in the general case [30]. However, the algorithmic procedure allows one to compute the space anyway.

We follow the strategy of the previous algorithm: we decompose $\mathfrak{g}$ into indecomposable $B$'s; we compute the quadratic spaces $\mathcal{Q}(B)$ for each of them and then glue these spaces together to get $\mathcal{Q}(\mathfrak{g})$.

4.2.1. Computing the Quadratic Space of Indecomposable Lie Algebras

In this step, we compute the quadratic space for all indecomposable pieces $B$'s of $\mathfrak{g}$, the simple (or one-dimensional) and the double extensions.

The quadratic space of a one-dimensional piece $B$ is the weighted multiplication, so the whole base field $\mathbb{F}$:

$$\mathcal{Q}(B) = \{Z_B = \alpha\mathbb{I} \quad | \quad \forall \alpha \in \mathbb{F}\} = \mathbb{F} \tag{23}$$

The quadratic space of a simple piece $B$ is the vector space spanned by the Killing form.

$$\mathcal{Q}(B) = \{Z_B = \alpha Z_{\text{Killing}} \quad | \quad \forall \alpha \in \mathbb{F}\} \tag{24}$$

The quadratic space of a double extension $B = W \oplus S \oplus S^*$, where the basis of $S$ and $S^*$ are chosen duals, is given by:

$$\mathcal{Q}(B) = \left\{ Z_B = \begin{pmatrix} Z_W & M & N \\ M^T & \alpha Z_{Killing} & \beta\mathbb{I} \\ N^T & \beta\mathbb{I} & (0) \end{pmatrix} \;\middle|\; \begin{array}{c} \forall \alpha, \beta \in \mathbb{F}, \forall Z_W \in \mathcal{Q}(W), \\ \forall M, N \text{ solutions of equations derived from (7)} \end{array} \right\} \tag{25}$$

We leave to the reader the computations of the equations derived from Equations (7) that $M$ and $N$ are solving. Because of the dimension reduction, these equations can be solved in a lot of interesting cases. In our computations on selected Lie groups in the next subsection, $N$ and $N$ are vectors or scalars, for example.

4.2.2. Computing the Quadratic Space of a Direct Sum

The second step is the computation of the quadratic space of a direct sum $\mathfrak{g} = B_1 \oplus_{\mathfrak{g}} ... \oplus_{\mathfrak{g}} B_N$, given the quadratic spaces of each of its indecomposable pieces $B_i$. This gives:

$$\mathcal{Q}(\mathfrak{g}) = \left\{ Z_{\mathfrak{g}} \in \text{Sym}(n) \quad | \quad \text{s.t. for } i \in \{1, ..N\} \text{ (block index):} \quad \begin{array}{rcll} Z_{\mathfrak{g}ii} & = & Z_{Bi} \in \mathcal{Q}(B_i) & \text{if } i = j \\ Z_{\mathfrak{g}ij} & = & M_{ij} \text{ if } i < j & \end{array} \right\} \tag{26}$$

where $M_{ij}$ is a matrix that solves the following equation, derived from Equations (7):

$$A(b_i)^T.M_{ij} + M_{ij}.A(b_j) = 0 \quad \forall b_i \in B_i, \forall b_j \in B_j \tag{27}$$

In summary, the problem of computing all bi-invariant pseudo-metrics of a given $\mathfrak{g}$ amounts to the resolution of a reduced number of algebraic equations of lower dimension.

### *4.3. Results of the Algorithm on Selected Lie Groups*

We run our algorithm manually to determine if a bi-invariant pseudo-metric exists on some real Lie groups for which there is a locally unique bi-invariant mean: $SE(n)$, $ST(n)$, $H$ and $UT(n)$, for $n \in \mathbb{N}^*$ [2].

We run the computations manually and illustrate them, for each example, with the corresponding progress through the tree of the algorithm. The results show that most of these Lie groups are not quadratic.

#### 4.3.1. Scalings and Translations $ST(n)$

The Lie group $ST(n)$ comprises uniform scalings together with translations of $\mathbb{R}^n$. It is the semi-direct product $\mathbb{R}_+^* \ltimes \mathbb{R}^n$, its elements being written $(\lambda, t)$. More precisely, $ST(n)$ is defined by its action on $\mathbb{R}^n$: $(\lambda, t).x = \lambda.x + t$. The group law and the group inversion are written as follows: $(\lambda_1, t_1) * (\lambda_2, t_2) = (\lambda_1.\lambda_2, \lambda_1 * t_2 + t_1)$ and $(\lambda, t)^{(-1)} = (1/\lambda, -t/\lambda)$.
The Lie algebra $\mathfrak{st}(n)$ comprises the $(\mu, u) \in \mathbb{R} \oplus \mathbb{R}^n$ with Lie bracket:

$$[(\mu_1, u_1), (\mu_2, u_2)] = (0, \mu_2.u_1 - \mu_1.u_2). \tag{28}$$

Input: We choose the basis $(D, \{P_a\}_{a=1}^n)$ defined as: $D = (1, 0)$ and $P_a = (0, e_a)$ with $(e_a)_{a=1}^n$ the canonical basis of $\mathbb{R}^n$. In this basis, the structure constants can be read in the following Lie brackets:

$$\begin{aligned} [P_a, P_b] &= 0, \\ [D, P_a] &= P_a, \\ [D, D] &= 0. \end{aligned}$$

Step 1: From the expression of the Lie brackets above, we can compute all ideals of $\mathfrak{st}(n)$ manually and find: Span$(P_1)$, ..., Span$(P_n)$ and their linear combinations. We remark that there is no ideal containing $D$. Thus, $\mathfrak{st}(n)$ cannot be written as the direct sum of ideals, *i.e.*, $\mathfrak{st}(n)$ is indecomposable.

Step 2: First, as $n \in \mathbb{N}^*$, we have $\dim(\mathfrak{st}(n)) > 1$. Thus, $\mathfrak{st}(n)$ is not one-dimensional. Then, as Span$(P_1)$, for example, is an ideal, $\mathfrak{st}(n)$ is not simple. We conclude that $\mathfrak{st}(n)$ is not of Type (1).

Step 3: We take $I =$ Span$(P_1)$, which is obviously a minimal abelian ideal. From the commutation relations given by the Lie brackets, we see that $C_{\mathfrak{st}(n)}(I) =$ Span$(\{P_a\}_{a=1}^n)$, and we are in the case $C_{\mathfrak{st}(n)}(I) \neq \mathfrak{st}(n)$. Thus, there is only one double extension candidate, with $J = C_{\mathfrak{st}(n)}(I)$. We define $S = \mathfrak{st}(n)/J =$ Span$(D)$ and $W = J/I =$ Span$(P_2, ..P_n)$. We call the

algorithm recursively on $W$, which decomposes into one-dimensional ideals on which we return the multiplication.

The $S$-representation on $S$ is the null representation: $[D, D] = 0$. The $S$-representation on $I$ is the trivial representation: $[D, P_1] = P_1$. Hence, $I$ and $S$ are not isomorphic $S$-representations, and $\mathrm{Hom}_S(S, I)$ is zero. We conclude that $\mathfrak{st}(n)$ is not of Type (2).

Output: We have found that $\mathfrak{st}(n)$ is indecomposable and neither of Type (1) nor of Type (2). Thus, $\mathfrak{st}(n)$ is not quadratic: there is no bi-invariant pseudo-metric $<,>$ on $\mathfrak{st}(n)$.

This reasoning is illustrated on Figure 6 through the tree representation of the algorithm.

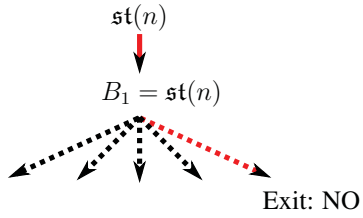



**Figure 6.** Schematical result for $ST(n)$. We see on the top level that $\mathfrak{st}(n)$ is indecomposable (it decomposes into itself). We see on the bottom level that $\mathfrak{st}(n)$ is neither one-dimensional, nor simple, nor a double extension, and therefore, we exit the algorithm: $\mathfrak{st}(n)$ is not quadratic.

4.3.2. Heisenberg Group H

The Heisenberg group $H$ comprises 3D upper triangular matrices $M$ of the form:

$$M = \begin{pmatrix} 1 & x & z \\ 0 & 1 & y \\ 0 & 0 & 1 \end{pmatrix}.$$

Thus, an element of this group can be written as $(x, y, z) \in \mathbb{R}^3$, with corresponding group law $(x_1, y_1, z_1) * (x_2, y_2, z_2) = (x_1 + x_2, y_1 + y_2, z_1 + z_2 + x_1 * y_2)$ and group inversion $(x, y, z)^{(-1)} = (-x, -y, -z + xy)$.

The Lie algebra $\mathfrak{h}$ comprises the nilpotent matrices:

$$N = \begin{pmatrix} 0 & p & c \\ 0 & 0 & q \\ 0 & 0 & 0 \end{pmatrix}.$$

Input: A basis for $\mathfrak{h}$ is thus $(P, Q, C)$ with clear notations. In this basis, the structure constants can be read in the following Lie brackets:

$$[C, P] = 0,$$
$$[C, Q] = 0,$$
$$[P, Q] = C.$$

Step 1: From the expression of the Lie brackets above, we can compute all ideals of $\mathfrak{h}$ manually, and we find: Span$(C)$, Span$(C, P)$ and Span$(C, Q)$. We remark that there is no ideal whose supplementary is also an ideal. Thus, $\mathfrak{h}$ is indecomposable.

Step 2: $\mathfrak{h}$ is obviously not one-dimensional. Moreover, as Span$(C)$, for example, is an ideal, $\mathfrak{h}$ is not simple. We conclude that $\mathfrak{h}$ is not of Type (1).

Step 3: We take $I = \text{Span}(C)$, which is a minimal abelian ideal of $\mathfrak{h}$. From the commutation relations given by the Lie brackets, we compute the commutator of $I$, and we see that we are in the case $C_{\mathfrak{h}}(I) = \mathfrak{h}$. Thus, we consider all maximal ideals of $\mathfrak{h}$ that are of codimension one and contain $I$. We get $J = \text{Span}(C, P)$ or $J = \text{Span}(C, Q)$; thus, we have two double extension candidates. By symmetry in $P \leftrightarrow Q$ (see the structure constants), we can consider $J = \text{Span}(C, P)$ only, without lost of generality. We define $S = \mathfrak{h}/J = \text{Span}(Q)$ and $W = J/I = \text{Span}(P)$. We call the algorithm recursively on $W$. As $W$ is one-dimensional, $W$ is quadratic, and we return $Z_W = \begin{pmatrix} 1 \end{pmatrix}$.

The $S$-representation on $S$ is given by the bracket $[Q, Q] = 0$: it is the null representation. The $S$-representation on $I$ is given by the bracket $[Q, C] = 0$: it is also the null representation. The isomorphism of vector spaces $L$ that maps $C$ on $Q$ is an isomorphism of representations, whose matricial form is the identity in our basis. The dimension of $\text{Hom}_S(S, I)$ is obviously one.

Thus, we plug:

$$Z_{W \oplus S \oplus I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

into Equation (6) to determine if it is bi-invariant. Computations show that it is not. We conclude that $\mathfrak{h}$ is not of Type (2).

Output: We have found that $\mathfrak{h}$ is indecomposable and neither of Type (1) nor of Type (2). Thus, $\mathfrak{h}$ is not quadratic: there is no bi-invariant pseudo-metric $<,>$ on $\mathfrak{h}$.

We try the algorithm on the general Heisenberg algebra $\mathfrak{h}_{2m+1}$, which is defined abstractly by the basis $\left\{C, \{P_i\}_{i=1}^m, \{Q_j\}_{j=1}^m\right\}$ and the Lie bracket:

$$\begin{aligned} [C, P_i] &= 0, \\ [C, Q_j] &= 0, \\ [P_i, Q_j] &= \delta_{ij} \end{aligned}$$

where $\delta$ is the Kronecker symbol. We are in the same situation as with $\mathfrak{h}$, except that $W$ is abelian (but not necessarily one-dimensional). We thus decompose $W$ into abelian one-dimensional ideals, and we return the following elementary bi-invariant pseudo-metric:

$$Z_W = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

However, we exit the algorithm as previously. Thus, the algorithm confirms that the general $\mathfrak{h}_{2m+1}$ has no bi-invariant pseudo-metric [20].

This reasoning is illustrated on the left hand side of Figure 7 through the tree representation of the algorithm.

### 4.3.3. The Group of Scaled Upper Unitriangular Matrices $UT(n)$

The group $UT(n)$ comprises the upper triangular matrices $M$ of the form: $M = \lambda.Id + N$, where $\lambda > 0$ and $N$ an upper triangular nilpotent matrix.

The Lie algebra $\mathfrak{ut}(n)$ comprises the matrices of the form $X = \mu.Id + Y$, where $\mu \in \mathbb{R}$ and $Y$ an upper triangular nilpotent matrix, the Lie bracket being the commutator of matrices.

Now, $\mathfrak{ut}(n)$ is decomposable into the one-dimensional Lie algebra generated by $\mathbb{I}$ and the Heisenberg algebra $\mathfrak{h}$. As $\mathfrak{h}$ has no bi-invariant pseudo-metric, neither does $\mathfrak{ut}(n)$.

This reasoning is illustrated on the right hand side of Figure 7 through the tree representation of the algorithm.

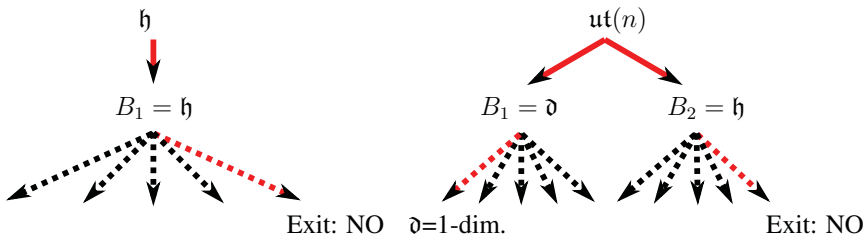



**Figure 7.** Schematical result for $H$ and $UT(n)$. The top level indicates the direct sum decomposition step. Thus, $\mathfrak{h}$ is indecomposable, and $\mathfrak{ut}(n)$ decomposes into $\mathfrak{d}$ and $\mathfrak{h}$. The bottom level for $\mathfrak{h}$ indicates that $\mathfrak{h}$ is neither one-dimensional, nor simple, nor a double extension, and therefore, we exit the algorithm: $\mathfrak{h}$ is not quadratic. The bottom level for $\mathfrak{ut}(n)$ indicates that $\mathfrak{d}$ is one-dimensional and therefore quadratic, but that $\mathfrak{h}$ is not quadratic: $\mathfrak{ut}(n)$ is not quadratic.

### 4.3.4. Rigid Body Transformations $SE(n)$

The group of isometries $SE(n)$ comprises rotations together with translations of $\mathbb{R}^n$. It is the semi-direct product $SO(n) \ltimes \mathbb{R}^n$, its elements being written $(R, t)$. More precisely, $SE(n)$ is defined by its action on $\mathbb{R}^n$ as $(R, t).x = R.x + t$. The group law and the group inversion are $(R_1, t_1) * (R_2, t_2) = (R_1.R_2, R_1 * t_2 + t_1)$ and $(R, t)^{(-1)} = (R^{(-1)}, R^{(-1)}.(-t))$.

The Lie algebra $\mathfrak{se}(n)$ comprises the $(A, u) \in Skew(n) \oplus \mathbb{R}^n$ with Lie bracket:

$$[(A_1, u_1), (A_2, u_2)] = (A_1.A_2 - A_2.A_1, A_1.u_2 - A_2.u_1) \tag{29}$$

Input: We choose the basis: $(\{J_{ij}\}_{1\leq i<j\leq n}, \{P_a\}_{a=1}^n)$ with $J_{ij} = e_{i.e.,j}^T - e_j.e_i^T$ and $\{P_a\}_{a=1}^n$ the canonical basis of $\mathbb{R}^n$. In this basis, the structure constants can be read in the following Lie brackets:

$$\begin{aligned}
[J_{ij}, J_{kl}] &= \delta_{ik}.J_{jl} - \delta_{jk}.J_{il} + \delta_{jl}.J_{ik} - \delta_{il}.J_{jk},\\
[J_{ij}, P_a] &= \delta_{aj}.P_i - \delta_{ai}.P_j,\\
[P_a, P_b] &= 0,
\end{aligned}$$

with $\delta$ the Kronecker symbol.

As preliminaries, we show that $P = \text{Span}(\{P_a\}_{a=1}^n)$ is the only proper ideal of $\mathfrak{se}(n)$. First, we see from the Lie brackets that $P$ is a proper ideal of $\mathfrak{se}(n)$. Suppose that $\mathfrak{se}(n)$ has another proper ideal $K$. Then, either $K \cap P$ is a proper ideal of $\mathfrak{se}(n)$ included in $P$ or $K \subset \mathfrak{so}(n)$ is a proper ideal of $\mathfrak{se}(n)$. $P$ does not contain any proper ideal of $\mathfrak{se}(n)$, because $\mathfrak{so}(n)$ acts transitively on $P$ with the Lie bracket. We can show that $\mathfrak{so}(n)$ does not contain any proper ideal of $\mathfrak{se}(n)$ (considering independently the case $n = 4$). Thus, $P$ is the only proper ideal of $\mathfrak{se}(n)$.

Step 1: The Lie algebra $\mathfrak{se}(n)$ has only one ideal $P$. Thus, $\mathfrak{se}(n)$ cannot be decomposed as a direct sum of ideals. We conclude that $\mathfrak{se}(n)$ is indecomposable.

Step 2: If $n = 1$, $\mathfrak{se}(1)$ is obviously one-dimensional. We return the multiplication, which is a bi-invariant pseudo-metric on $\mathfrak{se}(1)$. Otherwise, $\dim(\mathfrak{se}(n)) > 1$. As $P$ is an ideal of $\mathfrak{se}(n)$, $\mathfrak{se}(n)$ is not simple. We conclude that $\mathfrak{se}(1)$ is quadratic with the multiplication as the bi-invariant pseudo-metric and that $\mathfrak{se}(n)$ with $n > 1$ is not of Type(1). We go on with $n > 1$.

Step 3: We take $I = P$ and $J = C_{\mathfrak{se}(n)}(I) = P = I$. The necessary condition $\text{codim}(J) = \dim(I)$ is verified only for $n = 3$. We conclude that $\mathfrak{se}(n)$ is not of Type (2) if $n \neq 3$. We go on with $n = 3$. We compute $S = \mathfrak{se}(3)/P \sim \mathfrak{so}(3)$ and $W = P/P = \{0\}$.

In order to study the $S$-representations, we write the Lie bracket as:

$$\begin{aligned}
[J_m, J_n] &= \epsilon_{mnp}.J_p,\\
[J_m, P_a] &= \epsilon_{map}.P_p,\\
[P_a, P_b] &= 0
\end{aligned}$$

where we define $J_1 = J_{23}, J_2 = J_{31}$ and $J_3 = J_{12}$. The $S$-representation on $S$ is the adjoint representation: $[J_m, J_n] = \epsilon_{mnp}.J_p$. The $S$-representation on $I = P$ is given by: $[J_m, P_a] = \epsilon_{map}.P_p$. It is also the adjoint representation. The isomorphism of vector spaces $L$ that maps each $P_a$ on $J_a$ is an isomorphism of representations whose matricial form is the identity in our basis.

Hence, we write $Z_{\mathfrak{se}(3)}$ on the decomposition $S \oplus I = \mathfrak{so}(3) \oplus P$ with basis $(\{J_a\}_{a=1}^3, \{P_a\}_{a=1}^3)$ and get:

$$Z_{\mathfrak{se}(3)} = \begin{pmatrix} 0 & \mathbb{I}_3 \\ \mathbb{I}_3 & 0 \end{pmatrix}. \tag{30}$$

We plug it into Equations (7). Running the computation shows that the pseudo-metric $Z_{\mathfrak{se}(3)}$ is bi-invariant on $\mathfrak{se}(3)$. $Z_{\mathfrak{se}(3)}$ is actually known as the Klein form [51].

Output: $\mathfrak{se}(1)$ is quadratic; we return the multiplication, which is a bi-invariant pseudo-metric on $\mathfrak{se}(1)$. $\mathfrak{se}(3)$ is quadratic; we return the Klein form, which a bi-invariant pseudo-metric on $\mathfrak{se}(3)$. Otherwise, $\mathfrak{se}(n)$ is indecomposable and neither of Type (1) nor of Type (2): it is not quadratic.

This reasoning is illustrated on Figure 8 through the tree representation of the algorithm.
We can build the whole quadratic space of $\mathfrak{se}(3)$. This gives the two-dimensional vector space:

$$\mathcal{Q}(\mathfrak{se}(3)) = \left\{ \begin{pmatrix} \alpha Z_{\text{Killing}} & \beta.\mathbb{I} \\ \beta.\mathbb{I} & 0 \end{pmatrix} \quad \middle| \quad \forall \alpha, \beta \in \mathbb{F} \right\} \tag{31}$$

Moreover, we have recognized in $\mathfrak{se}(3)$ the special case of a double extension $K \oplus K^*$ of $W = \{0\}$ by a compact Lie algebra $K = \mathfrak{so}(3)$. Therefore, the dual structure presented in Section 3 can be used in practice. We recall that we can represent the elements of $SO(3)$ as unit quaternions. Thus, we can represent the elements of $SE(3)$ as unit dual quaternions [52]. A generalization of the theory of Riemannian statistics to a theory of dual Riemannian statistics would thus be useful for rigid body transformations, which are present in many different fields.

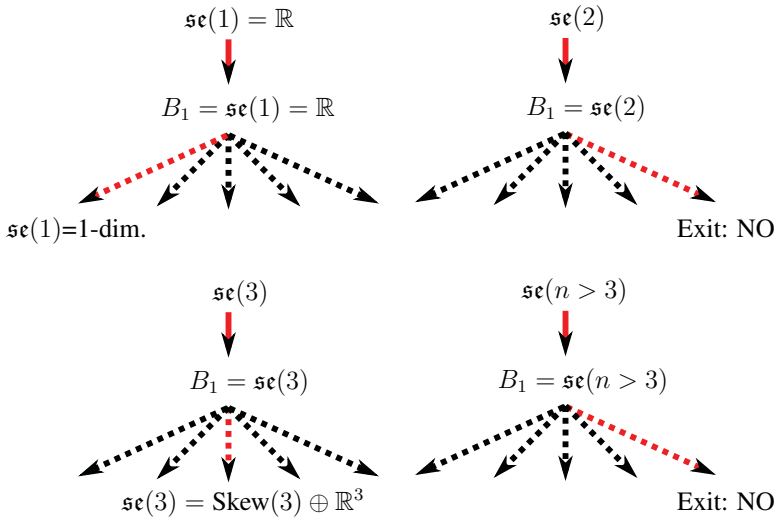



**Figure 8.** Schematical result for $SE(n)$. We recover the different cases depending on $n$.

## 5. Conclusions

In this paper, we have presented an algorithmic method to compute a bi-invariant pseudo-metric on a Lie group, in the case of existence. The method allows one to test simultaneously if the Lie group given as input is quadratic or not. We indicated how to compute all pf the bi-invariant pseudo-metrics on the given Lie group. First, the algorithm by itself represents a contribution to the field of computational Lie algebra.

Then, regarding statistics on Lie groups, which was our original motivation, we see two consequences of this article. First, it enables one to distinguish, from a practical point of view, Lie groups on which a future pseudo-Riemannian theory of statistics could be used and implemented. This is the case of $SE(3)$, the Lie group of rotations and translations of the 3D space, which is found in various fields.

Second, this paper shows that a general Lie group with bi-invariant mean does not admit a bi-invariant metric. Therefore, if one wants to define a general theory of statistics that works for all Lie groups, one needs to find a geometric framework beyond the Riemannian and the pseudo-Riemannian ones.

**Acknowledgments**

This work has been supported by an INRIA-CORDI (Contrat de recherche doctorale de l'INRIA) Fellowship. The authors would like to thank the reviewers for their comments, which considerably improved the manuscript.

**Author Contributions**

Xavier Pennec thought about the use of the pseudo-Riemannian framework for consistent statistics on Lie groups. In this context, Xavier Pennec suggested the preliminary study of the class of quadratic Lie groups while emphasizing the need of an efficient receipt to recognize them. Nina Miolane conducted the theoretical algebraic study on the characterization of quadratic Lie groups. Nina Miolane wrote the algorithm to recognize them and tested in on the Lie groups of interest. Both authors have read and approved the final manuscript.

**Conflicts of Interest**

The authors declare no conflict of interest.

**References**

1. Thompson, D.W.; Bonner, J.T. *On Growth and Form*; Cambridge University Press: Cambridge, UK, 1992.
2. Pennec, X.; Arsigny, V. Exponential Barycenters of the Canonical Cartan Connection and Invariant Means on Lie Groups. In *Matrix Information Geometry*; Springer: New York, NY, USA, 2012; pp. 123–168.
3. Fréchet, M. *L'intégrale abstraite d'une fonction abstraite d'une variable abstraite et son application a la moyenne d'un élément aléatoire de nature quelconque*; La Revue Scientifique: Paris, France, 1944.
4. Fréchet, M. Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l'institut Henri Poincaré* **1948**, *10*, 215–310.
5. Karcher, H. Riemannian center of mass and mollifier smoothing. *Commun. Pure Appl. Math.* **1977**, *30*, 509–541.
6. Kendall, W.S. Probability, Convexity, and Harmonic Maps with Small Image I: Uniqueness and Fine Existence. *Proc. Lond. Math. Soc.* **1990**, *s3-61*, 371–406.
7. Émery, M.; Mokobodzki, G. Sur le barycenter d'une probabilité dans une variété. *Séminaire de probabilités de Strasbourg* **1991**, *25*, 220–233.

8. Corcuera, J.M.; Kendall, W.S. Riemannian Barycentres and Geodesic Convexity. *Math. Proc. Camb. Philos. Soc.* **1998**, *127*, 253–269.
9. Huiling, L. Estimation of Riemannian Barycentres. *LMS J. Comput. Math.* **2004**, *7*, 193–200.
10. Yang, L. Riemannian median and its estimation. *LMS J. Comput. Math.* **2010**, *13*, 461–479.
11. Afsari, B. Riemannian $L^p$ center of mass: existence, uniqueness, and convexity. *Proc. Am. Math. Soc.* **2011**, *139*, 655–673.
12. Sternberg, S. *Lectures on Differential Geometry*; Prentice-Hall Mathematics Series; Prentice-Hall: Englewood Cliffs, NJ, USA, 1964;
13. Cartan, E. *Sur la Structure des Groupes de Transformations Finis et Continus*, 2nd ed.; Vuibert.: Paris, France, 1933; 157 S.
14. Tsou, S.T.; Walker, A.G. XIX. Metrisable Lie Groups and Algebras. *Proc. R. Soc. Edinb. Sect. A Math. Phys. Sci.* **1957**, *64*, 290–304.
15. Tsou, S.T. XI. On the Construction of Metrisable Lie Algebras. *Proc. R. Soc. Edinb. Sect. A Math. Phys. Sci.* **1962**, *66*, 116–127.
16. Astrakhantsev, V.V. Decomposability of metrizable Lie algebras. *Funct. Anal. Appl.* **1978**, *12*, 210–212.
17. Keith, V. On Invariant Bilinear Forms on Finite-dimensional Lie Algebras. Ph.D. Thesis, Tulane University, New Orleans, LA, USA, 1984.
18. Medina, A.; Revoy, P. Les groupes oscillateurs et leurs reseaux. *Manuscr. Math.* **1985**, *52*, 81–95.
19. Guts, A.K.; Levichev, A.V. On the Foundations of Relativity Theory. *Doklady Akademii Nauk SSSR* **1984**, *277*, 1299–1303.
20. Medina, A. Groupes de Lie munis de pseudo-métriques de Riemann bi-invariantes. *Sémin. géométrie différentielle 1981-1982, Montpellier 1982, Exp. No.6, 37 p. (1982).* **1982**.
21. Medina, A.; Revoy, P. Algèbres de Lie et produit scalaire invariant. *Annales scientifiques de l'École Normale Supérieure* **1985**, *18*, 553–561.
22. Hofmann, K.H.; Keith, V.S. Invariant quadratic forms on finite dimensional lie algebras. *Bull. Aust. Math. Soc.* **1986**, *33*, 21–36.
23. Bordemann, M. Nondegenerate invariant bilinear forms on nonassociative algebras. *Acta Mathematica Universitatis Comenianae. New Ser.* **1997**, *66*, 151–201.
24. Favre, G.; Santharoubane, L. Symmetric, invariant, non-degenerate bilinear form on a Lie algebra. *J. Algebra* **1987**, *105*, 451–464.
25. Campoamor-Stursberg, R. Quasi-Classical Lie Algebras and their Contractions. *Int. J. Theor. Phys.* **2008**, *47*, 583–598.
26. Benayadi, S.; Elduque, A. Classification of quadratic Lie algebras of low dimension. *J. Math. Phys.* **2014**, *55*, 081703.
27. Hilgert, J.; Hofmann, K. Lorentzian cones in real Lie algebras. *Monatsh. Math.* **1985**, *100*, 183–210.
28. Kath, I.; Olbrich, M. Metric Lie algebras with maximal isotropic centre. *Math. Z.* **2004**, *246*, 23–53.

29. Kath, I.; Olbrich, M. Metric Lie algebras and quadratic extensions. *Transform. Groups* **2006**, *11*, 87–131.
30. Duong, M.T. A New Invariant of Quadratic Lie Algebras and Quadratic Lie Superalgebras. Ph.D. Theses, Université de Bourgogne, Dijon, France, 2011.
31. Drinfeld, V.G. *Quantum Groups*; American Mathematics Society: Providence, RI, USA, 1987; pp. 798–820.
32. Belavin, A.; Drinfeld, V. Triangle Equations and Simple Lie Algebras. *Math. Phys. Rev.*, Harwood Academic: Newark, NJ, USA, 1998; *4*.
33. Delorme, P. Classification des triples de Manin pour les algebres de Lie reductives complexes: Avec un appendice de Guillaume Macey. *J. Algebra* **2001**, *246*, 97–174.
34. Rand, D. {PASCAL} programs for identification of Lie algebras: Part 1. Radical—A program to calculate the radical and nil radical of parameter-free and parameter-dependent lie algebras. *Comput. Phys. Commun.* **1986**, *41*, 105–125.
35. Rand, D.; Winternitz, P.; Zassenhaus, H. On the identification of a Lie algebra given by its structure constants. I. Direct decompositions, levi decompositions, and nilradicals. *Linear Algebra Appl.* **1988**, *109*, 197–246.
36. Cohen, A.M.; Graaf, W.A.D.; Rónyai, L. Computations in finite-dimensional Lie algebras. *Discret. Math. Theor. Comput. Sci.* **1997**, *1*, 129–138.
37. Ronyai, L.; Ivanyos, G.; Küronya, A.; de Graaf, W.A. Computing Levi Decompositions in Lie algebras. *Appl. Algebra Eng. Commun. Comput.* **1997**, *8*, 291–303.
38. De Graaf, W. *Lie Algebras: Theory and Algorithms*; North-Holland Mathematical Library; Elsevier: Amsterdam, The Netherlands, 2000.
39. Postnikov, M. *Geometry VI: Riemannian Geometry*; Encyclopaedia of Mathematical Sciences; Springer: New York, NY, USA, 2001.
40. Bourbaki, N. *Lie Groups and Lie Algebras*; Springer: Paris, France, 1989; Chapters 1–3.
41. Milnor, J. Curvatures of left invariant metrics on lie groups. *Adv. Math.* **1976**, *21*, 293–329.
42. Bartels, R.H.; Stewart, G.W. Solution of the Matrix Equation AX + XB = C [F4]. *Commun. ACM* **1972**, *15*, 820–826.
43. Kitagawa, G. An algorithm for solving the matrix equation X = FXF T + S. *Int. J. Control* **1977**, *25*, 745–753.
44. Grünwald, J. Über duale Zahlen und ihre Anwendung in der Geometrie. *Monatsh. Math. Phys.* **1906**, *17*, 81–136.
45. Levi, E. *Sulla struttura dei gruppi finiti e continui*; Atti della Reale Accademia delle Scienze di Torino: Turin, Italy, 1905; Volume 40, pp. 551–565.
46. Ceballos, M.; Núñez, J.; Tenorio, A.F. Algorithmic Method to Obtain Abelian Subalgebras and Ideals in Lie Algebras. *Int. J. Comput. Math.* **2012**, *89*, 1388–1411.
47. Motsak, O. Computation of the Central Elements and Centralizers of Sets of Elements in Non-Commutative Polynomial Algebras. Ph.D. Thesis, Technische Universität Kaiserslautern, Kaiserslautern, Germany, 2006.

48. Schur, I. *Neue Begründung der Theorie der Gruppencharaktere*; Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin: Berlin, Germany, 1905; pp. 406–432.
49. Frobenius, G. *Ueber lineare Substitutionen und bilineare Formen. J. Reine Angew. Math.* **1877**, *87*, 1–63.
50. Brooksbank, P.A.; Luks, E.M. Testing isomorphism of modules. *J. Algebra* **2008**, *320*, 4020–4029.
51. Karger, A.; Josef, N. *Space Kinematics and Lie Groups*; Gordon and Breach Science Publishers: New York, NY, USA, 1985. Translation of: Prostorová kinematika a Liehovy grupy.
52. Kenwright, B. A Beginners Guide to Dual-Quaternions: What They Are, How They Work, and How to Use Them for 3D Character Hierarchies. In Proceedings of the The 20th International Conference on Computer Graphics, Visualization and Computer Vision, Plzen, Czech, 25–28 June 2012; pp. 1–13.

# Koszul Information Geometry and Souriau Geometric Temperature/Capacity of Lie Group Thermodynamics

**Frédéric Barbaresco**

**Abstract:** The François Massieu 1869 idea to derive some mechanical and thermal properties of physical systems from "Characteristic Functions", was developed by Gibbs and Duhem in thermodynamics with the concept of potentials, and introduced by Poincaré in probability. This paper deals with generalization of this Characteristic Function concept by Jean-Louis Koszul in Mathematics and by Jean-Marie Souriau in Statistical Physics. The Koszul-Vinberg Characteristic Function (KVCF) on convex cones will be presented as cornerstone of "Information Geometry" theory, defining Koszul Entropy as Legendre transform of minus the logarithm of KVCF, and Fisher Information Metrics as hessian of these dual functions, invariant by their automorphisms. In parallel, Souriau has extended the Characteristic Function in Statistical Physics looking for other kinds of invariances through co-adjoint action of a group on its momentum space, defining physical observables like energy, heat and momentum as pure geometrical objects. In covariant Souriau model, Gibbs equilibriums states are indexed by a geometric parameter, the Geometric (Planck) Temperature, with values in the Lie algebra of the dynamical Galileo/Poincaré groups, interpreted as a space-time vector, giving to the metric tensor a null Lie derivative. Fisher Information metric appears as the opposite of the derivative of Mean "Moment map" by geometric temperature, equivalent to a Geometric Capacity or Specific Heat. We will synthetize the analogies between both Koszul and Souriau models, and will reduce their definitions to the exclusive Cartan "Inner Product". Interpreting Legendre transform as Fourier transform in (*Min*,+) algebra, we conclude with a definition of Entropy given by a relation mixing Fourier/Laplace transforms: *Entropy = (minus) Fourier*$_{(Min,+)}$ *o Log o Laplace*$_{(+,X)}$.

Reprinted from *Entropy*. Cite as: Barbaresco, F. Koszul Information Geometry and Souriau Geometric Temperature/Capacity of Lie Group Thermodynamics. *Entropy* **2014**, *16*, 4521ï 4565.

## 1. Introduction

The Koszul-Vinberg Characteristic Function (KVCF) is a dense knot in important mathematical fields such as Hessian Geometry, Kählerian Geometry and Affine Differential Geometry. As essence of Information Geometry, this paper develops KVCF as a transverse concept in Thermodynamics, in Statistical Physics and in Probability. From general KVCF definition, the paper introduces Koszul Entropy as the Legendre transform of minus the logarithm of KVCF, and compares both functions by analogy with the Dual Massieu-Duhem potentials in thermodynamics. This paper will also explore close inter-relations between these domains through geometric tools developed by Jean-Louis Koszul and Jean-Marie Souriau. The cornerstone of "Information Geometry" Theory will appear to be based on the fundamental property that derivatives of the

Koszul-Vinberg Characteristic Function Logarithm (KVCFL) $\log \psi_{\Omega}(x) = \log \int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi$ , defined on convex dual cone of $\Omega$ , are invariant by the automorphisms of $\Omega$ , and that its hessian defines a non-arbitrary Riemannian metric.

In thermodynamics, François Massieu [1–3] was the first to introduce the concept of characteristic function $\phi$ . This characteristic function or thermodynamic potential is able to provide all the body properties from their derivatives. In thermodynamics, Entropy *S* is one of the Massieu-Duhem potentials [4–8], derived from the Legendre-Moreau transform of the characteristic function logarithm $\phi$ : $S = \phi - \beta . \frac{\partial \phi}{\partial \beta}$ with $\beta = \frac{1}{kT}$ being the thermodynamic temperature. The most popular notion of "*characteristic function*" was introduced in a second step by Henri Poincaré in his lecture on probability [9,10], using the property that all moments of statistical laws could be deduced from its derivatives. Paul Levy then made systematic use of this concept in his 1925 book. We assume that Poincaré was influenced by his school fellow at Ecole des Mines de Paris, François Massieu, and his work on thermodynamic potentials (generalized by Pierre Duhem in an Energetic Theory). This assertion is corroborated by the observation that Poincaré added in his lecture on thermodynamics in the 2nd edition [9,10] in 1892, a chapter on the "Massieu characteristic function" with many developments and applications, before developing the concept in Probability [9,10], see Figure 1.

In Thermodynamics, Statistical Physics and Probability, we can observe that the "*characteristic function*" and its derivatives capture all information of system or physical model and random variable. Furthermore, the general notion of Entropy could be naturally defined by the Legendre Transform of minus the Koszul characteristic function logarithm. In the general case, Legendre transform of minus the logarithm of the KVCF will be designated in the following as "*Koszul Entropy*".

**M. Massieu a montré que, si l'on fait choix pour variables indépendantes de $v$ et de T ou de $p$ et de T, il existe une fonction, d'ailleurs inconnue, de laquelle les trois fonctions des variables, $p$, U et S dans le premier cas, $v$, U et S dans le second, peuvent se déduire facilement. M. Massieu a donné à cette fonction, dont la forme dépend du choix des variables, le nom de *fonction caractéristique*.**

[M. Massieu showed that, if we make choice for independent variables of v and T or of p and T, there is a function, moreover unknown, of which three functions of variables, p, U and S in the first case, v, U and S in the second, can be deducted easily. M. Massieu gave to this function, the form of which depends on the choice of variables, name of characteristic function.]

**Puisque des fonctions de M. Massieu on peut déduire les autres fonctions des variables, toutes les équations de la Thermodynamique pourront s'écrire de manière à ne plus renfermer que ces fonctions et leurs dérivées; il en résultera donc, dans certains cas, une grande simplification. Nous verrons bientôt une application importante de ces fonctions.**

[Because functions of M. Massieu, we can deduct the other functions of variables, all the equations of the Thermodynamics can be written not so as to contain more than these functions and their derivatives; it will thus result from it, in certain cases, a large simplification. We shall see soon an important application of these functions.]

**Figure 1.** Text of Poincaré Lecture on Thermodynamic with development of the concept of "Massieu Characteristic Function".

This general notion of "*characteristic function*" has been generalized by the French physicist Jean-Marie Souriau. In 1970, Souriau, that had followed the Elie Cartan Lecture at ENS Ulm in 1946 (one year after his aggregation), introduced the concept of co-adjoint action of a group on its momentum space (or "*moment map*": mapping induced by symplectic manifold symmetries), based on the orbit method works, that allows to define physical observables like energy, heat and momentum as pure geometrical objects (the moment map takes its values in a space attached to the group of symmetries in the dual space of its Lie algebra). The moment map is a constant of the motion and is associated to symplectic cohomology (assignment of algebraic invariants to a topological space that arises from the algebraic dualization of the homology construction). For Souriau, equilibrium states are indexed by a geometric parameter $\beta$ with values in the Lie algebra of the dynamical group (Galileo or Poincaré group). The Souriau approach generalizes the Gibbs equilibrium states, $\beta$ playing the role of temperature. The invariance with respect to the group, and the fact that the entropy $S$ is a convex function of $\beta$, imposes very strict conditions, that allow Souriau to interpret $\beta$ as a space-time vector (the vector-valued temperature of Planck), giving to the metric tensor a null Lie derivative. For Souriau, all the details of classical mechanics appear as geometric necessities (e.g., mass is the measure of the symplectic cohomology of the action of a Galileo group). We will synthetize the analogies between the Koszul and Souriau models in tables (the Information Geometry case being a particular case of Koszul Hessian geometry).

The Koszul-Vinberg characteristic function is a dense knot in mathematics and could be introduced in the framework of different geometries: Hessian Geometry (Jean-Louis Koszul's work), Homogeneous convex cones geometry (Ernest Vinberg's work [11]), Homogeneous Symmetric Bounded Domains Geometry [12,13] (Elie Cartan [14] and Carl Ludwig Siegel's works [15,16]), Symplectic Geometry [17,18] (Thomas von Friedrich [19] & Jean-Marie Souriau's work), Affine Geometry (Takeshi Sasaki and Eugenio Calabi's works) and Information Geometry (Calyampudi Rao and Nikolai Chentsov' works). Through Legendre duality, Contact Geometry (Vladimir Arnold's work) is considered as the odd-dimensional twin of symplectic geometry and could be used to understand Legendre mapping in Information Geometry. Fisher metrics of Information Geometry could be introduced as hessian metrics from minus Koszul-Vinberg characteristic function logarithm or from Koszul Entropy (Legendre transform of minus Koszul-Vinberg characteristic function logarithm). In a more general context, we can consider Information Geometry in the framework of "*Geometric Science of Information*", a new "corpus" that also covers probability in metric space (Maurice Fréchet's work), probability/geometry on structures (Yann Ollivier and Misha Gromov's works [20–23]) and probability on Riemannian manifold (Michel Emery and Marc Arnaudon's works). This link between "*Information Theory*" and "Geometry" is also deeply developed and influenced by fundamental works of Yann Ollivier [24,25] (initially described in his HDR report "*Randomness and Curvature*" in 2009 and more recent papers on IGO flow).

## 2. Legendre Duality and Projective Duality

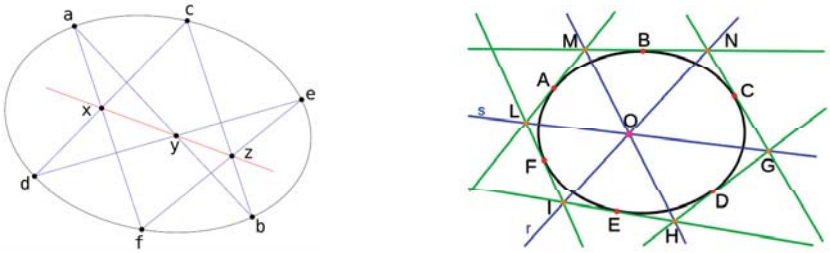
In following chapters, we will see that the minus Logarithm of the Characteristic Function and Entropy will be related by the Legendre transform, that can be considered in the context of projective duality. Duality is an old and very fruitful idea in mathematics that has been constantly generalized [26–38]. A duality translates concepts, theorems or mathematical structures into other concepts, theorems or structures, in a one-to-one fashion, often by means of an involution operation and sometimes with fixed points.

The simplest duality is linear duality in the plane with points and lines (two different points can be joined by a unique line. Two different lines meet in one point unless they are parallel). By adding some points at infinity (to avoid particular case of parallel lines) then we obtain the projective plane in which the duality is given symmetrical relationship between points and lines, and led to the classical principle of projective duality, where the dual theorem is also a theorem.

Most Famous example is given by *Pascal's theorem* (the Hexagrammum Mysticum Theorem) stating that:

- If the vertices of a simple hexagon are points of a point conic, then its diagonal points are collinear: *If an arbitrary six points are chosen on a conic (i.e., ellipse, parabola or hyperbola) and joined by line segments in any order to form a hexagon, then the three pairs of opposite sides of the hexagon (extended if necessary) meet in three points which lie on a straight line, called the Pascal line of the hexagon.*

The dual of Pascal's Theorem is known as *Brianchon's Theorem*, as illustrated in Figure 2:




(**a**) (**b**)

**Figure 2.** (**a**) Pascal's theorem, (**b**) Brianchon's theorem.

- If the sides of a simple hexagon are lines of a line conic, then the diagonal lines are concurrent.

The Legendre(-Moreau) transform [39,40] is an operation from convex functions on a vector space to functions on the dual space. The Legendre transform is related to projective duality and tangential coordinates in algebraic geometry, and to the construction of dual Banach spaces in analysis. Classical Legendre transform in Euclidean space is given by fixing a scalar product $\langle .,. \rangle$ on $R^n$. For a function $F : R^n \to R \cup \{\pm\infty\}$, let:

$$G(y) = LF(y) = \underset{x}{Sup}\left\{\langle y, x\rangle - F(x)\right\} \tag{1}$$

The Legendre transform is illustrated in Figure 3.

This is an involution on the class of convex lower semi-continuous functions on $R^n$. There are two dual possibilities to describe a function. We can either use a function, or we may regard the curve as the envelope of its tangent planes. We give in Appendix A1 the historical context of Legendre Transform introduction on a Minimal Surface problem considered initially by Gaspard Monge.

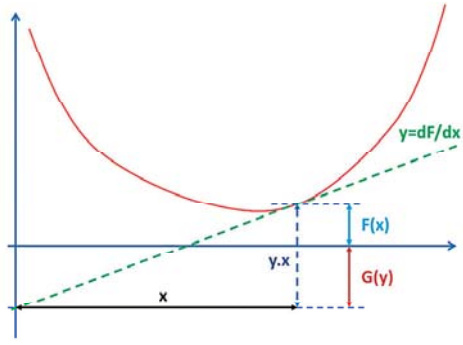



**Figure 3.** Legendre Transform *G(y)* of *F(x)*.

The Legendre Transform is very important in Information Geometry [39], which uses mutually dual (conjugate) affine connections, dual potentials in dual coordinates systems and dual metrics that are studied in the framework of Hessian or affine differential geometry.

To illustrate the role of Legendre transform in Information Geometry, we provide a canonical example, with the relations for the Multivariate Normal Gaussian Law $N(m,R)$:

- Dual Coordinates systems:

$$\begin{cases} \tilde{\Theta} = (\theta, \Theta) = \left(R^{-1}m, (2R)^{-1}\right) \\ \tilde{H} = (\eta, H) = \left(m, -R + mm^T\right) \end{cases} \tag{2}$$

- Dual potential functions:

$$\begin{cases} \tilde{\Psi}\left(\tilde{\Theta}\right) = 2^{-1}Tr\left(\Theta^{-1}\theta\theta^T\right) - 2^{-1}\log\left(\det\Theta\right) + 2^{-1}n\log(2\pi e) \\ \tilde{\Phi}\left(\tilde{H}\right) = -2^{-1}\log\left(1 + \eta^T H^{-1}\eta\right) - 2^{-1}\log\left(\det(-H)\right) - 2^{-1}n\log\left(2\pi e\right) \end{cases} \tag{3}$$

related by Legendre transform:

$$\tilde{\Phi}\left(\tilde{H}\right) = \left\langle\tilde{\Theta}, \tilde{H}\right\rangle - \tilde{\Psi}\left(\tilde{\Theta}\right) \text{ with } \left\langle\tilde{\Theta}, \tilde{H}\right\rangle = Tr\left(\theta\eta^T + \Theta H^T\right) \tag{4}$$

where dual coordinate systems are given by derivatives of dual potential functions:

$$\begin{cases} \dfrac{\partial \tilde{\Psi}}{\partial \theta} = \eta \\ \dfrac{\partial \tilde{\Psi}}{\partial \Theta} = H \end{cases} \text{ and } \begin{cases} \dfrac{\partial \tilde{\Phi}}{\partial \eta} = \theta \\ \dfrac{\partial \tilde{\Phi}}{\partial H} = \Theta \end{cases} \tag{5}$$

with $\tilde{\Phi}(\tilde{H}) = E[\log p]$ being the Entropy.

In the theory of Information Geometry introduced by Rao and Chentsov, a Riemannian manifold is then defined by a metric tensor given by hessian of these dual potential functions:

$$g_{ij} = \frac{\partial^2 \tilde{\Psi}}{\partial \tilde{\Theta}_i \partial \tilde{\Theta}_j} \text{ and } g^*_{ij} = \frac{\partial^2 \tilde{\Phi}}{\partial \tilde{H}_i \partial \tilde{H}_j} \tag{6}$$

In this paper, we will develop the concept of "*Hessian Manifolds*" theory that was initially studied by Koszul in a more general framework. In the next section, we will expose the theory of the Koszul-Vinberg characteristic function on convex sharp cones that will be presented as a general framework of Information Geometry.

## 3. Koszul Characteristic Function/Entropy by Legendre Duality

We define the Koszul-Vinberg Hessian metric on a convex sharp cone, and observe that the Fisher information metric of Information Geometry coincides with the canonical Koszul Hessian metric (given by Koszul forms) [41–47]. We also observe, by Legendre duality (Legendre transform of minus Koszul characteristic function logarithm), that we are able to introduce a *Koszul Entropy*, that plays the role of the generalized Shannon Entropy.

### *3.1. Koszul-Vinberg Characteristic Function and Metric for Convex Sharp Cone*

Jean-Louis Koszul [41,42,47] and Ernest B. Vinberg [48,49] have introduced an affinely invariant hessian metric on a sharp convex cone $\Omega^*$ through its characteristic function $\psi$. In the following, $\Omega^*$ is a sharp open convex cone in a vector space $E$ of finite dimension on $R$ (a convex cone is sharp if it does not contain any full straight line). In dual space $E^*$ of $E$, $\Omega^*$ is the set of linear strictly positive forms on $\bar{\Omega}-\{0\}$. $\Omega^*$ is the dual cone of $\Omega$ and is a sharp open convex cone. If $\xi \in \Omega^*$, then the intersection $\Omega \cap \{x \in E / \langle x, \xi \rangle = 1\}$ is bounded. $G = Aut(\Omega)$ is the group of linear transform of $E$ that preserves $\Omega$. $G = Aut(\Omega)$ operates on $\Omega^*$ by $\forall g \in G = Aut(\Omega), \forall \xi \in E^*$ then $\tilde{g}.\xi = \xi \circ g^{-1}$.

*Koszul-Vinberg Characteristic Function Definition:*

*Let* $d\xi$ be the Lebesgue measure on $E^*$, the following integral:

$$\psi_\Omega(x) = \int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi \quad \forall x \in \Omega \tag{7}$$

with $\Omega^*$ the dual cone is an analytic function on $\Omega$, with $\psi_\Omega(x) \in ]0, +\infty[$, called the *Koszul-Vinberg characteristic function* of cone $\Omega$.

The Koszul-Vinberg Characteristic Function has the following properties:

- The Bergman kernel of $\Omega + iR^{n+1}$ is written as $K_\Omega(\mathrm{Re}(z))$ up to a constant where $K_\Omega$ is defined by the integral:

$$K_\Omega(x) = \int_{\Omega^*} e^{-\langle \xi, x\rangle} \psi_{\Omega^*}(\xi)^{-1} d\xi \tag{8}$$

- $\psi_\Omega$ is analytic function defined on the interior of $\Omega$ and $\psi_\Omega(x) \to +\infty$ as $x \to \partial\Omega$

If $g \in Aut(\Omega)$ then $\psi_\Omega(gx) = |\det g|^{-1} \psi_\Omega(x)$ and since $tI \in G = Aut(\Omega)$ for any $t > 0$, we have

$$\psi_\Omega(tx) = \psi_\Omega(x)/t^n \tag{9}$$

- $\psi_\Omega$ is logarithmically strictly convex, and $\phi_\Omega(x) = \log(\psi_\Omega(x))$ is strictly convex.

From the KVCF, could be introduced two forms defined by Koszul:

*Koszul 1-form* $\alpha$ : The differential 1-form

$$\alpha = d\phi_\Omega = d\log\psi_\Omega = d\psi_\Omega / \psi_\Omega \tag{10}$$

is invariant by all automorphisms $G = Aut(\Omega)$ of $\Omega$. If and $u \in E$ then

$$\langle \alpha_x, u\rangle = -\int_{\Omega^*} \langle \xi, u\rangle . e^{-\langle \xi, x\rangle} d\xi \text{ and } \alpha_x \in -\Omega^* \tag{11}$$

and:

*Koszul 2-form* $\beta$*:* The symmetric differential 2-form:

$$\beta = D\alpha = d^2 \log \psi_\Omega \tag{12}$$

is a positive definite symmetric bilinear form on $E$ invariant under $G = Aut(\Omega)$. $D\alpha > 0$

This positivity is given by Schwarz inequality and:

$$d^2 \log\psi_\Omega(u,v) = \int_{\Omega^*} \langle \xi, u\rangle \langle \xi, v\rangle e^{-\langle \xi, u\rangle} d\xi \tag{13}$$

We can then introduce the Koszul metric based on previous definitions:

*Koszul Metric:* $D\alpha$ defines a Riemanian structure invariant by $Aut(\Omega)$, and then the Riemanian metric is given by $g = d^2 \log\psi_\Omega$

$$\left(d^2 \log\psi(x)\right)(u) = \frac{1}{\psi(u)^2}\left[\int_{\Omega^*} F(\xi)^2 d\xi . \int_{\Omega^*} G(\xi)^2 d\xi - \left(\int_{\Omega^*} F(\xi).G(\xi) d\xi\right)^2\right] > 0 \tag{14}$$

with $F(\xi) = e^{-\frac{1}{2}\langle x, \xi\rangle}$ and $G(\xi) = e^{-\frac{1}{2}\langle x, \xi\rangle}\langle u, \xi\rangle$

This result is obtained using Schwarz inequality, $d\log\psi = \frac{d\psi}{\psi}$ and $d^2\log\psi = \frac{d^2\psi}{\psi} - \left(\frac{d\psi}{\psi}\right)^2$ where $(d\psi(x))(u) = -\int_{\Omega^*} e^{-\langle x,\xi\rangle}\langle u,\xi\rangle d\xi$ and $(d^2\psi(x))(u) = -\int_{\Omega^*} e^{-\langle x,\xi\rangle}\langle u,\xi\rangle^2 d\xi$

A diffeomorphism is used to define dual coordinate:

$$x^* = -\alpha_x = -d\log\psi_\Omega(x) \tag{15}$$

with $\langle df(x),u\rangle = D_u f(x) = \left.\frac{d}{dt}\right|_{t=0} f(x+tu)$ . When the cone $\Omega$ is symmetric, the map $x \mapsto x^* = -\alpha_x$ is a bijection and an isometry with one unique fixed point (the manifold is a Riemannian Symmetric Space given by this isometry):

$$(x^*)^* = x \ , \ \langle x, x^*\rangle = n \text{ and } \psi_\Omega(x)\psi_{\Omega^*}(x^*) = cste \tag{16}$$

$x^*$ is characterized by $x^* = \arg\min\{\psi(y) / y \in \Omega^*, \langle x,y\rangle = n\}$ and $x^*$ *is the center of gravity of the cross section* $\{y \in \Omega^*, \langle x,y\rangle = n\}$ of $\Omega^*$:

$$x^* = \int_{\Omega^*} \xi . e^{-\langle \xi,x\rangle} d\xi \Big/ \int_{\Omega^*} e^{-\langle \xi,x\rangle} d\xi \text{ and } \langle -x^*, h\rangle = d_h \log\psi_\Omega(x) = -\int_{\Omega^*} \langle \xi, h\rangle e^{-\langle \xi,x\rangle} d\xi \Big/ \int_{\Omega^*} e^{-\langle \xi,x\rangle} d\xi \tag{17}$$

If we set $\Phi(x) = -\log\psi_\Omega(x)$ , Misha Gromov [20,21] has observed that $x^* = d\Phi(x)$ is an injection where the closure of the image equals the convex hull of the support and the volume of this hull is the the n-dimensional volume defined by the integral of the determinant of the hessian of this function $\Phi(x)$ , where the map $\Phi \mapsto M(\Phi) = \int_\Omega \det(Hess(\Phi(x)))dx$ obeys non-trivial convexity relation given by the Brunn-Minkowsky inequality $[M(\Phi_1 + \Phi_2)]^{1/n} \geq [M(\Phi_1)]^{1/n} + [M(\Phi_2)]^{1/n}$.

*3.2. Koszul Entropy and Its Barycenter*

From this last equation, we can deduce the "*Koszul Entropy*" defined as the Legendre Transform of $\Phi(x)$ minus logarithm of Koszul-Vinberg characteristic function:

$$\begin{gathered}\Phi^*(x^*) = \langle x, x^*\rangle - \Phi(x) \text{ with } x^* = D_x\Phi \text{ and } x = D_{x^*}\Phi^* \\ \text{where } \Phi(x) = -\log\psi_\Omega(x)\end{gathered} \tag{18}$$

$$\Phi^*(x^*) = \left\langle (D_x\Phi)^{-1}(x^*), x^*\right\rangle - \Phi\left[(D_x\Phi)^{-1}(x^*)\right] \quad \forall x^* \in \{D_x\Phi(x) / x \in \Omega\} \tag{19}$$

By the definition of the Koszul-Vinberg Characteristic function, and by using $-\langle \xi,x\rangle = \log e^{-\langle \xi,x\rangle}$, we can write:

$$-\langle x^*, x\rangle = \int_{\Omega^*} \log e^{-\langle \xi,x\rangle} . e^{-\langle \xi,x\rangle} d\xi \Big/ \int_{\Omega^*} e^{-\langle \xi,x\rangle} d\xi \tag{20}$$

and:

$$\Phi^*(x^*) = \langle x, x^* \rangle - \Phi(x) = -\int_{\Omega^*} \log e^{-\langle \xi, x \rangle}.e^{-\langle \xi, x \rangle} d\xi \Big/ \int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi + \log \int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi$$

$$\Phi^*(x^*) = \left[ \left( \int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi \right) . \log \int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi - \int_{\Omega^*} \log e^{-\langle \xi, x \rangle}.e^{-\langle \xi, x \rangle} d\xi \right] \Big/ \int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi$$

$$\Phi^*(x^*) = \left[ \log \int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi - \int_{\Omega^*} \log e^{-\langle \xi, x \rangle} . \frac{e^{-\langle \xi, x \rangle}}{\int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi} d\xi \right]$$

$$\Phi^*(x^*) = \left[ \log \int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi . \left( \int_{\Omega^*} \frac{e^{-\langle \xi, x \rangle}}{\int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi} d\xi \right) - \int_{\Omega^*} \log e^{-\langle \xi, x \rangle} . \frac{e^{-\langle \xi, x \rangle}}{\int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi} d\xi \right] \text{ with } \int_{\Omega^*} \frac{e^{-\langle \xi, x \rangle}}{\int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi} d\xi$$

$$\Phi^*(x^*) = \left[ -\int_{\Omega^*} \frac{e^{-\langle \xi, x \rangle}}{\int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi} . \log \left( \frac{e^{-\langle \xi, x \rangle}}{\int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi} \right) d\xi \right] \tag{21}$$

In this last equation, $p_x(\xi) = e^{-\langle \xi, x \rangle} / \int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi$ appears as a density, and the Legendre transform $\Phi^*()$ looks like the classical Shannon Entropy, named in the following *Koszul Entropy*:

$$\Phi^* = -\int_{\Omega^*} p_x(\xi) \log p_x(\xi) d\xi \tag{22}$$

with:

$$p_x(\xi) = e^{-\langle \xi, x \rangle} / \int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi = e^{-\langle x, \xi \rangle - \log \int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi} = e^{-\langle x, \xi \rangle + \Phi(x)} \text{ and } x^* = \int_{\Omega^*} \xi . p_x(\xi) d\xi \tag{23}$$

We will call $p_x(\xi) = \frac{e^{-\langle \xi, x \rangle}}{\int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi}$ the Koszul Density, with the property that:

$$\log p_x(\xi) = -\langle x, \xi \rangle - \log \int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi = -\langle x, \xi \rangle + \Phi(x) \tag{24}$$

and:

$$E_\xi[-\log p_x(\xi)] = \langle x, x^* \rangle - \Phi(x) \tag{25}$$

We can observe that:

$$\Phi(x) = -\log \int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi = -\log \int_{\Omega^*} e^{-[\Phi^*(\xi) + \Phi(x)]} d\xi = \Phi(x) - \log \int_{\Omega^*} e^{-\Phi^*(\xi)} d\xi$$
$$\Rightarrow \int_{\Omega^*} e^{-\Phi^*(\xi)} d\xi = 1 \tag{26}$$

But the development is not achieved and we have to make appear $x^*$ in $\Phi^*(x^*)$. For this objective, we have to write:

$$\begin{aligned}&\log p_x(\xi)=\log e^{-\langle x,\xi\rangle+\Phi(x)}=\log e^{-\Phi^*(\xi)}=-\Phi^*(\xi)\\&\Rightarrow\Phi^*=-\int_{\Omega^*}p_x(\xi)\log p_x(\xi)d\xi=\int_{\Omega^*}\Phi^*(\xi)p_x(\xi)d\xi=\Phi^*(x^*)\end{aligned}\tag{27}$$

The last equality is true if and only if we have the following relation:

$$\int_{\Omega^*}\Phi^*(\xi)p_x(\xi)d\xi=\Phi^*\left(\int_{\Omega^*}\xi.p_x(\xi)d\xi\right)\text{ as }x^*=\int_{\Omega^*}\xi.p_x(\xi)d\xi\tag{28}$$

This condition could be written more synthetically [50,51]:

$$E\left[\Phi^*(\xi)\right]=\Phi^*\left(E[\xi]\right),\ \xi\in\Omega^*\tag{29}$$

The meaning of this relation is that "*the Barycenter of Koszul Entropy is the Koszul Entropy of Barycenter*".

This condition is achieved for $x^*=D_x\Phi$ taking into account Legendre Transform property:

$$\begin{aligned}&\text{Legendre Transform}:\ \Phi^*(x^*)=\underset{x}{Sup}\left[\left\langle x,x^*\right\rangle-\Phi(x)\right]\\&\Rightarrow\begin{cases}\Phi^*(x^*)\geq\left\langle x,x^*\right\rangle-\Phi(x)\\\Phi^*(x^*)\geq\int_{\Omega^*}\Phi^*(\xi)p_x(\xi)d\xi\end{cases}\Rightarrow\begin{cases}\Phi^*(x^*)\geq E\left[\Phi^*(\xi)\right]\\\text{equality for }x^*=\dfrac{d\Phi}{dx}\end{cases}\end{aligned}\tag{30}$$

*3.3. Relation of Koszul Density with the Maximum Entropy Principle*

We will observe in this section that Koszul density is a solution of the Maximum Entropy. Classically, the density given by the Maximum Entropy Principle [52–58] is given by:

$$\underset{p_x(.)}{Max}\left[-\int_{\Omega^*}p_x(\xi)\log p_x(\xi)d\xi\right]\ \text{ such }\ \begin{cases}\int_{\Omega^*}p_x(\xi)d\xi=1\\\int_{\Omega^*}\xi.p_x(\xi)d\xi=x^*\end{cases}\tag{31}$$

If we take $q_x(\xi)=e^{-\langle\xi,x\rangle}/\int_{\Omega^*}e^{-\langle\xi,x\rangle}d\xi=e^{-\langle x,\xi\rangle-\log\int_{\Omega^*}e^{-\langle\xi,x\rangle}d\xi}$ such that:

$$\begin{cases}\int_{\Omega^*}q_x(\xi).d\xi=\int_{\Omega^*}e^{-\langle\xi,x\rangle}d\xi/\int_{\Omega^*}e^{-\langle\xi,x\rangle}d\xi=1\\\log q_x(\xi)=\log e^{-\langle x,\xi\rangle-\log\int_{\Omega^*}e^{-\langle\xi,x\rangle}d\xi}=-\left\langle x,\xi\right\rangle-\log\int_{\Omega^*}e^{-\langle x,\xi\rangle}d\xi\end{cases}\tag{32}$$

Then by using the fact that $\log x\geq\left(1-x^{-1}\right)$ with equality if and only if $x=1$, we find the following:

$$-\int_{\Omega^*} p_x(\xi)\log\frac{p_x(\xi)}{q_x(\xi)}d\xi \le -\int_{\Omega^*} p_x(\xi)\left(1-\frac{q_x(\xi)}{p_x(\xi)}\right)d\xi \tag{33}$$

We can then observe that:

$$\int_{\Omega^*} p_x(\xi)\left(1-\frac{q_x(\xi)}{p_x(\xi)}\right)d\xi = \int_{\Omega^*} p_x(\xi)d\xi - \int_{\Omega^*} q_x(\xi)d\xi = 0 \tag{34}$$

because $\int_{\Omega^*} p_x(\xi)d\xi = \int_{\Omega^*} q_x(\xi)d\xi = 1$

We can then deduce that:

$$-\int_{\Omega^*} p_x(\xi)\log\frac{p_x(\xi)}{q_x(\xi)}d\xi \le 0 \Rightarrow -\int_{\Omega^*} p_x(\xi)\log p_x(\xi)d\xi \le -\int_{\Omega^*} p_x(\xi)\log q_x(\xi)d\xi \tag{35}$$

If we develop the last inequality, using expression of $q_x(\xi)$:

$$-\int_{\Omega^*} p_x(\xi)\log p_x(\xi)d\xi \le -\int_{\Omega^*} p_x(\xi)\left[-\langle x,\xi\rangle - \log\int_{\Omega^*} e^{-\langle x,\xi\rangle}d\xi\right]d\xi \tag{36}$$

$$-\int_{\Omega^*} p_x(\xi)\log p_x(\xi)d\xi \le \left\langle x, \int_{\Omega^*} \xi.p_x(\xi)d\xi\right\rangle + \log\int_{\Omega^*} e^{-\langle x,\xi\rangle}d\xi \tag{37}$$

If we take $x^* = \int_{\Omega^*} \xi.p_x(\xi)d\xi$ and $\Phi(x) = -\log\int_{\Omega^*} e^{-\langle x,\xi\rangle}d\xi$, then we deduce that the Koszul density $q_x(\xi) = e^{-\langle \xi,x\rangle} / \int_{\Omega^*} e^{-\langle \xi,x\rangle}d\xi = e^{-\langle x,\xi\rangle - \log\int_{\Omega^*} e^{-\langle \xi,x\rangle}d\xi}$ is the Maximum Entropy solution constrained by $\int_{\Omega^*} p_x(\xi)d\xi = 1$ and $\int_{\Omega^*} \xi.p_x(\xi)d\xi = x^*$:

$$-\int_{\Omega^*} p_x(\xi)\log p_x(\xi)d\xi \le \langle x,x^*\rangle - \Phi(x) \tag{38}$$

$$-\int_{\Omega^*} p_x(\xi)\log p_x(\xi)d\xi \le \Phi^*(x^*) \tag{39}$$

We have then observed that Koszul Entropy provides density of Maximum Entropy:

$$p_{\overline{\xi}}(\xi) = \frac{e^{-\langle \xi,\Theta^{-1}(\overline{\xi})\rangle}}{\int_{\Omega^*} e^{-\langle \xi,\Theta^{-1}(\overline{\xi})\rangle}d\xi} \text{ with } x = \Theta^{-1}(\overline{\xi}) \text{ and } \overline{\xi} = \Theta(x) = \frac{d\Phi(x)}{dx} \tag{40}$$

where:

$$\overline{\xi} = \int_{\Omega^*} \xi.p_{\overline{\xi}}(\xi)d\xi \text{ and } \Phi(x) = -\log\int_{\Omega^*} e^{-\langle x,\xi\rangle}d\xi \tag{41}$$

We can then deduce the Maximum Entropy solution without solving the classical variational problem with Lagrangian hyperparameters, but only by inversing function $\overline{\xi} = \Theta(x) = \dfrac{d\Phi(x)}{dx}$. This remark was made by Jean-Souriau in the paper [59]. If we take vector with tensor components $\xi = \begin{pmatrix} z \\ z \otimes z \end{pmatrix}$, components of $\overline{\xi}$ will provide moments of 1st and 2nd order of the density of probability $p_{\overline{\xi}}(\xi)$, that is defined by Gaussian law. In this particular case, we can write:

$$\langle \xi, x \rangle = a^T z + \frac{1}{2} z^T H z \tag{42}$$

with $a \in R^n$ and $H \in Sym(n)$. By the change of variables given by $z' = H^{1/2} z + H^{-1/2} a$, we can then compute the logarithm of the Koszul characteristic function:

$$\Phi(x) = -\frac{1}{2}\left[a^T H^{-1} a + \log\det\left[H^{-1}\right] + n\log(2\pi)\right] \tag{43}$$

We can prove that the 1st moment is equal to $-H^{-1}a$ and that components of variance tensor are equal to elements of matrix $H^{-1}$, that induces the second moment. The Koszul Entropy, defined as the Legendre transform of the Koszul characteristic function, is then given by:

$$\Phi^*(\overline{\xi}) = = \frac{1}{2}\left[\log\det\left[H^{-1}\right] + n\log(2\pi.e)\right] \tag{44}$$

*3.4. Crouzeix Relation on Hessian of Dual Potentials and Its Consequences*

In previous sections, we have used the duality between dual potential functions that is recovered by this relation:

$$\Phi^*(x^*) + \Phi(x) = \langle x, x^* \rangle \text{ with } x^* = \frac{d\Phi}{dx} \text{ and } x = \frac{d\Phi^*}{dx^*} \text{ where } \Phi(x) = -\log \psi_\Omega(x) \tag{45}$$

If we develop relations, we can deduce that the hessian of one potential function is the inverse of the hessian of the dual potential function, then the Information Geometry metric could be given in two systems of dual coordinates:

$$\begin{cases} \dfrac{d\Phi}{dx} = x^* \\ \dfrac{d\Phi^*}{dx^*} = x \end{cases} \Rightarrow \begin{cases} \dfrac{d^2\Phi}{dx^2} = \dfrac{dx^*}{dx} \\ \dfrac{d^2\Phi^*}{dx^{*2}} = \dfrac{dx}{dx^*} \end{cases} \Rightarrow \frac{d^2\Phi}{dx^2} \cdot \frac{d^2\Phi^*}{dx^{*2}} = 1 \Rightarrow \frac{d^2\Phi}{dx^2} = \left[\frac{d^2\Phi^*}{dx^{*2}}\right]^{-1}$$
$$\Rightarrow ds^2 = -\frac{d^2\Phi}{dx^2} dx^2 = -\left[\frac{d^2\Phi^*}{dx^{*2}}\right]^{-1} \cdot \left[\frac{d^2\Phi^*}{dx^{*2}} . dx^*\right]^2 = -\frac{d^2\Phi^*}{dx^{*2}} . dx^{*2} \tag{46}$$

Gromov [22] observed that the hessian of the entropy $\Phi^*$ on the space of probability measure is positive definite by the Shannon inequality and defines a (non-complete) Riemannian metric, and that this metric is called the Fisher-Rao-Kramer, Antonelli-Strobeck, Svirezhev-Shahshahani, Karquist metric.

The relation $\frac{d^2\Phi}{dx^2}=\left[\frac{d^2\Phi^*}{dx^{*2}}\right]^{-1}$ has been established first by Crouzeix in 1977 in a short communication [60] for convex smooth functions and their Legendre transforms. This result has been extended for non-smooth function by Seeger [61] and Hiriart-Urruty [62], using a polarity relationship between the second-order sub-differentials. This relation was mentioned in texts of calculus of variations and theory of elastic materials (with work potentials) [62].

This last relation has also been used in the framework of the Monge-Ampere measure associated to a convex function, to prove equality with Lebesgue measure $\lambda$:

$$\begin{aligned} &m_\Phi(\Lambda)=\int_\Lambda \varphi(x)dx=\lambda\left(\left\{\nabla\phi(\mathrm{x})/\mathrm{x}\in\Lambda\right\}\right)\\ &\forall\Lambda\in B_\Omega\ (\text{Borel set in }\Omega)\text{ and }\varphi(x)=\det\left[\nabla^2\Phi(x)\right] \end{aligned} \tag{47}$$

That is proved using the Crouzeix relation $\nabla^2\Phi(x)=\nabla^2\Phi\left(\nabla\Phi^*(y)\right)=\left[\nabla^2\Phi^*(y)\right]^{-1}$:

$$\begin{aligned} &m_\Phi(\Lambda)=\int_\Lambda \varphi(x)dx=\int_\Lambda \det\left[\nabla^2\Phi(x)\right]dx\\ &m_\Phi(\Lambda)=\int_{\left(\nabla\Phi^*\right)^{-1}(A)} \det\left[\nabla^2\Phi\left(\nabla\Phi^*(y)\right)\right]\det\left[\nabla^2\Phi^*(y)\right]dy=\int_{\nabla\Phi(\Lambda)}1.dy=\lambda\left(\left\{\nabla\phi(\mathrm{x})/\mathrm{x}\in\Lambda\right\}\right) \end{aligned} \tag{48}$$

*3.5. Fisher Information Geometry Metric as a Particular Case of Koszul Metric*

To make the link with the classical Fisher metric given by Fisher Information matrix $I(x)$ in Information Geometry, we can observe that the second derivative of $\log p_x(\xi)$ is given by:

$$\begin{aligned} &p_x(\xi)=e^{-\langle\xi,x\rangle}/\int_{\Omega^*}e^{-\langle\xi,x\rangle}d\xi=e^{-\langle x,\xi\rangle-\log\int_{\Omega^*}e^{-\langle\xi,x\rangle}d\xi}\Rightarrow\log p_x(\xi)=-\langle x,\xi\rangle-\log\int_{\Omega^*}e^{-\langle\xi,x\rangle}d\xi\\ &\text{with}\quad \Phi(x)=-\log\int_{\Omega^*}e^{-\langle\xi,x\rangle}d\xi=-\log\psi_\Omega(x) \end{aligned} \tag{49}$$

$$\frac{\partial^2\log p_x(\xi)}{\partial x^2}=\frac{\partial^2\Phi(x)}{\partial x^2} \tag{50}$$

$$\Rightarrow I(x)=-E_\xi\left[\frac{\partial^2\log p_x(\xi)}{\partial x^2}\right]=-\frac{\partial^2\Phi(x)}{\partial x^2}=\frac{\partial^2\log\psi_\Omega(x)}{\partial x^2} \tag{51}$$

We could then deduce the close interrelation between Fisher metric and hessian of minus Koszul-Vinberg characteristic logarithm, that are totally equivalent. Information Geometry then appears as a particular case of Koszul Hessian Geometry.

We can also observed that the Fisher metric or hessian of KVCF logarithm is related to the variance of $\xi$:

$$\log\Psi_\Omega(x)=\log\int_{\Omega^*}e^{-\langle\xi,x\rangle}d\xi\Rightarrow\frac{\partial\log\Psi_\Omega(x)}{\partial x}=-\frac{1}{\int_{\Omega^*}e^{-\langle\xi,x\rangle}d\xi}\int_{\Omega^*}\xi.e^{-\langle\xi,x\rangle}d\xi \tag{52}$$

$$\frac{\partial^2 \log \Psi_{\Omega}(x)}{\partial x^2} = -\frac{1}{\left(\int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi\right)^2} \left[ -\int_{\Omega^*} \xi^2 . e^{-\langle \xi, x \rangle} d\xi . \int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi + \left( \int_{\Omega^*} \xi . e^{-\langle \xi, x \rangle} d\xi \right)^2 \right] \tag{53}$$

$$\frac{\partial^2 \log \Psi_{\Omega}(x)}{\partial x^2} = \int_{\Omega^*} \xi^2 . \frac{e^{-\langle \xi, x \rangle}}{\int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi} d\xi - \left( \int_{\Omega^*} \xi . \frac{e^{-\langle \xi, x \rangle}}{\int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi} d\xi \right)^2 = \int_{\Omega^*} \xi^2 . p_x(\xi) d\xi - \left( \int_{\Omega^*} \xi . p_x(\xi) d\xi \right)^2 \tag{54}$$

$$I(x) = -E_{\xi}\left[ \frac{\partial^2 \log p_x(\xi)}{\partial x^2} \right] = \frac{\partial^2 \log \psi_{\Omega}(x)}{\partial x^2} = E_{\xi}\left[\xi^2\right] - E_{\xi}[\xi]^2 = Var(\xi) \tag{55}$$

The Inverse of the Fisher/Information Matrix $I(x)$ defines the lower bound of statistical estimators. Classically, this Lower bound is called Cramer-Rao Bound because it was described in the Rao's paper of 1945 [63]. Historically, this bound has been published first by Maurice Fréchet in 1939 in his winter "Mathematical Statistics" Lecture at the Institut Henri Poincaré during winter 1939–1940. Maurice Fréchet has published these elements in a paper as early as 1943 [64]. We can read at the bottom of the first page of his paper [64]:

> "*Le contenu de ce mémoire a formé une partie de notre cours de statistique mathématique a l'Institut Henri Poincaré pendant l'hiver 1939–1940. Il constitue l'un des chapitres du deuxième cahier (en préparation) de nos «Leçons de Statistique Mathématique», dont le premier cahier, «Introduction: Exposé préliminaire de Calcul des Probabilités" (119 pages in-quarto, dactylographiées) vient de paraitre au «Centre de Documentation Universitaire, Tournois et Constans. Paris».*"
>
> [*The contents of this report formed a part of our lecture of mathematical statistics at the Henri Poincaré institute during winter 1939–1940. It constitutes one of the chapters of the second exercise book (in preparation) of our "Lessons of Mathematical Statistics", the first exercise book of which, "Introduction: preliminary Presentation of Probability theory" (119 pages quarto, typed) has just been published in the "Centre de Documentation Universitaire, Tournois et Constans. Paris".*]

### *3.6. Extended Results by Koszul, Vey and Sasaki*

Koszul [41,65] and Vey [66,67] have developed extended results with the following theorem for connected hessian manifolds:

Koszul-Vey Theorem*:* Let *M* be a connected hessian manifold with hessian metric $g$. Suppose that *M* admits a closed 1-form $\alpha$ such that $D\alpha = g$ and there exists a group *G* of affine automorphisms of *M* preserving $\alpha$:

- If $M/G$ is quasi-compact, then the universal covering manifold of *M* is affinely isomorphic to a convex domain $\Omega$ of an affine space not containing any full straight line.
- If $M/G$ is compact, then $\Omega$ is a sharp convex cone.

On this basis, Koszul has given a Lie Group construction of a homogeneous cone that has been developed and applied in Information Geometry by Shima [68,69] and Boyom [70] in the framework of Hessian Geometry.

After the pioneering work of Koszul, Sasaki has developed the study of hessian manifolds in Affine Geometry [71,72]. He has denoted by $S_c$ the level surface of $\psi_\Omega$ : $S_c = \{\psi_\Omega(x) = c\}$which is a non-compact sub-manifold in $\Omega$, and by $\omega_c$ the induced metric of $d^2 \log \psi_\Omega$ on $S_c$, then assuming that the cone $\Omega$ is homogeneous under $G(\Omega)$, he proved that $S_c$ is a homogeneous hyperbolic affine hypersphere and every such hyperspheres can be obtained in this way .Sasaki also remarks that $\omega_c$ is identified with the affine metric and $S_c$ is a global Riemannian symmetric space when $\Omega$ is a self-dual cone. He concludes that, let $\Omega$ be a regular convex cone and let $g = d^2 \log \psi_\Omega$ be the canonical Hessian metric, then each level surface of the characteristic function $\psi_\Omega$ is a minimal surface of the Riemannian manifold $(\Omega, g)$.

*3.7. Geodesics Equation for the Koszul Hessian Metric*

The last contribution has been given by Rothaus [73] who studied the construction of geodesics for this hessian metric geometry, using the following property:

$$\Gamma^i_{jk} = \frac{1}{2} g^{il} \left( \frac{\partial g_{ij}}{\partial x_k} + \frac{\partial g_{lk}}{\partial x_j} - \frac{\partial g_{jk}}{\partial x_l} \right) = \frac{1}{2} g^{il} \frac{\partial^3 \log \psi_\Omega(x)}{\partial x_j \partial x_k \partial x_l} \text{ with } g_{ij} = \frac{\partial^2 \log \psi_\Omega(x)}{\partial x_i \partial x_j} \tag{56}$$

or expressed also according the Christoffel symbol of the first kind:

$$\Gamma_{ijk} = \frac{1}{2} \left( \frac{\partial g_{jk}}{\partial x_i} + \frac{\partial g_{ki}}{\partial x_j} - \frac{\partial g_{ij}}{\partial x_k} \right) = \frac{1}{2} \frac{\partial^3 \log \psi_\Omega(x)}{\partial x_j \partial x_k \partial x_l} \tag{57}$$

Then geodesic is given by:

$$\frac{d^2 x_k}{ds^2} + \Gamma^k_{ij} \frac{dx_i}{ds} \frac{dx_j}{ds} = g_{kl} \frac{d^2 x_k}{ds^2} + \Gamma_{ijk} \frac{dx_i}{ds} \frac{dx_j}{ds} = 0 \tag{58}$$

that could be developed with previous relation:

$$\frac{d^2 x_k}{ds^2} \frac{\partial^2 \log \psi_\Omega}{\partial x_k \partial x_l} + \frac{1}{2} \frac{dx_i}{ds} \frac{dx_j}{ds} \frac{\partial^3 \log \psi_\Omega}{\partial x_l \partial x_i \partial x_j} = 0 \tag{59}$$

We can then observe that:

$$\frac{d^2}{ds^2} \left[ \frac{\partial \log \psi_\Omega}{\partial x_l} \right] = \frac{dx_i}{ds} \frac{dx_j}{ds} \frac{\partial^3 \log \psi_\Omega}{\partial x_i \partial x_j \partial x_l} + \frac{d^2 x_k}{ds^2} \frac{\partial^2 \log \psi_\Omega}{\partial x_k \partial x_l} \tag{60}$$

The geodesic equation can then be rewritten:

$$\frac{d^2 x_k}{ds^2} \frac{\partial^2 \log \psi_\Omega}{\partial x_k \partial x_l} + \frac{d^2}{ds^2} \left[ \frac{\partial \log \psi_\Omega}{\partial x_l} \right] = 0 \tag{61}$$

That we can put in vector form using notations $x^* = -d\log\psi_\Omega$ and Fisher matrix $I(x) = d^2\log\psi_\Omega$:

$$I(x)\frac{d^2x}{ds^2} - \frac{d^2x^*}{ds^2} = 0 \text{ or } I(x) = \left[\frac{d^2x}{ds^2}\right]^{-1}\frac{d^2x^*}{ds^2} \tag{62}$$

*3.8. Koszul Metric for Siegel Homogeneous Domains*

Koszul [42] has developed his previously described theory for Homogenous Siegel Domains *SD*. He has proved that there is a subgroup *G* in the group of the complex affine automorphisms of these domains (Iwasawa subgroup), such that *G* acts on *SD* simply transitively. The Lie algebra $g$ of *G* has a structure that is an algebraic translation of the Kähler structure of *SD*. There is an integrable almost complex structure *J* on, *g* and there exists $\eta \in g^*$ such that $\langle X,Y\rangle_\eta = \langle [JX,Y],\eta\rangle$ defines a *J-invariant positive definite inner product on* $g$**.** Koszul has proposed as admissible form $\eta \in g^*$, the form $\xi$:

$$\Psi(X) = \langle X,\xi\rangle = Tr[ad(JX) - J.ad(X)] \quad \forall X \in g \tag{63}$$

Koszul has proved that $\langle X,Y\rangle_\xi$ coincides, up to a positive number multiple with the real part of the Hermitian inner product obtained by the Bergman metric of *SD* by identifying *g* with the tangent space of *SD*. The First Koszul form is then given by:

$$\alpha = -\frac{1}{4}d\Psi(X) \tag{64}$$

We can illustrate this new Koszul expression for Poincaré's Upper Half Plane $V = \{z = x + iy \,/\, y > 0\}$ (most simple symmetric homogeneous bounded domain).

Define vector fields $X = y\frac{d}{dx}$ and $Y = y\frac{d}{dy}$, and *J* an almost complex structure on *V* defined by $JX = Y$

As:

$$[X,Y] = -Y \text{ and } ad(Y).Z = [Y,Z] \text{ then } \begin{cases} Tr[ad(JX) - Jad(X)] = 2 \\ Tr[ad(JY) - Jad(Y)] = 0 \end{cases} \tag{65}$$

The Koszul 1-form and then the Koszul/Poincaré metric is given by:

$$\Psi(X) = 2\frac{dx}{y} \Rightarrow \alpha = -\frac{1}{4}d\Psi = -\frac{1}{2}\frac{dx\wedge dy}{y^2} \Rightarrow ds^2 = \frac{dx^2+dy^2}{2y^2} \tag{66}$$

This could be also applied for Siegel's Upper Half Space $V = \{Z = X + iY \,/\, X,Y \in Sym(p), Y > 0\}$ (more natural extension of Poincaré Upper-half plane, and general notion of symmetric bounded homogeneous domains studied by Elie Cartan and Carl-Ludwig Siegel):

$$\begin{cases} SZ = (AZ + B)D^{-1} \\ A^T D = I, B^T D = D^T B \end{cases} \quad \text{with} \quad S = \begin{pmatrix} A & B \\ 0 & D \end{pmatrix} \quad \text{and} \quad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \tag{67}$$

$$\Psi(dX + idY) = \frac{3p+1}{2} Tr\left(Y^{-1} dX\right) \Rightarrow \begin{cases} \alpha = -\frac{1}{4} d\Psi = \frac{3p+1}{8} Tr\left(Y^{-1} dZ \wedge Y^{-1} d\bar{Z}\right) \\ ds^2 = \frac{(3p+1)}{8} Tr\left(Y^{-1} dZ Y^{-1} d\bar{Z}\right) \end{cases} \tag{68}$$

To recover the metric of the space of Symmetric Positive Definite (HPD) matrices, we take $Z = iR$ (with $X = 0$), and obtain the metric $ds^2 = Tr\left[\left(R^{-1} dR\right)^2\right]$. In the context of Information Geometry, this metric is the metric for multivariate Gaussian law of covariance matrix $R$ and zero mean. For more development and application for Radar signal processing, we give reference to author papers [74–77].

## 4. Souriau Geometric Temperature and Covariant Definition of Thermodynamic Equilibriums

Souriau, a student of Elie Cartan [78] at ENS Ulm in 1946, has given in [59,79–87] a covariant definition of thermodynamic equilibriums and has formulated statistical mechanics [88–90] and thermodynamics in the framework of Symplectic Geometry [59] by use of symplectic moments and distribution-tensor concepts, giving a geometric status for temperature, heat and entropy. This work has been extended by Vallée and de Saxcé [91–94], Iglésias [95,96] and Dubois [97]. Other recent works address equilibrium states on manifolds of negative curvature and could be analyzed in the framework of Information Geometry [98–103].

Other directions related to polarized surface have been developed by Donaldson, Guillemin and Abreu, in which invariant Kähler metrics correspond to convex functions on the moment polytope of a toric variety [104–108] based on precursor work of Atiyah and Bott [109] on moment map and its convexity by Bruguières [110], Condevaux [111], Delzant [112], Guillemin and Sternberg [113] and Kirwan [114]. More recently, Mikhail Kapranov has also given a thermodynamical interpretation of the moment map for toric varieties [115]. Readers may consult the tutorial paper of Biquard [116].

The first general definition of the "*moment map"* (constant of the motion for dynamical systems) was introduced by Souriau during 1970s, with geometric generalization of such earlier notions as the Hamiltonian and the invariant theorem of Noether describing the connection between symmetries and invariants (it is the moment map for a one-dimensional Lie group of symmetries). In symplectic geometry the analog of Noether's theorem is the statement that the moment map of a Hamiltonian action which preserves a given time evolution is itself conserved by this time evolution. The conservation of the moment of a Hamilotnian action was called by Souriau the "*Symplectic or Geometric Noether theorem*" (considering phases space as symplectic manifold, cotangent fiber of configuration space with canonical symplectic form, if Hamiltonian has Lie

algebra, moment map is constant along system integral curves. Noether theorem is obtained by considering independently each component of moment map).

In previous approach based on Koszul's work, we have defined two convex functions $\Phi(x)$ and $\Phi^*(x^*)$ with dual system of coordinates $x$ and $x^*$ on dual cones $\Omega$ and $\Omega^*$:

$$\Phi(x) = -\log \int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi \quad \forall x \in \Omega \text{ and } \Phi^*(x^*) = \left\langle x, x^* \right\rangle - \Phi(x) = -\int_{\Omega^*} p_x(\xi) \log p_x(\xi) d\xi \tag{69}$$

where:

$$x^* = \int_{\Omega^*} \xi . p_x(\xi) d\xi \text{ and } p_x(\xi) = e^{-\langle \xi, x \rangle} / \int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi = e^{-\langle x, \xi \rangle - \log \int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi} = e^{-\langle x, \xi \rangle + \Phi(x)} \tag{70}$$

with

$$x^* = \frac{\partial \Phi(x)}{\partial x} \text{ and } x = \frac{\partial \Phi^*(x^*)}{\partial x^*} \tag{71}$$

Souriau introduced these relations in the framework of variational problems to extend them with a covariant definition. Let $M$ be a differentiable manifold with a continuous positive density $d\omega$ and let $E$ a finite vector space and $U(\xi)$ a continuous function defined on $M$ with values in $E$. A continuous positive function $p(\xi)$ solution of this problem with respect to calculus of variations:

$$\underset{p(\xi)}{ArgMin} \left[ s = -\int_M p(\xi) \log p(\xi) d\omega \right] \text{ such that } \begin{cases} \int_M p(\xi) d\omega = 1 \\ \int_M U(\xi) p(\xi) d\omega = Q \end{cases} \tag{72}$$

is given by:

$$p(\xi) = e^{\Phi(\beta) - \beta . U(\xi)} \text{ with } \Phi(\beta) = -\log \int_M e^{-\beta . U(\xi)} d\omega \text{ and } Q = \frac{\int_M U(\xi) e^{-\beta . U(\xi)} d\omega}{\int_M e^{-\beta . U(\xi)} d\omega} \tag{73}$$

Entropy $s = -\int_M p(\xi) \log p(\xi) d\omega$ can be stationary only if there exist a scalar $\Phi$ and an element $\beta$ belonging to the dual of $E$, where $\Phi$ and $\beta$ are Lagrange parameters associated to the previous constraints. Entropy appears naturally as Legendre transform of $\Phi$ :

$$s(Q) = \beta . Q - \Phi(\beta) \tag{74}$$

This value is a strict minimum of $s$, and the equation $Q = \frac{\int_M U(\xi) e^{-\beta . U(\xi)} d\omega}{\int_M e^{-\beta . U(\xi)} d\omega}$ has a maximum of one solution for each value of $Q$. The function $\Phi(\beta)$ is differentiable and we can write $d\Phi = d\beta . Q$ and identifying $E$ with its dual:

$$Q = \frac{\partial \Phi}{\partial \beta} \tag{75}$$

Uniform convergence of $\int_M U(\xi) \otimes U(\xi) e^{-\beta . U(\xi)} d\omega$ proves that $-\frac{\partial^2 \Phi}{\partial \beta^2} > 0$ and that $-\Phi(\beta)$ is convex. Then, $Q(\beta)$ and $\beta(Q)$ are mutually inverse and differentiable, where $ds = \beta . dQ$.

Identifying *E* with its bidual:

$$\beta = \frac{\partial s}{\partial Q} \tag{76}$$

Classically, if we take $U(\xi) = \begin{pmatrix} \xi \\ \xi \otimes \xi \end{pmatrix}$, components of $Q$ will provide moments of first and second order of the density of probability $p(\xi)$, that is defined by Gaussian law.

Souriau has applied this approach for classical statistical mechanic systems. Considering a mechanical system with *n* parameters $q_1, \cdots, q_n$, its movement could be defined by its phase at arbitrary time *t* on a manifold of dimension *2n*: $q_1, \cdots, q_n, p_1, \cdots, p_n$.

The Liouville theorem shows that coordinate changes have a Jacobian equal to unity, and a Liouville density could be defined on manifold $M$ : $d\omega = dq_1 \cdots dq_n dp_1 \cdots dp_n$ that will not depend on choice to *t*.

A system state is one point on *2n*-Manifold *M* and a statistical state is a law of probability defined on $M$ such that $\int_M p(\xi) d\omega(\xi) = 1$, and its time evolution is driven by:

$$\frac{\partial p}{\partial t} = \sum \frac{\partial p}{\partial p_j} \frac{\partial H}{\partial q_j} - \frac{\partial p}{\partial q_j} \frac{\partial H}{\partial p_j} \tag{77}$$

where *H* is the Hamiltonian.

A thermodynamic equilibrium is a statistical state that maximizes the entropy:

$$s = -\int_M p(\xi) \log p(\xi) d\omega \tag{78}$$

among all states giving the mean value of energy Q:

$$\int_M H(\xi) . p(\xi) d\omega = Q \tag{79}$$

Applying this for free particles, for an ideal gas, equilibrium is given for $\beta = \frac{1}{kT}$ (with k being the Boltzmann constant) and if we set $S = k.s$, the previous relation $dS = \frac{dQ}{T}$ provides: $S = \int \frac{dQ}{T}$ and $S = \int \frac{dQ}{T}$ and $\Phi(\beta)$ is identified with the Massieu-Duhem Potential. We recover also the Maxwell Speed law:

$$p(\xi) = cste . e^{-\frac{H}{kT}} \tag{80}$$

The main discovery of Jean-Marie Souriau is that *previous thermodynamic equilibrium is not covariant on a relativity point of view*. Then, he has proposed a covariant definition of thermodynamic equilibrium where the previous definition is a particular case. In previous formalization, manifold $M$ was solution of the calculus of variations problem:

$$d\int_{t_0}^{t_1} l\left(t, q_j, \frac{dq_j}{dt}\right)dt = 0 \text{ with } p_j = \frac{\partial l}{\partial q_j} \tag{81}$$

We can then consider the time variable *t* like other variables $q_j$ through an arbitrary parameter $\tau$, and define the new calculus of variations problem by:

$$d\int_{t_0}^{t_1} L\left(q_J, \dot{q}_J\right)d\tau = 0 \text{ with } t = q_{n+1},\ \dot{q}_J = \frac{dq_J}{d\tau} \text{ and } J = 1, 2, ..., n+1 \tag{82}$$

where:

$$L\left(q_J, \dot{q}_J\right) = l\left(t, q_j, \frac{\dot{q}_j}{\dot{t}}\right)\dot{t} \tag{83}$$

Variables $p_j$ are not changed and we have the relation:

$$p_{n+1} = l - \sum_j p_j . \frac{dq_j}{dt} \tag{84}$$

If we compare with classical mechanic, we have:

$$p_{n+1} = -H \text{ with } H = \sum_j p_j . \frac{dq_j}{dt} - l \text{ ($H$ is Legendre transform of $l$)} \tag{85}$$

*H* is the energy of the system that is conservative if the Lagrangian doesn't depend explicitly of time *t*. It is a particular case of Noether Theorem:

If Lagrangian L is invariant by an infinitesimal transform $dQ_J = F_J(Q_K)$ ,then $u = \sum_J p_J dQ_J$ is first integral of variations equations.

As energy is not the conjugate variable of time *t*, or the value provided by Noether theorem by system invariance to time translation, the thermodynamic equilibrium is not covariant. Then, Souriau proposes a new covariant definition of thermodynamic equilibrium:

Let a mechanical system with a Lagrangian invariant by a Lie Group G. Equilibrium states by Group G are statistical states that maximizes the Entropy, while providing given mean values to all variables associated by Noether theorem to infinitesimal transforms of group G.

Neither theorem allows associating to all system movement $\xi$ , a value $U(\xi)$ belonging to the vector space dual of Lie Algebra $\mathfrak{g}$ of group *G*. $U(\xi)$ is called *the moment* of the group.

For each derivation $\delta$ of this Lie algebra [83], we take:

$$U(\xi)(\delta) = \sum_J p_J . \delta Q_J \tag{86}$$

With previous development, as $\mathfrak{g}^*$ is dual of $\mathfrak{g}$, value $\beta$ belongs to this Lie algebra $\mathfrak{g}$, geometric generalization of thermodynamic temperature. Value Q is a geometric generalization of heat and belongs to $\mathfrak{g}^*$, the dual of $\mathfrak{g}$.

An Equilibrium state exists having the largest entropy, with a distribution function $p(\xi)$ that is the exponential of an affine function of $U$ [83]:

$$p(\xi) = e^{\Phi(\beta)-\beta.U(\xi)} \text{ with } \Phi(\beta) = -\log \int_M e^{-\beta.U(\xi)} d\omega \text{ and } Q = \frac{\int_M U(\xi) e^{-\beta.U(\xi)} d\omega}{\int_M e^{-\beta.U(\xi)} d\omega} \tag{87}$$

with:

$$s(Q) = \beta.Q - \Phi(\beta),\ d\Phi = d\beta.Q \text{ and } ds = \beta.dQ \tag{88}$$

A statistical state $p(\xi)$ is invariant by $\delta$ if $\delta[p(\xi)] = 0$ for all $\xi$ (then $p(\xi)$ is invariant by finite transform of $G$ generated by $\delta$).

Jean-Marie Souriau gave the following theorem:

*Souriau Theorem 1*: An equilibrium state allowed by a group $G$ is invariant by an element $\delta$ of Lie Algebra $\mathfrak{g}$, if and only if $[\delta, \beta] = 0$ (with [.], the Lie Bracket), with $\beta$ the generalized equilibrium temperature.

For classical thermodynamic, where $G$ is an Abelian group of translation with respect to time $t$, all equilibrium states are invariant under $G$. For Group of transformation of Space-Time, elements of Lie Algebra of $G$ could be defined as vector fields in Space-Time. The generalized temperature $\beta$ previously defined, would be also defined as a vector field. For each point of manifold $M$, we could then define:

- Temperature Vector:

$$\beta_M = \frac{V}{kT} \tag{89}$$

with:

- Unitary Mean Speed:

$$\text{Unitary Mean Speed: } V = \frac{\beta_M}{\|\beta_M\|} \text{ with } \|V\| = 1 \tag{90}$$

- Eigen Absolute Temperature:

$$T = \frac{1}{k.\|\beta_M\|} \tag{91}$$

Classical formula of thermodynamics are thus generalized, but variables are defined with a geometrical status, like the geometrical temperature $\beta_M$ an element of the Lie algebra of the Galileo or Poincaré groups, interpreted as the field of space-time vectors. Souriau proved that in relativistic version $\beta_M$ is a *time like vector* with an orientation that characterizes *the arrow of time*. The temperature vector and entropy flux are in duality. Souriau said "$\beta$, *c'est la flèche qui nous*

*indique dans quel sens coule le temps*" [$\beta$, it is the arrow that informs about the flow of time direction].

## 5. Souriau-Gibbs Canonical Ensemble of Dynamical Group and Lie Group Thermodynamics

In statistical mechanics, a canonical ensemble [117–121] is the statistical ensemble that is used to represent the possible states of a mechanical system that is being maintained in thermodynamic equilibrium. Souriau has defined this Gibbs canonical ensemble on Symplectic manifold *M* for a Lie group action on *M*.

In classical statistical mechanics, a state is given by the solution of Liouville equation on the phase space, the partition function. The seminal idea of Lagrange was to consider that a statistical state is simply a probability measure on the manifold of motions, as in the Souriau approach, where one movement of a dynamical system (classical state) is a point on manifold of movements. For statistical mechanics, the movement variable is replaced by a random variable where a statistical state is probability law on this manifold. As symplectic manifolds have a completely continuous measure, invariant by diffeomorphisms, the Liouville measure $\lambda$, all statistical states will be the product of Liouville measure by the scalar function given by the generalized partition function $e^{\Phi-\beta.U}$ defined by the generalized energy *U* (the moment that is defined in dual of Lie Algebra of this dynamical group) and the geometric temperature $\beta$, where $\Phi$ is a normalizing constant such the mass of probability is equal to 1, $\Phi=-\log\int_M e^{-\beta.U}d\omega$. Souriau then generalizes the Gibbs equilibrium state to all symplectic manifolds that have a dynamical group. To ensure that all integrals, that will be defined, could converge, *the canonical Gibbs ensemble is the largest open proper subset (in Lie algebra) where these integrals are convergent. This canonical Gibbs ensemble is convex*. The derivative of $\Phi$, $Q=\frac{\partial\Phi}{\partial\beta}$ is equal to the mean value of the energy *U* (heat in thermodynamic). The minus derivative of this generalized heat *Q*, $-\frac{\partial Q}{\partial\beta}$ is symmetric and positive (it is a generalization of heat capacity). Entropy $s$ is then defined by Legendre transform of $\Phi$, $s=\beta.Q-\Phi$. If this approach is applied for the group of time translation, this is the classical thermodynamic theory. But Souriau has observed that if we apply this theory for non-commutative group (Galileo or Poincaré groups), the symmetry has been broken. Classical Gibbs equilibrium states are no longer invariant by this group. This symmetry breaking provides new equations, discovered by Souriau.

For each temperature $\beta$, Souriau has introduced a tensor $f_\beta$, equal to the sum of cocycle *f* and Heat coboundary (with [.,.] Lie bracket):

$$f_\beta(Z_1,Z_2)=f(Z_1,Z_2)+Q.Ad_{Z_1}(Z_2)\quad\text{with}\quad Ad_{Z_1}(Z_2)=[Z_1,Z_2] \tag{92}$$

This tensor $f_\beta$ has the following properties:

- $f$ is a symplectic cocycle (we refer to books of Sympectic geometry for cocycle definition)
- $\beta \in Ker\ f_\beta$
- The following symmetric tensor $g_\beta$, defined on all values of $Ad_\beta(.)$ is positive definite:

$$g_\beta([\beta, Z_1], [\beta, Z_2]) = f_\beta(Z_1, [\beta, Z_2]) \tag{93}$$

These equations are universal, because they are not dependent of the symplectic manifold but only of the dynamical group G, its symplectic cocycle $f$, the temperature $\beta$ and the heat $Q$. Souriau called this model "*Lie Groups Thermodynamics*". We can read in his paper this prophetical sentence "*Peut-être cette thermodynamique des groups de Lie a-t-elle un intérêt mathématique"[Maybe this thermodynamics of Lie groups has a mathematical interest]*. He explains that for dynamic Galileo group (rotation and translation) with only one axe of rotation, this thermodynamic theory is the theory of centrifuge where the temperature vector dimension is equal to 2 (sub-group of invariance of size 2), used to make "butter", "uranium 235" and "ribonucleic acid". The physical meaning of these 2 dimensions for vector-valued temperature are "thermic conduction" and "viscosity". Souriau said that the model unifies "heat conduction" and "viscosity" (Fourier and Navier equations) in the same theory of irreversible process. Souriau has applied this theory in details for relativistic ideal gas with Poincaré group for dynamical group.

We will give in the following the two others main theorems of Souriau on this Lie Group Thermodynamics.

*Souriau Theorem 2*. Let $\Omega$ be the largest open proper subset of $\mathfrak{g}$, Lie algebra of G, such that $\int_M e^{-\beta.U(\xi)} d\omega$ and $\int_M \xi . e^{-\beta.U(\xi)} d\omega$ are convergent integrals, this set $\Omega$ is convex and is invariant under every transformation $\overline{a}_\mathfrak{g}$, where $a \mapsto \overline{a}_\mathfrak{g}$ is the adjoint representation of G. Then, the variables are changed according to:

$$\beta \to \overline{a}_\mathfrak{g}(\beta) \tag{94}$$

$$\Phi \to \Phi - \theta\left(a^{-1}\right)\beta = \Phi + \theta(a).\overline{a}_\mathfrak{g}(\beta) \tag{95}$$

$$s \to s \tag{96}$$

$$Q \to \overline{a}_{\mathfrak{g}^*}(Q) + \theta(a) = \overline{a}_{\mathfrak{g}_\theta^*}(Q) \tag{97}$$

$$\varsigma \to \overline{a}_M^+(\varsigma) \tag{98}$$

where $\theta$ is the cocycle associated with the group *G* and the moment, and $\overline{a}_M^+(\varsigma)$ is the image under $\overline{a}_M$ of the probability measure $\varsigma$.

We observe that the entropy *s* is unchanged, and $\Phi$ is changed but with linear dependence to $\beta$, with consequence that Fisher Information Geometry metric is unchanged by the dynamical group:

$$I\left(\overline{a}_\mathfrak{g}(\beta)\right) = -\frac{\partial^2\left[\Phi - \theta\left(a^{-1}\right)\beta\right]}{\partial\beta^2} = -\frac{\partial^2\Phi}{\partial\beta^2} = I(\beta) \tag{99}$$

These transformations have been geometrically interpreted by Souriau in Figure 4:

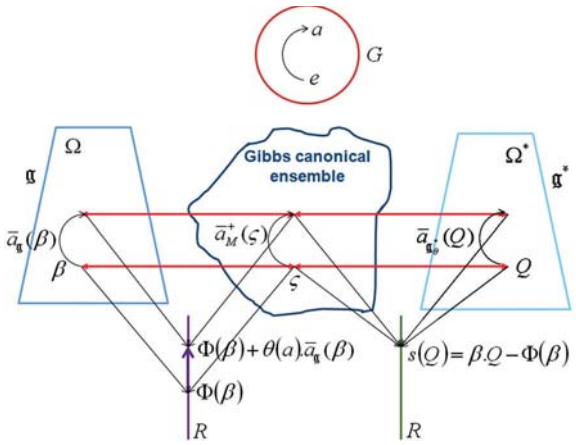



**Figure 4.** Souriau figure on Lie Groups Thermodynamics.

In previous notation, $a \mapsto \bar{a}_{\mathfrak{g}}$ the adjoint representation of $G$ can be written:

$$\bar{a}_{\mathfrak{g}}(Z) = \delta\left[a \times b \times a^{-1}\right] \quad \text{with} \quad b = e \ , \ \delta b = Z \quad \text{and} \quad \delta a = 0 \tag{100}$$

$a \mapsto \bar{a}_{\mathfrak{g}}$ defines an action of $G$ on its Lie algebra $\mathfrak{g}$, with $\bar{a}_{\mathfrak{g}}$ is called the adjoint representation, that is a linear representation of $G$ on its Lie algebra $\mathfrak{g}$.

Let $a$ be an arbitrary element of $G$ and $\bar{a}_M$ action of $a$ on the manifold $M$. Since $\bar{a}_M^{-1}$ is a symplectomorphism, the image under $\bar{a}_M^{-1}$ of the Liouville measure $\lambda$ is equal to $\lambda$. The integral $\int_M e^{-\beta.U(\xi)}.d\omega$ is equal with invariance property of Liouville measure to the integral $\int_M e^{-\beta.U\left(\bar{a}_M^{-1}(\xi)\right)}.d\omega$:

$$\int_M e^{-\beta.U(\xi)}.d\omega = \int_M e^{-\beta.U\left(\bar{a}_M^{-1}(\xi)\right)}.d\omega \tag{101}$$

We can then use the following relation:

$$U\left(\bar{a}_M^{-1}(\xi)\right) = \bar{a}_{\mathfrak{g}^*}^{-1}\left(U(\xi)\right) + \theta\left(a^{-1}\right) \tag{102}$$

with $\theta$ a symplectic cocycle of $G$. This cocycle is defined for:

$$\begin{aligned} U : M &\to \mathfrak{g}^* \\ \xi &\mapsto \mu \end{aligned} \tag{103}$$

there exist then a differential map $\theta$ defined by:

$$\begin{aligned} \theta : G &\to \mathfrak{g}^* \\ a &\mapsto U\left(\bar{a}_M(\xi)\right) - \bar{a}_{\mathfrak{g}^*}\left(U(\xi)\right) \end{aligned} \tag{104}$$

This differential map $\theta$ satisfy the condition:

$$\theta(a\times b)=\theta(a)+\overline{a}_{\mathfrak{g}^*}(\theta(b)) \tag{105}$$

and its derivative $f=D(\theta)(e)$ where $e$ is the identity element of $G$, is a 2-form on the Lie algebra $\mathfrak{g}$ of $G$ which satisfies:

$$f\left(Z_1,[Z_2,Z_3]\right)+f\left(Z_2,[Z_3,Z_1]\right)+f\left(Z_3,[Z_1,Z_2]\right)=0 \quad , \quad \forall Z_1,Z_2,Z_3\in\mathfrak{g} \tag{106}$$

and the following identities:

$$D(U)(\xi,Z_M(\xi))=U(\xi).Ad_Z(.)+f(Z,Z) \tag{107}$$

where $Z_M(\xi)$ is the fundamental vector field on the manifold $M$ associated to $Z\in\mathfrak{g}$:

$$Z_M(\xi)=\delta\left[\overline{a}_M(\xi)\right] \text{ for } a=e \ , \ \delta a=Z \text{ and } \delta\xi=0 \tag{108}$$

$$\sigma\left(Z_{1,M}(\xi),Z_{2,M}(\xi)\right)=\mu.[Z_1,Z_2]+f(Z_1,Z_2) \tag{109}$$

with $\sigma$ the Lagrange form.

If we use previous relation $U\left(\overline{a}_M^{-1}(\xi)\right)=\overline{a}_{\mathfrak{g}^*}^{-1}(U(\xi))+\theta\left(a^{-1}\right)$, and the property that $\overline{a}_{\mathfrak{g}^*}(U(\xi))=U(\xi).\overline{a}_{\mathfrak{g}}^{-1}$, by defining:

$$\beta'=\overline{a}_{\mathfrak{g}}(\beta) \tag{110}$$

the integral is then defined by:

$$\int_M e^{-\beta'.U(\xi)}.d\omega=\int_M e^{-\overline{a}_{\mathfrak{g}}(\beta).U\left(\overline{a}_M^{-1}(\xi)\right)}.d\omega=\int_M e^{-\overline{a}_{\mathfrak{g}}(\beta).\left[\overline{a}_{\mathfrak{g}^*}^{-1}(U(\xi))+\theta\left(a^{-1}\right)\right]}.d\omega=e^{\theta\left(a^{-1}\right).\beta}\int_M e^{-\beta.U(\xi)}.d\omega \tag{111}$$

We can then deduce the equation of Souriau theorem on $\Phi$:

$$\Phi'=\Phi(\beta')=\Phi\left(\overline{a}_{\mathfrak{g}}(\beta)\right)=-\log\int_M e^{-\beta'.U(\xi)}.d\omega=-\log\left(e^{\theta\left(a^{-1}\right).\beta}\int_M e^{-\beta.U(\xi)}.d\omega\right)=\Phi(\beta)-\theta\left(a^{-1}\right)\beta \tag{112}$$

The equation of Souriau theorem on $Q$ uses the relation $\overline{a}_{\mathfrak{g}^*}(Q)=Q.\overline{a}_{\mathfrak{g}}^{-1}$:

$$\Phi'=\Phi(\beta')=\Phi\left(\overline{a}_{\mathfrak{g}}(\beta)\right)=-\log\int_M e^{-\beta'.U(\xi)}.d\omega=-\log\left(e^{\theta\left(a^{-1}\right).\beta}\int_M e^{-\beta.U(\xi)}.d\omega\right)=\Phi(\beta)-\theta\left(a^{-1}\right)\beta \tag{113}$$

Finally, using $\overline{a}_{\mathfrak{g}^*}(Q)=Q.\overline{a}_{\mathfrak{g}}^{-1}$, we can prove that the Entropy is invariant:

$$s'=\beta'.Q'-\Phi'=\overline{a}_{\mathfrak{g}}(\beta).\left(\overline{a}_{\mathfrak{g}^*}(Q)+\theta(a)\right)-\left(\Phi+\theta(a)\overline{a}_{\mathfrak{g}}(\beta)\right)=\overline{a}_{\mathfrak{g}}(\beta).\overline{a}_{\mathfrak{g}^*}(Q)-\Phi=\overline{a}_{\mathfrak{g}}^{-1}\overline{a}_{\mathfrak{g}}(\beta).Q-\Phi=\beta.Q-\Phi=s \tag{114}$$

Considering the density of probability $p_\beta(\xi)=e^{-\beta.U(\xi)+\Phi(\beta)}$ with $\beta'=\overline{a}_{\mathfrak{g}}(\beta)$, then: $p_{\beta'}(\xi)=e^{-\overline{a}_{\mathfrak{g}}(\beta).U(\xi)+\Phi(\beta)-\theta\left(a^{-1}\right)\beta}$.

From which, we can recover $\overline{a}_M^+$ the image under $\overline{a}_M$ of the probability measure.

The last Souriau theorem is given by:

*Souriau Theorem 3*. Let $f = D(\theta)(e)$ be the derivative of $\theta$ (symplectic cocycle of G) at the identity element and let us define:

$$\forall \beta \in \Omega,\ \ f_\beta(Z_1,Z_2) = f(Z_1,Z_2) + Q.Ad_{Z_1}(Z_2) \ \text{ with } \ Ad_{Z_1}(Z_2) = [Z_1,Z_2] \tag{115}$$

Then

$f_\beta$ *is a symplectic cocycle of* $\mathfrak{G}$ *, that is independent of the moment of G*

$$f_\beta(\beta,\beta) = 0 \ \ , \ \ \forall \beta \in \Omega \ (\beta \in Ker\, f_\beta) \tag{116}$$

- There exists a symmetric tensor $g_\beta$ defined on the image of $Ad_\beta(.) = [.,\beta]$ such that:

$$g_\beta([\beta,Z_1],Z_2) = f_\beta(Z_1,Z_2)\ ,\ \forall Z_1 \in \mathfrak{G}, \forall Z_2 \in \mathrm{Im}(Ad_\beta(.)) \tag{117}$$

and:

$$g_\beta(Z_1,Z_2) \geq 0 \ \ , \ \ \forall Z_1,Z_2 \in \mathrm{Im}(Ad_\beta(.)) \tag{118}$$

Last equation gives the structure of a positive Euclidean space.

$f_\beta(\beta,\beta) = 0$ could be deduced by differentiating $\Phi(\bar{a}_{\mathfrak{G}}(\beta)) = \Phi + \theta(a).\bar{a}_{\mathfrak{G}}(\beta)$ and taking $a = \mathrm{e}$, $\delta a = Z_2$ and $\delta Z_1 = 0$. As $Z_M(\xi) = \delta[\bar{a}_M(\xi)]$ and $Z_{\mathfrak{G}} = -Ad_Z(.)$, we have $Q[Z_1,Z_2] = -f(Z_1,Z_2)$.

If we differentiate $Q(\bar{a}_{\mathfrak{G}}(\beta))\bar{a}_{\mathfrak{G}^*}(Q) + \theta(a)$ , *the following relation* $\frac{\partial Q}{\partial \beta}(-[Z_1,\beta]) = f(Z_1,Z_1) + Q.Ad_{Z_1}(.) = f_\beta(Z_1,Z_1)$ appears. Then, writing $\delta\beta = [\beta,Z_1] = Z_2$ , we have $\delta Q.\delta\beta \geq 0 \Rightarrow f_\beta(Z_1,Z_2) \geq 0$.

See more details in appendix A.3.

## 6. Synthesis of Analogies Between the Koszul Information Geometry Model and Souriau Statistical Physics Model

### *6.1. Comparison of Koszul and Souriau Models*

We will synthetize in Table 1 results of previous chapters with Koszul Hessian Structure of Information Geometry and the Souriau model of Statistical Physics with the general concepts of geometric temperature, heat and capacity. Analogies between models will deal with characteristic function, Entropy, Legendre Transform, density of probability, dual coordinate systems, Hessian Metric and Fisher metric.

As $Q = \frac{\partial \Phi}{\partial \beta}$, we observe that the Information Geometry metric $I(\beta) = -\frac{\partial^2 \Phi(\beta)}{\partial \beta^2} = -\frac{\partial Q}{\partial \beta}$ could be considered as a *generalization of "Heat Capacity"*. Souriau called it $K$ the "*Geometric Capacity*". When $\beta = \frac{1}{kT}$ , $K = -\frac{\partial Q}{\partial \beta} = -\frac{\partial Q}{\partial T}\left(\frac{\partial \frac{1}{kT}}{\partial T}\right) = \frac{1}{kT^2}\frac{\partial Q}{\partial T}$ , then this geometric capacity is related to calorific capacity. $Q$ is related to the mean, and $K$ is related to the variance of $U$ [122]:

$$I(\beta) = -\frac{\partial Q}{\partial \beta} = \mathrm{var}(U) = \int_M U(\xi)^2 . p_\beta(\xi) d\omega - \left( \int_M U(\xi) . p_\beta(\xi) d\omega \right)^2 \quad (119)$$

**Table 1.** Synthesis of Koszul and Souriau models.

| | **Koszul Information Geometry Model** | **Souriau Lie Groups Thermodynamics Model** |
|---|---|---|
| Characteristic function | $\Phi(x) = -\log \int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi \quad \forall x \in \Omega$ | $\Phi(\beta) = -\log \int_M e^{-\beta . U(\xi)} d\omega \quad \forall \beta \in \mathfrak{g}$ |
| Entropy | $\Phi^*(x^*) = -\int_{\Omega^*} p_x(\xi) \log p_x(\xi) d\xi$ | $s = -\int_M p(\xi) \log p(\xi) d\omega$ |
| Legendre Transform | $\Phi^*(x^*) = \langle x, x^* \rangle - \Phi(x)$ | $s(Q) = \beta . Q - \Phi(\beta)$ |
| Density of probability | $p_x(\xi) = e^{-\langle x, \xi \rangle + \Phi(x)}$<br>$p_x(\xi) = \frac{e^{-\langle \xi, x \rangle}}{\int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi}$ | $p_\beta(\xi) = e^{-\beta . U(\xi) + \Phi(\beta)}$<br>$p_\beta(\xi) = \frac{e^{-\beta . U(\xi)}}{\int_M e^{-\beta . U(\xi)} d\omega}$ |
| Dual Coordinate Systems | $x \in \Omega$ and $x^* \in \Omega^*$<br>$x^* = \int_{\Omega^*} \xi . p_x(\xi) d\xi = \frac{\int_{\Omega^*} \xi . e^{-\langle \xi, x \rangle} d\xi}{\int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi}$<br>$x^* = \frac{\partial \Phi(x)}{\partial x}$ and $x = \frac{\partial \Phi^*(x^*)}{\partial x^*}$ | $\beta \in \mathfrak{g}$ and $Q \in \mathfrak{g}^*$<br>$Q = \int_M U(\xi) . p_\beta(\xi) d\omega = \frac{\int_M U(\xi) e^{-\beta . U(\xi)} d\omega}{\int_M e^{-\beta . U(\xi)} d\omega}$<br>$\beta$ : Souriau Geometric Temperature<br>$U$ : Souriau Moment map<br>$Q$ : Mean of Souriau Moment Map or Geometric heat<br>$Q = \frac{\partial \Phi}{\partial \beta}$ and $\beta = \frac{\partial s}{\partial Q}$ |
| Hessian Metric | $ds^2 = -d^2 \Phi(x)$ | $ds^2 = -d^2 \Phi(\beta)$ |
| Fisher metric | $I(x) = -E_\xi \left[ \frac{\partial^2 \log p_x(\xi)}{\partial x^2} \right]$<br>$I(x) = = -\frac{\partial^2 \Phi(x)}{\partial x^2} = \frac{\partial^2 \log \int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi}{\partial x^2}$ | $I(\beta) = -E_\xi \left[ \frac{\partial^2 \log p_\beta(\xi)}{\partial \beta^2} \right]$<br>$I(\beta) = -\frac{\partial^2 \Phi(\beta)}{\partial \beta^2} = \frac{\partial^2 \log \int_M e^{-\beta . U(\xi)} d\omega}{\partial \beta^2}$<br>$I(\beta) = -\frac{\partial^2 \Phi(\beta)}{\partial \beta^2} = -\frac{\partial Q}{\partial \beta}$<br>$K = -\frac{\partial Q}{\partial \beta}$ : Souriau Geometric Capacity |

*6.2. Invariances in Koszul and Souriau Models*

We have observed in previous chapters the main invariances characterizing the Koszul Model and the Souriau Model. We will synthetize these invariances in Table 2.

In both the Koszul and Souriau models, the Information Geometry Metric and the Entropy are invariant respectively to the automosphisms $g$ of the convex cone $\Omega$ and to $\overline{a}_{\mathfrak{g}}$ adjoint

representation of Dynamical group $G$ acting on $\Omega$, the convex cone considered as largest open subset of $\mathfrak{g}$, Lie algebra of G, such that $\int_M e^{-\beta.U(\xi)}d\omega$ and $\int_M \xi.e^{-\beta.U(\xi)}d\omega$ are convergent integrals.

*6.3. Souriau Thermometer*

Souriau has built a thermometer (θερμός) device principle that could measure the Geometric Temperature using "*Relative Ideal Gas Thermometer*" based on a theory of Dynamical Group Thermometry, and has also recovered the Laplace barometric law $p(\vec{r}) \propto e^{-m\beta\langle \vec{g},\vec{r}\rangle}$.

**Table 2.** Comparison of invariances for the Koszul and Souriau models.

| | **Koszul Information Geometry Model** | **Souriau Lie Groups Thermodynamics Model** |
|---|---|---|
| Convex Cone | $x \in \Omega$<br>$\Omega$ *convex cone* | $\beta \in \Omega$<br>$\Omega$ *convex cone: largest open subset of* $\mathfrak{g}$, *Lie algebra of G, such that* $\int_M e^{-\beta.U(\xi)}d\omega$ *and* $\int_M \xi.e^{-\beta.U(\xi)}d\omega$ *are convergent integrals* |
| Transformation | $x \to gx \quad \text{with} \quad g \in Aut(\Omega)$ | $\beta \to \overline{a}_{\mathfrak{g}}(\beta)$ |
| Transformation of Potential (non invariant) | $\Phi_\Omega(x) \to \Phi_\Omega(gx) = \Phi_\Omega(x) + \log\left(\lvert\det g\rvert\right)$ | $\Phi(\beta) \to \Phi\left(\overline{a}_{\mathfrak{g}}(\beta)\right) = \Phi(\beta) - \theta\left(a^{-1}\right)\beta$ |
| Transformation of Entropy (invariant) | $\Phi^*_{\Omega^*}(x^*) \to \Phi^*_{\Omega^*}\left(\frac{\partial \Phi_\Omega(gx)}{\partial x}\right) = \Phi^*_{\Omega^*}(x^*)$<br>with $x^* = \frac{\partial \Phi_\Omega(x)}{\partial x}$ | $s(Q) \to s'(Q') = \beta'.Q' - \Phi' = \beta.Q - \Phi = s(Q)$<br>.with<br>$\beta' = \overline{a}_{\mathfrak{g}}(\beta)$<br>$Q' = \frac{\partial \Phi'}{\partial \beta'} = \frac{\partial\left(\Phi + \theta(a)\overline{a}_{\mathfrak{g}}(\beta)\right)}{\partial \overline{a}_{\mathfrak{g}}(\beta)} = \overline{a}_{\mathfrak{g}^*}(Q) + \theta(a)$<br>$\Phi' = \Phi(\beta') = \Phi\left(\overline{a}_{\mathfrak{g}}(\beta)\right) = \Phi(\beta) - \theta\left(a^{-1}\right)\beta$ |
| Information Geometry Metric (invariant) | $I(gx) = -\frac{\partial^2\left[\Phi_\Omega(x) + \log\left(\lvert\det g\rvert\right)\right]}{\partial x^2}$<br>$I(gx) = -\frac{\partial^2 \Phi_\Omega(x)}{\partial x^2} = I(x)$ | $I\left(\overline{a}_{\mathfrak{g}}(\beta)\right) = -\frac{\partial^2\left[\Phi(\beta) - \theta\left(a^{-1}\right)\beta\right]}{\partial \beta^2} = -\frac{\partial^2 \Phi(\beta)}{\partial \beta^2} = I(\beta)$ |

## 7. From Characteristic Function to Generative Inner Product

Cartan's works have greatly influenced Koszul (Koszul's PhD thesis extended previous work of Cartan) and Souriau (Souriau was a student of Elie Cartan at ENS, the year after his aggregation). We have shown that "Information Geometry" could be considered as a particular application domain of Hessian Geometry through Koszul's work (Koszul-Vinberg metric deduced from the associated characteristic function having the main property of being invariant to all automorphisms of the convex cone), that could be extended in the framework of Souriau's theory, as an extension towards "Lie Group Thermodynamics" with vector-valued geometric temperature (providing a geometric extension of Noether's theorem). Should we deduce that the "essence" of Information Geometry is limited to the

"Koszul Characteristic Function"? This notion seems to not be the more general one, and we will explore the notion of Generative Inner Products. We will reduce Koszul's and Souriau's definitions to exclusive "Inner Product" selection using symmetric bilinear "Cartan-Killing form" introduced by Cartan in 1894.

In Koszul Geometry, we have two convex dual functions $\Phi(x)$ and $\Phi^*(x^*)$ with dual system of coordinates $x$ and $x^*$ defined on dual cones $\Omega$ and $\Omega^*$: $\Phi(x) = -\log \int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi \quad \forall x \in \Omega$ and $\Phi^*(x^*) = \langle x, x^* \rangle - \Phi(x)$. We can then remark that if we can define an Inner Product $\langle .,. \rangle$, we will be able to build a convex function $\Phi(x) = -\log \psi_\Omega(x)$ and its dual by Legendre transform because both are only dependent of the Inner product, and dual coordinate is also defined by $x^* = \arg\min \{ \psi_\Omega(y) / y \in \Omega^*, \langle x, y \rangle = n \} = \int_{\Omega^*} \xi . e^{-\langle \xi, x \rangle} d\xi \,/ \int_{\Omega^*} e^{-\langle \xi, x \rangle} d\xi$ where $x^*$ is also the center of gravity of the cross section $\{ y \in \Omega^*, \langle x, y \rangle = n \}$ of $\Omega^*$ (with notation: $\Phi(x) = -\log \psi_\Omega(x)$).

It is not possible to define an ad(g)-invariant inner product for any two elements of a Lie Algebra, but a symmetric bilinear form, called "Cartan-Killing form", could be introduced. This form has been introduced first by Cartan in 1894 in his PhD thesis. This form is defined according to the adjoint endomorphism $Ad_x$ of $g$ that is defined for every element $x$ of $g$ with the help of the Lie bracket:

$$Ad_x(y) = [x, y] \tag{120}$$

The trace of the composition of two such endomorphisms defines a bilinear form, the Cartan-Killing form:

$$B(x, y) = Tr\left( Ad_x Ad_y \right) \tag{121}$$

The Cartan-Killing form is symmetric:

$$B(x, y) = B(y, x) \tag{122}$$

and has the associativity property:

$$B([x, y], z) = B(x, [y, z]) \tag{123}$$

given by:

$$B([x, y], z) = Tr\left( Ad_{[x,y]} Ad_z \right) = Tr\left( [Ad_x, Ad_y] Ad_z \right) = Tr\left( Ad_x [Ad_y, Ad_z] \right) = B(x, [y, z]) \tag{124}$$

Elie Cartan has proved that if $g$ is a simple Lie algebra (the Killing form is non-degenerate) then any invariant symmetric bilinear form on $g$ is a scalar multiple of the Cartan-Killing form. The Cartan-Killing form is invariant under automorphisms $\sigma \in Aut(g)$ of the algebra $g$:

$$B(\sigma(x), \sigma(y)) = B(x, y) \tag{125}$$

To prove this invariance, we have to consider:

$$\begin{cases}\sigma[x,y]=[\sigma(x),\sigma(y)]\\ z=\sigma(y)\end{cases}\Rightarrow\sigma\left[x,\sigma^{-1}(z)\right]=[\sigma(x),z]\quad\text{rewritten}\quad Ad_{\sigma(x)}=\sigma\circ Ad_x\circ\sigma^{-1}\tag{126}$$

Then:

$$B\left(\sigma(x),\sigma(y)\right)=Tr\left(Ad_{\sigma(x)}Ad_{\sigma(y)}\right)=Tr\left(\sigma\circ Ad_xAd_y\circ\sigma^{-1}\right)=Tr\left(Ad_xAd_y\right)=B(x,y)\tag{127}$$

A natural *G*-invariant inner product could be then introduced by Cartan-Killing form.

*Cartan Generative Inner Product:* The following Inner product defined by Cartan-Killing form is invariant by automorphisms of the algebra

$$\langle x,y\rangle=-B\left(x,\theta(y)\right)\tag{128}$$

where $\theta\in g$ is a Cartan involution (an involution on $g$ is a Lie algebra automorphism $\theta$ of $g$ whose square is equal to the identity).

From the Cartan Inner Product, we can generate logarithm of the Koszul Characteristic Function, and its Legendre Transform to define Koszul Entropy, Koszul Density and Koszul Metric, as explained in the following Figure 5:

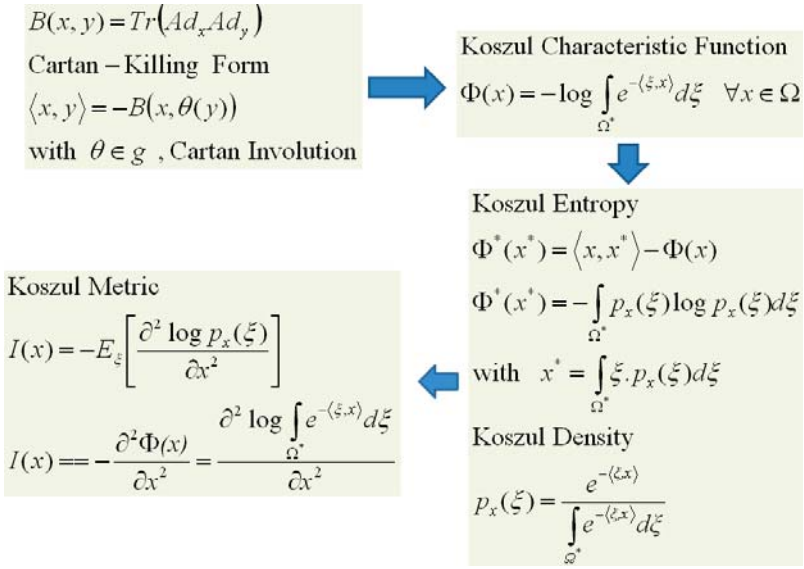



**Figure 5.** Generation of Koszul elements from Cartan Inner Product.

In Appendix A2, we give the definition of another inner product, Gromov Inner product, in CAT(−1) space, that could be also used to generalize Koszul definition of Characteristic Function.

On the concept of generative structure, we could also explore the notion of Generative Function [123–126] and come back to seminal paper of Chentsov about axiomatization of Information Geometry [127].

## 8. Conclusions on General Definition of Entropy by Legendre Transform

Definition of Entropy has been widely debated [128,129]. Based on the cornerstone concept of the Koszul Vinberg Characteristic Function, we have introduced Koszul Entropy as the Legendre transform of its logarithm. This definition of Entropy could be extended by interpretating Legendre transform as Fourier transform in (Min,+) algebra [130,131].

As we have observed previously, Koszul Entropy has a Shannon Entropy structure:

$$\begin{aligned}
&\Phi^*(x^*) = \Phi^*(E[\xi]) = -\int_{\Omega^*} p_x(\xi)\log p_x(\xi)d\xi = \int_{\Omega^*} \Phi^*(\xi)p_x(\xi)d\xi = E\left[\Phi^*(\xi)\right] \\
&\text{with}\quad \Phi^*(\xi) = -\log p_x(\xi) \text{ and } x^* = \int_{\Omega^*} \xi.p_x(\xi)d\xi = E[\xi] \\
&\text{where } p_x(\xi) = \frac{e^{-\langle \xi,x\rangle}}{\int_{\Omega^*} e^{-\langle \xi,x\rangle}d\xi} = e^{-\langle x,\xi\rangle - \log\int_{\Omega^*} e^{-\langle \xi,x\rangle}d\xi} = e^{-\langle x,\xi\rangle + \Phi(x)}
\end{aligned} \tag{129}$$

In last equation, variable $x$ could be defined by $\overline{\xi} = E[\xi] = x^*$ if function $\frac{d\Phi(x)}{dx}$ could be inverted:

$$p_{\overline{\xi}}(\xi) = \frac{e^{-\langle \xi,\Theta^{-1}(\overline{\xi})\rangle}}{\int_{\Omega^*} e^{-\langle \xi,\Theta^{-1}(\overline{\xi})\rangle}d\xi} \text{ with } x = \Theta^{-1}(\overline{\xi}) \text{ and } \overline{\xi} = \Theta(x) = \frac{d\Phi(x)}{dx} \tag{130}$$

where:

$$\overline{\xi} = \int_{\Omega^*} \xi.p_{\overline{\xi}}(\xi)d\xi \text{ and } \Phi(x) = -\log\int_{\Omega^*} e^{-\langle x,\xi\rangle}d\xi \tag{131}$$

In previous chapters, a definition of Koszul Entropy $\Phi^*(x^*)$ through Legendre transform of Koszul-Vinberg characteristic function $\Phi(x)$ has been given:

$$\begin{aligned}
&\Phi^*(x^*) = \langle x,x^*\rangle - \Phi(x) \\
&\text{with}\quad \Phi(x) = -\log\int_{\Omega^*} e^{-\langle \xi,x\rangle}d\xi \quad \forall x \in \Omega
\end{aligned} \tag{132}$$

where $\Phi(x)$ could be interpreted as opposite of logarithm of Laplace transform [132,133]:

$$Entropy = Legendre\left[-\log\left[Laplace\right]\right] \tag{133}$$

that we will write synthetically as:

$$Ent = -Leg \circ Log \circ Lapl \tag{134}$$

The function $Leg \circ Log \circ Lapl$ is sometimes called "Cramer transform".

If we remark that the Legendre transform is closely related to the idempotent analogue of the Fourier transform [130,131,134–136], we could then give a new definition of Entropy.

If we consider the semiring $R_{\min} = R \cup \{+\infty\}$ with the operations $\oplus = Min$ and $\bullet = +$. In $R_{\min} = R \cup \{+\infty\}$ the idempotent analogues of integration on $R^N$ is given by the formula:

$$I(f)=\overset{\oplus}{\int_{R^N}} f(x)dx=\underset{x\in R^N}{Inf}\, f(x) \tag{135}$$

Then, the Legendre transform is equivalent to the Fourier transform in $(\oplus,\bullet)=(Min,+)$ algebra [130]:

$$\Phi^*(\xi)=\underset{x\in\Omega}{Sup}\left[\langle x,\xi\rangle-\Phi(x)\right]=-\overset{\oplus}{\int_{\Omega}}\left(-\langle x,\xi\rangle\right)\bullet\Phi(x)dx=Four_{(Min,+)}\left[\Phi(x)\right] \tag{136}$$

The Legendre transform generates an idempotent version of harmonic analysis for the space of convex functions. We can then give a general definition of Entropy:

$$Ent=-Four_{(Min,+)}\circ Log\circ Lapl_{(+,\times)} \tag{137}$$

We can also observe the following properties deduced from the Laplace and Legendre transforms' characteristics:

$$Ent(\mu\otimes\gamma)=Ent(\mu)\bullet Ent(\gamma) \tag{138}$$

where * is the convolution operator and $\otimes$ the inf-convolution operator (see [130] for the definition of inf-convolution) defined by:

$$[f\bullet g](z)=\underset{x}{Inf}\left[f(x)+g(y-x)\right] \tag{139}$$

with $f$ and $g$, two functions $R\to R_{\min}$.

"*La théorie cinétique des gaz laisse encore subsister bien des points embarrassants pour ceux qui sont accoutumés à la rigueur mathématique… L'un des points qui m'embarrassaient le plus était le suivant: il s'agit de démontrer que l'entropie va en diminuant, mais le raisonnement de Gibbs semble supposer qu'après avoir fait varier les conditions extérieures on attend que le régime soit établi avant de les faire varier à nouveau. Cette supposition est-elle essentielle, ou en d'autres termes, pourrait-on arriver à des résultats contraires au principe de Carnot en faisant varier les conditions extérieures trop vite pour que le régime permanent ait le temps de s'établir?* "

Henri Poincaré « Réflexions sur la théorie cinétique des gaz », 1906

[*The kinetic theory of gases leaves awkward points for those who are accustomed to mathematical rigor … One of the points which embarrassed me most was the following one: it is a question of demonstrating that the entropy keeps decreasing, but the reasoning of Gibbs seems to suppose that having made vary the outside conditions we wait that the regime is established before making them vary again. Is this supposition essential, or in other words, we could arrive at opposite results to the principle of Carnot by making vary the outside conditions too fast so that the permanent regime has time to become established* ?]

Henri Poincaré "Reflection on The kinetic theory of gases", 1906

"*Quel est l'objet de l'art ? Si la réalité venait frapper directement nos sens et notre conscience, si nous pouvions entrer en communication immédiate avec les choses et avec nous-mêmes, je crois bien que l'art serait inutile, ou plutôt que nous serions tous artistes, car notre âme vibrerait alors continuellement à l'unisson de la nature.* "

Henri Bergson, Le rire, p.115, Éd. P.U.F

[*What is the object of art? Could reality come into direct contact with sense and consciousness, could we enter into immediate communion with things and with ourselves, probably art would be useless, or rather we should all be artists, for then our soul would continually vibrate in perfect accord with nature.*]

Henri Bergson, Laughter

**Acknowledgments**

Many thanks are due to members of the Leon Brillouin seminar with very fruitful discussions on Geometric Science of Information and Information Geometry, initiated since December 2009. Souriau Models were more clearly understood with the help of Claude Vallée that sent me unpublished chapters of Souriau's book [86]. Jean-Louis Koszul has participated to the first GSI conference at Ecole des Mines in 2013, where Hirohiko Shima gave a keynote lecture on Koszul Hessian Geometry.

**Appendix**

*A1. Legendre Transform and Minimal Surface*

Laplace contribution to probability was around 1774 [137]. At almost the same period, in 1787, Adrien-Marie Legendre has introduced the "Legendre Transform" [138] to solve the Minimal Surface Problem equation introduced by Lagrange and partially solved by Gaspard Monge in 1784 [139]. In 1768, Lagrange considered the variational problem of least area surface stretched across a given closed contour. Based on Euler-Lagrange equation, Lagrange has introduced the equation of *Minimal Surface* $z(x,y)$:

$$\left(1+q^2\right)\frac{d^2z}{dx^2}-2pq\frac{d^2z}{dxdy}+\left(1+p^2\right)\frac{d^2z}{dy^2}=0 \quad \text{with} \quad \frac{dz}{dx}=p \quad \text{and} \quad \frac{dz}{dy}=q \tag{140}$$

Lagrange has observed that affine functions $z(x,y)=a.x+b.y+c$ are solutions of this equation and minimal surfaces are planes.

Jean-Baptiste Marie Meusnier de La Place, a student of Monge, has observed that for this surface, two curvature radiuses are everywhere equal but directed in opposite direction, because first equation is equal to two times the mean curvature $H_z$:

$$2H_z = \frac{d}{dx}\left(\frac{\frac{dz}{dx}}{\sqrt{1+\left(\frac{dz}{dx}\right)^2+\left(\frac{dz}{dy}\right)^2}}\right)+\frac{d}{dy}\left(\frac{\frac{dz}{dy}}{\sqrt{1+\left(\frac{dz}{dx}\right)^2+\left(\frac{dz}{dy}\right)^2}}\right)=\frac{\left(1+q^2\right)\frac{d^2z}{dx^2}-2pq\frac{d^2z}{dxdy}+\left(1+p^2\right)\frac{d^2z}{dy^2}}{\left(1+\left(\frac{dz}{dx}\right)^2+\left(\frac{dz}{dy}\right)^2\right)^{3/2}} \quad (141)$$

Gaspard Monge integrated this equation in [139] but with a non-rigorous approach and asked Legendre to find a more classical solution. For this task, Legendre has introduced a change of variable that is the nowadays well-known "Legendre transform". Adrien-Marie Legendre said "*J'y suis parvenu simplement par un changement de variables qui peut être utile dans d'autres occasions*"(*"I reached there simply by a change of variables which can be useful in other opportunities"*).

Legendre reduced the problem to solve to determine $p$ and $q$ as functions of $x$ and $y$ such that:

$$p.dx+q.dy \quad \text{and} \quad \frac{p.dy-q.dx}{\sqrt{1+p^2+q^2}} \quad (142)$$

are exact differentials. If we set $1+p^2+q^2=u^2$, then these other expressions are complete differentials:

$$x.dp+y.dq \quad \text{and} \quad y.d\left(\frac{p}{u}\right)+x.d\left(\frac{q}{u}\right) \quad (143)$$

Legendre considered $x$ and $y$ as functions of $p$ and $q$:

$$x.dp+y.dq=d\omega \quad \text{with} \quad x=\frac{d\omega}{dp} \quad \text{and} \quad y=\frac{d\omega}{dq} \quad (144)$$

If we then develop $y.d\left(\frac{p}{u}\right)+x.d\left(\frac{q}{u}\right)$, we have:

$$\left[\left(1+q^2\right)y+pq.x\right]\frac{dp}{u^3}-\left[\left(1+p^2\right)x+pq.y\right]\frac{dq}{u^3} \quad (145)$$

That should be an exact differential. By replacing $x$ and $y$, we have a new equation:

$$\left(1+q^2\right)\frac{d^2\omega}{dq^2}+2pq.\frac{d^2\omega}{dpdq}+\left(1+p^2\right)\frac{d^2\omega}{dp^2}=0 \quad (146)$$

This new equation is very similar to the previous one, but simpler because it depends on $p$ and $q$ and not their partial differentials of first order. When the function $\omega$ will be known, then functions $x$, $y$ and $z$ will be also defined according to $p$ and $q$ thanks to "*Legendre transform*":

$$z(x,y)=p.x+q.y-\omega(p,q)$$
$$\text{with} \quad x=\frac{d\omega}{dp} \quad \text{and} \quad y=\frac{d\omega}{dq} \quad (147)$$

About this Legendre transform, Darboux [140] gave an interpretation by Chasles "*Ce qui revient suivant une remarque de M. Chasles, à substituer à la surface sa polaire réciproque par rapport à*

*un paraboloïde*" [*What is equivalent according to M. Chasles's remark, to substitute for the surface its mutual polar with regard to a paraboloid*]. This equation could be also written as classical "*Legendre transform*" with our previous notations:

$$s(Q)=\beta.Q-\Phi(\beta)=\langle\beta,Q\rangle-\Phi(\beta)$$

$$\text{with}\quad \begin{cases}\Phi(\beta)=z(x,y)\\ s(Q)=\omega(p,q)\end{cases},\quad \begin{cases}Q=\begin{bmatrix}Q_1\\Q_2\end{bmatrix}=\begin{bmatrix}p\\q\end{bmatrix}\\ \beta=\begin{bmatrix}\beta_1\\\beta_2\end{bmatrix}=\begin{bmatrix}x\\y\end{bmatrix}\end{cases}\quad\text{and}\quad \begin{cases}Q=\begin{bmatrix}\frac{d\Phi}{d\beta_1}\\ \frac{d\Phi}{d\beta_2}\end{bmatrix}=\frac{d\Phi}{d\beta}\\ \beta=\begin{bmatrix}\frac{ds}{dQ_1}\\ \frac{ds}{dQ_2}\end{bmatrix}=\frac{ds}{dQ}\end{cases} \tag{148}$$

In the following relation, we recover the definition of Entropy $ds=\beta.dQ=\frac{dQ}{T}$:

$$\begin{cases}x.dp+y.dq=d\omega\\ x=\frac{d\omega}{dp}\quad\text{and}\quad y=\frac{d\omega}{dq}\end{cases}\Rightarrow\begin{cases}\beta.dQ=ds\\ \beta=\frac{ds}{dQ}\end{cases} \tag{149}$$

The equation of the surface is characterized by the following equation:

$$\frac{d}{d\beta}\left(\frac{Q}{\sqrt{1+\|Q\|^2}}\right)=\frac{d}{d\beta}\left(\frac{\frac{d\Phi}{d\beta}}{\sqrt{1+\left\|\frac{d\Phi}{d\beta}\right\|^2}}\right)=2.H_\Phi\quad\text{or}\quad div_\beta\left(\frac{Q}{\sqrt{1+\|Q\|^2}}\right)=div_\beta\left(\frac{\nabla\Phi}{\sqrt{1+\|\nabla\Phi\|^2}}\right)=2.H_\Phi \tag{150}$$

We can then observed that when $\|Q\|<<1$, $\frac{d}{d\beta}\left(\frac{Q}{\sqrt{1+\|Q\|^2}}\right)\approx\frac{dQ}{d\beta}=-I(\beta)=2.H_\Phi$.

We can also characterized Entropy with this 2nd equation:

$$\left(1+Q_2^2\right)\frac{d^2s(Q)}{dQ_2^2}+2pq.\frac{d^2s(Q)}{dQ_1dQ_2}+\left(1+Q_1^2\right)\frac{d^2s(Q)}{dQ_1^2}=0 \tag{151}$$

We can also find direct equations for *x*, *y* and *z*, based on "Legendre transform" and Equation (146):

$$\left(1+q^2\right)\frac{d^2x}{dq^2}+2pq.\frac{d^2x}{dpdq}+\left(1+p^2\right)\frac{d^2x}{dp^2}+2q\frac{dx}{dp}+2p\frac{dx}{dq}=0 \tag{152}$$

We have exactly same equations for *y* and *z*.

Legendre then solved Equations (145) and (148), by determining two constants *a* and *b* given by double integral of the equation:

$$\left(1+q^2\right)dp^2-2pq.dp.dq+\left(1+p^2\right)dp^2=0 \tag{153}$$

By selecting $p=aq+A$ with a and A two constants. Previous equation gives $1+a^2+A^2=0$. Then *a* will be let an arbitrary function and $A=\pm\sqrt{-1-a^2}$. Two integrals of Equation (129) will be:

$$\begin{cases} p=aq+\sqrt{-1-a^2}=aq+A \\ p=bq-\sqrt{-1-b^2}=bq+B \end{cases} \tag{154}$$

with *a* and *b* two arbitrary constants, roots of the following Equation:

$$\begin{gathered} \left(1+q^2\right)v^2-2pq..v+\left(1+p^2\right)=0 \\ \text{with} \quad \begin{cases} a+b=\dfrac{2pq}{1+q^2} \\ ab=\dfrac{1+p^2}{1+q^2} \end{cases} \end{gathered} \tag{155}$$

Equations (145) and (148) could be then simplified:

$$\begin{gathered} (a-b)\frac{d^2\omega}{dadb}-\frac{A}{B}.\frac{d\omega}{da}+\frac{B}{A}.\frac{d\omega}{db}=0 \\ \frac{d^2x}{dadb}=0 \end{gathered} \tag{156}$$

Then Legendre deduced that three coordinates could be given by two arbitrary functions:

$$\begin{cases} x=\dfrac{d\varphi}{da}+\dfrac{d\psi}{db} \\ y=\varphi-a\dfrac{d\varphi}{da}+\psi-b\dfrac{d\psi}{db} \\ z=-\int A\dfrac{d^2\varphi}{da^2}da+\int B\dfrac{d^2\psi}{db^2}db \end{cases} \tag{157}$$

This is the integral solution of "*Minimal surface*" Lagrange equation (Legendre recovered the solution given by Monge in 1784).

*A2. Gromov Inner Product*

As other generalization of inner product, we can consider for specific case *CAT(-1)-space*[141,142] (generalization of simply connected Riemannian manifold of negative curvature lower than unity) or for an Homogeneous Symmetric Bounded domains, a "generative" Gromov Inner Product between points y–z (relatively to *x*) that is defined by the distance [143,144]:

$$\langle y,z\rangle_x=\frac{1}{2}\left(d(x,y)+d(x,z)-d(y,z)\right) \tag{158}$$

with *d(.,.)* the distance in CAT(−1). This Gromov Inner Product is illustrated in Figure 6. Intuitively, this inner product measures the distance of *x* to the geodesics between *y* to *z*. This Inner

product could be also defined for points on the Shilov Boundary of the domain through Busemann distance:

$$\langle \xi, \xi' \rangle_x = \frac{1}{2}\left(B_\xi(x,p) + B_{\xi'}(x,p)\right) \tag{159}$$

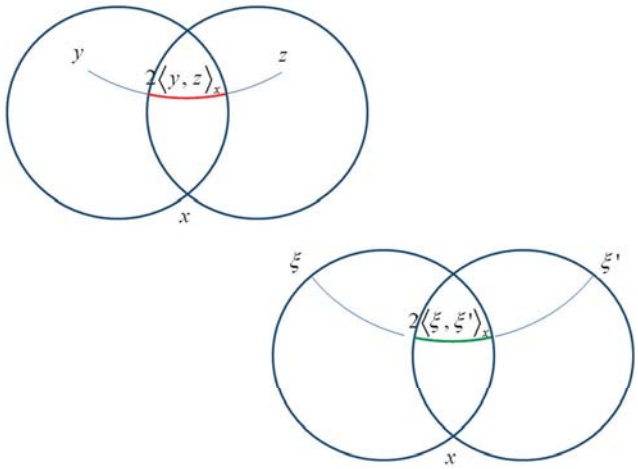



**Figure 6.** Gromov Inner product in homogeneous bounded domains and its Shilov boundary.

Independent of *p*, where $B_\xi(x,y) = \underset{t\to+\infty}{Lim}\left[|x-r(t)|-|y-r(t)|\right]$ is the horospheric distance, from *x* to *y* relatively to $\xi$ , with $r(t)$ geodesic ray. We have the property that:

$$\langle \xi, \xi' \rangle_x = \underset{\substack{y\to\xi \\ y'\to\xi'}}{Lim} \langle y, y' \rangle_x \tag{160}$$

We can then define a visual metric on the Shilov boundary by:

$$\begin{aligned} d_x(\xi,\xi') &= e^{-\langle \xi,\xi' \rangle_x} \quad \text{if} \quad \xi \neq \xi' \\ d_x(\xi,\xi') &= 0 \quad \text{otherwise} \end{aligned} \tag{161}$$

We can then define the characteristic function according to the origin *0*:

$$\Phi(x) = -\log \int_{\Omega^*} e^{-\langle x,\gamma \rangle_0} d\gamma \text{ or } \Phi_\Omega(x) = -\log \int_{\Omega^*} e^{-\frac{1}{2}(d(0,x)+d(0,\gamma)-d(x,\gamma))} d\gamma \tag{162}$$

and:

$$\Phi^*(x^*) = \langle x, x^* \rangle_0 - \Phi(x) = \frac{1}{2}\left(d(0,x) + d(0,x^*) - d(x,x^*)\right) - \Phi(x) \tag{163}$$

$$d(x,x^*) = \left(d(0,x^*) - 2\Phi^*(x^*)\right) + \left(d(0,x) - 2\Phi(x)\right) \tag{164}$$

with the center of gravity:

$$x^* = \int_{\Omega^*} \gamma . e^{-\langle x,\gamma \rangle_0} d\gamma \Big/ \int_{\Omega^*} e^{-\langle x,\gamma \rangle_0} d\gamma \tag{165}$$

All these relations are also true on the Shilov Boundary:

$$\Phi(\xi) = -\log \int_{\partial\Omega^*} e^{-\langle \xi,\xi' \rangle_0} d\xi' = -\log \int_{\partial\Omega^*} d_0(\xi,\xi').d\xi' \tag{166}$$

where $\int_{\partial\Omega^*} d_0(\xi,\xi').d\xi'$ is the functional of Busemann barycenter on the Shilov Boundary $\partial\Omega^*$ (existence and unicity of this barycenter have been proved by Cartan [14] for Cartan-Hadamard Spaces).

*A3. The Cohomology of a Dynamical Group*

In the following, we give some details of Souriau development about the Moment of the *G* action (see Figure 7) and the Cohomology of a dynamical group (see Figure 8). Other details about Symplectic Geometry could be found in [145] or [146].

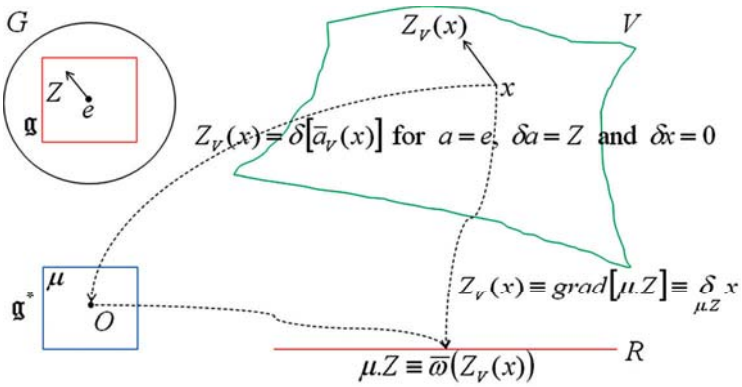



**Figure 7.** Moment of the *G* action.

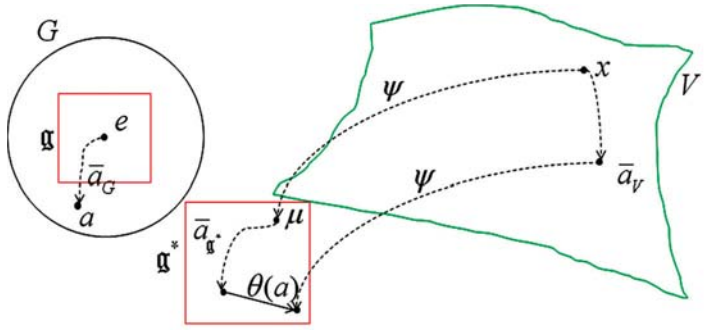



**Figure 8.** The Cohomology of a dynamical group.

If $G$ is a dynamical group of a symplectic manifold $V$, torsor $\mu$ is called a moment of the $G$-action, if there is a differential map $x \mapsto \mu$ from $V$ to $\mathfrak{g}^*$ such that:

$$\sigma(Z_V(x)) \equiv -d[\mu.Z] \tag{167}$$

To every torsor $\mu$, there corresponds a field $[x \mapsto \varpi]$ of 1-forms (Maurer-Cartan forms) on $G$ which is invariant under right translation and which takes the value $\mu$ when $x$ is the identity element:

$$\sigma_V = d\varpi \tag{168}$$

Using the moment of the G action, Souriau has introduced the following theorem on the Cohomology of a dynamical group:

*Theorem*. Let $V$ be a connected symplectic manifold and let $G$ be a dynamical group of $V$ possessing a moment $\mu$. Finally, let $\psi$ denote the map $x \mapsto \mu$ from $V$ to the space $\mathfrak{g}^*$ of torsor of $G$:

There exists a differential map $\theta : G \to \mathfrak{g}^*$:

$$\theta(a) \equiv \psi(\overline{a}_V(x)) - \overline{a}_{\mathfrak{g}^*}(\psi(x)) \tag{169}$$

The derivative $f = D(\theta)(e)$ is a 2-form on the Lie algebra $\mathfrak{g}$ of $G$:

$$f(Z)([Z', Z'']) + f(Z')([Z'', Z]) + f(Z'')([Z, Z']) \equiv 0 \tag{170}$$

Then, the following identities hold:

$$\sigma(Z_V(x))(Z'_V(x)) \equiv \mu.[Z, Z'] + f(Z)(Z') \tag{171}$$

$$D(\psi)(x)(Z_V(x)) \equiv \psi(x).ad(Z) + f(Z) \tag{172}$$

## Conflicts of Interest

The author declares no conflict of interest.

## References


1. Massieu, F. Sur les Fonctions caractéristiques des divers fluides. *Comptes Rendus de l'Académie des Sciences* **1869**, *69*, 858–862. (In French)
2. Massieu, F. Addition au précédent Mémoire sur les Fonctions caractéristiques. *Comptes Rendus de l'Académie des Sciences* **1869**, *69*, 1057–1061. (In French)
3. Massieu, F. Thermodynamique: Mémoire sur les Fonctions Caractéristiques des Divers Fluides et sur la Théorie des Vapeurs; Académie des Sciences: Paris, France, 1876; p. 92. (In French)
4. Duhem, P. Sur les équations générales de la Thermodynamique. *Annales Scientifiques de l'Ecole Normale Supérieure* **1891**, *8*, 231–266. (In French)

5. Duhem, P. Commentaire aux principes de la Thermodynamique—Première partie. *Journal de Mathématiques pures et appliquées* **1892**, *8*, 269–330. (In French)
6. Duhem, P. Commentaire aux principes de la Thermodynamique—Troisième partie. *Journal de Mathématiques pures et appliquées* **1894**, *10*, 207–286. (In French)
7. Duhem, P. Les théories de la chaleur. *Revue des deux Mondes* **1895**, *130*, 851–868.
8. Gibbs, J.W. *Graphical Methods in the Thermodynamics of Fluids*. In *The Scientific Papers of J. Willard Gibbs*; Bumstead, H.A., van Name, R.G.; Eds.; Dover: New York, NY, USA, 1961.
9. Poincaré, H. *Calcul des Probabilités*; Gauthier-Villars: Paris, France, 1896. (In French)
10. Poincaré, H. *Thermodynamique, Cours de Physique Mathématique*; Carré, G., Ed.; 1892; Available online: http://gallica.bnf.fr/ark:/12148/bpt6k2048983 (accessed on 30 July 2014; In French)
11. Vinberg, E.B. Structure of the Group of Automorphisms of a Homogeneous Convex Cone. *Trudy Moskovskogo Matematicheskogo Obshchestva* **1965**, *13*, 56–83.
12. Faraut, J.; Koranyi, A. *Analysis on Symmetric Cones, Oxford Mathematical Monographs*; The Clarendon Press; Oxford University Press: New York, NY, USA, 1994.
13. Lichnerowicz, A. Espaces homogènes Kähleriens. In *Colloque de Géométrie Différentielle*; Pub. du CNRSP: Paris, France, 1953; pp. 171–184. (In French)
14. Cartan, E. Sur les domaines bornés de l'espace de n variables complexes. *Abh. Math. Seminar Hamburg* **1935**, *1*, 116–162. (In French)
15. Siegel, C.L. Über der analytische Theorie der quadratischen Formen. *Ann. Math.* **1935**, *36*, 527–606. (In German)
16. Siegel, C.L. Symplectic geometry. *Amer. J. Math*. **1943**, *65*, 1–86.
17. Marle, C.M. On mechanical systems with a Lie group as configuration space. In *Jean Leray'99 Conference*, Proceedings of the Karlskrona Conference in the Honor of Jean Leray, Kluwer, Dordrecht, The Netherlands, 2003; de Gosson, M., Ed.; Springer: Berlin/Heidelberg, Germany, 2003; pp. 183–203.
18. Marle, C.M. On Henri Poincaré's note "Sur une forme nouvelle des équations de la mécanique". *J. Geom. Symmetry Phys.* **2013**, *29*, 1–38.
19. Friedrich, T. Die Fisher-Information und symplektische Strukturen. *Math. Nachr*. **1991**, *153*, 273–296. (In German)
20. Gromov, M. In a Search for a Structure, Part 1: On Entropy. Available online: http://www.ihes.fr/~gromov/PDF/structre-serch-entropy-july5-2012.pdf (accessed on 23 June 2013).
21. Gromov, M. Convex Sets and Kähler Manifolds. In *Advances in Differential Geometry and Topology*; Tricerri, F., Ed.; World Scientific: Singapore, Singapore, 1990; pp. 1–38.
22. Gromov, M. Entropy and Isoperimetry for Linear and non-Linear Group Actions. *Groups Geom. Dyn*. **2008**, *2*, 499–593.
23. Gromov, M. Carnot-Carathéodory spaces seen from within. In *Progress in Mathematics*; Springer: Berlin/Heidelberg, Germany, 1996; Volume 144.

24. Ollivier, Y.; Akimoto, Y. Objective Improvement in Information-Geometric Optimization. In *Foundations of Genetic Algorithms XII*; Springer: Berlin/Heidelberg, Germnay, 2013.
25. Bensadon, J. Black-box optimization using geodesics in statistical manifolds. **2013**, arXiv:1309.7168.
26. Bennequin, D. Dualités de champs et de cordes. *Séminaire N. Bourbaki* **2003**, *899*, 117–148. (In French)
27. Bennequin, D. Dualité Physique-Géométrie et Arithmétique. Avalable online: http://archive.numdam.org/ARCHIVE/SB/SB_2001-2002__44_/SB_2001-2002__44__117_0/SB_2001-2002__44__117_0.pdf (accessed on 16 June 2014).
28. Chasles, M. Aperçu Historique sur L'origine et le Développement des Méthodes en Géométrie; Gauthier-Villars: Paris, France, 1837. (In French)
29. Gergonne, J.D. Polémique mathématique. Réclamation de M. le capitaine Poncelet (extraite du bulletin universel des annonces et nouvelles scientifiques); avec des notes. *Annales de Mathématiques Pures et Appliquées* **1827**–**1828**, *18*, 125. (In French)
30. Poncelet, J.V. *Traité des propriétés projectives des figures*; Gauthier-Villars: Paris, France, 1822. (In French)
31. André, Y. Dualités, Sixième séance, ENS, Séminaire MaxMux. In *Leçons de Mathématiques contemporaines à l'IRCAM*; ENS: Paris, France, 2009; Chap.6. (In French)
32. Atiyah, M.F. Duality in mathematics and physics. Available online: https://www.fme.upc.edu/arxius/butlleti-digital/riemann/071218_conferencia_atiyah-d_article.pdf (accessed on 20 June 2014).
33. Von Oettingen, A.J. Das duale System der Harmonie (Part 1). *Annalen der Naturphilosophie* **1902**, *1*, 62–75. (In German)
34. Von Oettingen, A.J. Das duale System der Harmonie (Part 2). *Annalen der Naturphilosophie* **1903**, *2*, 375–403. (In German)
35. Von Oettingen, A.J. Das duale System der Harmonie (Part 3). *Annalen der Naturphilosophie* **1904**, *3*, 241–269. (In German)
36. Von Oettingen, A.J. Das duale System der Harmonie (Part 4). *Annalen der Naturphilosophie* **1905**, *4*, 116–152. (In German)
37. Von Oettingen, A.J. Das duale System der Harmonie (Part 5). *Annalen der Naturphilosophie* **1906**, *5* 449–503. (In German)
38. Von Oettingen, A.J. *Das duale Harmoniesystem; C.F.W. Siegel's musikalienhandlung*; R. Linnemann: Leipzig, Germany, 1913. (In German)
39. Moreau, J.J. Fonctions convexes duales et points proximaux dans un espace hilbertien. *Comptes Rendus de l'Académie des Sciences* **1962**, *255*, 2897–2899. (In French)
40. Zia, R.K.P.; Redish, E.F.; McKay, S.R. Making Sense of the Legendre Transform. **2009**, arXiv:0806.1147.
41. Koszul, J.L. Variétés localement plates et convexité. *Osaka. J. Math.* **1965**, *2*, 285–290. (In French)

42. Koszul, J.L. Exposés sur les Espaces Homogènes Symétriques; Publicação da Sociedade de Matematica de São Paulo: São Paulo, Brazil, 1959. (In French)
43. Koszul, J.L. Sur la forme hermitienne canonique des espaces homogènes complexes. *Can. J. Math.* **1955**, *7*, 562–576. (In French)
44. Koszul, J.L. *Lectures on Groups of Transformations*; Tata Institute of Fundamental Research: Bombay, India, 1965.
45. Koszul, J.L. Domaines bornées homogènes et orbites de groupes de transformations affines. *Bull. Soc. Math. Fr.* **1961**, *89*, 515–533. (In French)
46. Koszul, J.L. Ouverts convexes homogènes des espaces affines. *Math. Z.* **1962**, *79*, 254–259. (In French)
47. Koszul, J.L. Déformations des variétés localement plates. *Ann. Inst. Fourier* **1968**, *18*, 103–114. (In French)
48. Vinberg, E.B. Homogeneous convex cones. *Trans. Mosc. Math. Soc.* **1963**, *12*, 340–363.
49. Vinberg, E.B. The Theory of Homogeneous Convex Cones. *Trudy Moskovskogo Ma tematicheskogo Obshchestva* **1963**, *12*, 303–358.
50. Jensen, J.L.W. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Math*. **1906**, *30*, 175–193. (In French)
51. Needham, T. A Visual Explanation of Jensen's Inequality. *Am. Math. Mon*. **1993**, *8*, 768–777.
52. Jaynes, E.T. Information Theory and Statistical Mechanics. *Phys. Rev. Ser. II* **1957**, *106*, 620–630.
53. Jaynes, E.T. Information Theory and Statistical Mechanics II. *Phys. Rev. Ser*. **1957**, *108*, 171–190.
54. Jaynes, E.T. Prior Probabilities. *IEEE Trans. Syst. Sci. Cybern.* **1968**, *4*, 227–241.
55. Dacunha-Castelle, D.; Gamboa, F. Maximum d'entropie et problèmes des moments. *Ann. Inst. H. Poincaré Prob. Stat.* **1990**, *26*, 567–596. (In French)
56. Gamboa, F.; Gassiat, E. Maximum d'entropie et problème des moments: Cas multidimensionnel. *Probab. Math. Stat*. **1991**, *12*, 67–83. (In French)
57. Dacunha-Castelle, D.; Gamboa, F. Maximum de l'entropie sous contraintes non linéaires. *Ann. Inst H. Poincaré Prob. Stat.* **1990**, *4*, 567–596. (In French)
58. Krein, M.G.; Nudelman, A.A. *The Markov Moment Problem And Extremal Problems*; American Mathematical Society: New York, NY, USA, 1977.
59. Souriau, J.M. D*éfinition covariante* des équilibres thermodynamiques. *Suppl. Nuov. Cimento* **1966**, *1*, 203–216. (In French)
60. Crouzeix, J.P. A Relationship Between The Second Derivatives of a Convex Function and of Its Conjugate. *Math. Program*. **1977**, *3*, 364–365.
61. Seeger, A. Second Derivative of a Convex Function and of Its Legendre-Fenchel Transformate. *SIAM J. Optim*. **1992**, *2*, 405–424.
62. Hiriart-Urruty, J.B. A new set-valued second-order derivative for convex functions. In *Mathematics for Optimization*; Mathematical Studies Series 129; Elsevier: Amsterdam, The Netherlands, 1986.

63. Rao, C.R. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **1945**, *37*, 81–89.
64. Fréchet, M. Sur l'extension de certaines évaluations statistiques au cas de petits échantillons. *Revue de l'Institut International de Statistique* **1943**, *11*, 182–205. (In French)
65. Koszul, J.L. Trajectoires Convexes de Groupes Affines Unimodulaires. In *Essays on Topology and Related Topics*; Springer: Berlin, Germany, 1970; pp. 105–110.
66. Vey, J. Sur les Automorphismes Affines des Ouverts Convexes Saillants. *Annali della Scuola Normale Superiore di Pisa, Classe di Science* **1970**, *24*, 641–665. (In French)
67. Vey, J. Sur une notion d'hyperbolicité des variables localement plates. *Thèse de Troisième Cycle de Mathématiques Pures*; Faculté des sciences de l'université de Grenoble: Grenoble, France, 1969. (In French)
68. Shima, H. Geometry of Hessian Structures. In *Springer Lecture Notes in Computer Science*; Nielsen, F., Frederic, B., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; Volume 8085, pp. 37–55.
69. Shima, H. The Geometry of Hessian Structures; World Scientific: Singapore, Singapore, 2007.
70. Byande, P.M.; Ngakeu, F.; Nguiffo Boyom, M.; Wolak, R. KV-Cohomology and Differential Geometry of Affinely Flat Manifolds. Information Geometry. *Afr. Diaspora J. Math*. **2012**, *14*, 197–226.
71. Sasaki, T. A Note on Characteristic Functions and Projectively Invariant Metrics on a Bounded Convex Domain. *Tokyo J. Math.* **1985**, *8*, 49–79.
72. Sasaki, T. Hyperbolic Affine Hyperspheres. *Nagoya Math. J.* **1980**, *77*, 107–123.
73. Rothaus, O.S. The Construction of Homogeneous Convex Cones. *Ann. Math.* **1966**, *83*, 358–376.
74. Barbaresco, F. *Information Geometry of Covariance Matrix: Cartan-Siegel Homogeneous Bounded Domains, Mostow/Berger Fibration and Fréchet Median, Matrix Information Geometry*; Bhatia, R., Nielsen, F., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 199–256.
75. Arnaudon, M.; Barbaresco, F.; Yang, L. Riemannian medians and means with applications to radar signal processing. *IEEE J. Sel. Top. Signal Process*. **2013**, *7*, 595–604.
76. Yang, L. Médianes de Mesures de Probabilité dans les Variétés Riemanniennes et Applications à la Détection de Cibles Radar. Thales Ph.D, Thèse de l'Université de Poitiers, Poitiers, France, 2012. (In French)
77. Barbaresco, F. Information/Contact Geometries and Koszul Entropy, Geometric Science of Information. In *Lecture Notes in Computer Science*; Nielsen, F., Barbaresco F., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; Volume 8085, pp. 604–611.
78. Cartan, E. *Le rôle de la Théorie des Groupes de Lie Dans L'évolution de la Géométrie Moderne*; International Mathematical Union: Berlin, Germany, 1936. (In French)

79. Souriau, J.M. Thermodynamique et géométrie. In *Differential Geometrical Methods in Mathematical Physics II*; Bleuler, K., Reetz, A., Petry, H.R., Eds.; Springer: Berlin/Heidelberg, Germany, 1978; pp. 369–397.
80. Souriau, J.M. Géométrie de l'espace de phases. *Comm. Math. Phys.* **1966**, *374*, 1–30 (In French)
81. Souriau J.M. On Geometric Mechanics. *Discret. Cont. Dyn. Syst. J.* **2007**, *19*, 595–607.
82. Souriau, J.M. *Structure des systèmes dynamiques*; Editions Jacques Gabay: Paris, France, 1970. (In French)
83. Souriau, J.M. Structure of Dynamical Systems, volume 149 of Progress in Mathematics. In *A Symplectic View of Physics*; Birkhäuser Boston Inc.: Boston, MA, USA, 1997.
84. Souriau, J.M. Thermodynamique Relativiste des Fluides; Centre de Physique Théorique: Marseille, France, 1977. (In French)
85. Souriau, J.M.; Iglesias, P. *Heat Cold and Geometry. Differential Geometry and Mathematical Physics, Mathematical Physics Studies Volume*; Springer: Amsterdam, The Netherlands, 1983; pp. 37–68.
86. Souriau, J.M. Dynamic Systems Structure (chap.16 convexité, chap. 17 Mesures, chap. 18 Etats statistiques, Chap. 19 Thermodynamique), available in Souriau archive (document sent by C. Vallée), unpublished technical notes, 1980.
87. Souriau, J.M. Géométrie *Symplectique et Physique Mathématique*; Éditions du C.N.R.S.: Paris, France, 1975. (In French)
88. Ruelle, D. *Statistical Mechanics. Rigorous Results*; Imperial College Press: London, UK, 1999.
89. Ruelle, D. *Hasard et Chaos*; Editions Odile Jacob: Paris, France, 1991.
90. Ruelle, D. *Thermodynamic Formalism: The Mathematical Structure of Equilibrium Statistical Mechanics*, 2nd ed.; Cambridge Mathematical Library; Cambridge University Press: Cambridge, UK, 2004.
91. Vallée, C. Lois de Comportement des Milieux Continus Dissipatifs Compatibles Avec la Physique Relativiste. Ph.D Thesis, Poitiers University, Poitiers, France, 1987. (In French)
92. Vallée, C. Relativistic thermodynamics of continua. *Int. J. Eng. Sci*. **1981**, *19*, 589–601.
93. De Saxcé, G.; Vallée, C. Bargmann group, momentum tensor and Galilean invariance of Clausius-Duhem inequality. *Int. J. Eng. Sci*. **2012**, *50*, 216–232.
94. Vallée, C.; Hjiaj, M.; Fortuné, D.; de Saxcé, G. Canonical and Anti-Canonical Transformations Preserving Convexity of Potentials. *J. Elast*. **2011**, *103*, 247–267.
95. Iglésias, P. Equilibre statistiques et géométrie symplectique en relativité générale. A*nnales de l'institut Henri Poincaré (A) Physique théorique* **1982**, *36*, 257–270. (In French)
96. Iglésias, P. Essai de thermodynamique rationnelle des milieux continus. A*nnales de l'institut Henri Poincaré (A) Physique théorique*. **1981**, *34*, 1–24. (In French)
97. Dubois, F. Conservation Laws Invariants for Galileo Group; Cemracs Preliminary Results. Available online: http://hal.archives-ouvertes.fr/docs/00/55/53/13/PDF/dubois-cemracs99.janv2011.pdf (accessed on 20 June 2014).

98. Roblin, T. Ergodicité et équidistribution en courbure négative. *Mémoire de la Société Mathématique de France* **2003**, *95*, 96. (In French)
99. Paulin, F.; Pollicott, M.; Schapira, B. Equilibrium states in negative curvature. **2013**, arXiv:1211.6242.
100. Coudène, Y. Gibbs measures on negatively curved manifolds. *J. Dynam. Control Syst.* **2003**, *9*, 89–101.
101. Coudène, Y.; Schapira, B. Generic measures for geodesic flows on nonpositively curved manifolds. *Dyn. Syst.* **2014**, submitted.
102. Haydn, N.T.A.; Ruelle, D. Equivalence of Gibbs and equilibrium states for homeomorphisms satisfying expansiveness and specification. *Comm. Math. Phys.* **1992**, *148*, 155–167.
103. Mohsen, O. Le bas du spectre d'une variété hyperbolique est un point selle. *Ann. Sci. École Norm. Sup.* **2007**, *40*, 191–207. (In French)
104. Donaldson, S.K. Scalar Curvature and Stability of Toric Variety. *J. Differ. Geom*. **2002**, *62*, 289–349.
105. Abreu, M. Kähler geometry of toric varieties and extremal metrics. *Int. J. Math.* **1998**, *9*, 641–651.
106. Guan, D. On modified Mabuchi functional and Mabuchi moduli space of kahler metrics on toric bundles. *Math. Res. Lett*. **1999**, *6*, 547–555.
107. Guillemin, V. Kaehler structures on toric varieties. *J. Differ. Geom*. **1994**, *40*, 285–309.
108. Guillemin, V. *Moment Maps and Combinatorial Invariants of Hamiltonian Tn-Spaces*; Birkhau-ser: Basel, Switzerland, 1994.
109. Atiyah, M.; Bott, R. The moment map and equivariant cohomology. *Topology* **1984**, *23*, 1–28.
110. Bruguières, A. Propriétés de convexité de l'application moment. *Séminaire Bourbaki* **1985**, *28*, 63–87. (In French)
111. Condevaux, M.; Dazord, P.; Molino, P. *Géométrie du moment. Dans Travaux du séminaire Sud Rhodanien de Géométrie*; University Lyon: Lyon, France, 1988; pp. 131–160.
112. Delzant, T. Hamiltoniens périodiques et images convexes de l'application moment. *Bull. Soc. Math. Fr.* **1988**, *116*, 315–339. (In French)
113. Guillemin, V.; Sternberg, S. Convexity properties of the moment mapping. *Inv. Math*. **1982**, *67*, 491–513.
114. Kirwan, F. Convexity properties of the moment mapping. *Inv.Math*. **1984**, *77*, 547–552.
115. Kapranov, M. Thermodynamics and the moment map. **2011**, arXiv:1108.3472v1.
116. Biquard, O. *Métriques Kählériennes Extrémales sur les Surfaces Toriques*; Société Mathématique de France: Paris, France, **2011**; Volume 1018. (In French)
117. Pavlov, V.P.; Sergeev, V.M. Thermodynamics from the Differential Geometry Standpoint. *Theor. Math. Phys*. **2008**, *157*, 1484–1490.
118. Kozlov, V.V. Heat Equilibrium by Gibbs and Poincaré; *Dokl. Ross. Akad. Nauk* **2002**, *382*, 602–606. (In French)
119. Berezin, F.A. *Lectures on Statistical Physics*; World Scientific: Singapore, Singapore, 2007.

120. Poincaré, H. Réflexions sur la théorie cinétique des gaz. *J. Phys. Theor. Appl.* **1906**, *5*, 369–403.
121. Carathéodory, C. Untersuchungen über die Grundlagen der Thermodynamik. *Math. Ann*. **1909**, *67*, 355–386. (In German)
122. Souriau, J.M. Mécanique Classique et Géométrie Symplectique; Report ref. CPT-84/PE-1695 CNRS Centre de Physique Théorique: Marseille, France, 1984. (In French)
123. Viterbo, C. Generating Functions, Symplectic Geometry and Applications. *Proc. Intern. Congr. Math., Zurich* **1994**, *1*, 537–547.
124. Viterbo, C. Symplectic topology as the geometry of generating functions. *Math. Ann*. **1992**, *292*, 685–710.
125. Hörmander, L. Fourier integral operators I. *Acta Math.* **1971**, *127*, 79–183.
126. Théret, D. A complete proof of Viterbo's uniqueness theorem on generating functions. *Topol. Appl*. **1999**, *96*, 249–266.
127. Chentsov, N.N. Statistical Decision Rules and Optimal Inferences (Transactions of Mathematics Monograph); American Mathematical Society: Providence, RI, USA, 1982; Volume 53.
128. Villani, C. (Ir)rréversibilité et entropie/(Ir)reversibility and entropy. In *Séminaire Poincaré Le temps*; École Polytechnique: Paris, France, 2010. (In French)
129. Ollivier, Y. Aspects de l'entropie en mathématiques et en physique. Available online: http://www.yann-ollivier.org/entropie/entropie.pdf (accessed on 1 January 2014). (In French)
130. Avantaggiati, A.; Loreti, P. On Fourier Transform, Parseval Equality, and the Inversion Formula in Idempotent Analysis. In Proceedings of the 2013 European Control Conference (ECC), Zürich, Switzerland, 17–19 July 2013.
131. Litvinov, G.L. The Maslov Dequantization, Idempotent and Tropical Mathematics: A brief Introduction. *J. Math. Sci.* **2007**, *140*, 426–444.
132. Leray, J. Un prolongement de la transformation de Laplace qui transforme la solution unitaire d'un opérateur hyperbolique en sa solution élémentaire. (Problème de Cauchy. IV.). *Bull. Soc. Math. Fr.* **1962**, *90*, 39–156. (In French)
133. Leray, J. Le calcul différentiel et intégral sur une variété analytique complexe Problème de Cauchy, III. *Bull. Soc. Math. Fr*. **1959**, *82*, 6–180. (In French)
134. Maslov, V.P. *Operational Methods*; MIR: Moscow, Russia, 1976.
135. Del Moral, P.; Doisy, M. Maslov idempotent probability calculus, I, II. *Theory Probab. Appl.* **2000**, *44*, 319–332.
136. Del Moral, P.; Doisy, M. On the applications of Maslov optimization theory. *Math. Notes* **2001**, *69*, 232–244.
137. Laplace P.S. Mémoire sur la probabilité des causes sur les évènements. In *Mémoires de Mathématique et de Physique*; De l'Imprimerie Royale: Paris, France, 1774. (In French)
138. Legendre, A.M. *Mémoire Sur L'intégration de Quelques Equations aux Différences Partielles*; Mémoires de l'Académie des Sciences: Paris, France, 1787; pp. 309–351. (In French)

139. Monge, G. *Sur le Calcul Intégral des Equations aux Différences Partielles*; Mémoires de l'Académie des Sciences: Paris, France, 1784; pp. 118–192. (In French)
140. Darboux, G. *Leçons sur la Théorie Générale des Surfaces et les Applications Géométriques du Calcul Infinitésimal: Premiere Partie (Généralités, Coordonnées Curvilignes, Surface Minima)*; Gauthier-Villars: Paris, France, 1887. (In French)
141. Hadamard, J. Les surfaces à courbures opposées et leurs lignes géodésiques. *Journal de Mathématiques Pures et Appliquées* **1898**, *4*, 27–74. (In French)
142. Vesentini, E. *Geometry of Homogeneous Bounded Domains*; Springer-Verlag: Berlin/Heidelberg, Germany, 2011.
143. Bourdon, M. Structure conforme au bord et flot géodésique d'un CAT(−1)-espace. *L'Enseignement Mathématique* **1995**, *41*, 63–102. (In French)
144. Deza, E.; Deza, M.M. Dictionary of Distances; Elsevier: Amsterdam, The Netherlands, 2006.
145. Cartan, E. *Leçons sur les Invariants Intégraux*; Hermann: Paris, France, 1922. (In French)
146. Libermann, P.; Marle, C.M. *Symplectic Geometry and Analytical Mechanics*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 1987.

# The Entropy-Based Quantum Metric

**Roger Balian**

**Abstract:** The von Neumann entropy $S(\hat{D})$ generates in the space of quantum density matrices $\hat{D}$ the Riemannian metric $\mathrm{d}s^2 = -\mathrm{d}^2 S(\hat{D})$, which is physically founded and which characterises the amount of quantum information lost by mixing $\hat{D}$ and $\hat{D} + \mathrm{d}\hat{D}$. A rich geometric structure is thereby implemented in quantum mechanics. It includes a canonical mapping between the spaces of states and of observables, which involves the Legendre transform of $S(\hat{D})$. The Kubo scalar product is recovered within the space of observables. Applications are given to equilibrium and non equilibrium quantum statistical mechanics. There the formalism is specialised to the relevant space of observables and to the associated reduced states issued from the maximum entropy criterion, which result from the exact states through an orthogonal projection. Von Neumann's entropy specialises into a relevant entropy. Comparison is made with other metrics. The Riemannian properties of the metric $\mathrm{d}s^2 = -\mathrm{d}^2 S(\hat{D})$ are derived. The curvature arises from the non-Abelian nature of quantum mechanics; its general expression and its explicit form for q-bits are given, as well as geodesics.

Reprinted from *Entropy*. Cite as: Balian, R. The Entropy-Based Quantum Metric. *Entropy* **2014**, *16*, 3878–3888.

## 1. A Physical Metric for Quantum States

Quantum physical quantities pertaining to a given system, termed as "observables" $\hat{O}$, behave as non-commutative random variables and are elements of a C*-algebra. We will consider below systems for which these observables can be represented by *n*-dimensional Hermitean matrices in a finite-dimensional Hilbert space $\mathcal{H}$. In quantum (statistical) mechanics, the "state" of such a system encompasses the expectation values of all its observables [1]. It is represented by a density matrix $\hat{D}$, which plays the rôle of a probability distribution, and from which one can derive the expectation value of $\hat{O}$ in the form

$$< \hat{O} >= \mathrm{Tr}\, \hat{D}\hat{O} = (\hat{D}; \hat{O}) \,. \tag{1}$$

Density matrices should be Hermitean ($< \hat{O} >$ is real for $\hat{O} = \hat{O}^\dagger$), normalised (the expectation value of the unit observable is $\mathrm{Tr}\, \hat{D} = 1$) and non-negative (variances $< \hat{O}^2 > - < \hat{O} >^2$ are non-negative). They depend on $n^2 - 1$ real parameters. If we keep aside the multiplicative structure of the set of operators and focus on their linear vector space structure, Equation (1) appears as a linear mapping of the space of observables onto real numbers. We can therefore regard the observables and the density operators $\hat{D}$ as elements of two dual vector spaces, and expectation values (1) appear as scalar products.

It is of interest to define a metric in the space of states. For instance, the distance between an exact state $\hat{D}$ and an approximation $\hat{D}_{\mathrm{app}}$ would then characterise the quality of this approximation. However, all physical quantities come out in the form (1) which lies astride the two dual spaces of observables and states. In order to build a metric having physical relevance, we need to rely on another meaningful quantity which pertains only to the space of states.

We note at this point that quantum states are probabilistic objects that gather information about the considered system. Then, the amount of missing information is measured by von Neumann's entropy

$$S(\hat{D}) \equiv -\operatorname{Tr} \hat{D} \ln \hat{D} \,. \tag{2}$$

Introduced in the context of quantum measurements, this quantity is identified with the thermodynamic entropy when $\hat{D}$ is an equilibrium state. In non-equilibrium statistical mechanics, it encompasses, in the form of "relevant entropy" (see Section 5 below), various entropies defined through the maximum entropy criterion. It is also introduced in quantum computation. Alternative entropies have been introduced in the literature, but they do not present all the distinctive and natural features of von Neumann's entropy, such as additivity and concavity.

As $S(\hat{D})$ is a concave function, and as it is the sole physically meaningful quantity apart from expectation values, it is natural to rely on it for our purpose. We thus define [2] the distance $\mathrm{d}s$ between two neighbouring density matrices $\hat{D}$ and $\hat{D} + \mathrm{d}\hat{D}$ as the square root of

$$\mathrm{d}s^2 = -\mathrm{d}^2 S(\hat{D}) = \operatorname{Tr} \mathrm{d}\hat{D} \mathrm{d} \ln \hat{D} \,. \tag{3}$$

This Riemannian metric is of the Hessian form since the metric tensor is generated by taking second derivatives of the function $S(\hat{D})$ with respect to the $n^2 - 1$ coordinates of $\hat{D}$. We may take for such coordinates the real and imaginary parts of the matrix elements, or equivalently (Section 6) some linear transform of these (keeping aside the norm $\operatorname{Tr} \hat{D} = 1$).

## 2. Interpretation in the Context of Quantum Information

The simplest example, related to quantum information theory, is that of a q-bit (two-level system or spin $\frac{1}{2}$) for which $n = 2$. Its states, represented by $2 \times 2$ Hermitean normalised density matrices $\hat{D}$, can conveniently be parameterised, on the basis of Pauli matrices, by the components $r_\mu = D_{12} + D_{21}$, $i(D_{12} - D_{21})$, $D_{11} - D_{22}$ ($\mu = 1, 2, 3$) of a 3-dimensional vector $\mathbf{r}$ lying within the unit Poincaré–Bloch sphere ($r \leq 1$). From the corresponding entropy

$$S = \frac{1+r}{2} \ln \frac{2}{1+r} + \frac{1-r}{2} \ln \frac{2}{1-r} \,, \tag{4}$$

we derive the metric

$$\mathrm{d}s^2 = \frac{1}{1-r^2} \left( \frac{\mathbf{r} \cdot \mathrm{d}\mathbf{r}}{r} \right)^2 + \frac{1}{2r} \ln \frac{1+r}{1-r} \left\| \frac{\mathbf{r} \times \mathrm{d}\mathbf{r}}{r} \right\|^2 \,, \tag{5}$$

which is a natural Riemannian metric for q-bits, or more generally for positive $2 \times 2$ matrices. The metric tensor characterizing (5) diverges in the vicinity of pure states $r = 1$, due to the singularity of the entropy (2) for vanishing eigenvalues of $\hat{D}$. However, the distance between two arbitrary (even pure) states $\hat{D}'$ and $\hat{D}''$ measured along a geodesic is always finite. We shall see (Equation (29)) that for $n = 2$ the geodesic distance $s$ between two neighbouring pure states $\hat{D}'$ and $\hat{D}''$, represented by unit vectors $\mathbf{r}'$ and $\mathbf{r}''$ making a small angle $\delta\varphi \sim |\mathbf{r}' - \mathbf{r}''|$, behaves as $\delta s^2 \sim \delta\varphi^2 \ln(4\sqrt{\pi}/\delta\varphi)$. The singularity of the metric tensor manifests itself through this logarithmic factor.

Identifying von Neumann's entropy to a measure of missing information, we can give a simple interpretation to the distance between two states. Indeed, the concavity of entropy expresses that some information is lost when two statistical ensembles described by different density operators merge. By mixing two equal size populations described by the neighbouring distributions $\hat{D}' = \hat{D} + \frac{1}{2}\delta\hat{D}$ and $\hat{D}'' = \hat{D} - \frac{1}{2}\delta\hat{D}$ separated by a distance $\delta s$, we lose an amount of information given by

$$\Delta S \equiv S\left(\hat{D}\right) - \frac{S(\hat{D}') + S(\hat{D}'')}{2} \sim \frac{\delta s^2}{8}\,, \tag{6}$$

and thereby directly related to the distance $\delta s$ defined by (3). The proof of this equivalence relies on the expansion of the entropies $S(\hat{D}')$ and $S(\hat{D}'')$ around $\hat{D}$, and is valid when $\mathrm{Tr}\,\delta\hat{D}^2$ is negligible compared to the smallest eigenvalue of $\hat{D}$. If $\hat{D}'$ and $\hat{D}''$ are distant, the quantity $8\Delta S$ cannot be regarded as the square of a distance that would be generated by a local metric. The equivalence (6) for neighbouring states shows that $\mathrm{d}s^2$ is the metric that is the best suited to measure losses of information my mixing.

The singularity of $\delta s^2$ at the edge of the positivity domain of $\hat{D}$ may suggest that the result (6) holds only within this domain. In fact, this equivalence remains nearly valid even in the limit of pure states because $\Delta S$ itself involves a similar singularity. Indeed, if the states $\hat{D}' = |\psi' >< \psi'|$ and $\hat{D}'' = |\psi'' >< \psi''|$ are pure and close to each other, the loss of information $\Delta S$ behaves as $8\Delta S \sim \delta\varphi^2 \ln(4/\delta\varphi)$ where $\delta\varphi^2 \sim 2\,\mathrm{Tr}\,\delta D^2$. This result should be compared to various geodesic distances between pure quantum states, which behave as $\delta s^2 \sim \delta\varphi^2 \ln(4\sqrt{\pi}/\delta\varphi$ for the present metric, and as $\delta s^2_{\mathrm{BH}} = 4\delta s^2_{\mathrm{FS}} \sim \delta\varphi^2 \sim \mathrm{Tr}(\hat{D}' - \hat{D}'')^2$ for the Bures – Helstrom and the quantum Fubini – Study metrics, respectively (see Section 7; these behaviours hold not only for $n = 2$ but for arbitrary $n$ since only the space spanned by $|\psi' >$ and $|\psi'' >$ is involved). Thus, among these metrics, only $\mathrm{d}s^2 = -\mathrm{d}^2 S$ can be interpreted in terms of information loss, whether the states $\hat{D}'$ and $\hat{D}''$ are pure or mixed.

At the other extreme, around the most disordered state $\hat{D} = \hat{I}/n$, in the region $\parallel n\hat{D} - \hat{I} \parallel \ll 1$, the metric becomes Euclidean since $\mathrm{d}s^2 = \mathrm{Tr}\,\mathrm{d}\hat{D}\mathrm{d}\ln\hat{D} \sim n\,\mathrm{Tr}(\mathrm{d}\hat{D})^2$ (for $n = 2$, $\mathrm{d}s^2 = \mathrm{d}\mathbf{r}^2$). For a given shift $\mathrm{d}\hat{D}$, the qualitative change of a state $\hat{D}$, as measured by the distance $\mathrm{d}s$, gets larger and larger as the state $\hat{D}$ becomes purer and purer, that is, when the information contents of $\hat{D}$ increases.

### 3. Geometry of Quantum Statistical Mechanics

A rich geometric structure is generated for both states and observables by von Neumann's entropy through introduction of the metric $\mathrm{d}s^2 = -\mathrm{d}^2 S$. Now, this metric (3) supplements the algebraic structure of the set of observables and the above duality between the vector spaces of states and of observables, with scalar product (1). Accordingly, we can define naturally within the space of states scalar products, geodesics, angles, curvatures.

We can also regard the coordinates of $\mathrm{d}\hat{D}$ and $\mathrm{d}\ln\hat{D}$ as covariant and contravariant components of the same infinitesimal vector (Section 6). To this aim, let us introduce the mapping

$$\hat{D} \equiv \frac{e^{\hat{X}}}{\mathrm{Tr}\,e^{\hat{X}}} \tag{7}$$

between $\hat{D}$ in the space of states and $\hat{X}$ in the space of observables. The operator $\hat{X}$ appears as a parameterisation of $\hat{D}$. (The normalisation of $\hat{D}$ entails that $\hat{X}$, defined within an arbitrary additive constant operator $X_0\,\hat{I}$, also depends on $n^2-1$ independent real parameters.) The metric (3) can then be re-expressed in terms of $\hat{X}$ in the form

$$\mathrm{d}s^2 = \mathrm{Tr}\,\mathrm{d}\hat{D}\mathrm{d}\hat{X} = \mathrm{Tr}\int_0^1 \mathrm{d}\xi\,\hat{D}e^{-\xi\hat{X}}\mathrm{d}\hat{X}e^{\xi\hat{X}}\mathrm{d}\hat{X} - (\mathrm{Tr}\,\hat{D}\mathrm{d}\hat{X})^2 = \mathrm{d}^2\ln\mathrm{Tr}\,e^{\hat{X}} = \mathrm{d}^2F\,, \qquad (8)$$

where we introduced the function

$$F(\hat{X}) \equiv \ln\mathrm{Tr}\,e^{\hat{X}} \qquad (9)$$

of the observable $\hat{X}$(The addition of $X_0\hat{I}$ to $\hat{X}$ results in the addition of the irrelevant constant $X_0$ to $F$). This mapping provides us with a natural metric in the space of observables, from which we recover the scalar product between $\mathrm{d}\hat{X}_1$ and $\mathrm{d}\hat{X}_2$ in the form of a Kubo correlation in the state $\hat{D}$. The metric (8) has been quoted in the literature under the names of Bogoliubov–Kubo–Mori.

## 4. Covariance and Legendre Transformation

We can recover the above geometric mapping (7) between $\hat{D}$ and $\hat{X}$, or between the covariant and contravariant coordinates of $\mathrm{d}\hat{D}$, as the outcome of a Legendre transformation, by considering the function $F(\hat{X})$. Taking its differential $\mathrm{d}F = \mathrm{Tr}\,e^{\hat{X}}\mathrm{d}\hat{X}/\,\mathrm{Tr}\,e^{\hat{X}}$, we identify the partial derivatives of $F(\hat{X})$ with the coordinates of the state $\hat{D} = e^{\hat{X}}/\,\mathrm{Tr}\,e^{\hat{X}}$, so that $\hat{D}$ appears as conjugate to $\hat{X}$ in the sense of Legendre transformations. Expressing then $\hat{X}$ as function of $\hat{D}$ and inserting into $F - \mathrm{Tr}\,\hat{D}\hat{X}$, we recognise that the Legendre transform of $F(\hat{X})$ is von Neumann's entropy $F - \mathrm{Tr}\,\hat{D}\hat{X} = S(\hat{D}) = -\,\mathrm{Tr}\,\hat{D}\ln\hat{D}$. The conjugation between $\hat{D}$ and $\hat{X}$ is embedded in the equations

$$\mathrm{d}F = \mathrm{Tr}\,\hat{D}\mathrm{d}\hat{X}\;;\quad \mathrm{d}S = -\,\mathrm{Tr}\,\hat{X}\mathrm{d}\hat{D}\,. \qquad (10)$$

Legendre transformations are currently used in equilibrium thermodynamics. Let us show that they come out in this context directly as a special case of the present general formalism. The entropy of thermodynamics is a function of the extensive variables, energy, volume, particle numbers, etc. Let us focus for illustration on the energy $U$, keeping the other extensive variables fixed. The thermodynamic entropy $S(U)$, a function of the single variable $U$, generates the inverse temperature as $\beta = \partial S/\partial U$. Its Legendre transform is the Massieu potential $F(\beta) = S - \beta U$. In order to compare these properties with the present formalism, we recall how thermodynamics comes out in the framework of statistical mechanics. The thermodynamic entropy $S(U)$ is identified with the von Neumann entropy (2) of the Boltzmann–Gibbs canonical equilibrium state $\hat{D}$, and the internal energy with $U = \mathrm{Tr}\,\hat{D}\hat{H}$. In the relation (7), the operator $\hat{X}$ reads $\hat{X} = -\beta\hat{H}$ (within an irrelevant additive constant). By letting $U$ or $\beta$ vary, we select within the spaces of states and of observables a one-dimensional subset. In these restricted subsets, $\hat{D}$ is parameterised by the single coordinate $U$, and the corresponding $\hat{X}$ by the coordinate $-\beta$.

By specialising the general relations (10) to these subsets, we recover the thermodynamic relations $\mathrm{d}F = -U\mathrm{d}\beta$ and $\mathrm{d}S = \beta\mathrm{d}U$. We also recover, by restricting the metric (3) or (8) to these subsets, the current thermodynamic metric $\mathrm{d}s^2 = -(\partial^2 S/\partial U^2)\mathrm{d}U^2 = -\mathrm{d}U\mathrm{d}\beta$.

More generally, we can consider the Boltzmann–Gibbs states of equilibrium statistical mechanics as the points of a manifold embedded in the full space of states. The thermodynamic extensive variables, which parameterise these states, are the expectation values of the conserved macroscopic observables, that is, they are a subset of the expectation values (1) which parameterise arbitrary density operators. Then the standard geometric structure of thermodynamics simply results from the restriction of the general metric (3) to this manifold of Boltzmann–Gibbs states. The commutation of the conserved observables simplifies the reduced thermodynamic metric, which presents the same features as a Fisher metric (see Section 6).

## 5. Relevant Entropy and Geometry of the Projection Method

The above ideas also extend to non-equilibrium quantum statistical mechanics [2–4]. When introducing the metric (3), we indicated that it may be used to estimate the quality of an approximation. Let us illustrate this point with the Nakajima–Zwanzig–Mori–Robertson projection method, best introduced through maximum entropy. Consider some set $\{\hat{A}_k\}$ of "relevant observables", whose time-dependent expectation values $a_k \equiv < \hat{A}_k > = \mathrm{Tr}\, \hat{D}\hat{A}_k$ we wish to follow, discarding all other variables. The exact state $\hat{D}$ encodes the variables $\{a_k\}$ that we are interested in, but also the expectation values (1) of the other observables that we wish to eliminate. This elimination is performed by associating at each time with $\hat{D}$ a "reduced state" $\hat{D}_{\mathrm{R}}$ which is equivalent to $\hat{D}$ as regards the set $a_k = \mathrm{Tr}\, \hat{D}_{\mathrm{R}}\hat{A}_k$, but which provides no more information than the values$\{a_k\}$. The former condition provides the constraints $< \hat{A}_k > = a_k$, and the latter condition is implemented by means of the maximum entropy criterion: One expresses that, within the set of density matrices compatible with these constraints, $\hat{D}_{\mathrm{R}}$ is the one which maximises von Neumann's entropy (2), that is, which contains solely the information about the relevant variables $a_k$. The least biased state $\hat{D}_{\mathrm{R}}$ thus defined has the form $\hat{D}_{\mathrm{R}} = e^{\hat{X}_{\mathrm{R}}}/\,\mathrm{Tr}\, e^{\hat{X}_{\mathrm{R}}}$, where $\hat{X}_{\mathrm{R}} \equiv \sum_k \lambda_k \hat{A}_k$ involves the time-dependent Lagrange multipliers $\lambda_k$, which are related to the set $a_k$ through $\mathrm{Tr}\, \hat{D}_{\mathrm{R}}\hat{A}_k = a_k$.

The von Neumann entropy $S(\hat{D}_{\mathrm{R}}) \equiv S_{\mathrm{R}}\{a_k\}$ of this reduced state $\hat{D}_{\mathrm{R}}$ is called the "relevant entropy" associated with the considered relevant observables $\hat{A}_k$. It measures the amount of missing information, when only the values $\{a_k\}$ of the relevant variables are given. During its evolution, $\hat{D}$ keeps track of the initial information about all the variables $< \hat{O} >$ and its entropy $S(\hat{D})$ remains constant in time. It is therefore smaller than the relevant entropy $S(\hat{D}_{\mathrm{R}})$ which accounts for the loss of information about the irrelevant variables. Depending on the choice of relevant observables $\{\hat{A}_k\}$, the corresponding relevant entropies $S_{\mathrm{R}}\{a_k\}$ encompass various current entropies, such as the non-equilibrium thermodynamic entropy or Boltzmann's H-entropy.

The same structure as the one introduced above for the full spaces of observables and states is recovered in this context. Here, for arbitrary values of the parameters $\lambda_k$, the exponents $\hat{X}_{\mathrm{R}} = \sum_k \lambda_k \hat{A}_k$ constitute a subspace of the full vector space of observables, and the parameters $\{\lambda_k\}$ appear as the coordinates of $\hat{X}_{\mathrm{R}}$ on the basis $\{\hat{A}_k\}$. The corresponding states $\hat{D}_{\mathrm{R}}$, parameterised by the set $\{a_k\}$, constitute a subset of the space of states, the manifold $\mathcal{R}$ of "reduced states"(Note that this manifold is not a hyperplane, contrary to the space of relevant observables; it is embedded in the

full vector space of states, but does not constitute a subspace). By regarding $S_{\rm R}\{a_k\}$ as a function of the coordinates $\{a_k\}$, we can define a metric $\mathrm{d}s^2 = -\mathrm{d}^2 S_{\rm R}\{a_k\}$ on the manifold $\mathcal{R}$, which is the restriction of the metric (3).
Its alternative expression $\mathrm{d}s^2 = \sum_k \mathrm{d}a_k \mathrm{d}\lambda_k = \mathrm{d}^2 F_{\rm R}\{\lambda_k\}$, where $F_{\rm R}\{\lambda_k\} \equiv \ln \mathrm{Tr} \exp \sum_k \lambda_k \hat{A}_k$, is a restriction of (8). The correspondence between the two parameterisations $\{a_k\}$ and $\{\lambda_k\}$ is again implemented by the Legendre transformation which relates $S_{\rm R}\{a_k\}$ and $F_{\rm R}\{\lambda_k\}$.

The projection method relies on the mapping $\hat{D} \mapsto \hat{D}_{\rm R}$ which associates $\hat{D}_{\rm R}$ to $\hat{D}$. It consists in replacing the Liouville–von Neumann equation of motion for $\hat{D}$ by the corresponding dynamical equation for $\hat{D}_{\rm R}$ on the manifold $\mathcal{R}$, or equivalently for the coordinates $\{a_k\}$ or for the coordinates $\{\lambda_k\}$, a programme that is in practice achieved through some approximations. This mapping is obviously a projection in the sense that $\hat{D} \mapsto \hat{D}_{\rm R} \mapsto \hat{D}_{\rm R}$, but moreover the introduction of the metric (3) shows that the vector $\hat{D} - \hat{D}_{\rm R}$ in the space of states is perpendicular to the manifold $\mathcal{R}$ at the point $\hat{D}_{\rm R}$. This property is readily shown by writing, in this metric, the scalar product $\mathrm{Tr}\, \mathrm{d}\hat{D}\, \mathrm{d}\hat{X}'$ of the vector $\mathrm{d}\hat{D} = \hat{D} - \hat{D}_{\rm R}$ by an arbitrary vector $\mathrm{d}\hat{D}'$ in the tangent plane of $\mathcal{R}$. The latter is conjugate to any combination $\mathrm{d}\hat{X}'$ of observables $\hat{A}_k$, and this scalar product vanishes because $\mathrm{Tr}\, \hat{D}\hat{A}_k = \mathrm{Tr}\, \hat{D}_{\rm R}\hat{A}_k$. Thus the mapping $\hat{D} \mapsto \hat{D}_{\rm R}$ appears as an orthogonal projection, so that the relevant state $\hat{D}_{\rm R}$ associated with $\hat{D}$ may be regarded as its best possible approximation on the manifold $\mathcal{R}$.

## 6. Properties of the Metric

The metric tensor can be evaluated explicitly in a basis where the matrix $\hat{D}$ is diagonal. Denoting by $D_i$ its eigenvalues and by $\mathrm{d}D_{ij}$ the matrix elements of its variations, we obtain from (3)

$$\mathrm{d}s^2 = \mathrm{Tr} \int_0^\infty \mathrm{d}\xi \left( \frac{\mathrm{d}\hat{D}}{\hat{D} + \xi} \right)^2 = \sum_{ij} \frac{\ln D_i - \ln D_j}{D_i - D_j} \mathrm{d}D_{ij} \mathrm{d}D_{ji} \,. \tag{11}$$

(For $D_i = D_j$,whether or not $i = j$, the ratio is defined as $1/D_i$ by continuity.) In the same basis, the form (8) of the metric reads

$$\mathrm{d}s^2 = \frac{1}{Z} \sum_{ij} \frac{e^{X_i} - e^{X_j}}{X_i - X_j} \mathrm{d}X_{ij} \mathrm{d}X_{ji} - \left( \frac{\sum_i e^{X_i} \mathrm{d}X_{ii}}{Z} \right)^2 \,, \tag{12}$$

with $Z = \sum_i e^{X_i}$(For $X_i = X_j$, the ratio is $e^{X_i}$). The singularity of the metric (11) in the vicinity of vanishing eigenvalues of $\hat{D}$, in particular near pure states (end of Section 2), is not apparent in the representation (12) of this metric, because the mapping from $\hat{D}$ to $\hat{X}$ sends the eignevalue $X_i$ to $-\infty$ when $D_i$ tends to zero.

Let us compare the expression (11) with the corresponding classical metric, which is obtained by starting from Shannon's entropy instead of von Neumann's entropy. For discrete probabilities $p_i$, we have then $S\{p_i\} = -\sum_i p_i \ln p_i$ and hence the same definition $\mathrm{d}s^2 = -\mathrm{d}^2 S\{p_i\}$ as above of an entropy-based metric yields $\mathrm{d}s^2 = \sum_i \mathrm{d}p_i^2/p_i$, which is identified with the Fisher information metric. The present metric thus appears as the extension to quantum statistical mechanics of the Fisher metric

when the latter is interpreted in terms of entropy. In fact, the terms of (11) which involve the diagonal elements $i = j$ of the variations $\mathrm{d}\hat{D}$ reduce to $\mathrm{d}D_{ii}^2/D_i$. This result was expected since density matrices behave as probability distributions if both $\hat{D}$ and $\mathrm{d}\hat{D}$ are diagonal.

Let us more generally consider in (11), instead of solely diagonal variations $\mathrm{d}D_{ii}$, variations $\mathrm{d}D_{ij}$ with indices $i$ and $j$ such that $|D_i - D_j| \ll D_i + D_j$. The expansion of $D_i$ and $D_j$ around $\frac{1}{2}(D_i + D_j)$ in the corresponding ratios of (11) yields $(\ln D_i - \ln D_j)/(D_i - D_j) \sim 2/(D_i + D_j)$. The considered terms of (11) are therefore the same as in the Bures–Helstrom metric

$$\mathrm{d}s_{\mathrm{BH}}^2 = \sum_{ij} \frac{2}{D_i + D_j} \mathrm{d}D_{ij} \mathrm{d}D_{ji} \,, \tag{13}$$

introduced long ago as an extension to matrices of the Fisher metric [5]. We thus recover this Bures–Helstrom metric as an approximation of the present entropy-based metric $\mathrm{d}s^2 = -\mathrm{d}^2 S(\hat{D})$. For $n = 2$, $\mathrm{d}s_{\mathrm{BH}}^2$ is obtained from the expression (5) of $\mathrm{d}s^2$ by omitting the factor $\tanh^{-1} r/r$ entering the second term.

In order to express the properties of the Riemannian metric (3) in a general form, which will exhibit the tensor structure, we use a Liouville representation. There, the observables $\hat{O} = O_\mu \hat{\Omega}^\mu$, regarded as elements of a vector space, are represented by their coordinates $O_\mu$ on a complete basis $\hat{\Omega}^\mu$ of $n^2$ observables. The space of states is spanned by the dual basis $\hat{\Sigma}_\mu$, such that $\mathrm{Tr}\, \hat{\Omega}^\nu \hat{\Sigma}_\mu = \delta_\mu^\nu$, and the states $\hat{D} = D^\mu \hat{\Sigma}_\mu$ are represented by their coordinates $D_\mu$. Thus, the expectation value (1) is the scalar product $D^\mu O_\mu$. In the matrix representation which appears as a special case, $\mu$ denotes a pair of indices $i, j$, $\hat{\Omega}^\mu$ stands for $| \, j >< i \, |$, $\hat{\Sigma}_\mu$ for $| \, i >< j \, |$, $O_\mu$ denotes the matrix element $O_{ji}$ and $D^\mu$ the element $D_{ij}$. For the q-bit ($n = 2$) considered in Section 2, we have chosen the Pauli operators $\hat{\sigma}^\mu$ as basis $\hat{\Omega}^\mu$ for observables, and $\frac{1}{2}\hat{\sigma}_\mu$ as dual basis $\hat{\Sigma}_\mu$ for states, so that the coordinates $D^\mu = \mathrm{Tr}\, \hat{D}\hat{\Omega}^\mu$ of $\hat{D} = \frac{1}{2}(\hat{I} + r^\mu \hat{\sigma}_\mu)$ are the components $r^\mu$ of the vector $\mathbf{r}$ (The unit operator $\hat{I}$ is kept aside since $\hat{D}$ is normalised and since constants added to $\hat{X}$ are irrelevant). The function $F\{X\} = \ln \mathrm{Tr}\, e^{\hat{X}}$ of the coordinates $X_\mu$ of the observable $\hat{X}$, and the von Neumann entropy $S\{D\}$ as function of the coordinates $D^\mu$ of the state $\hat{D}$, are related by the Legendre transformation $F = S + D^\mu X_\mu$, and the relations (10) are expressed by $D^\mu = \partial F/\partial X_\mu$, $X_\mu = -\partial S/\partial D^\mu$. The metric tensor is given by

$$g^{\mu\nu} = \frac{\partial^2 F}{\partial X_\mu \partial X_\nu}, \quad g_{\mu\nu} = -\frac{\partial^2 S}{\partial D^\mu \partial D^\nu} \,, \tag{14}$$

and the correspondence issued from (7) between covariant and contravariant infinitesimal variations of $\hat{X}$ and $\hat{D}$ is implemented as $\mathrm{d}D^\mu = g^{\mu\nu}\mathrm{d}X_\nu$, $\mathrm{d}X_\mu = g_{\mu\nu}\mathrm{d}D^\nu$.

These expressions exhibit the Hessian nature of the metric. This property simplifies the expression of the Christoffel symbol, which reduces to

$$\Gamma_{\mu\nu\rho} = -\frac{1}{2}\frac{\partial^3 S}{\partial D^\mu \partial D^\nu \partial D^\rho} \,, \tag{15}$$

and which provides a parametric representation $\hat{D}(t)$ of the geodesics in the space of states through

$$\frac{\mathrm{d}^2 D^\mu}{\mathrm{d}t^2} + g^{\mu\sigma}\Gamma_{\sigma\nu\rho}\frac{\mathrm{d}D^\nu}{\mathrm{d}t}\frac{\mathrm{d}D^\rho}{\mathrm{d}t} = 0 \,. \tag{16}$$

Then, the Riemann curvature tensor comes out as

$$R_{\mu\rho\ \nu\sigma} = g^{\xi\zeta}(\Gamma_{\mu\sigma\xi}\Gamma_{\nu\rho\zeta} - \Gamma_{\mu\nu\xi}\Gamma_{\rho\sigma\zeta}) \,, \tag{17}$$

the Ricci tensor and the scalar curvature as

$$R_{\mu\nu} = g^{\rho\sigma}R_{\mu\rho\ \nu\sigma}, \quad R = g^{\mu\nu}R_{\mu\nu} \,, \tag{18}$$

We have noted that the classical equivalent of the entropy-based metric $\mathrm{d}s^2 = -\mathrm{d}^2 S$ is the Fisher metric $\sum_i \mathrm{d}p_i^2/p_i$, which as regards the curvature is equivalent to a Euclidean metric. While the space of classical probabilities is thus flat, the above equations show that the space of quantum states is curved. This curvature arises from the non-commutation of the observables, it vanishes for the completely disordered state $\hat{D} = \hat{I}/n$. Curvature can thus be used as a measure of the degree of classicality of a state.

## 7. Geometry of the Space of q-Bits

In the illustrative example of a q-bit, the operator $\hat{X} = \chi_\mu \hat{\sigma}^\mu$ associated with $\hat{D}$ is parameterised by the 3 components of the vector $\chi_\mu$ ($\mu = 1, 2, 3$), related to $\mathbf{r}$ by $\chi = \tanh^{-1} r$ and $\chi_\mu/\chi = r^\mu/r$. The metric tensor given by (5) is expressed as

$$\begin{aligned} g_{\mu\nu} &= K r_\mu r_\nu + \frac{\chi}{r}\delta_{\mu\nu} \,, \quad K \equiv \frac{1}{r}\frac{\mathrm{d}}{\mathrm{d}r}\frac{\chi}{r} = \frac{1}{r^2}\left(\frac{1}{1-r^2} - \frac{\chi}{r}\right) \,, \\ g^{\mu\nu} &= (1-r^2)p^{\mu\nu} + \frac{r}{\chi}q^{\mu\nu} \,. \end{aligned} \tag{19}$$

(We have defined $r_\mu = r^\mu$, $\delta_{\mu\nu} = \delta^\mu{}_\nu = \delta^{\mu\nu}$ so as to introduce the projectors $r^\mu r^\nu/r^2 \equiv p^{\mu\nu} \equiv \delta^{\mu\nu} - q^{\mu\nu}$ in the Euclidean 3-dimensional space, and thus to simplify the subsequent calculations.) In polar coordinates $\mathbf{r} = (r, \theta, \varphi)$, the infinitesimal distance takes the form

$$\mathrm{d}s^2 = \mathrm{d}r\mathrm{d}\chi + r\chi(\mathrm{d}\theta^2 + \sin^2\theta \mathrm{d}\varphi^2) \,. \tag{20}$$

We determine from (15) and (19) the explicit form

$$\Gamma_{\mu\nu\rho} = \frac{K}{2}\left(r_\mu\delta_{\nu\rho} + r_\nu\delta_{\mu\rho} + r_\rho\delta_{\mu\nu}\right) + \frac{1}{2r}\frac{\mathrm{d}K}{\mathrm{d}r} r_\mu r_\nu r_\rho \tag{21}$$

of the Christoffel symbol. By raising its first index with $g^{\mu\nu}$ and using polar coordinates, we obtain from (16) the equations of geodesics for $n = 2$. Within the Poincaré–Bloch sphere the geodesics are deduced by rotations from a one-parameter family of curves which lie in the $\theta = \frac{1}{2}\pi$, $|\varphi| \leq \frac{1}{2}\pi$ half-plane and which are symmetric with respect to the $\varphi = 0$ axis. This family is characterized by the equations (where $\chi = \tanh^{-1} r$):

$$\frac{\mathrm{d}^2 r}{\mathrm{d}t^2} + \frac{r}{1-r^2}\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2 - \frac{r}{2}\left[1 + \frac{\chi}{r}\left(1-r^2\right)\right]\left(\frac{\mathrm{d}\varphi}{\mathrm{d}t}\right)^2 = 0 \,, \tag{22}$$

$$\frac{\mathrm{d}^2\varphi}{\mathrm{d}t^2} + \frac{1}{r}\frac{\mathrm{d}r}{\mathrm{d}t}\frac{\mathrm{d}\varphi}{\mathrm{d}t} + \frac{1}{\chi}\frac{\mathrm{d}\chi}{\mathrm{d}t}\frac{\mathrm{d}\varphi}{\mathrm{d}t} = 0 \,, \tag{23}$$

and the boundary conditions at $t = 0$:

$$r\,(0) = a\,, \quad \varphi\,(0) = 0\,, \quad \frac{\mathrm{d}r\,(0)}{\mathrm{d}t} = 0\,, \quad \frac{\mathrm{d}\varphi\,(0)}{\mathrm{d}t} = \frac{1}{k}\,, \quad k^2 = a\tanh^{-1} a\,. \tag{24}$$

Equation (23) provides, using the boundary conditions (24):

$$\frac{\mathrm{d}\varphi}{\mathrm{d}t} = \frac{k}{r\chi}\,. \tag{25}$$

Insertion of (25) into (22) gives rise to an equation for $r\,(t)$, which can be integrated by regarding $t$ as a function of $\zeta = \arcsin r$. One obtains:

$$\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2 = \left(1 - r^2\right)\left(1 - \frac{k^2}{r\chi}\right)\,. \tag{26}$$

The scale of $t$ has been fixed by relating to $r\,(0)$ the boundary condition (24) for $\mathrm{d}\varphi\,(0)\,/\mathrm{d}t$, a choice which ensures that $\mathrm{d}s^2 = \mathrm{d}r\mathrm{d}\chi + r\chi\mathrm{d}\varphi^2 = \mathrm{d}t^2$, and hence that the parameter $t$ measures the distance along geodesics.

For $k = 0$, we obtain $r = |\sin t|$, $\varphi = \pm\pi/2$. Thus, the longest geodesics are the diameters of the Poincaré–Bloch sphere. We find the value $\pi$ for their "length", that is, for the geodesic distance between two orthogonal pure states. At the other extreme, when the middle point $r = a$, $\varphi = 0$ of a geodesic lies close to the surface $r = 1$ of the sphere, the asymptotic form of the equation (26) is solved as

$$t = \pm 2k\sqrt{\pi}\mathrm{e}^{-k^2}\,\mathrm{erf}\,\xi\,, \quad \xi = \sqrt{\frac{1}{2}\ln\frac{1-a}{1-r}}\,, \quad k^2 = \frac{1}{2}\ln\frac{2}{1-a} \tag{27}$$

(by taking $\xi$ as variable instead of $r$). The determination of the explicit equations of such short geodesic curves is achieved by integrating (25) into

$$\varphi = \frac{t}{k} = \pm 2\sqrt{\pi}\mathrm{e}^{-k^2}\,\mathrm{erf}\,\xi\,. \tag{28}$$

From (27) and (28) we can determine the geodesic distance between two neighbouring pure states $\hat{D}' = |\psi' >< \psi'|$ and $\hat{D}'' = |\psi'' >< \psi''|$ represented by the points $r_{\max} = 1$, $\varphi_{\max} = \pm\frac{1}{2}\delta\varphi$ with $\delta\varphi$ small. At these two points, we have $\xi \to \infty$, $\mathrm{erf}\,\xi = 1$, and this determines $k$ in terms of $\frac{1}{2}\delta\varphi$ through (28). The length of the geodesic that joins them, given by (27), is:

$$\delta s^2 = \delta\varphi^2 \ln\frac{4\sqrt{\pi}}{\delta\varphi}\,, \quad \delta\varphi = \arccos|< \psi' \mid \psi'' >|\,. \tag{29}$$

Thus, in spite of its singularity for $r = 1$, the present 3-dimensional metric (5) in the space $r$, $\theta$, $\varphi$ defines distances between pure states represented by points on the surface $r = 1$ of the Poincaré–Bloch sphere. However, It should be noted that the presence of the logarithmic factor in (29) forbids such distances to be generated by a 2-dimensional metric in the space $\theta$, $\varphi$. In fact, the distance (29) is measured along a geodesic that penetrates the sphere $r = 1$, because no geodesic is tangent to the surface of this sphere nor lies on its surface.

In contrast, all geodesics produced by the Bures–Helstrom metric are tangent to the surface of the sphere, or are its great circles. They are given by Equations (25) and (26), where $\chi$ is replaced by $r$ and $k$ by $a$; the solution of these equations provides the ellipses

$$r\cos\varphi = a\cos t\ , \quad r\sin\varphi = \sin t\ . \tag{30}$$

Here as above, the largest distance $\pi$ is reached for orthogonal pure states represented by opposite points on the sphere, but now a peculiarity occurs. Whereas the metric $\mathrm{d}s^2 = -\mathrm{d}^2S$ produces a single geodesic, the diameter joining these two points (with "length" $\pi$), the Bures metric produces a double infinity of geodesics, the half-ellipses (30) having as long axis this diameter, and having all the same "length" $\pi$. Other pairs of pure states are joined by geodesics which are arcs of great circles, and their Bures distance $\delta s_{\mathrm{BH}} = \delta\varphi$ is identified with the ordinary length of the arc. Here for $n = 2$ as in the general case, the 3-dimensional Bures–Helstrom metric admits a restriction to pure states generated by a 2-dimensional metric, which is identified with the quantum Fubini–Study metric, itself defined only for pure states by $s_{\mathrm{FS}} = \arccos|<\psi'\mid\psi''>| = \frac{1}{2}s_{\mathrm{BH}}$.

Returning to the metric $\mathrm{d}s^2 = \mathrm{d}^2S$, the Riemann curvature is obtained from (17) as

$$\begin{aligned} R^{\mu}{}_{\rho\,\nu\sigma} = \frac{K}{4}\Big[&(r^2+\frac{r}{\chi}-1)(q^{\mu}{}_{\sigma}q_{\nu\rho}-q^{\mu}{}_{\nu}q_{\rho\sigma})+(r^2-\frac{r}{\chi}+1)(p^{\mu}{}_{\sigma}q_{\nu\rho}-p^{\mu}{}_{\nu}q_{\rho\sigma}) \\ &+\frac{r}{\chi}\frac{1}{1-r^2}(r^2-\frac{r}{\chi}+1)(q^{\mu}{}_{\sigma}p_{\nu\rho}-q^{\mu}{}_{\nu}p_{\rho\sigma})\Big]\ . \end{aligned} \tag{31}$$

Contracting with $g^{\rho\sigma}$ the indices of (30) as in (18), we finally derive the Ricci curvature

$$R^{\mu}{}_{\nu} = -\frac{Kr}{2\chi}\left(r^2\delta^{\mu}{}_{\nu}+\frac{\chi-r}{\chi}p^{\mu}{}_{\nu}\right)\ , \tag{32}$$

and the scalar curvature

$$R = -\frac{Kr}{2\chi}\left(3r^2+\frac{\chi-r}{\chi}\right)\ . \tag{33}$$

Both are negative in the whole Poincaré sphere. In the limit $r \to 0$, the curvature $R$ vanishes as $R \sim -\frac{10}{9}r^2$, as expected from the general argument of Section 2: a weakly polarised spin behaves classically. At the other extreme $r \to 1$, $R$ behaves as $R \sim -2\left[(1-r)\mid\ln(1-r)\mid\right]^{-1}$; it diverges, again as expected: pure states have the largest quantum nature.

The metric $\mathrm{d}s^2 = -\mathrm{d}^2S$, introduced above in the context of quantum mechanics for mixed states (and their pure limit) and information theory, might more generally be useful to characterise distances in spaces of positive matrices.

**Conflicts of Interest**

The author declares no conflict of interest.

## References


1. Thirring, W. Quantum Mechanics of Large Systems. In *A Course of Mathematical Physics*; Volume 4; Springler-Verlag: New York, NY, USA, 1983.
2. Balian, R.; Alhassid, Y.; Reinhardt, H. Dissipation in many-body systems: A geometric approach based on information theory. *Phys. Rep.* **1986**, *131*, 1–146.
3. Balian, R. Incomplete descriptions and relevant entropies. *Am. J. Phys.* **1999**, *67*, 1078–1090.
4. Balian, R. Information in statistical physics. *Stud. Hist. Philos. Mod. Phys.* **2005**, *36*, 323–353.
5. Bures, D. An extension of Kakutani's theorem. *Trans. Am. Math. Soc.* **1969**,*135*, 199–212.

# Geometry of Fisher Information Metric and the Barycenter Map

**Mitsuhiro Itoh and Hiroyasu Satoh**

**Abstract:** Geometry of Fisher metric and geodesics on a space of probability measures defined on a compact manifold is discussed and is applied to geometry of a barycenter map associated with Busemann function on an Hadamard manifold $X$. We obtain an explicit formula of geodesic and then several theorems on geodesics, one of which asserts that any two probability measures can be joined by a unique geodesic. Using Fisher metric and thus obtained properties of geodesics, a fibre space structure of barycenter map and geodesical properties of each fibre are discussed. Moreover, an isometry problem on an Hadamard manifold $X$ and its ideal boundary $\partial X$—for a given homeomorphism $\Phi$ of $\partial X$ find an isometry of $X$ whose $\partial X$-extension coincides with $\Phi$—is investigated in terms of the barycenter map.

Reprinted from *Entropy*. Cite as: Itoh, M.; Satoh, H. Geometry of Fisher Information Metric and the Barycenter Map. *Entropy* **2015**, *17*, 1814–1849.

## 1. Introduction

The aim of this article is to deal with two subjects related with information geometry. One is Fisher metric $G$ defined on a space $\mathcal{P}(M)$ of probability measures having continuous positive density function over a connected, compact manifold $M$, and another one is barycenter map from $\mathcal{P}(\partial X)$ to an Hadamard manifold $X$, where $\partial X$ is the ideal boundary of $X$. This article is an extended version of [1] presented at MaxEnt 2014, Amboise, France.

The Fisher metric $G$, remarkably important in information geometry, is defined in a natural way. The metric $G$ is push-forward invariant, and has an explicit formula of Levi–Civita connection and its sectional curvature is constant $1/4$, as shown in [2] by T. Friedrich.

Before introducing main results, we will explain motivation and background of our study.

An $n$-dimensional Hadamard manifold $(X, g)$ is diffeomorphic to $\mathbf{R}^n$, and hence to an open ball $D^n$, whose actual boundary is $S^{n-1}$. $X$ admits also the ideal boundary $\partial X$ as a quotient space of oriented geodesics on $X$. Then, we are able to consider Dirichlet problem at boundary $\partial X$; given a $f \in C^0(\partial X)$, find a solution $u = u(x)$ on $X$ satisfying $\Delta u = 0, \quad u|_{\partial X} = f$. Using the fundamental solution $P = P(x, \theta)$, called Poisson kernel, when its existence is guaranteed, the solution is described as

$$u(x) = \int_{\theta \in \partial X} P(x, \theta)\, d\theta, \quad x \in X. \tag{1}$$

Refer to [3] for precise definition of Poisson kernel. We obtain then a probability measure $P(x, \theta)d\theta$ on $\partial X$ parametrized in $x \in X$ and have a map, called Poisson kernel map $\Theta : X \to \mathcal{P}(\partial X)$.

**Theorem 1** ([4–6]). *Let $(X, g)$ be an $n$-dimensional Damek-Ricci space. Then the map $\Theta$ is homothetic with respect to the Fisher metric $G$ and $g$; $\Theta^* G = \frac{Q}{n} g$ where $Q$ is volume entropy of $(X, g)$. Further $\Theta$ is a harmonic map.*

Here, for volume entropy $Q$ refer to §4. The quantity $Q$ is an invariant of Riemannian geometry which is closely related to the topological entropy of geodesic flow ([7,8]). Refer to [9] with respect to volume entropy treated in a framework of information geometry.

In the theorem a Damek-Ricci space is a solvable Lie group of a left invariant metric, one dimensional extension of a generalized Heisenberg group. Refer to [10] for details. A Damek-Ricci space is a harmonic, Einstein Hadamard manifold and any rank one symmetric space of non-compact type, namely hyperbolic spaces over the real numbers $\mathbf{R}$, the complex numbers $\mathbf{C}$, the quaternions $\mathbf{H}$ and $16$-dimensional one over Cayley numbers $\mathbf{O}$ are also Damek-Ricci spaces. With respect to Theorem 1, we have the following theorem.

**Theorem 2** ([5]). *Let $(X, g)$ be an Hadamard manifold which is equipped with Poisson kernel $P(x, \theta)$. Assume that the map $\Theta : X \to \mathcal{P}(\partial X)$ is homothetic; $\Theta^* G = C\, g$, $C > 0$, and harmonic with respect to the metrics $G$ and $g$. Then $(X, g)$ is asymptotically harmonic and satisfies visibility axiom. Moreover, $C = Q/n$ and the Poisson kernel has the form $P(x, \theta) = \exp\{-Q\, B_\theta(x)\}$ in terms of Busemann function $B_\theta$ on $X$.*

The terminology with respect to asymptotical harmonicity, visibility axiom and Busemann function will be explained in the subsequent sections.

Remark that the equality $C = Q/n$ is derived from asymptotical formula related with mean curvature of geodesic spheres and mean curvature of corresponding horospheres, level hypersurfaces of Busemann function ([11]).

With respect to these theorems we are interested in characterization of Damek-Ricci space from information geometry, especially from a viewpoint of Fisher metric $G$, since a Damek-Ricci space is a counterexample of Lichnerowicz conjecture of non-compact version ([12]) and its characterization is only given by Heber in [13] by Lie group theory argument. By approaching from a viewpoint of the ideal boundary $\partial X$, we focus on barycenter of probability measures on $\partial X$ with respect to Busemann function and shed a light on information geometry of barycenter map $\mathrm{bar} : \mathcal{P}(\partial X) \to X$.

## 2. Main Results and Conclusive Remarks

Before entering into the detailed argument, we give an outline of main results and remarks.

In Section 3 we deal with several results on Fisher metric $G$ and also on geodesic, a basic notion of geometry, defined on $(\mathcal{P}(M), G)$. We give a formula of geodesic $\mu(t) = \exp_\mu t\tau$ on $\mathcal{P}(M)$ in a simple form (Theorem 9);

$$\mu(t) = \left( \cos \frac{t}{2} + \sin \frac{t}{2}\, \frac{d\tau}{d\mu}(x) \right)^2 \mu, \quad x \in M$$

for an initial condition; $\mu(0) = \mu, \dot{\mu}(0) = \tau$ ($|\tau|_{G,\mu} = 1$). Here, $(d\tau/d\mu)(x)$ denotes Radon-Nikodym derivative of $\tau$ with respect to $\mu$. From this, it is concluded in Corollary 2 that any geodesic is periodic, of period $2\pi$, while not definable over $\mathbf{R}$. Moreover, from this formula which is an improvement of the formula given by T. Friedrich ([2]) we obtain

**Theorem 3.** *Let $\mu$ and $\mu^*$ be arbitrary distinct probability measures in $\mathcal{P}(M)$. Then, a curve $t \in \mathbf{R} \mapsto \mu(t) \in \mathcal{P}(M)$ defined by*

$$\mu(t) = \exp_\mu t\tau = \left(\cos\frac{t}{2} + \sin\frac{t}{2}\frac{d\tau}{d\mu}(x)\right)^2 \mu \tag{2}$$

*is a unique geodesic such that $\mu(0) = \mu$ and $\mu(\ell) = \mu^*$. Here $\ell = \ell(\mu, \mu^*)$ is defined by (4) and $\tau$ is a unit tangent vector at $\mu$ given by*

$$\tau = \frac{1}{\sin\frac{\ell}{2}}\left(\sqrt{\frac{d\mu^*}{d\mu}}(x) - \int_{y\in M}\sqrt{\frac{d\mu^*}{d\mu}}(y)\, d\mu(y)\right)\mu(x). \tag{3}$$

This theorem asserts that any $\mu$, $\mu^*$, $\mu \neq \mu^*$ can be joined by a unique geodesic. The quantity $\ell = \ell(\mu, \mu^*)$, $0 < \ell < \pi$ is defined as

$$\cos\frac{\ell}{2} = \int_{x\in M}\sqrt{\frac{d\mu^*}{d\mu}}(x)\, d\mu(x), \tag{4}$$

giving an apparent length of a geodesic joining $\mu$ and $\mu^*$. Notice that the RHS is an $f$-divergence-like quantity with respect to $f(u) = \sqrt{u}$ (refer to [14]).

Another main subject is information geometry of barycenter map by applying results of Fisher information geometry, thus obtained in Section 3. Related results on barycenter map will be explained in Section 4 and Section 5.

Let $(X, g)$ be an Hadamard manifold with a Riemannian metric $g$, a simply connected, complete Riemannian manifold of non-positive curvature. Then, it admits the ideal boundary $\partial X$ and by using a probability measure defined on $\partial X$, we consider a function, a $\mu$-average Busemann function $\mathbf{B}_\mu : X \to \mathbf{R}$;

$$\mathbf{B}_\mu(x) = \int_{\theta\in\partial X} B_\theta(x)\, d\mu(\theta)$$

whose critical point is called a barycenter of a probability measure $\mu$ so that we have a map, barycenter map, from a space $\mathcal{P}(\partial X)$ of probability measures on $\partial X$ to an Hadamard manifold $X$. Here, the integrand is a normalized Busemann function (for its detailed argument see Section 4).

Recall, here, an original definition of a barycenter, a center of mass, as follows. Let $y_1, \ldots, y_n$ be points of a Euclidean space $\mathbf{R}^3$ and $\mu_1, \ldots, \mu_n$ be non-negative real numbers satisfying $\sum_i \mu_i = 1$. A point $p$ of $\mathbf{R}^3$ is called a barycenter of $y_i, i = 1, \ldots, n$ of weights $\mu_i, i = 1, \ldots, n$, when $p$ satisfies

$$p = \sum_{i=1}^n \mu_i y_i \quad \text{or} \quad \sum_{i=1}^n \mu_i (y_i - p) = 0.$$

A barycenter is defined also by a critical point of a function on $\mathbf{R}^3$; $f : \mathbf{R}^3 \to \mathbf{R}$; $f(q) = \sum_{i=1}^n \mu_i d^2(q, y_i)$.

This definition of barycenter for a finite points of $\mathbf{R}^3$ with weights with respect to the square-distance can be generalized as one for points of $\mathbf{R}^3$ distributed continuously over a bounded set $D$ of $\mathbf{R}^3$;

$$f : \mathbf{R}^3 \to \mathbf{R}; \quad f(q) = \int_{\mathbf{R}^3} d^2(q, x)\, \mu(x)\, dx,$$

where $\mu = \mu(x)$ is a non-negative function with $\mathrm{supp}(\mu) \subset D$ satisfying $\int_{\mathbf{R}^3} \mu(x)\, dx = 1$. A critical point of $f$ can be regarded as a barycenter of a probability measure $\mu(x)\, dx$. A famous theorem of E. Cartan is regarded as a barycenter theorem ([15]). A choice of testing function $d^2(x, y)$ is not essential. Convexity of testing function is crucial in a theory of barycenter. In our study we deal with barycenter with respect to Busemann function, a convex testing function, by following the idea of Douady, Earle ([16]) and Besson, Courtois and Gallot ([8,17]). Refer to [18,19] for studies and results on barycenter of square-distance and of distance over a Riemannian manifold. Refer also to [20] in this direction, which is a reference comment due to Professor M. Gromov at the conference.

In our situation, the existence of barycenter for any $\mu \in \mathcal{P}(\partial X)$ is assured in Theorem 12, when $(X, g)$ satisfies visibility axiom (for precise definition see Definition 4 and refer to [21]) and Busemann function $B_\theta(x)$ on $X$ is continuous with respect to $\theta \in \partial X$. Uniqueness of barycenter for any $\mu$ is assured, when, for some $\mu_0$, average Hessian $\nabla d\mathbf{B}_{\mu_0}$ of $\mathbf{B}_{\mu_0}$ is positive definite everywhere on $X$ (Proposition 6). Thus, we have the barycenter map $\mathrm{bar} : \mathcal{P}(\partial X) \to X$; $\mu \mapsto y$, by assigning to $\mu$ a barycenter point $y$ of $\mu$ . This map turns out to be surjective, when $(X, g)$ admits Busemann-Poisson kernel $P(x, \theta) = \exp\{-Q\, B_\theta(x)\}$, Poisson kernel of Busemann type. Denote by $\mu_x$ the probability measure $P(x, \theta)\, d\theta$. Then, $\mathrm{bar}(\mu_x) = x$ for any $x \in X$. Busemann-Poisson kernel ensures also the uniqueness of barycenter for any $\mu$ from the identity (see Theorem 14)

$$(\nabla d\mathbf{B}_{\mu_x})_x(u, v) = Q\, G_{\mu_x}(\nu_x^{\mu_x} u, \nu_x^{\mu_x} v), \;\; u, v \in T_x X,\, x \in X$$

where, $G_{\mu_x}$ is Fisher metric at the tangent space $T_{\mu_x}\mathcal{P}(\partial X)$, and $\nu_x^{\mu_x} : T_x X \to T_{\mu_x}\mathcal{P}(\partial X)$ is an injective linear map associated to $\mu$ and a point $x = \mathrm{bar}(\mu)$ (for its definition see Section 4).

The map $\mathrm{bar} : \mathcal{P}(\partial X) \to X$, being surjective gives us a projection of a fibre space whose total space is $\mathcal{P}(\partial X)$ and base space is $X$ with fibres $\{\mathrm{bar}^{-1}(x);\, x \in X\}$. The image of the linear map $\nu_x^\mu$ for $\mu$ and $x = \mathrm{bar}(\mu)$ yields a subspace of $T_\mu\mathcal{P}(\partial X)$, normal to $T_\mu \mathrm{bar}^{-1}(x)$, the subspace tangent to a fibre $\mathrm{bar}^{-1}(x)$ so that $T_\mu\mathcal{P}(\partial X)$ splits into a $G$-orthogonal direct sum (Theorem 15);

$$T_\mu\mathcal{P}(\partial X) = T_\mu \mathrm{bar}^{-1}(x) \oplus \mathrm{Im}\ \nu_x^\mu.$$

$T_\mu \mathrm{bar}^{-1}(x)$ and $\mathrm{Im}\ \nu_x^\mu$ are the vertical, horizontal subspaces of $T_\mu\mathcal{P}(\partial X)$, respectively. Here, $\dim \mathrm{Im}\ \nu_x^\mu = \dim X$. Remark that the fibration asserted here is infinitesimal.

Each fibre $\mathrm{bar}^{-1}(x)$, $x \in X$ is a path-connected submanifold of $\mathcal{P}(\partial X)$. Its geometry is investigated in terms of the second fundamental form;

$$H_\mu : T_\mu \mathrm{bar}^{-1}(x) \times T_\mu \mathrm{bar}^{-1}(x) \to \mathrm{Im}\ \nu_x^\mu; \quad H_\mu(\tau, \tau_1) = (\nabla_\tau \tau_1)^\perp, \tag{5}$$

which is the normal component of $\nabla_\tau \tau_1$, the covariant derivative of $\tau_1$ in direction to $\tau$ with respect to the Levi–Civita connection $\nabla$. Refer to [22,23] for definition of the second fundamental form. Applying the results concerning geodesics on $\mathcal{P}(\partial X)$ given in Section 3 to a submanifold $\mathrm{bar}^{-1}(x)$, we are able to determine a geodesic $\mu(t) = \exp_\mu t\tau$ which is entirely contained in $\mathrm{bar}^{-1}(x)$ as

**Theorem 4.** *Let $\mu(t) = \exp_\mu t\tau$ be a unit speed geodesic, of $\mu(0) = \mu$, $\dot{\mu}(0) = \tau$, $|\tau|_{G,\mu} = 1$. Then, $\mu(t)$ lies entirely on fibre $\mathrm{bar}^{-1}(x)$ if and only if, $\mu \in \mathrm{bar}^{-1}(x)$, $\tau \in T_\mu \mathrm{bar}^{-1}(x)$ and $H_\mu(\tau, \tau) = 0$.*

Remark that the equation $H_\mu(\tau, \tau) = 0$ on $\tau$ is written down in a manner of information geometry as $\int_{\theta\in\partial X} (dB_\theta)_x(u)\, (d\tau/d\mu)^2(\theta)\, d\mu(\theta) = 0$.

Moreover, by applying Theorem 11 in Section 3, it is possible to assert the following theorem.

**Theorem 5.** *Let $\mu$, $\mu^* \in \mathrm{bar}^{-1}(x)$. Then, a geodesic joining $\mu$ and $\mu^*$ is contained completely in the same fibre $\mathrm{bar}^{-1}(x)$ if and only if $\mu$ and $\mu^*$ fulfill*

$$\int_{\theta\in\partial X} (dB_\theta)_x(u) \sqrt{\frac{d\mu^*}{d\mu}}(\theta) d\mu = 0$$

*for any $u \in T_x X$.*

Let $\phi$ be an isometry of an Hadamard manifold $(X, g)$. Then, a $\mu$-average Busemann function $\mathbf{B}_\mu$ satisfies a cocycle formula with respect to $\phi$;

$$\mathbf{B}_\mu(\phi^{-1}x) = \mathbf{B}_{(\hat{\phi}_\sharp)\mu}(x) + \mathbf{B}_\mu(\phi^{-1}x_o), \quad x \in X,\ \mu \in \mathcal{P}(\partial X),$$

where $\phi^{-1}$ is the inverse of $\phi$, and $\hat{\phi} : \partial X \to \partial X$ is a $\partial X$-extension of $\phi$ and $\hat{\phi}_\sharp$ is a push-forward induced by $\hat{\phi}$. See Theorem 18. From this formula we have

$$\mathrm{bar}(\hat{\phi}_\sharp \mu) = \phi(\mathrm{bar}(\mu)), \quad \mu \in \mathcal{P}(\partial X)$$

from which each fibre $\mathrm{bar}^{-1}(x)$ is mapped by $\hat{\phi}_\sharp$ to a fibre $\mathrm{bar}^{-1}(\phi x)$ over $\phi x$.

By the aid of information geometry we are able to apply above results to an isometry problem; given a homeomorphism $\Phi$ of $\partial X$, find an isometry $\phi$ of $(X, g)$ whose $\partial X$-extension coincides with $\Phi$. With respect to this problem we consider a bijective map $\phi$ of $X$ satisfying $\mathrm{bar}(\Phi_\sharp \mu) = \phi(\mathrm{bar}(\mu))$ for any $\mu \in \mathcal{P}(\partial X)$ ( we call such a map $\phi$ as barycentrically associated to $\Phi$ ).

The following theorem gives us an answer to this problem, even partial, provided there exists a cross section of the fibre space $\mathrm{bar} : \mathcal{P}(\partial X) \to X$ enjoying commutativity properties.

**Theorem 6.** *Let $\varphi : X \to X$ be a $C^1$-map barycentrically associated to a homeomorphism $\Phi : \partial X \to \partial X$. Assume that there exists a cross section $\Sigma : X \to \mathcal{P}(\partial X)$ of the fibre space $\mathrm{bar} : \mathcal{P}(\partial X) \to X$, a map satisfying $\mathrm{bar} \circ \Sigma = id_X$ such that a fibrewise diagram commutes*

$$\begin{array}{ccc} \mathcal{P}(\partial X) & \xrightarrow{\Phi_\sharp} & \mathcal{P}(\partial X) \\ \uparrow \Sigma & & \uparrow \Sigma \\ X & \xrightarrow{\varphi} & X \end{array}$$

*and a diagram of tangent space level commutes*

$$\begin{array}{ccc} T_xX & \stackrel{\nu_x^{\mu_x}}{\longrightarrow} & T_{\mu_x}\mathcal{P}(\partial X) \\ \downarrow\ (\varphi_*)_x & & \downarrow\ \Phi_\sharp \\ T_{\varphi x}X & \stackrel{\nu_{\varphi x}^{\mu_{\varphi x}}}{\longrightarrow} & T_{\mu_{\varphi x}}\mathcal{P}(\partial X) \end{array} \tag{6}$$

*where we denote by* $\mu_x := \Sigma(x) \in \mathcal{P}(\partial X)$. *Then,* $\varphi$ *is an isometry of* $(X, g)$ *and* $\partial X$*-extension* $\hat{\varphi}$ *of* $\varphi$ *coincides with* $\Phi$.

A particularly significant cross section $\Sigma$ is given by a Poisson kernel map; $\Theta : X \to \mathcal{P}(\partial X)$; $x \mapsto P(x, \theta)d\theta = \exp\{-QB_\theta(x)\}\ d\theta$, where $P(x, \theta)$ is a Busemann-Poisson kernel on $(X, g)$. The differential map $(\Theta_*)_x$ of $\Theta$ fulfills $(\Theta_*)_x(u) = -\ Q\ \nu_x^{\mu_x}(u)$, $u \in T_xX$ in terms of the linear map $\nu_x^{\mu_x}$, $\mu_x := P(x, \theta)d\theta$, so that we have

**Corollary 1** ([24]). *Let* $\Phi : \partial X \to \partial X$ *be a homeomorphism of* $\partial X$ *and* $\varphi : X \to X$ *be a* $C^1$*-map. Assume the following diagram commutes with respect to Poisson kernel map* $\Theta$*;*

$$\begin{array}{ccc} \mathcal{P}(\partial X) & \stackrel{\Phi_\sharp}{\longrightarrow} & \mathcal{P}(\partial X) \\ \uparrow\ \Theta & & \uparrow\ \Theta \\ X & \stackrel{\varphi}{\longrightarrow} & X \end{array} \tag{7}$$

*Then,* $\varphi$ *is an isometry of* $(X, g)$ *and its* $\partial X$*-extension coincides with* $\Phi$ *of* $\partial X$.

Theorem 6 is a generalization of Corollary 1, a main result of [24,25].

The article is organized as follows. In Section 3 we introduce basic notions of information geometry of a space $\mathcal{P}(M)$ of probability measures on a compact manifold $M$ and define Fisher metric $G$ on it. We show several useful theorems on the Levi–Civita connection and geodesics on $\mathcal{P}(M)$ with detailed proofs. We derive in Section 4 fundamental properties of Busemann function on an Hadamard manifold, preliminarily. By using them, we investigate existence and uniqueness of barycenter of a probability measure, following a proof given in [8]. We define the barycenter map and develop information geometry of this map. A fibration theorem is similar to our earlier paper [25]. However, geodesical arguments on fibres develop further the arguments of [25], by applying the results of geodesics on $\mathcal{P}(M)$ in Section 3. Finally, in Section 5, we treat an isometry problem for an Hadamard manifold and give a proof of Theorem 6.

## 3. A Space of Probability Measures and Fisher Metric

### *3.1. A Space of Probability Measures*

Let $M$ be a connected, compact smooth manifold. Let $\mathcal{B}(M)$ be the family of Borel sets of $M$. Here $\mathcal{B}(M)$ is a family of subsets of $M$ which satisfies the following; (i) $\mathcal{B}(M)$ is a $\sigma$-family of $M$, (ii) every open subset of $M$ is an element of $\mathcal{B}(M)$ and (iii) if $\mathcal{F}$ is a family of subsets of $M$ satisfying (i), (ii), then $\mathcal{F} \subset \mathcal{B}(M)$.

A function $P : \mathcal{B}(M) \to \mathbf{R}$ is called a probability measure of a measurable space $(M, \mathcal{B}(M))$, or a probability measure on $M$, when $P$ satisfies

(i) $P(A) \geq 0$ for any $A \in \mathcal{B}(M)$, $P(M) = 1$ and $P(\emptyset) = 0$.

(ii) Let $\{E_j \mid j = 1, \ldots, \}$ be a countable sequence of sets of $\mathcal{B}(M)$ satisfying $E_i \cap E_j = \emptyset$ for any $i, j$, $i \neq j$. Then

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i).$$

A smooth manifold $M$, even in an unorientable case, admits a measure induced by the volume measure of $M$. We normalize this measure and denote this normalized measure by $d\,\theta$. The measure $d\,\theta$ is a probability measure on $M$.

For example, let $M = S^{n-1} = \{x \in \mathbf{R}^n \ ; \ |x| = 1\}$ be a unit $(n-1)$-sphere and let $d\omega$ be the standard $(n-1)$-spherical volume measure on $S^{n-1}$. Then $d\theta = (1/A_{n-1})\, d\omega$ is the normalized measure, where $A_{n-1}$ is the volume of $S^{n-1}$.

Let $\mu$, $\mu_1$ be probability measures on $M$. $\mu$ is called to be absolutely continuous with respect to $\mu_1$, if $\mu(A) = 0$ for any $A \in \mathcal{B}(M)$, whenever $\mu_1(A) = 0$.

Let $\mu$ be a probability measure on $M$, absolutely continuous with respect to $d\theta$. Then, from Radon-Nikodym theorem ([26]) there exists a $p \in L^1(M, d\theta)$ such that $\mu$ is represented as $\mu = p\, d\theta$, namely $\mu$ satisfies

$$\mu(A) = \int_{x \in A} p(x)\, d\theta(x), \quad \forall A \in \mathcal{B}(M).$$

The probability density function $p = p(x)$ is called Radon-Nikodym derivative of $\mu$ with respect to $d\theta$, written by $p = d\mu/d\theta$.

We denote by $\mathcal{P}(M)$ a space of probability measures $\mu$ on $M$, $d\theta$-absolutely continuous (denoted by $\mu \ll d\theta$) such that $\mu$ has a positive continuous density function $p = p(x)$, *i.e.*, $p \in C^0(M)$, $p(x) > 0$ for any $x \in M$.

A manifold $M$ admits an $L^2$-function space $L^2(M, d\theta)$ as

$$L^2(M, d\theta) = \left\{ h : M \to \mathbf{R} \, ; \, \int_M h^2(x)\, d\theta(x) < \infty \right\}.$$

We notice that there exists a natural embedding

$$\rho : \mathcal{P}(M) \to L^2(M, d\theta); \quad \mu = p\, d\theta \mapsto \sqrt{p} = \sqrt{\frac{d\mu}{d\theta}}. \tag{8}$$

By using this embedding we induce a topology on $\mathcal{P}(M)$. Remark that a sequence $\{\mu_i\}$ of $\mathcal{P}(M)$ does not necessarily admit a limit inside $\mathcal{P}(M)$.

Let $\mu$, $\mu_1$ be probability measures in $\mathcal{P}(M)$. Then we can join $\mu$ and $\mu_1$ by a path $\mu(t) = (1-t)\mu + t\mu_1$, $t \in [0, 1]$ inside $\mathcal{P}(M)$.

Differentiate $\mu(t)$ as a curve in $\mathcal{P}(M)$ to have

$$\frac{d}{dt}\ ((1-t)\mu + t\mu_1)) = \mu_1 - \mu,$$

which is a measure on $M$, represented as $\mu_1 - \mu = (p_1 - p)\, d\theta$ satisfying

$$\int_M d(\mu_1 - \mu) = 0.$$

Based on this fact, a tangent space $T_\mu\mathcal{P}(M)$ of $\mathcal{P}(M)$ at $\mu$ is defined as

$$T_\mu\mathcal{P}(M) = \left\{ \tau = q(x)\, d\theta(x)\ ;\ q \in C^0(M),\ \int_M q(x)\, d\theta(x) = 0 \right\} \tag{9}$$

Notice that the RHS (right hand side) of (9) is independent of $\mu$. So, if we denote by $\mathcal{V}$ the RHS of (9), then $\mathcal{V}$ is an infinite dimensional vector space and an arbitrary $\tau \in \mathcal{V}$ induces at any $\mu \in \mathcal{P}(M)$ a curve $\mu + t\tau \in \mathcal{P}(M)$ for $t \in (-\varepsilon, \varepsilon)$ with a sufficiently small $\varepsilon$.

We define a curve $c : (a, b) \to \mathcal{P}(M)$ as

$$c(t) = p(x, t)\, d\theta,$$

where $p(x, t)$ is of $C^1$ in $t$ for any fixed $x \in M$. So, $c = c(t)$ has velocity vector field along $c$

$$\frac{dc}{dt}(t) = \frac{\partial}{\partial t} p(x, t)\ d\theta \in T_{c(t)}\mathcal{P}(M),\ \ t \in (a, b).$$

*3.2. Fisher Metric*

**Definition 1.** *A positive definite inner product $G_\mu$ on $T_\mu\mathcal{P}(M)$ at $\mu \in \mathcal{P}(M)$ is defined as*

$$G_\mu : T_\mu\mathcal{P}(M) \times T_\mu\mathcal{P}(M) \to \mathbf{R}; \quad G_\mu(\tau, \tau_1) = \int_{x \in M} \frac{d\tau}{d\mu}(x)\, \frac{d\tau_1}{d\mu}(x)\, d\mu(x).$$

*A family $G = \{G_\mu \mid \mu \in \mathcal{P}(M)\}$ is called Fisher metric. $\sqrt{G_\mu(\tau, \tau)}$ is denoted by $|\tau|_{G,\mu}$.*

The Fisher metric is a generalization of Fisher information matrices appeared in parametric models.

The following is one of remarkable properties of the Fisher metric $G$.

Let $\Phi : M \to M$ be a homeomorphism of $M$. Then, $\Phi$ induces a push-forward

$$\Phi_\sharp : \mathcal{P}(M) \to \mathcal{P}(M); \quad (\Phi_\sharp(\mu))(A) := \mu(\Phi^{-1}(A)), \quad A \in \mathcal{B}(M). \tag{10}$$

Here $\mu(A) = \int_{x \in A} d\mu(x)$. See [27]. We represent (10) in an integral form as

$$\int_{x \in A} q(x)\, d(\Phi_\sharp\mu)(x) = \int_{x \in \Phi^{-1}(A)} q(\Phi(x))\, d\mu(x)$$

for any measurable function $q : M \to \mathbf{R}$.

When $\Phi$ is self-diffeomorphism of $M$, $\Phi_\sharp\mu$ coincides with $(\Phi^{-1})^*\mu$, the pull-back of $\mu$ by the inverse diffeomorphism $\Phi^{-1}$. Notice that

$$(\Phi \circ \Psi)_\sharp = \Phi_\sharp \circ \Psi_\sharp$$

for homeomorphisms $\Phi$, $\Psi$ of $M$ and that the push-forward $\Phi_\sharp : \mathcal{P}(M) \to \mathcal{P}(M)$ has differential map

$$(d\Phi_\sharp)_\mu : T_\mu\mathcal{P}(M) \to T_{\Phi_\sharp\mu}\mathcal{P}(M); \quad (d\Phi_\sharp)_\mu(\tau) = \Phi_\sharp(\tau).$$

Here $\Phi_\sharp\tau$ is defined similarly as (10). In fact, we have

$$(d\Phi_\sharp)_\mu(\tau) = \frac{d}{dt}\Big|_{t=0}(\Phi_\sharp(\mu + t\tau)) = \frac{d}{dt}\Big|_{t=0}(\Phi_\sharp(\mu) + t\Phi_\sharp(\tau)) = \Phi_\sharp(\tau).$$

**Theorem 7** ([2])**.** *Let $\Phi_\sharp$ be a push-forward. Then it acts on $\mathcal{P}(M)$ isometrically with respect to the Fisher metric $G$. Namely,*

$$G_{\Phi_\sharp\mu}(\Phi_\sharp\tau, \Phi_\sharp\tau_1) = G_\mu(\tau, \tau_1), \quad \forall\tau, \tau_1 \in T_\mu\mathcal{P}(M), \forall\mu \in \mathcal{P}(M).$$

**Proof.** We write $\mu = p(x)\, d\theta(x)$ and $\tau = q(x)\, d\theta(x)$, $\tau_1 = q_1(x)\, d\theta(x)$. Set $\sigma = \Phi_\sharp\mu$. We have then from definition of push-forward

$$\sigma = p(\Phi^{-1}(x))\, \Phi_\sharp d\theta(x). \tag{11}$$

This follows in fact from

$$\begin{aligned}\int_M h(x)\, d(\Phi_\sharp\mu)(x) &= \int_M h(\Phi(x))\, d\mu(x) = \int_M h(\Phi(x))\, p(x)\, d\theta(x) \\ &= \int_M \left(h \times (p \circ \Phi^{-1})\right)(\Phi(x))\, d\theta(x)\end{aligned} \tag{12}$$

for any measurable function $h$ on $M$, which, by definition of push-forward, coincides with $\int_M h(x)\, (p \circ \Phi^{-1})(x)\, \Phi_\sharp d\theta(x)$ and the above is obtained.

In a similar way to (11) we have

$$\Phi_\sharp\tau = q(\Phi^{-1}(x))\, \Phi_\sharp d\theta(x), \quad \Phi_\sharp\tau_1 = q_1(\Phi^{-1}(x))\, \Phi_\sharp d\theta(x),$$

so that

$$\begin{aligned}G_\sigma(\Phi_\sharp\tau, \Phi_\sharp\tau_1) &= \int_{x\in M} \left(\frac{d\Phi_\sharp\tau}{d\sigma}\right)(x)\, \left(\frac{d\Phi_\sharp\tau_1}{d\sigma}\right)(x)\, d\sigma(x) \\ &= \int_{x\in M} \frac{q(\Phi^{-1}(x))}{p(\Phi^{-1}(x))}\, \frac{q_1(\Phi^{-1}(x))}{p(\Phi^{-1}(x))}\, d(\Phi_\sharp\mu)(x) \\ &= \int_{x\in M} \frac{d\tau}{d\mu}(\Phi^{-1}(x))\, \frac{d\tau_1}{d\mu}(\Phi^{-1}(x))\, d(\Phi_\sharp\mu)(x).\end{aligned}$$

Set $F(x) = \frac{d\tau}{d\mu}(\Phi^{-1}(x))\, \frac{d\tau_1}{d\mu}(\Phi^{-1}(x))$ and write the above as

$$\int_M F(x)\, d(\Phi_\sharp\mu)(x) = \int_M F(\Phi(x))\, d\mu(x) = \int_M \frac{d\tau}{d\mu}(x)\, \frac{d\tau_1}{d\mu}(x)\, d\mu(x)$$

which is the inner product $G_\mu(\tau, \tau_1)$ of $\tau$ and $\tau_1$ at $\mu$. $\square$

**Note.** The following is known. For any $\mu \in \mathcal{P}(M)$ there exists a homeomorphism $\Phi : M \to M$ satisfying $\mu = \Phi_\sharp d\theta$ (refer to [28,29]). This fact implies that the action of $\mathrm{Homeo}(M)$, which is the group of homeomorphisms on $M$, on the space $\mathcal{P}(M)$ is isometric and transitive.

**Remark 1.** *The embedding $\rho$ given in (8) satisfies $\rho^*\langle\cdot,\cdot\rangle_{L^2} = \frac{1}{4}\,G$, that is,*

$$\langle(\rho_*)_\mu\tau, (\rho_*)_\mu\tau_1\rangle_{L^2} = \frac{1}{4}G_\mu(\tau,\tau_1), \quad \forall\tau,\tau_1 \in T_\mu\mathcal{P}(M),\ \mu\in\mathcal{P}(M).$$

*where $\langle\cdot,\cdot\rangle_{L^2}$ is the $L^2$-inner product of $L^2(M,d\theta)$;*

$$\langle f_1, f_2\rangle_{L^2} := \int_{x\in M} f_1(x)f_2(x)d\theta(x), \quad f_1, f_2 \in L^2(M,d\theta).$$

*3.3. Levi–Civita Connection*

The Fisher metric provides the space $\mathcal{P}(M)$ a Riemannian metric as above and then induces on $\mathcal{P}(M)$ the Levi–Civita connection $\nabla$ and the Riemannian curvature tensor $R$. To derive their formulae we will introduce a constant vector fields on $\mathcal{P}(M)$.

Let $\tau \in \mathcal{V}$. Then, $\tau$ is considered as a constant vector field on $\mathcal{P}(M)$ by defining a vector field $\{\tau_\mu \mid \mu\in\mathcal{P}(M)\}$, $\tau_\mu = \frac{d}{dt}\big|_{t=0}\mu(t) \in T_\mu\mathcal{P}(M)$, that is, $\tau$ is a velocity vector of a curve $\mu(t) = \mu + t\tau$ at $t = 0$. Notice that an integral curve of a constant vector field $\tau$ passing through $\mu \in \mathcal{P}(M)$ is given by the curve $\mu(t)$.

**Theorem 8.** *Let $\tau$, $\tau_1$ be constant vector fields on $\mathcal{P}(M)$. Then, the Levi–Civita connection of the Fisher metric $G$ at $\mu \in \mathcal{P}(M)$ is represented as*

$$\begin{aligned}\nabla_\tau\tau_1 =& -\frac{1}{2}\left(\frac{d\tau}{d\mu}(x)\,\frac{d\tau_1}{d\mu}(x) - G_\mu(\tau,\tau_1)\right)\mu \\ =& -\frac{1}{2}\left(\frac{d\tau}{d\mu}(x)\,\frac{d\tau_1}{d\mu}(x) - \int_{y\in M}\frac{d\tau}{d\mu}(y)\frac{d\tau_1}{d\mu}(y)d\mu(y)\right)\mu.\end{aligned} \tag{13}$$

For this formula see [2].

**Proof.** Recall that on a Riemannian manifold $N$ with a Riemannian metric $g$ the Levi–Civita connection $\nabla$ of $g$ is an affine connection on $N$, that is, $\nabla$ is a bilinear map; $(Y,Z) \mapsto \nabla_Y Z$ satisfying

$$\begin{aligned}\nabla_{fY}Z =& f\,\nabla_Y Z, \\ \nabla_Y(fZ) =& Yf\cdot Z + f\nabla_Y Z\end{aligned} \tag{14}$$

for smooth vector fields $Y$, $Z$ on $N$ and a smooth function $f$ on $N$, which satisfies $(\nabla_Y g)(Z,W) = 0$ together with a symmetry condition, that is, the torsion tensor $T(Y,Z) := \nabla_Y Z - \nabla_Z Y - [Y,Z]$ vanishes. Then, the Levi–Civita connection $\nabla$ exists uniquely and one has Koszul's formula for $\nabla$

$$g(\nabla_Y Z, W) = \frac{1}{2}\{Yg(Z,W) + Zg(W,Y) - Wg(Y,Z) + g([Y,Z],W) - g([Z,W],Y) - g([Y,W],Z)\}.$$

Here $Y$, $Z$, $W$ are smooth vector fields on $N$ ([22]).

We give a reference comment on a metric connection with non-trivial torsion, appeared in information geometry. A non-trivial torsion $T$ implies geometrically a breaking of the symmetry in connection coefficients; $\Gamma^k_{ij} = \Gamma^k_{ji}$. In a framework of classical parametric model there are very few study of a metric connection with non-trivial torsion. However, as far as the authors know, the $e$-connection developed in a quantum model has non-trivial torsion. Refer to Chapter 7 of [14] and references cited there.

Now we return back to our situation, that is, to the space $(\mathcal{P}(M), G)$ in which we have for constant vector fields $\tau$, $\tau_1$ and $\tau_2$

$$G(\nabla_{\tau_2}\tau, \tau_1) = \frac{1}{2}\{\tau_2 G(\tau, \tau_1) + \tau G(\tau_1, \tau_2) - \tau_1 G(\tau, \tau_2)\},$$

since $[\tau, \tau_1] = [\tau, \tau_2] = [\tau_1, \tau_2] = 0$.

Let $\mu(t) = \mu + t\,\tau_2$ be a curve in $\mathcal{P}(M)$ of $\mu(0) = \mu$ and $\dot{\mu}(t) = \tau_2$. We have then

$$\begin{aligned}
(\tau_2)_\mu G(\tau, \tau_1) &= \frac{d}{dt}\Big|_{t=0} G_{\mu(t)}(\tau, \tau_1) \\
&= \frac{d}{dt}\Big|_{t=0} \int_M \frac{d\tau}{d\mu(t)}(x)\, \frac{d\tau_1}{d\mu(t)}(x)\, d\mu(t)(x) \\
&= \int_M \frac{\partial}{\partial t}\Big|_{t=0} \left( \frac{d\tau}{d\mu(t)}(x)\, \frac{d\tau_1}{d\mu(t)}(x)\, d\mu(t)(x) \right)
\end{aligned}$$

in which the integrand is

$$\begin{aligned}
&\frac{\partial}{\partial t}\Big|_{t=0} \left( \frac{d\tau}{d\mu(t)}(x)\, \frac{d\tau_1}{d\mu(t)}(x)\, d\mu(t)(x) \right) \\
&= \frac{\partial}{\partial t} \left( \frac{d\tau}{d\mu(t)}(x) \right)\Big|_{t=0} \frac{d\tau_1}{d\mu}(x)\, d\mu(x) + \frac{d\tau}{d\mu}(x)\ \frac{\partial}{\partial t} \left( \frac{d\tau_1}{d\mu(t)}(x) \right)\Big|_{t=0} d\mu(x) \\
&\quad + \frac{d\tau}{d\mu}(x)\, \frac{d\tau_1}{d\mu}(x)\ \frac{\partial}{\partial t}(d\mu(t)(x))\Big|_{t=0}.
\end{aligned}$$

The partial derivative term becomes

$$\frac{\partial}{\partial t} \left( \frac{d\tau}{d\mu(t)}(x) \right)\Big|_{t=0} = -\frac{d\tau}{d\mu}(x)\, \frac{d\tau_2}{d\mu}(x),$$

since by calculation

$$\begin{aligned}
\frac{\partial}{\partial t} \left( \frac{d\tau}{d\mu(t)}(x) \right)\Big|_{t=0} &= \frac{\partial}{\partial t} \left( \frac{q(x)}{p(x) + tq_2(x)} \right)\Big|_{t=0} = -\left. \frac{q(x)q_2(x)}{(p(x) + tq_2(x))^2} \right|_{t=0} \\
&= -\frac{q(x)q_2(x)}{p^2(x)} = -\frac{d\tau}{d\mu}(x)\, \frac{d\tau_2}{d\mu}(x).
\end{aligned}$$

Similarly

$$\frac{\partial}{\partial t} \left( \frac{d\tau_1}{d\mu(t)}(x) \right)\Big|_{t=0} = -\frac{d\tau_1}{d\mu}(x)\, \frac{d\tau_2}{d\mu}(x).$$

Since

$$\frac{\partial}{\partial t}(d\mu(t)(x))\bigg|_{t=0} = d\tau_2(x),$$

one obtains

$$\begin{aligned}(\tau_2)_\mu G(\tau,\tau_1) = &\int_M \left(-\frac{d\tau}{d\mu}(x)\frac{d\tau_2}{d\mu}(x)\right)\frac{d\tau_1}{d\mu}(x)\,d\mu(x)\\ &+\int_M \frac{d\tau}{d\mu}(x)\left(-\frac{d\tau_1}{d\mu}(x)\frac{d\tau_2}{d\mu}(x)\right)\,d\mu(x)+\int_M\left(\frac{d\tau}{d\mu}(x)\frac{d\tau_1}{d\mu}(x)\right)\,d\tau_2(x).\end{aligned}$$

Here $d\tau_2(x) = \frac{d\tau_2}{d\mu}(x)\,d\mu(x)$. So,

$$(\tau_2)_\mu G(\tau,\tau_1) = -\int_M \frac{d\tau}{d\mu}(x)\frac{d\tau_1}{d\mu}(x)\frac{d\tau_2}{d\mu}(x)\,d\mu(x).$$

One obtains similar formulae for the terms $\tau_\mu G(\tau_1,\tau_2)$, $(\tau_1)_\mu G(\tau,\tau_2)$ and then finally

$$G_\mu(\nabla_{\tau_2}\tau,\tau_1) = -\frac{1}{2}\int_M \frac{d\tau}{d\mu}(x)\frac{d\tau_2}{d\mu}(x)\frac{d\tau_1}{d\mu}(x)\,d\mu(x).$$

On the other hand, one observes

$$\int_M G_\mu(\tau,\tau_2)\,\frac{d\tau_1}{d\mu}(x)\,d\mu(x) = G_\mu(\tau,\tau_2)\int_M d\tau_1(x) = 0$$

and hence

$$\begin{aligned}G_\mu(\nabla_{\tau_2}\tau,\tau_1) =& -\frac{1}{2}\int_M\left(\frac{d\tau}{d\mu}(x)\frac{d\tau_2}{d\mu}(x)-G_\mu(\tau,\tau_2)\right)\frac{d\tau_1}{d\mu}(x)\,d\mu(x)\\ =&G_\mu\left(-\frac{1}{2}\left(\frac{d\tau}{d\mu}(x)\frac{d\tau_2}{d\mu}(x)-G_\mu(\tau,\tau_2)\right)\mu,\tau_1\right).\end{aligned}$$

Since $\tau_1$ is arbitrary, (13) is derived. □

**Theorem 9.** *The Riemannian curvature tensor $R$ of the Fisher metric $G$ satisfies*

$$R_\mu(\tau_1,\tau_2)\tau = \frac{1}{4}\left(G_\mu(\tau,\tau_2)\tau_1 - G_\mu(\tau,\tau_1)\tau_2\right)$$

*for constant vector fields $\tau$, $\tau_1$, $\tau_2$. Hence, sectional curvature of any section $\tau\wedge\tau_1$ is $K(\tau\wedge\tau_1)=\frac{1}{4}$.*

Refer to [2] for this theorem. We omit proving this theorem. In general, a finite dimensional Riemannian manifold of constant sectional curvature $1/4$ is (locally) isometric to a sphere of radius $2$. So, the space $\mathcal{P}(M)$ with the metric $G$ is considered to be isometrically an infinite dimensional sphere of radius $2$.

As is shown in the next section, this infinite dimensional Riemannian manifold $(\mathcal{P}(M),G)$ is not geodesically complete, in other words, every geodesic is not necessarily extended over $\mathbf{R}$.

*3.4. Geodesics*

**Theorem 10.** *Let $\mu \in \mathcal{P}(M)$ and $\tau \in T_\mu\mathcal{P}(M)$. Assume $\tau$ is a unit tangent vector at $\mu$,* i.e., *$|\tau|_{G,\mu} = 1$. Then, the geodesic $\mu(t)$, denoted by $\exp_\mu t\tau$, with $\mu(0) = \mu$, $\dot\mu(0) = \tau$ has the form represented by*

$$\mu(t) = \left(\cos\frac{t}{2} + \sin\frac{t}{2}\frac{d\tau}{d\mu}(x)\right)^2 \mu, \tag{15}$$

*in other words,*

$$\mu(t) = \left(\cos\frac{t}{2} + \sin\frac{t}{2}\frac{q(x)}{p(x)}\right)^2 p(x)\, d\theta(x) \tag{16}$$

*where $\mu = p(x)\, d\theta$ and $\tau = q(x)\, d\theta$ are density function representation of $\mu$, $\tau$.*

**Note.** Set $t = \pi$ into (15). Then $\mu(\pi) = (\frac{d\tau}{d\mu}(x))^2\mu = \frac{q(x)^2}{p(x)}\, d\theta$. However, $\tau$ is a tangent vector to $\mathcal{P}(M)$ so $\tau$ satisfies $\int_{x\in M} q(x)d\theta(x) = 0$ from which there exists a point $x_o \in M$ with $q(x_o) = 0$ and then the density function of $\mu(\pi)$ vanishes at point $x_o$ and then $\mu(\pi) \notin \mathcal{P}(M)$.

To prove Theorem 10, we will show the following lemma, obtained by T. Friedrich ([2]).

**Lemma 1.** *Let $\mu(t) = p(x,t)\, d\theta$ be a geodesic such that $\mu(0) = \mu = p_0(x)\, d\theta$, $\dot\mu(0) = \tau = q(x)\, d\theta \in T_\mu\mathcal{P}^+(M)$, $|\tau|_{G,\mu} = 1$. Then,*

$$p(x,t) = \frac{1}{1+\tan^2\frac{t}{2}}\left\{p_0(x) + 2\tan\frac{t}{2}\, q(x) + \tan^2\frac{t}{2}\,\frac{q^2(x)}{p_0(x)}\right\}. \tag{17}$$

From this lemma it is easy to see that $\mu(t) = p(x,t)\, d\theta$ has the above form (15).

**Proof of Lemma 1.** A proof is given in [2]. However, we will give a proof for a later convenience in proving Theorems 16 and 17. So, for simplicity we write by abbreviation $\mu(t) = p(t)\, d\theta$ and $\dot\mu(t) = \dot p(t)\, d\theta$. Letting $\tau$ be a constant vector field, we have

$$G(\nabla_{\dot\mu(t)}\dot\mu(t), \tau) = \dot\mu(t)\,(G(\dot\mu(t), \tau)) - G(\dot\mu(t), \nabla_{\dot\mu(t)}\tau), \tag{18}$$

since $\nabla$ preserves the metric $G$.

Notice that from the rule (14) of Levi–Civita connection, the tangent vector $\dot\mu(t) \in T_{\mu(t)}\mathcal{P}(M)$ of $\nabla_{\dot\mu(t)}$, appeared in the second term of (18) can be extended as a constant vector field, denoted by the same symbol. So, we can apply (13) to the above and have

$$\nabla_{\dot\mu(t)}\tau = -\frac{1}{2}\left(\frac{d\dot\mu(t)}{d\mu(t)}\,\frac{d\tau}{d\mu(t)} - G_{\mu(t)}(\dot\mu(t), \tau)\right)\mu(t).$$

Thus, the Radon-Nikodym derivative of $\nabla_{\dot\mu(t)}\tau$ with respect to $\mu(t)$ is

$$\frac{d\,\nabla_{\dot\mu(t)}\tau}{d\mu(t)} = -\frac{1}{2}\left(\frac{d\dot\mu(t)}{d\mu(t)}\,\frac{d\tau}{d\mu(t)} - G_{\mu(t)}(\dot\mu(t), \tau)\right).$$

Therefore, we have

$$\begin{aligned} G(\dot{\mu}(t), \nabla_{\dot{\mu}(t)}\tau) &= \int_M \left(\frac{d\dot{\mu}(t)}{d\mu(t)}\right) \left\{ -\frac{1}{2} \left( \frac{d\dot{\mu}(t)}{d\mu(t)} \, \frac{d\tau}{d\mu(t)} - G_{\mu(t)}(\dot{\mu}(t), \tau) \right) \right\} d\mu(t) \\ &= -\frac{1}{2} \int_M \left(\frac{d\dot{\mu}(t)}{d\mu(t)}\right)^2 \frac{d\tau}{d\mu(t)} \, d\mu(t) + \frac{1}{2} G_{\mu(t)}(\dot{\mu}(t), \tau) \int_M \left(\frac{d\dot{\mu}(t)}{d\mu(t)}\right) d\mu(t) \\ &= -\frac{1}{2} \int_M \left(\frac{d\dot{\mu}(t)}{d\mu(t)}\right)^2 d\tau + \frac{1}{2} G_{\mu(t)}(\dot{\mu}(t), \tau) \int_M d\dot{\mu}(t) \\ &= -\frac{1}{2} \int_M \left(\frac{d\dot{\mu}(t)}{d\mu(t)}\right)^2 d\tau = -\frac{1}{2} \int_M \left(\frac{\dot{p}(t)}{p(t)}\right)^2 d\tau. \end{aligned}$$

On the other hand, the first term of (18) is

$$\dot{\mu}(t) G(\dot{\mu}(t), \tau) = \frac{d}{dt} \, G_{\mu(t)}(\dot{\mu}(t), \tau) = \frac{d}{dt} \int_M \left(\frac{\dot{p}(t)}{p(t)}\right) d\tau = \int_M \frac{\partial}{\partial t} \left(\frac{\dot{p}(t)}{p(t)}\right) d\tau.$$

Then (18) becomes

$$G(\nabla_{\dot{\mu}(t)}\dot{\mu}(t), \tau) = \int_M \left\{ \frac{\partial}{\partial t} \left(\frac{\dot{p}(t)}{p(t)}\right) + \frac{1}{2} \left(\frac{\dot{p}(t)}{p(t)}\right)^2 \right\} d\tau.$$

Here, $\left\{ \frac{\partial}{\partial t} \left(\frac{\dot{p}(t)}{p(t)}\right) + \frac{1}{2} \left(\frac{\dot{p}(t)}{p(t)}\right)^2 \right\} \mu(t)$ is not necessarily a tangent vector at $\mu(t)$. We choose a real valued function $C(t)$ of $t$, independent of $x \in M$ satisfying

$$\left\{ \frac{\partial}{\partial t} \left(\frac{\dot{p}(t)}{p(t)}\right) + \frac{1}{2} \left(\frac{\dot{p}(t)}{p(t)}\right)^2 + C(t) \right\} \mu(t) \in T_{\mu(t)} \mathcal{P}^+(M).$$

In fact, we define $C(t)$ as

$$\begin{aligned} C(t) &= -\int_M \left\{ \frac{\partial}{\partial t} \left(\frac{\dot{p}(t)}{p(t)}\right) + \frac{1}{2} \left(\frac{\dot{p}(t)}{p(t)}\right)^2 \right\} p(t) \, d\theta \\ &= -\int_M \frac{\partial}{\partial t} \left(\frac{d\dot{\mu}(t)}{d\mu(t)}\right) d\mu(t) - \frac{1}{2} G(\dot{\mu}(t), \dot{\mu}(t)). \end{aligned}$$

Hence, we have

$$G(\nabla_{\dot{\mu}(t)}\dot{\mu}(t), \tau) = G\left( \left\{ \frac{\partial}{\partial t} \left(\frac{\dot{p}(t)}{p(t)}\right) + \frac{1}{2} \left(\frac{\dot{p}(t)}{p(t)}\right)^2 + C(t) \right\} \mu(t), \tau \right)$$

for an arbitrary constant vector field $\tau$.

Therefore we have

$$\nabla_{\dot{\mu}(t)}\dot{\mu}(t) = \left\{ \frac{\partial}{\partial t} \left(\frac{\dot{p}(t)}{p(t)}\right) + \frac{1}{2} \left(\frac{\dot{p}(t)}{p(t)}\right)^2 + C(t) \right\} \mu(t)$$

Thus, it is concluded that $\mu(t) = p(t) \, d\theta$ is a geodesic if and only if $p(t) = p(x, t)$ satisfies

$$\frac{\partial}{\partial t} \left(\frac{\dot{p}(t)}{p(t)}\right) + \frac{1}{2} \left(\frac{\dot{p}(t)}{p(t)}\right)^2 + C(t) = 0, \quad \int_M \dot{p}(t) \, d\theta = 0. \tag{19}$$

**Remark 2.** *Equation (19) is expressed as*

$$\frac{\partial}{\partial t}\left(\frac{\dot{\mu}(t)}{\mu(t)}\right)+\frac{1}{2}\left(\frac{\dot{\mu}(t)}{\mu(t)}\right)^2+C(t)=0,\quad \dot{\mu}(t)\in T_{\mu(t)}\mathcal{P}(M).$$

Now we will solve these equations.

Notice that if $\mu(t)$ is a geodesic, $G_{\mu(t)}(\dot{\mu}(t),\dot{\mu}(t))$ is constant along $\mu(t)$ so from the initial condition $G_{\mu(t)}(\dot{\mu}(t),\dot{\mu}(t))\equiv 1$. Therefore, we can write (19) as

$$\begin{gathered}\frac{\partial}{\partial t}\left(\frac{\dot{p}(t)}{p(t)}\right)+\frac{1}{2}\left(\frac{\dot{p}(t)}{p(t)}\right)^2-\int_M\frac{\partial}{\partial t}\left(\frac{\dot{p}(t)}{p(t)}\right)p(t)\,d\theta-\frac{1}{2}=0,\\ \int_M\frac{\dot{p}^2(t)}{p(t)}d\theta=1,\qquad \int_M\dot{p}(t)d\theta=0.\end{gathered}\tag{20}$$

The second equation means that $|\dot{\mu}(t)|_{G,\mu(t)}=1$. Set $g(t):=\int_M\frac{\partial}{\partial t}\left(\frac{\dot{p}(t)}{p(t)}\right)p(t)\,d\theta$. We have then

$$\begin{aligned}g(t)=&\frac{d}{dt}\left(\int_M\left(\frac{\dot{p}(t)}{p(t)}\right)p(t)\,d\theta\right)-\int_M\left(\frac{\dot{p}(t)}{p(t)}\right)\dot{p}(t)\,d\theta\\ =&\frac{d}{dt}\left(\int_M d\dot{\mu}(t)\right)-G_{\mu(t)}(\dot{\mu}(t),\dot{\mu}(t))=0-1=-1.\end{aligned}$$

So Equations (20) reduce to

$$\frac{\partial}{\partial t}\left(\frac{\dot{p}(t)}{p(t)}\right)+\frac{1}{2}\left(\frac{\dot{p}(t)}{p(t)}\right)^2+\frac{1}{2}=0,\quad \int_M\left(\frac{\dot{p}(t)}{p(t)}\right)^2p(t)d\theta=1,\quad \int_M\dot{p}(t)d\theta=0.$$

To solve these, we set $w(t)=\frac{\dot{p}(t)}{p(t)}$. Then $w(t)$ satisfies

$$\dot{w}(t)+\frac{1}{2}\left(w^2(t)+1\right)=0.\tag{21}$$

By solving (21), we have $w(t)=\tan\left(-\frac{1}{2}t+A(x)\right),\ x\in M$, where $A(x)$ is an integral constant depending on $x$. By integrating $w(t)=\dot{p}(t)/p(t)$, where $p(t)=p(x,t)$,

$$\log p(x,t)=2\log\cos\left(-\frac{1}{2}t+A(x)\right)+B_1(x),$$

and hence

$$p(x,t)=B(x)\cos^2\left(-\frac{1}{2}t+A(x)\right).$$

Here $B(x)$ is an integral constant. The constants $A(x)$, $B(x)$ are given as

$$\begin{aligned}A(x)=&\arctan\left(\frac{\dot{p}(x,0)}{p(x,0)}\right),\\ B(x)=&\frac{p^2(x,0)+\dot{p}^2(x,0)}{p(x,0)},\end{aligned}$$

where $p(x,0)=p_0(x)$, $\dot{p}(x,0)=q(x)$. $\square$

**Corollary 2** ([2]). *Every geodesic on $(\mathcal{P}(M), G)$ is periodic, of period $2\pi$.*

In fact, from (15) a geodesic $\mu(t)$ is represented as

$$\mu(t) = \left( \cos^2 \frac{t}{2} + 2 \sin \frac{t}{2} \cos \frac{t}{2} \frac{d\tau}{d\mu}(x) + \sin^2 \frac{t}{2} \left( \frac{d\tau}{d\mu} \right)^2 (x) \right) \mu$$

and $\cos^2 \frac{t}{2} = \frac{1}{2}(1 + \cos t)$. We have then the corollary.

**Definition 2.** *Define $\ell : \mathcal{P}(M) \times \mathcal{P}(M) \to [0, \pi); (\mu, \mu^*) \mapsto \ell = \ell(\mu, \mu^*)$ by*

$$\cos \frac{\ell}{2} = \int_{x \in M} \sqrt{\frac{d\mu^*}{d\mu}}(x) \, d\mu(x). \tag{22}$$

One sees $\ell(\mu, \mu^*) = \ell(\mu^*, \mu)$ and that $\ell = 0$ if and only if $\mu = \mu^*$. $\cos \frac{\ell}{2}$ is an $f$-divergence-like quantity with respect to $f(u) = \sqrt{u}$ (See [14]).

**Remark 3.** *In* [2] *T. Friedrich remarked that the distance between $\mu$ and $\mu^*$ in $\mathcal{P}(M)$ is given by $\ell = \ell(\mu, \mu^*)$.*

**Theorem 11.** *Let $\mu$ and $\mu^*$ be arbitrary probability measures in $\mathcal{P}(M)$, $\mu \neq \mu^*$. Then, there exists a unique geodesic $\mu(t)$,* i.e.*, a curve; $t \in I \subset \mathbf{R} \mapsto \mu(t) \in \mathcal{P}(M)$ with $\mu(0) = \mu$, $\mu(\ell) = \mu^*$, where $\ell = \ell(\mu, \mu^*)$ is given by (22) and $I$ is an open interval;*

$$\mu(t) = \exp_\mu t\tau = \left( \cos \frac{t}{2} + \sin \frac{t}{2} \frac{d\tau}{d\mu}(x) \right)^2 \mu, \tag{23}$$

*where $\tau$ is a unit tangent vector at $\mu$ represented by*

$$\tau = \frac{1}{\sin \frac{\ell}{2}} \left( \sqrt{\frac{d\mu^*}{d\mu}}(x) - \int_{y \in M} \sqrt{\frac{d\mu^*}{d\mu}}(y) \, d\mu(y) \right) \mu(x). \tag{24}$$

**Proof.** First we will show

**Assertion 1.** The measure $\tau$ given by (24) is a unit tangent vector to $\mathcal{P}(M)$ at $\mu$.

In fact,

$$\begin{aligned} \int_M d\tau =& \frac{1}{\sin \frac{\ell}{2}} \int_{x \in M} \left( \sqrt{\frac{d\mu^*}{d\mu}}(x) - \int_{y \in M} \sqrt{\frac{d\mu^*}{d\mu}}(y) \, d\mu(y) \right) d\mu(x) \\ =& \frac{1}{\sin \frac{\ell}{2}} \left( \int_{x \in M} \sqrt{\frac{d\mu^*}{d\mu}}(x) \, d\mu(x) - \int_{y \in M} \sqrt{\frac{d\mu^*}{d\mu}}(y) \, d\mu(y) \right) = 0, \end{aligned}$$

so that $\tau$ is a tangent vector to $\mathcal{P}(M)$. Moreover, $\tau$ is unit, *i.e.*, $G_\mu(\tau, \tau) = 1$, as we compute straightforward

$$G_\mu(\tau, \tau) = \int_M \left( \frac{d\tau}{d\mu}(x) \right)^2 d\mu(x) \tag{25}$$

and substitute (24) into (25) to have

$$\begin{aligned}
G_\mu(\tau,\tau) =& \frac{1}{\sin^2\frac{\ell}{2}} \int_{x\in M} \left( \sqrt{\frac{d\mu^*}{d\mu}}(x) - \int_{y\in M} \sqrt{\frac{d\mu^*}{d\mu}}(y) d\mu(y) \right)^2 d\mu(x) \\
=& \frac{1}{\sin^2\frac{\ell}{2}} \int_{x\in M} \left\{ \frac{d\mu^*}{d\mu}(x) - 2 \left( \int_{y\in M} \sqrt{\frac{d\mu^*}{d\mu}}(y)\, d\mu(y) \right) \sqrt{\frac{d\mu^*}{d\mu}}(x) \right. \\
& \left. + \left( \int_{y\in M} \sqrt{\frac{d\mu^*}{d\mu}}(y) d\mu(y) \right)^2 \right\} d\mu(x) \\
=& \frac{1}{\sin^2\frac{\ell}{2}} \left\{ \int_{x\in M} d\mu^*(x) - 2 \left( \int_{y\in M} \sqrt{\frac{d\mu^*}{d\mu}}(y)\, d\mu(y) \right)^2 + \left( \int_{y\in M} \sqrt{\frac{d\mu^*}{d\mu}}(y)\, d\mu(y) \right)^2 \right\} \\
=& \frac{1}{\sin^2\frac{\ell}{2}} \left( 1 - \cos^2\frac{\ell}{2} \right) = 1.
\end{aligned}$$

**Assertion 2.** A geodesic defined by (23) satisfies $\mu(0) = \mu$ and $\mu(\ell) = \mu^*$.

It is seen $\mu(0) = \mu$ from (23). At $t = \ell$

$$\mu(\ell) = \left( \cos\frac{\ell}{2} + \sin\frac{\ell}{2} \frac{d\tau}{d\mu}(x) \right)^2 \mu$$

and from (25)

$$\sin\frac{\ell}{2} \frac{d\tau}{d\mu}(x) = \sqrt{\frac{d\mu^*}{d\mu}}(x) - \int_{y\in M} \sqrt{\frac{d\mu^*}{d\mu}}(y)\, d\mu(y) = \sqrt{\frac{d\mu^*}{d\mu}}(x) - \cos\frac{\ell}{2}.$$

Hence, we find $\mu(\ell) = \mu^*$ as follows;

$$\mu(\ell) = \left\{ \cos\frac{\ell}{2} + \left( \sqrt{\frac{d\mu^*}{d\mu}}(x) - \cos\frac{\ell}{2} \right) \right\}^2 \mu = \left( \sqrt{\frac{d\mu^*}{d\mu}}(x) \right)^2 \mu = \frac{d\mu^*}{d\mu}(x)\, \mu = \mu^*.$$

**Assertion 3.** A geodesic joining $\mu$ and $\mu^*$ is unique for $\mu \neq \mu^*$.

To verify this assertion let $\mu(t) = \exp_\mu t\tau$ and $\tilde{\mu}(t) = \exp_\mu t\tilde{\tau}$ be unit speed geodesics satisfying $\mu(0) = \tilde{\mu}(0) = \mu$ and $\mu(\ell) = \tilde{\mu}(\ell) = \mu^*$. From the latter condition we have by using (15)

$$\left( \cos\frac{\ell}{2} + \sin\frac{\ell}{2} \frac{d\tilde{\tau}}{d\mu}(x) \right)^2 \mu = \left( \cos\frac{\ell}{2} + \sin\frac{\ell}{2} \frac{d\tau}{d\mu}(x) \right)^2 \mu, \quad \forall x \in M$$

from which

$$\cos\frac{\ell}{2} + \sin\frac{\ell}{2} \frac{d\tilde{\tau}}{d\mu}(x) = \pm \left( \cos\frac{\ell}{2} + \sin\frac{\ell}{2} \frac{d\tau}{d\mu}(x) \right)$$

for any $x \in M$.

To assert

$$\cos\frac{\ell}{2} + \sin\frac{\ell}{2} \frac{d\tilde{\tau}}{d\mu}(x) = \cos\frac{\ell}{2} + \sin\frac{\ell}{2} \frac{d\tau}{d\mu}(x), \quad \forall x \in M$$

define subsets $M_\pm$ of $M$ by

$$M_\pm = \left\{ x \in M \,;\, \cos\frac{\ell}{2} + \sin\frac{\ell}{2}\frac{d\tilde{\tau}}{d\mu}(x) = \pm\left(\cos\frac{\ell}{2} + \sin\frac{\ell}{2}\frac{d\tau}{d\mu}(x)\right)\right\}.$$

Then, $M = M_+ \cup M_-$. Moreover $M_+ \cap M_- = \emptyset$. This is because if otherwise, $M_+ \cap M_- \neq \emptyset$, then at a point $x \in M_+ \cap M_-$ it holds

$$\cos\frac{\ell}{2} + \sin\frac{\ell}{2}\frac{d\tilde{\tau}}{d\mu}(x) = 0.$$

So, at $x$, $\tilde{\mu}(\ell) = \left(\cos\frac{\ell}{2} + \sin\frac{\ell}{2}\frac{d\tilde{\tau}}{d\mu}(x)\right)^2 \mu = 0$. However, it must coincide at $x$ with $\mu^*$ which has a positive density function. So $M_+ \cap M_- = \emptyset$.

We see that $M_+ = \{x \in M \,;\, \tilde{\tau}(x) = \tau(x)\}$ and $M_- = \{x \in M \,;\, q(x) + \tilde{q}(x) = -2\cot\frac{\ell}{2}\}$. Here we write $\tau = q(x)\,d\theta$ and $\tilde{\tau} = \tilde{q}(x)\,d\theta$. Then, $M_+$ and $M_-$ are closed in $M$.

Suppose $M_- \neq \emptyset$. Then, $M_-$ must be a non-empty, proper subset of $M$. This is because, otherwise if $M_- = M$ is assumed, then, since $\tilde{\tau}$, $\tau$ are tangent to $\mathcal{P}(M)$ we see, from $0 < \ell < \pi$

$$\int_M d\tilde{\tau} = -\int_M d\tau - 2\cot\frac{\ell}{2}\int_M d\theta = -2\cot\frac{\ell}{2} \neq 0.$$

This is a contradiction. So, $M_-$ is a proper subset and hence $M_+ = M \setminus M_-$ is a non-empty closed, but open subset of $M$. Therefore, since $M$ is connected, $M_+ = M$, namely $\tilde{\tau}(x) = \tau(x)$ for any $x \in X$, from which the assertion is proved.

From these assertions Theorem 11 is verified. □

**Remark 4.** *For the $\ell$ of (22)*

$$\mu^* \neq \mu \iff \sin\frac{\ell}{2} \neq 0.$$

*It suffices for this to show*

$$\mu^* = \mu \iff \ell = 0,$$

*since, for $\ell \in [0, \pi)$, $\sin \ell/2 = 0$ if and only if $\ell = 0$. With respect to the embedding $\rho : \mathcal{P}(M) \to L^2(M, d\theta)$, given in (8), we have*

$$||\rho(\mu^*) - \rho(\mu)||_{L^2}^2 = 2 - 2\cos\frac{\ell}{2}$$

*from which it follows that $\ell = 0$ implies $||\rho(\mu^*) - \rho(\mu)||_{L^2} = 0$ and hence $\mu^* = \mu$, since $\rho$ is an embedding. Conversely, if $\mu^* = \mu$, then $\sqrt{d\mu^*/d\mu}(x) = 1$ so $\cos \ell/2 = \int_M \sqrt{d\mu^*/d\mu}(x) d\mu = 1$ and thus $\ell = 0$.*

To guarantee completeness of geodesics we must extend the space $\mathcal{P}(M)$, for example, to the space of probability measures on $M$, absolutely continuous with respect to $d\theta$ and with non-negative density function.

## 4. Hadamard Manifolds and Barycenter Map

### *4.1. Hadamard Manifolds and Ideal Boundary*

Let $(X, g)$ be an Hadamard manifold. Then the ideal boundary $\partial X$ of $(X, g)$ is defined by taking quotient of space of geodesics of $X$ and is homeomorphic to an $(n-1)$-sphere $S^{n-1}$. For any $\theta \in \partial X$ Busemann function $B_\theta$ normalized at some point and parametrized in $\theta \in \partial X$ provides a $\mu$-average Busemann function $\mathbf{B}_\mu$ on $X$ in terms of a probability measure $\mu$ on $\partial X$. Under some geometrical assumptions which $X$ fulfills, $\mathbf{B}_\mu$ admits a unique critical point so that we have a barycenter map $\mathrm{bar} : \mathcal{P}(\partial X) \to X$ by assigning to an arbitrary probability measure $\mu$ on $\partial X$ a point in $X$ as its barycenter, a critical point of $\mathbf{B}_\mu$.

Let $(X, g)$ be an Hadamard manifold, a simply connected, complete Riemannian manifold with a metric $g = \langle \cdot, \cdot \rangle$ of non-positive curvature. By Cartan-Hadamard theorem an Hadamard manifold is diffeomorphic to a Euclidean space of same dimension. Refer, for this theorem, to [30,31]

A Euclidean space together with a real hyperbolic space $H^n(\mathbf{R})$ are typical examples of Hadamard manifold. Geometrical properties which an Hadamard manifold enjoys are the following;

(i) Any two points on $X$ can be joined by a unique geodesic.

(ii) Let $\Delta$ be a geodesic triangle in $X$ with interior angles $\alpha_1$, $\alpha_2$, $\alpha_3$ and lengths of opposite side, $\ell_1$, $\ell_2$, $\ell_3$. Then, we have a law of cosines;

$$\ell_3^2 \ \geq \ \ell_1^2 + \ell_2^2 - 2\ell_1\ell_2 \cos\alpha_3.$$

(iii) The distance function from a fixed point $x_o \in X$; $d_{x_o} : X \to \mathbf{R}$, $x \mapsto d(x, x_o)$ is convex, *i.e.*, for any geodesic $\gamma$ in $X$; $t \mapsto \gamma(t)$ the restricted function $d_{x_o} \circ \gamma : t \mapsto d_{x_o}(\gamma(t))$ is convex on $\mathbf{R}$.

Here a function $f : \mathbf{R} \to \mathbf{R}$ is convex, if it satisfies

$$f(at_1 + (1-a)t_2) \leq a\ f(t_1) + (1-a)\ f(t_2), \quad \forall t_1,\ t_2 \in \mathbf{R},\ 0 \leq a \leq 1. \tag{26}$$

Let us define for an Hadamard manifold $(X, g)$ the ideal boundary $\partial X$, or a boundary at infinity.

Let $\gamma,\ \sigma; \mathbf{R} \to X$ be unit speed geodesics on $X$. We say that $\gamma$ is asymptotically equivalent with $\sigma$, denoted by $\gamma \sim \sigma$, when there exists a constant $C > 0$ such that $d(\gamma(t), \sigma(t)) \leq C$ for any $t \geq 0$. The relation $\sim$ is an equivalence relation on the space $\mathrm{Geo}(X)$ of all oriented, unit speed geodesics on $X$. The quotient space $\mathrm{Geo}(X)/\sim$ is called the ideal boundary of $X$, denoted by $\partial X$. An equivalence class represented by $\gamma \in \mathrm{Geo}(X)$ is called an asymptotic class, denoted by $[\gamma]$ or $\gamma(\infty)$. Notice that all geodesics on $X$ are assumed to be of unit speed and oriented.

Let $x \in X$ be an arbitrary point of $X$ and $S_xX$ the space of unit tangent vectors at $x$;

$$S_xX = \{v \in T_xX\ ;\ ||v|| = 1\}.$$

Then we define a map

$$\beta = \beta_x : S_xX \to \partial X; \quad v \mapsto [\gamma], \tag{27}$$

where $\gamma \in \mathrm{Geo}(X)$ is a geodesic given by $\gamma(t) = \exp_x(tv),\ t \in \mathbf{R}$.

**Proposition 1** ([30]). *The map $\beta_x$ is bijective.*

Moreover, we equip the space $X \cup \partial X$ with a topology, called a cone topology as follows. For $x \in X$ and $\theta_1,\ \theta_2 \in X \cup \partial X$ ($x \neq \theta_1,\ x \neq \theta_2$) we define $\angle_x(\theta_1, \theta_2) = \angle(\dot\gamma_1(0), \dot\gamma_2(0))$ angles between a geodesic $\gamma_1$ from $x$ to $\theta_1$ and a geodesic $\gamma_2$ from $x$ to $\theta_2$. For $x \in X$ and $\theta \in \partial X$, $\varepsilon > 0$ let $C_x(\theta, \varepsilon) = \{\theta_1 \in X \cup \partial X\,;\, \theta_1 \neq x,\ \angle_x(\theta, \theta_1) < \varepsilon\}$ be a cone. Further, let $T_x(\theta, \varepsilon) = C_x(\theta, \varepsilon) \setminus B(x, r)$ be a truncated cone ($B(x, r) = \{y \in X | d(y, x) \leq r\}$ is a closed geodesic ball). Then, a topology generated by open geodesic balls in $X$ and such truncated cones is called a cone topology of $X \cup \partial X$. Notice that thus defined cone topology when restricted to $\partial X$ is homeomorphic to the usual topology on $S_xX$ via the mapping $\beta_x$. Refer to [31] for the detail.

Let $(d\theta)_x$ be a standard volume measure on $S_xX$, normalized by $(d\theta)_x(S_xX) = 1$. Through $\beta_x$ we obtain a measure $(\beta_x)_\sharp(d\theta)_x$ on $\partial X$, denoted by $d\theta$.

*4.2. Normalized Busemann Function*

Busemann function on $X$ is introduced to define an average Busemann function in terms of a probability measure on $\partial X$.

Let $\gamma : \mathbf{R} \to X$ be a geodesic on $X$. Define a function $f_t : X \to \mathbf{R}$ for $t > 0$ by

$$f_t(x) = d_x(\gamma(t)) - t = d(x, \gamma(t)) - d(\gamma(0), \gamma(t)).$$

For any $x \in X$ a limit $\lim_{t\to\infty} f_t(x)$ exists, as we will see in Proposition 2. We write this limit as $f_\infty(x)$ and define a function on $X$; $x \mapsto f_\infty(x)$, called Busemann function, denoted by $B_\gamma : X \to \mathbf{R}$; $x \mapsto f_\infty(x)$.

Each level set of $B_\gamma$ is called a horosphere, important in studying geometry of Hadamard manifolds. See [11], for example.

**Example 1.** *In a Euclidean space $(X, g) = (\mathbf{R}^n, g_o)$ let $\gamma$ be a geodesic, $\gamma(t) = (t, 0, \ldots, 0)$. Then, $B_\gamma(\mathbf{x}) = -x^1$ for $\mathbf{x} = (x^1, \ldots, x^n) \in \mathbf{R}^n$ from the following*

$$f_t(\mathbf{x}) = d(\mathbf{x}, \gamma(t)) - t = \sqrt{(x^1 - t)^2 + \sum_{i=2}^{n}(x^i)^2} - t = \frac{-2tx^1 + \sum_{i=2}^n (x^i)^2}{\sqrt{(x^1 - t)^2 + \sum_{i=2}^n (x^i)^2} + t} \to -x^1,$$

*as $t \to \infty$.*

**Example 2.** *Busemann function on $H^n(\mathbf{R})$, an $n$-dimensional real hyperbolic space with standard hyperbolic metric, normalized at $o$, is given*

$$B_\gamma(x) = -\log \frac{1 - |x|^2}{|x - \theta|^2},$$

*where* $\theta = \gamma(\infty) \in S^{n-1}$.

**Proposition 2.** *The functions* $f_t : X \to \mathbf{R}$, $t > 0$, *introduced above, have a limit* $\lim_{t\to\infty} f_t(x)$ *for each* $x \in X$.

From the triangle inequality we have

$$t_1 < t_2 \implies f_{t_1}(x) \geq f_{t_2}(x), \quad \forall x \in X.$$

In fact, since $d(\gamma(t_1), \gamma(t_2)) = t_2 - t_1$, we see

$$d(x, \gamma(t_2)) \leq d(x, \gamma(t_1)) + d(\gamma(t_1), \gamma(t_2)) = d(x, \gamma(t_1)) + (t_2 - t_1)$$

from which the above is derived. On the other hand, we observe uniform boundedness of $f_t$, that is $|f_t(x)| \leq d(\gamma(0), x)$ for any $x \in X$ and $t > 0$ as follows;

$$f_t(x) = d(x, \gamma(t)) - t \leq d(x, \gamma(0)) + d(\gamma(0), \gamma(t)) - t = d(x, \gamma(0))$$

and

$$\begin{aligned} t - d(x, \gamma(t)) &= d(\gamma(0), \gamma(t)) - d(x, \gamma(t)) \\ &\leq d(\gamma(0), x) + d(x, \gamma(t)) - d(x, \gamma(t)) = d(x, \gamma(0)) \end{aligned}$$

so $-d(x, \gamma(0)) \leq f_t(x) \leq d(x, \gamma(0))$.

Therefore, the sequence $\{f_t(x) | t > 0\}$ is bounded and decreasing and then has a limit as $t \to \infty$.

**Proposition 3.** *Let* $\gamma$ *and* $\sigma$ *be geodesics. If* $\gamma \sim \sigma$, *then*

$$B_\gamma(x) - B_\sigma(x) = c, \quad \forall x \in X,$$

*for a constant* $c$.

See [30,31] for this proposition, from which, for $\theta \in \partial X$ Busemann function $B_\gamma$ associated with a geodesic $\gamma$, $[\gamma] = \theta$, gives us a same function on $X$ modulo additive constant. So, let $x_o \in X$ be an arbitrary point of $X$ as a base point. Then, from non-positive curvature of $X$ there exists a unique geodesic $\gamma$ such that $\gamma(0) = x_o$ and $[\gamma] = \theta$.

**Definition 3.** *Let* $x_o \in X$ *and* $\theta \in \partial X$. *Let* $\gamma : \mathbf{R} \to X$ *be a geodesic satisfying* $\gamma(0) = x_o$ *and* $[\gamma] = \theta$. *The Busemann function* $B_\gamma$ *associated to* $\gamma$ *is called normalized Busemann function, denoted by* $B_\theta$.

**Properties of (normalized) Busemann function**:

(i) $B_\theta(x_o) = 0$ for any $\theta \in \partial X$,

(ii) $B_\theta(\gamma(t)) = -t$, $t \in \mathbf{R}$, for any $\theta \in \partial X$, where $\gamma$ is a geodesic satisfying $\gamma(0) = x_o$, $[\gamma] = \theta$.

(iii) Busemann function is Lipschitz continuous;

$$|B_\theta(x) - B_\theta(y)| \le d(x,y), \quad x, y \in X.$$

(iv) Busemann function is of class $C^2$ (refer to [32]).

(v) Gradient vector field $\nabla B_\theta$ satisfies $|(\nabla B_\theta)_x| \equiv 1$ for any $x \in X$ and $\theta \in \partial X$. Here $(\nabla B_\theta)_x \in T_xX$ is defined by $\langle (\nabla B_\theta)_x, v\rangle = v(B_\theta)$, directional derivative of $B_\theta$ with respect to $v \in T_xX$. An integral curve $x(t)$ of $\nabla B_\theta$ passing through a point $x$ is obtained by $x(t) = \sigma(-t)$, where $\sigma$ is a geodesic of $\sigma(0) = x$ and $[\sigma] = \theta$. Moreover, for any $x \in X$ and any vector $v \in S_xX$ there exists a $\theta \in \partial X$ such that $v = -(\nabla B_\theta)_x$ so that $\beta_x(v) = \theta$.

(vi) Busemann function is convex (see (26)), since it is a limit of convex functions.

(vii) From (vi), the Hessian of Busemann function $(\nabla dB_\theta)_x : T_x \times T_x \to \mathbf{R}$ is positive semi-definite at any point $x \in X$, *i.e.*, $(\nabla dB_\theta)_x(v,v) \ge 0$, for any $v \in T_xX$ and $x \in X$, and satisfies

$$(\nabla dB_\theta)_x((\nabla B_\theta)_x, v) = 0, \quad v \in T_xX.$$

Here, for a $C^2$-function $f$ on $X$ the Hessian $\nabla df$ is a symmetric bilinear form, defined by

$$(\nabla df)_x(u,v) = u\,(Vf) - (\nabla_u V)f, \quad u, v \in T_xX,\ x \in X$$

where $V$ is a smooth vector field, an extension of $v$. Notice that for a unit vector $u \in S_xX$

$$(\nabla df)_x(u,u) = \left.\frac{d^2}{dt^2}\right|_{t=0} f(\gamma(t))$$

with respect to a geodesic $\gamma$ such that $\gamma(0) = x$, $\dot\gamma(0) = u$.

**Example 3.** *On a Euclidean space $\nabla dB_\theta = 0$ for any $\theta \in \partial X$(due to Example 1).*

**Example 4.** *On a real hyperbolic space $H^n(\mathbf{R})$, $n \ge 2$,*

$$(\nabla dB_\theta)_x(u,v) = \langle u, v\rangle - \langle \nabla B_\theta, u\rangle \langle \nabla B_\theta, u\rangle, \quad u, v \in T_xX.$$

*See for this* [8].

Let $(X, g)$ be an Hadamard manifold and $\phi : X \to X$ be an isometry of $X$, *i.e.*, a smooth transformation of $X$ satisfying $\phi^* g = g$. An isometry preserves the distance $d$ of $X$, *i.e.*, $d(\phi(x), \phi(y)) = d(x,y)$, $x, y \in X$ and transforms a geodesic $\sigma$ into a new geodesic $\phi \circ \sigma$ so that, if $\sigma \sim \gamma$, then $\phi \circ \sigma \sim \phi \circ \gamma$. Therefore, $\phi$ induces a transformation $\hat\phi$ of $\partial X$, a $\partial X$-extension of $\phi$ as

$$\hat\phi : \partial X \to \partial X; \quad \theta = [\gamma] \mapsto [\phi \circ \gamma].$$

Notice that $(\hat\phi)^{-1} = \widehat{\phi^{-1}}$ for the inverse $\phi^{-1}$ of $\phi$. $\hat\phi$ is a homeomorphism of $\partial X$ in terms of cone topology.

**Proposition 4** (Busemann cocycle formula [33])**.** *Any normalized Busemann function enjoys a cocycle formula with respect to an isometry $\phi$ of $X$;*

$$B_\theta(\phi(x)) = B_{\hat{\phi}^{-1}(\theta)}(x) + B_\theta(\phi(x_o)). \tag{28}$$

**Proof.** Let $\gamma : \mathbf{R} \to X$ be a geodesic, $\gamma(0) = x_o$, $[\gamma] = \theta$. Notice that $\phi \circ \gamma$ is a geodesic with $\phi \circ \gamma(0) = \phi(x_o)$, which, in general, does not coincide with the base point $x_o$. For the Busemann function $B_{\phi^{-1}\circ\gamma}(x)$ with respect to a geodesic $\phi^{-1} \circ \gamma$ we have

$$\begin{aligned} B_{\phi^{-1}\circ\gamma}(x) &= \lim_{t\to\infty} \left(d(x, \phi^{-1} \circ \gamma(t)) - t\right) \\ &= \lim_{t\to\infty} \left(d(\phi(x), \gamma(t)) - t\right) = B_\gamma(\phi(x)). \end{aligned} \tag{29}$$

On the other hand, $\phi^{-1} \circ \gamma$ belongs to $\widehat{\phi}^{-1}(\theta)$ and $(\phi^{-1} \circ \gamma)(0) = \phi^{-1}(x_o)$. Let $\sigma$ be a geodesic such that $[\sigma] = \widehat{\phi}^{-1}(\theta)$, $\sigma(0) = x_o$. Then, the normalized Busemann function $B_{\widehat{\phi}^{-1}(\theta)}$ is given by $B_\sigma$. Since $\phi^{-1} \circ \gamma$ and $\sigma$ belong to the same $\widehat{\phi}^{-1}(\theta)$, from (29) $B_\sigma - B_{\phi^{-1}\circ\gamma}$ is a constant function on $X$. This constant is given from the above by $\left(B_\sigma - B_{\phi^{-1}\circ\gamma}\right)(x_o) = -B_{\phi^{-1}\circ\gamma}(x_o) = -B_\gamma(\phi(x_o))$ so that on $X$

$$B_\sigma(x) - B_{\phi^{-1}\circ\gamma}(x) \equiv -B_\gamma(\phi(x_o)). \tag{30}$$

Since $B_\theta$ is given by $B_\gamma$, (28) is obtained from (30). □

In what follows, every normalized Busemann function $B_\theta$ on $X$ is assumed to be continuous with respect to $\theta \in \partial X$ for each fixed point $x \in X$. This assumption is guaranteed by a real hyperbolic space. See Example 3. Rank one symmetric spaces of non-compact type and Damek-Ricci spaces satisfy this assumption, as is seen in [25].

**Definition 4** ([21])**.** *An Hadamard manifold $(X, g)$ is said to satisfy visibility axiom, if for any distinct ideal point $\theta$, $\theta_1$ of $\partial X$ there exists a geodesic $\gamma : \mathbf{R} \to X$ such that $\gamma(+\infty) = \theta$ and $\gamma(-\infty) = \theta_1$. Here $\gamma(-\infty) \in \partial X$ defined by $[\gamma^-]$, where $\gamma^-$ is the geodesic of reversed orientation given by $\gamma^-(t) = \gamma(-t)$, $t \in \mathbf{R}$.*

A Euclidean space does not satisfy visibility axiom.

**Proposition 5.** *Let $(X, g)$ be an Hadamard manifold. $(X, g)$ satisfies visibility axiom if and only if, for any $\theta \in \partial X$*

$$\lim_{x\to\theta_1} B_\theta(x) = +\infty,$$

*provided $\theta_1 \neq \theta$. Refer to* [31] *for this.*

Notice $B_\theta(x) = -\infty$, if $x \to \theta$, from property (i) of normalized Busemann function.

*4.3. Average Busemann Function and Barycenter*

In what follows, an Hadamard manifold satisfies visibility axiom and Busemann function $B_\theta(x)$ is continuous with respect to every $\theta$.

Let $\partial X$ be, as before, the ideal boundary of an Hadamard manifold $(X, g)$, diffeomorphic to $S^{n-1}$, $n = \dim X$ and $d\theta$ a normalized standard measure on $\partial X$.

Denote by $\mathcal{P}(\partial X)$ a space of probability measures $\mu$ on $\partial X$ which is absolutely continuous with respect to $d\theta$ ($\mu \ll d\theta$) whose density function $p = p(\theta)$ is of $C^0$ and positive;

$$\mathcal{P}(\partial X) = \left\{ \mu = p(\theta)\, d\theta\,;\ \int_{\partial X} p(\theta)\, d\theta = 1,\ p \in C^0(\partial X),\ p(\theta) > 0\ (\forall \theta \in \partial X) \right\}.$$

**Definition 5.** *Let $\mu \in \mathcal{P}(\partial X)$. Then, a function $\mathbf{B}_\mu : X \to \mathbf{R}$, called $\mu$-average Busemann function, is defined by*

$$\mathbf{B}_\mu(x) = \int_{\theta \in \partial X} B_\theta(x)\, d\mu(\theta).$$

Average Busemann function for any $\mu \in \mathcal{P}(\partial X)$ fulfills the following;

(i) For any $\mu \in \mathcal{P}(\partial X)$ each $\mathbf{B}_\mu$ is convex on $X$ and $\mathbf{B}_\mu(x_o) = 0$.

(ii) $\mathbf{B}_\mu(\gamma(t)) \to +\infty$, as $t \to \infty$, where $\gamma : \mathbf{R} \to X$ is an arbitrary geodesic in $X$ (Theorem 12).

(iii) $\mathbf{B}_\mu$ is Lipschitz continuous, in fact, $|\mathbf{B}_\mu(x) - \mathbf{B}_\mu(y)| \leq |d(x, y)|$ for $x, y \in X$.

(iv) The gradient vector field $\nabla \mathbf{B}_\mu$ is defined on $X$ as

$$(\nabla \mathbf{B}_\mu)_x = \int_{\partial X} (\nabla B_\theta)_x\, d\mu(\theta), \quad x \in X$$

and satisfies $|(\nabla \mathbf{B}_\mu)_x| \leq 1,\ x \in X$.

(v) the Hessian $\nabla d\mathbf{B}_\mu$ can be defined as $\mu$-average Hessian;

$$(\nabla d\mathbf{B}_\mu)_x(u, v) = \int_{\partial X} (\nabla dB_\theta)_x(u, v)\, d\mu(\theta), \quad u, v \in T_x X\, x \in X,$$

provided $(X, g)$ is of bounded Ricci curvature and moreover $\Delta B_\theta$, the Laplacian of $B_\theta$ together with $d(\Delta B_\theta)$ are uniformly bounded with respect to $x \in X$ and $\theta \in \partial X$. Here $\Delta f = -\mathrm{trace}\ \nabla df$ for a $C^2$-function $f$ on $X$. This is derived from Bochner formula (see [23]). If $(X, g)$ is asymptotically harmonic ([34]), *i.e.*, $\Delta B_\theta \equiv c$ for any $\theta$, and of bounded Ricci curvature, the average Hessian $\nabla d\mathbf{B}_\mu$, $\mu \in \mathcal{P}(\partial X)$ is defined.

**Definition 6.** *Let $\mu \in \mathcal{P}(\partial X)$. A critical point of $\mu$-average Busemann function $\mathbf{B}_\mu$ is called a barycenter of $\mu$.*

For a $C^1$-function $f : X \to \mathbf{R}$, $y \in X$ is called a critical point of $f$, if one of the following equivalent conditions holds;

(i) the differential of $f$ at $y$ vanishes along all directional vector, *i.e.*,

$$\left.\frac{d}{dt}\right|_{t=0} f(x(t)) = 0$$

for any $C^1$-curve $x(t)$ of $x(0) = y$,

(ii) the one-form $df$, or the gradient vector field $\nabla f$ vanishes at $y$.

**Observation.** For $\mu = p(\theta)\, d\theta \in \mathcal{P}(\partial X)$, $x \in X$ is a barycenter of $\mu$ if and only if $(d\mathbf{B}_\mu)_x(u) = 0$ for any $u \in T_xX$, which is equivalent to stating that a measure $\tau$ defined by $\tau = (dB_\theta)_x(u)\, d\mu = \langle (\nabla B_\theta)_x, u\rangle\, p(\theta)\, d\theta$ is a tangent vector to $\mathcal{P}(\partial X)$ at $\mu$ for each $u \in T_xX$.

**Theorem 12.** *If, as is assumed, an Hadamard manifold $(X, g)$ satisfies visibility axiom and Busemann function is continuous with respect to any $\theta \in \partial X$. Then, every $\mu \in \mathcal{P}(\partial X)$ admits a barycenter.*

**Proof.** This theorem is proved by Besson, Courtois and Gallot in [8] by showing $\mathbf{B}_\mu(\gamma(t)) \to \infty$ as $t \to +\infty$ along any geodesic $\gamma$ of $X$. However, they assume that all probability measures on $\partial X$ are without atom and an Hadamard manifold $(X, g)$ is of special type, *i.e.*, a rank one symmetric space of non-compact type. We restrict the space of probability measures as $\mathcal{P}(\partial X)$. However, we relax the assumptions concerning an Hadamard manifold $(X, g)$ and then, assume only that $(X, g)$ satisfies visibility axiom and Busemann function is continuous with respect to any $\theta \in \partial X$.

Let $C > 0$ be a constant and set $A_C = \{y \in X\,;\, \mathbf{B}_\mu(y) \le C\}$. $A_C$ is a convex set and $x_o \in A_C$, since $\mathbf{B}_\mu$ is convex and $\mathbf{B}_\mu(x_o) = 0$. Note $A_C$ contains a geodesic ball $\{y \in X\,;\, d(y, x_o) \le C/2\}$ since $\mathbf{B}_\mu$ is Lipschitz. Let $\mu \in \mathcal{P}(\partial X)$ and $\gamma$ be a geodesic satisfying $\gamma(0) = x_o$ and $[\gamma] = \theta$. Then it is possible to verify $\lim_{t\to+\infty} \mathbf{B}_\mu(\gamma(t)) = +\infty$ in the following steps;

**Step I.** Since Busemann function $B_\theta$ is convex and $B_\theta(x_0) = 0$ for any $\theta$, we have

$$d(\gamma(t_1), x_0)\, B_\theta(\gamma(t) \ge d(\gamma(t), x_0)\, B_\theta(\gamma(t_1) \tag{31}$$

in other words

$$t_1\, B_\theta(\gamma(t)) \ge t\, B_\theta(\gamma(t_1)) \quad (0 \le t_1 \le t).$$

In fact, if we set $a = t_1/t$, then, $0 \le a \le 1$ and we have $t_1 = (1-a)\,0 + a\,t$. The Convex function $B_\theta(\gamma(t))$ fulfills

$$B_\theta(\gamma(t_1)) \le (1-a)\, B_\theta(\gamma(0)) + a\, B_\theta(\gamma(t)) = a\, B_\theta(\gamma(t)),$$

that is

$$B_\theta(\gamma(t_1)) \le a\, B_\theta(\gamma(t)) \tag{32}$$

which is just (31).

**Step II.** Fix $t_1 > 0$ of Step I. Take an arbitrary $\theta_0 \in \partial X$ and fix it. Let $\gamma$ be a geodesic of $\gamma(0) = x_0$ and $[\gamma] = \theta_0$. For any $t > 0$ set

$$J_{\theta_0}(t) := \{\theta \in \partial X\,;\, B_\theta(\gamma(t)) \le 0\}.$$

We show that there exists a $t_1 \in (0,\infty)$ such that $\mu(J_{\theta_0}(t_1)) < 1$, as follows.

Since $B_\theta(x)$ is continuous with respect to $\theta$, the set $J_{\theta_0}(t)$ is compact in $\partial X$. We see $\theta_0 \in J_{\theta_0}(t)$. From (32) it holds

$$J_{\theta_0}(t) \subset J_{\theta_0}(t_1) \quad (0 \le t_1 \le t). \tag{33}$$

It follows then

$$\bigcap_{t\in[0,\infty)} J_{\theta_0}(t) = \{\theta_0\},$$

because from the visibility axiom, we have from Proposition 3.4 for any $\theta \in \partial X$ such that $\theta \neq \theta_0$ $\lim_{t\to\infty} B_\theta(\gamma(t)) = +\infty$. Moreover, from (33) for the $\mu$ we find

$$\lim_{t\to\infty} \mu(J_{\theta_0}(t)) = \mu\left(\bigcap_{t\in[0,\infty)} J_{\theta_0}(t)\right) = \mu(\{\theta_0\}) = 0. \tag{34}$$

Therefore, we have a $t_1 \in (0,\infty)$ such that $\mu(J_{\theta_0}(t_1)) < 1$. Here, notice $\mu(J_{\theta_0}(t)) \le \mu(J_{\theta_0}(t_1)) < 1$ for any $t \ge t_1 > 0$.

**Step III.** Let $K$ be a compact subset of $\partial X \setminus J_{\theta_0}(t_1)$ satisfying $\mu(K) > 0$. It is possible to choose such a $K$. Then, from (33) it holds for any $t \ge t_1$ that $K \subset \partial X \setminus J_{\theta_0}(t)$, and $B_\theta(\gamma(t)) \ge 0$ for any $\theta \in \partial X \setminus J_{\theta_0}(t)$. So,

$$\begin{aligned}\int_{\partial X} B_\theta(\gamma(t))\, d\mu(\theta) &= \int_{J_{\theta_0}(t)} B_\theta(\gamma(t))\, d\mu(\theta) + \int_{\partial X\setminus J_{\theta_0}(t)} B_\theta(\gamma(t))\, d\mu(\theta)\\ &\ge \int_{J_{\theta_0}(t)} B_\theta(\gamma(t))\, d\mu(\theta) + \int_K B_\theta(\gamma(t))\, d\mu(\theta).\end{aligned}$$

Since $K$ is compact, we choose a constant $C > 0$ satisfying

$$B_\theta(\gamma(t_1)) \ge C > 0, \quad \forall \theta \in K.$$

From (32), we have

$$\int_{\partial X} B_\theta(\gamma(t))\, d\mu(\theta) \ge \frac{t}{t_1} \int_{J_{\theta_0}(t)} B_\theta(\gamma(t_1))\, d\mu(\theta) + C\,\frac{t}{t_1}\,\mu(K).$$

To estimate the first term of the RHS we choose $D \ge 0$ satisfying

$$B_\theta(\gamma(t_1)) \ge -\sup\{|B_\theta(\gamma(t_1))|\,;\, \theta \in \partial X\} = -D.$$

In fact, since $\partial X$ is compact, $B_\theta(\gamma(t_1))$, as a continuous function of $\theta$, is bounded with respect to $\theta$. Therefore, the above is written as

$$\int_{\partial X} B_\theta(\gamma(t))\, d\mu(\theta) \;\ge\; \frac{t}{t_1}\,\left(-D\,\mu(J_{\theta_0}(t)) + C\,\mu(K)\right).$$

We let $t \to +\infty$ and then, from (34) we have

$$\lim_{t\to\infty} \mathbf{B}_\mu(\gamma(t)) = \lim \int_{\partial X} B_\theta(\gamma(t))\, d\mu(\theta) = +\infty$$

from which it follows that the closed set $A_C$ is bounded and hence is compact. Therefore, $\mathbf{B}_\mu$ admits a minimal point $x \in X$, namely, $x$ is a barycenter of $\mu$. $\square$

**Proposition 6.** *Let $(X, g)$ be an Hadamard manifold of bounded Ricci curvature. If $(X, g)$ is asymptotically harmonic, then the following holds; If there exists $\mu_0 \in \mathcal{P}(\partial X)$ such that $\mu$-average Hessian $\nabla d\mathbf{B}_{\mu_0}$ is positive definite at every point in $X$, then, for any $\mu \in \mathcal{P}(\partial X)$ $\mu$-average Hessian $\nabla d\mathbf{B}_{\mu}$ is also positive definite at every point in $X$.*

**Proof.** Let $x \in X$ and $u \in T_xX$. Then, for a geodesic $\gamma$ in $X$, $\gamma(0) = x$, $\dot{\gamma}(0) = u$ we have

$$\begin{aligned}(\nabla d\mathbf{B}_{\mu_0})_x(u,u) &= \frac{d^2}{dt^2}\Big|_{t=0} \mathbf{B}_{\mu_0}(\gamma(t)) = \int_{\theta\in\partial X} \frac{d^2}{dt^2}\Big|_{t=0} B_\theta(\gamma(t)) d\mu_0(\theta) \\ &= \int_{\partial X} (\nabla dB_\theta)_x(u,u) d\mu_0(\theta).\end{aligned}$$

Similarly for any $\mu \in \mathcal{P}(\partial X)$, we have

$$\begin{aligned}(\nabla d\mathbf{B}_{\mu})_x(u,u) &= \frac{d^2}{dt^2}\Big|_{t=0} \mathbf{B}_{\mu}(\gamma(t)) = \int_{\theta\in\partial X} \frac{\partial^2}{\partial t^2}\Big|_{t=0} B_\theta(\gamma(t))\, d\mu(\theta) \\ &= \int_{\partial X} (\nabla dB_\theta)_x(u,u)\, d\mu(\theta).\end{aligned} \tag{35}$$

Let $C = \min_{\theta\in\partial X} \frac{d\mu}{d\mu_0}(\theta) > 0$. It is concluded then from the above

$$(\nabla d\mathbf{B}_{\mu})_x(u,u) \geq C(\nabla d\mathbf{B}_{\mu_0})_x(u,u) > 0, \quad \forall u \in T_xX(\neq 0), x \in X.$$

□

**Theorem 13.** *Let $(X, g)$ be an Hadamard manifold satisfying the above assumptions. Then, any $\mu \in \mathcal{P}(\partial X)$ admits a unique barycenter.*

In fact, Theorem 12 asserts an existence of a barycenter for any $\mu$. From Proposition 6 a barycenter must be unique, since, if, otherwise, $\mu$ admits barycenters $y_1$, $y_2$, $y_1 \neq y_2$, then $f(t) := \mathbf{B}_\mu(\gamma(t))$ along a geodesic $\gamma : \mathbf{R} \to X$ joining $y_1$ and $y_2$ satisfies $f'(0) = f'(d) = 0$ ($d = d(y_1, y_2)$) and $f''(t) > 0$, $t \in [0, d]$ because of the positive definiteness of $\mu$-average Hessian (35). So, $f(t)$ must be constant along $\gamma$. This contradicts property (ii) of $\mu$-average Busemann function. Hence uniqueness is proved.

**Proposition 7** (average Busemann cocycle formula)**.** *Let $\phi$ be an isometry of an Hadamard manifold $(X, g)$. Then for any $\mu \in \mathcal{P}(\partial X)$*

$$\mathbf{B}_\mu(\phi^{-1}x) = \mathbf{B}_{(\hat{\phi})_\sharp\mu}(x) + \mathbf{B}_\mu(\phi^{-1}x_o). \tag{36}$$

**Proof.** Integrate the Busemann cocycle formula (28)

$$B_\theta(\phi^{-1}(x)) = B_{\hat{\phi}(\theta)}(x) + B_\theta(\phi^{-1}x_o)$$

for the inverse isometry $\phi^{-1}$ with respect to a measure $\mu$. We then get (36). □

From Theorem 13 we define a map, called barycenter map

$$\mathrm{bar} \,:\, \mathcal{P}(\partial X) \to X; \quad \mu \mapsto y, \tag{37}$$

by assigning a barycenter $y$ to $\mu$.

**Example 5.** *The standard measure $d\theta$ has $\mathrm{bar}(d\theta) = x_o$, the base point as its barycenter. In fact, we observe*

$$\int_{\partial X} \langle (\nabla B_\theta)_{x_o}, u\rangle \, d\theta = 0, \quad \forall u \in T_{x_o}X,$$

*since $d\theta = (\beta_{x_o})_\sharp (d\theta)_{x_o}$ is the push-forward of the spherical measure $(d\theta)_{x_o}$ of $S_{x_o}X$, where $\beta_{x_o}$ is a map given in (27), and one has $\beta_{x_o}(-(\nabla B_\theta)_{x_o}) = \theta$ so that $\beta_{x_o}(v) = \theta$ implies $v = -(\nabla B_\theta)_{x_o}$. Then the LHS of the above is written as*

$$\int_{v\in S_{x_o}X} \langle (\nabla B_{\beta_{x_o}(v)})_{x_o}, u\rangle d\theta_{x_o}(v) = -\int_{v\in S_{x_o}X} \langle v, u\rangle d\theta_{x_o}(v) = 0.$$

*Here, the last integration is derived from a standard formula on $S^{n-1}$ which is described as $\sum_{i=1}^n (\theta^i)^2 = 1$ with respect to the standard coordinates $\theta = (\theta^1, \ldots, \theta^n) \in \mathbf{R}^n$;*

$$\int_{\theta\in S^{n-1}} \theta^i (d\theta)_{S^{n-1}} = 0, \quad i = 1, \cdots, n.$$

*4.4. Barycenter Map*

In this subsection we will verify that the barycenter map (37) enjoys a fibration over an Hadamard manifold $X$ in terms of Fisher metric $G$. Before giving a detailed argument, we prepare some special probability measures on $\partial X$ which play a crucial role in the barycenter map, that is, Poisson kernel measures. Here, Poisson kernel is a fundamental solution of Dirichlet problem at $\partial X$; given a $C^0$-function $f$ on $\partial X$, find a function $u$ on $X$ which satisfies the Laplace equation $\Delta u = 0$ and the boundary condition $\lim_{x\to\theta} u(x) = f(\theta)$ for $\theta \in \partial X$.

**Definition 7** ([3,35])**.** *A function $P_\theta(x) = P(x, \theta)$ on $X$ is called a Poisson kernel, normalized at $x_o$, for $\theta \in \partial X$ if it satisfies*

*(i)* $\Delta P(x, \theta) = 0$ *and* $P(x, \theta) > 0$ *for any* $x \in X$ *and* $\theta \in \partial X$.

*(ii)* $P(x_o, \theta) = 1$ *for any* $\theta \in \partial X$.

*(iii) for any* $\theta \in \partial X$, $P(x, \theta) \in C^0(X \cup \partial X \setminus \{\theta\})$ *as an extension function on* $X \cup \partial X$ *and* $\lim_{x\to\theta_1} P(x, \theta) = 0$ *for* $\theta_1 \neq \theta$.

The solution $u = u(x)$ of the Dirichlet problem on $\partial X$ is described as an integration form;

$$u(x) = \int_{\theta\in\partial X} P(x, \theta)\, f(\theta)\, d\theta$$

so, $P(x, \theta)\, d\theta \in \mathcal{P}(\partial X)$ for each $x \in X$.

**Example 6.** *On a real hyperbolic space $H^n(\mathbf{R})$ of standard hyperbolic metric of Poicaré ball model, the Poisson kernel is given by*

$$P(x,\theta) = \left(\frac{1-|x|^2}{|x-\theta|^2}\right)^{n-1}, \quad \theta \in \partial X = S^{n-1}.$$

**Example 7.** *The Poisson integral formula, well known in potential theory, is for a bounded harmonic function $h = h(z)$, $z = re^{i\varphi} \in \{z \in \mathbf{C} | |z| \leq 1\}$*

$$h(re^{i\varphi}) = \frac{1}{2\pi} \int_{0\leq\theta\leq 2\pi} \frac{1-r^2}{1-2r\cos(\varphi-\theta)+r^2} f(e^{i\theta})\, d\theta,$$

*where $f$ is a bounded function on $S^1$. The kernel function $(1-r^2)/(1-2r\cos(\varphi-\theta)+r^2)$ is just the Poisson kernel $P(z,\theta) = (1-|z|^2)/|z-\theta|^2$ of the hyperbolic plane $H^2(\mathbf{R})$. See, for example* [20].

**Definition 8.** *A Poisson kernel on an Hadamard manifold $(X, g)$ is called Busemann-Poisson kernel, when it has the following form*

$$P(x,\theta) = \exp\{-Q\, B_\theta(x)\}, \quad x \in X,\ \theta \in \partial X,$$

*where $Q > 0$ is volume entropy of $(X, g)$, the exponential growth rate of the volume of $(X, g)$*

$$Q = \lim_{r\to\infty} \frac{1}{r} \log \mathrm{vol}\, B(x,r)$$

*for a geodesic ball $B(x, r)$.*

**Remark 5.** *For volume entropy refer to* [8] *in which the following theorem, Theorem of Manning, (*[7]*) is cited; if $Q_{top}$ denotes the topological entropy of a compact Riemannian manifold $Y$ of non-positive curvature, then one has*

*(i) $Q(\tilde{Y}) \leq Q_{top}(Y)$,*

*(ii) $Q(\tilde{Y}) = Q_{top}(Y)$, provided the curvature of $Y$ is negative or zero.*

*Here, $\tilde{Y}$ is the universal covering space of $Y$ and the topological entropy $Q_{top}(Y)$ is defined by*

$$Q_{top}(Y) = \lim_{R\to\infty} \frac{1}{R} \log(\sharp\{\gamma \,|\, \ell(\gamma) \leq R\}),$$

*where $\gamma$ denotes a periodic geodesic in $Y$ of length $\ell(\gamma)$ and $\sharp\{\gamma \,|\, \ell(\gamma) \leq R\}$ denotes the number of periodic geodesics of length not greater than $R$.*

For example $Q = 0$ for a Euclidean space and $Q = n-1$ for a real hyperbolic space $H^n(\mathbf{R})$ of standard hyperbolic metric.

**Remark 6.** *Any Damek-Ricci space admits a Busemann-Poisson kernel (refer to* [4]*). See also* [8] *for a rank one symmetric space of non-compact type which is just a member of Damek-Ricci spaces, as observed by using Iwasawa decomposition of isometry groups.*

**Theorem 14.** *Let $(X, g)$ be an Hadamard manifold satisfying the assumptions in Theorem 12 and Proposition 6. If $(X, g)$ admits a Busemann-Poisson kernel, then, for $\mu_x := P(x, \theta)\, d\theta \in \mathcal{P}(\partial X)$*

*(i)* $\mathrm{bar}(\mu_x) = x$ *for any* $x \in X$ *and*

*(ii) at any point* $y \in X$, $(\nabla d\mathbf{B}_{\mu_x})_y$ *is positive definite.*

The statement (i) is shown in [8]. From (i) the barycenter map $\mathrm{bar}$ is onto.

**Definition 9.** *Let $\mu = p(\theta)\, d\theta \in \mathcal{P}(\partial X)$ and $x \in X$ be a barycenter of $\mu$. We define a linear map*

$$\nu_x^\mu : T_xX \to T_\mu \mathcal{P}(\partial X); \quad u \mapsto \nu_x^\mu(u) = (dB_\theta)_x(u)\ \mu = \langle (\nabla B_\theta)_x, u\rangle\ p(\theta) d\theta. \tag{38}$$

Notice that the map $\nu_x^\mu$ is injective.

**Proof of Theorem 14.** (i) Let $u \in T_xX$ and $x(t)$ a $C^1$-curve in $X$ such that $x(0) = x$, $\dot{x}(0) = u$. Differentiate $\int_{\partial X}\ P(x(t), \theta)\, d\theta \equiv 1$ as

$$\begin{aligned} 0 &= \frac{d}{dt}\Big|_{t=0} \int P(x(t), \theta)\, d\theta \\ &= \int \frac{\partial}{\partial t}\Big|_{t=0} \exp\{-QB_\theta(x(t))\}\, d\theta \\ &= \int -Q(dB_\theta)_x(\dot{x}(0))\ \exp\{-QB_\theta(x(0))\} d\theta \\ &= -\,Q\ \frac{d}{dt}\Big|_{t=0} \int_{\partial X} B_\theta(x(t))\, d\mu_x = -Q(d\mathbf{B}_{\mu_x})_x(u). \end{aligned}$$

So, $x = \mathrm{bar}\ \mu_x$.

For a proof of (ii) we first show

**Assertion 4.** The measure $\mu_x$ satisfies

$$(\nabla d\mathbf{B}_{\mu_x})_x(u, v) = Q\ G_{\mu_x}(\nu_x^{\mu_x}(u), \nu_x^{\mu_x}(v)), \quad u, v \in T_xX \tag{39}$$

in terms of the Fisher metric $G$, where $\nu_x^{\mu_x}$ is the linear map defined in (38).

It suffices to show this in case of $u = v$. Let $\gamma$ be a geodesic in $X$ satisfying $\gamma(0) = x$, $\dot{\gamma}(0) = u$. Then we have

$$0 = \frac{d^2}{dt^2}\Big|_{t=0} \int_{\partial X} P(\gamma(t), \theta)\, d\theta.$$

However, this is

$$\begin{aligned} &\int_{\partial X} \frac{\partial^2}{\partial t^2}\Big|_{t=0} \exp\{-Q\, B_\theta(\gamma(t))\}\, d\theta \\ &= -\,Q \int_{\partial X} \left\{(\nabla dB_\theta)_x(u, u) - Q\, \{(dB_\theta)_x(u)\}^2\right\} \exp\{-Q\ B_\theta(x)\}\, d\theta \\ &= -\,Q\, \{(\nabla d\mathbf{B}_{\mu_x})_x(u, u) - Q\, G_{\mu_x}(\nu_x^{\mu_x}(u), \nu_x^{\mu_x}(u))\} \end{aligned}$$

showing (39).

From this assertion (ii) is proved as follows. At $y \in X$ we have, since $(\nabla dB_\theta)_y(\cdot,\cdot)$ is positive semi-definite,

$$(\nabla d\mathbf{B}_{\mu_x})_y(u,u) = \int (\nabla dB_\theta)_y(u,u)\, d\mu_x(\theta) = \int (\nabla dB_\theta)_y(u,u) P(x,\theta)\, d\theta$$
$$\geq C \int (\nabla dB_\theta)_y(u,u) P(y,\theta)\, d\theta = C\, (\nabla d\mathbf{B}_{\mu_y})_y(u,u)$$

for any $u \in T_yX$, where $C = \inf_{\theta\in\partial X} P(x,\theta)/P(y,\theta) > 0$. $\square$

Now we will investigate the map $\mathrm{bar} : \mathcal{P}(\partial X) \to X$.

**Theorem 15.** *The barycenter map* $\mathrm{bar} : \mathcal{P}(\partial X) \to X$ *gives a projection of a fibre space whose total space is* $\mathcal{P}(\partial X)$ *and base space is* $X$ *with fibres* $\mathrm{bar}^{-1}(x)$ *over* $x \in X$. *In fact, let* $x \in X$ *and* $\mu \in \mathrm{bar}^{-1}(x)$. *Then*

$$T_\mu \mathcal{P}(\partial X) = T_\mu \mathrm{bar}^{-1}(x) \oplus \mathrm{Im}\ \nu_x^\mu \quad (\dim \mathrm{Im}\ \nu_x^\mu = n),$$

*as an orthogonal direct sum of the vertical subspace* $T_\mu \mathrm{bar}^{-1}(x)$ *and the horizontal subspace* $\mathrm{Im}\ \nu_x^\mu$ *with respect to Fisher metric* $G_\mu$.

This orthogonal decomposition indicates that $N = \{N_\mu = \mathrm{Im}\ \nu_x^\mu\ ;\ \mu \in \mathrm{bar}^{-1}(x)\}$ distributes a normal bundle to each fibre $\mathrm{bar}^{-1}(x)$, $x \in X$. Notice that $\mathrm{bar}^{-1}(x)$ is path-connected, since, for $\mu$, $\mu_1 \in \mathrm{bar}^{-1}(x)$ $(1-t)\mu + t\mu_1$, $0 \leq t \leq 1$, also belongs to $\mathrm{bar}^{-1}(x)$.

We will show that the vertical subspace $T_\mu \mathrm{bar}^{-1}(x)$ is orthogonal to $N_\mu = \mathrm{Im}\ \nu_x^\mu$. Let $u \in T_xX$ and $\tau \in T_\mu \mathrm{bar}^{-1}(x)$ and take $\mu(t) = \mu + t\tau$ for sufficiently small $|t|$. Then, $\mu(t) \in \mathrm{bar}^{-1}(x)$. So, for a sufficiently small $|t|$, we have

$$0 = \int (dB_\theta)_x(u) d\mu(t)(\theta) = \int (dB_\theta)_x(u) d(\mu + t\tau)(\theta)$$
$$= \int (dB_\theta)_x(u) d\mu(\theta) + t \int (dB_\theta)_x(u) d\tau(\theta)$$
$$= t \int (dB_\theta)_x(u) d\tau(\theta) = t\, G_\mu(\nu_x^\mu(u), \tau).$$

This means the orthogonality of $T_\mu \mathrm{bar}^{-1}(x)$ and $N_\mu = \mathrm{Im}\ \nu_x^\mu$.

Since the image $N_\mu = \mathrm{Im}\ \nu_x^\mu$ is a finite dimensional subspace of $T_\mu \mathcal{P}(\partial X)$, the direct sum decomposition is easily shown and so we skip.

*4.5. Fibres* $\mathrm{bar}^{-1}(x)$ *and Geodesics*

We discussed in Section 3 several properties and propositions of geodesics on a space of probability measures. In this section we will investigate under which condition a geodesic of $\mathcal{P}(\partial X)$ is contained in a fibre $\mathrm{bar}^{-1}(x)$.

**Theorem 16.** *Let $(X, g)$ be an Hadamard manifold satisfying the assumptions in Theorem 12 and Proposition 6, and admitting a Busemann-Poisson kernel.*

*Let $\mu \in \mathrm{bar}^{-1}(x)$ and $\tau \in T_\mu \mathrm{bar}^{-1}(x)$, $|\tau|_{G,\mu} = 1$. Then a geodesic $\mu(t) = \exp_\mu t\tau$ entirely belongs to $\mathrm{bar}^{-1}(x)$ for any $t$ at which $\mu(t)$ is well-defined, if and only if $\tau$ fulfills $H_\mu(\tau, \tau) = 0$.*

Here $H$ is the second fundamental form of a submanifold $\mathrm{bar}^{-1}(x)$ of the ambient space $\mathcal{P}(\partial X)$ (see Equation (5) in Section 2).

**Proof.** From Theorem 10, Section 3 the geodesic $\mu(t)$ is given by

$$\mu(t) = \left(\cos\frac{t}{2} + \sin\frac{t}{2}\,\frac{d\tau}{d\mu}(\theta)\right)^2 \mu.$$

Then $\mu(t) \in \mathrm{bar}^{-1}(x)$ for all $t$ if and only if for any $u \in T_x X$

$$0 = \int_{\theta\in\partial X} (dB_\theta)_x(u)\, d\mu(t)(\theta).$$

However the RHS is

$$\cos^2\frac{t}{2}\int (dB_\theta)_x(u)\, d\mu(\theta) + 2\cos\frac{t}{2}\sin\frac{t}{2}\int (dB_\theta)_x(u)\frac{d\tau}{d\mu}(\theta)\, d\mu(\theta) + \sin^2\frac{t}{2}\int (dB_\theta)_x(u)\left(\frac{d\tau}{d\mu}\right)^2(\theta)\, d\mu(\theta)$$

for all $t$. Since $\mu \in \mathrm{bar}^{-1}(x)$ and $\tau \in T_\mu \mathrm{bar}^{-1}(x)$, this is equivalent to

$$0 = \int (dB_\theta)_x(u)\left(\frac{d\tau}{d\mu}\right)^2(\theta)\, d\mu(\theta)$$

which is reduced by the aid of Levi–Civita connection formula to

$$0 = \int (dB_\theta)_x(u)\ \left(\frac{d\nabla_\tau \tau}{d\mu}\right)(\theta)\, d\mu(\theta) = -2\, G_\mu(\nu_x^\mu(u), \nabla_\tau \tau)$$

which means that $H(\tau, \tau) = 0$. Conversely, if $\tau$ satisfies $H(\tau, \tau) = 0$, then it is easy to see that $\mu(t) = \exp_\mu t\tau$ belongs to the fibre $\mathrm{bar}^{-1}(x)$ by following reversely the above argument. □

**Theorem 17.** *Let $\mu$, $\mu^* \in \mathrm{bar}^{-1}(x)$, $x \in X$ ($\mu \neq \mu^*$). Then, a geodesic $\mu(t)$ joining $\mu$ and $\mu^*$ lies entirely on $\mathrm{bar}^{-1}(x)$ if and only if*

$$\int_{\partial X} (dB_\theta)_x(u)\sqrt{\frac{d\mu^*}{d\mu}}(\theta)\, d\mu(\theta) = 0, \quad \forall u \in T_x X. \tag{40}$$

**Proof.** The geodesic $\mu(t)$ joining $\mu$ and $\mu^*$ is written from Theorem 11 by $\exp_\mu t\tau$ of an initial vector

$$\tau = \frac{1}{\sin\frac{\ell}{2}}\left\{\sqrt{\frac{d\mu^*}{d\mu}}(\theta) - \cos\frac{\ell}{2}\right\}\mu, \qquad \ell = \ell(\mu, \mu*) > 0. \tag{41}$$

Then, $\mu(t)$ lies on $\mathrm{bar}^{-1}(x)$ if and only if the following conditions hold, that is, $\tau$ is tangent to $\mathrm{bar}^{-1}(x)$, namely,

$$G_\mu(\nu_x^\mu(u), \tau) = 0 \tag{42}$$

for any $u \in T_xX$, and that

$$H(\tau,\tau) = -\frac{1}{2}\int_{\partial X}(dB_\theta)_x(u)\left(\frac{d\tau}{d\mu}\right)^2(\theta)\,d\mu(\theta) = 0. \tag{43}$$

Equation (42) is equivalent to (40), since $\tau$ is given by (41). On the other hand, (43) is written as

$$\begin{aligned} 0 &= \int_{\partial X}(dB_\theta)_x(u)\left(\sqrt{\frac{d\mu^*}{d\mu}}(\theta) - \cos\frac{\ell}{2}\right)^2 d\mu(\theta) \\ &= -2\cos\frac{\ell}{2}\int_{\partial X}(dB_\theta)_x(u)\sqrt{\frac{d\mu^*}{d\mu}}(\theta)\,d\mu(\theta) \end{aligned}$$

for any $u \in T_xX$. This condition is also (40), so we get Theorem 17. □

**Example 8.** *Let $\mu = d\theta$. Then, $\mathrm{bar}(d\theta) = x_o$, as seen in Example 5. We exhibit tangent vectors $\tau$, $\tau_1$ at $d\theta$ satisfying $H(\tau,\tau) = 0$, whereas $H(\tau_1,\tau_1) \neq 0$, as follows;*

*(i) Identify $\partial X$ with $S_{x_o}X \cong S^{n-1}$ via $\beta_{x_o}$, and $d\theta$ with $(d\theta)_{x_o}$. Choose on $S^{n-1}$ a function $q = q(\theta) = \theta^i\theta^j$, $i \neq j$ and define $\tau = q(\theta)\,d\theta$ as a measure on $\partial X$. Then, $\tau \in T_{d\theta}\mathcal{P}(\partial X)$. Moreover, $\tau \in T_{d\theta}\mathrm{bar}^{-1}(x_o)$, since $G_{d\theta}(\nu^{d\theta}_{x_o}(u),\tau) = 0$ for any $u \in T_{x_o}X$ and $H(\tau,\tau) = 0$. These are directly from the integral formulae; $\int_{S^{n-1}}\theta^i\theta^j\theta^k(d\theta)_{x_o} = 0$, $\int_{S^{n-1}}(\theta^i\theta^j)^2\theta^k(d\theta)_{x_o} = 0$ for any $k = 1,\ldots,n$. By normalizing $\tau' = \tau/|\tau|_G$ in terms of $G$, from Theorem 16 $\gamma(t) = \exp_{d\theta} t\tau'$ gives a geodesic lying on $\mathrm{bar}^{-1}(x_o)$.*

*(ii) Let $q_1 = q_1(\theta)$ is a function on $S^{n-1}$, $n \geq 3$, defined by $q_1(\theta) = \theta^1\theta^2\theta^3 + \theta^2\theta^3$ and set $\tau_1 = q_1(\theta)\,d\theta$. Then $(\gamma_1)(t) = \exp_{d\theta} t\tau_1'$, $\tau_1' = \tau_1/|\tau_1|_G$, is a geodesic being not completely on the fibre $\mathrm{bar}^{-1}(x_o)$.*

## 5. Barycentrically Associated Maps

Let $\phi$ be an isometry of an Hadamard manifold $(X,g)$. Then, from the average Busemann cocycle formula (36).

**Theorem 18** ([8])**.** *For any isometry $\phi$ of $(X,g)$, we have*

$$\mathrm{bar}(\hat{\phi}_\sharp\mu) = \phi(\mathrm{bar}(\mu)), \quad \mu \in \mathcal{P}(\partial X). \tag{44}$$

**Proof.** Let $y = \mathrm{bar}(\hat{\phi}_\sharp\mu)$. Then $(d\mathbf{B}_{\hat{\phi}_\sharp\mu})_y(u) = 0$ for any $u \in T_yX$, namely, due to (36) $\phi^{-1}y$ turns out to be a critical point of $\mathbf{B}_\mu$, that is, $y = \phi(\mathrm{bar}(\mu))$, so (44) is obtained. □

**Definition 10.** *Let $\Phi : \partial X \to \partial X$ be a homeomorphism of $\partial X$. Then, a bijective map $\phi : X \to X$ is said to be barycentrically associated to $\Phi$, if $\Phi$ and $\phi$ satisfy the relation $\mathrm{bar} \circ \Phi = \phi \circ \mathrm{bar}$, that is, $\mathrm{bar}(\Phi(\mu)) = \phi(\mathrm{bar}(\mu))$ for any $\mu \in \mathcal{P}(\partial X)$.*

Now we are ready to give a proof of Theorem 6.

**Proof of Theorem 6.** From the statement of the theorem, diagram (6) asserts for any $x \in X$, *i.e.*,

$$\Phi_\sharp\left(\nu_x^{\mu_x}(u)\right) = \nu_{\varphi x}^{\mu_{\varphi x}}((\varphi_*)_x(u)), \quad \forall u \in T_xX, \tag{45}$$

namely

$$\Phi_\sharp\left((dB_\theta)_x(u)\mu_x(\theta)\right) = (dB_\theta)_{\varphi x}\left((\varphi_*)_x u\right)\mu_{\varphi x}(\theta) \tag{46}$$

for $\mu_x = \Sigma(x)$, where $\Sigma : X \to \mathcal{P}(\partial X)$ is a cross section whose existence is assumed in the theorem. We write (46) as

$$(dB_{\Phi^{-1}\theta})_x(u)\ \Phi_\sharp\mu_x = (dB_\theta)_{\varphi x}((\varphi_*)_x u)\ \mu_{\varphi x}.$$

Since another diagram (6) implies $\Phi_\sharp\mu_x = \mu_{\varphi x}$, we have

$$(dB_{\Phi^{-1}\theta})_x(u)\ \mu_{\varphi x} = (dB_\theta)_{\varphi x}((\varphi_*)_x u)\ \mu_{\varphi x}, \tag{47}$$

that is,

$$(dB_{\Phi^{-1}\theta})_x(u) = (dB_\theta)_{\varphi x}((\varphi_*)_x u), \quad u \in T_xX,\ \forall\theta \in \partial X,$$

or

$$\langle(\nabla B_{\Phi^{-1}\theta})_x, u\rangle = \langle(\nabla B_\theta)_{\varphi x}, (\varphi_*)_x u\rangle \quad u \in T_xX,\ \forall\theta \in \partial X.$$

Using the formal adjoint $((\varphi_*)_x)^*$ of $(\varphi_*)_x$, we write the above as

$$(\nabla B_{\Phi^{-1}\theta})_x = ((\varphi_*)_x)^*(\nabla B_\theta)_{\varphi x} \quad u \in T_xX,\ \forall\theta \in \partial X.$$

Let $v \in S_{\varphi x}X$ be a unit tangent vector at $\varphi x$ and choose $\theta \in \partial X$ such that $(\nabla B_\theta)_{\varphi x} = v$ so that the above is written as $(\varphi_*)_x^* v = (\nabla B_{\Phi^{-1}\theta})_x$ and thus from property (v) in Section 4.2 it is concluded that $|(\varphi_*)_x^* v| = |(\nabla B_{\Phi^{-1}\theta})_x| = 1$ which implies that $(\varphi_*)_x^*$ and consequently $(\varphi_*)_x$ is a linear isometry. Since $x \in X$ is arbitrary, $\varphi$ turns out to be an isometry of $(X, g)$.

To show that $\Phi$ coincides with $\partial X$-extension $\hat{\varphi}$ we make use of (47) together with the following

$$(dB_{\hat{\varphi}^{-1}\theta})_x(u)\ \mu_{\varphi x} = (dB_\theta)_{\varphi x}((\varphi_*)_x u)\ \mu_{\varphi x} \tag{48}$$

which is derived by differentiating the Busemann cocycle formula (28) to get for any $u \in T_xX$, $x \in X$

$$(dB_{\hat{\varphi}^{-1}\theta})_x(u) = (dB_{\Phi^{-1}\theta})_x(u). \tag{49}$$

Namely, we have $d(B_{\hat{\varphi}^{-1}\theta} - B_{\Phi^{-1}\theta}) = 0$ on $X$ for any $\theta \in \partial X$. Since $X$ is connected, $B_{\hat{\varphi}^{-1}\theta}(x) - B_{\Phi^{-1}\theta}(x) = C$ for a constant $C$ which depends on $\theta$. From this it follows that $\hat{\varphi} = \Phi$. In fact, assume $\hat{\varphi}^{-1}\theta \neq \Phi^{-1}\theta$ for some $\theta \in \partial X$, otherwise, and let $x \to \Phi^{-1}\theta$. Then, from the visibility axiom (see Proposition 5) $B_{\hat{\varphi}^{-1}\theta}(x) - B_{\Phi^{-1}\theta}(x) \to -\infty$ contradicting that $C$ is constant. $\square$

## Author Contributions

Mitsuhiro Itoh conceived the idea and analysis. Mitsuhiro Itoh and Hiroyasu Satoh together did derivation and wrote the paper. Both authors have read and approved the final manuscript.

## Conflicts of Interest

The authors declare no conflict of interest.

## References


1. Itoh, M.; Satoh, H. Fisher Information Geometry of The Barycenter Map. *AIP Conf. Proc.* **2015**, *1641*, http://dx.doi.org/10.1063/1.4905967.
2. Friedrich, T. Die Fisher-Information und symplektische Strukturen. *Math. Nachr.* **1991**, *153*, 273–296.
3. Schoen, R.; Yau, S.-T. *Lectures on Differential Geometry*; Intern. Press: Boston, MA, USA, 1994.
4. Itoh, M.; Satoh, H. Information geometry of Poisson Kernels on Damek-Ricci spaces. *Tokyo J. Math.* **2010**, *33*, 129–144.
5. Itoh, M.; Satoh, H. Fisher Information Geometry, Poisson Kernel and Asymtotical Harmonicity. *Differ. Geom. Appl*. **2011**, *29*, S107–S115.
6. Itoh, M.; Shishido, Y. Fisher Information Metric and Poisson Kernels, *Differ. Geom. Appl*. **2008**, *26*, 347–356.
7. Manning, A.: Topological entropy for geodesic flows, *Ann. Math*., **1979**, *110*, 567-573.
8. Besson, G.; Courtois, G.; Gallot, S. Entropies et Rigidités des espaces localement symétriques de courbure strictement négative. *Geom. Func. Anal*. **1995**, *5*, 731–799.
9. Cafaro, C., Ali, S.A.: Jacobi fields on statistical manifolds of negative curvature, *Physica*, **2007**, *D234*, 70-92.
10. Berndt, J.; Tricerri, F.; Vanhecke, L. *Generalized Heisenberg Groups and Damek-Ricci Harmonic Spaces*; Lecture Notes in Mathematics, Volume 1598; Springer: Heidelberg, Germany, 1991.
11. Itoh, M.; Satoh, H.; Suh, Y.-J. Horospheres and Hyperbolicity of Hadamard manifolds. *Differ. Geom. Appl*. **2014**, *35*, Supplement, 50–68.
12. Szabo, Z. The Lichnerowicz conjecture on harmonic manifolds. *J. Diff. Geom*. **1990**, *31*, 1–28.
13. Heber, J. On harmonic and asymptotically harmonic homogeneous spaces, *Geom. Funct. Anal*. **2006**, *16*, 869–890.
14. Amari, S.; Nagaoka, H. *Methods of Information Geometry*; Translations of Mathematical Monographs, Volume 191; American Mathematical Society: Providence, RI, USA, 2000.
15. Helgason, S. *Differential Geometry and Symmetric Spaces*; Academic Press: New York, NY, USA, 1962.

16. Douady, E.; Earle, C. Conformally natural extension of homeomorphisms of the circle. *Acta Math*. **1986**, *157*, 23–48.
17. Besson, G.; Courtois, G.; Gallot, S. Minimal entropy and Mostow's rigidity theorems. *Erg. Th. Dyn. Sys*. **1996**, *16*, 623–649.
18. Arnaudon, M.; Barbaresco, F.; Yang, L. Medians and Means in Riemannian Geometry, Existence, Uniqueness and Computation. In *Matrix Information Geometry*; Nielson, F., Bhatia, R., Eds.; Springer: Heidelberg, Germany, 2013; pp.169–197.
19. Barbaresco, F.: Information Geometry of Covariance Matrix: Cartan-Siegel Homogeneous Bounded Domains, Mostow/Berger Fibration and Fréchet Median. In *Matrix Information Geometry*; Nielson, F., Bhatia, R., Eds.; Springer: Heidelberg, Germany, 2013; pp.199–255.
20. Furstenberg, H. A Poisson formula for semi-simple Lie groups. *Ann. Math*. **1963**, *77*, 335–386.
21. Eberlein, P.; O'Neill, B. Visibility manifolds. *Pac. J. Math*. **1973**, *46*, 45–110.
22. Do Carmo, M. *Riemannian Geometry*; Birkhäuser: Boston, MA, USA, 1992.
23. Gallot, S.; Hulin, D.; Lafontaine, J. *Riemannian Geometry*, 2nd ed.; Springer-Verlag: Berlin, Germany, 1980.
24. Itoh, M.; Satoh, H. Information Geometry of Barycenter Map. In *Real and Complex Submanifolds*; Suh, Y.J., Berndt, J., Ohnita, Y., Kim, B.H., Eds.; Springer Proceedings in Mathematics and Statistics, Volume 106; Springer: Tokyo, Japan, 2014; pp. 79–88.
25. Itoh, M.; Satoh, H. Information Geometry of Busemann-Barycenter for Probability Measures. *Int. J. Math.* **2015**, in press.
26. Halmos, P. *Measure Theory*; Graduate Texts in Mathematics, Volume 18; Springer: New York, NY, USA, 1950.
27. Villani, C. *Topics in Optimal Transformation*; Gradudate Study in Mathematics, Volume 58; American Mathematical Society: Providence, RI, USA, 2003.
28. Fathi, A. Structure of the group of homeomorphisms preserving a good measure on a compact manifold. *Ann. scient. Éc. Norm. Sup* **1980**, *13*, 45–93.
29. Oxtoby, C.; Ulam, S.M. Measure preserving homeomorphisms and metrical transitivity. *Ann. Math*. **1941**, *42*, 874–920.
30. Sakai, T. *Riemannian Geometry*; American Mathematical Society: Providence, RI, USA, 2000.
31. Ballmann, W.; Gromov, M.; Schroeder, V. *Manifolds of Nonpositive Curvature*; Progress in Mathematics, Volume 61; Birkhäuser: Boston, MA, USA, 1985.
32. Heintze, E.; Im Hof, H.-C. Geometry of horospheres. *J. Diff. Geom*. **1977**, *12*, 481–491.
33. Guivarc'h, Y.; Ji, L.; Taylor, J.C. *Compactifications of Symmetric Spaces*; Progress in Mathematics, Volume 156; Birkhäuser: Boston, MA, USA, 1997.
34. Ledrappier, F. Harmonic measures and Bowen-Margulis measures. *Israel J. Math*. **1990**, *71*, 275–287.
35. Eberlein, P. Geodesic Flows in Manifolds of Nonpositive Curvature. *Proc. Symp. Pure Math.* **2001**, *69*, 525–571.

# Chapter 3:
# Applications of Information/Entropy Geometric Structures

# Entropy, Information Theory, Information Geometry and Bayesian Inference in Data, Signal and Image Processing and Inverse Problems

**Ali Mohammad-Djafari**

**Abstract:** The main content of this review article is first to review the main inference tools using Bayes rule, the maximum entropy principle (MEP), information theory, relative entropy and the Kullback–Leibler (KL) divergence, Fisher information and its corresponding geometries. For each of these tools, the precise context of their use is described. The second part of the paper is focused on the ways these tools have been used in data, signal and image processing and in the inverse problems, which arise in different physical sciences and engineering applications. A few examples of the applications are described: entropy in independent components analysis (ICA) and in blind source separation, Fisher information in data model selection, different maximum entropy-based methods in time series spectral estimation and in linear inverse problems and, finally, the Bayesian inference for general inverse problems. Some original materials concerning the approximate Bayesian computation (ABC) and, in particular, the variational Bayesian approximation (VBA) methods are also presented. VBA is used for proposing an alternative Bayesian computational tool to the classical Markov chain Monte Carlo (MCMC) methods. We will also see that VBA englobes joint maximum *a posteriori* (MAP), as well as the different expectation-maximization (EM) algorithms as particular cases.

Reprinted from *Entropy*. Cite as: Mohammad-Djafari, A. Entropy, Information Theory, Information Geometry and Bayesian Inference in Data, Signal and Image Processing and Inverse Problems. *Entropy* **2015**, *17*, 3989–4027.

## 1. Introduction

As this paper is an overview and an extension of my tutorial paper in MaxEnt 2014 workshop [1], this Introduction gives a summary of the content of this paper.

The qualification Bayesian refers to the influence of Thomas Bayes [2], who introduced what is now known as Bayes' rule, even if the idea had been developed independently by Pierre-Simon de Laplace [3]. For this reason, I am asking a question of the community if we shall change Bayes to Laplace and Bayesian to Laplacian or at least mention them both. Whatever the answer, we assume that the reader knows what probability means in a Bayesian or Laplacian framework. The main idea is that a probability law $P(X)$ assigned to a quantity $X$ represents our state of knowledge that we have about it. If $X$ is a discrete valued variable, $\{P(X = x_n) = p_n, n = 1, \cdots, N\}$ with mutually exclusive values $x_n$ is its probability distribution. When $X$ is a continuous valued variable, $p(x)$ is its probability density function from which we can compute $P(a \leq X < b) = \int_a^b p(x)\,\mathrm{d}x$ or any other probabilistic quantity, such as its mode, mean, median, region of high probabilities, *etc*.

In science, it happens very often that a quantity cannot be directly observed or, even when it can be observed, the observations are uncertain (commonly said to be noisy), by uncertain or noisy, here, I mean that, if we repeat the experiences with the same practical conditions, we obtain different data. However, in the Bayesian approach, for a given experiment, we have to use the data as they are, and we want to infer it from those observations. Before starting the observation and gathering new data, we have very incomplete knowledge about it. However, this incomplete knowledge can be translated in probability theory via an *a priori* probability law. We will discuss this point later on regarding how to do this. For now, we assume that this can be done. When a new observation (data $D$) on $X$ becomes available (direct or indirect), we gain some knowledge via the likelihood $P(D|X)$. Then, our state of knowledge is updated combining $P(D|X)$ and $P(X)$ to obtain an *a posteriori* law $P(X|D)$, which represents the new state of knowledge on $X$. This is the main esprit of the Bayes rule, which can be summarized as:

$$P(X|D) = P(D|X)P(X)/P(D). \tag{1}$$

As $P(X|D)$ has to be a probability law, we have:

$$P(D) = \sum_X P(D|X)P(X). \tag{2}$$

This relation can be extended to the continuous case. Some more details will be given in Section 2.

Associated with a probability law is the quantity of information it contains. Shannon [4] introduced the notion of quantity of information $I_n$ associated with one of the possible values of $x_n$ of $X$ with probabilities $P(X = x_n) = p_n$ to be $I_n = \ln \frac{1}{p_n} = -\ln p_n$ and the entropy $H$ as the expected value of $I_n$:

$$H = -\sum_{n=1}^{N} p_n \ln p_n. \tag{3}$$

The word entropy has also its roots in thermodynamics and physics. However, this notion of entropy has no direct link with entropy in physics, even if for a particular physical system, we may attribute a probability law to a quantity of interest of that system and then define its entropy. This information definition of Shannon entropy became the main basis of information theory in many data analyses and the science of communication. More details and extensions about this subject will be given in Section 3.

As we can see up to now, we did not yet discuss how to assign a probability law to a quantity. For the discrete value variable, when $X$ can take one of the $N$ values $\{x_1, \cdots, x_N\}$ and when we do not know anything else about it, Laplace proposed the "*Principe d'indifférence*", where $P(X = x_n) = p_n = \frac{1}{N}, \forall n = 1, \cdots, N$, a uniform distribution. However, what if we know more, but not enough to be able to assign the probability law $\{p_1, \cdots, p_N\}$ completely?

For example, if we know that the expected value is $\sum_n x_n p_n = d$, this problem can be handled by considering this equation as a constraint on the probability distribution $\{p_1, \cdots, p_N\}$. If we have a sufficient number of constraints (at least $N$), then we may obtain a unique solution. However, very often, this is not the case. The question now is how to assign a probability distribution $\{p_1, \cdots, p_N\}$

that satisfies the available constraints. This question is an ill-posed problem in the mathematical sense of Hadamard [5] in the sense that the solution is not unique. We can propose many probability distributions that satisfy the constraint imposed by this problem. To answer this question, Jaynes [6–8] introduced the maximum entropy principle (MEP) as a tool for assigning a probability law to a quantity on which we have some incomplete or macroscopic (expected values) information. Some more details about this MEP, the mathematical optimization problem, the expression of the solution and the algorithm to compute it will be given in Sections 3 and 4.

Kullback [9] was interested in comparing two probability laws and introduced a tool to measure the increase of information content of a new probability law with respect to a reference one. This tool is called either the Kullback–Leibler (KL) divergence, cross entropy or relative entropy. It has also been used to update a prior law when new pieces of information in the form of expected values are given. As we will see later, this tool can also be used as an extension to the MEP of Jaynes. Furthermore, as we will see later, this criterion of comparison of two probability laws is not symmetric: one of the probability laws has to be chosen to be the reference, and then, the second is compared to this reference. Some more details and extensions will be given in Section 5.

Fisher [10] wanted to measure the amount of information that a random variable $X$ carries about an unknown parameter $\theta$ upon which its probability law $p(x|\theta)$ depends. The partial derivative with respect to $\theta$ of the logarithm of this probability law, called the log-likelihood function for $\theta$, is called the score. He showed that the first order moment of the score is zero, but its second order moment is positive and is also equivalent to the expected values of the second derivative of log-likelihood function with respect to $\theta$. This quantity is called the Fisher information. It is also been shown that for the small variations of $\theta$, the Fisher information induces locally a distance in the space of parameters $\Theta$, if we had to compare two very close values of $\theta$. In this way, the notion of the geometry of information is introduced [11,12]. We must be careful here that this geometrical property is related to the space of the parameters $\Theta$ for small changes of the parameter for a given family of parametric probability law $p(x|\theta)$ and not in the space of probabilities. However, for two probability laws $p_1(x) = p(x|\theta_1)$ and $p_2(x) = p(x|\theta_2)$ in the same exponential family, the Kullback–Leibler divergence KL $[p_1 : p_2]$ induces a Bregman divergence B$[\theta_1 : \theta_2]$ between the two parameters [13,14]. More details will be given in Section 8.

At this stage, we have almost introduced all of the necessary tools that we can use for different levels of data, signal and image processing. In the following, we give some more details for each of these tools and their inter-relations. Then, we review a few examples of their use in different applications. As examples, we demonstrate how these tools can be used in independent components analysis (ICA) and source separation, data model selection, in spectral analysis of the signals and in inverse problems, which arise in many sciences and engineering applications. At the end, we focus more on the Bayesian approach for inverse problems. We present some details concerning unsupervised methods, where the hyper parameters of the problem have to be estimated jointly with the unknown quantities (hidden variables). Here, we will see how the Kullback–Leibler divergence can help approximate Bayesian computation (ABC). In particular, some original materials concerning variational Bayesian approximation (VBA) methods are presented.

## 2. Bayes Rule

Let us introduce things very simply. If we have two discrete valued related variables $X$ and $Y$, for which we have assigned probability laws $P(X)$ and $P(Y)$, respectively, and their joint probability law $P(X,Y)$, then from the sum and product rule, we have:

$$P(X,Y) = P(X|Y)P(Y) = P(Y|X)P(X) \tag{4}$$

where $P(X,Y)$ is the joint probability law, $P(X) = \sum_Y P(X,Y)$ and $P(Y) = \sum_X P(X,Y)$ are the marginals and $P(X|Y) = P(X,Y)/P(Y)$ and $P(Y|X) = P(X,Y)/P(X)$ are the conditionals. Now, consider the situation where $Y$ can be observed, but not $X$. Because these two quantities are related, we may want to infer $X$ from the observations on $Y$. Then, we can use:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \tag{5}$$

which is called the Bayes rule.

This relation is extended to the continuous valued variables using the measure theory [15,16]:

$$p(x|y) = \frac{p(y|x)\,p(x)}{p(y)} \tag{6}$$

with:

$$p(y) = \int p(y|x)\,p(x)\,\mathrm{d}x. \tag{7}$$

More simply, the Bayes rule is often written as:

$$p(x|y) \propto p(y|x)\,p(x). \tag{8}$$

This writing can be used when we want to use $p(x|y)$ to compute quantities that are only dependent on the shape of $p(x|y)$, such as the mode, the median or quantiles. However, we must be careful that the denominator is of importance in many other cases, for example when we want to compute expected values. There is no need for more sophisticated mathematics here if we want to use this approach.

As we mentioned, the main use of this rule is in particular when $X$ can not be observed (unknown quantity), but $Y$ is observed and we want to infer $X$. In this case, the term $p(y|x)$ is called the likelihood (of unknown quantity $X$ in the observed data $y$), $p(x)$ is called *a priori* and $p(x|y)$ *a posteriori*. The likelihood is assigned using the link between the observed $Y$ and the unknown $X$, and $p(x)$ is assigned using the prior knowledge about it. The Bayes rule then is a way to do state of knowledge fusion. Before taking into account any observation, our state of knowledge is represented by $p(x)$, and after the observation of $Y$, it becomes $p(x|y)$.

However, in this approach, two steps are very important. The first step is the assigning of $p(x)$ and $p(y|x)$ before being able to use the Bayes rule. As noted in the Introduction and as we will see later, we need other tools for this step. The second important step is after: how to use $p(x|y)$ to summarize

it. When $X$ is just a scalar value variable, we can do this computation easily. For example, we can compute the probability that $X$ is in the interval $[a, b]$ via:

$$P(a \leq X < b|y) = \int_a^b p(x|y)\,\mathrm{d}x. \tag{9}$$

However, when the unknown becomes a high dimensional vectorial variable $\boldsymbol{X}$, as is the case in many signal and image processing applications, this computation becomes very costly [17]. We may then want to summarize $p(x|y)$ by a few interesting or significant point estimates. For example, compute the maximum *a posteriori* (MAP) solution:

$$\widehat{x}_{\mathrm{MAP}} = \arg\max_x \left\{p(x|y)\right\}, \tag{10}$$

the expected *a posteriori* (EAP) solution:

$$\widehat{x}_{\mathrm{EAP}} = \int x\, p(x|y)\,\mathrm{d}x, \tag{11}$$

the domains of $X$ which include an integrated probability mass of more than some desired value (0.95 for example):

$$[x_1, x_2] : \int_{x_1}^{x_2} p(x|y)\,\mathrm{d}x = .95, \tag{12}$$

or any other questions, such as median or any $\alpha$-quantiles:

$$x_q : \int_{-\infty}^{x_q} p(x|y)\,\mathrm{d}x = (1-\alpha). \tag{13}$$

As we see, computation of MAP needs an optimization algorithm, while these last three cases need integration, which may become very complicated for high dimensional cases [17].

We can also just explore numerically the whole space of the distribution using the Markov chain Monte Carlo (MCMC) [18–26] or any other sampling techniques [17]. In the scalar case (one dimension), all of these computations can be done numerically very easily. For the vectorial case, when the dimensions become large, we need to develop specialized approximation methods, such as VBA and algorithms to do these computations. We give some more details about these when using this approach for inverse problems in real applications.

**Remarks** on notation used for the expected value in this paper: For a variable $X$ with the probability density function (pdf) $p(x)$ and any regular function $h(X)$, we use indifferently:

$$\mathrm{E}\left\{X\right\} = \mathrm{E}_p\left\{X\right\} =< X >=< X >_p = \int x\, p(x)\,\mathrm{d}x$$

and:

$$\mathrm{E}\left\{h(X)\right\} = \mathrm{E}_p\left\{h(X)\right\} =< h(X) >=< h(X) >_p = \int h(x)\, p(x)\,\mathrm{d}x.$$

As an example, as we will say later, the entropy of $p(x)$ is noted indifferently:

$$H[p] = \mathrm{E}\left\{-\ln(p(X))\right\} = \mathrm{E}_p\left\{-\ln p(X)\right\} =< -\ln p(X) >=< -\ln p(X) >_p = -\int p(x)\,\ln p(x)\,\mathrm{d}x.$$

For any conditional probability density function (pdf) $p(x|y)$ and any regular function $h(X)$, we use indifferently:

$$\mathrm{E}\left\{X|y\right\} = \mathrm{E}_{p(x|y)}\left\{X\right\} =< X|y >=< X >_{p(x|y)}= \int x\, p(x|y)\, \mathrm{d}x$$

and:

$$\mathrm{E}\left\{h(X)|y\right\} = \mathrm{E}_{p(x|y)}\left\{h(X)\right\} =< h(X)|y >=< h(X) >_{p(x|y)}= \int h(x)\, p(x|)\, \mathrm{d}x.$$

As another example, as we will see later, the relative entropy of $p(x)$ over $q(x)$ is noted indifferently:

$$D[p|q] = \mathrm{E}_p\left\{-\ln\frac{p(X)}{q(X)}\right\} =< -\ln\frac{p(X)}{q(X)} >_{p(x)}= -\int p(x)\ \ln\frac{p(x)}{q(x)}\,\mathrm{d}x$$

and when there is not any ambiguity in the integration variable, we may omit it. For example, we may note:

$$D[p|q] = \mathrm{E}_p\left\{-\ln\frac{p}{q}\right\} =< -\ln\frac{p}{q} >_p= -\int p\ \ln\frac{p}{q}.$$

Finally, when we have two variables $X$ and $Y$ with their joint pdf $p(x,y)$, their marginals $p(x)$ and $p(y)$ and their conditionals $p(x|y)$ and $p(y|x)$, we may use the following notations:

$$\mathrm{E}\left\{h(X)|y\right\} = \mathrm{E}_{p(x|y)}\left\{h(X)\right\} = \mathrm{E}_{X|Y}\left\{h(X)\right\} =< h(X)|y >=< h(X) >_{p(x|y)}= \int h(x)\, p(x|y)\, \mathrm{d}x.$$

## 3. Quantity of Information and Entropy

### *3.1. Shannon Entropy*

To introduce the quantity of information and the entropy, Shannon first considered a discrete valued variable $X$ taking values $\{x_1, \cdots, x_N\}$ with probabilities $\{p_1, \cdots, p_N\}$ and defined the quantities of information associated with each of them as $I_n = \ln\frac{1}{p_n} = -\ln p_n$ and its expected value as the entropy:

$$\mathrm{H}\left[X\right] = -\sum_{i=1}^{N} p_i\ \ln p_i. \tag{14}$$

Later, this definition is extended to the continuous case by:

$$\mathrm{H}\left[X\right] = -\int p(x)\ \ln p(x)\,\mathrm{d}x. \tag{15}$$

By extension, if we consider two related variables $(X, Y)$ with the probability laws, joint $p(x,y)$, marginals, $p(x)$, $p(y)$, and conditionals, $p(y|x)$, $p(x|y)$, we can define, respectively, the joint entropy:

$$\mathrm{H}\left[X,Y\right] = -\iint p(x,y)\ \ln p(x,y)\,\mathrm{d}x\,\mathrm{d}y, \tag{16}$$

as well as $\mathrm{H}\left[X\right]$, $\mathrm{H}\left[Y\right]$, $\mathrm{H}\left[Y|X\right]$ and $\mathrm{H}\left[X|Y\right]$.

Therefore, for any well-defined probability law, we can have an expression for its entropy. $\mathrm{H}\left[X\right]$, $\mathrm{H}\left[Y\right]$, $\mathrm{H}\left[Y|X\right]$, $\mathrm{H}\left[X|Y\right]$ and $\mathrm{H}\left[X,Y\right]$, which should better be noted as $\mathrm{H}\left[p(x)\right]$, $\mathrm{H}\left[p(y)\right]$, $\mathrm{H}\left[p(y|x)\right]$, $\mathrm{H}\left[p(x|y)\right]$ and $\mathrm{H}\left[p(x,y)\right]$.

*3.2. Thermodynamical Entropy*

Entropy is also a property of thermodynamical systems introduced by Clausius [27]. For a closed homogeneous system with reversible transformation, the differential entropy $\delta S$ is related to $\delta Q$ the incremental reversible transfer of heat energy into that system by $\delta S = \delta Q/T$ with $T$ being the uniform temperature of the closed system.

It is very hard to establish a direct link between these two entropies. However, in statistical mechanics, thanks to Boltzmann, Gibbs and many others, we can establish some link if we consider the microstates (for example, the number, positions and speeds of the particles) and the macrostates (for example, the temperature $T$, pressure $P$, volume $V$ and energy $E$) of the system and if we assign a probability law to microstates and consider the macrostates as the average (expected values) of some functions of those microstates. Let us give a very brief summary of some of those interpretations.

*3.3. Statistical Mechanics Entropy*

The interpretation of entropy in statistical mechanics is the measure of uncertainty that remains about the state of a system after its observable macroscopic properties, such as temperature ($T$), pressure ($P$) and volume ($V$), have been taken into account. For a given set of macroscopic variables $T$, $P$ and $V$, the entropy measures the degree to which the probability of the system is spread out over different possible microstates. In contrast to the macrostate, which characterizes plainly observable average quantities, a microstate specifies all atomic details about the system, including the position and velocity of every atom. Entropy in statistical mechanics is a measure of the number of ways in which the microstates of the system may be arranged, often taken to be a measure of "disorder" (the higher the entropy, the higher the disorder). This definition describes the entropy as being proportional to the natural logarithm of the number of possible microscopic configurations of the system (microstates), which could give rise to the observed macroscopic state (macrostate) of the system. The proportionality constant is the Boltzmann constant.

*3.4. Boltzmann Entropy*

Boltzmann described the entropy as a measure of the number of possible microscopic configurations $\Omega$ of the individual atoms and molecules of the system (microstates) that comply with the macroscopic state (macrostate) of the system. Boltzmann then went on to show that $k \ln \Omega$ was equal to the thermodynamic entropy. The factor $k$ has since been known as Boltzmann's constant.

In particular, Boltzmann showed that the entropy $S$ of an ideal gas is related to the number of states of the molecules (microstates $\Omega$) with a given temperature (macrostate):

$$S = k \ln \Omega \tag{17}$$

*3.5. Gibbs Entropy*

The macroscopic state of the system is defined by a distribution on the microstates that are accessible to a system in the course of its thermal fluctuations. Therefore, the entropy is defined over two different levels of description of the given system. The entropy is given by the Gibbs entropy formula, named after J. Willard Gibbs. For a classical system (*i.e.*, a collection of classical particles) with a discrete set of microstates, if $E_i$ is the energy of microstate $i$ and $p_i$ is its probability that it occurs during the system's fluctuations, then the entropy of the system is:

$$S = -k \sum_{i=1}^{N} p_i \, \ln p_i. \tag{18}$$

where $k$ is again the physical constant of Boltzmann, which, like the entropy, has units of heat capacity. The logarithm is dimensionless. It is interesting to note that Relation (17) can be obtained from Relation (18) when the probability distribution is uniform over the volume $\Omega$ [28–30].

## 4. Relative Entropy or Kullback–Leibler Divergence

Kullback wanted to compare the relative quantity of information between two probability laws $p_1$ and $p_2$ on the same variable $X$. Two related notions have been defined:

- Relative Entropy of $p_1$ with respect to $p_2$:

$$D\left[p_1 : p_2\right] = -\int p_1(x) \, \ln \frac{p_1(x)}{p_2(x)} \, \mathrm{d}x \tag{19}$$

- Kullback–Leibler divergence of $p_1$ with respect to $p_2$:

$$\mathrm{KL}\left[p_1 : p_2\right] = -D\left[p_1 : p_2\right] = \int p_1(x) \, \ln \frac{p_1(x)}{p_2(x)} \, \mathrm{d}x \tag{20}$$

We may note that:

- $\mathrm{KL}\left[q : p\right] \geq 0$,
- $\mathrm{KL}\left[q : p\right] = 0$, if $q = p$ and
- $\mathrm{KL}\left[q : p_0\right] \geq \mathrm{KL}\left[q : p_1\right] + \mathrm{KL}\left[p_1 : p_0\right]$.
- $\mathrm{KL}\left[q : p\right]$ is invariant with respect to a scale change, but is not symmetric.
- A symmetric quantity can be defined as:

$$\mathrm{J}\left[q, p\right] = \frac{1}{2} \left(\mathrm{KL}\left[q : p\right] + \mathrm{KL}\left[p : q\right]\right). \tag{21}$$

## 5. Mutual Information

The purpose of mutual information is to compare two related variables $Y$ and $X$. It can be defined as the expected amount of information that one gains about Xif we observe the value of Y, and *vice versa*. Mathematically, the mutual information between $X$ and $Y$ is defined as:

$$\mathrm{I}\left[Y,X\right]=\mathrm{H}\left[X\right]-\mathrm{H}\left[X|Y\right]=\mathrm{H}\left[Y\right]-\mathrm{H}\left[Y|X\right] \tag{22}$$

or equivalently as:

$$\mathrm{I}\left[Y,X\right]=D\left[p(X,Y):p(X)p(Y)\right]. \tag{23}$$

With this definition, we have the following properties:

$$\mathrm{H}\left[X,Y\right]=\mathrm{H}\left[X\right]+\mathrm{H}\left[Y|X\right]=\mathrm{H}\left[Y\right]+\mathrm{H}\left[X|Y\right]=\mathrm{H}\left[X\right]+\mathrm{H}\left[Y\right]-\mathrm{I}\left[Y,X\right] \tag{24}$$

and:

$$\begin{aligned}\mathrm{I}\left[Y,X\right] &= \mathrm{E}_X\left\{D\left[p(Y|X):p(Y)\right]\right\} \stackrel{\triangle}{=} \int D\left[p(y|x):p(y)\right]p(x)\,\mathrm{d}x \\ &= \mathrm{E}_Y\left\{D\left[p(X|Y):p(X)\right]\right\} \stackrel{\triangle}{=} \int D\left[p(x|y):p(x)\right]p(y)\,\mathrm{d}y.\end{aligned} \tag{25}$$

We may also remark on the following property:

- $\mathrm{I}\left[Y,X\right]$ is a concave function of $p(y)$ when $p(x|y)$ is fixed and a convex function of $p(x|y)$ when $p(y)$ is fixed.
- $\mathrm{I}\left[Y,X\right]\geq 0$ with equality only if $X$ and $Y$ are independent.

## 6. Maximum Entropy Principle

The first step before applying any probability rules for inference is to assign a probability law to a quantity. Very often, the available knowledge on that quantity can be described mathematically as the constraints on the desired probability law. However, in general, those constraints are not enough to determine in a unique way that probability law. There may exist many solutions that satisfy those constraints. We need then a tool to select one.

Jaynes introduced the MEP [8], which can be summarized as follows: When we do not have enough constraints to determine a probability law that satisfies those constraints, we may select between them the one with maximum entropy.

Let us be now more precise. Let us assume that the available information on that quantity $X$ is the form of:

$$\mathrm{E}\left\{\phi_k(X)\right\}=d_k,\quad k=1,\ldots,K. \tag{26}$$

where $\phi_k$ are any known functions. First, we assume that such probability laws exist by defining:

$$\mathcal{P}=\left\{p(x):\int\phi_k(x)p(x)\,\mathrm{d}x=d_k,\quad k=0,\ldots,K\right\}$$

with $\phi_0 = 1$ and $d_0 = 1$ for the normalization purpose. Then, the MEP is written as an optimization problem:

$$p_{ME}(x) = \arg\max_{p \in \mathcal{P}} \left\{ \mathrm{H}\,[p] = -\int p(x)\, \ln p(x)\, \mathrm{d}x \right\} \tag{27}$$

whose solution is given by:

$$p_{ME}(x) = \frac{1}{Z(\boldsymbol{\lambda})} \exp\left[ -\sum_{k=1}^{K} \lambda_k \phi_k(x) \right] \tag{28}$$

where $Z(\boldsymbol{\lambda})$, called the partition function, is given by: $Z(\boldsymbol{\lambda}) = \int \exp[-\sum_{k=1}^{K} \lambda_k \phi_k(x)]\, \mathrm{d}x$ and $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_K]'$ have to satisfy:

$$-\frac{\partial \ln Z(\boldsymbol{\lambda})}{\partial \lambda_k} = d_k, \quad k = 1, \ldots, K \tag{29}$$

which can also be written as $-\nabla_{\boldsymbol{\lambda}} \ln Z(\boldsymbol{\lambda}) = \boldsymbol{d}$. Different algorithms have been proposed to compute numerically the ME distributions. See, for example, [31–37]

The maximum value of entropy reached is given by:

$$H_{\max} = \ln Z(\boldsymbol{\lambda}) + \boldsymbol{\lambda}' \boldsymbol{d}. \tag{30}$$

This optimization can easily be extended to the use of relative entropy by replacing $H(p)$ by $D[p : q]$ where $q(x)$ is a given reference of *a priori* law. See [9,38,39] and [34,40–42] for more details.

## 7. Link between Entropy and Likelihood

Consider the problem of the parameter estimation $\boldsymbol{\theta}$ of a probability law $p(x|\boldsymbol{\theta})$ from an $n$-element sample of data $\boldsymbol{x} = \{x_1, \cdots, x_n\}$.

The log-likelihood of $\boldsymbol{\theta}$ is defined as:

$$L(\boldsymbol{\theta}) = \ln \prod_{i=1}^{n} p(x_i|\theta) = \sum_{i=1}^{n} \ln p(x_i|\theta). \tag{31}$$

Maximizing $L(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ gives what is called the maximum likelihood (ML) estimate of $\theta$.

Noting that $L(\boldsymbol{\theta})$ depends on $n$, we may consider $\frac{1}{n} L(\boldsymbol{\theta})$ and define:

$$\bar{L}(\boldsymbol{\theta}) = \lim_{n \mapsto \infty} \frac{1}{n} L(\boldsymbol{\theta}) = \mathrm{E}\left\{ \ln p(x|\boldsymbol{\theta}) \right\} = \int p(x|\boldsymbol{\theta}^*)\, \ln p(x|\boldsymbol{\theta})\, \mathrm{d}x, \tag{32}$$

where $\boldsymbol{\theta}^*$ is the right answer and $p(x|\boldsymbol{\theta}^*)$ its corresponding probability law. We may then remark that:

$$D\left[p(x|\boldsymbol{\theta}^*) : p(x|\boldsymbol{\theta})\right] = -\int p(x|\boldsymbol{\theta}^*)\, \ln \frac{p(x|\boldsymbol{\theta})}{p(x|\boldsymbol{\theta}^*)}\, \mathrm{d}x = -\int p(x|\boldsymbol{\theta}^*)\, \ln p(x|\boldsymbol{\theta}^*)\, \mathrm{d}x + \bar{L}(\boldsymbol{\theta}). \tag{33}$$

The first term in the right-hand side being a constant, we derive that:

$$\arg\max_{\boldsymbol{\theta}} \left\{ D\left[ p(x|\boldsymbol{\theta}^*) : p(x|\boldsymbol{\theta}) \right] \right\} = \arg\max_{\boldsymbol{\theta}} \left\{ \bar{L}(\boldsymbol{\theta}) \right\}.$$

In this way, there is a link between the maximum likelihood and maximum relative entropy solutions [24].

There is also a link between the maximum relative entropy and the Bayes rule. See, for example, [43,44] and their corresponding references.

## 8. Fisher Information, Bregman and Other Divergences

Fisher [10] was interested in measuring the amount of information that samples of a variable $X$ carries about an unknown parameter $\theta$ upon which its probability law $p(x|\theta)$ depends. For a given sample of observation $x$ and its probability law $p(x|\theta)$, the function $\mathcal{L}(\theta) = p(x|\theta)$ is called the likelihood of $\theta$ in the sample $x$. He called the score of $x$ over $\theta$ the partial derivative with respect to $\theta$ of the logarithm of this function:

$$S(x|\theta) = \frac{\partial \ln p(x|\theta)}{\partial \theta} \tag{34}$$

He also showed that the first order moment of the score is zero:

$$\mathrm{E}\left\{ S(X|\theta) \right\} = \mathrm{E}\left\{ \frac{\partial \ln p(x|\theta)}{\partial \theta} \right\} = 0 \tag{35}$$

but its second order moment is positive and is also equivalent to the expected values of the second derivative of the log-likelihood function with respect to $\theta$.

$$\mathrm{E}\left\{ S^2(X|\theta) \right\} = \mathrm{E}\left\{ \left| \frac{\partial \ln p(x|\theta)}{\partial \theta} \right|^2 \right\} = \mathrm{E}\left\{ \frac{\partial^2 \ln p(x|\theta)}{\partial \theta^2} \right\} = F \tag{36}$$

This quantity is called the Fisher information [14].

It is also shown that for the small variations of $\theta$, the Fisher information induces locally a distance in the space of parameters $\Theta$, if we had to compare two very close values of $\theta$. In this way, the notion of the geometry of information is introduced. The main steps for introducing this notion are the following: Consider $D\left[ p(x|\boldsymbol{\theta}^*) : p(x|\boldsymbol{\theta}^* + \Delta\boldsymbol{\theta}) \right]$ and assume that $\ln p(x|\boldsymbol{\theta})$ can be developed in a Taylor series. Then, keeping the terms up to the second order, we obtain:

$$D\left[ p(x|\boldsymbol{\theta}^*) : p(x|\boldsymbol{\theta}^* + \Delta\boldsymbol{\theta}) \right] \simeq \frac{1}{2} \Delta\boldsymbol{\theta}' \boldsymbol{F}(\theta^*) \Delta\boldsymbol{\theta}. \tag{37}$$

where $\boldsymbol{F}$ is the Fisher information:

$$\boldsymbol{F}(\theta^*) = \mathrm{E}\left\{ \frac{\partial^2 \ln p(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}' \partial \boldsymbol{\theta}} |_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \right\}. \tag{38}$$

We must be careful here that this geometry property is related to the space of the parameters $\Theta$ for a given family of parametric probability law $p(x|\theta)$ and not in the space of probabilities. However,

for two probability laws $p_1(x) = p(x|\theta_1)$ and $p_2(x) = p(x|\theta_2)$ in the same exponential family, the Kullback–Leibler divergence $\mathrm{KL}\,[p_1 : p_2]$ induces a Bregman divergence $\mathrm{B}[\theta_1|\theta_2]$ between the two parameters [14,45–48].

To go further into detail, let us extend the discussion about the link between Fisher information and KL divergence, as well as other divergences, such as $f$-divergences, Rényi's divergences and Bregman divergences.

- $f$-divergences:

  The $f$-divergences, which are a general class of divergences, indexed by convex functions $f$, that include the KL divergence as a special case. Let $f : (0, \infty) \mapsto \mathbf{R}$ be a convex function for which $f(1) = 0$. The $f$-divergence between two probability measures $P$ and $Q$ is defined by:

  $$D_f[P : Q] = \int q\, f\left(\frac{p}{q}\right) \tag{39}$$

  Every $f$-divergence can be viewed as a measure of distance between probability measures with different properties. Some important special cases are:

  - $f(x) = x \ln x$ gives KL divergence: $\mathrm{KL}\,[P : Q] = \int p\, \ln(\frac{p}{q})$.
  - $f(x) = |x - 1|/2$ gives total variation distance: $\mathrm{TV}[P, Q] = \int |p - q|/2$.
  - $f(x) = (\sqrt{x} - 1)^2$ gives the square of the Hellinger distance: $H^2[P, Q] = \int (\sqrt{p} - \sqrt{q})^2$.
  - $f(x) = (x - 1)^2$ gives the chi-squared divergence: $\chi^2[P : Q] = \int \frac{(p-q)^2}{q}$.

- Rényi divergences:

  These are another generalization of the KL divergence. The Rényi divergence between two probability distributions $P$ and $Q$ is:

  $$D_\alpha[P : Q] = \frac{1}{\alpha - 1} \ln \int p^\alpha\, q^{1-\alpha}. \tag{40}$$

  When $\alpha = 1$, by a continuity argument, $D_\alpha[P : Q]$ converges to $\mathrm{KL}\,[P : Q]$.

  $D_{1/2}[P, Q] = -2 \ln \int \sqrt{p\, q}$ is called Bhattacharyya divergence (closely related to Hellinger distance). Interestingly, this quantity is always smaller than KL:

  $$D_{1/2}[P : Q] \leq \mathrm{KL}\,[P : Q]\,. \tag{41}$$

  As a result, it is sometimes easier to derive risk bounds with $D_{1/2}$ as the loss function as opposed to KL.

- Bregman divergences:

  The Bregman divergences provide another class of divergences that are indexed by convex functions and include both the Euclidean distance and the KL divergence as special cases. Let $\phi$ be a differentiable strictly convex function. The Bregman divergence $B_\phi$ is defined by:

  $$B_\phi[\boldsymbol{x} : \boldsymbol{y}] = \phi(\boldsymbol{x}) - \phi(\boldsymbol{y}) - \langle \boldsymbol{x} - \boldsymbol{y}, \nabla\phi(\boldsymbol{y}) \rangle \tag{42}$$

where $< \boldsymbol{x}, \boldsymbol{y} >= \sum_j x_i\, y_j$ here means the scalar product of $\boldsymbol{x}$ and $\boldsymbol{y}$ and where the domain of $\phi$ is a space where convexity and differentiability make sense (e.g., whole or a subset of $\mathbf{R}^d$ or an $L_p$ space). For example, $\phi(\boldsymbol{x}) = \|\boldsymbol{x}\|^2$ on $\mathbf{R}^d$ gives the Euclidean distance:

$$B_\phi[\boldsymbol{x} : \boldsymbol{y}] = \phi(\boldsymbol{x}) - \phi(\boldsymbol{y}) - \langle \boldsymbol{x} - \boldsymbol{y}, \nabla\phi(\boldsymbol{y}) \rangle = \|\boldsymbol{x}\|^2 - \|\boldsymbol{y}\|^2 - \langle \boldsymbol{x} - \boldsymbol{y}, 2\boldsymbol{y} \rangle = \|\boldsymbol{x} - \boldsymbol{y}\|^2 \quad (43)$$

and $\phi(\boldsymbol{x}) = \sum_j x_j \ln x_j$ on the simplex in $\mathbf{R}^d$ gives the KL divergence:

$$B_\phi[\boldsymbol{x} : \boldsymbol{y}] = \sum_j x_j \ln x_j - \sum_j y_j \ln y_j - \sum_j (x_j - y_j)(1 + \ln y_j) = \sum_j x_j \ln \frac{x_j}{y_j} = \mathrm{KL}\,[\boldsymbol{x} : \boldsymbol{y}] \quad (44)$$

where it is assumed $\sum_j x_j = \sum_j y_j = 1$.

Let $X$ be a quantity taking values in the domain of $\phi$ with a probability distribution function $p(x)$. Then, $\mathrm{E}_{p(x)}\{B_\phi(X, m)\}$ is minimized over $m$ in the domain of $\phi$ at $m = \mathrm{E}\{X\}$:

$$\widehat{m} = \arg\min_m \{B_\phi(X, m)\} = \mathrm{E}\{X\}\,.$$

Moreover, this property characterizes Bregman divergence. When applied to the Bayesian approach, this means that, using the Bregman divergence as the loss function, the Bayes estimator is the posterior mean. This point is detailed in the following.

Links between all of these through an example:
Let us consider the Bayesian parameter estimation where we have some data $\boldsymbol{y}$, a set of parameters $\boldsymbol{x}$, a likelihood $p(\boldsymbol{y}|\boldsymbol{x})$ and a prior $\pi(\boldsymbol{x})$, which gives the posterior $p(\boldsymbol{x}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{x})\,\pi(\boldsymbol{x})$. Let us also consider a cost function $C[\boldsymbol{x}, \widetilde{\boldsymbol{x}}]$ in the parameter space $\boldsymbol{x} \in \mathcal{X}$. The classical Bayesian point estimation of $\boldsymbol{x}$ is expressed as the minimizer of an expected risk:

$$\widehat{\boldsymbol{x}} = \arg\min_{\widetilde{\boldsymbol{x}}} \left\{\bar{C}(\widetilde{\boldsymbol{x}})\right\} \quad (45)$$

where:

$$\bar{C}(\widetilde{\boldsymbol{x}}) = \mathrm{E}_{p(\boldsymbol{x}|\boldsymbol{y})}\{C[\boldsymbol{x}, \widetilde{\boldsymbol{x}}]\} = \int C[\boldsymbol{x}, \widetilde{\boldsymbol{x}}]\, p(\boldsymbol{x}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{x}$$

It is very well known that the mean squared error estimator, which corresponds to $C[\boldsymbol{x}, \widetilde{\boldsymbol{x}}] = \|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|^2$, is the posterior mean. It is now interesting to know that choosing $C[\boldsymbol{x}, \widetilde{\boldsymbol{x}}]$ to be any Bregman divergence $B_\phi[\boldsymbol{x}, \widetilde{\boldsymbol{x}}]$, we obtain also the posterior mean:

$$\widehat{\boldsymbol{x}} = \arg\min_{\widetilde{\boldsymbol{x}}} \left\{\bar{B}_\phi(\widetilde{\boldsymbol{x}})\right\} = \mathrm{E}_{p(\boldsymbol{x}|\boldsymbol{y})} \left\{ \int \boldsymbol{x}\, p(\boldsymbol{x}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{x} \right\} \quad (46)$$

where:

$$\bar{B}_\phi(\widetilde{\boldsymbol{x}}) = \mathrm{E}_{p(\boldsymbol{x}|\boldsymbol{y})}\{D_\phi[\boldsymbol{x}, \widetilde{\boldsymbol{x}}]\} = \int B_\phi[\boldsymbol{x}, \widetilde{\boldsymbol{x}}]\, p(\boldsymbol{x}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{x}$$

Consider now that we have two prior probability laws $\pi_1(\boldsymbol{x})$ and $\pi_2(\boldsymbol{x})$, which give rise to two posterior probability laws $p_1(\boldsymbol{x}|\boldsymbol{y})$ and $p_2(\boldsymbol{x}|\boldsymbol{y})$. If the prior laws and the likelihood are in the exponential families, then the posterior laws are also in the exponential family. Let us note them as $p_1(\boldsymbol{x}|\boldsymbol{y}; \boldsymbol{\theta}_1)$ and $p_2(\boldsymbol{x}|\boldsymbol{y}; \boldsymbol{\theta}_2)$, where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are the parameters of those posterior laws. We then have the following properties:

- KL $[p_1 : p_2]$ is expressed as a Bregman divergence $B[\boldsymbol{\theta}_1 : \boldsymbol{\theta}_2]$.
- A Bregman divergence $B[\boldsymbol{x}_1 : \boldsymbol{x}_2]$ is induced when KL $[p_1 : p_2]$ is used to compare the two posteriors.

## 9. Vectorial Variables and Time Indexed Process

The extension of the scalar variable to the finite dimensional vectorial case is almost immediate. In particular, for the Gaussian case $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{R})$, the mean becomes a vector $\boldsymbol{\mu} = \mathrm{E}\{\boldsymbol{X}\}$, and the variances are replaced by a covariance matrix: $\boldsymbol{R} = \mathrm{E}\{(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})'\}$; and almost all of the quantities can be defined immediately. For example, for a Gaussian vector $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{0}, \boldsymbol{R})$, the entropy is given by [49]:

$$H = \frac{n}{2} \ln(2\pi) + \frac{1}{2} \ln(|\det(\boldsymbol{R})|) \tag{47}$$

and the relative entropy of $\mathcal{N}(\boldsymbol{x}|\boldsymbol{0}, \boldsymbol{R})$ with respect to $\mathcal{N}(\boldsymbol{x}|\boldsymbol{0}, \boldsymbol{S})$ is given by:

$$D = -\frac{1}{2}\left(\mathrm{tr}\left(\boldsymbol{R}\boldsymbol{S}^{-1}\right) - \log\frac{|\det(\boldsymbol{R})|}{|\det(\boldsymbol{S})|} - n\right). \tag{48}$$

The notion of time series or processes needs extra definitions. For example, for a random time series $X(t)$, we can define $p(X(t)), \forall t$, the expected value time series $\bar{x}(t) = \mathrm{E}\{X(t)\}$ and what is called the autocorrelation function $\Gamma(t, \tau) = \mathrm{E}\{X(t)\,X(t+\tau)\}$. A time series is called stationary when these quantities does not depend on $t$, *i.e.*, $\bar{x}(t) = m$ and $\Gamma(t, \tau) = \Gamma(\tau)$ [50]. Another quantity of interest for a stationary time series is its power spectral density (PSD) function:

$$S(\omega) = \mathrm{FT}\{\Gamma(\tau)\} = \int \Gamma(\tau) \exp[-j\omega\tau]\, \mathrm{d}\tau. \tag{49}$$

When $X(t)$ is observed on times $t = n\Delta T$ with $\Delta T = 1$, we have $X(n)$, and for a sample $\{X(1), \cdots, X(N)\}$, we may define the mean $\boldsymbol{\mu} = \mathrm{E}\{\boldsymbol{X}\}$ and the covariance matrix $\boldsymbol{\Sigma} = \mathrm{E}\{(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})'\}$.

With these definitions, it can easily been shown that the covariance matrix of a stationary Gaussian process is Toeplitz [49]. It is also possible to show that the entropy of such a process can be expressed as a function of its PSD function:

$$\lim_{n\longrightarrow\infty} \frac{1}{n} H(p) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln S(\omega)\, \mathrm{d}\omega. \tag{50}$$

For two stationary Gaussian processes with two spectral density functions $S_1(\omega)$ and $S_2(\omega)$, we have:

$$\lim_{n\longrightarrow\infty} \frac{1}{n} D[p_1 : p_2] = \frac{1}{4\pi} \int_{-\pi}^{\pi} \left(\frac{S_1(\omega)}{S_2(\omega)} - \ln\frac{S_1(\omega)}{S_2(\omega)} - 1\right) \mathrm{d}\omega \tag{51}$$

where we find the Itakura–Saito distance in the spectral analysis literature [50–53].

These definitions and expressions have often been used in time series analysis. In what follows, we give a few examples of the different ways these notions and quantities have been used in different applications of data, signal and image processing.

## 10. Entropy in Independent Component Analysis and Source Separation

Given a vector of time series $\boldsymbol{x}(t)$, the independent component analysis (ICA) consists of finding a separating matrix $\boldsymbol{B}$, such that the components $\boldsymbol{y}(t) = \boldsymbol{B}\boldsymbol{x}(t)$ are as independent as possible. The notion of entropy is used here as a measure of independence. For example, to find $\boldsymbol{B}$, we may choose $D\left[p(\boldsymbol{y}) : \prod_j p_j(y_j)\right]$ as a criterion of independence of the components $y_j$. The next step is to choose a probability law $p(\boldsymbol{x})$ from which we can find an expression for $p(\boldsymbol{y})$ from which we can find an expression for $D\left[p(\boldsymbol{y}) : \prod_j p_j(y_j)\right]$ as a function of the matrix $\boldsymbol{B}$, which can be optimized to obtain it.

The ICA problem has a tight link with the source separation problem, where it is assumed that the measured time series $\boldsymbol{x}(t)$ is a linear combination of the sources $\boldsymbol{s}(t)$, *i.e.*, $\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{s}(t)$, with $\boldsymbol{A}$ being the mixing matrix. The objective of source separation is then to find the separating matrix $\boldsymbol{B} = \boldsymbol{A}^{-1}$.

To see how the entropy is used here, let us note $\boldsymbol{y} = \boldsymbol{B}\boldsymbol{x}$. Then,

$$p_Y(\boldsymbol{y}) = \frac{1}{|\partial \boldsymbol{y}/\partial \boldsymbol{x}|} p_X(\boldsymbol{x}) \longrightarrow H(\boldsymbol{y}) = -\mathrm{E}\left\{\ln p_Y(\boldsymbol{y})\right\} = \mathrm{E}\left\{\ln |\partial \boldsymbol{y}/\partial \boldsymbol{x}|\right\} - H(\boldsymbol{x}). \tag{52}$$

$H(\boldsymbol{y})$ is used as a criterion for ICA or source separation. As the objective in ICA is to obtain $\boldsymbol{y}$ in such a way that its components become as independent as possible, the separating matrix $\boldsymbol{B}$ has to maximize $H(\boldsymbol{y})$. Many ICA algorithms are based on this optimization [54–65]

## 11. Entropy in Parametric Modeling and Model Selection

Determining the order of a model, *i.e.*, the dimension of the vector parameter $\boldsymbol{\theta}$ in a probabilistic model $p(\boldsymbol{x}|\boldsymbol{\theta})$, is an important subject in many data and signal processing problems. As an example, in autoregressive (AR) modeling:

$$x(n) = \sum_{k=1}^{K} \theta_k x(n-k) + \epsilon(n) \tag{53}$$

where $\boldsymbol{\theta} = \{\theta_1, \cdots, \theta_K\}$, we may want to compare two models with two different values of $K$.

When the order $K$ is fixed, the estimation of the parameters $\boldsymbol{\theta}$ is a very well-known problem, and there are likelihood based [66] or Bayesian approaches for that [67]. The determination of the order is however more difficult [68]. Between the tools, we may mention here the Bayesian methods [69–74], but also the use of relative entropy $D\left[p(\boldsymbol{x}|\boldsymbol{\theta}^*) : p(\boldsymbol{x}|\boldsymbol{\theta})\right]$, where $\boldsymbol{\theta}^*$ represents the vector of the parameters of dimension $K^*$ and $\boldsymbol{\theta}$ and the vector $\boldsymbol{\theta}$ with dimension $K \leq K^*$. In such cases, even if the two probability laws to be compared have parameters with different dimensions, we can always use the KL $\left[p(\boldsymbol{x}|\boldsymbol{\theta}^*) : p(\boldsymbol{x}|\boldsymbol{\theta})\right]$ to compare them. The famous criterion of Akaike [75–78] uses this quantity to determine the optimal order. For a linear parameter model with Gaussian probability laws and likelihood-based methods, there are analytic solutions for it [68].

## 12. Entropy in Spectral Analysis

Entropy and MEP have been used in different ways in the spectral analysis problem. It has been an important subject of signal processing for the decades. Here, we are presenting, in a brief way, these different approaches.

### *12.1. Burg's Entropy-Based Method*

A classical one is Burg's entropy method [79], which can be summarized as follows: Let $X(n)$ be a stationary, centered process, and assume we have as data a finite number of samples (lags) of its autocorrelation function:

$$r(k) = \mathrm{E}\left\{X(n)X(n+k)\right\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} S(\omega)\exp\left[\mathrm{j}k\omega\right]\,\mathrm{d}\omega, \quad k = 0,\ldots,K. \tag{54}$$

The task is then to estimate its power spectral density function:

$$S(\omega) = \sum_{k=-\infty}^{\infty} r(k)\exp\left[-\mathrm{j}k\omega\right]$$

As we can see, due to the fact that we have only the elements of the right-hand for $k = -K,\cdots,+K$, the problem is ill posed. To obtain a probabilistic solution, we may start by assigning a probability law $p(\boldsymbol{x})$ to the vector $\underline{X} = [X(0),\ldots,X(N-1)]'$. For this, we can use the principle of maximum entropy (PME) with the data as constraints (54). As these constraints are the second order moments, the PME solution is a Gaussian probability law: $\mathcal{N}(\boldsymbol{x}|\mathbf{0},\boldsymbol{R})$. For a stationary Gaussian process, when the number of samples $N \longrightarrow \infty$, the expression of the entropy becomes:

$$H = \int_{-\pi}^{\pi} \ln S(\omega)\,\mathrm{d}\omega. \tag{55}$$

This expression is called Burg's entropy [79]. Thus, Burg's method consists of maximizing $H$ subject to the constraints (54). The solution is:

$$S(\omega) = \frac{1}{\left|\displaystyle\sum_{k=-K}^{K}\lambda_k\exp\left[\mathrm{j}k\omega\right]\right|^2}, \tag{56}$$

where $\boldsymbol{\lambda} = [\lambda_0,\cdots,\lambda_K]'$, the Lagrange multipliers associated with the constraints (54), are here equivalent to the AR modeling of the Gaussian process $X(n)$.

We may note that, in this particular case, we have an analytical expression for $\boldsymbol{\lambda}$, which provides the possibility to give an analytical expression for $S(\omega)$ as a function of the data $\{r(k), k = 0,\cdots,K\}$:

$$S(\omega) = \frac{\boldsymbol{\delta}\,\boldsymbol{\Gamma}^{-1}\boldsymbol{\delta}}{\boldsymbol{e}\,\boldsymbol{\Gamma}^{-1}\boldsymbol{e}}, \tag{57}$$

where $\boldsymbol{\Gamma} = \text{Toeplitz}(r(0),\cdots,r(K))$ is the correlation matrix and $\boldsymbol{\delta}$ and $\boldsymbol{e}$ are two vectors defined by $\boldsymbol{\delta} = [1,0,\cdots,0]'$ and $\boldsymbol{e} = [1,e^{-\mathrm{j}\omega},e^{-\mathrm{j}2\omega},\cdots,e^{-\mathrm{j}K\omega}]'$.

We may note that we first used MEP to choose a probability law for $X(n)$. With the prior knowledge that we have second order moments, the MEP results in a Gaussian probability density function. Then, as for a stationary Gaussian process, the expression of the entropy is related to the power spectral density $S(\omega)$, and as this is related to the correlation data by a Fourier transform, an ME solution could be computed easily.

*12.2. Extensions to Burg's Method*

The second approach consists of maximizing the relative entropy $D\left[p(\boldsymbol{x}) : p_0(\boldsymbol{x})\right]$ or minimizing $\mathrm{KL}\left[p(\boldsymbol{x}) : p_0(\boldsymbol{x})\right]$ where $p_0(\boldsymbol{x})$ is an *a priori* law. The choice of the prior is important. Choosing a uniform $p_0(\boldsymbol{x})$, we retrieve the previous case [77].

However, choosing a Gaussian law for $p_0(\boldsymbol{x})$, the expression to maximize becomes:

$$D\left[p(\boldsymbol{x}) : p_0(\boldsymbol{x})\right] = \frac{1}{4\pi} \int_{-\pi}^{\pi} \left( \frac{S(\omega)}{S_0(\omega)} - \ln \frac{S(\omega)}{S_0(\omega)} - 1 \right) \mathrm{d}\omega \tag{58}$$

when $N \mapsto \infty$ and where $S_0(\omega)$ corresponds to the power spectral density of the reference process $p_0(\boldsymbol{x})$. Now, the problem becomes: minimize $D\left[p(\boldsymbol{x}) : p_0(\boldsymbol{x})\right]$ subject to the constraints (54).

*12.3. Shore and Johnson Approach*

Another approach is to decompose first the process $X(n)$ on the Fourier basis $\{\cos k\omega t,\ \sin k\omega t\}$, to consider $\omega$ to be the variable of interest and $S(\omega)$, normalized properly, to be considered as its probability distribution function. Then, the problem can be reformulated as the determination of the $S(\omega)$, which maximizes the entropy:

$$-\int_{-\pi}^{\pi} S(\omega)\, \ln S(\omega)\, \mathrm{d}\omega \tag{59}$$

subject to the linear constraints (54). The solution is in the form of:

$$S(\omega) = \exp\left[ \sum_{k=-K}^{K} \lambda_k \exp\left[\mathrm{j}k\omega\right] \right]. \tag{60}$$

which can be considered as the most uniform power spectral density that satisfies those constraints.

*12.4. ME in the Mean Approach*

In this approach, we consider $S(\omega)$ as the expected value $Z(\omega)$ for which we have a prior law $\mu(z)$, and we are looking to assign $p(z)$, which maximizes the relative entropy $D\left[p(z) : \mu(z)\right]$ subject to the constraints (54).

When $p(z)$ is determined, the solution is given by:

$$S(\omega) = \mathrm{E}\left\{Z(\omega)\right\} = \int Z(\omega) p(z)\, \mathrm{d}z. \tag{61}$$

The expression of $S(\omega)$ depends on $\mu(z)$. When $\mu(z)$ is Gaussian, we obtain the Rényi entropy:

$$H = \int_{-\pi}^{\pi} S^2(\omega)\,\mathrm{d}\omega. \tag{62}$$

If we choose a Poisson measure for $\mu(z)$, we obtain the Shannon entropy:

$$H = -\int_{-\pi}^{\pi} S(\omega)\ln S(\omega)\,\mathrm{d}\omega, \tag{63}$$

and if we choose a Lebesgue measure over $[0,\infty]$, we obtain Burg's entropy:

$$H = \int_{-\pi}^{\pi} \ln S(\omega)\,\mathrm{d}\omega. \tag{64}$$

When this step is done, the next step becomes maximizing these entropies subject to the constraints of the correlations. The obtained solutions are very different. For more details, see [39,79–85].

## 13. Entropy-Based Methods for Linear Inverse Problems

### *13.1. Linear Inverse Problems*

A general way to introduce inverse problems is the following: Infer an unknown signal $f(t)$, image $f(x,y)$ or any multi-variable function $f(\boldsymbol{r})$ through an observed signal $g(t)$, image $g(x,y)$ or any multi-variable observable function $g(\boldsymbol{s})$, which are related through an operator $\mathcal{H}: f \mapsto g$. This operator can be linear or nonlinear. Here, we consider only linear operators $g = \mathcal{H}f$:

$$g(\boldsymbol{s}) = \int h(\boldsymbol{r},\boldsymbol{s}) f(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r} \tag{65}$$

where $h(\boldsymbol{r},\boldsymbol{s})$ is the response of the measurement system. Such linear operators are very common in many applications of signal and image processing. We may mention a few examples of them:

- Convolution operations $g = h * f$ in 1D (signal):

$$g(t) = \int h(t-t')f(t')\,\mathrm{d}t' \tag{66}$$

  or in 2D (image):

$$g(x,y) = \iint h(x-x',y-y')f(x',y')\,\mathrm{d}x'\,\mathrm{d}y' \tag{67}$$

- Radon transform (RT) in computed tomography (CT) in the 2D case [86]:

$$g(r,\phi) = \int\int \delta(r - x\cos\phi - y\sin\phi)\, f(x,y)\,\mathrm{d}x\,\mathrm{d}y \tag{68}$$

- Fourier transform (FT) in the 2D case:

$$g(u,v) = \int\int \exp\left[-j(ux+vy)\right]\, f(x,y)\,\mathrm{d}x\,\mathrm{d}y \tag{69}$$

  which arise in magnetic resonance imaging (MRI), in synthetic aperture radar (SAR) imaging or in microwave and diffraction optical tomography (DOT) [86–90].

No matter the category of the linear transforms, when the problem is discretized, we arrive at the relation:

$$\boldsymbol{g} = \boldsymbol{H}\boldsymbol{f} + \boldsymbol{\epsilon}, \tag{70}$$

where $\boldsymbol{f} = [f_1, \cdots, f_n]'$ represents the unknowns, $\boldsymbol{g} = [g_1, \cdots, g_m]'$ the observed data, $\boldsymbol{\epsilon} = [\epsilon_1, \cdots, \epsilon_m]'$ the errors of modeling and measurement and $\boldsymbol{H}$ the matrix of the system response.

*13.2. Entropy-Based Methods*

Let us consider first the simple no noise case:

$$\boldsymbol{g} = \boldsymbol{H}\boldsymbol{f}, \tag{71}$$

where $\boldsymbol{H}$ is a matrix of dimensions $(M \times N)$, which is in general singular or very ill conditioned. Even if the cases $M > N$ or $M = N$ may appear easier, they have the same difficulties as those of the underdetermined case $M < N$ that we consider here. In this case, evidently the problem has an infinite number of solutions, and we need to choose one.

Between the numerous methods, we may mention the minimum norm solution, which consists of choosing between all of the possible solutions:

$$\mathcal{F} = \{\boldsymbol{f} : \boldsymbol{H}\boldsymbol{f} = \boldsymbol{g}\} \tag{72}$$

the one that has the minimum norm:

$$\Omega(\boldsymbol{f}) = \|\boldsymbol{f}\|_2^2 = \sum_j f_j^2. \tag{73}$$

This optimization problem can be solved easily in this case, and we obtain:

$$\widehat{\boldsymbol{f}}_{NM} = \arg\min_{\boldsymbol{f}\in\mathcal{F}} \left\{\Omega(\boldsymbol{f}) = \|\boldsymbol{f}\|_2^2\right\} = \boldsymbol{H}'(\boldsymbol{H}\boldsymbol{H}')^{-1}\boldsymbol{g}. \tag{74}$$

In fact, we may choose any other convex criterion $\Omega(\boldsymbol{f})$ and satisfy the uniqueness of the solution. For example:

$$\Omega(\boldsymbol{f}) = -\sum_j f_j \ln f_j \tag{75}$$

which can be interpreted as the entropy when $f_j > 0$ and $\sum f_j = 1$, thus considering $f_j$ as a probability distribution $f_j = P(U = u_j)$. The variable $U$ can correspond (or not) to a physical quantity. $\Omega(\boldsymbol{f})$ is the entropy associated with this variable.

If we consider $f_j > 0$ to represent the power spectral density of a physical quantity, then the entropy becomes:

$$\Omega(\boldsymbol{f}) = \sum_j \ln f_j \tag{76}$$

and we can use it as criterion to select a solution to the problem (71).

As we can see, any convex criterion $\Omega(\boldsymbol{f})$ can be used. Here, we mentioned four of them with different interpretations.

- $L_2$ or quadratic:

$$\Omega(\boldsymbol{f}) = \sum_j f_j^2 \tag{77}$$

which can be interpreted as the Rényi's entropy with $q = 2$.

- $L_\beta$ :

$$\Omega(\boldsymbol{f}) = \sum_j |f_j|^\beta \tag{78}$$

When $\beta < 1$ the criterion is not bounded at zero. When $\beta \geq 1$ the criterion is convex.

- Shannon entropy:

$$\Omega(\boldsymbol{f}) = -\sum_j f_j \ln f_j \tag{79}$$

which has a valid interpretation if $0 < f_j < 1$,

- The Burg entropy:

$$\Omega(\boldsymbol{f}) = \sum_j \ln f_j \tag{80}$$

which needs $f_j > 0$.

Unfortunately, only for the first case, there is an analytical solution for the problem, which is $\widehat{\boldsymbol{f}} = \boldsymbol{H}'(\boldsymbol{H}\boldsymbol{H}')\boldsymbol{g}$. For all of the other cases, we may need an optimization algorithm to obtain a numerical solution [91–95].

*13.3. Maximum Entropy in the Mean Approach*

A second approach consists of considering $f_j = \mathrm{E}\{U_j\}$ or $\boldsymbol{f} = \mathrm{E}\{\boldsymbol{U}\}$ [41,41,42]. Again, here, $U_j$ or $\boldsymbol{U}$ can, but need not, correspond to some physical quantities. In any case, we now want to assign a probability law $\widehat{p}(\boldsymbol{u})$ to it. Noting that the data $\boldsymbol{g} = \boldsymbol{H}\boldsymbol{f} = \boldsymbol{H}\mathrm{E}\{\boldsymbol{U}\} = \mathrm{E}\{\boldsymbol{H}\boldsymbol{U}\}$ can be considered as the constraints on it, we may need again a criterion to determine $\widehat{p}(\boldsymbol{u})$. Assuming then having some prior $\mu(\boldsymbol{u})$, we may maximize the relative entropy as that criterion. The mathematical problem then becomes:

$$\text{minimize } D[p(\boldsymbol{u}) : \mu(\boldsymbol{u})] \text{ subject to } \int \boldsymbol{H}\boldsymbol{u}\, p(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} = \boldsymbol{g} \tag{81}$$

The solution is:

$$\widehat{p}(\boldsymbol{u}) = \frac{1}{Z(\boldsymbol{\lambda})}\mu(\boldsymbol{u}) \exp\left[-\boldsymbol{\lambda}'\boldsymbol{H}\boldsymbol{u}\right] \tag{82}$$

where:

$$Z(\boldsymbol{\lambda}) = \int \mu(\boldsymbol{u}) \exp\left[-\boldsymbol{\lambda}'\boldsymbol{H}\boldsymbol{u}\right]\, \mathrm{d}\boldsymbol{u}. \tag{83}$$

When $\widehat{p}(\boldsymbol{u})$ is obtained, we may be interested in computing:

$$\widehat{\boldsymbol{f}} = \mathrm{E}\{\boldsymbol{U}\} = \int \boldsymbol{u}\widehat{p}(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} \tag{84}$$

which is the required solution.

Interestingly, if we focus on $\widehat{\boldsymbol{f}} = \mathrm{E}\{\boldsymbol{U}\}$, we will see that its expression depends on the choice of the prior $\mu(\boldsymbol{u})$. When $\mu(\boldsymbol{u})$ is separable: $\mu(\boldsymbol{u}) = \prod_j \mu_j(u_j)$, the expression of $\widehat{p}(\boldsymbol{u})$ will also be separable.

To go a little more into the details, let us introduce $\boldsymbol{s} = \boldsymbol{H}'\boldsymbol{\lambda}$ and define:

$$G(\boldsymbol{s}) = \ln \int \mu(\boldsymbol{u}) \exp\left[-\boldsymbol{s}'\boldsymbol{u}\right] \mathrm{d}\boldsymbol{u} \tag{85}$$

and its conjugate convex:

$$F(\boldsymbol{f}) = \sup_{\boldsymbol{s}} \left\{\boldsymbol{f}'\boldsymbol{s} - G(\boldsymbol{s})\right\}. \tag{86}$$

It can be shown easily that $\widehat{\boldsymbol{f}} = \mathrm{E}\{\boldsymbol{U}\}$ can be obtained either via the dual $\widehat{\boldsymbol{\lambda}}$ variables:

$$\widehat{\boldsymbol{f}} = G'(\boldsymbol{H}'\widehat{\boldsymbol{\lambda}}) \tag{87}$$

where $\widehat{\boldsymbol{\lambda}}$ is obtained by:

$$\widehat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}} \left\{D(\boldsymbol{\lambda}) = \ln Z(\boldsymbol{\lambda}) + \boldsymbol{\lambda}'\boldsymbol{g}\right\}, \tag{88}$$

or directly:

$$\widehat{\boldsymbol{f}} = \arg\min_{\{\boldsymbol{f}\,:\,\boldsymbol{H}\boldsymbol{f}=\boldsymbol{g}\}} \left\{F(\boldsymbol{f})\right\}. \tag{89}$$

$D(\boldsymbol{\lambda})$ is called the dual criterion and $F(\boldsymbol{f})$ primal. However, it is not always easy to obtain an analytical expression for $G(\boldsymbol{s})$ and its gradient $G'(\boldsymbol{s})$. The functions $F(\boldsymbol{f})$ and $G(\boldsymbol{s})$ are conjugate convex.

For the computational aspect, unfortunately, the cases where we may have analytical expressions for $Z(\boldsymbol{\lambda})$ or $G(\boldsymbol{s}) = \ln Z$ or $F(\boldsymbol{f})$ are very limited. However, when there is analytical expressions for them, the computations can be done very easily. In Table 1, we summarizes some of those solutions:

**Table 1.** Analytical solutions for different measures $\mu(\boldsymbol{u})$

| | | |
|---|---|---|
| $\mu(\boldsymbol{u}) \propto \exp[-\frac{1}{2}\sum_j u_j^2]$ | $\widehat{\boldsymbol{f}} = \boldsymbol{H}'\boldsymbol{\lambda}$ | $\widehat{\boldsymbol{f}} = \boldsymbol{H}'(\boldsymbol{H}\boldsymbol{H}')^{-1}\boldsymbol{g}$ |
| $\mu(\boldsymbol{u}) \propto \exp[-\sum_j \lvert u_j\rvert]$ | $\widehat{\boldsymbol{f}} = \mathbf{1}./(\boldsymbol{H}'\boldsymbol{\lambda} \pm \mathbf{1})$ | $\boldsymbol{H}\widehat{\boldsymbol{f}} = \boldsymbol{g}$ |
| $\mu(\boldsymbol{u}) \propto \exp[-\sum_j u_j^{\alpha-1} \exp\left[-\beta u_j\right]], \quad u_j > 0$ | $\widehat{\boldsymbol{f}} = \alpha\mathbf{1}./(\boldsymbol{H}'\boldsymbol{\lambda} + \beta\mathbf{1})$ | $\boldsymbol{H}\widehat{\boldsymbol{f}} = \boldsymbol{g}$ |

## 14. Bayesian Approach for Inverse Problems

In this section, we present in a brief way the Bayesian approach for the inverse problems in signal and image processing.

*14.1. Simple Bayesian Approach*

The different steps to find a solution to an inverse problem using the Bayesian approach can be summarized as follows:

- Assign a prior probability law $p(\boldsymbol{\epsilon})$ to the modeling and observation errors, here $\boldsymbol{\epsilon}$. From this, find the expression of the likelihood $p(\boldsymbol{g}|\boldsymbol{f},\boldsymbol{\theta}_1)$. As an example, consider the Gaussian case:

$$p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, v_\epsilon \boldsymbol{I}) \longrightarrow p(\boldsymbol{g}|\boldsymbol{f}) = \mathcal{N}(\boldsymbol{g}|\boldsymbol{H}\boldsymbol{f}, v_\epsilon \boldsymbol{I}). \tag{90}$$

  $\theta_1$ in this case is the noise variance $v_\epsilon$.

- Assign a prior probability law $p(\boldsymbol{f}|\boldsymbol{\theta}_2)$ to the unknown $\boldsymbol{f}$ to translate your prior knowledge on it. Again, as an example, consider the Gaussian case:

$$p(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{f}|\mathbf{0}, v_f \boldsymbol{I}) \tag{91}$$

  $\theta_2$ in this case is the variance $v_f$.

- Apply the Bayes rule to obtain the expression of the posterior law:

$$p(\boldsymbol{f}|\boldsymbol{g},\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = \frac{p(\boldsymbol{g}|\boldsymbol{f},\boldsymbol{\theta}_1)\, p(\boldsymbol{f}|\boldsymbol{\theta}_2)}{p(\boldsymbol{g}|\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)} \propto p(\boldsymbol{g}|\boldsymbol{f},\boldsymbol{\theta}_1)\, p(\boldsymbol{f}|\boldsymbol{\theta}_2), \tag{92}$$

  where the sign $\propto$ stands for "proportionality to", $p(\boldsymbol{g}|\boldsymbol{f},\boldsymbol{\theta}_1)$ is the likelihood, $p(\boldsymbol{f}|\boldsymbol{\theta}_2)$ the prior model, $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2]'$ their corresponding parameters (often called the hyper-parameters of the problem) and $p(\boldsymbol{g}|\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ is called the evidence of the model.

- Use $p(\boldsymbol{f}|\boldsymbol{g},\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ to infer any quantity dependent of $\boldsymbol{f}$.

For the expressions of likelihood in (90) and the prior in (91), we obtain very easily the expression of the posterior:

$$p(\boldsymbol{f}|\boldsymbol{g}, v_\epsilon, v_f) = \mathcal{N}(\boldsymbol{f}|\widehat{\boldsymbol{f}}, \widehat{\boldsymbol{V}}) \text{ with } \widehat{\boldsymbol{V}} = (\boldsymbol{H}'\boldsymbol{H} + \frac{v_\epsilon}{v_f}\boldsymbol{I})^{-1} \text{ and } \widehat{\boldsymbol{f}} = \widehat{\boldsymbol{V}}\boldsymbol{H}'\boldsymbol{g} \tag{93}$$

When the hyper-parameters $\boldsymbol{\theta}$ can be fixed *a priori*, the problem is easy. In practice, we may use some summaries, such as:

- MAP:

$$\widehat{\boldsymbol{f}}_{\text{MAP}} = \arg\max_{\boldsymbol{f}} \{p(\boldsymbol{f}|\boldsymbol{g},\boldsymbol{\theta})\} \tag{94}$$

- EAP or posterior mean (PM):

$$\widehat{\boldsymbol{f}}_{EAP} = \int \boldsymbol{f}\, p(\boldsymbol{f}|\boldsymbol{g},\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{f} \tag{95}$$

For the Gaussian case of (91), the MAP and EAP are the same and can be obtained by noting that:

$$\widehat{\boldsymbol{f}}_{\text{MAP}} = \arg\min_{\boldsymbol{f}} \{J(\boldsymbol{f})\} \text{ with } J(\boldsymbol{f}) = \|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\|_2^2 + \lambda\|\boldsymbol{f}\|_2^2, \text{ where } \lambda = v_\epsilon / v_f. \tag{96}$$

However, in real applications, the computation of even these simple point estimators may need efficient algorithm:

- For MAP, we need optimization algorithms, which can handle the huge dimensional criterion $J(\boldsymbol{f}) = -\ln p(\boldsymbol{f}|\boldsymbol{g}, \boldsymbol{\theta})$. Very often, we may be limited to using gradient-based algorithms.
- For EAP, we need integration algorithms, which can handle huge dimensional integrals. The most common tool here is the MCMC methods [24]. However, for real applications, very often, the computational costs are huge. Recently, different methods, called approximate Bayesian computation (ABC) [96–100] or VBA, have been proposed [74,96,98,101–107].

*14.2. Full Bayesian: Hyperparameter Estimation*

When the hyperparameters $\boldsymbol{\theta}$ have also to be estimated, a prior $p(\boldsymbol{\theta})$ is assigned to them, and the expression of the joint posterior:

$$p(\boldsymbol{f}, \boldsymbol{\theta}|\boldsymbol{g}) = \frac{p(\boldsymbol{g}|\boldsymbol{f}, \boldsymbol{\theta}_1)\, p(\boldsymbol{f}|\boldsymbol{\theta}_2)\, p(\boldsymbol{\theta})}{p(\boldsymbol{g})} \tag{97}$$

is obtained, which can then be used to infer them jointly. Very often, the expression of this joint posterior law is complex, and any computation may become very costly. The VBA methods try to approximate $p(\boldsymbol{f}, \boldsymbol{\theta}|\boldsymbol{g})$ by a simpler distribution, which can be handled more easily. Two particular and extreme cases are:

- Bloc separable, such as $q(\boldsymbol{f}, \boldsymbol{\theta}) = q_1(\boldsymbol{f})\, q_2(\boldsymbol{\theta})$ or
- Completely separable, such as $q(\boldsymbol{f}, \boldsymbol{\theta}) = \prod_j q_{1j}(f_j) \prod_k q_{2k}(\theta_k)$.

Any mixed solution is also valid. For example, the one we have chosen is:

$$q(\boldsymbol{f}, \boldsymbol{\theta}) = q_1(\boldsymbol{f}) \prod_k q_{2k}(\theta_k) \tag{98}$$

Obtaining the expressions of these approximated separable probability laws has to be done via a criterion. The natural criterion with some geometrical interpretation for the probability law manifolds is the Kullback–Leibler (KL) criterion:

$$\text{KL}\,[q : p] = \int q \ln \frac{q}{p} = \left\langle \ln \frac{q}{p} \right\rangle_q . \tag{99}$$

For hierarchical prior models with hidden variables $\boldsymbol{z}$, the problem becomes more complex, because we have to give the expression of the joint posterior law:

$$p(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}|\boldsymbol{g}) \propto p(\boldsymbol{g}|\boldsymbol{f}, \boldsymbol{\theta}_1)\, p(\boldsymbol{f}|\boldsymbol{z}, \boldsymbol{\theta}_2)\, p(\boldsymbol{z}|\boldsymbol{\theta}_3)\, p(\boldsymbol{\theta}) \tag{100}$$

and then approximate it by separable ones:

$$q(\boldsymbol{f},\boldsymbol{z},\boldsymbol{\theta}|\boldsymbol{g}) = q_1(\boldsymbol{f})\, q_2(\boldsymbol{z})\, q_3(\boldsymbol{\theta}) \text{ or } q(\boldsymbol{f},\boldsymbol{\theta}) = \prod_j q_{1j}(f_j|z_{f_j}) \prod_j q_{2j}(z_{f_j}) \prod_k q_{3k}(\theta_k) \tag{101}$$

and then use them for estimation. See more discussions in [9,31,38,108–110]

In the following, first the general VBA method is detailed for the inference problems with hierarchical prior models. Then, a particular class of prior model (Student *t*) is considered, and the details of VBA algorithms for that are given.

## 15. Basic Algorithms of the Variational Bayesian Approximation

To illustrate the basic ideas and tools, let us consider a vector $\boldsymbol{X}$ and its probability density function $p(\boldsymbol{x})$, which we want to approximate by $q(\boldsymbol{x}) = \prod_j q_j(x_j)$. Using the KL criterion:

$$\begin{aligned}
\mathrm{KL}\left[q:p\right] &= \int q(\boldsymbol{x}) \ln \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\, \mathrm{d}\boldsymbol{x} = \int q(\boldsymbol{x}) \ln q(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} - \int q(\boldsymbol{x}) \ln p(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \\
&= \sum_j \int q_j(x_j) \ln q_j(x_j)\, \mathrm{d}x_j - \left\langle \ln p(\boldsymbol{x}) \right\rangle_q \\
&= \sum_j \int q_j(x_j) \ln q_j(x_j)\, \mathrm{d}x_j - \int q_j(x_j) < \ln p(\boldsymbol{x}) >_{q_{-j}}\, \mathrm{d}x_j
\end{aligned} \tag{102}$$

where we used the notation: $\left\langle \ln p(\boldsymbol{x}) \right\rangle_q = \int q(\boldsymbol{x}) \ln p(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}$ and $q_{-j}(\boldsymbol{x}) = \prod_{i \neq j} q_i(x_i)$.

From here, trying to find the solution $q_i$, the basic method is an alternate optimization algorithm:

$$q_j(x_j) \propto \exp\left[< \ln p(\boldsymbol{x}) >_{q_{-j}}\right]. \tag{103}$$

As we can see, the expression of $q_j(x_j)$ depends on $q_i(x_i), i \neq j$. It is not always possible to obtain analytical expressions for $q_j(x_j)$. It is however possible to show that, if $p(\boldsymbol{x})$ is a member of exponential families, then $q_j(x_j)$ are also members of exponential families. These iterations then become much simpler, because at each iteration, we need to update the parameters of the exponential families. To go a little more into the details, let us consider some particular simple cases.

### *15.1. Case of Two Gaussian Variables*

In the case of two variables $\boldsymbol{x} = [x_1, x_2]'$, we have:

$$\begin{cases} q_1(x_1) \propto \exp\left[< \ln p(\boldsymbol{x}) >_{q_2(x_2)}\right] \\ q_2(x_2) \propto \exp\left[< \ln p(\boldsymbol{x}) >_{q_1(x_1)}\right] \end{cases} \tag{104}$$

As an illustrative example, consider the case where we want to approximate $p(x_1, x_2)$ by $q(x_1, x_2) = q_1(x_1)\, q_2(x_2)$ to be able to compute the expected values:

$$\begin{cases} m_1 = \mathrm{E}\left\{x_1\right\} = \int\int x_1\, p(x_1, x_2)\, \mathrm{d}x_1\, \mathrm{d}x_2 \\ m_2 = \mathrm{E}\left\{x_2\right\} = \int\int x_2\, p(x_1, x_2)\, \mathrm{d}x_1\, \mathrm{d}x_2 \end{cases} \tag{105}$$

which need double integrations when $p(x_1, x_2)$ is not separable in its two variables. If we can do that separable approximation, then, we can compute:

$$\begin{cases} \widetilde{\mu}_1 = \mathrm{E}\{x_1\} = \int x_1\, q_1(x_1)\, \mathrm{d}x_1 \\ \widetilde{\mu}_2 = \mathrm{E}\{x_2\} = \int x_2\, q_2(x_2)\, \mathrm{d}x_2 \end{cases} \tag{106}$$

which needs only 1D integrals. Let us see if $(\widetilde{\mu}_1, \widetilde{\mu}_2)$ will converge to $(m_1, m_2)$. To illustrate this, let us consider the very simple case of the Gaussian:

$$p(x_1, x_2) = \mathcal{N}\left( \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \middle| \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}, \begin{bmatrix} v_1 & \rho\sqrt{v_1 v_2} \\ \rho\sqrt{v_1 v_2} & v_2 \end{bmatrix} \right). \tag{107}$$

It is then easy to see that $q_1(x_1) = \mathcal{N}(x_1|\widetilde{\mu}_1, \widetilde{v}_1)$ and $q_2(x_2) = \mathcal{N}(x_2|\widetilde{\mu}_2, \widetilde{v}_2)$ and that:

$$\begin{cases} q_1^{(k+1)}(x_1) = p(x_1|x_2 = \widetilde{\mu}_2^{(k)}) = \mathcal{N}\left(x_1|\widetilde{\mu}_1^{(k)}, \widetilde{v}_1^{(k)}\right) \\ q_2^{(k+1)}(x_2) = p(x_2|x_1 = \widetilde{\mu}_1^{(k)}) = \mathcal{N}\left(x_1|\widetilde{\mu}_2^{(k)}, \widetilde{v}_2^{(k)}\right) \end{cases} \tag{108}$$

with:

$$\begin{cases} \widetilde{\mu}_1^{(k+1)} = m_1 + \rho\sqrt{v_1/v_2}(\widetilde{\mu}_2^{(k)} - m_2) \\ \widetilde{v}_1^{(k+1)} = (1-\rho^2)v_1 \\ \widetilde{\mu}_2^{(k+1)} = m_2 + \rho\sqrt{v_2/v_1}(\widetilde{\mu}_1^{(k)} - m_1) \\ \widetilde{v}_2^{(k+1)} = (1-\rho^2)v_2 \end{cases}. \tag{109}$$

See [111] for details and where we showed that, initializing the algorithm with $\widetilde{\mu}_1^{(0)} = 0$ and $\widetilde{\mu}_2^{(0)} = 0$, the means converges to the right values $m_1$ and $m_2$, However, we may be careful about the convergence of the variances.

### *15.2. Case of Exponential Families*

As we could see, to be able to use such an algorithm in practical cases, we need to be able to compute $< \ln p(\boldsymbol{x}) >_{q_2(x_2)}$ and $< \ln p(\boldsymbol{x}) >_{q_1(x_1)}$. Only for a few cases can we can do this analytically. Different algorithms can be obtained depending on the choice of a particular family for $q_j(x_j)$ [103,112–120].

To show this, let us consider the exponential family:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = g(\boldsymbol{\theta})\, \exp\left[\boldsymbol{\theta}'\boldsymbol{u}(\boldsymbol{x})\right] \tag{110}$$

where $\boldsymbol{\theta}$ is a vector of parameter and $g(\boldsymbol{\theta})$ and $\boldsymbol{u}(\boldsymbol{x})$ are known functions.

This parametric exponential family has the following conjugacy property: For a given prior $p(\boldsymbol{\theta})$ in the family:

$$p(\boldsymbol{\theta}|\eta, \boldsymbol{\nu}) = h(\eta, \boldsymbol{\nu})\, g(\boldsymbol{\theta})^{\eta}\, \exp\left[\boldsymbol{\nu}'\boldsymbol{\theta}\right] \tag{111}$$

the corresponding posterior:

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{x}) &\propto p(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\eta, \boldsymbol{\nu}) \\ &\propto g(\boldsymbol{\theta})^{\eta+1} \exp\left[[\boldsymbol{\nu} + \boldsymbol{u}(\boldsymbol{x})]'\boldsymbol{\theta}\right] \\ &\propto p(\boldsymbol{\theta}|\eta + 1, \boldsymbol{\nu} + \boldsymbol{u}(\boldsymbol{x})) \end{aligned} \tag{112}$$

is in the same family.

For this family, we have:

$$\langle \ln p(\boldsymbol{x}|\boldsymbol{\theta}) \rangle_q = \ln g(\boldsymbol{\theta}) + \boldsymbol{\theta}' \langle \boldsymbol{u}(\boldsymbol{x}) \rangle_q . \tag{113}$$

It is then easy to show that:

$$q_j(x_j) \propto g(\boldsymbol{\theta}) \exp\left[\boldsymbol{\theta}' \langle \boldsymbol{u}(\boldsymbol{x}) \rangle_{q_{-j}}\right] \tag{114}$$

which are in the same exponential family. This simplifies greatly the computations, thanks to the fact that, in each iteration, we only need to compute $\widetilde{\boldsymbol{u}}(\boldsymbol{x}) = \langle \boldsymbol{u}(\boldsymbol{x}) \rangle_{q_{-j}}$ and update the parameters.

Now, if we consider:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = g(\boldsymbol{\theta}) \, \exp\left[\boldsymbol{\theta}' \boldsymbol{u}(\boldsymbol{x})\right] \tag{115}$$

with a prior on $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta}|\eta, \boldsymbol{\nu}) = h(\eta, \boldsymbol{\nu}) \, g(\boldsymbol{\theta})^{\eta} \, \exp\left[\boldsymbol{\nu}' \boldsymbol{\theta}\right] \tag{116}$$

and the joint $p(\boldsymbol{x}, \boldsymbol{\theta}|\eta, \boldsymbol{\nu}) = p(\boldsymbol{x}|\boldsymbol{\theta}) \, p(\boldsymbol{\theta}|\eta, \boldsymbol{\nu})$, which is not separable in $\boldsymbol{x}$ and $\boldsymbol{\theta}$, and we want to approximate it with the separable $q(\boldsymbol{x}, \boldsymbol{\theta}) = q_1(\boldsymbol{x}) \, q_2(\boldsymbol{\theta})$, then we will have:

$$\left\{ \begin{array}{l} q(\boldsymbol{\theta}) = h(\tilde{\eta}, \widetilde{\boldsymbol{\nu}}) \, g(\boldsymbol{\theta})^{\tilde{\eta}} \, \exp\left[\widetilde{\boldsymbol{\nu}}' \boldsymbol{\theta}\right] \\ q(\boldsymbol{x}) = g(\widetilde{\boldsymbol{\theta}}) \exp\left[\widetilde{\boldsymbol{\theta}}' \boldsymbol{u}(\boldsymbol{x})\right] \end{array} \right. \quad \text{with} \quad \left\{ \begin{array}{ll} \tilde{\eta} & = \eta + 1 \\ \widetilde{\boldsymbol{\nu}} & = \boldsymbol{\nu} + \widetilde{\boldsymbol{u}}(\boldsymbol{x}) \\ \widetilde{\boldsymbol{\theta}} & = \widetilde{\boldsymbol{\nu}} \end{array} \right. \tag{117}$$

where $\widetilde{\boldsymbol{u}} = \langle \boldsymbol{u}(\boldsymbol{x}) \rangle_{q_1(\boldsymbol{x})}$.

## 16. VBA for the Unsupervised Bayesian Approach to Inverse Problems

Before going into the details and for similarity with the notations in the next sections, we replace $\boldsymbol{x}$ by $\boldsymbol{f}$, such that now we are trying to approximate $p(\boldsymbol{f}, \boldsymbol{\theta}) = p(\boldsymbol{f}|\boldsymbol{\theta}) \, p(\boldsymbol{\theta})$ by a separable $q(\boldsymbol{f}, \boldsymbol{\theta}) = q_1(\boldsymbol{f}) \, q_2(\boldsymbol{\theta})$. Interestingly, depending on the choice of the family laws for $q_1$ and $q_2$, we obtain different algorithms:

- $q_1(\boldsymbol{f}) = \delta(\boldsymbol{f} - \widetilde{\boldsymbol{f}})$ and $q_2(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})$. In this case, we have:

$$\left\{ \begin{array}{l} q_1(\boldsymbol{f}) \propto \exp\left[< \ln p(\boldsymbol{f}, \boldsymbol{\theta}) >_{q_2}\right] \propto \exp\left[\ln p(\boldsymbol{f}, \widetilde{\boldsymbol{\theta}})\right] \propto p(\boldsymbol{f}, \boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}) \propto p(\boldsymbol{f}|\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}) \\ q_2(\boldsymbol{\theta}) \propto \exp\left[< \ln p(\boldsymbol{f}, \boldsymbol{\theta}) >_{q_1}\right] \propto \exp\left[\ln p(\widetilde{\boldsymbol{f}}, \boldsymbol{\theta})\right] \propto p(\boldsymbol{f} = \widetilde{\boldsymbol{f}}, \boldsymbol{\theta}) \propto p(\boldsymbol{\theta}|\boldsymbol{f} = \widetilde{\boldsymbol{f}}) \end{array} \right. \tag{118}$$

  and so:

$$\left\{ \begin{array}{l} \widetilde{\boldsymbol{f}} = \arg\max_{\boldsymbol{f}} \left\{ p(\boldsymbol{f}, \boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}) \right\} \\ \widetilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \left\{ p(\boldsymbol{f} = \widetilde{\boldsymbol{f}}, \boldsymbol{\theta}) \right\} \end{array} \right. \tag{119}$$

  which can be interpreted as an alternate optimization algorithm for obtaining the JMAPestimates:

$$(\widetilde{\boldsymbol{f}}, \widetilde{\boldsymbol{\theta}}) = \arg\max_{(\boldsymbol{f}, \boldsymbol{\theta})} \left\{ p(\boldsymbol{f}, \boldsymbol{\theta}) \right\} . \tag{120}$$

  The main drawback here is that the uncertainties of the $\boldsymbol{f}$ are not used for the estimation of $\boldsymbol{\theta}$ and the uncertainties of $\boldsymbol{\theta}$ are not used for the estimation of $\boldsymbol{f}$.

- $q_1(\boldsymbol{f})$ is free form and $q_2(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})$. In the same way, this time we obtain:

$$\left\{\begin{array}{l} < \ln p(\boldsymbol{f}, \boldsymbol{\theta}) >_{q_2(\boldsymbol{\theta})} = \ln p(\boldsymbol{f}, \widetilde{\boldsymbol{\theta}}) \\ < \ln p(\boldsymbol{f}, \boldsymbol{\theta}) >_{q_1(\boldsymbol{f})} = < \ln p(\boldsymbol{f}, \boldsymbol{\theta}) >_{q_1(\boldsymbol{f}|\widetilde{\boldsymbol{\theta}})} = Q(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}) \end{array}\right. \tag{121}$$

  which leads to:

$$\left\{\begin{array}{l} q_1(\boldsymbol{f}) \propto \exp\left[\ln p(\boldsymbol{f}, \boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}})\right] \propto p(\boldsymbol{f}, \widetilde{\boldsymbol{\theta}}) \\ q_2(\boldsymbol{\theta}) \propto \exp\left[Q(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}})\right] \longrightarrow \widetilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \left\{Q(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}})\right\} \end{array}\right. \tag{122}$$

  which can be compared with the Bayesian expectation maximization (BEM) algorithm. The E-step is the computation of the expectation $Q(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}})$ in (121), and the M-step is the maximization in (122). Here, the uncertainties of the $\boldsymbol{f}$ are used for the estimation of $\boldsymbol{\theta}$, but the uncertainties of $\boldsymbol{\theta}$ are not used for the estimation of $\boldsymbol{f}$.

- $q_1(\boldsymbol{f}) = \delta(\boldsymbol{f} - \widetilde{\boldsymbol{f}})$ and $q_2(\boldsymbol{\theta})$ is free form. In the same way, this time we obtain:

$$\left\{\begin{array}{l} < \ln p(\boldsymbol{f}, \boldsymbol{\theta}) >_{q_1(\boldsymbol{f})} = \ln p(\boldsymbol{f} = \widetilde{\boldsymbol{f}}, \boldsymbol{\theta}) \\ < \ln p(\boldsymbol{f}, \boldsymbol{\theta}) >_{q_2(\boldsymbol{\theta})} = < \ln p(\boldsymbol{f}, \boldsymbol{\theta}) >_{p(\boldsymbol{\theta}|\boldsymbol{f}=\widetilde{\boldsymbol{f}})} = Q(\widetilde{\boldsymbol{f}}, \boldsymbol{f}) \end{array}\right. \tag{123}$$

$$\left\{\begin{array}{l} q_2(\theta) \propto \ln p(\boldsymbol{f} = \widetilde{\boldsymbol{f}}, \boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{f} = \widetilde{\boldsymbol{f}}) \\ q_1(\boldsymbol{f}) \propto \exp\left[Q(\widetilde{\boldsymbol{f}}, \boldsymbol{\theta})\right] \longrightarrow \widetilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \left\{Q(\boldsymbol{f} = \widetilde{\boldsymbol{f}}, \boldsymbol{\theta})\right\} \end{array}\right. \tag{124}$$

  which can be compared with the classical EM algorithm. Here, the uncertainties of the $\boldsymbol{f}$ are used for the estimation of $\boldsymbol{\theta}$, but the uncertainties of $\boldsymbol{\theta}$ are not used for the estimation of $\boldsymbol{f}$.

- Both $q_1(\boldsymbol{f})$ and $q_2(\boldsymbol{\theta})$ have free form. The main difficulty here is that, at each iteration, the expression of $q_1$ and $q_2$ may change. However, if $p(\boldsymbol{f}, \boldsymbol{\theta})$ is in the generalized exponential family, the expressions of $q_1(\boldsymbol{f})$ and $q_2(\boldsymbol{\theta})$ will also be in the same family, and we have only to update the parameters at each iteration.

## 17. VBA for a Linear Inverse Problem with Simple Gaussian Priors

As a simple example, consider the Gaussian case where $p(\boldsymbol{g}|\boldsymbol{f}, \theta_1) = \mathcal{N}(\boldsymbol{g}|\boldsymbol{H}\boldsymbol{f}, (1/\theta_1)\boldsymbol{I})$, $p(\boldsymbol{f}|\theta_2) = \mathcal{N}(\boldsymbol{f}|\boldsymbol{0}, (1/\theta_2)\boldsymbol{I})$ and $p(\theta_1) = \mathcal{G}(\theta_1|\alpha_{10}, \beta_{10})$ $p(\theta_2) = \mathcal{G}(\theta_2|\alpha_{20}, \beta_{20})$, and so, we have:

$$\begin{aligned} \ln p(\boldsymbol{f}, \theta_1, \theta_2|\boldsymbol{g}) = & \frac{M}{2}\ln\theta_1 - \frac{\theta_1}{2}\|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\|_2^2 + \frac{N}{2}\ln\theta_2 - \frac{\theta_2}{2}\|\boldsymbol{f}\|_2^2 \\ & + (\alpha_{10} - 1)\ln\theta_1 - \beta_{10}\theta_1 + (\alpha_{20} - 1)\ln\theta_2 - \beta_{20}\theta_2. \end{aligned} \tag{125}$$

From this expression $J(\boldsymbol{f},\theta_1,\theta_2)=\ln p(\boldsymbol{f},\theta_1,\theta_2|\boldsymbol{g})$, it is easy to obtain the equations of an alternate JMAP algorithm by computing the derivatives of it with respective to its arguments and equating them to zero:

$$\begin{aligned}
&\frac{\partial J}{\partial \boldsymbol{f}}=0 \longrightarrow \boldsymbol{f}=(\boldsymbol{H}'\boldsymbol{H}+\lambda\boldsymbol{I})^{-1}\boldsymbol{H}'\boldsymbol{g} \text{ with } \lambda=\frac{\theta_2}{\theta_1}\\
&\frac{\partial J}{\partial \theta_1}=0 \longrightarrow \theta_1=\frac{\widetilde{\alpha}_1}{\widetilde{\beta}_1} \text{ with } \widetilde{\alpha}_1=(\alpha_{10}-1)+\frac{M}{2} \text{ and } \widetilde{\beta}_1=\beta_{10}+\frac{1}{2}\|\boldsymbol{g}-\boldsymbol{H}\boldsymbol{f}\|_2^2\\
&\frac{\partial J}{\partial \theta_2}=0 \longrightarrow \theta_1=\frac{\widetilde{\alpha}_2}{\widetilde{\beta}_2} \text{ with } \widetilde{\alpha}_2=(\alpha_{20}-1)+\frac{M}{2} \text{ and } \widetilde{\beta}_2=\beta_{20}+\frac{1}{2}\|\boldsymbol{f}\|_2^2
\end{aligned} \tag{126}$$

From the expression of the joint probability law $p(\boldsymbol{f},\theta_1,\theta_2|\boldsymbol{g})$, we can also obtain the expressions of the conditionals:

$$\left\{\begin{array}{l}
p(\boldsymbol{f}|\boldsymbol{g},\theta_1,\theta_2)=\mathcal{N}(\boldsymbol{f}|\widetilde{\boldsymbol{f}},\widetilde{\boldsymbol{V}})\\
\quad \text{with} \quad \widetilde{\boldsymbol{V}}=(\boldsymbol{H}'\boldsymbol{H}+\lambda\boldsymbol{I})^{-1}, \quad \widetilde{\boldsymbol{f}}=\widetilde{\boldsymbol{V}}\boldsymbol{H}'\boldsymbol{g}, \quad \lambda=\frac{\theta_2}{\theta_1}\\
p(\theta_1|\boldsymbol{g},\boldsymbol{f},\theta_2)=\mathcal{G}(\theta_1|\widetilde{\alpha}_1,\widetilde{\beta}_1)\\
\quad \text{with} \quad \widetilde{\alpha}_1=(\alpha_{10}-1)+\frac{M}{2}, \quad \widetilde{\beta}_1=\beta_{10}+\frac{1}{2}\|\boldsymbol{g}-\boldsymbol{H}\boldsymbol{f}\|_2^2\\
p(\theta_2|\boldsymbol{g},\boldsymbol{f},\theta_1)=\mathcal{G}(\theta_2|\widetilde{\alpha}_2,\widetilde{\beta}_2)\\
\quad \text{with} \quad \widetilde{\alpha}_2=(\alpha_{20}-1)+\frac{M}{2}, \quad \widetilde{\beta}_2=\beta_{20}+\frac{1}{2}\|\boldsymbol{f}\|_2^2
\end{array}\right. \tag{127}$$

However, obtaining analytical expressions of the marginals $p(\boldsymbol{f}|\boldsymbol{g})$, $p(\theta_1|\boldsymbol{g})$ and $p(\theta_2|\boldsymbol{g})$ is not easy. We can then obtain approximate expressions $q_1(\boldsymbol{f}|\boldsymbol{g})$, $q_2(\theta_1|\boldsymbol{g})$ and $q_3(\theta_2|\boldsymbol{g})$ using the VBA method. For this case, thanks to the conjugacy property, we have:

$$\left\{\begin{array}{l}
q(\boldsymbol{f})=\mathcal{N}(\boldsymbol{f}|\widetilde{\boldsymbol{f}},\widetilde{\boldsymbol{V}})\\
\quad \text{with} \quad \widetilde{\boldsymbol{V}}=(\boldsymbol{H}'\boldsymbol{H}+\widetilde{\lambda}\boldsymbol{I})^{-1}, \quad \widetilde{\boldsymbol{f}}=\widetilde{\boldsymbol{V}}\boldsymbol{H}'\boldsymbol{g}, \quad \widetilde{\lambda}=\frac{<\theta_2>}{<\theta_1>};\\
q(\theta_1)=\mathcal{G}(\theta_1|\widetilde{\alpha}_1,\widetilde{\beta}_1)\\
\quad \text{with} \quad \widetilde{\alpha}_1=(\alpha_{10}-1)+\frac{M}{2}, \quad \widetilde{\beta}_1=\beta_{10}+\frac{1}{2}<\|\boldsymbol{g}-\boldsymbol{H}\boldsymbol{f}\|_2^2>\\
p(\theta_2|\boldsymbol{g},\boldsymbol{f})=\mathcal{G}(\theta_2|\widetilde{\alpha}_2,\widetilde{\beta}_2)\\
\quad \text{with} \quad \widetilde{\alpha}_2=(\alpha_{20}-1)+\frac{N}{2}, \quad \widetilde{\beta}_2=\beta_{20}+\frac{1}{2}<\|\boldsymbol{f}\|_2^2>
\end{array}\right. \tag{128}$$

We can then compare the three algorithms in Table 2:

**Table 2.** Comparision of three algorithms: JMAP, BEM and VBA

| JMAP | BEM | VBA |
|---|---|---|
| $q(\boldsymbol{f}) = \delta(\boldsymbol{f} - \widetilde{\boldsymbol{f}})$ | $q(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{f}\|\widetilde{\boldsymbol{f}}, \widetilde{\boldsymbol{V}})$ | $q(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{f}\|\widetilde{\boldsymbol{f}}, \widetilde{\boldsymbol{V}})$ |
| $\widetilde{\boldsymbol{V}} = (\boldsymbol{H}'\boldsymbol{H} + \widetilde{\lambda}\boldsymbol{I})^{-1}$ | $\widetilde{\boldsymbol{V}} = (\boldsymbol{H}'\boldsymbol{H} + \widetilde{\lambda}\boldsymbol{I})^{-1}$ | $\widetilde{\boldsymbol{V}} = (\boldsymbol{H}'\boldsymbol{H} + \widetilde{\lambda}\boldsymbol{I})^{-1}$ |
| $\widetilde{\boldsymbol{f}} = \widetilde{\boldsymbol{V}}\boldsymbol{H}'\boldsymbol{g}$ | $\widetilde{\boldsymbol{f}} = \widetilde{\boldsymbol{V}}\boldsymbol{H}'\boldsymbol{g}$ | $\widetilde{\boldsymbol{f}} = \widetilde{\boldsymbol{V}}\boldsymbol{H}'\boldsymbol{g}$ |
| $q(\theta_1) = \delta(\theta_1 - \widetilde{\theta}_1)$ | $q(\theta_1) = \delta(\theta_1 - \widetilde{\theta}_1)$ | $q(\theta_1) = \mathcal{G}(\theta_1\|\widetilde{\alpha}_1, \widetilde{\beta}_1)$ |
| $\widetilde{\alpha}_1 = (\alpha_{10} - 1) + \frac{M}{2}$ | $\widetilde{\alpha}_1 = (\alpha_{10} - 1) + \frac{M}{2}$ | $\widetilde{\alpha}_1 = (\alpha_{10} - 1) + \frac{M}{2}$ |
| $\widetilde{\beta}_1 = \beta_{10} + \frac{1}{2}\|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\|_2^2$ | $\widetilde{\beta}_1 = \beta_{10} + \frac{1}{2} < \|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\|_2^2 >$ | $\widetilde{\beta}_1 = \beta_{10} + \frac{1}{2} < \|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\|_2^2 >$ |
| $\widetilde{\theta}_1 = \frac{\widetilde{\alpha}_1}{\widetilde{\beta}_1}$ | $\widetilde{\theta}_1 = \frac{\widetilde{\alpha}_1}{\widetilde{\beta}_1}$ | $\widetilde{\theta}_1 = \frac{\widetilde{\alpha}_1}{\widetilde{\beta}_1}$ |
| $q(\theta_2) = \delta(\theta_2 - \widetilde{\theta}_2)$ | $q(\theta_2) = \delta(\theta_2 - \widetilde{\theta}_2)$ | $q(\theta_2) = \mathcal{G}(\theta_2\|\widetilde{\alpha}_2, \widetilde{\beta}_2)$ |
| $\widetilde{\alpha}_2 = (\alpha_{20} - 1) + \frac{M}{2}$ | $\widetilde{\alpha}_2 = (\alpha_{20} - 1) + \frac{M}{2}$ | $\widetilde{\alpha}_2 = (\alpha_{20} - 1) + \frac{N}{2}$ |
| $\widetilde{\beta}_2 = \beta_{10} + \frac{1}{2}\|\boldsymbol{f}\|_2^2$ | $\widetilde{\beta}_2 = \beta_{20} + \frac{1}{2} < \|\boldsymbol{f}\|_2^2 >$ | $\widetilde{\beta}_2 = \beta_{20} + \frac{1}{2} < \|\boldsymbol{f}\|_2^2 >$ |
| $\widetilde{\theta}_2 = \frac{\widetilde{\alpha}_2}{\widetilde{\beta}_2}$ | $\widetilde{\theta}_2 = \frac{\widetilde{\alpha}_2}{\widetilde{\beta}_2}$ | $\widetilde{\theta}_2 = \frac{\widetilde{\alpha}_2}{\widetilde{\beta}_2}$ |
| $\widetilde{\lambda} = \frac{\widetilde{\theta}_2}{\widetilde{\theta}_1}$ | $\widetilde{\lambda} = \frac{\widetilde{\theta}_2}{\widetilde{\theta}_1}$ | $\widetilde{\lambda} = \frac{\widetilde{\theta}_2}{\widetilde{\theta}_1}$ |

It is important to remark that, in JMAP, the computation of $\boldsymbol{f}$ can be done via the optimization of the criterion $J(\boldsymbol{f}, \theta_1, \theta_2) = \ln p(\boldsymbol{f}, \theta_1, \theta_2|\boldsymbol{g})$, which does not need explicitly the matrix inversion of $\widetilde{\boldsymbol{V}} = (\boldsymbol{H}'\boldsymbol{H} + \widetilde{\lambda}\boldsymbol{I})^{-1}$. However, in BEM and VBA, we need to compute it due to the following requirements:

$$
\begin{aligned}
< \boldsymbol{f} >_q &= \widetilde{\boldsymbol{f}}, \\
< \|\boldsymbol{f}\|^2 >_q &= \text{tr}\left( < \widetilde{\boldsymbol{f}}\widetilde{\boldsymbol{f}}' >_q \right) = \text{tr}\left( \widetilde{\boldsymbol{f}}\widetilde{\boldsymbol{f}}' + \widetilde{\boldsymbol{V}} \right) = \|\widetilde{\boldsymbol{f}}\|^2 + \text{tr}\left( \widetilde{\boldsymbol{V}} \right), \\
< f_j^2 >_q &= [\widetilde{\boldsymbol{V}}]_{jj} + \widetilde{f}_j^2, \\
< \|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\|^2 >_q &= [\boldsymbol{g}'\boldsymbol{g} - 2\left\langle \boldsymbol{f} \right\rangle_q' \boldsymbol{H}'\boldsymbol{g} + \boldsymbol{H}' \left\langle \boldsymbol{f}'\boldsymbol{f} \right\rangle_q \boldsymbol{H}] \\
&= [\boldsymbol{g}'\boldsymbol{g} - 2\widetilde{\boldsymbol{f}}'\boldsymbol{H}'\boldsymbol{g} + \boldsymbol{H}'(\widetilde{\boldsymbol{V}} + \widetilde{\boldsymbol{f}}\widetilde{\boldsymbol{f}}')\boldsymbol{H}] \\
&= \|\boldsymbol{g} - \boldsymbol{H}\widetilde{\boldsymbol{f}}\|^2 + \text{tr}\left( \boldsymbol{H}'\widetilde{\boldsymbol{V}}\boldsymbol{H} \right)]
\end{aligned}
\tag{129}
$$

For some extensions and more details, see [111].

## 18. Bayesian Variational Approximation with Hierarchical Prior Models

For a linear inverse problem:

$$
\mathcal{M} : \boldsymbol{g} = \boldsymbol{H}\boldsymbol{f} + \boldsymbol{\epsilon} \tag{130}
$$

with an assigned likelihood $p(\boldsymbol{g}|\boldsymbol{f},\boldsymbol{\theta}_1)$, when a hierarchical prior model $p(\boldsymbol{f}|\boldsymbol{z},\boldsymbol{\theta}_2)\,p(\boldsymbol{z}|\boldsymbol{\theta}_3)$ is used and when the estimation of the hyper-parameters $\boldsymbol{\theta} = [\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\boldsymbol{\theta}_3]'$ has to be considered, the joint posterior law of all the unknowns becomes:

$$p(\boldsymbol{f},\boldsymbol{z},\boldsymbol{\theta}|\boldsymbol{g}) = \frac{p(\boldsymbol{f},\boldsymbol{z},\boldsymbol{\theta}|\boldsymbol{g})}{p(\boldsymbol{g})} = \frac{p(\boldsymbol{g}|\boldsymbol{f},\boldsymbol{\theta}_1)\,p(\boldsymbol{f}|\boldsymbol{z},\boldsymbol{\theta}_2)\,p(\boldsymbol{z}|\boldsymbol{\theta}_3)\,p(\boldsymbol{\theta})}{p(\boldsymbol{g})}. \tag{131}$$

The main idea behind the VBA is to approximate this joint posterior by a separable one, for example: $q(\boldsymbol{f},\boldsymbol{z},\boldsymbol{\theta}|\boldsymbol{g}) = q_1(\boldsymbol{f})\,q_2(\boldsymbol{z})\,q_3(\boldsymbol{\theta})$ and where the expressions of $q(\boldsymbol{f},\boldsymbol{z},\boldsymbol{\theta}|\boldsymbol{g})$ are obtained by minimizing the Kullback–Leibler divergence (99), as explained in previous section. This approach can also be used for model selection based on the evidence of the model $\ln p(\boldsymbol{g})$ [121] where:

$$p(\boldsymbol{g}) = \int\int\int p(\boldsymbol{f},\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{g})\,\mathrm{d}\boldsymbol{f}\,\mathrm{d}\boldsymbol{z}\,\mathrm{d}\boldsymbol{\theta}. \tag{132}$$

Interestingly, it is easy to show that:

$$\ln p(\boldsymbol{g}) = \mathrm{KL}\left[q:p\right] + \mathcal{F}(q) \tag{133}$$

where $\mathcal{F}(q)$ is the free energy associated with $q$ defined as:

$$\mathcal{F}(q) = \left\langle \ln \frac{p(\boldsymbol{f},\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{g})}{q(\boldsymbol{f},\boldsymbol{z},\boldsymbol{\theta})} \right\rangle_q \tag{134}$$

Therefore, for a given model $\mathcal{M}$, minimizing $\mathrm{KL}\left[q:p\right]$ is equivalent to maximizing $\mathcal{F}(q)$ and when optimized, $\mathcal{F}(q^*)$ gives a lower bound for $\ln p(\boldsymbol{g})$. Indeed, the name variational approximation is due to the fact that $\ln p(\boldsymbol{g}) \geq \mathcal{F}(q)$, and so, $\mathcal{F}(q)$ is a lower bound to the evidence $\ln p(\boldsymbol{g})$.

Without any other constraint than the normalization of $q$, an alternate optimization of $\mathcal{F}(q)$ with respect to $q_1$, $q_2$ and $q_3$ results in:

$$\left\{\begin{array}{l} q_1(\boldsymbol{f}) \propto \exp\left[-\left\langle \ln p(\boldsymbol{f},\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{g})\right\rangle_{q(\boldsymbol{z})q(\boldsymbol{\theta})}\right], \\ q_2(\boldsymbol{z}) \propto \exp\left[-\left\langle \ln p(\boldsymbol{f},\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{g})\right\rangle_{q(\boldsymbol{f})q(\boldsymbol{\theta})}\right], \\ q_3(\boldsymbol{\theta}) \propto \exp\left[-\left\langle \ln p(\boldsymbol{f},\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{g})\right\rangle_{q(\boldsymbol{f})q(\boldsymbol{z})}\right]. \end{array}\right. \tag{135}$$

Note that these relations represent an implicit solution for $q_1(\boldsymbol{f})$, $q_2(\boldsymbol{z})$ and $q_3(\boldsymbol{\theta})$, which need, at each iteration, the expression of the expectations in the right hand of exponentials. If $p(\boldsymbol{g}|\boldsymbol{f},\boldsymbol{z},\boldsymbol{\theta}_1)$ is a member of an exponential family and if all of the priors $p(\boldsymbol{f}|\boldsymbol{z},\boldsymbol{\theta}_2)$, $p(\boldsymbol{z}|\boldsymbol{\theta}_3)$, $p(\boldsymbol{\theta}_1)$, $p(\boldsymbol{\theta}_2)$ and $p(\boldsymbol{\theta}_3)$ are conjugate priors, then it is easy to see that these expressions lead to standard distributions for which the required expectations are easily evaluated. In that case, we may note:

$$q(\boldsymbol{f},\boldsymbol{z},\boldsymbol{\theta}) = q_1(\boldsymbol{f}|\widetilde{\boldsymbol{z}},\widetilde{\boldsymbol{\theta}})\,q_2(\boldsymbol{z}|\widetilde{\boldsymbol{f}},\widetilde{\boldsymbol{\theta}})\,q_3(\boldsymbol{\theta}|\widetilde{\boldsymbol{f}},\widetilde{\boldsymbol{z}}) \tag{136}$$

where the tilded quantities $\widetilde{\boldsymbol{z}}$, $\widetilde{\boldsymbol{f}}$ and $\widetilde{\boldsymbol{\theta}}$ are, respectively, functions of $(\widetilde{\boldsymbol{f}},\widetilde{\boldsymbol{\theta}})$, $(\widetilde{\boldsymbol{z}},\widetilde{\boldsymbol{\theta}})$ and $(\widetilde{\boldsymbol{f}},\widetilde{\boldsymbol{z}})$. This means that the expression of $q_1(\boldsymbol{f}|\widetilde{\boldsymbol{z}},\widetilde{\boldsymbol{\theta}})$ depends on $(\widetilde{\boldsymbol{f}},\widetilde{\boldsymbol{\theta}})$, the expression of $q_2(\boldsymbol{z}|\widetilde{\boldsymbol{f}},\widetilde{\boldsymbol{\theta}})$ depends on $(\widetilde{\boldsymbol{z}},\widetilde{\boldsymbol{\theta}})$ and the expression of $q_3(\boldsymbol{\theta}|\widetilde{\boldsymbol{f}},\widetilde{\boldsymbol{z}})$ depends on $(\widetilde{\boldsymbol{f}},\widetilde{\boldsymbol{z}})$. With this notation, the alternate

optimization results in alternate updating of the parameters $(\widetilde{\boldsymbol{z}}, \widetilde{\boldsymbol{\theta}})$ of $q_1$, the parameters $(\widetilde{\boldsymbol{f}}, \widetilde{\boldsymbol{\theta}})$ of $q_2$ and the parameters $(\widetilde{\boldsymbol{f}}, \widetilde{\boldsymbol{z}})$ of $q_3$. Finally, we may note that, to monitor the convergence of the algorithm, we may evaluate the free energy:

$$\begin{aligned}\mathcal{F}(q) &= \left\langle \ln p(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{g}) \right\rangle_q - \left\langle \ln q(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}) \right\rangle_q \\ &= \left\langle \ln p(\boldsymbol{g}|\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}) \right\rangle_q + \left\langle \ln p(\boldsymbol{f}|\boldsymbol{z}, \boldsymbol{\theta}) \right\rangle_q + \left\langle \ln p(\boldsymbol{z}|\boldsymbol{\theta}) \right\rangle_q + \left\langle \ln p(\boldsymbol{\theta}) \right\rangle_q \\ &\quad - \left\langle \ln q(\boldsymbol{f}) \right\rangle_q - \left\langle \ln q(\boldsymbol{z}) \right\rangle_q - \left\langle \ln q(\boldsymbol{\theta}) \right\rangle_q .\end{aligned} \tag{137}$$

Other decompositions for $q(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta})$ are also possible. For example: $q(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}) = q_1(\boldsymbol{f}|\boldsymbol{z})\, q_2(\boldsymbol{z})\, q_3(\boldsymbol{\theta})$ or even: $q(\boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\theta}) = \prod_j q_{1j}(f_j) \prod_j q_{2j}(z_{f_j}) \prod_l q_{3l}(\theta_l)$. Here, we consider the first case and give some more details on it.

## 19. Bayesian Variational Approximation with Student *t* Priors

The Student *t* model is:

$$p(\boldsymbol{f}|\nu) = \prod_j \mathcal{S}t(f_j|\nu) \text{ with } \mathcal{S}t(f_j|\nu) = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \left(1 + f_j^2/\nu\right)^{-(\nu+1)/2} \tag{138}$$

The Cauchy model is obtained when $\nu = 1$. Knowing that:

$$\mathcal{S}t(f_j|\nu) = \int_0^\infty \mathcal{N}(f_j|0, 1/z_{f_j})\, \mathcal{G}(z_{f_j}|\nu/2, \nu/2)\, \mathrm{d}z_{f_j} \tag{139}$$

we can write this model via the positive hidden variables $z_{f_j}$:

$$\left\{ \begin{array}{ll} p(f_j|z_{f_j}) & = \mathcal{N}(f_j|0, 1/z_{f_j}) \propto \exp\left[-\frac{1}{2} z_{f_j} f_j^2\right] \\ p(z_{f_j}|\alpha, \beta) & = \mathcal{G}(z_{f_j}|\alpha, \beta) \propto z_{f_j}^{(\alpha-1)} \exp\left[-\beta z_{f_j}\right] \end{array} \right. \tag{140}$$

Now, let us consider the forward model $\boldsymbol{g} = \boldsymbol{H}\boldsymbol{f} + \boldsymbol{\epsilon}$ and assign a Gaussian law with unknown variance $v_{\epsilon_i}$ to the noise $\epsilon_i$, which results in $p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{g}|\boldsymbol{0}, \boldsymbol{V}_\epsilon)$ with $\boldsymbol{V}_\epsilon = \mathrm{diag}\,[\boldsymbol{v}_\epsilon]$ with $\boldsymbol{v}_\epsilon = [v_{\epsilon_1}, \cdots, v_{\epsilon_M}]$, and so:

$$p(\boldsymbol{g}|\boldsymbol{f}, \boldsymbol{v}_\epsilon) = \mathcal{N}(\boldsymbol{g}|\boldsymbol{H}\boldsymbol{f}, \boldsymbol{V}_\epsilon) \propto \exp\left[-\frac{1}{2}(\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f})\boldsymbol{V}_\epsilon^{-1}(\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f})\right]. \tag{141}$$

Let us also note by $z_{\epsilon_i} = 1/v_{\epsilon_i}$, $\boldsymbol{z}_\epsilon = [z_{\epsilon_1}, \cdots, z_{\epsilon_M}]$ and $\boldsymbol{Z}_\epsilon = \mathrm{diag}\,[\boldsymbol{z}_\epsilon] = \boldsymbol{V}_\epsilon^{-1}$ and assign a prior on it $p(v_{\epsilon_i}|\alpha_{\epsilon_0}, \beta_{\epsilon_0}) = \mathcal{IG}(v_{\epsilon_i}|\alpha_{\epsilon_0}, \beta_{\epsilon_0})$ or equivalently:

$$p(z_{\epsilon_i}|\alpha_{\epsilon_0}, \beta_{\epsilon_0}) = \mathcal{G}(z_{\epsilon_i}|\alpha_{\epsilon_0}, \beta_{\epsilon_0}) \quad \text{and} \quad p(\boldsymbol{z}_\epsilon|\alpha_{\epsilon_0}, \beta_{\epsilon_0}) = \prod_i \mathcal{G}(z_{\epsilon_i}|\alpha_{\epsilon_0}, \beta_{\epsilon_0}). \tag{142}$$

Let us also note $\boldsymbol{v}_f = [v_{f_1}, \cdots, v_{f_N}]$, $\boldsymbol{V}_f = \mathrm{diag}\,[\boldsymbol{v}_f]$, $z_{f_j} = 1/v_{f_j}$, $\boldsymbol{Z}_f = \mathrm{diag}\,[\boldsymbol{z}_f] = \boldsymbol{V}_f^{-1}$ and note:

$$p(\boldsymbol{f}|\boldsymbol{v}_f) = \prod_j p(f_j|v_{f_j}) = \prod_j \mathcal{N}\left(f_j|0, v_{f_j}\right) = \mathcal{N}(\boldsymbol{f}|\boldsymbol{0}, \boldsymbol{V}_f) \tag{143}$$

and finally,

$$p(\boldsymbol{v}_f|\alpha_{f_0},\beta_{f_0}) = \prod_j \mathcal{G}(v_{f_j}|\alpha_{f_0},\beta_{f_0}). \tag{144}$$

Then, we obtain the following expressions for the VBA:

$$\left\{\begin{array}{l} q_1(\boldsymbol{f}|\widetilde{\boldsymbol{\mu}},\widetilde{\boldsymbol{V}}_f) = \mathcal{N}(\boldsymbol{f}|\widetilde{\boldsymbol{\mu}},\widetilde{\boldsymbol{V}}) \text{ with } \widetilde{\boldsymbol{\mu}} = \widetilde{\boldsymbol{V}}\boldsymbol{H}'\boldsymbol{g},\ \widetilde{\boldsymbol{V}} = (\boldsymbol{H}'\widetilde{\boldsymbol{V}}_\epsilon^{-1}\boldsymbol{H} + \widetilde{\boldsymbol{Z}}_f)^{-1}; \\ q_{2j}(z_{f_j}) = \mathcal{G}(z_{f_j}|\widetilde{\alpha}_j,\widetilde{\beta}_j) \text{ with } \widetilde{\alpha}_j = \alpha_{00} + 1/2,\ \widetilde{\beta}_j = \beta_{00} + < f_j^2 > /2; \\ q_3(z_{\epsilon_i}) = \mathcal{G}(z_{\epsilon_i}|\widetilde{\alpha}_{\epsilon_i},\widetilde{\beta}_{\epsilon_i}) \text{ with } \widetilde{\alpha}_{\epsilon_i} = \alpha_{\epsilon_0} + (N+1)/2,\ \widetilde{\beta}_{\epsilon_i} = \beta_{\epsilon 0} + \frac{1}{2} < |g_i - [\boldsymbol{H}\boldsymbol{f}]_i|^2 >; \end{array}\right. \tag{145}$$

where:

$$\begin{aligned} &< |g_i - [\boldsymbol{H}\boldsymbol{f}]_i|^2 >= |g_i - \boldsymbol{H} < \boldsymbol{f} >]_i|^2 + [\boldsymbol{H}'\widetilde{\boldsymbol{V}}\boldsymbol{H}]_{ii}, \\ &< \boldsymbol{f} >= \widetilde{\boldsymbol{\mu}}, \quad < \boldsymbol{f}\boldsymbol{f}' >= \widetilde{\boldsymbol{V}} + \widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}', \\ &< f_j^2 >= [\widetilde{\boldsymbol{V}}]_{jj} + \widetilde{\mu}_j^2 \end{aligned}$$

We have implemented these algorithms for many linear inverse problems [102], such as periodic components estimation in time series [122] or computed tomography [123], blind deconvolution [124], blind image separation [125,126] and blind image restoration [89].

## 20. Conclusions

The main conclusions of this paper can be summarized as follows:

- A probability law is a tool for representing our state of knowledge about a quantity.
- The Bayes or Laplace rule is an inference tool for updating our state of knowledge about an inaccessible quantity when another accessible, related quantity is observed.
- Entropy is a measure of information content in a variable with a given probability law.
- The maximum entropy principle can be used to assign a probability law to a quantity when the available information about it is in the form of a limited number of constraints on that probability law.
- Relative entropy and Kullback–Leibler divergence are tools for updating probability laws in the same context.
- When a parametric probability law is assigned to a quantity and we want to measure the amount of information gain about the parameters when some direct observations of that quantity is available, we can use the Fisher information. The structure of the Fisher information geometry in the space of parameters is derived from the relative entropy by a second order Taylor series approximation.
- All of these rules and tools are used currently in different ways in data and signal processing. In this paper, a few examples of the ways these tools are used in data and signal processing

problems are presented. One main conclusion is that each of these tools has to be used in appropriate contexts. The example in spectral estimation shows that it is very important to define the problems very clearly at the beginning and to use appropriate tools and interpret the results appropriately.

- The Laplacian or Bayesian inference is the appropriate tool for proposing satisfactory solutions to inverse problems. Indeed, the expression of the posterior probability law represents the combination of the state of the knowledge in the forward model and the data and the state of the knowledge before using the data.
- The Bayesian approach can also easily be used to propose unsupervised methods for the practical application of these methods.
- One of the main limitation of those sophisticated methods is the computational cost. For this, we proposed to use VBA as an alternative to MCMC methods to propose realistic algorithms in huge dimensional inverse problems where we want to estimate an unknown signal (1D), image (2D), volume (3D) or even more (3D + time or 3D + wavelength), *etc*.

**Acknowledgments**

The author would like to thank the reviewers who, by their true review work and their extensive comments and remarks, helped to improve this review paper greatly.

**Conflicts of Interest**

The author declares no conflict of interest.

**References**

1. Mohammad-Djafari, A. Bayesian or Laplacian inference, entropy and information theory and information geometry in data and signal processing. *AIP Conf. Proc.* **2014**, *1641*, 43–58.
2. Bayes, T. An Essay toward Solving a Problem in the Doctrine of Chances. *Philos. Trans.* **1763**, *53*, 370–418. By the late Rev. Mr. Bayes communicated by Mr. Price, in a Letter to John Canton.
3. De Laplace, P. S. Mémoire sur la probabilité des causes par les évènements. *Mémoires de l'Academie Royale des Sciences Presentés par Divers Savan* **1774**, *6*, 621–656.
4. Shannon, C. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.
5. Hadamard, J. *Mémoire sur le problème d'analyse relatif à l'équilibre des plaques élastiques encastrées*; Mémoires présentés par divers savants à l'Académie des sciences de l'Institut de France: Imprimerie nationale, 1908.
6. Jaynes, E.T. Information Theory and Statistical Mechanics. *Phys. Rev.* **1957**, *106*, 620–630.
7. Jaynes, E.T. Information Theory and Statistical Mechanics II. *Phys. Rev.* **1957**, *108*, 171–190.

8. Jaynes, E.T. Prior Probabilities. *IEEE Trans. Syst. Sci. Cybern.* **1968**, *4*, 227–241.
9. Kullback, S. *Information Theory and Statistics*; Wiley: New York, NY, USA, 1959.
10. Fisher, R. On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Stat. Soc. A* **1922**, *222*, 309–368.
11. Rao, C. Information and accuracy attainable in the estimation of statistical parameters. *Bull. Culcutta Math. Soc.* **1945**, *37*, 81–91.
12. Sindhwani, V.; Belkin, M.; Niyogi, P. The Geometric basis for Semi-supervised Learning. In *Semi-supervised Learning*; Chapelle, O., Schölkopf, B., Zien, A., Eds.; MIT press: Cambridge, MA, USA, 2006; pp. 209–226.
13. Lin, J. Divergence Measures Based on the Shannon Entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151.
14. Johnson, O.; Barron, A.R. Fisher Information Inequalities and the Central Limit Theorem. *Probab. Theory Relat. Fields* **2004**, *129*, 391–409.
15. Berger, J. *Statistical Decision Theory and Bayesian Analysis*, 2nd ed.; Springer-Verlag: New York, NY, USA, 1985.
16. Gelman, A.; Carlin, J.B.; Stern, H.S.; Rubin, D.B. *Bayesian Data Analysis*, 2nd ed.; Chapman & Hall/CRC Texts in Statistical Science; Chapman and Hall/CRC: Boca Raton, FL, USA, 2003.
17. Skilling, J. Nested Sampling. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, Proceedings of 24th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Garching, Germany, 25–30 July 2004; Fischer, R., Preuss, R., Toussaint, U.V., Eds.; pp. 395–405.
18. Metropolis, N.; Rosenbluth, A.W.; Rosenbluth, M.N.; Teller, A.H.; Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**, *21*, 1087–1092.
19. Hastings, W.K. Monte Carlo Sampling Methods using Markov Chains and their Applications. *Biometrika* **1970**, *57*, 97–109.
20. Gelfand, A.E.; Smith, A.F.M. Sampling-Based Approaches to Calculating Marginal Densities. *J. Am. Stat. Assoc.* **1990**, *85*, 398–409.
21. Gilks, W.R.; Richardson, S.; Spiegelhalter, D.J. Introducing Markov Chain Monte Carlo. In *Markov Chain Monte Carlo in Practice*; Gilks, W.R., Richardson, S., Spiegelhalter, D.J., Eds.; Chapman and Hall: London, UK, 1996; pp. 1–19.
22. Gilks, W.R. Strategies for Improving MCMC. In *Markov Chain Monte Carlo in Practice*; Gilks, W.R., Richardson, S., Spiegelhalter, D.J., Eds.; Chapman and Hall: London, UK, 1996; pp. 89–114.
23. Roberts, G.O. Markov Chain Concepts Related to Sampling Algorithms. In *Markov Chain Monte Carlo in Practice*; Gilks, W.R., Richardson, S., Spiegelhalter, D.J., Eds.; Chapman and Hall: London, UK, 1996; pp. 45–57.
24. Tanner, M.A. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*; Springer series in Statistics; Springer: New York, NY, USA, 1996.

25. Djurić, P.M., Godsill, S.J., Eds. *Special Issue on Monte Carlo Methods for Statistical Signal Processing*; IEEE: New York, NY, USA, 2002.
26. Andrieu, C.; de Freitas, N.; Doucet, A.; Jordan, M.I. An Introduction to MCMC for Machine Learning. *Mach. Learn.* **2003**, *50*, 5–43.
27. Clausius, R. *On the Motive Power of Heat, and on the Laws Which Can be Deduced From it for the Theory of Heat*; Poggendorff's Annalen der Physick,LXXIX, Dover Reprint: New York, NY, USA, 1850; ISBN 0-486-59065-8.
28. Caticha, A. Maximum Entropy, fluctuations and priors. Presented at MaxEnt 2000, the 20th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Gif-sur-Yvette, Paris, France, 8–13 July 2000.
29. Giffin, A.; Caticha, A. Updating Probabilities with Data and Moments. In Proceedings of the 27th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, The Saratoga Hotel Saratoga Springs, New York, NY, USA, 8–13 July 2007.
30. Caticha, A.; Preuss, R. Maximum Entropy and Bayesian Data Analysis: Entropic Priors Distributions. *Phys. Rev. E* **2004**, *70*, 046127.
31. Akaike, H. On Entropy Maximization Principle. In *Applications of Statistics*; Krishnaiah, P.R., Ed.; North-Holland: Amsterdam, The Netherlands, 1977; pp. 27–41.
32. Agmon, N.; Alhassid, Y.; Levine, D. An Algorithm for Finding the Distribution of Maximal Entropy. *J. Comput. Phys.* **1979**, *30*, 250–258.
33. Jaynes, E.T. Where do we go from here? In *Maximum-Entropy and Bayesian Methods in Inverse Problems*; Smith, C.R., Grandy, W.T., Jr., Eds.; Springer: Dordrecht, The Netherlands, 1985; pp. 21–58.
34. Borwein, J.M.; Lewis, A.S. Duality relationships for entropy-like minimization problems. *SIAM J. Control Optim.* **1991**, *29*, 325–338.
35. Elfwing, T. On some Methods for Entropy Maximization and Matrix Scaling. *Linear Algebra Appl.* **1980**, *34*, 321–339.
36. Eriksson, J. A note on Solution of Large Sparse Maximum Entropy Problems with Linear Equality Constraints. *Math. Program.* **1980**, *18*, 146–154.
37. Erlander, S. Entropy in linear programs. *Math. Program.* **1981**, *21*, 137–151.
38. Jaynes, E.T. On the Rationale of Maximum-Entropy Methods. *Proc. IEEE* **1982**, *70*, 939–952.
39. Shore, J.E.; Johnson, R.W. Properties of Cross-Entropy Minimization. *IEEE Trans. Inf. Theory* **1981**, *27*, 472–482.
40. Mohammad-Djafari, A. Maximum d'entropie et problèmes inverses en imagerie. *Traitement Signal* **1994**, *11*, 87–116.
41. Bercher, J. Développement de critères de nature entropique pour la résolution des problèmes inverses linéaires. Ph.D. Thesis, Université de Paris–Sud, Orsay, France, 1995.

42. Le Besnerais, G. Méthode du maximum d'entropie sur la moyenne, critère de reconstruction d'image et synthèse d'ouverture en radio astronomie. Ph.D. Thesis, Université de Paris-Sud, Orsay, France, 1993.
43. Caticha, A.; Giffin, A. Updating Probabilities. Presented at MaxEnt 2006, the 26th International Workshop on Bayesian Inference and Maximum Entropy Methods, Paris, France, 8–13 July 2006; doi:10.1063/1.2423258.
44. Caticha, A. Entropic Inference. Presented at MaxEnt 2010, the 30th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Chamonix, France, 4–9 July 2010.
45. Costa, S.I.R.; Santos, S.A.; Strapasson, J.E. Fisher information distance: A geometrical reading. **2012**, arXiv:1210.2354.
46. Rissanen, J. Fisher Information and Stochastic Complexity. *IEEE Trans. Inf. Theory* **1996**, *42*, 40–47.
47. Shimizu, R. On Fisher's amount of information for location family. In *A Modern Course on Statistical Distributions in Scientific Work*; D. Reidel: Dordrecht, The Netherlands, 1975; Volume 3, pp. 305–312.
48. Nielsen, F.; Nock, R. Sided and Symmetrized Bregman Centroids. *IEEE Trans. Inf. Theory* **2009**, *55*, 2048–2059.
49. Rasmussen, C.E.; Williams, C.K.I. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, USA, 2006.
50. Schroeder, M.R. Linear prediction, entropy and signal analysis. *IEEE ASSP Mag.* **1984**, *1*, 3–11.
51. Itakura, F.; Saito, S. A Statistical Method for Estimation of Speech Spectral Density and Formant Frequencies. *Electron. Commun. Jpn.* **1970**, *53-A*, 36–43.
52. Kitagawa, G.; Gersch, W. *Smoothness Priors Analysis of Time Series*; Lecture Notes in Statistics, Volume 116; Springer: New York, NY, USA, 1996.
53. Rue, H.; Held, L. *Gaussian Markov Random Fields: Theory and Applications*; CRC Press: New York, NY, USA, 2005.
54. Amari, S.; Cichocki, A.; Yang, H.H. A new learning algorithm for blind source separation. In *Advances in Neural Information Processing Systems 8*, Proceedings of the Conference on Neural Information Processing Systems 1995 (NIPS 1995), Denver, CO, USA, 27–30 November 1995; pp. 757–763.
55. Amari, S. Neural learning in structured parameter spaces—Natural Riemannian gradient. In *Advances in Neural Information Processing Systems 9*, Proceedings of the Conference on Neural Information Processing Systems 1995 (NIPS 1996), Denver, CO, USA, 2–5 December 1996; pp. 127–133.
56. Amari, S. Natural gradient works efficiently in learning. *Neural Comput.* **1998**, *10*, 251–276.
57. Knuth, K.H. Bayesian source separation and localization. *SPIE Proc.* **1998**, *3459*, doi:10.1117/12.323794.

58. Knuth, K.H. A Bayesian approach to source separation. In Proceedings of the First International Workshop on Independent Component Analysis and Signal Separation (ICA'99), Aussios, France, 11–15 January 1999; Cardoso, J.-F., Loubaton, P., Eds.; pp. 283–288.
59. Attias, H. Independent Factor Analysis. *Neural Comput.* **1999**, *11*, 803–851.
60. Mohammad-Djafari, A. A Bayesian approach to source separation. Presented at MaxEnt 99, the 19th International Workshop on Bayesian Inference and Maximum Entropy Methods, Boise State University, Boise, ID, USA, 2–6 August 1999; pp. 221–244.
61. Choudrey, R.A.; Roberts, S. Variational Bayesian Mixture of Independent Component Analysers for Finding Self-Similar Areas in Images. In Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), Nara, Japan, 1–4 April 2003; pp. 107–112.
62. Lopes, H.F.; West, M. Bayesian Model Assessment in Factor Analysis. *Statsinica* **2004**, *14*, 41–67.
63. Ichir, M.; Mohammad-Djafari, A.; Bayesian Blind Source Separation of Positive Non Stationary Sources. In Proceedings of MaxEnt 2004, 23rd International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Max-Planck Institiute, Garching, Germany, 25–30 July 2004; pp. 493–500
64. Mohammad-Djafari, A. Bayesian Source Separation: Beyond PCA and ICA. In Proceedings of 14th European Symposium on Artificial Neural Networks (ESANN 2006), Bruges, Belgium, 26–28 April 2006.
65. Comon, P., Jutten, C., Eds. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*; Academic Press: Burlington, MA, USA, 2010.
66. Yuan, M.; Lin, Y. Model selection and estimation in the Gaussian graphical model. *Biometrika* **2007**, *94*, 19–35.
67. Fitzgerald, W. Markov Chain Monte Carlo methods with Applications to Signal Processing. *Signal Process.* **2001**, *81*, 3–18.
68. Matsuoka, T.; Ulrych, T. Information theory measures with application to model identification. *IEEE Trans. Acoust. Speech Signal Process.* **1986**, *34*, 511–517.
69. Bretthorst, G.L. Bayesian Model Selection: Examples Relevant to NMR. In *Maximum Entropy and Bayesian Methods*; Springer: Dordrecht, The Netherlands, 1989; pp. 377–388.
70. Gelfand, A.E.; Dey, D.K. Bayesian model choice: Asymptotics and exact calculations. *J. R. Stat. Soc. Ser. B* **1994**, *56*, 501–514.
71. Mohammad-Djafari, A. Model selection for inverse problems: Best choice of basis function and model order selection. Presented at MaxEnt 1999, the 19th International Workshop on Bayesian Inference and Maximum Entropy Methods, Boise, Idaho, USA, 2–6 August 1999.
72. Clyde, M.A.; Berger, J.O.; Bullard, F.; Ford, E.B.; Jefferys, W.H.; Luo, R.; Paulo, R.; Loredo, T. Current Challenges in Bayesian Model Choice. In *Statistical Challenges in Modern Astronomy IV*, Proceedings of Conference on Statistical Challenges in Modern

Astronomy, Penn State University, PA, USA, 12–15 June 2006; Babu, G.J., Feigelson, E.D., Eds.; Volume 71, pp. 224–240.
73. Wyse, J.; Friel, N. Block clustering with collapsed latent block models. *Stat. Comput.* **2012**, *22*, 415–428.
74. Giovannelli, J.F.; Giremus, A. Bayesian noise model selection and system identification based on approximation of the evidence. In Proceedings of 2014 IEEE Statistical Signal Processing Workshop (SSP), Jupiters TBD, Gold Coast, Australia, 29 June–2 July 2014; pp. 125–128.
75. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Automat. Control* **1974**, *AC-19*, 716–723.
76. Akaike, H. Power spectrum estimation through autoregressive model fitting. *Ann. Inst. Stat. Math.* **1969**, *21*, 407–419.
77. Farrier, D. Jaynes' principle and maximum entropy spectral estimation. *IEEE Trans. Acoust. Speech Signal Process.* **1984**, *32*, 1176–1183.
78. Wax, M. Detection and Estimation of Superimposed Signals. Ph.D. Thesis, Standford University, CA, USA, March, 1985.
79. Burg, J.P. Maximum Entropy Spectral Analysis. In Proceedings of the 37th Annual International Meeting of Society of Exploration Geophysicists, Oklahoma City, OK, USA, 31 October 1967.
80. McClellan, J.H. Multidimensional spectral estimation. *Proc. IEEE* **1982**, *70*, 1029–1039.
81. Lang, S.; McClellan, J.H. Multidimensional MEM spectral estimation. *IEEE Trans. Acoust. Speech Signal Process.* **1982**, *30*, 880–887.
82. Johnson, R.; Shore, J. Which is Better Entropy Expression for Speech Processing:-SlogS or logS? *IEEE Trans. Acoust. Speech Signal Process.* **1984**, *ASSP-32*, 129–137.
83. Wester, R.; Tummala, M.; Therrien, C. Multidimensional Autoregressive Spectral Estimation Using Iterative Methods. In Proceedings of 1990 Conference Record Twenty-Fourth Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 5–7 November 1990; Volume 1, doi:10.1109/ACSSC.1990.523376.
84. Picinbono, B.; Barret, M. Nouvelle présentation de la méthode du maximum d'entropie. *Traitement Signal* **1990**, *7*, 153–158.
85. Borwein, J.M.; Lewis, A.S. Convergence of best entropy estimates. *SIAM J. Optim.* **1991**, *1*, 191–205.
86. Mohammad-Djafari, A., Ed. *Inverse Problems in Vision and 3D Tomography*; digital signal and image processing series; ISTE: London, UK and Wiley: Hoboken, NJ, USA, 2010.
87. Mohammad-Djafari, A.; Demoment, G. Tomographie de diffraction and synthèse de Fourier à maximum d'entropie. *Rev. Phys. Appl. (Paris)* **1987**, *22*, 153–167.
88. Féron, O.; Chama, Z.; Mohammad-Djafari, A. Reconstruction of piecewise homogeneous images from partial knowledge of their Fourier transform. In Proceedings of MaxEnt 2004, 23rd International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Max-Planck Institiute, Garching, Germany, 25–30 July 2004; pp.68–75.

89. Ayasso, H.; Mohammad-Djafari, A. Joint NDT Image Restoration and Segmentation Using Gauss–Markov–Potts Prior Models and Variational Bayesian Computation. *IEEE Trans. Image Process.* **2010**, *19*, 2265–2277.
90. Ayasso, H.; DuchÃłne, B.; Mohammad-Djafari, A. Bayesian inversion for optical diffraction tomography. *J. Mod. Opt.* **2010**, *57*, 765–776.
91. Burch, S.; Gull, S.F.; Skilling, J. Image Restoration by a Powerful Maximum Entropy Method. *Comput. Vis. Graph. Image Process.* **1983**, *23*, 113–128.
92. Gull, S.F.; Skilling, J. Maximum entropy method in image processing. *IEE Proc. F* **1984**, *131*, 646–659.
93. Gull, S.F. Developments in maximum entropy data analysis. In *Maximum Entropy and Bayesian Methods*; Skilling, J., Ed.; Springer: Dordrecht, The Netherlands, 1989; pp. 53–71.
94. Jones, L.K.; Byrne, C.L. General entropy criteria for inverse problems with application to data compression, pattern classification and cluster analysis. *IEEE Trans. Inf. Theory* **1990**, *36*, 23–30.
95. Macaulay, V.A.; Buck, B. Linear inversion by the method of maximum entropy. *Inverse Probl.* **1989**, *5*, doi:10.1088/0266-5611/5/5/013.
96. Rue, H.; Martino, S. Approximate Bayesian inference for hierarchical Gaussian Markov random field models. *J. Stat. Plan. Inference* **2007**, *137*, 3177–3192.
97. Wilkinson, R. Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. **2009**, arXiv:0811.3355.
98. Rue, H.; Martino, S.; Chopin, N. Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations. *J. R. Stat. Soc. Ser. B* **2009**, *71*, 319–392.
99. Fearnhead, P.; Prangle, D. Constructing Summary Statistics for Approximate Bayesian Computation: Semi-automatic ABC. **2011**, arxiv:1004.1112v2.
100. Turner, B.M.; van Zandt, T. A tutorial on approximate Bayesian computation. *J. Math. Psych.* **2012**, *56*, 69–85.
101. MacKay, D.J.C. A Practical Bayesian Framework for Backpropagation Networks. *Neural Comput.* **1992**, *4*, 448–472.
102. Mohammad-Djafari, A. Variational Bayesian Approximation for Linear Inverse Problems with a hierarchical prior models. In *Geometric Science of Information*, Proceedings of First International Conference on Geometric Science of Information (GSI 2013), Paris, France, 28–30 August 2013; Lecture Notes in Computer Science, Volume 8085; pp. 669–676.
103. Likas, C.L.; Galatsanos, N.P. A Variational Approach For Bayesian Blind Image Deconvolution. *IEEE Trans. Signal Process.* **2004**, *52*, 2222–2233.
104. Beal, M.; Ghahramani, Z. Variational Bayesian learning of directed graphical models with hidden variables. *Bayesian Stat.* **2006**, *1*, 793–832.
105. Kim, H.; Ghahramani, Z. Bayesian Gaussian Process Classification with the EM-EP Algorithm. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 1948–1959.

106. Jordan, M.I.; Ghahramani, Z.; Jaakkola, T.S.; Saul, L.K. An introduction to variational methods for graphical models. *Mach. Learn.* **2006**, *37*, 183–233.
107. Forbes, F.; Fort, G. Combining Monte Carlo and Mean-Field-Like Methods for Inference in Hidden Markov Random Fields. *IEEE Trans. Image Process.* **2007**, *16*, 824–837.
108. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum Likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. (B)* **1977**, *39*, 1–38.
109. Miller, M.I.; Snyder, D.L. The Role of Likelihood and Entropy in Incomplete-Data Problems: Applications to Estimating Point-Process Intensities and Toeplitz Constrained Covariances. *Proc. IEEE* **1987**, *75*, 892–907.
110. Snoussi, H.; Mohammad-Djafari, A. Information geometry of Prior Selection. In *Bayesian Inference and Maximum Entropy Methods*, Proceedings of 22nd International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, University of Idaho, Moscow, Idaho, ID, USA, 3–7 August 2002; Williams, C., Ed.; AIP Conference Proceedings 570.
111. Mohammad-Djafari, A. Approche variationnelle pour le calcul bayésien dans les problémes inverses en imagerie. **2009**, arXiv:0904.4148.
112. Beal, M. Variational Algorithms for Approximate Bayesian Inference. Ph.D. Thesis, Gatsby Computational Neuroscience Unit, University College London, UK, 2003.
113. Winn, J.; Bishop, C.M.; Jaakkola, T. Variational message passing. *J. Mach. Learn. Res.* **2005**, *6*, 661–694.
114. Chatzis, S.; Varvarigou, T. Factor Analysis Latent Subspace Modeling and Robust Fuzzy Clustering Using t-DistributionsClassification of binary random Patterns. *IEEE Trans. Fuzzy Syst.* **2009**, *17*, 505–517.
115. Park, T.; Casella, G. The Bayesian Lasso. *J. Am. Stat. Assoc.* **2008**, *103*, 681–686.
116. Mohammad-Djafari, A. A variational Bayesian algorithm for inverse problem of computed tomography. In *Mathematical Methods in Biomedical Imaging and Intensity-Modulated Radiation Therapy (IMRT)*; Censor, Y., Jiang, M., Louis, A.K., Eds.; Publications of the Scuola Normale Superiore/CRM Series; Edizioni della Normale: Rome, Italy, 2008; pp. 231–252.
117. Mohammad-Djafari, A.; Ayasso, H. Variational Bayes and mean field approximations for Markov field unsupervised estimation. In Proceedings of IEEE International Workshop on Machine Learning for Signal Processing (MLSP), Grenoble, France, 2–4 September 2009; pp. 1–6.
118. Tipping, M.E. Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* **2001**, *1*, 211–244.
119. He, L.; Chen, H.; Carin, L. Tree-Structured Compressive Sensing With Variational Bayesian Analysis. *IEEE Signal Process. Lett.* **2010**, *17*, 233–236.
120. Fraysse, A.; Rodet, T. A gradient-like variational Bayesian algorithm. In Proceedings of 2011 IEEE Conference on Statistical Signal Processing Workshop (SSP), Nice, France, 28–30 June 2011; pp. 605–608.

121. Johnson, V.E. On Numerical Aspects of Bayesian Model Selection in High and Ultrahigh-dimensional Settings. *Bayesian Anal.* **2013**, *8*, 741–758.
122. Dumitru, M.; Mohammad-Djafari, A. Estimating the periodic components of a biomedical signal through inverse problem modeling and Bayesian inference with sparsity enforcing prior. *AIP Conf. Proc.* **2015**, *1641*, 548–555.
123. Wang, L.; Gac, N.; Mohammad-Djafari, A. Bayesian 3D X-ray computed tomography image reconstruction with a scaled Gaussian mixture prior model. *AIP Conf. Proc.* **2015**, *1641*, 556–563.
124. Mohammad-Djafari, A. Bayesian Blind Deconvolution of Images Comparing JMAP, EM and VBA with a Student-t a priori Model. In Proceedings of International Workshops on Electrical and Computer Engineering Subfields, Koc University, Istanbul, Turkey, 22–23 August 2014; pp. 98–103.
125. Su, F.; Mohammad-Djafari, A. An Hierarchical Markov Random Field Model for Bayesian Blind Image Separation. In Proceedings of International Congress on Image and Signal Processing (CISP2008), Sanya, China, 27–30 May 2008.
126. Su, F.; Cai, S.; Mohammad-Djafari, A. Bayesian blind separation of mixed text patterns. In Proceedings of IEEE International Conference on Audio, Language and Image Processing (ICALIP 2008), Shanghai, China, 7–9 July 2008; pp. 1373–1378.

# Black-Box Optimization Using Geodesics in Statistical Manifolds

**Jérémy Bensadon**

**Abstract:** Information geometric optimization (IGO) is a general framework for stochastic optimization problems aiming at limiting the influence of arbitrary parametrization choices: the initial problem is transformed into the optimization of a smooth function on a Riemannian manifold, defining a parametrization-invariant first order differential equation and, thus, yielding an approximately parametrization-invariant algorithm (up to second order in the step size). We define the geodesic IGO update, a fully parametrization-invariant algorithm using the Riemannian structure, and we compute it for the manifold of Gaussians, thanks to Noether's theorem. However, in similar algorithms, such as CMA-ES (Covariance Matrix Adaptation - Evolution Strategy) and xNES (exponential Natural Evolution Strategy), the time steps for the mean and the covariance are decoupled. We suggest two ways of doing so: twisted geodesic IGO (GIGO) and blockwise GIGO. Finally, we show that while the xNES algorithm is not GIGO, it is an instance of blockwise GIGO applied to the mean and covariance matrix separately. Therefore, xNES has an almost parametrization-invariant description.

Reprinted from *Entropy*. Cite as: Bensadon, J. Black-Box Optimization Using Geodesics in Statistical Manifolds. *Entropy* **2015**, *17*, 304–345.

## 1. Introduction

Consider an objective function $f\colon X \to \mathbb{R}$ to be minimized. We suppose we have absolutely no knowledge about $f$: the only thing we can do is ask for its value at any point $x \in X$ (black-box optimization) and that the evaluation of $f$ is a costly operation. We are going to study algorithms that can be described in the IGO framework (see [1]).

We consider the following optimization procedure:

We choose $(P_\theta)_{\theta\in\Theta}$ a family of probability distributions (which will be given a Riemannian manifold structure, following [2]) on $X$ and an initial probability distribution $P_{\theta^0}$. Now, we replace $f$ by $F\colon \Theta \to \mathbb{R}$ (for example $F(\theta) = E_{x\sim P_\theta}[f(x)]$), and we optimize $F$ by gradient descent, corresponding to the gradient flow:

$$\frac{\mathrm{d}\theta^t}{\mathrm{d}t} = -\nabla_\theta E_{x\sim P_\theta}[f(x)]. \tag{1}$$

However, because of the gradient, this equation depends entirely on the parametrization we chose for $\Theta$, which is disturbing: we do not want to have two different updates, because we chose different parameters to represent the objects with which we are working. Moreover, in the case of a function with several local minima, changing the parametrization can change the attained optimum (see [3], for example). That is why invariance is a design principle behind IGO. More precisely, we want

invariance with respect to monotone transformations of $f$ and invariance under reparametrization of $\Theta$.

The IGO framework uses the geometry of the family $\Theta$, which is given by the Fisher metric to provide a differential equation on $\theta$ with the desired properties, but because of the discretization of time needed to obtain an explicit algorithm, we lose invariance under reparametrization of $\theta$: two IGO algorithms applied to the same function to be optimized, but with different parametrizations, coincide only at first order in the step size. A possible solution to this problem is geodesic IGO (GIGO), introduced here (see also IGO-Maximum Likelihoodin [1], for example.): the initial direction of the update at each step of the algorithm remains the same as in IGO, but instead of moving straight for the chosen parametrization, we use the Riemannian manifold structure of our family of probability distributions (see [2]) by following its geodesics.

Finding the geodesics of a Riemannian manifold is not always easy, but Noether's theorem will allow us to obtain quantities that are preserved along the geodesics, thus allowing, in the case of Gaussian distributions, one to obtain a first order differential equation satisfied by the geodesics, which makes their computation easier.

Although the geodesic IGO algorithm is not, strictly speaking, parametrization-invariant when no closed form for the geodesics is known, it is possible to compute them at arbitrary precision without increasing the numbers of objective function calls.

The first two sections are preliminaries: in Section 2, we recall the IGO algorithm, introduced in [1], and in Section 3, after a reminder about Riemannian geometry, we state Noether's theorem, which will be our main tool to compute the GIGO update for Gaussian distributions.

In Section 4, we consider Gaussian distributions with a covariance matrix proportional to the identity matrix: this space is isometric to the hyperbolic space, and the geodesics of the latter are known.

In Section 5.1, we consider the general Gaussian case, and we use Noether's theorem to obtain two different sets of equations to compute the GIGO update. The equations are known (see [4–6]), but the connection with Noether's theorem has not been mentioned. We then give the explicit solution for these equations, from [5].

In Section 6, we recall quickly the xNES and CMA-ESupdates, and we introduce a slight modification of the IGO algorithm to incorporate the direction-dependent learning rates used in CMA-ESand xNES. We then compare these different algorithms and prove that xNES is not GIGO in general, and we finally introduce a new family of algorithms extending GIGO and recovering xNES from abstract principles.

Finally, Section 7 presents numerical experiments, which suggest that when using GIGO with Gaussian distributions, the step size must be chosen carefully.

## 2. Definitions: IGO, GIGO

In this section, we recall what the IGO framework is and we define the geodesic IGO update. Consider again Equation (1):

$$\frac{\mathrm{d}\theta^t}{\mathrm{d}t} = -\nabla_\theta E_{x\sim P_\theta}[f(x)].$$

As we saw in the Introduction:

- The gradient depends on the parametrization of our space of probability distributions (see Section 2.3 for an example).
- The equation is not invariant under monotone transformations of $f$. For example, the optimization for $10f$ moves ten times faster than the optimization for $f$.

In this section, we recall how IGO deals with this (see [1] for a better presentation).

### *2.1. Invariance under Reparametrization of $\theta$: Fisher Metric*

In order to achieve invariance under reparametrization of $\theta$, it is possible to turn our family of probability distributions into a Riemannian manifold (this is the main topic of information geometry; see [2]), which allows us to use a canonical, parametrization-invariant gradient (called the natural gradient).

**Definition 1.** *Let $P, Q$ be two probability distributions on $X$. The Kullback–Leibler divergence of $Q$ from $P$ is defined by:*

$$\mathrm{KL}(Q\|P) = \int_X \ln(\frac{Q(x)}{P(x)})\mathrm{d}Q(x). \tag{2}$$

By definition, it does not depend on the parametrization. It is not symmetrical, but if for all $x$, the application $\theta \mapsto P_\theta(x)$ is $C^2$, then a second-order expansion yields:

$$\mathrm{KL}(P_{\theta+\mathrm{d}\theta}\|P_\theta) = \frac{1}{2}\sum_{i,j} I_{ij}(\theta)\mathrm{d}\theta_i\mathrm{d}\theta_j + o(\mathrm{d}\theta^2), \tag{3}$$

where:

$$I_{ij}(\theta) = \int_X \frac{\partial \ln P_\theta(x)}{\partial\theta_i}\frac{\partial \ln P_\theta(x)}{\partial\theta_j}\mathrm{d}P_\theta(x) = -\int_X \frac{\partial^2 \ln P_\theta(x)}{\partial\theta_i\partial\theta_j}\mathrm{d}P_\theta(x). \tag{4}$$

This is enough to endow the family $(P_\theta)_{\theta\in\Theta}$ with a Riemannian manifold structure: a Riemannian manifold $M$ is a differentiable manifold, which can be seen as pieces of $\mathbb{R}^n$ glued together, with a metric. The metric at $x$ is a symmetric positive-definite quadratic form on the tangent space of $M$ at $x$: it indicates how expensive it is to move in a given direction on the manifold. We will think of the updates of the algorithms that we will be studying as paths on $M$.

The matrix $I(\theta)$ is called the "Fisher information matrix", and the metric it defines is called the "Fisher metric".

Given a metric, it is possible to define a gradient attached to this metric; the key property of the gradient is that for any smooth function $f$:

$$f(x+h) = f(x) + \sum_i h_i\frac{\partial f}{\partial x_i} + o(\|h\|) = f(x) + \langle h, \nabla f(x)\rangle + o(\|h\|), \tag{5}$$

where $\langle x, y\rangle = x^T I y$ is the dot product in metric $I$. Therefore, in order to keep the property of Equation (5), we must have $\nabla f = I^{-1}\frac{\partial f}{\partial x}$.

We have therefore the following gradient (called the "natural gradient"; see [2]):

$$\tilde{\nabla}_\theta = I^{-1}(\theta)\frac{\partial}{\partial\theta}, \tag{6}$$

and since the Kullback–Leibler divergence does not depend on the parametrization, neither does the natural gradient.

Later in this paper, we will study families of Gaussian distributions. The following proposition gives the Fisher metric for these families.

**Proposition 1.** *Let $(P_\theta)_{\theta\in\Theta}$ be a family of normal probability distributions: $P_\theta = \mathcal{N}(\mu(\theta), \Sigma(\theta))$. If $\mu$ and $\Sigma$ are $C^1$, the Fisher metric is given by:*

$$I_{i,j}(\theta) = \frac{\partial\mu}{\partial\theta_i}^T\Sigma^{-1}\frac{\partial\mu}{\partial\theta_j} + \frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_i}\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_j}\right). \tag{7}$$

**Proof.** This is a non-trivial calculation. See [7] or [8] for more details. □

As we will often be working with Gaussian distributions, we introduce the following notation:

**Notation 1.** *$\mathbb{G}_d$ is the manifold of Gaussian distributions in dimension $d$, equipped with the Fisher metric. $\tilde{\mathbb{G}}_d$ is the manifold of Gaussian distributions in dimension $d$, with the covariance matrix proportional to identity in the canonical basis of $\mathbb{R}^d$, equipped with the Fisher metric.*

### *2.2. IGO Flow, IGO Algorithm*

In IGO [1], invariance with respect to monotone transformations is achieved by replacing $f$ by the following transform; we set:

$$q(x) = P_{x'\sim P_\theta}(f(x') \leqslant f(x)), \tag{8}$$

a non-increasing function $w\colon [0;1] \to \mathbb{R}$ is chosen (the selection scheme), and finally, $W_\theta^f(x) = w(q(x))$ (this definition has to be slightly changed if the probability of a tie is not zero, see [1] for more details). By performing a gradient descent on $E_{x\sim P_\theta}[W_{\theta^t}^f(x)]$, we obtain the "IGO flow":

$$\frac{\mathrm{d}\theta^t}{\mathrm{d}t} = \tilde{\nabla}_\theta\int_X W_{\theta^t}^f(x)P_\theta(\mathrm{d}x) = \int_X W_{\theta^t}^f(x)\tilde{\nabla}_\theta\ln P_\theta(x)P_{\theta^t}(\mathrm{d}x). \tag{9}$$

Notice that the function we are optimizing is $E_{x\sim P_\theta}[W_{\theta^t}^f(x)]$ and not $E_{x\sim P_\theta}[W_\theta^f(x)]$ (the second function is constant and always equal to $\int_0^1 w$). In particular, the function for which we are performing the gradient descent changes at each step, although their optimum (a Dirac at the minimum of $f$) does not: the IGO flow is not a gradient flow; it is simply a vector flow given by the gradient of interrelated functions.

For practical implementation, the integral in (9) has to be approximated. For the integral itself, the Monte-Carlo method is used; $N$ values $(x_1, ..., x_N)$ are sampled from the distribution $P_{\theta^t}$, and the integral becomes:

$$\frac{1}{N}\sum_{i=1}^{N} W_{\theta^t}^f(x_i)\tilde{\nabla}_\theta \ln P_\theta(x_i) \tag{10}$$

and we approximate $\frac{1}{N}W_\theta^f(x_i) = \frac{1}{N}w(q(x_i))$ by $\hat{w}_i = \frac{1}{N}w(\frac{\mathrm{rk}(x_i)+1/2}{N})$, where $\mathrm{rk}(x_i) = |\{j, f(x_j) < f(x_i)\}|$: it can be proven (see [1]) that $\lim_{N\to\infty} N\hat{w}_i = W_f^{\theta^t}(x_i)$ (here again, we are assuming that there are no ties).

We now have an algorithm that can be used in practice if the Fisher information matrix is known.

**Definition 1.** *The IGO update associated with parametrization $\theta$, sample size $N$, step size $\delta t$ and selection scheme $w$ is given by the following update rule:*

$$\theta^{t+\delta t} = \theta^t + \delta t I^{-1}(\theta^t)\sum_{i=1}^{N}\hat{w}_i\frac{\partial \ln P_\theta(x_i)}{\partial\theta}. \tag{11}$$

*We call IGO speed the vector $I^{-1}(\theta^t)\sum_{i=1}^{N}\hat{w}_i\frac{\partial \ln P_\theta(x_i)}{\partial\theta}$.*

Notice that one could start directly with the $\hat{w}_i$ rather than $w$, as we will do later.

Replacing $f$ by its expected value under a probability distribution $P_\theta$ and optimizing over $\theta$ using the natural gradient have already been discussed. For example, in the case of a function $f$ defined on $\{0,1\}^n$, IGO with the Bernoulli distributions yields the algorithm, PBIL[9]. Another similar approach (stochastic relaxation) is given in [10]. For a continuous function, as we will see later, the IGO framework recovers several known ranked-based natural gradient algorithms, such as pure rank-$\mu$ CMA-ES [11], xNES or SNES (Separable Natural Evolution Strategies) [12]. See [13] or [14] for other, not necessarily gradient-based, optimization algorithms on manifolds.

### *2.3. Geodesic IGO*

Although the IGO flow associated with a family of probability distributions is intrinsic (it only depends on the family itself, not the parametrization we choose for it), the IGO update is not. However, the difference between two steps of IGO that differ only by the parametrization is only $O(\delta t^2)$, whereas the different between two vanilla gradient descents with different parametrizations is $O(\delta t)$.

Intuitively, the reason for this difference is that two IGO algorithms start at the same point and follow "straight lines" with the same initial speed, but the definition of "straight lines" changes with the parametrization.

For instance, in the case of Gaussian distributions, let us consider two different IGO updates with Gaussian distributions in dimension one, the first one with parametrization $(\mu, \sigma)$ and the second one with parametrization $(\mu, c := \sigma^2)$. We suppose that the IGO speed for the first algorithm is $(\dot{\mu}, \dot{\sigma})$. The corresponding IGO speed in the second parametrization is given by the identity $\dot{c} = 2\sigma\dot{\sigma}$.

Therefore, the first algorithm gives the standard deviation $\sigma_{\text{new},1} = \sigma_{\text{old}} + \delta t\dot{\sigma}$ and the variance $c_{\text{new},1} = (\sigma_{\text{new},1})^2 = c_{\text{old}} + 2\delta t\sigma_{\text{old}}\dot{\sigma} + \delta t^2\dot{\sigma}^2 = c_{\text{new},2} + \delta t^2\dot{\sigma}^2$.

The geodesics of a Riemannian manifold are the generalization of the notion of a straight line: they are curves that locally minimize length. In particular, given two points $a$ and $b$ on the Riemannian manifold $M$, the shortest path from $a$ to $b$ is always a geodesic (the converse is not true, though). The notion will be explained precisely in Section 3, but let us define the geodesic IGO algorithm, which follows the geodesics of the manifold instead of following the straight lines for an arbitrary parametrization.

**Definition 2** (GIGO)**.** *The geodesic IGO update (GIGO) associated with sample size $N$, step size $\delta t$ and selection scheme $w$ is given by the following update rule:*

$$\theta^{t+\delta t} = \exp_{\theta^t}(Y\delta t) \tag{12}$$

*where:*

$$Y = I^{-1}(\theta^t)\sum_{i=1}^{N}\hat{w}_i\frac{\partial \ln P_\theta(x_i)}{\partial\theta}, \tag{13}$$

*is the IGO speed and $\exp_{\theta^t}$ is the exponential of the Riemannian manifold $\Theta$. Namely, $\exp_{\theta^t}(Y\delta t)$ is the endpoint of the geodesic of $\Theta$ starting at $\theta^t$, with initial speed $Y$, after a time $\delta t$. By definition, this update does not depend on the parametrization $\theta$.*

Notice that while the GIGO update is compatible with the IGO flow (in the sense that when $\delta t \to 0$ and $N \to \infty$, a parameter $\theta^t$ updated according to the GIGO algorithm is a solution of Equation (9), the equation defining the IGO flow), it not necessarily an IGO update. More precisely, the GIGO update is an IGO update if and only if the geodesics of $\Theta$ are straight lines for some parametrization (by Beltrami's theorem, this is equivalent to $\Theta$ having constant curvature).

This is a particular case of a retraction [14]: a map from the tangent bundle of a manifold to the manifold itself satisfying a rigidity condition. Arguably, the Riemannian exponential is the most natural retraction, since it depends only on the Riemannian manifold itself and not on any decomposition. However, in general, the geodesics are difficult to compute.

In the next section, we will state Noether's theorem, which will be our main tool to compute the GIGO update for Gaussian distributions.

## 3. Riemannian Geometry, Noether's Theorem

### *3.1. Riemannian Geometry*

The goal of this section is to state Noether's theorem. See [15] for the proofs and [16] or [17] for a more detailed presentation. Noether's theorem states that if a system has symmetries, then there are invariants attached to these symmetries. Firstly, we need some definitions.

**Definition 3** (Motion in a Lagrangian system)**.** *Let $M$ be a differentiable manifold, $TM$ the set of tangent vectors on $M$ (a tangent vector is identified by the point at which it is tangent and a vector in*

*the tangent space) and* $\begin{array}{rccc}\mathcal{L} & : & TM & \to & \mathbb{R} \\ & & (q,v) & \mapsto & \mathcal{L}(q,v)\end{array}$ *a differentiable function called the Lagrangian function (in general, it could depend on* $t$*). A "motion in the Lagrangian system* $(M,\mathcal{L})$ *from* $x$ *to* $y$*" is map* $\gamma\colon [t_0,t_1]\to M$*, such that:*

- $\gamma(t_0)=x$
- $\gamma(t_1)=y$
- $\gamma$ *is a local extremum of the functional:*

$$\Phi(\gamma)=\int_{t_0}^{t_1}\mathcal{L}(\gamma(t),\dot{\gamma}(t))\mathrm{d}t, \tag{14}$$

*among all curves* $c\colon [t_0,t_1]\to M$*, such that* $c(t_0)=x$*, and* $c(t_1)=y$*.*

For example, when $(M,g)$ is a Riemannian manifold, the length of a curve $\gamma$ between $\gamma(t_0)$ and $\gamma(t_1)$ is:

$$\int_{t_0}^{t_1}\sqrt{g(\dot{\gamma}(t),\dot{\gamma}(t))}\mathrm{d}t. \tag{15}$$

The curves that follow the shortest path between two points $x,y\in M$ are therefore the minima $\gamma$ of the functional (15), such that $\gamma(t_0)=x$ and $\gamma(t_1)=y$, and the corresponding Lagrangian function is $(q,v)\mapsto\sqrt{g(v,v)}$. However, any curve following the shortest trajectory will have minimum length. For example, if $\gamma_1:[a,b]\to M$ is a curve of the shortest path, so is $\gamma_2: t\mapsto\gamma_1(t^2)$: these two curves define the same trajectory in $M$, but they do not travel along this trajectory at the same speed. This leads us to the following definition:

**Definition 4** (Geodesics)**.** *Let* $I$ *be an interval of* $\mathbb{R}$ *and* $(M,g)$ *be a Riemannian manifold. A curve* $\gamma: I\to M$ *is called a geodesic if for all* $t_0,t_1\in I$*,* $\gamma|_{[t_0,t_1]}$ *is a motion in the Lagrangian system* $(M,\mathcal{L})$ *from* $\gamma(t_0)$ *to* $\gamma(t_1)$*, where:*

$$\mathcal{L}(\gamma)=\int_{t_0}^{t_1}g(\dot{\gamma}(t),\dot{\gamma}(t))\mathrm{d}t. \tag{16}$$

It can be shown (see [16]) that geodesics are curves that locally minimize length, with constant velocity, in the sense that $\frac{dg(\dot{\gamma}(t),(\dot{\gamma}(t))}{dt}=0$. In particular, given a starting point and a starting speed, the geodesic is unique. This motivates the definition of the exponential of a Riemannian manifold.

**Definition 5.** *Let* $(M,g)$ *be a Riemannian manifold. We call the exponential of* $M$ *the application:*

$$\begin{array}{rccc}\exp & : & TM & \to & M \\ & & (x,v) & \mapsto & \exp_x(v),\end{array}$$

*such that for any* $x\in M$*, if* $\gamma$ *is the geodesic of* $M$ *satisfying* $\gamma(0)=x$ *and* $\gamma'(0)=v$*, then* $\exp_x(v)=\gamma(1)$*.*

In order to find an extremal of a functional, the most commonly-used result is called the "Euler–Lagrange equations" (see [15], for example); a motion $\gamma$ in the Lagrangian system $(M, \mathcal{L})$ must satisfy:

$$\frac{\partial \mathcal{L}}{\partial x}(\gamma(t)) - \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial \mathcal{L}}{\partial \dot{x}}(\dot{\gamma}(t))\right) = 0. \tag{17}$$

By applying this equation with the Lagrangian given by (16), it is possible to show that the geodesics of a Riemannian manifold follow the "geodesic equations":

$$\ddot{x}^k + \Gamma^k_{ij}\dot{x}^i\dot{x}^j = 0, \tag{18}$$

where the

$$\Gamma^k_{ij} = \frac{1}{2}g^{lk}\left(\frac{\partial g_{jl}}{\partial q_i} + \frac{\partial g_{li}}{\partial q_j} - \frac{\partial g_{ij}}{\partial q_l}\right) \tag{19}$$

are called "Christoffel symbols" of the metric $g$. However, these coefficients are tedious (and sometimes difficult) to compute, and (18) is a second order differential equation. Noether's theorem will give us a first order equation to compute the geodesics.

### *3.2. Noether's Theorem*

**Definition 6.** *Let $h\colon M \to M$, a diffeomorphism. We say that the Lagrangian system $(M, \mathcal{L})$ admits the symmetry $h$ if for any $(q, v) \in TM$,*

$$\mathcal{L}\left(h(q), \mathrm{d}h(v)\right) = \mathcal{L}(q, v), \tag{20}$$

*where $\mathrm{d}h$ is the differential of $h$.*

*If $M$ is clear in the context, we will sometimes say that $\mathcal{L}$ is invariant under $h$.*

An example will be given in the proof of Theorem 3.

We can now state Noether's theorem (see, for example, [15]).

**Theorem 1** (Noether's Theorem)**.** *If the Lagrangian system $(M, \mathcal{L})$ admits the one-parameter group of symmetries $h^s\colon M \to M$, $s \in \mathbb{R}$, then the following quantity remains constant during motions in the system $(M, \mathcal{L})$. Namely,*

$$I(\gamma(t), \dot{\gamma}(t)) = \frac{\partial \mathcal{L}}{\partial v}\left(\frac{\mathrm{d}h^s(\gamma(t))}{\mathrm{d}s}|_{s=0}\right) \tag{21}$$

*does not depend on $t$ if $\gamma$ is a motion in $(M, \mathcal{L})$.*

Now, we are going to apply this theorem to our problem: computing the geodesics of Riemannian manifolds of Gaussian distributions.

## 4. GIGO in $\tilde{\mathbb{G}}_d$

If we force the covariance matrix to be either diagonal or proportional to the identity matrix, the geodesics have a simple expression that we give below. In the former case, the manifold we are considering is $(\mathbb{G}_1)^d$, and in the latter case, it is $\tilde{\mathbb{G}}_d$.

The geodesics of $(\mathbb{G}_1)^d$ are given by:

**Proposition 2.** *Let $M$ be a Riemannian manifold; let $d \in \mathbb{N}$; let $\Phi$ be the Riemannian exponential of $M^d$; and let $\phi$ be the Riemannian exponential of $M$. We have:*

$$\Phi_{(x_1,\dots,x_n)}((v_1,\dots,v_n)) = (\phi_{x_1}(v_1),\dots,\phi_{x_n}(v_n)) \tag{22}$$

*In particular, knowing the geodesics of $\mathbb{G}_1$ is enough to compute the geodesics of $(\mathbb{G}_1)^d$.*

This is true, because a block of the product metric does not depend on variables of the other blocks.

Consequently, a GIGO update with a diagonal covariance matrix with the sample $(x_i)$ is equivalent to $d$ separate one-dimensional GIGO updates using the same samples. Moreover, $\mathbb{G}_1 \cong \tilde{\mathbb{G}}_1$, the geodesics of which are given below.

We will show that $\tilde{\mathbb{G}}_d$ and the "hyperbolic space", of which the geodesics are known, are isometric.

### *4.1. Preliminaries: Poincaré Half-Plane, Hyperbolic Space*

In dimension two, the hyperbolic space is called the "hyperbolic plane" or the Poincaré half-plane. We recall its definition:

**Definition 7** (Poincaré half-plane)**.** *We call the "Poincaré half-plane" the Riemannian manifold:*

$$\mathcal{H} = \{(x, y) \in \mathbb{R}^2,\ y > 0\},$$

*with the metric $ds^2 = \frac{dx^2+dy^2}{y^2}$.*

We also recall the expression of its geodesics (see, for example, [18]):

**Proposition 3** (Geodesics of the Poincaré half-plane)**.** *The geodesics of the Poincaré half-plane are exactly the:*

$$t \mapsto (\mathrm{Re}(z(t)), \mathrm{Im}(z(t)))\,,$$

*where:*

$$z(t) = \frac{aie^{vt} + b}{cie^{vt} + d}, \tag{23}$$

*with $ad - bc = 1$ and $v > 0$.*

The geodesics are half-circles perpendicular to the line $y = 0$ and vertical lines, as shown in Figure 1 below.

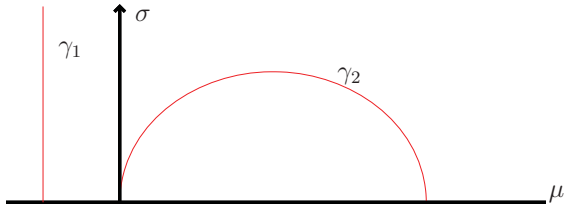



**Figure 1.** Geodesics of the Poincaré half-plane.

The generalization to the higher dimension is the following:

**Definition 8** (Hyperbolic space)**.** *We call the "hyperbolic space of dimension $n$" the Riemannian manifold:*

$$\mathcal{H}_n = \{(x_1, ..., x_{n-1}, y) \in \mathbb{R}^n, y > 0\},$$

*with the metric $ds^2 = \frac{dx_1^2+...+dx_{n-1}^2+dy^2}{y^2}$ (or equivalently, the metric given by the matrix $\mathrm{Diag}(\frac{1}{y^2})$ ).*

The Lagrangian for the geodesics is invariant under all translations along the $x_i$, so by Noether's theorem, its geodesics stay in a plane containing the direction $y$ and the initial speed . The induced metric on this plane is the metric of the Poincaré half-plane. The geodesics are therefore given by the following proposition:

**Proposition 4** (Geodesics of the hyperbolic space)**.** *If $\gamma\colon t \mapsto (x_1(t), ..., x_{n-1}(t), y(t)) = (\boldsymbol{x}(t), y(t))$ is a geodesic of $\mathcal{H}_n$, then there exists $a, b, c, d \in \mathbb{R}$, such that $ad - bc = 1$, and $v > 0$, such that $\boldsymbol{x}(t) = \boldsymbol{x}(0) + \frac{\dot{\boldsymbol{x}}_0}{\|\dot{\boldsymbol{x}}_0\|}\tilde{x}(t)$, $y(t) = \mathrm{Im}(\gamma_{\mathbb{C}}(t))$, with $\tilde{x}(t) = \mathrm{Re}(\gamma_{\mathbb{C}}(t))$ and:*

$$\gamma_{\mathbb{C}}(t) := \frac{aie^{vt} + b}{cie^{vt} + d}. \tag{24}$$

*4.2. Computing the GIGO Update in $\tilde{\mathbb{G}}_d$*

If we want to implement the GIGO algorithm in $\tilde{\mathbb{G}}_d$, we need to compute the natural gradient in $\tilde{\mathbb{G}}_d$ and to be able to compute the Riemannian exponential of $\tilde{\mathbb{G}}_d$.

Using Proposition 1, we can compute the metric of $\tilde{\mathbb{G}}_d$ in the parametrization $(\mu, \sigma) \mapsto \mathcal{N}(\mu, \sigma^2 I)$. We find:

$$\begin{pmatrix} \frac{1}{\sigma^2} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \frac{1}{\sigma^2} & 0 \\ 0 & \dots & 0 & \frac{2d}{\sigma^2} \end{pmatrix}. \tag{25}$$

Since this matrix is diagonal, it is easy to invert, and we immediately have the natural gradient and, consequently, the IGO speed.

**Proposition 5.** *In $\tilde{\mathbb{G}}_d$, the IGO speed $Y$ is given by:*

$$Y_\mu = \sum_i \hat{w}_i (x_i - \mu), \tag{26}$$

$$Y_\sigma = \sum_i \hat{w}_i \left( \frac{(x_i - \mu)^T (x_i - \mu)}{2d\sigma} - \frac{\sigma}{2} \right). \tag{27}$$

**Proof.** We recall the IGO speed is defined by $Y = I^{-1}(\theta^t) \sum_{i=1}^N \hat{w}_i \frac{\partial \ln P_\theta(x_i)}{\partial \theta}$. Since $P_{\mu,\sigma}(x) = (2\pi\sigma^2)^{-d/2} \exp(-\frac{(x-\mu)^T(x-\mu)}{2\sigma^2})$, we have:

$$\frac{\partial \ln P_{\mu,\sigma}(x)}{\partial \mu} = x - \mu,$$

$$\frac{\partial \ln P_{\mu,\sigma}(x)}{\partial \sigma} = -\frac{d}{\sigma} + \frac{(x-\mu)^T(x-\mu)}{\sigma^3}.$$

The result follows. □

The metric defined by Equation (25) is not exactly the metric of the hyperbolic space, but with the substitution $\mu \leftarrow \frac{\mu}{\sqrt{2d}}$, the metric becomes $\frac{2d}{\sigma^2} I$, which is proportional to the metric of the hyperbolic space and, therefore, defines the same geodesics.

**Theorem 2** (Geodesics of $\tilde{\mathbb{G}}_d$)**.** *If $\gamma \colon t \mapsto \mathcal{N}(\mu(t), \sigma(t)^2 I)$ is a geodesic of $\tilde{\mathbb{G}}_d$, then there exists $a, b, c, d \in \mathbb{R}$, such that $ad - bc = 1$, and $v > 0$, such that:*

*$\mu(t) = \mu(0) + \sqrt{2d} \frac{\dot{\mu}_0}{\|\dot{\mu}_0\|} \tilde{r}(t)$, $\sigma(t) = \mathrm{Im}(\gamma_{\mathbb{C}}(t))$, with $\tilde{r}(t) = \mathrm{Re}(\gamma_{\mathbb{C}}(t))$ and*

$$\gamma_{\mathbb{C}}(t) := \frac{aie^{vt} + b}{cie^{vt} + d}. \tag{28}$$

Now, in order to implement the corresponding GIGO algorithm, we only need to be able to find the coefficients $a, b, c, d, v$ corresponding to an initial position $(\mu_0, \sigma_0)$ and an initial speed $(\dot{\mu}_0, \dot{\sigma}_0)$. This is a tedious but easy computation, the result of which is given in Proposition 17.

The pseudocode of GIGO in $\tilde{\mathbb{G}}_d$ is also given in the Appendix: it is obtained by concatenating Algorithms 1 and 7 (Proposition 17 and the pseudocode in the Appendix allow the metric to be slightly modified; see Section 6.2).

## 5. GIGO in $\mathbb{G}_d$

### *5.1. Obtaining a First Order Differential Equation for the Geodesics of $\mathbb{G}_d$*

In the case where both the covariance matrix and the mean can vary freely, the equations of the geodesics have been computed in [4] and [5]. However, these articles start with the equations of the geodesics obtained with the Christoffel symbols, then partially integrate them . These equations are in fact a consequence of Noether's theorem and can be found directly.

**Theorem 3.** *Let $\gamma : t \mapsto \mathcal{N}(\mu_t, \Sigma_t)$ be a geodesic of $\mathbb{G}_d$. Then, the following quantities do not depend on $t$:*

$$J_\mu = \Sigma_t^{-1}\dot{\mu}_t, \tag{29}$$

$$J_\Sigma = \Sigma_t^{-1}(\dot{\mu}_t\mu_t^T + \dot{\Sigma}_t). \tag{30}$$

**Proof.** This is a direct application of Noether's theorem, with suitable groups of diffeomorphisms. By Proposition 1, the Lagrangian associated with the geodesics of $\mathbb{G}_d$ is:

$$\mathcal{L}(\mu, \Sigma, \dot{\mu}, \dot{\Sigma}) = \dot{\mu}^T\Sigma^{-1}\dot{\mu} + \frac{1}{2}\operatorname{tr}(\dot{\Sigma}\Sigma^{-1}\dot{\Sigma}\Sigma^{-1}). \tag{31}$$

Its derivative is:

$$\frac{\partial\mathcal{L}}{\partial\dot{\theta}} = \left[(h, H) \mapsto 2\dot{\mu}^T\Sigma^{-1}h + \operatorname{tr}(H\Sigma^{-1}\dot{\Sigma}\Sigma^{-1})\right]. \tag{32}$$

Let us show that this Lagrangian is invariant under affine changes of basis (thus illustrating Definition 6).

The general form of an affine change of basis is $\phi_{\mu_0,A} : (\mu, \Sigma) \mapsto (A\mu + \mu_0, A\Sigma A^T)$, with $\mu_0 \in \mathbb{R}^d$ and $A \in \mathrm{GL}_d(\mathbb{R})$.

We have:

$$\mathcal{L}(\phi_{\mu_0,A}(\mu, \Sigma), d\phi_{\mu_0,A}(\dot{\mu}, \dot{\Sigma})) = \dot{\overline{A\mu}}^T(A\Sigma A^T)^{-1}\dot{\overline{A\mu}} + \frac{1}{2}\operatorname{tr}\left(\dot{\overline{A\Sigma A^T}}(A\Sigma A^T)^{-1}\dot{\overline{A\Sigma A^T}}(A\Sigma A^T)^{-1}\right), \tag{33}$$

and since $\dot{\overline{A\mu}} = A\dot{\mu}$ and $\dot{\overline{A\Sigma A^T}} = A\dot{\Sigma}A^T$, we find easily that:

$$\mathcal{L}(\phi_{\mu_0,A}(\mu, \Sigma), d\phi_{\mu_0,A}(\dot{\mu}, \dot{\Sigma})) = \mathcal{L}(\mu, \Sigma, \dot{\mu}, \dot{\Sigma}), \tag{34}$$

or in other words: $\mathcal{L}$ is invariant under $\phi_{\mu_0,A}$ for any $\mu_0 \in \mathbb{R}^d, A \in GL_d(\mathbb{R})$.

In order to use Noether's theorem, we also need one-parameter groups of transformations. We choose the following:

(1) Translations of the mean vector. For any $i \in [1, d]$, let $h_i^s : (\mu, \Sigma) \mapsto (\mu + se_i, \Sigma)$, where $e_i$ is the $i$-th basis vector. We have $\frac{\mathrm{d}h_i^s}{ds}|_{s=0} = (e_i, 0)$, so by Noether's theorem,

$$\frac{\partial\mathcal{L}}{\partial\dot{\theta}}(e_i, 0) = 2\dot{\mu}^T\Sigma^{-1}e_i = 2e_i^T\Sigma^{-1}\dot{\mu}$$

remains constant for all $i$. The fact that $J_\mu$ is an invariant immediately follows.

(2) Linear base changes. For any $i, j \in [1, d]$, let $h_{i,j}^s : (\mu, \Sigma) \mapsto (\exp(sE_{ij})\mu, \exp(sE_{ij})\Sigma\exp(sE_{ji}))$, where $E_{ij}$ is the matrix with a one at position $(i, j)$ and zeros elsewhere. We have:

$$\frac{\mathrm{d}h_{E_{ij}}^s}{\mathrm{d}s}|_{s=0} = (E_{ij}\mu, E_{ij}\Sigma + \Sigma E_{ji}).$$

Therefore, by Noether's theorem, we then obtain the following invariants:

$$J_{ij} := \frac{\partial\mathcal{L}}{\partial\dot{\theta}}(E_{ij}\mu, E_{ij}\Sigma + \Sigma E_{ji}) \tag{35}$$

$$= 2\,\dot{\mu}^T\Sigma^{-1}E_{ij}\mu + \operatorname{tr}((E_{ij}\Sigma + \Sigma E_{ji})\Sigma^{-1}\dot{\Sigma}\Sigma^{-1}) \tag{36}$$

$$= 2\,(\Sigma^{-1}\dot{\mu})^T E_{ij}\mu + \operatorname{tr}(E_{ij}\dot{\Sigma}\Sigma^{-1}) + \operatorname{tr}(E_{ji}\Sigma^{-1}\dot{\Sigma}) \tag{37}$$

$$= 2(J_\mu\mu^T)_{ij} + 2(\Sigma^{-1}\dot{\Sigma})_{ij}, \tag{38}$$

and the coefficients of $J_\Sigma$ in (30) are the $(J_{ij}/2)$.

□

This leads us to first order equations satisfied by the geodesics mentioned in [4–6].

**Theorem 4** (GIGO-$\Sigma$). *$t \mapsto \mathcal{N}(\mu_t, \Sigma_t)$ is a geodesic of $\mathbb{G}_d$ if and only if $\mu : t \mapsto \mu_t$ and $\Sigma : t \mapsto \Sigma_t$ satisfy the equations:*

$$\dot{\mu}_t = \Sigma_t J_\mu \tag{39}$$

$$\dot{\Sigma}_t = \Sigma_t (J_\Sigma - J_\mu \mu_t^T) = \Sigma_t J_\Sigma - \dot{\mu}_t \mu_t^T, \tag{40}$$

*where:*

$$J_\mu = \Sigma_0^{-1} \dot{\mu}_0,$$

*and:*

$$J_\Sigma = \Sigma_0^{-1} \left( \dot{\mu}_0 \mu_0^T + \dot{\Sigma}_0 \right).$$

**Proof.** This is an immediate consequence of Proposition 3. □

These equations can be solved analytically (see [5]); however, usually, that is not the case, and they have to be solved numerically, for example with the Euler method (the corresponding algorithm, which we call GIGO-$\Sigma$, is described in the Appendix). The goal of the remainder of the subsection is to show that having to use the Euler method is fine.

To avoid confusion, we will call the step size of the GIGO algorithm ($\delta t$ in Proposition 2) "GIGO step size" and the step size of the Euler method (inside a step of the GIGO algorithm) "Euler step size".

Having to solve our equations numerically brings two problems:

The first one is a theoretical problem: the main reason to study GIGO is its invariance under reparametrization of $\theta$, and we lose this invariance property when we use the Euler method. However, GIGO can get arbitrarily close to invariance by decreasing the Euler step size. In other words, the difference between two different IGO algorithms is $O(\delta t^2)$, and the difference between two different implementations of the GIGO algorithm is $O(h^2)$, where $h$ is the Euler step size; it is easier to reduce the latter. Still, without a closed form for the geodesics of $\mathbb{G}_d$, the GIGO update is rather expensive to compute, but it can be argued that most of the computation time will still be the computation of the objective function $f$.

The second problem is purely numerical: we cannot guarantee that the covariance matrix remains positive-definite along the Euler method. Here, apart from finding a closed form for the geodesics, we have two solutions.

We can enforce this *a posteriori*: if the covariance matrix we find is not positive-definite after a GIGO step, we repeat the failed GIGO step with a reduced Euler step size (in our implementation, we divided it by four; see Algorithm 2 in the Appendix.).

The other solution is to obtain differential equations on a square root of the covariance matrix (*any* matrix $A$, such that $\Sigma = AA^T$).

**Theorem 5** (GIGO-$A$)**.** *If $\mu : t \mapsto \mu_t$ and $A : t \mapsto A_t$ satisfy the equations:*

$$\dot{\mu}_t = A_t A_t^T J_\mu, \tag{41}$$

$$\dot{A}_t = \frac{1}{2}(J_\Sigma - J_\mu \mu_t^T)^T A_t, \tag{42}$$

*where:*

$$J_\mu = (A_0^{-1})^T A_0^{-1} \mu_0$$

*and:*

$$J_\Sigma = (A_0^{-1})^T A_0^{-1} (\dot{\mu}_0 \mu_0^T + \dot{A}_0 A_0^T + A_0 \dot{A}_0^T),$$

*then $t \mapsto \mathcal{N}(\mu_t, A_t A_t^T)$ is a geodesic of $\mathbb{G}_d$.*

**Proof.** This is a simple rewriting of Theorem 4: if we write $\Sigma := AA^T$, we find that $J_\mu$ and $J_\Sigma$ are the same as in Theorem 4, and we have:

$$\dot{\mu} = \Sigma J_\mu,$$

and:

$$\dot{\Sigma} = (\dot{A}A^T + A\dot{A}^T) = \frac{1}{2}(J_\Sigma - J_\mu \mu^T)^T AA^T + \frac{1}{2} AA^T (J_\Sigma - J_\mu \mu^T)$$

$$= \frac{1}{2}(J_\Sigma - J_\mu \mu^T)^T \Sigma + \frac{1}{2}\Sigma(J_\Sigma - J_\mu \mu^T) = \frac{1}{2}\Sigma(J_\Sigma - J_\mu \mu^T) + \frac{1}{2}[\Sigma(J_\Sigma - J_\mu \mu^T)]^T.$$

By Theorem 4, $\Sigma(J_\Sigma - J_\mu \mu^T)$ is symmetric (since $\dot{\Sigma}$ has to be symmetric). Therefore, we have $\dot{\Sigma} = \Sigma(J_\Sigma - J_\mu \mu^T)$, and the result follows. □

Notice that Theorem 5 gives an equivalence, whereas Theorem 4 does not. The reason is that the square root of a symmetric positive-definite matrix is not unique. Still, it is canonical; see the discussion in Section 6.1.2.

As for Theorem 4, we can solve Equations (41) and (42) numerically, and we obtain another algorithm (Algorithm 3 in the Appendix; we will call it GIGO-$A$), with a behavior similar to the previous one (with Equations (39) and (40)). For both of them, numerical problems can arise when the covariance matrix is almost singular.

We have not managed to find any example where one of these two algorithms converged to the minimum of the objective function, whereas the other did not, and their behavior is almost the same.

More interestingly, the performances of these two algorithms are also the same as the performances of the exact GIGO algorithm, using the equations of Section 5.2.

Notice that even though GIGO-$A$ directly maintains a square root of the covariance matrix, which makes sampling new points easier (to sample a point from $\mathcal{N}(\mu, \Sigma)$, a square root of $\Sigma$ is needed), both GIGO-$\Sigma$ and GIGO-$A$ still have to invert the covariance matrix (or its square root) at each step, which is as costly as the decomposition, so one of these algorithms is roughly as expensive to compute as the other.

*5.2. Explicit Form of the Geodesics of $\mathbb{G}_d$ (from [5])*

We now give the exact geodesics of $\mathbb{G}_d$: the following results are a rewriting of Theorem 3.1 and its first corollary in [5].

**Theorem 6.** *Let $(\dot{\mu}_0, \dot{\Sigma}_0) \in T_{\mathcal{N}(0,I)}\mathbb{G}_d$. The geodesic of $\mathbb{G}_d$ starting from $\mathcal{N}(0,1)$ with initial speed $(\dot{\mu}_0, \dot{\Sigma}_0)$ is given by:*

$$\exp_{\mathcal{N}(0,I)}(s\dot{\mu}_0, s\dot{\Sigma}_0) = \mathcal{N}\left(2R(s)\mathrm{sh}(\frac{sG}{2})G^{-}\dot{\mu}_0, R(s)R(s)^T\right), \tag{43}$$

*where* exp *is the Riemannian exponential of $\mathbb{G}_d$, $G$ is any matrix satisfying:*

$$G^2 = \dot{\Sigma}_0^2 + 2\dot{\mu}_0\dot{\mu}_0^T, \tag{44}$$

$$R(s) = \left(\left(\mathrm{ch}(\frac{sG}{2}) - \dot{\Sigma}_0 G^{-}\mathrm{sh}(\frac{sG}{2})\right)^{-1}\right)^T \tag{45}$$

*and $G^{-}$ is a pseudo-inverse of $G$*

In [5], the existence of $G$ (as a square root of $\dot{\Sigma}_0^2 + 2\dot{\mu}_0\dot{\mu}_0^T$) is proven. Notice that, anyway, in the expansions of (43) and (45), only even powers of $G$ appear.

Additionally, since, for all $A \in GL_d(\mathbb{R})$, for all $\mu_0 \in \mathbb{R}^d$, the application:

$$\begin{array}{rccc} \phi & : & \mathbb{G}_d & \to & \mathbb{G}_d \\ & & \mathcal{N}(\mu, \Sigma) & \mapsto & \mathcal{N}(A\mu + \mu_0, A\Sigma A^T) \end{array} \tag{46}$$

preserves the geodesics, we find the general expression for the geodesics of $\mathbb{G}_d$.

**Corollary 1.** *Let $\mu_0 \in \mathbb{R}^d$, $A \in GL_d(\mathbb{R})$ and $(\dot{\mu}_0, \dot{\Sigma}_0) \in T_{\mathcal{N}(\mu_0, A_0A_0^T)}\mathbb{G}_d$. The geodesic of $\mathbb{G}_d$ starting from $\mathcal{N}(\mu, \Sigma)$ with initial speed $(\dot{\mu}_0, \dot{\Sigma}_0)$ is given by:*

$$\exp_{\mathcal{N}(\mu_0, A_0A_0^T)}(s\dot{\mu}_0, s\dot{\Sigma}_0) = \mathcal{N}(\mu_1, A_1A_1^T), \tag{47}$$

*with:*

$$\mu_1 = 2A_0R(s)\mathrm{sh}(\frac{sG}{2})G^{-}A_0^{-1}\dot{\mu}_0 + \mu_0, \tag{48}$$

$$A_1 = A_0R(s), \tag{49}$$

*where* exp *is the Riemannian exponential of $\mathbb{G}_d$, $G$ is any matrix satisfying:*

$$G^2 = A_0^{-1}(\dot{\Sigma}_0\Sigma_0^{-1}\dot{\Sigma}_0 + 2\dot{\mu}_0\dot{\mu}_0^T)(A_0^{-1})^T, \tag{50}$$

$$R(s) = \left(\left(\mathrm{ch}(\frac{sG}{2}) - A_0^{-1}\dot{\Sigma}_0(A_0^{-1})^T G^{-}\mathrm{sh}(\frac{sG}{2})\right)^{-1}\right)^T, \tag{51}$$

*and $G^{-}$ is a pseudo-inverse of $G$.*

It should be noted that the final values for mean and covariance do not depend on the choice of $G$ as a square root of:

$$A_0^{-1}(\dot{\Sigma}_0\Sigma_0^{-1}\dot{\Sigma}_0 + 2\dot{\mu}_0\dot{\mu}_0^T)(A_0^{-1})^T.$$

The reason for this is that $\mathrm{ch}(G)$ is a Taylor series in $G^2$, and so are $\mathrm{sh}(G)G^{-}$ and $G^{-}\mathrm{sh}(G)$.

For our practical implementation, we actually used these Taylor series instead of the expression of the corollary.

## 6. Comparing GIGO, xNES and Pure Rank-$\mu$ CMA-ES

### *6.1. Definitions*

In this section, we recall the xNES and pure rank-$\mu$ CMA-ES, and we describe them in the IGO framework, thus allowing a reasonable comparison with the GIGO algorithms.

#### 6.1.1. xNES

We recall a restriction of the xNES algorithm, introduced in [19] (this restriction is sufficient to describe the numerical experiments in [19]).

**Definition 9** (xNES algorithm)**.** *The xNES algorithm with sample size $N$, weights $w_i$ and learning rates $\eta_\mu$ and $\eta_\Sigma$ updates the parameters $\mu \in \mathbb{R}^d$, $A \in M_d(\mathbb{R})$ with the following rule: At each step, $N$ points $x_1, ..., x_N$ are sampled from the distribution $\mathcal{N}(\mu, AA^T)$. Without loss of generality, we assume $f(x_1) < ... < f(x_N)$. The parameter is updated according to:*

$$\mu \leftarrow \mu + \eta_\mu A G_\mu,$$

$$A \leftarrow A \exp(\eta_\Sigma G_M/2),$$

*where, setting $z_i = A^{-1}(x_i - \mu)$:*

$$G_\mu = \sum_{i=1}^{N} w_i z_i,$$

$$G_M = \sum_{i=1}^{N} w_i(z_i z_i^T - I).$$

The more general version decomposes the matrix $A$ as $\sigma B$, where $\det B = 1$, and uses two different learning rates for $\sigma$ and for $B$. We gave the version where these two learning rates are equal (in particular, for the default parameters in [19], these two learning rates are equal). This restriction of the xNES algorithm can be described in the IGO framework, provided all of the learning rates are equal (most of the elements of the proof can be found in [19] (the proposition below essentially states that xNES is a natural gradient update) or in [1]):

**Proposition 6** (xNES as IGO)**.** *The xNES algorithm with sample size $N$, weights $w_i$ and learning rates $\eta_\mu = \eta_\Sigma = \delta t$ coincides with the IGO algorithm with sample size $N$, weights $w_i$, step size $\delta t$ and in which, given the current position $(\mu_t, A_t)$, the set of Gaussians is parametrized by:*

$$\phi_{\mu_t, A_t} : (\delta, M) \mapsto \mathcal{N}\left(\mu_t + A_t\delta, \left(A_t \exp(\frac{1}{2}M)\right)\left(A_t \exp(\frac{1}{2}M)\right)^T\right),$$

*with $\delta \in \mathbb{R}^m$ and $M \in \mathrm{Sym}(\mathbb{R}^m)$.*

*The parameters maintained by the algorithm are $(\mu, A)$, and the $x_i$ are sampled from $\mathcal{N}(\mu, AA^T)$.*

**Proof.** Let us compute the IGO update in the parametrization $\phi_{\mu_t, A_t}$: we have $\delta^t = 0$, $M^t = 0$, and by using Proposition 1, we can see that for this parametrization, the Fisher information matrix at $(0, 0)$ is the identity matrix. The IGO update is therefore,

$$(\delta, M)^{t+\delta t} = (\delta, M)^t + \delta t Y_\delta(\delta, M) + \delta t Y_M(\delta, M) = \delta t Y_\delta(\delta, M) + \delta t Y_M(\delta, M),$$

where:

$$Y_\delta(\delta, M) = \sum_{i=1}^{N} w_i \nabla_\delta \ln(p(x_i|(\delta, M))$$

and:

$$Y_M(\delta, M) = \sum_{i=1}^{N} w_i \nabla_M \ln(p(x_i|(\delta, M)).$$

Since $\mathrm{tr}(M) = \log(\det(\exp(M)))$, we have:

$$\ln P_{\delta,M}(x) = -\frac{d}{2}\ln(2\pi) - \ln(\det A) - \frac{1}{2}\,\mathrm{tr}\, M - \frac{1}{2}\|\exp(-\frac{1}{2}M)A^{-1}(x - \mu - A\delta)\|^2,$$

and a straightforward computation yields:

$$Y_\delta(\delta, M) = \sum_{i=1}^{N} w_i z_i = G_\mu,$$

and:

$$Y_M(\delta, M) = \frac{1}{2}\sum_{i=1}^{N} w_i (z_i z_i^T - I) = G_M.$$

Therefore, the IGO update is:

$$\delta(t + \delta t) = \delta(t) + \delta t G_\mu,$$
$$M(t + \delta t) = M(t) + \delta t G_M,$$

or, in terms of mean and covariance matrix:

$$\mu(t + \delta t) = \mu(t) + \delta t A(t) G_\mu$$
$$A(t + \delta t) = A(t)\exp(\delta t G_M/2),$$

or:

$$\Sigma(t + \delta t) = A(t)\exp(\delta t G_M)A(t)^T.$$

This is the xNES update. □

6.1.2. Using a Square Root of the Covariance Matrix

Firstly, we recall that the IGO framework (on $\mathbb{G}_d$, for example) emphasizes the Riemannian manifold structure on $\mathbb{G}_d$. All of the algorithms studied here (including GIGO, which is not strictly speaking an IGO algorithm) define a trajectory in $\mathbb{G}_d$ (a new point for each step), and to go from a point $\theta$ to the next one ($\theta'$), we follow some curve $\gamma : [0, \delta t] \to \mathbb{G}_d$, with $\gamma(0) = \theta$, $\gamma(\delta t) = \theta'$ and $\dot{\gamma}(0)$ given by the natural gradient ($\dot{\gamma}(0) = \sum_{i=1}^{N} \hat{w}_i \tilde{\nabla}_\theta P_\theta(x_i) \in T_\theta \mathbb{G}_d$).

To be compatible with this point of view, an algorithm giving an update rule for a square root (any matrix $A$ such that $\Sigma = AA^T$: since we do not force $A$ to be symmetric, the decomposition is not unique) of the covariance matrix $A$ has to satisfy the following condition: for a given initial speed, the covariance matrix $\Sigma^{t+\delta t}$ after one step must depend only on $\Sigma^t$ and not on the square root $A^t$ chosen for $\Sigma^t$.

The xNES algorithm does satisfy this condition: consider two xNES algorithms, with the same learning rates, respectively, at $(\mu, A_1^t)$ and $(\mu, A_2^t)$, with $A_1^t (A_1^t)^T = A_2^t (A_2^t)^T$ (*i.e.*, they define the same $\Sigma^t$), using the same samples $x_i$ to compute the natural gradient update , then we will have $\Sigma_1^{t+\delta t} = \Sigma_2^{t+\delta t}$. Using the definitions of Section 6.3, we have just shown that what we will call the "xNES trajectory" is well defined.

It is also important to notice that, in order to be well defined, a natural gradient algorithm updating a square root of the covariance matrix has to specify more conditions than simply following the natural gradient.

The reason for this is that the natural gradient is a vector tangent to $\mathbb{G}_d$: it lives in a space of dimension $d(d+3)/2$ (the dimension of $\mathbb{G}_d$), whereas the vector $(\mu, A)$ lives in a space of dimension $d(d+1)$ (the dimension of $\mathbb{R}^n \times GL_n(\mathbb{R})$), which is too large: there exists infinitely many applications $t \mapsto A_t$, such that a given curve $\gamma : t \mapsto \mathcal{N}(\mu_t, \Sigma_t)$ can be written $\gamma(t) = \mathcal{N}(\mu_t, A_t A_t^T)$. This is why Theorem 5 is simply an implication, whereas Theorem 4 is an equivalence.

More precisely, let us consider $A$ in $\mathrm{GL}_d(\mathbb{R})$ and $v_A$, $v'_A$ two infinitesimal updates of $A$. Since $\Sigma = AA^T$, the infinitesimal update of $\Sigma$ corresponding to $v_A$ (resp. $v'_A$) is $v_\Sigma = Av_A^T + v_A A^T$ (resp. $v'_\Sigma = Av'^T_A + v'_A A^T$).

It is now easy to see that $v_A$ and $v'_A$ define the same direction for $\Sigma$ (*i.e.*, $v_\Sigma = v'_\Sigma$) if and only if $AM^T + MA^T = 0$, where $M = v_A - v'_A$. This is equivalent to $A^{-1}M$ antisymmetric.

For any $A \in \mathrm{M}_d(\mathbb{R})$, let us denote by $T_A$ the space of the matrices $M$, such that $A^{-1}M$ is antisymmetric or, in other words, $T_A := \{u \in \mathrm{M}_d(\mathbb{R}),\ Au^T + uA^T = 0\}$. Having a subspace $S_A$ in direct sum with $T_A$ for all $A$ is sufficient (but not necessary) to have a well-defined update rule. Namely, consider the (linear) application:

$$\begin{array}{rccc} \phi_A & : & \mathrm{M}_d(\mathbb{R}) & \to & \mathrm{S}_d(\mathbb{R}) \\ & & v_A & \mapsto & Av_A^T + v_A A^T \end{array},$$

sending an infinitesimal update of $A$ to the corresponding update of $\Sigma$. It is not bijective, but as we have seen before, $\mathrm{Ker}\, \phi_A = T_A$, and therefore, if we have, for some $U_A$,

$$\mathrm{M}_d(\mathbb{R}) = U_A \oplus T_A, \tag{52}$$

then $\phi_A|_{U_A}$ is an isomorphism. Let $v_\Sigma$ be an infinitesimal update of $\Sigma$. We choose the following update of $A$ corresponding to $v_\Sigma$:

$$v_A := (\phi_A|_{U_A})^{-1}(v_\Sigma). \tag{53}$$

Any $U_A$, such that $U_A \oplus T_A = \mathrm{M}_d(\mathbb{R})$, is a reasonable choice to pick $v_A$ for a given $v_\Sigma$. The choice $S_A = \{u \in \mathrm{M}_d(\mathbb{R}),\ Au^T - uA^T = 0\}$ has an interesting additional property; it is the orthogonal of $T_A$ for the norm:

$$\|v_A\|_\Sigma^2 := \mathrm{Tr}(v_A^T \Sigma^{-1} v_A) = \mathrm{Tr}((A^{-1}v_A)^T A^{-1} v_A). \tag{54}$$

and consequently, it can be defined without referring to the parametrization, which makes it a canonical choice. To prove this, remark that $T_A = \{M \in \mathrm{M}_d(\mathbb{R}), A^{-1}M \text{ antisymmetric}\}$ and $S_A = \{M \in \mathrm{M}_d(\mathbb{R}), A^{-1}M \text{ symmetric}\}$ and that if $M$ is symmetric and $N$ is antisymmetric, then

$$\mathrm{Tr}(M^T N) = \sum_{i,j=1}^{d} m_{ij} n_{ij} = \sum_{i=1}^{d} m_{ii} n_{ii} + \sum_{1 \leqslant i < j \leqslant d} m_{ij}(n_{ij} + n_{ji}) = 0. \tag{55}$$

Let us now show that this is the choice made by xNES and GIGO-$A$ (which are well-defined algorithms updating a square root of the covariance matrix).

**Proposition 7.** *Let $A \in M_n(\mathbb{R})$. The $v_A$ given by the xNES and GIGO-A algorithms lies in $S_A = \{u \in \mathrm{M}_d(\mathbb{R}),\ Au^T - uA^T = 0\} = S_A$.*

**Proof.** For xNES, let us write $\dot\gamma(0) = (v_\mu, v_\Sigma)$ and $v_A := \frac{1}{2} A G_M$. We have $A^{-1}v_A = \frac{1}{2}G_M$, and therefore, forcing $M$ (and $G_M$) to be symmetric in xNES is equivalent to $A^{-1}v_A = (A^{-1}v_A)^T$, which can be rewritten as $Av_A^T = v_A A^T$. For GIGO-$A$, Equation (40) shows that $\Sigma_t(J_\Sigma - J_\mu \mu_t^T)$ is symmetric, and with this fact in mind, Equation (42) shows that we have $Av_A^T = v_A A^T$ ($v_A$ is $\dot A_t$). □

### 6.1.3. Pure Rank-$\mu$ CMA-ES

We now recall the pure rank-$\mu$ CMA-ES algorithm. The general CMA-ES algorithm is described in [21].

**Definition 10** (Pure rank-$\mu$ CMA-ES algorithm)**.** *The pure rank-$\mu$ CMA-ES algorithm with sample size $N$, weights $w_i$ and learning rates $\eta_\mu$ and $\eta_\Sigma$ is defined by the following update rule: At each step, $N$ points $x_1, ..., x_N$ are sampled from the distribution $\mathcal{N}(\mu, \Sigma)$. Without loss of generality, we assume $f(x_1) < ... < f(x_N)$. The parameter is updated according to:*

$$\mu \leftarrow \mu + \eta_\mu \sum_{i=1}^{N} w_i (x_i - \mu),$$

$$\Sigma \leftarrow \Sigma + \eta_\Sigma \sum_{i=1}^{N} w_i ((x_i - \mu)(x_i - \mu)^T - \Sigma).$$

The pure rank-$\mu$ CMA-ES can also be described in the IGO framework; see, for example, [20].

**Proposition 8** (Pure rank-$\mu$ CMA-ES as IGO)**.** *The pure rank-$\mu$ CMA-ES algorithm with sample size $N$, weights $w_i$ and learning rates $\eta_\mu = \eta_\Sigma = \delta t$ coincides with the IGO algorithm with sample size $N$, weights $w_i$, step size $\delta t$ and the parametrization $(\mu, \Sigma)$.*

### *6.2. Twisting the Metric*

As we can see, the IGO framework does not allow one to recover the learning rates for xNES and pure rank-$\mu$ CMA-ES, which is a problem, since usually, the covariance learning rate is set much smaller than the mean learning rate (see either [19] or [21]).

A way to recover these learning rates is to incorporate them directly into the metric (see also blockwise GIGO, in Section 6.4). More precisely:

**Definition 11** (Twisted Fisher metric)**.** *Let $\eta_\mu, \eta_\Sigma \in \mathbb{R}$, and let $(P_\theta)_{\theta\in\Theta}$ be a family of normal probability distributions: $P_\theta = \mathcal{N}(\mu(\theta), \Sigma(\theta))$, with $\mu$ and $\Sigma$ $C^1$. We call the "$(\eta_\mu, \eta_\Sigma)$-twisted Fisher metric" the metric defined by:*

$$I_{i,j}(\eta_\mu, \eta_\Sigma)(\theta) = \frac{1}{\eta_\mu}\frac{\partial \mu^T}{\partial \theta_i}\Sigma^{-1}\frac{\partial \mu}{\partial \theta_j} + \frac{1}{\eta_\Sigma}\frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}\frac{\partial \Sigma}{\partial \theta_i}\Sigma^{-1}\frac{\partial \Sigma}{\partial \theta_j}\right). \tag{56}$$

All of the remainder of this section is simply a rewriting of the work in Section 2 with the twisted Fisher metric instead of the regular Fisher metric. We will use the term "twisted geodesic" instead of "geodesic for the twisted metric".

This approach seems to be somewhat arbitrary: arguably, the mean and the covariance play a "different role" in the definition of a Gaussian (only the covariance can affect diversity, for example), but we lack a reasonable intrinsic characterization that would make this choice of twisting more natural. This construction can be slightly generalized (see the Appendix).

The IGO flow and the IGO algorithms can be modified to take into account the twisting of the metric; the $(\eta_\mu, \eta_\Sigma)$-twisted IGO flow reads:

$$\frac{\mathrm{d}\theta^t}{\mathrm{d}t} = I(\eta_\mu, \eta_\Sigma)^{-1}(\theta)\int_X W^f_{\theta^t}(x)\nabla_\theta \ln P_\theta(x) P_{\theta^t}(\mathrm{d}x). \tag{57}$$

The only difference with (9) is that $I^{-1}(\theta)$ has been replaced by $I(\eta_\mu, \eta_\Sigma)^{-1}(\theta)$.

This leads us to the twisted IGO algorithms.

**Definition 12.** *The $(\eta_\mu, \eta_\Sigma)$-twisted IGO algorithm associated with parametrization $\theta$, sample size $N$, step size $\delta t$ and selection scheme $w$ is given by the following update rule:*

$$\theta^{t+\delta t} = \theta^t + \delta t I(\eta_\mu, \eta_\Sigma)^{-1}(\theta^t)\sum_{i=1}^N \hat{w}_i \frac{\partial \ln P_\theta(x_i)}{\partial \theta}.$$

**Definition 13.** *The $(\eta_\mu, \eta_\Sigma)$-twisted geodesic IGO algorithm associated with sample size $N$, step size $\delta t$ and selection scheme $w$ is given by the following update rule:*

$$\theta^{t+\delta t} = \exp_{\theta^t}(Y\delta t) \tag{58}$$

*where:*

$$Y = I(\eta_\mu, \eta_\Sigma)^{-1}(\theta^t) \sum_{i=1}^{N} \hat{w}_i \frac{\partial \ln P_\theta(x_i)}{\partial \theta}. \tag{59}$$

*By definition, the twisted geodesic IGO algorithm does not depend on the parametrization (but it does depend on $\eta_\mu$ and $\eta_\Sigma$).*

There is some redundancy between $\delta t$, $\eta_\mu$ and $\eta_\Sigma$: the only values actually appearing in the equations are $\delta t \eta_\mu$ and $\delta t \eta_\Sigma$. More formally:

**Proposition 9.** *Let $k, d, N \in \mathbb{N}$, $\eta_\mu, \eta_\Sigma, \delta t, \lambda_1, \lambda_2 \in \mathbb{R}$ and $w \colon [0;1] \to \mathbb{R}$.*

*The $(\eta_\mu, \eta_\Sigma)$-twisted IGO algorithm with sample size $N$, step size $\delta t$ and selection scheme $w$ coincides with the $(\lambda_1 \eta_\mu, \lambda_1 \eta_\Sigma)$-twisted IGO algorithm with sample size $N$, step size $\lambda_2 \delta t$ and selection scheme $\frac{1}{\lambda_1 \lambda_2} w$. The same is true for geodesic IGO.*

In order to obtain the twisted algorithms, the Fisher metric in IGO has to be replaced by the metric from Definition 11. In practice, the equations found by twisting the metric are exactly the equations without twisting, except that we have "forced" the learning rates $\eta_\mu$, $\eta_\Sigma$ to appear by multiplying the increments of $\mu$ and $\Sigma$ by $\eta_\mu$ and $\eta_\Sigma$.

We can now describe pure rank-$\mu$ CMA-ES and xNES with separate learning rates as twisted IGO algorithms:

**Proposition 10** (xNES as IGO)**.** *The xNES algorithm with sample size $N$, weights $w_i$ and learning rates $\eta_\mu, \eta_\sigma = \eta_B = \eta_\Sigma$ coincides with the $\frac{\eta_\mu}{\delta t}, \frac{\eta_\Sigma}{\delta t}$-twisted IGO algorithm with sample size $N$, weights $w_i$, step size $\delta t$ and in which, given the current position $(\mu_t, A_t)$, the set of Gaussians is parametrized by:*

$$(\delta, M) \mapsto \mathcal{N}\left(\mu_t + A_t \delta, \left(A_t \exp(\frac{1}{2} M)\right)\left(A_t \exp(\frac{1}{2} M)\right)^T\right),$$

*with $\delta \in \mathbb{R}^m$ and $M \in \mathrm{Sym}(\mathbb{R}^m)$.*

*The parameters maintained by the algorithm are $(\mu, A)$, and the $x_i$ are sampled from $\mathcal{N}(\mu, AA^T)$.*

**Proposition 11** (Pure rank-$\mu$ CMA-ES as IGO)**.** *The pure rank-$\mu$ CMA-ES algorithm with sample size $N$, weights $w_i$ and learning rates $\eta_\mu$ and $\eta_\Sigma$ coincides with the $(\frac{\eta_\mu}{\delta t}, \frac{\eta_\Sigma}{\delta t})$-twisted IGO algorithm with sample size $N$, weights $w_i$, step size $\delta t$ and the parametrization $(\mu, \Sigma)$.*

The proofs of these two statements are an easy rewriting of their non-twisted counterparts: one can return to the non-twisted metric (up to a $\eta_\Sigma$ factor) by changing $\mu$ to $\frac{\sqrt{\eta_\sigma}}{\sqrt{\eta_\mu}}\mu$.

We give the equations of the twisted geodesics of $\mathbb{G}_d$ in the Appendix.

### *6.3. Trajectories of Different IGO Steps*

As we have seen, two different IGO algorithms (or an IGO algorithm and the GIGO algorithm) coincide at first order in $\delta t$ when $\delta t \to 0$. In this section, we study the differences between pure

rank-$\mu$ CMA-ES, xNES and GIGO by looking at the second order in $\delta t$, and in particular, we show that xNES and GIGO do not coincide in the general case.

We view the updates done by one step of the algorithms as paths on the manifold $\mathbb{G}_d$, from $(\mu(t), \Sigma(t))$ to $(\mu(t+\delta t), \Sigma(t+\delta t))$, where $\delta t$ is the time step of our algorithms, seen as IGO algorithms. More formally:

**Definition 14.** *(1) We call the GIGO update trajectory the application:*

$$T_{\text{GIGO}} \colon (\mu, \Sigma, v_\mu, v_\Sigma) \mapsto \left(\delta t \mapsto \exp_{\mathcal{N}(\mu, AA^T)}(\delta t \eta_\mu v_\mu, \delta t \eta_\Sigma v_\Sigma)\right).$$

*(*$\exp$ *is the exponential of the Riemannian manifold* $\mathbb{G}_d(\eta_\mu, \eta_\Sigma)$*)*

*(2) We call the xNES update trajectory the application:*

$$T_{\text{xNES}} \colon (\mu, \Sigma, v_\mu, v_\Sigma) \mapsto \left(\delta t \mapsto \mathcal{N}(\mu + \delta t \eta_\mu v_\mu, A \exp[\eta_\Sigma \delta t A^{-1} v_\Sigma (A^{-1})^T] A^T)\right),$$

*with* $AA^T = \Sigma$*. The application above does not depend on the choice of a square root* $A$*.*

*(3) We call the CMA-ES update trajectory the application:*

$$T_{\text{CMA}} \colon (\mu, \Sigma, v_\mu, v_\Sigma) \mapsto \left(\delta t \mapsto \mathcal{N}(\mu + \delta t \eta_\mu v_\mu, AA^T + \delta t \eta_\Sigma v_\Sigma)\right).$$

*These applications map the set of tangent vectors to* $\mathbb{G}_d$ *(*$T\mathbb{G}_d$*) to the curves in* $\mathbb{G}_d(\eta_\mu, \eta_\Sigma)$*.*

*We will also use the following notation:* $\mu_{\text{GIGO}} := \phi_\mu \circ T_{GIGO}$*,* $\mu_{\text{xNES}} := \phi_\mu \circ T_{xNES}$*,* $\mu_{\text{CMA}} := \phi_\mu \circ T_{CMA}$*,* $\Sigma_{\text{GIGO}} := \phi_\Sigma \circ T_{GIGO}$*,* $\Sigma_{\text{xNES}} := \phi_\Sigma \circ T_{xNES}$ *and* $\Sigma_{\text{CMA}} := \phi_\Sigma \circ T_{CMA}$*, where* $\phi_\mu$ *(resp.* $\phi_\Sigma$*) extracts the* $\mu$*-component (resp. the* $\Sigma$*-component) of a curve.*

*In particular,* $\text{Im}(\phi_\mu) \subset \mathbb{R}^d$ *and* $\text{Im}(\phi_\Sigma) \subset P_d$*, where* $P_d$ *(the set of real symmetric positive-definitematrices of dimension* $d$*) is seen as a subset of* $\mathbb{R}^{d^2}$*.*

For instance, $T_{\text{GIGO}}(\mu, \Sigma, v_\mu, v_\Sigma)(\delta t)$ gives the position (mean and covariance matrix) of the GIGO algorithm after a step of size $\delta t$, while $\mu_{\text{GIGO}}$ and $\Sigma_{\text{GIGO}}$ give, respectively, the mean component and the covariance component of this position.

This formulation ensures that the trajectories we are comparing had the same initial position and the same initial speed, which is the case provided the sampled points (the values directly sampled from $\mathcal{N}(\mu, \Sigma)$, not from $\mathcal{N}(0, I)$ and transformed) are the same.

Different IGO algorithms coincide at first order in $\delta t$. The following proposition gives the second order expansion of the trajectories of the algorithms.

**Proposition 12** (Second derivatives of the trajectories)**.** *We have:*

$$\mu_{\text{GIGO}}(\mu, \Sigma, v_\mu, v_\Sigma)''(0) = \eta_\mu \eta_\Sigma v_\Sigma \Sigma_0^{-1} v_\mu,$$

$$\mu_{\text{xNES}}(\mu, \Sigma, v_\mu, v_\Sigma)''(0) = \mu_{\text{CMA}}(\mu, \Sigma, v_\mu, v_\Sigma)''(0) = 0,$$

$$\Sigma_{\text{GIGO}}(\mu, \Sigma, v_\mu, v_\Sigma)''(0) = \eta_\Sigma^2 v_\Sigma \Sigma^{-1} v_\Sigma - \eta_\mu \eta_\Sigma v_\mu v_\mu^T,$$

$$\Sigma_{\text{xNES}}(\mu, \Sigma, v_\mu, v_\Sigma)''(0) = \eta_\Sigma^2 v_\Sigma \Sigma^{-1} v_\Sigma,$$

$$\Sigma_{\text{CMA}}(\mu, \Sigma, v_\mu, v_\Sigma)''(0) = 0.$$

**Proof.** We can immediately see that the second derivatives of $\mu_{\text{xNES}}$, $\mu_{\text{CMA}}$ and $\Sigma_{\text{CMA}}$ are zero. Next, we have:

$$\begin{aligned}\Sigma_{\text{xNES}}(\mu, \Sigma, v_\mu, v_\Sigma)(t) &= A \exp[tA^{-1}\eta_\Sigma v_\Sigma (A^{-1})^T] A^T \\ &= AA^T + t\eta_\Sigma v_\Sigma + \frac{t^2}{2}\eta_\Sigma^2 v_\Sigma (A^{-1})^T A^{-1} v_\Sigma + o(t^2) \\ &= \Sigma + t\eta_\Sigma v_\Sigma + \frac{t^2}{2}\eta_\Sigma^2 v_\Sigma \Sigma^{-1} v_\Sigma + o(t^2).\end{aligned}$$

The expression of $\Sigma_{\text{xNES}}(\mu, \Sigma, v_\mu, v_\Sigma)''(0)$ follows.

Now, for GIGO, let us consider the geodesic starting at $(\mu_0, \Sigma_0)$ with initial speed $(\eta_\mu v_\mu, \eta_\Sigma v_\Sigma)$. By writing $J_\mu(0) = J_\mu(t)$, we find $\dot{\mu}(t) = \Sigma(t)\Sigma_0^{-1}\dot{\mu}_0$. We then easily have $\ddot{\mu}(0) = \dot{\Sigma}_0\Sigma_0^{-1}\dot{\mu}_0$. In other words:

$$\mu_{\text{GIGO}}(\mu, \Sigma, v_\mu, v_\Sigma)''(0) = \eta_\mu \eta_\Sigma v_\Sigma \Sigma_0^{-1} v_\mu.$$

Finally, by using Theorem 4 and differentiating, we find:

$$\ddot{\Sigma} = \eta_\Sigma \dot{\Sigma}(J_\Sigma - J_\mu \mu^T) - \eta_\Sigma \Sigma J_\mu \dot{\mu}^T,$$

$$\ddot{\Sigma}_0 = \eta_\Sigma \dot{\Sigma}_0 \frac{1}{\eta_\Sigma}\Sigma_0^{-1}\dot{\Sigma}_0 - \frac{\eta_\Sigma}{\eta_\mu}\dot{\mu}_0\dot{\mu}_0^T = \eta_\Sigma^2 v_\Sigma \Sigma_0^{-1} v_\Sigma - \eta_\Sigma \eta_\mu v_\mu v_\mu^T.$$

□

In order to interpret these results, we will look at what happens in dimension one. In higher dimensions, we can suppose that the algorithms exhibit a similar behavior, but an exact interpretation is more difficult for GIGO in $\mathbb{G}_d$.

- In [19], it has been noted that xNES converges to quadratic minima slower than CMA-ES and that it is less subject to premature convergence. That fact can be explained by observing that the mean update is exactly the same for CMA-ES and xNES, whereas xNES tends to have a higher variance (Proposition 12 shows this at order two, and it is easy to see that in dimension one, for any $\mu$, $\Sigma$, $v_\mu$, $v_\Sigma$, we have $\Sigma_{\text{xNES}}(\mu, \Sigma, v_\mu, v_\Sigma) > \Sigma_{\text{CMA}}(\mu, \Sigma, v_\mu, v_\Sigma)$).
- At order two, GIGO moves the mean faster than xNES and CMA-ES if the standard deviation is increasing and more slowly if it is decreasing. This seems to be a reasonable behavior (if the covariance is decreasing, then the algorithm is presumably close to a minimum, and it should not leave the area too quickly). This remark holds only for isolated steps, because we do not take into account the evolution of the variance.
- The geodesics of $\mathbb{G}_1$ are half-circles (see Figure 2 below; we recall that $\mathbb{G}_1$ is the Poincaré half-plane). Consequently, if the mean is supposed to move (which always happens), then $\sigma \to 0$ when $\delta t \to \infty$. For example, a step whose initial speed has no component on the standard deviation will always decrease it. See also Proposition 15, about the optimization of a linear function.
- For the same reason, for a given initial speed, the update of $\mu$ always stays bounded as a function of $\delta t$: it is not possible to make one step of the GIGO algorithm go further than a fixed point by increasing $\delta t$. Still, the geodesic followed by GIGO changes at each step, so the mean of the overall algorithm is not bounded.

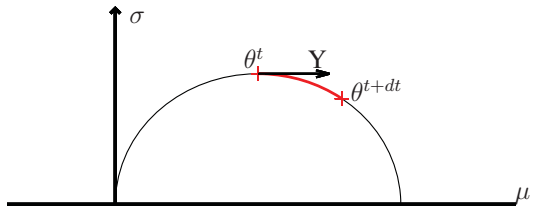



**Figure 2.** One step of the geodesic IGO (GIGO) update.

We now show that xNES follows the geodesics of $\mathbb{G}_d$ if the mean is fixed, but that xNES and GIGO do not coincide otherwise.

**Proposition 13** (xNES is not GIGO in the general case)**.** *Let $\mu, v_\mu \in \mathbb{R}^d$, $A \in \mathrm{GL}_d$, $v_\Sigma \in \mathrm{M}_d$.*

*Then, the GIGO and xNES updates starting at $\mathcal{N}(\mu, \Sigma)$ with initial speeds $v_\mu$ and $v_\Sigma$ follow the same trajectory if and only if the mean remains constant. In other words:*

$T_{\mathrm{GIGO}}(\mu, \Sigma, v_\mu, v_\Sigma) = T_{\mathrm{xNES}}(\mu, \Sigma, v_\mu, v_\Sigma)$ *if and only if* $v_\mu = 0$.

**Proof.** If $v_\mu = 0$, then we can compute the GIGO update by using Theorem 4: since $J_\mu = 0$, $\dot{\mu} = 0$, and $\mu$ remains constant. Now, we have $J_\Sigma = \Sigma^{-1}\dot{\Sigma}$; this is enough information to compute the update. Since this quantity is also preserved by the xNES algorithm (see, for example, the proof of Proposition 14), the two updates coincide.

If $v_\mu \neq 0$, then $\Sigma_{\mathrm{xNES}}(\mu, \Sigma, v_\mu, v_\Sigma)''(0) - \Sigma_{\mathrm{GIGO}}(\mu, \Sigma, v_\mu, v_\Sigma)''(0) = \eta_\mu \eta_\Sigma v_\mu v_\mu^T \neq 0$ and, in particular, $T_{\mathrm{GIGO}}(\mu, \Sigma, v_\mu, v_\Sigma) \neq T_{\mathrm{xNES}}(\mu, \Sigma, v_\mu, v_\Sigma)$. □

### *6.4. Blockwise GIGO*

Although xNES is not GIGO, it is possible to define a family of algorithms extending GIGO and including xNES, by decomposing our family of probability distributions as a product and by following the restricted geodesics simultaneously.

**Definition 15** (Splitting)**.** *Let $\Theta$ be a Riemannian manifold. A splitting of $\Theta$ is $n$ manifolds $\Theta_1, ..., \Theta_n$ and a diffeomorphism $\Theta \cong \Theta_1 \times ... \times \Theta_n$. If for all $x \in \Theta$, for all $1 \leqslant i < j \leqslant n$, we also have $T_{i,x}M \perp T_{j,x}M$ as subspaces of $T_xM$ (see Notation 2), then the splitting is said to be compatible with the Riemannian structure. If the Riemannian manifold is not ambiguous, we will simply write a "compatible splitting".*

We now give some notation, and we define the blockwise GIGO update:

**Notation 2.** *Let $\Theta$ be a Riemannian manifold, $\Theta_1, ..., \Theta_n$ a splitting of $\Theta$, $\theta = (\theta_1, ..., \theta_n) \in \Theta$, $Y \in T_\theta\Theta$ and $1 \leqslant i \leqslant n$.*

- *We denote by $\Theta_{\theta,i}$ the Riemannian manifold*

$$\{\theta_1\} \times ... \times \{\theta_{i-1}\} \times \Theta_i \times \{\theta_{i+1}\} \times ... \times \{\theta_n\},$$

*with the metric induced from $\Theta$. There is a canonical isomorphism of vector spaces $T_\theta\Theta = \oplus_{i=1}^n T\Theta_{\theta,i}$. Moreover, if the splitting is compatible, it is an isomorphism of Euclidean spaces.*
- *We denote by $\Phi_{\theta,i}$ the exponential at $\theta$ of the manifold $\Theta_{\theta,i}$.*

**Definition 16** (Blockwise GIGO update)**.** *Let $\Theta_1, ..., \Theta_n$ be a compatible splitting. The blockwise GIGO algorithm in $\Theta$ with splitting $\Theta_1, ..., \Theta_n$ associated with sample size $N$, step sizes $\delta t_1, ..., \delta t_n$ and selection scheme $w$ is given by the following update rule:*

$$\theta \leftarrow (\theta_1^{t+\delta t_1}, ..., \theta_n^{t+\delta t_n}) \tag{60}$$

*where:*

$$Y = I^{-1}(\theta^t) \sum_{i=1}^N \hat{w}_i \frac{\partial \ln P_\theta(x_i)}{\partial \theta}, \tag{61}$$

$$\theta_k^{t+\delta t_k} = \Phi_{\theta^t,k}(\delta t_k Y_k), \tag{62}$$

*with $Y_k$ the $T\Theta_{\theta,k}$-component of $Y$. This update only depends on the splitting (and not on the parametrization inside each $\Theta_k$).*

The compatibility condition ensures that the natural gradient of $W_{\theta^t}^f$ (defined in Section 2.2) in the whole manifold $\Theta$ really is the sum of the gradients of this same function in the submanifolds $\Theta_k$. A practical consequence is that the $Y_k$ in Equation (62) can be computed simply by taking the natural gradient in $\Theta_k$:

$$Y_k = I_k^{-1}(\theta_i^t) \sum_{i=1}^N \hat{w}_i \frac{\partial \ln P_\theta(x_i)}{\partial \theta_k}, \tag{63}$$

where $I_k$ is the metric of $\Theta_k$.

Since blockwise GIGO only depends on the splitting (and the tunable parameters: sample size, step sizes and selection scheme), it can be thought of as almost parametrization-invariant.

Notice that blockwise GIGO updates and twisted GIGO updates are two different things: firstly, blockwise GIGO can be defined on any manifold with a compatible splitting, whereas twisted GIGO (and twisted IGO) are only defined for Gaussians. However, even in $\mathbb{G}_d(\eta_\mu, \eta_\Sigma)$, with the splitting $(\mu, \Sigma)$, these two algorithms are different: for instance, if $\eta_\mu = \eta_\Sigma$ and $\delta t = 1$, then the twisted GIGO is the regular GIGO algorithm, whereas blockwise GIGO is not (actually, we will prove that it is the xNES algorithm). The only thing blockwise GIGO and twisted GIGO have in common is that they are compatible with the $(\eta_\mu, \eta_\Sigma)$-twisted IGO flow Equation (57): a parameter $\theta^t$ following these updates with $\delta t \to 0$ and $N \to \infty$ is a solution of Equation (57).

We now have a new description of the xNES algorithm:

**Proposition 14** (xNES is a Blockwise GIGO algorithm)**.** *The Blockwise GIGO algorithm in $\mathbb{G}_d$ with splitting $\Phi : \mathcal{N}(\mu, \Sigma) \mapsto (\mu, \Sigma)$, sample size $N$, step sizes $\delta t_\mu, \delta t_\Sigma$ and selection scheme $w$ coincides with the xNES algorithm with sample size $N$, weights $w_i$ and learning rates $\eta_\mu = \delta t_\mu, \eta_\sigma = \eta_B = \delta t_\Sigma$.*

**Proof.** Firstly, notice that the splitting $(\mu, \Sigma)$ is compatible, by Proposition 1.

Now, let us compute the Blockwise GIGO update: we have $\mathbb{G}_d \cong \mathbb{R}^d \times P_d$, where $P_d$ is the space of real positive-definite matrices of dimension $d$. We have $\Theta_{\theta^t,1} = (\mathbb{R}^d \times \{\Sigma^t\}) \hookrightarrow \mathbb{G}_d$, $\Theta_{\theta^t,2} = (\{\mu^t\} \times P_d) \hookrightarrow \mathbb{G}_d$. The induced metric on $\Theta_{\theta^t,1}$ is the Euclidean metric, so we have:

$$\mu \leftarrow \mu^t + \delta t_1 Y_\mu.$$

Since we have already shown (using the notation in Definition 9) that $Y_\mu = AG_\mu$ (in the proof of Proposition 6), we find:

$$\mu \leftarrow \mu^t + \delta t_1 AG_\mu.$$

On $\Theta_{\theta^t,2}$, we have the following Lagrangian for the geodesics:

$$\mathcal{L}(\Sigma, \dot{\Sigma}) = \frac{1}{2}\operatorname{tr}(\dot{\Sigma}\Sigma^{-1}\dot{\Sigma}\Sigma^{-1}).$$

By applying Noether's theorem, we find that

$$J_\Sigma = \Sigma^{-1}\dot{\Sigma}$$

is invariant along the geodesics of $\Theta_{\theta^t,2}$, so they are defined by the equation $\dot{\Sigma} = \Sigma J_\Sigma = \Sigma\Sigma_0^{-1}\dot{\Sigma}_0$ (and therefore, any update preserving the invariant $J_\Sigma$ will satisfy this first-order differential equation and follow the geodesics of $\Theta_{\theta^t,2}$). The xNES update for the covariance matrix is given by $A(t) = A_0 \exp(tG_M/2)$. Therefore, we have $\Sigma(t) = A_0 \exp(tG_M)A_0^T$, $\Sigma^{-1}(t) = (A_0^{-1})^T \exp(-tG_M)A_0^{-1}$, $\dot{\Sigma}(t) = A_0 \exp(tG_M)G_M A_0^T$ and, finally, $\Sigma^{-1}(t)\dot{\Sigma}(t) = (A_0^{-1})^T G_M A_0^T = \Sigma_0^{-1}\dot{\Sigma}_0$. Therefore, xNES preserves $J_\Sigma$, and therefore, xNES follows the geodesics of $\Theta_{\theta^t,2}$ (notice that we had already proven this in Proposition 13, since we are looking at the geodesics of $\mathbb{G}_d$ with a fixed mean). □

Although blockwise GIGO is somewhat "less natural" than GIGO, it can be easier to compute for some splittings (as we have just seen), and in the case of the Gaussian distributions, the mean-covariance splitting seems reasonable.

## 7. Numerical Experiments

We conclude this article with some numerical experiments to compare the behavior of GIGO, xNES and pure rank-$\mu$ CMA-ES (we give the pseudocodes for these algorithms in the Appendix). We made two series of tests. The first one is a performance test, using classical benchmark functions and the settings from [19]. The goal of the second series of tests is to illustrate the computations in Section 6.3 by plotting the trajectories (standard deviation *versus* mean) of these three algorithms in dimension one.

The source code is available at [22].

*7.1. Benchmarking*

For the first series of experiments, presented in Figure 3, we used the following parameters, taken from [19] (we recall that xNES and pure rank-$\mu$ CMA-ES are seen as IGO algorithms):

- Varying dimension.
- Sample size: $\lfloor 4 + 3\log(d) \rfloor$.
- Weights: $w_i = \frac{\max(0,\log(\frac{n}{2}+1)-\log(i)}{\sum_{j=1}^{N}\max(0,\log(\frac{n}{2}+1)-\log(j)} - \frac{1}{N}$.
- IGO step size and learning rates: $\delta t = 1, \eta_\mu = 1, \eta_\Sigma = \frac{3}{5}\frac{3+\log(d)}{d\sqrt{d}}$..
- Initial position: $\theta^0 = \mathcal{N}(x_0, I)$, where $x_0$ is a random point of the circle with center zero, and radius $10$.
- Euler method for GIGO: Number of steps: $100$. We used the GIGO-A variant of the algorithm. No significant difference was noticed with GIGO-$\Sigma$ or with the exact GIGO algorithm. The only advantage of having an explicit solution of the geodesic equations is that the update is quicker to compute.
- We chose not to use the exact expression of the geodesics for this benchmarking to show that having to use the Euler method is fine. However, we did run the tests, and the results are basically the same as GIGO-$A$.

We plot the median number of runs to achieve target fitness ($10^{-8}$). Each algorithm has been tested in dimension $2$, $4$, $8$, $16$, $32$ and $64$: a missing point means that all runs converged prematurely.

7.1.1. Failed Runs

In Figure 3, a point is plotted even if only one run was successful. Below is the list of the settings for which at least one run converged prematurely.

- Only one run reached the optimum for the cigar-tablet function with CMA-ES in dimension eight.
- Seven runs (out of $24$) reached the optimum for the Rosenbrock function with CMA-ES in dimension $16$.
- About half of the runs reached the optimum for the sphere function with CMA-ES in dimension four.

| | | |
|---|---|---|
| Dimension | $d$ | From 2 to 64 |
| Sample size | $N$ | $4 + 3\log(d)$ |
| Weights | $(w_i)_{i\in[1,N]}$ | $\frac{\max(0,\log(\frac{n}{2}+1)-\log(i)}{\sum_{j=1}^{N}\max(0,\log(\frac{n}{2}+1)-\log(j)} - \frac{1}{N}$ |
| IGO step size | $\delta t$ | 1 |
| Mean learning rate | $\eta_\mu$ | 1 |
| Covariance learning rate | $\eta_\Sigma$ | $\frac{3}{5}\frac{3+\log(d)}{d\sqrt{d}}$ |
| Euler step-size (for GIGO only) | $h$ | 0.01(100 steps) |
| GIGO implementation | | GIGO-$A$ |
| Sphere function | | $x \mapsto \sum_{i=1}^{d} x_i^2$ |
| Cigar-tablet | | $x \mapsto x_1^2 + \sum_{i=2}^{d-1} 10^4 x_i^2 + 10^8 x_d^2$ |
| Rosenbrock | | $x \mapsto \sum_{i=1}^{d-1}(100(x_i^2 - x_{i+1})^2 + (x_i - 1)^2)$ |
| $x$-axis | | Dimension |
| $y$-axis | | Number of function calls to reach fitness $10^{-8}$. |

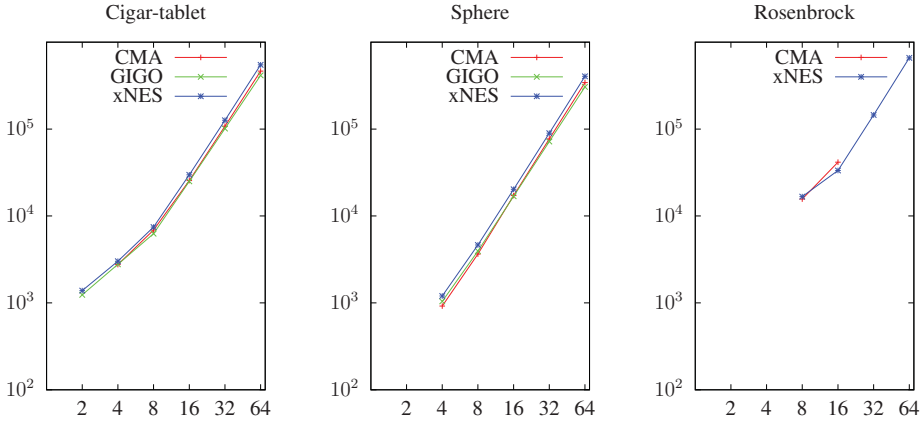



**Figure 3.** Median number of function calls to reach $10^{-8}$ fitness on 24 runs for: sphere function, cigar-tablet function and Rosenbrock function. Initial position $\theta^0 = \mathcal{N}(x_0, I)$, with $x_0$ uniformly distributed on the circle of center zero and radius 10. We recall that the "CMA-ES" algorithm here is using the so-called pure rank-$\mu$ CMA-ES update.

For the following settings, all runs converged prematurely.

- GIGO did not find the optimum of the Rosenbrock function in any dimension.
- CMA-ES did not find the optimum of the Rosenbrock function in dimension 2, 4, 32 and 64.
- All of the runs converged prematurely for the cigar-tablet function in dimension two with CMA-ES, for the sphere function in dimension two for all algorithms and for the Rosenbrock function in dimension two and four for all algorithms.

7.1.2. Discussion

As the last item in Section 7.1.1 shows, all of the algorithms converge prematurely in a low dimension, probably because the covariance learning rate has been set too high (or because the sample size is too small). This is different from the results in [19].

This remark aside, as noted in [19], the xNES algorithm shows more robustness than CMA-ES and GIGO: it is the only algorithm able to find the minimum of the Rosenbrock function in high dimensions. However, its convergence is consistently slower.

In terms of performance, when both of them work, pure rank-$\mu$ CMA-ES (or equivalently, IGO in the parametrization $(\mu, \Sigma)$) and GIGO are extremely close (GIGO is usually a bit better). An advantage of GIGO is that it is theoretically defined for any $\delta t$, $\eta_\Sigma$, whereas the covariance matrix maintained by CMA-ES (not only pure rank-$\mu$ CMA-ES) can stop being positive definite if $\eta_\Sigma \delta t > 1$. However, in that case, the GIGO algorithm is prone to premature convergence (remember Figure 2 and see Proposition 15 below), and in practice, the learning rates are much smaller.

*7.2. Plotting Trajectories in* $\mathbb{G}_1$

We want the second series of experiments to illustrate the remarks about the trajectories of the algorithms in Section 6.3, so we decided to take a large sample size to limit randomness, and we chose a fixed starting point for the same reason. We use the weights below because of the property of quantile improvement proven in [23]: the $1/4$-quantile will improve at each step. The parameters we used were the following:

- Sample size: $\lambda = 5,000$
- Dimension one only.
- Weights: $w = 4\mathbf{1}_{q \leqslant 1/4}$ $(w_i = 4.\mathbf{1}_{i \leqslant 1,250})$
- IGO step size and learning rates: $\eta_\mu = 1, \eta_\Sigma = \frac{3}{5}\frac{3+\log(d)}{d\sqrt{d}} = 1.8$, varying $\delta t$.
- Initial position: $\theta^0 = \mathcal{N}(10, 1)$
- Dots are placed at $t = 0, 1, 2 \ldots$ (except for the graph $\delta t = 1.5$, for which there is a dot for each step).

Figures 4–8 show the optimization of $x \mapsto x^2$, and Figures 9–11 show the optimization of $x \mapsto -x$.

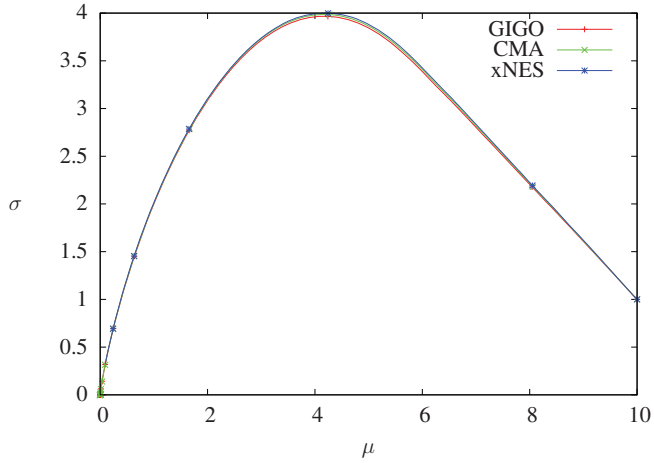



**Figure 4.** Trajectories of GIGO, CMA and xNES optimizing $x \mapsto x^2$ in dimension one with $\delta t = 0.01$, sample size $5000$, weights $w_i = 4.\mathbf{1}_{i \leqslant 1250}$ and learning rates $\eta_\mu = 1$, $\eta_\Sigma = 1.8$. One dot every $100$ steps. All algorithms exhibit a similar behavior

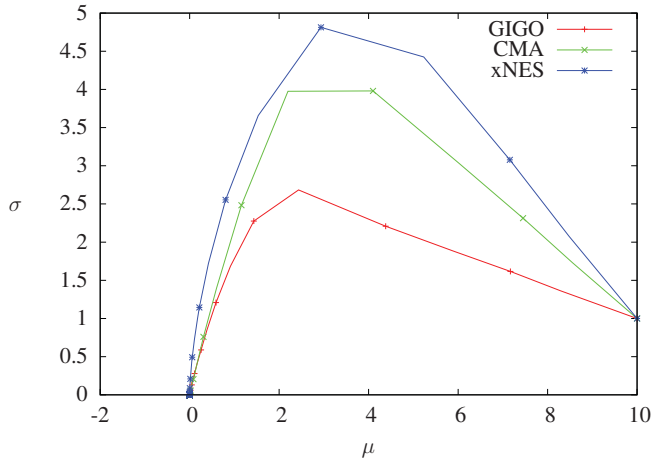



**Figure 5.** Trajectories of GIGO, CMA and xNES optimizing $x \mapsto x^2$ in dimension one with $\delta t = 0.5$, sample size $5000$, weights $w_i = 4.\mathbf{1}_{i \leqslant 1250}$ and learning rates $\eta_\mu = 1$, $\eta_\Sigma = 1.8$. One dot every two steps. Stronger differences. Notice that after one step, the lowest mean is still GIGO ($\sim 8.5$, whereas xNES is around $8.75$), but from the second step, GIGO has the highest mean, because of the lower variance.

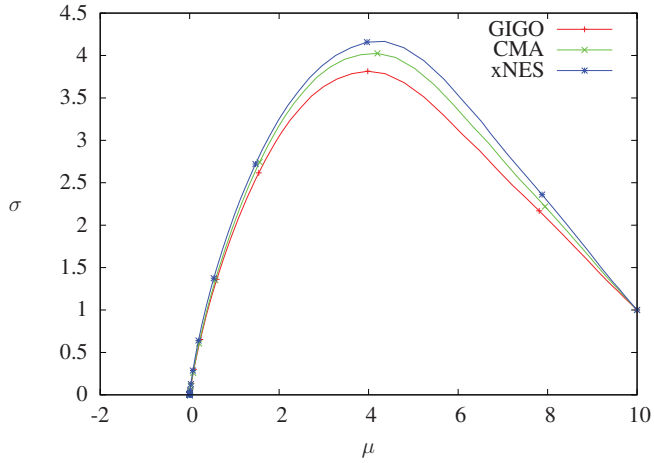



**Figure 6.** Trajectories of GIGO, CMA and xNES optimizing $x \mapsto x^2$ in dimension one with $\delta t = 0.1$, sample size $5000$, weights $w_i = 4.\mathbf{1}_{i\leqslant 1250}$ and learning rates $\eta_\mu = 1$, $\eta_\Sigma = 1.8$. One dot every $10$ steps. All algorithms exhibit a similar behavior, and differences start to appear. It cannot be seen on the graph, but the algorithm closest to zero after $400$ steps is CMA ($\sim 1.10^{-16}$, followed by xNES ($\sim 6.10^{-16}$) and GIGO ($\sim 2.10^{-15}$).

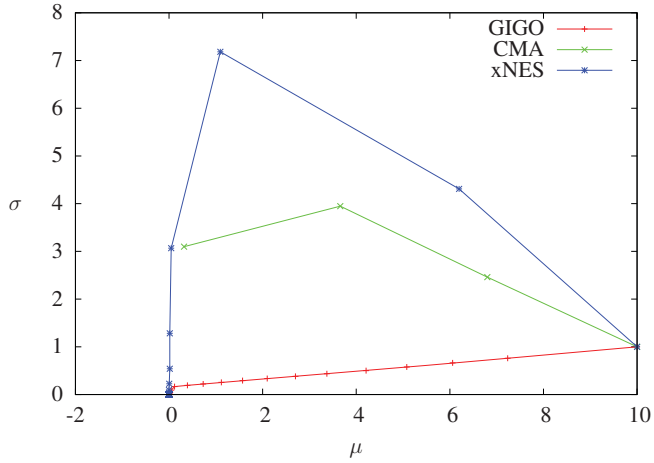



**Figure 7.** Trajectories of GIGO, CMA and xNES optimizing $x \mapsto x^2$ in dimension one with $\delta t = 1$, sample size $5000$, weights $w_i = 4.\mathbf{1}_{i\leqslant 1250}$ and learning rates $\eta_\mu = 1$, $\eta_\Sigma = 1.8$. One dot per step. The CMA-ES algorithm fails here, because at the fourth step, the covariance matrix is not positive definite anymore (it is easy to see that the CMA-ES update is always defined if $\delta t \eta_\Sigma < 1$, but this is not the case here). Furthermore, notice (see also Proposition 15) that at the first step, GIGO decreases the variance, whereas the $\sigma$-component of the IGO speed is positive.

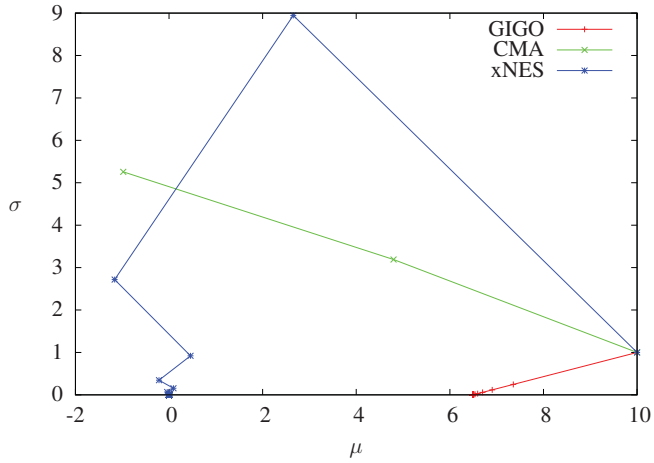



**Figure 8.** Trajectories of GIGO, CMA and xNES optimizing $x \mapsto x^2$ in dimension one with $\delta t = 1.5$, sample size $5000$, weights $w_i = 4.\mathbf{1}_{i\leqslant 1250}$ and learning rates $\eta_\mu = 1$, $\eta_\Sigma = 1.8$. One dot per step. Same as $\delta t = 1$ for CMA. GIGO converges prematurely.

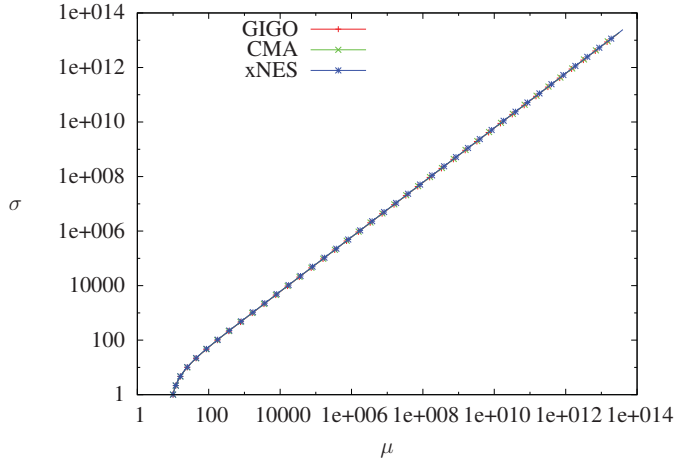



**Figure 9.** Trajectories of GIGO, CMA and xNES optimizing $x \mapsto -x$ in dimension one with $\delta t = 0.01$, sample size $5000$, weights $w_i = 4.\mathbf{1}_{i\leqslant 1250}$ and learning rates $\eta_\mu = 1$, $\eta_\Sigma = 1.8$. One dot every $100$ steps. Almost the same for all algorithms.

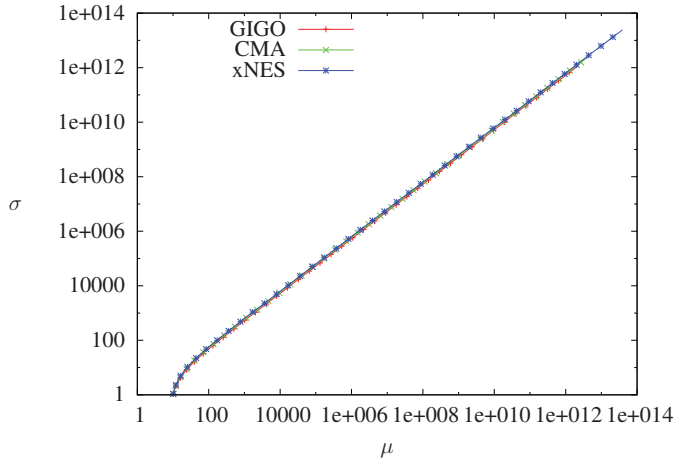



**Figure 10.** Trajectories of GIGO, CMA and xNES optimizing $x \mapsto -x$ in dimension one with $\delta t = 0.1$, sample size $5000$, weights $w_i = 4.\mathbf{1}_{i \leqslant 1250}$ and learning rates $\eta_\mu = 1$, $\eta_\Sigma = 1.8$. One dot every $10$ steps. It is not obvious on the graph, but xNES is faster than CMA, which is faster than GIGO.

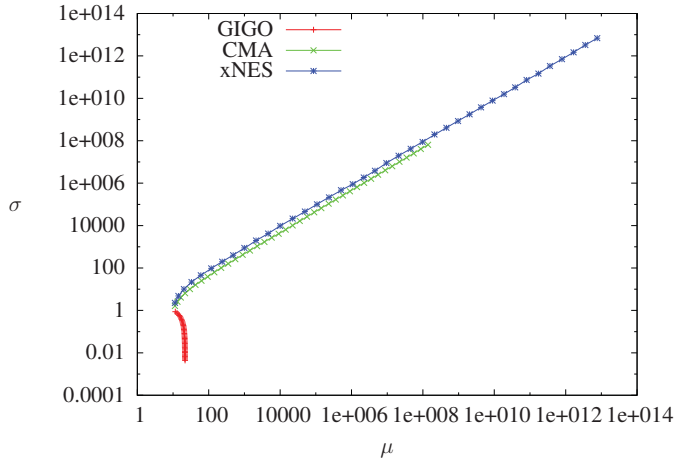



**Figure 11.** Trajectories of GIGO, CMA and xNES optimizing $x \mapsto -x$ in dimension one with $\delta t = 1$, sample size $5,000$, weights $w_i = 4.\mathbf{1}_{i \leqslant 1,250}$ and learning rates $\eta_\mu = 1$, $\eta_\Sigma = 1.8$. One dot per step. GIGO converges, for the reasons discussed earlier.

Figures 7, 8 and 11 show that when $\delta t \geqslant 1$, GIGO reduces the covariance, even at the first step. More generally, when using the GIGO algorithm in $\tilde{\mathbb{G}}_d$ for the optimization of a linear function, there exists a critical step size $\delta t_{\mathrm{cr}}$ (depending on the learning rates $\eta_\mu, \eta_\sigma$ and on the weights $w_i$),

above which, GIGO will converge, and we can compute its value when the weights are of the form $\mathbf{1}_{q\leqslant q_0}$ (for $q_0 \geqslant 0.5$, the discussion is not relevant, because in that case, even the IGO flow converges prematurely. Compare with the critical $\delta t$ of the smoothed cross entropy method and IGO-ML in [1]).

**Proposition 15.** *Let $d \in \mathbb{N}$, $k$, $\eta_\mu$, $\eta_\sigma \in \mathbb{R}_+^*$; let $w = k.\mathbf{1}_{q\leqslant q_0}$; and let*

$$\begin{array}{rccc} g & : & \mathbb{R}^d & \to & \mathbb{R} \\ & & x & \mapsto & -x_1 \end{array}.$$

*Let $\mu_n$ be the first coordinate of the mean, and let $\sigma_n^2$ be the variance (at step $n$) maintained by the $(\eta_\mu, \eta_\sigma)$-twisted geodesic IGO algorithm in $\tilde{\mathbb{G}}_d$ associated with selection scheme $w$, sample size $\infty$ and step size $\delta t$, when optimizing $g$ ("sample size $\infty$" meaning the limit of the update when the sample size tends to infinity, which is deterministic [1]).*

*There exists $\delta t_{cr}$, such that:*

- *if $\delta t > \delta t_{cr}$, $(\sigma_n)$ converges to zero with exponential speed and $(\mu_n)$ converges.*
- *if $\delta t = \delta t_{cr}$, $(\sigma_n)$ remains constant and $(\mu_n)$ tends to $\infty$ with linear speed.*
- *if $0 < \delta t < \delta t_{cr}$, both $(\sigma_n)$ and $\mu_n$ tend to $\infty$ with exponential speed.*

The proof and the expression of $\delta t_{\mathrm{cr}}$ can be found in the Appendix.

In the case corresponding to $k = 4$, $n = 1$, $q_0 = 1/4$, $\eta_\mu = 1$, $\eta_\sigma = 1.8$, we find:

$$\delta t_{\mathrm{cr}} \approx 0.84. \tag{64}$$

## 8. Conclusions

We introduced the geodesic IGO algorithm, and we showed that in the case of Gaussian distributions, Noether's theorem directly gives a first order equation satisfied by the geodesics. In terms of performance, the GIGO algorithm is similar to pure rank-$\mu$ CMA-ES, which is rather encouraging: it would be interesting to test GIGO on real problems. Moreover, GIGO is a reasonable and totally parametrization-invariant algorithm (provided we can compute the solution of the equations of the geodesics), and as such, it should be studied for other families of probability distributions, like Bernoulli distributions (although in that case, the Riemannian exponential is not defined if the step size is too large, because the length of the geodesics is finite). Noether's theorem could be a crucial tool for this.

We also showed that xNES and GIGO are not the same algorithm, and we defined blockwise GIGO, a simple extension of the GIGO algorithm, showing that xNES has a special status, as it admits a definition that is "almost" parametrization-invariant.

## Acknowledgments

I would like to thank Yann Ollivier for his numerous remarks about this article and Frédéric Barbaresco for finding [5].

## Appendix A

**Proof of Proposition 15.** Let us first consider the case $k = 1$.

When optimizing a linear function, the non-twisted IGO flow in $\tilde{\mathbb{G}}_d$ with the selection function $w : q \mapsto \mathbf{1}_{q \leqslant q_0}$ is known [1], and in particular, we have:

$$\mu_t = \mu_0 + \frac{\beta(q_0)}{\alpha(q_0)}\sigma_t, \tag{65}$$

$$\sigma_t = \sigma_0 \exp(\alpha(q_0)t), \tag{66}$$

where, if we denote by $\mathcal{N}$ a random vector following a standard normal distribution and $\mathcal{F}$ the cumulative distribution of a standard normal distribution,

$$\alpha(q_0, d) = \frac{1}{2d}\left(\int_0^{q_0} \mathcal{F}^{-1}(u)^2 du - q_0\right), \tag{67}$$

and:

$$\beta(q_0) = \mathbb{E}(\mathcal{N}\mathbf{1}_{\mathcal{N} \leqslant \mathcal{F}^{-1}(q_0)}). \tag{68}$$

In particular, $\alpha := \alpha(\frac{1}{4}, 1) \approx 0.107$ and $\beta := \beta(\frac{1}{4}) \approx -0.319$.

With a minor modification of the proof in [1], we find that the $(\eta_\mu, \eta_\sigma)$-twisted IGO flow is given by:

$$\mu_t = \mu_0 + \frac{\beta(q_0)}{\alpha(q_0)}\sigma_0 \exp(\eta_\mu \alpha(q_0)t), \tag{69}$$

$$\sigma_t = \sigma_0 \exp(\eta_\sigma \alpha(q_0)t), \tag{70}$$

Notice that Equation (69) shows that the assertions about the convergence of $(\sigma_n)$ immediately imply the assertions about the convergence of $(\mu_n)$.

Let us now consider a step of the GIGO algorithm: The twisted IGO speed is $Y = (\eta_\mu \beta \sigma_0, \eta_\sigma \alpha \sigma_0)$, with $\alpha\sigma_0 > 0$ (*i.e.*, the variance should be increased: this is where we need $q_0 < 0.5$).

Proposition 17 shows that the covariance at the end of the step is (using the same notation):

$$\sigma(\delta t) = \sigma(0)\mathrm{Im}\left(\frac{die^{v\delta t} - c}{cie^{v\delta t} + d}\right) = \sigma(0)\frac{e^{v\delta t}(d^2 + c^2)}{c^2 e^{2v\delta t} + d^2} =: \sigma(0) f(\delta t), \tag{71}$$

and it is easy to see that $f$ only depends on $\delta t$ (and on $q_0$). In other words, $f(\delta t)$ will be the same at each step of the algorithm. The existence of $\delta t_{\mathrm{cr}}$ easily follows (furthermore, recall Figure 1 in Section 4.1), and $\delta t_{\mathrm{cr}}$ is the positive solution of $f(x) = 1$.

After a quick computation, we find:

$$\exp(v\delta t_{\mathrm{cr}}) = \frac{\sqrt{1 + u^2} + 1}{\sqrt{1 + u^2} - 1}. \tag{72}$$

where:

$$u := \sqrt{\frac{\eta_\mu}{2n\eta_\sigma}}\frac{\beta}{\alpha}, \tag{73}$$

and:

$$v := \sqrt{\eta_\sigma^2 \alpha^2 + \frac{\eta_\mu \eta_\sigma}{2n} \beta^2}. \tag{74}$$

Finally, for $w = k.\mathbf{1}_{q \leqslant q_0}$, Proposition 9 shows that:

$$\delta t_{\text{cr}} = \frac{1}{k} \frac{1}{v} \ln \left( \frac{\sqrt{1+u^2}+1}{\sqrt{1+u^2}-1} \right). \tag{75}$$

□

### A1. Generalization of the Twisted Fisher Metric

The following definition is a more general way to introduce the twisted Fisher metric.

**Definition 17.** *Let $(\Theta, g)$ be a Riemannian manifold, $(\Theta_1, g|_{\Theta_1}), ..., (\Theta_n, g|_{\Theta_n})$, a splitting (as defined in Section 6.4) of $\Theta$ compatible with the metric $g$.*

*We call $(\eta_1, ..., \eta_n)$-twisted metric on $(\Theta, g)$ for the splitting $\Theta_1, ..., \Theta_n$ the metric $g'$ on $\Theta$ defined by $g'|_{\Theta_i} = \frac{1}{\eta_i} g|_{\Theta_i}$ for $1 \leqslant i \leqslant n$, and $\Theta_i \perp \Theta_j$ for $i \neq j$.*

**Proposition 16.** *The $(\eta_\mu, \eta_\Sigma)$-twisted metric on $\mathbb{G}_d$ with the Fisher metric for the splitting $\mathcal{N}(\mu, \Sigma) \mapsto (\mu, \Sigma)$ coincides with the $(\eta_\mu, \eta_\Sigma)$-twisted Fisher metric from Definition 11.*

**Proof.** It is easy to see that the $(\eta_\mu, \eta_\Sigma)$-twisted Fisher metric satisfies the condition in Definition 17. □

### A2. Twisted Geodesics

The following theorem can be used to compute the twisted geodesics from the non twisted geodesics. It is a simple calculation.

**Theorem 7.** *Let $\eta_\mu, \eta_\Sigma \in \mathbb{R}$, $\mu_0 \in \mathbb{R}^d$, $A_0 \in GL_d(\mathbb{R})$, and $(\dot{\mu}_0, \dot{\Sigma}_0) \in T_{\mathcal{N}(\mu_0, A_0 A_0^T)} \mathbb{G}_d$. Let*

$$\begin{array}{rccc} h & : & \mathbb{G}_d & \to & \mathbb{G}_d \\ & & \mathcal{N}(\mu, \Sigma) & \mapsto & \mathcal{N}(\sqrt{\frac{\eta_\mu}{\eta_\Sigma}}\mu, \Sigma) \end{array}. \tag{76}$$

*We denote by $\phi$ (resp. $\psi$) the Riemannian exponential of $\mathbb{G}_d$ (resp. $\mathbb{G}_d$ with the $(\eta_\mu, \eta_\Sigma)$-twisted Fisher metric) at $\mathcal{N}(\sqrt{\frac{\eta_\mu}{\eta_\Sigma}}\mu_0, A_0 A_0^T)$ (resp. $\mathcal{N}(\mu_0, A_0 A_0^T)$). We have:*

$$\psi(\dot{\mu}_0, \dot{\Sigma}_0) = h \circ \phi(\sqrt{\frac{\eta_\Sigma}{\eta_\mu}}\dot{\mu}_0, \dot{\Sigma}_0) \tag{77}$$

**Proof.** Let us denote by: $\begin{pmatrix} I_\mu & 0 \\ 0 & I_\Sigma \end{pmatrix}$ the Fisher metric in the parametrization $\mu, \Sigma$, and consider the following parametrization of $\mathbb{G}_d$: $(\tilde{\mu}, \Sigma) \mapsto \mathcal{N}(\frac{\sqrt{\eta_\Sigma}}{\sqrt{\eta_\mu}}\tilde{\mu}, \Sigma)$.

The Riemannian exponential at $\mathcal{N}(\mu_0, A_0A_0^T)$ in this parametrization is:

$$h \circ \phi \circ (\mathrm{d}h(\mu_0, A_0A_0^T))^{-1} \tag{78}$$

However, in this parametrization, the Fisher metric reads:

$$\begin{pmatrix} \frac{\eta_\Sigma}{\eta_\mu} I_\mu & 0 \\ 0 & I_\Sigma \end{pmatrix}, \tag{79}$$

which is proportional to the $(\eta_\mu, \eta_\Sigma)$-twisted Fisher metric up to a factor $\frac{1}{\eta_\Sigma}$. Consequently, the Christoffel symbols are the same as the Christoffel symbols of the $(\eta_\mu, \eta_\Sigma)$-twisted Fisher metric, and so are the geodesics. Therefore, we have:

$$\psi = h \circ \phi \circ (\mathrm{d}h(\mu_0, A_0A_0^T))^{-1}, \tag{80}$$

which is what we wanted. □

For the remainder of this section, we fix $\eta_\mu$ and $\eta_\Sigma$; $\mathbb{G}_d$ is endowed with the $(\eta_\mu, \eta_\Sigma)$-twisted Fisher metric, and $\tilde{\mathbb{G}}_d$ is endowed with the induced metric. The proofs of the propositions below are a simple rewriting of their non-twisted counterparts that can be found in Sections 4 and 5.1 and can be seen as corollaries of Theorem 7.

**Theorem 8.** *If $\gamma\colon t \mapsto \mathcal{N}(\mu(t), \sigma(t)^2 I)$ is a twisted geodesic of $\tilde{\mathbb{G}}_d$, then there exists $a, b, c, d \in \mathbb{R}$, such that $ad - bc = 1$, and $v > 0$, such that*
$\mu(t) = \mu(0) + \sqrt{\frac{2d\eta_\mu}{\eta_\sigma}} \frac{\dot{\mu}_0}{\|\dot{\mu}_0\|} \tilde{r}(t)$, $\sigma(t) = \mathrm{Im}(\gamma_{\mathbb{C}}(t))$, *with* $\tilde{r}(t) = \mathrm{Re}(\gamma_{\mathbb{C}}(t))$ *and:*

$$\gamma_{\mathbb{C}}(t) := \frac{aie^{vt} + b}{cie^{vt} + d}. \tag{81}$$

**Proposition 17.** *Let $n \in \mathbb{N}$, $v_\mu \in \mathbb{R}^n$, $v_\sigma$, $\eta_\mu$, $\eta_\sigma$, $\sigma_0 \in \mathbb{R}$, with $\sigma_0 > 0$.*
*Let $v_r := \|v_\mu\|$, $\lambda = \sqrt{\frac{2n\eta_\mu}{\eta_\sigma}}$ $v := \sqrt{\frac{\frac{1}{\lambda^2}v_r^2 + v_\sigma^2}{\sigma_0^2}}$, $M_0 := \frac{1}{\lambda}\frac{v_r}{v\sigma_0^2}$ and $S_0 := \frac{v_\sigma}{v\sigma_0^2}$.*
*Let $c := \left(\frac{\sqrt{M_0^2 + S_0^2} - S_0}{2}\right)^{\frac{1}{2}}$ and $d := \left(\frac{\sqrt{M_0^2 + S_0^2} + S_0}{2}\right)^{\frac{1}{2}}$.*
*Let $\gamma_{\mathbb{C}}(t) := \sigma_0 \frac{die^{vt} - c}{cie^{vt} + d}$.*
*Then:*

$$\gamma : t \mapsto \mathcal{N}\left(\mu_0 + \lambda \frac{v_\mu}{\|v_\mu\|} \mathrm{Re}(\gamma_{\mathbb{C}}(t)), \mathrm{Im}(\gamma_{\mathbb{C}}(t))\right) \tag{82}$$

*is the twisted geodesic of $\tilde{\mathbb{G}}_n$ satisfying $\gamma(0) = (\mu_0, \sigma_0)$ and $\dot{\gamma}(0) = (v_\mu, v_\sigma)$. The regular geodesics of $\tilde{\mathbb{G}}_n$ are obtained with $\eta_\mu = \eta_\sigma = 1$.*

**Theorem 9.** *Let $\gamma : t \mapsto \mathcal{N}(\mu_t, \Sigma_t)$ be a twisted geodesic of $\mathbb{G}_d$. Then, the following quantities are invariant:*

$$J_\mu = \frac{1}{\eta_\mu} \Sigma_t^{-1} \dot{\mu}_t, \tag{83}$$

$$J_\Sigma = \Sigma_t^{-1} \left(\frac{1}{\eta_\mu} \dot{\mu}_t \mu_t^T + \frac{1}{\eta_\Sigma} \dot{\Sigma}_t\right). \tag{84}$$

**Theorem 10.** *If $\mu : t \mapsto \mu_t$ and $\Sigma : t \mapsto \Sigma_t$ satisfy the equations:*

$$\dot{\mu}_t = \eta_\mu \Sigma_t J_\mu \tag{85}$$

$$\dot{\Sigma}_t = \eta_\Sigma \Sigma_t (J_\Sigma - J_\mu \mu_t^T) = \eta_\Sigma \Sigma_t J_\Sigma - \frac{\eta_\Sigma}{\eta_\mu} \dot{\mu}_t \mu_t^T, \tag{86}$$

*where:*

$$J_\mu = \frac{1}{\eta_\mu} \Sigma_0^{-1} \dot{\mu}_0,$$

*and:*

$$J_\Sigma = \Sigma_0^{-1} \left( \frac{1}{\eta_\mu} \dot{\mu}_0 \mu_0^T + \frac{1}{\eta_\Sigma} \dot{\Sigma}_0 \right).$$

*then $t \mapsto \mathcal{N}(\mu_t, \Sigma_t)$ is a twisted geodesic of $\mathbb{G}_d$.*

**Theorem 11.** *If $\mu : t \mapsto \mu_t$ and $A : t \mapsto A_t$ satisfy the equations:*

$$\dot{\mu} = \eta_\mu A_t A_t^T J_\mu, \tag{87}$$

$$\dot{A}_t = \frac{\eta_\Sigma}{2} (J_\Sigma - J_\mu \mu_t^T)^T A_t, \tag{88}$$

*where:*

$$J_\mu = \frac{1}{\eta_\mu} (A_0^{-1})^T A_0^{-1} \dot{\mu}_0$$

*and:*

$$J_\Sigma = (A_0^{-1})^T A_0^{-1} \left( \frac{1}{\eta_\mu} \dot{\mu}_0 \mu_0^T + \frac{1}{\eta_\Sigma} \dot{A}_0 A_0^T + \frac{1}{\eta_\Sigma} A_0 \dot{A}_0^T \right),$$

*then $t \mapsto \mathcal{N}(\mu_t, A_t A_t^T)$ is a twisted geodesic of $\mathbb{G}_d$.*

## A3. Pseudocodes

### A3.1. For All Algorithms

All studied algorithms have a common part, given here:

Variables: $\mu, \Sigma$ (or $A$ such that $\Sigma = AA^T$).

List of parameters: $f \colon \mathbb{R}^d \to \mathbb{R}$, step size $\delta t$, learning rates $\eta_\mu, \eta_\Sigma$, sample size $\lambda$, weights $(w_i)_{i \in [1,\lambda]}$, $N$ number of steps for the Euler method, $r$ Euler step size reduction factor (for GIGO-$\Sigma$ only).

**Algorithm 1** For all algorithms.

$\mu \leftarrow \mu_0$
**if** The algorithm updates $\Sigma$ directly **then**
  $\Sigma \leftarrow \Sigma_0$
  Find some $A$, such that $\Sigma = AA^T$
**else** {The algorithm updates a square root $A$ of $\Sigma$}
  $A \leftarrow A_0$
  $\Sigma = AA^T$
**end if**
**while** NOT (Termination criterion) **do**
  **for** $i = 1$ to $\lambda$ **do**
    $z_i \sim \mathcal{N}(0, I)$
    $x_i = Az_i + \mu$
  **end for**
  Compute the IGO initial speed, and update the mean and the covariance (the updates are Algorithms 2 to 6).
**end while**

Notice that we always need a square root $A$ of $\Sigma$ to sample the $x_i$, but the decomposition $\Sigma = AA^T$ is not unique. Two different decompositions will give two algorithms, such that one is a modification of the other as a stochastic process: same law (the $x_i$ are abstractly sampled from $\mathcal{N}(\mu, \Sigma)$), but different trajectories (for given $z_i$, different choices for the square root will give different $x_i$). For GIGO-$\Sigma$, since we have to invert the covariance matrix, we used the Cholesky decomposition ($A$ lower triangular. The the other implementation directly maintains a square root of $\Sigma$). Usually, in CMA-ES, the square root of $\Sigma$ ($\Sigma = AA^T$, $A$ symmetric) is used.

A3.2. Updates

When describing the different updates, $\mu$, $\Sigma$, $A$, the $x_i$ and the $z_i$ are those defined in Algorithm 1.

For Algorithm 2 (GIGO-$\Sigma$), when the covariance matrix after one step is not positive-definite, we compute the update again, with a step size divided by $r$ for the Euler method (we have no reason to recommend any particular value of $r$, the only constraint is $r > 1$).

**Algorithm 2** GIGO Update, one step, updating the covariance matrix.

1. Compute the IGO speed:

$v_\mu = A \sum_{i=1}^{\lambda} w_i z_i,$

$v_\Sigma = A \sum_{i=1}^{\lambda} w_i \left( z_i z_i^T - I \right) A^T.$

2. Compute the Noether invariants:

$J_\mu \leftarrow \Sigma^{-1} v_\mu,$

$J_\Sigma \leftarrow \Sigma^{-1} ({v_\mu}^t \mu + v_\Sigma).$

3. Solve numerically the equations of the geodesics:

Unhappy $\leftarrow$ true

$\mu_0 \leftarrow \mu$

$\Sigma_0 \leftarrow \Sigma$

$k = 0$

**while** Unhappy **do**

  $\mu \leftarrow \mu_0$

  $\Sigma \leftarrow \Sigma_0$

  $h \leftarrow \delta t / (N r^k)$

  **for** $i = 1$ to $N r^k$ **do**

    $\mu \leftarrow \mu + h \eta_\mu \Sigma J_\mu$

    $\Sigma \leftarrow \Sigma + h \eta_\Sigma \Sigma (J_\Sigma - J_\mu \mu^T)$

  **end for**

  **if** $\Sigma$ positive-definite **then**

    Unhappy $\leftarrow$ false

  **end if**

  $k \leftarrow k + 1$

**end while**

**return** $\mu, \Sigma$

**Algorithm 3** GIGO Update, one step, updating a square root of the covariance matrix.

1. Compute the IGO speed:

$$v_\mu = A \sum_{i=1}^{\lambda} w_i z_i,$$

$$v_\Sigma = A \sum_{i=1}^{\lambda} w_i \left( z_i z_i^T - I \right) A^T.$$

2. Compute the Noether invariants:

$J_\mu \leftarrow \Sigma^{-1} v_\mu,$

$J_\Sigma \leftarrow \Sigma^{-1}({v_\mu}^t \mu + v_\Sigma).$

3. Solve numerically the equations of the geodesics:

$h \leftarrow \delta t / N$

**for** $i = 1$ to $N$ **do**

$\mu \leftarrow \mu + h \eta_\mu A A^T J_\mu$

$A \leftarrow A + \frac{h}{2} \eta_\Sigma (J_\Sigma - J_\mu \mu^T)^T A$

**end for**

**return** $\mu$, $A$

**Algorithm 4** Exact GIGO, one step. Not exactly our implementation; see the discussion after Corollary 1.

1. Compute the IGO speed:

$$v_\mu = A \sum_{i=1}^{\lambda} w_i z_i,$$

$$v_\Sigma = A \sum_{i=1}^{\lambda} w_i \left( z_i z_i^T - I \right) A^T.$$

2. Learning rates

$\lambda \leftarrow \sqrt{\frac{\eta_\Sigma}{\eta_\mu}}$

$\mu \leftarrow \lambda \mu$

$v_\mu \leftarrow \eta_\mu \lambda v_\mu$

$v_\Sigma \leftarrow \eta_\Sigma v_\Sigma$

3. Intermediate computations.

$G^2 \leftarrow A^{-1} \left( v_\Sigma (A^{-1})^T A^{-1} v_\Sigma + 2 v_\mu v_\mu^T \right) (A^{-1})^T$

$C_1 \leftarrow \mathrm{ch}(\frac{G}{2})$

$C_2 \leftarrow \mathrm{sh}(\frac{G}{2}) G^{-1}$

$R \leftarrow \left( (C_1 - A^{-1} v_\Sigma (A^{-1})^T C_2)^{-1} \right)^T$

4. Actual update

$\mu \leftarrow \mu + 2 A R C_2 A^{-1} v_\mu$

$A \leftarrow AR$

5. Return to the "real" $\mu$

$\mu \leftarrow \frac{\mu}{\lambda}$

**return** $\mu$, $A$

**Algorithm 5** xNES update, one step.

1. Compute $G_\mu$ and $G_M$ (equivalent to the computation of the IGO speed):

$$G_\mu = \sum_{i=1}^{\lambda} w_i z_i$$

$$G_M = \sum_{i=1}^{\lambda} w_i \left( z_i z_i^T - I \right)$$

2. Actual update:

$\mu \leftarrow \mu + \eta_\mu A G_\mu$

$A \leftarrow A + A \exp(\eta_\Sigma G_M / 2)$

**return** $\mu$, $A$

**Algorithm 6** pure rank-$\mu$ CMA-ES update, one step

1. Computation of the IGO speed:

$$v_\mu = \sum_{i=1}^{\lambda} w_i (x_i - \mu)$$

$$v_\Sigma = \sum_{i=1}^{\lambda} w_i \left( (x_i - \mu)(x_i - \mu)^T - \Sigma \right)$$

2. Actual update:

$\mu \leftarrow \mu + \eta_\mu v_\mu$

$\Sigma \leftarrow \Sigma + \eta_\Sigma v_\Sigma$

**return** $\mu$, $\Sigma$

**Algorithm 7** GIGO in $\tilde{\tilde{\mathbb{G}}}_d$, one step.

1. Compute the IGO speed:

$$Y_\mu = \sum_{i=1}^{\lambda} w_i (x_i - \mu) \; ; \; Y_\sigma = \sum_{i=1}^{\lambda} w_i \left( \frac{(x_i - \mu)^T (x_i - \mu)}{2d\sigma} - \frac{\sigma}{2} \right)$$

2. Better parametrization:

$$\lambda := \sqrt{\frac{2d\eta_\mu}{\eta_\sigma}}$$

$$v_r := \frac{\eta_\mu}{\lambda} \|Y_\mu\| \; ; \; v_\sigma := \eta_\sigma Y_\sigma$$

3. Find $a, b, c, d, v$ corresponding to $\mu, \sigma, \dot{\mu}, \dot{\sigma}$:

$$v = \sqrt{\frac{v_r^2 + v_\sigma^2}{\sigma^2}}$$

$$S_0 := \frac{v_\sigma}{v\sigma^2} \; ; \; M_0 := \frac{v_r}{v\sigma^2}$$

$$C := \frac{\sqrt{S_0^2 + M_0^2} - S_0}{2} \; ; \; D := \frac{\sqrt{S_0^2 + M_0^2} + S_0}{2}$$

$$c := \sqrt{C} \; ; \; d := \sqrt{D}$$

4. Actual Update:

$$z := \sigma \frac{die^{v\delta t} - c}{cie^{v\delta t} + d}$$

$$\mu := \mu + \lambda \mathrm{Re}(z) \frac{Y_\mu}{\|Y_\mu\|} \; ; \; \sigma := \mathrm{Im}(z)$$

**return** $\mu$, $\sigma$

## Conflicts of Interest

The author declares no conflict of interest.

## References


1. Ollivier, Y.; Arnold, L.; Auger, A.; Hansen, N. Information-geometric optimization algorithms: A unifying picture via invariance principles. **2011**, arXiv:1106.3708.
2. Amari, S.-I.; Nagaoka, H. *Methods of Information Geometry (Translations of Mathematical Monographs)*; American Mathematical Society: Providence, RI, USA, 2007.
3. Malagò, L.; Pistone, G. Combinatorial optimization with information geometry: The Newton method. *Entropy* **2014**, *16*, 4260–4289.
4. Eriksen, P. *Geodesics Connected with the Fisher Metric on the Multivariate Normal Manifold*; Technical Report 86-13; Institute of Electronic Systems, Aalborg University: Aalborg, Denmark, 1986.
5. Calvo, M.; Oller, J.M. An Explicit Solution of Information Geodesic Equations for the Multivariate Normal Model. *Stat. Decis.* **1991**, *9*, 119–138.
6. Imai, T.; Takaesu, A.; Wakayama, M. Remarks on geodesics for multivariate normal models. *J. Math-for-Industry* **2011**, *3*, 125–130.
7. Skovgaard, L.T. A Riemannian geometry of the multivariate normal model. *Scand. J. Stat.* **1981**, *11*, 211–223.
8. Porat, B.; Friedlander, B. Computation of the Exact Information Matrix of Gaussian Time Series with Stationary Random Components. *IEEE Trans. Acoust. Speech Signal Process.* **1986**, *34*, 118–130.
9. Baluja, S.; Caruana, R. *Removing the Genetics from the Standard Genetic Algorithm*; Technical Report CMU-CS-95-141; Morgan Kaufmann Publishers: Burlington, MA, USA, 1995, pp. 38–46.
10. Malagò, L.; Matteucci, M.; Pistone, G. Towards the geometry of estimation of distribution algorithms based on the exponential family. In Proceedings of the 11th Workshop Proceedings on Foundations of Genetic Algorithms, Schwarzenberg, Austria, 5–9 January 2011; pp. 230–242.
11. Kern, S.; Müller, S.D.; Hansen, N.; Büche, D.; Ocenasek, J.; Koumoutsakos, P. Learning probability distributions in continuous evolutionary algorithms—A comparative review. *Nat. Comput.* **2003**, *3*, 77–112.
12. Wierstra, D.; Schaul, T.; Glasmachers, T.; Sun, Y.; Peters, J.; Schmidhuber, J. Natural evolution strategies. *J. Mach. Learn. Res.* **2014**, *15*, 949–980.
13. Huang, W. Optimization Algorithms on Riemannian Manifolds with Applications. Ph.D. Thesis, Florida State University, Tallahassee, FL, USA, 2013.
14. Absil, P.A.; Mahony, R.; Sepulchre, R. *Optimization Algorithms on Matrix Manifolds*; Princeton University Press: Princeton, NJ, USA, 2008.

15. Arnold, V.; Vogtmann, K.; Weinstein, A. *Mathematical Methods of Classical Mechanics (Graduate Texts in Mathematics)*; Springer: New York, NY, USA, 1989.
16. Bourguignon, J. *Calcul variationnel*; Ecole Polytechnique: Palaiseau, France, 2007. (in French)
17. Jost, J.; Li-Jost, X. *Calculus of Variations (Cambridge Studies in Advanced Mathematics)*; Cambridge University Press: Cambridge, UK, 1998.
18. Gallot, S.; Hulin, D.; LaFontaine, J. *Riemannian Geometry (Universitext)*, 3rd ed.; Springer: Berlin/Heidelberg, Germany, 2004.
19. Glasmachers, T.; Schaul, T.; Yi, S.; Wierstra, D.; Schmidhuber, J. Exponential natural evolution strategies. In Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation, Portland, OR, USA, 7–11 July 2010.
20. Akimoto, Y.; Nagata, Y.; Ono, I.; Kobayashi, S. Bidirectional relation between CMA evolution strategies and natural evolution strategies. In *Parallel Problem Solving from Nature, PPSN XI*; Schaefer, R., Cotta, C., Kołodziej, J., Rudolph, G., Eds.; Springer: New York, NY, USA, 2010.
21. Hansen, N. The CMA evolution strategy: A tutorial. Available online: https://www.lri.fr/∼hansen/
cmatutorial.pdf (accessed on 1 January 2015).
22. Bensadon, J. Source Code. Available online: https://www.lri.fr/~bensadon/ (accessed on 13 January 2015).
23. Akimoto, Y.; Ollivier, Y. Objective improvement in information-geometric optimization. In Proceedings of the twelfth workshop on Foundations of genetic algorithms XII, Adelaide, Australia, 16–20 January 2013.

# Natural Gradient Flow in the Mixture Geometry of a Discrete Exponential Family

**Luigi Malagò and Giovanni Pistone**

**Abstract:** In this paper, we study Amari's natural gradient flows of real functions defined on the densities belonging to an exponential family on a finite sample space. Our main example is the minimization of the expected value of a real function defined on the sample space. In such a case, the natural gradient flow converges to densities with reduced support that belong to the border of the exponential family. We have suggested in previous works to use the natural gradient evaluated in the mixture geometry. Here, we show that in some cases, the differential equation can be extended to a bigger domain in such a way that the densities at the border of the exponential family are actually internal points in the extended problem. The extension is based on the algebraic concept of an exponential variety. We study in full detail a toy example and obtain positive partial results in the important case of a binary sample space.

Reprinted from *Entropy*. Cite as: Malagò, L.; Pistone, G. Natural Gradient Flow in the Mixture Geometry of a Discrete Exponential Family. *Entropy* **2015**, *17*, 4215–4254.

## 1. Introduction

For the purpose of obtaining a clear presentation of our approach to the geometry of statistical models, we start with a recap of nonparametric statistical manifold; see, e.g., the review paper [1]. However, we will shortly move to the actual setup of the present paper, *i.e.*, the finite state space case.

Let $(\Omega, \mathcal{A}, \mu)$ be a measured space of sample points $\boldsymbol{x} \in \Omega$. We denote by $\mathcal{P}_{\geq} \subset L^1(\mu)$ the simplex of (probability) densities and by $\mathcal{P}_{>} \subset \mathcal{P}_{\geq}$ the convex set of strictly positive densities. If $\Omega$ is finite, then $\mathcal{P}_{>}$ is the topological interior of $\mathcal{P}_{\geq}$. We denote by $\mathcal{P}^1$ the affine space generated by $\mathcal{P}_{\geq}$.

The set $\mathcal{P}_{>}$ holds the exponential geometry, which is an affine geometry, whose geodesics are curves of the form $t \mapsto p_t \propto p_0^{1-t} p_1^t$. The set $\mathcal{P}^1$ holds the mixture geometry, whose geodesics are of the form $t \mapsto p_t = (1-t)p_0 + tp_1$. A proper definition of the exponential and mixture geometry, where probability densities are considered points, requires the definition of the proper tangent space to hold the vectors representing the velocity of a curve. In both cases, the tangent space $T_p$ at a point $p$ is a space of random variables $V$ with zero expected value, $\mathrm{E}_p[V] = 0$. On the tangent space $T_p$, a natural scalar product is defined, $\langle U, V \rangle_p = \mathrm{E}_p[UV]$, so that a pseudo-Riemannian structure is available. Note that the Riemannian structure is a third geometry, different from both the exponential and the mixture geometries. Note also that both the expected value and the covariance can be naturally extended to be defined on $\mathcal{P}^1$.

For each lower bounded objective function $f\colon \Omega \to \mathbb{R}$ and each statistical model $\mathcal{M} \subset \mathcal{P}_{>}$, the (stochastic) relaxation of $f$ to $\mathcal{M}$ is the function $F(p) = \mathrm{E}_p[f] \in \mathbb{R}$, $p \in \mathcal{M}$; *cf.* [2]. The

minimization of the stochastic relaxation as a tool to minimize the objective function has been studied by many authors [3–7].

If we have a parameterization $\boldsymbol{\xi} \mapsto p_{\boldsymbol{\xi}}$ of $\mathcal{M}$, the parametric expression of the relaxed function is $\hat{F}(\boldsymbol{\xi}) = \mathrm{E}_{p_{\boldsymbol{\xi}}}[f]$. Under integrability and differentiability conditions on both $\boldsymbol{\xi} \mapsto p_{\boldsymbol{\xi}}$ and $\boldsymbol{x} \mapsto f(\boldsymbol{x})$, $\hat{F}$ is differentiable, with $\partial_j \hat{F}(\boldsymbol{\xi}) = \mathrm{E}_{p_{\boldsymbol{\xi}}}[\partial_j \log(p_{\boldsymbol{\xi}})\, f]$ and $\mathrm{E}_{p_{\boldsymbol{\xi}}}[\partial_j \log(p_{\boldsymbol{\xi}})] = 0$; see [1,8]. In order to properly describe the gradient flow of a relaxed random variable, these classical computations are better cast into the formal language of information geometry (see [9]) and, even better, in the language of non-parametric differential geometry [10] that was used in [11]. The previous computations suggest to take the Fisher score $\partial_j \log(p_{\boldsymbol{\xi}})$ as the definition of a tangent vector at the $j$-th coordinate curve. While the development of this analogy in the finite state space case does not require a special setup, in the non-finite state space, some care has to be taken.

In this paper, we follow the non-parametric setup discussed in [1] and, in particular, the notion of an exponential family $\mathcal{E}$ and the identification of the tangent space at each $p \in \mathcal{E}$ with a space of $p$-centered random variables.

The paper is organized as follows. We discuss in Section 2 the generalities of the finite state space case; in particular, we carefully define the various notions of the Fisher information matrix and natural gradient that arise from a given parameterization. In Section 3, we discuss a toy example in order to introduce the construction of an algebraic variety extending the exponential family from positive probabilities $\mathcal{P}_{>}$ to signed probabilities $\mathcal{P}^1$; this construction is applied to the natural gradient flow in the expectation parameters; moreover, it is shown that this model has a variety that is ruled. The last Section 4 is devoted to the treatment of the special important case when the sample space is binary.

The present paper is a development of the paper [12], which was presented as a poster at the MaxEnt Conference 2014. While the topic is the same, the actual overlapping between the two papers is minimal and concerns mainly the generalities that are repeated for the convenience of the reader.

## 2. Gradient Flow of Relaxed Optimization

Let $\Omega$ be a finite set of points $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\mu$ the counting measure of $\Omega$. In this case, a density $p \in \mathcal{P}_{\geq}$ is a probability function, *i.e.*, $p\colon \Omega \to \mathbb{R}_+$, such that $\sum_{\boldsymbol{x}\in\Omega} p(\boldsymbol{x}) = 1$.

Let $\mathcal{B} = \{T_1, \ldots, T_d\}$ be a set of random variables, such that, if $\sum_{j=1}^{d} c_j T_j$ is constant, then $c_1 = \cdots = c_d = 0$; for instance consider $\mathcal{B}$ such that $\sum_{\boldsymbol{x}\in\Omega} T_j(\boldsymbol{x}) = 0$, $j = 0, \ldots, d$, and $\mathcal{B}$ is a linear basis. We say that $\mathcal{B}$ is a set of affinely independent random variables. If $\mathcal{B}$ is a linear basis, it is affinely independent if and only if $\{1, T_1, \ldots, T_d\}$ is a linear basis.

We consider the statistical model $\mathcal{E}$ whose elements are uniquely identified by the natural parameters $\boldsymbol{\theta}$ in the exponential family with sufficient statistics $\mathcal{B}$, namely:

$$p_{\boldsymbol{\theta}} \in \mathcal{E} \quad \Leftrightarrow \quad \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{i=1}^{d} \theta_i T_i(\boldsymbol{x}) - \psi(\boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \mathbb{R}^d\,,$$

see [13].

The proper convex function $\psi\colon \mathbb{R}^d$,

$$\boldsymbol{\theta} \mapsto \psi(\boldsymbol{\theta}) = \log \sum_{\boldsymbol{x}\in\Omega} \mathrm{e}^{\boldsymbol{\theta}\cdot\boldsymbol{T}(\boldsymbol{x})} = \boldsymbol{\theta}\cdot \mathrm{E}_{p_{\boldsymbol{\theta}}}\left[\boldsymbol{T}\right] - \mathrm{E}_{p_{\boldsymbol{\theta}}}\left[\log\left(p_{\boldsymbol{\theta}}\right)\right]$$

is the cumulant generating function of the sufficient statistics $\boldsymbol{T}$, in particular,

$$\nabla\psi(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}\left[\boldsymbol{T}\right], \quad \operatorname{Hess}\psi(\boldsymbol{\theta}) = \operatorname{Cov}_{\boldsymbol{\theta}}\left(\boldsymbol{T},\boldsymbol{T}\right) .$$

Moreover, the entropy of $p_{\boldsymbol{\theta}}$ is:

$$H(p_{\boldsymbol{\theta}}) = -\mathrm{E}_{p_{\boldsymbol{\theta}}}\left[\log\left(p_{\boldsymbol{\theta}}\right)\right] = \psi(\boldsymbol{\theta}) - \boldsymbol{\theta}\cdot\nabla\psi(\boldsymbol{\theta}) .$$

The mapping $\nabla\psi$ is one-to-one onto the interior $M^\circ$ of the marginal polytope, that is the convex span of the values of the sufficient statistics $M = \{\boldsymbol{T}(\boldsymbol{x})|\boldsymbol{x}\in\Omega\}$. Note that no extra condition is required, because on a finite state space, all random variables are bounded. Nonetheless, even in this case, the proof is not trivial; see [13].

Convex conjugation applies [14] (Section 25) with the definition:

$$\psi_*(\boldsymbol{\eta}) = \sup\left\{\boldsymbol{\theta}\in\mathbb{R}^d \middle| \boldsymbol{\theta}\cdot\boldsymbol{\eta} - \psi(\boldsymbol{\theta})\right\}, \quad \boldsymbol{\eta}\in\mathbb{R}^d .$$

The concave function $\boldsymbol{\theta} \mapsto \boldsymbol{\eta}\cdot\boldsymbol{\theta} - \psi(\boldsymbol{\theta})$ has divergence mapping $\boldsymbol{\theta} \mapsto \boldsymbol{\eta} - \nabla\psi(\boldsymbol{\theta})$, and the equation $\boldsymbol{\eta} = \nabla\psi(\boldsymbol{\theta})$ has a solution if and only if $\boldsymbol{\eta}$ belongs to the interior $M^\circ$ of the marginal polytope. The restriction $\phi = \psi_*|_{M^\circ}$ is the Legendre conjugate of $\psi$, and it is computed by:

$$\phi\colon M^\circ \ni \boldsymbol{\eta} \mapsto \in (\nabla\psi)^{-1}(\boldsymbol{\eta})\cdot\boldsymbol{\eta} - \psi\circ(\nabla\psi)^{-1}(\boldsymbol{\eta}) \in \mathbb{R} .$$

The Legendre conjugate $\phi$ is such that $\nabla\phi = (\nabla\psi)^{-1}$, and it provides an alternative parameterization of $\mathcal{E}$ with the so-called expectation or mixture parameter $\boldsymbol{\eta} = \nabla\psi(\boldsymbol{\theta})$,

$$p_{\boldsymbol{\eta}} = \exp\left((\boldsymbol{T} - \boldsymbol{\eta})\cdot\nabla\phi(\boldsymbol{\eta}) + \phi(\boldsymbol{\eta})\right) . \tag{1}$$

While in the $\boldsymbol{\theta}$ parameters, the entropy is $H(p_{\boldsymbol{\theta}}) = \psi(\boldsymbol{\theta}) - \boldsymbol{\theta}\cdot\nabla\psi(\boldsymbol{\theta})$, in the $\boldsymbol{\eta}$ parameters, the $\phi$ function gives the negative entropy: $-H(p_{\boldsymbol{\eta}}) = \mathrm{E}_{p_{\boldsymbol{\eta}}}\left[\log p_{\boldsymbol{\eta}}\right] = \phi(\boldsymbol{\eta})$.

**Proposition 1.**

1. $\operatorname{Hess}\phi(\boldsymbol{\eta}) = \left(\operatorname{Hess}\psi(\boldsymbol{\theta})\right)^{-1}$ *when* $\boldsymbol{\eta} = \nabla\psi(\boldsymbol{\theta})$.
2. *The Fisher information matrix of the statistical model given by the exponential family in the* $\boldsymbol{\theta}$ *parameters is* $I_e(\boldsymbol{\theta}) = \operatorname{Cov}_{p_{\boldsymbol{\theta}}}\left(\nabla\log p_{\boldsymbol{\theta}}, \nabla\log p_{\boldsymbol{\theta}}\right) = \operatorname{Hess}\psi(\boldsymbol{\theta})$.
3. *The Fisher information matrix of the statistical model given by the exponential family in the* $\boldsymbol{\eta}$ *parameters is* $I_m(\boldsymbol{\theta}) = \operatorname{Cov}_{p_{\boldsymbol{\eta}}}\left(\nabla\log p_{\boldsymbol{\eta}}, \nabla\log p_{\boldsymbol{\eta}}\right) = \operatorname{Hess}\phi(\boldsymbol{\eta})$.

**Proof.** Derivation of the equality $\nabla\phi = (\nabla\psi)^{-1}$ gives the first item. The second item is a property of the cumulant generating function $\psi$. The third item follows from Equation (1). □

*2.1. Statistical Manifold*

The exponential family $\mathcal{E}$ is an elementary manifold in either the $\boldsymbol{\theta}$ or the $\boldsymbol{\eta}$ parameterization, named respectively exponential or mixture parameterization. We discuss now the proper definition of the tangent bundle $T\mathcal{E}$.

**Definition 1** (Velocity)**.** *If $I \ni t \mapsto p_t$, $I$ open interval, is a differentiable curve in $\mathcal{E}$, then its velocity vector is identified with its Fisher score:*

$$\frac{D}{dt}p(t) = \frac{d}{dt}\log\left(p_t\right) \ .$$

The capital $D$ notation is taken from differential geometry; see the classical monograph [15].

**Definition 2** (Tangent space)**.** *In the expression of the curve by the exponential parameters, the velocity is:*

$$\frac{D}{dt}p(t) = \frac{d}{dt}\log\left(p_t\right) = \frac{d}{dt}\left(\boldsymbol{\theta}(t)\cdot\boldsymbol{T} - \psi(\boldsymbol{\theta}(t))\right) = \dot{\boldsymbol{\theta}}(t)\cdot\left(\boldsymbol{T} - \mathrm{E}_{\boldsymbol{\theta}(t)}\left[\boldsymbol{T}\right]\right), \tag{2}$$

*that is it equals the statistics whose coordinates are $\dot{\boldsymbol{\theta}}(t)$ in the basis of the sufficient statistics centered at $p_t$. As a consequence, we identify the tangent space at each $p \in \mathcal{E}$ with the vector space of centered sufficient statistics, that is:*

$$T_p\mathcal{E} = \mathrm{Span}\left(T_j - \mathrm{E}_p\left[T_j\right] | j = 1, \ldots, d\right) \ .$$

In the mixture parameterization of Equation (1), the computation of the velocity is:

$$\begin{aligned}\frac{D}{dt}p(t) = \frac{d}{dt}\log\left(p_t\right) = \frac{d}{dt}\left(\nabla\phi(\boldsymbol{\eta}(t))\cdot\left(\boldsymbol{T} - \boldsymbol{\eta}(t)\right) + \phi(\boldsymbol{\eta}(t))\right) = \\ \left(\mathrm{Hess}\,\phi(\boldsymbol{\eta}(t))\dot{\boldsymbol{\eta}}(t)\right)\cdot\left(\boldsymbol{T} - \boldsymbol{\eta}(t)\right) = \dot{\boldsymbol{\eta}}(t)\cdot\left[\mathrm{Hess}\,\phi(\boldsymbol{\eta}(t))\left(\boldsymbol{T} - \boldsymbol{\eta}(t)\right)\right].\end{aligned} \tag{3}$$

The last equality provides the interpretation of $\dot{\boldsymbol{\eta}}(t)$ as the coordinate of the velocity in the conjugate vector basis $\mathrm{Hess}\,\phi(\boldsymbol{\eta}(t))\left(\boldsymbol{T} - \boldsymbol{\eta}(t)\right)$, that is the basis of velocities along the $\boldsymbol{\eta}$ coordinates.

In conclusion, the first order geometry is characterized as follows.

**Definition 3** (Tangent bundle $T\mathcal{E}$)**.** *The tangent space at each $p \in \mathcal{E}$ is a vector space of random variables $T_p\mathcal{E} = \mathrm{Span}\left(T_j - \mathrm{E}_p\left[T_j\right] | j = 1, \ldots, d\right)$, and the tangent bundle $T\mathcal{E} = \{(p, V) | p \in \mathcal{E}, V \in T_p\mathcal{E}\}$, as a manifold, is defined by the chart:*

$$T\mathcal{E} \ni \left(\mathrm{e}^{\boldsymbol{\theta}\cdot\boldsymbol{T} - \psi(\boldsymbol{\theta})}, \boldsymbol{v}\cdot\left(\boldsymbol{T} - \mathrm{E}_{\boldsymbol{\theta}}\left[\boldsymbol{T}\right]\right)\right) \mapsto (\boldsymbol{\theta}, \boldsymbol{v}). \tag{4}$$

**Proposition 2.**

1. *If $V = \boldsymbol{v}\cdot\left(\boldsymbol{T} - \boldsymbol{\eta}\right) \in T_{p_{\boldsymbol{\eta}}}\mathcal{E}$, then $V$ is represented in the conjugate basis as:*

$$\begin{aligned}V = \boldsymbol{v}\cdot\left(\boldsymbol{T} - \boldsymbol{\eta}\right) = \boldsymbol{v}\cdot\left(\mathrm{Hess}\,\phi(\boldsymbol{\eta})\right)^{-1}\mathrm{Hess}\,\phi(\boldsymbol{\eta})\left(\boldsymbol{T} - \boldsymbol{\eta}\right) = \\ \left(\left(\mathrm{Hess}\,\phi(\boldsymbol{\eta})\right)^{-1}\boldsymbol{v}\right)\cdot\mathrm{Hess}\,\phi(\boldsymbol{\eta})\left(\boldsymbol{T} - \boldsymbol{\eta}\right).\end{aligned} \tag{5}$$

2. *The mapping* $(\operatorname{Hess}\phi(\boldsymbol{\eta}))^{-1}$ *maps the coordinates* $\boldsymbol{v}$ *of a tangent vector* $V \in T_{p_{\boldsymbol{\eta}}}\mathcal{E}$ *with respect to the basis of centered sufficient statistics to the coordinates* $\boldsymbol{v}^*$ *with respect to the conjugate basis.*

3. *In the* $\boldsymbol{\theta}$ *parameters, the transformation is* $\boldsymbol{v} \mapsto \boldsymbol{v}^* = \operatorname{Hess}\psi(\boldsymbol{\theta})\boldsymbol{v}$.

**Remark 1.** *In the finite state space case, it is not necessary to go on to the formal construction of a dual tangent bundle, because all finite dimensional vector spaces are isomorphic. However, this step is compulsory in the infinite state space case, as was done in* [1]. *Moreover, the explicit construction of natural connections and natural parallel transports of the tangent and dual tangent bundle is unavoidable when considering the second-order calculus, as was done in* [1,8]*, in order to compute Hessians and implement Newton methods of optimization. However, the scope of the present paper is restricted to a basic study of gradient flows; hence, from now on, we focus on the Riemannian structure and disregard all second-order topics.*

**Proposition 3** (Riemannian metric)**.** *The tangent bundle has a Riemannian structure with the natural scalar product of each* $T_p\mathcal{E}$, $\langle V, W\rangle_p = \mathrm{E}_p[VW]$. *In the basis of sufficient statistics, the metric is expressed by the Fisher information matrix* $I(p) = \operatorname{Cov}_p(\boldsymbol{T}, \boldsymbol{T})$, *while in the conjugate basis, it is expressed by the inverse Fisher matrix* $I^{-1}(p)$.

**Proof.** In the basis of the sufficient statistics, $V = \boldsymbol{v}\cdot(\boldsymbol{T} - \mathrm{E}_p[\boldsymbol{T}])$, $W = \boldsymbol{w}\cdot(\boldsymbol{T} - \mathrm{E}_p[\boldsymbol{T}])$, so that:

$$\langle V, W\rangle_p = \boldsymbol{v}'\mathrm{E}_p\left[(\boldsymbol{T} - \mathrm{E}_p[\boldsymbol{T}])(\boldsymbol{T} - \mathrm{E}_p[\boldsymbol{T}])'\right]\boldsymbol{w} = \boldsymbol{v}'\operatorname{Cov}_p(\boldsymbol{T}, \boldsymbol{T})\,\boldsymbol{w} = \boldsymbol{v}' I(p)\boldsymbol{w}\,, \tag{6}$$

where $I(p) = \operatorname{Cov}_p(\boldsymbol{T}, \boldsymbol{T})$ is the Fisher information matrix.

If $p = p_{\boldsymbol{\theta}} = p_{\boldsymbol{\eta}}$, the conjugate basis at $p$ is:

$$\operatorname{Hess}\phi(\boldsymbol{\eta})(\boldsymbol{T} - \boldsymbol{\eta}) = \operatorname{Hess}\psi(\boldsymbol{\theta})^{-1}(\boldsymbol{T} - \nabla\phi(\boldsymbol{\theta})) = I^{-1}(p)(\boldsymbol{T} - \mathrm{E}_p[\boldsymbol{T}]), \tag{7}$$

so that for elements of the tangent space expressed in the conjugate basis, we have $V = \boldsymbol{v}^*\cdot I^{-1}(p)(\boldsymbol{T} - \mathrm{E}_p[\boldsymbol{T}])$, $W = \boldsymbol{w}^*\cdot I^{-1}(p)(\boldsymbol{T} - \mathrm{E}_p[\boldsymbol{T}])$; thus:

$$\langle V, W\rangle_p = \boldsymbol{v}^{*\prime}\mathrm{E}_p\left[I^{-1}(p)\cdot(\boldsymbol{T} - \mathrm{E}_p[\boldsymbol{T}])(\boldsymbol{T} - \mathrm{E}_p[\boldsymbol{T}])' I^{-1}(p)\right]\boldsymbol{w}^* = \boldsymbol{v}^{*\prime} I^{-1}(p)\boldsymbol{w}^*. \tag{8}$$

□

### *2.2. Gradient*

For each $C^1$ real function $F\colon \mathcal{E} \to \mathbb{R}$, its gradient is defined by taking the derivative along a $C^1$ curve $I \mapsto p(t)$, $p = p(0)$, and writing it with the Riemannian metrics,

$$\frac{d}{dt}\hat{F}(\boldsymbol{\theta}(t))\bigg|_{t=0} = \left\langle \nabla F(p), \frac{D}{dt}p(t)\bigg|_{t=0}\right\rangle_p, \quad \nabla F(p) \in T_p\mathcal{E}\,. \tag{9}$$

If $\boldsymbol{\theta} \mapsto \hat{F}(\boldsymbol{\theta})$ is the expression of $F$ in the parameter $\boldsymbol{\theta}$ and $t \mapsto \boldsymbol{\theta}(t)$ is the expression of the curve, then $\frac{d}{dt}\hat{F}(\boldsymbol{\theta}(t)) = \nabla\hat{F}(\boldsymbol{\theta}(t)) \cdot \dot{\boldsymbol{\theta}}(t)$, so that at $p = p_{\boldsymbol{\theta}(0)}$, with velocity $V = \frac{D}{dt}p(t)\big|_{t=0} = \dot{\boldsymbol{\theta}}(0) \cdot (\boldsymbol{T} - \nabla\psi(\boldsymbol{\theta}(0))$, so that we obtain the celebrated Amari's natural gradient of [16]:

$$\langle \nabla F(p), V \rangle_p = \Big(\operatorname{Hess}\psi(\boldsymbol{\theta}(0))^{-1}\nabla\hat{F}(\boldsymbol{\theta}(0)\Big)' \operatorname{Hess}\psi(\boldsymbol{\theta}(0))\dot{\boldsymbol{\theta}}(0)\ . \tag{10}$$

If $\boldsymbol{\eta} \mapsto \check{F}(\boldsymbol{\eta})$ is the expression of $F$ in the parameter $\boldsymbol{\eta}$ and $t \mapsto \boldsymbol{\eta}(t)$ is the expression of the curve, then $\frac{d}{dt}\check{F}(\boldsymbol{\eta}(t)) = \nabla\check{F}(\boldsymbol{\eta}(t)) \cdot \dot{\boldsymbol{\eta}}(t)$ so that at $p = p_{\boldsymbol{\eta}(0)}$, with velocity $V = \frac{d}{dt}\log\left(p(t)\right)\big|_{t=0} = \dot{\boldsymbol{\eta}}(0) \cdot \operatorname{Hess}\phi(\boldsymbol{\eta}(0))(\boldsymbol{T} - \boldsymbol{\eta}(0))$,

$$\langle \nabla F(p), V \rangle_p = (\operatorname{Hess}\phi(\boldsymbol{\eta}(0))^{-1}\nabla\hat{F}(\boldsymbol{\eta}(0))' \operatorname{Hess}\phi(\boldsymbol{\eta}(0))\dot{\boldsymbol{\eta}}(0). \tag{11}$$

We summarize all notions of gradient in the following definition.

**Definition 4** (Gradients)**.**

1. *The random variable $\nabla F(p)$ uniquely defined by Equation* (9) *is called the (geometric) gradient of $F$ at $p$. The mapping $\nabla F\colon \mathcal{E} \ni p \mapsto \nabla F(p)$ is a vector field of $T\mathcal{E}$.*

2. *The vector $\widetilde{\nabla}\,\hat{F}(\boldsymbol{\theta}) = \operatorname{Hess}\psi(\boldsymbol{\theta})^{-1}\nabla\hat{F}(\boldsymbol{\theta})$ of Equation* (10) *is the expression of the geometric gradient in the $\boldsymbol{\theta}$ in the basis of sufficient statistics, and it is called the natural gradient, while $\nabla\hat{F}(\boldsymbol{\theta})$, which is the expression in the conjugate basis of the sufficient statistics, is called the vanilla gradient.*

3. *The vector $\widetilde{\nabla}\,\check{F}(\boldsymbol{\eta}) = \operatorname{Hess}\phi(\boldsymbol{\eta})^{-1}\nabla\check{F}(\boldsymbol{\eta})$ of Equation* (10) *is the expression of the geometric gradient in the $\boldsymbol{\eta}$ parameter and in the conjugate basis of sufficient statistics, and it is called the natural gradient, while $\nabla\check{F}(\boldsymbol{\eta})$, which is the expression in the basis of sufficient statistics, is called the vanilla gradient.*

Given a vector field of $\mathcal{E}$, *i.e.*, a mapping $G$ defined on $\mathcal{E}$, such that $G(p) \in T_p\mathcal{E}$, which is called a section of the tangent bundle in the standard differential geometric language, an integral curve from $p$ is a curve $I \ni t \mapsto p(t)$, such that $p(0) = p$ and $\frac{D}{dt}p(t) = G(p(t))$. In the $\boldsymbol{\theta}$ parameters, $G(p_{\boldsymbol{\theta}}) = \hat{\boldsymbol{G}}(\boldsymbol{\theta}) \cdot (\boldsymbol{T} - \nabla\psi(\boldsymbol{\theta}))$, so that the differential equation is expressed by $\dot{\boldsymbol{\theta}}(t) = \hat{\boldsymbol{G}}(\boldsymbol{\theta}(t))$. In the $\boldsymbol{\eta}$ parameters, $G(p_{\boldsymbol{\eta}}) = \check{\boldsymbol{G}}(\boldsymbol{\eta}) \cdot \operatorname{Hess}\phi(\boldsymbol{\eta})(\boldsymbol{T} - \boldsymbol{\eta})$, and the differential equation is $\dot{\boldsymbol{\eta}}(t) = \check{\boldsymbol{G}}(\boldsymbol{\eta}(t))$.

**Definition 5** (Gradient flow)**.** *The gradient flow of the real function $F\colon \mathcal{E}$ is the flow of the differential equation $\frac{D}{dt}p(t) = \nabla F(p(t))$,* i.e., *$\frac{d}{dt}p(t) = p(t)\nabla F(p(t))$. The expression in the $\boldsymbol{\theta}$ parameters is $\dot{\boldsymbol{\theta}}(t) = \widetilde{\nabla}\,\hat{F}(\boldsymbol{\theta}(t))$, and the expression in the $\boldsymbol{\eta}$ parameters is $\dot{\boldsymbol{\eta}}(t) = \widetilde{\nabla}\,\check{F}(\boldsymbol{\eta}(t))$.*

The cases of gradient computation we have discussed above are just a special case of a generic argument. Let us briefly study the gradient flow in a general chart $f\colon \boldsymbol{\zeta} \mapsto p_{\boldsymbol{\zeta}}$. Consider the change of parametrization from $\boldsymbol{\zeta}$ to $\boldsymbol{\theta}$,

$$\boldsymbol{\zeta} \mapsto p_{\boldsymbol{\zeta}} \mapsto \boldsymbol{\theta}(p_{\boldsymbol{\zeta}}) = I(p_{\boldsymbol{\zeta}})^{-1}\operatorname{Cov}_{p_{\boldsymbol{\zeta}}}\left(\boldsymbol{T}, \log p_{\boldsymbol{\zeta}}\right)\ ,$$

and denote the Jacobian matrix of the parameters' change by $J(\boldsymbol{\zeta})$. We have:

$$\begin{aligned}\log p_{\boldsymbol{\zeta}} &= \boldsymbol{T}\cdot\boldsymbol{\theta}(\boldsymbol{\zeta}) - \psi(\boldsymbol{\theta}(\boldsymbol{\zeta})) \\ &= \boldsymbol{T}\cdot I(p_{\boldsymbol{\zeta}})^{-1}\operatorname{Cov}_{p_{\boldsymbol{\zeta}}}(\boldsymbol{T},\log p_{\boldsymbol{\zeta}}) - \psi\left(I(p_{\boldsymbol{\zeta}})^{-1}\operatorname{Cov}_{p_{\boldsymbol{\zeta}}}(\boldsymbol{T},\log p_{\boldsymbol{\zeta}})\right)\ ,\end{aligned}$$

and the $\boldsymbol{\zeta}$ coordinate basis of the tangent space $T_{p_{\boldsymbol{\zeta}}}\mathcal{E}$ consists of the components of the gradient with respect to $\boldsymbol{\zeta}$,

$$\nabla(\boldsymbol{\zeta}\mapsto\log p_{\boldsymbol{\zeta}}) = J^{-1}(\boldsymbol{\zeta})\left(\boldsymbol{T}-\mathrm{E}_{p_{\boldsymbol{\zeta}}}[\boldsymbol{T}]\right)$$

It should be noted that in this case, the expression of the Fisher information matrix does not have the form of a Hessian of a potential function. In fact, the case of the exponential and the mixture parameters point to a special structure, which is called the Hessian manifold; see [17].

*2.3. Gradient Flow in the Mixture Geometry*

From now on, we are going to focus on the expression of the gradient flow in the $\boldsymbol{\eta}$ parameters. From Definition 4, we have:

$$\widetilde{\nabla}\check{F}(\boldsymbol{\eta}) = \operatorname{Hess}\phi(\boldsymbol{\eta})^{-1}\nabla\check{F}(\boldsymbol{\eta}) = \operatorname{Hess}\psi(\nabla\phi(\boldsymbol{\eta}))\nabla\check{F}(\boldsymbol{\eta}) = I(p_{\boldsymbol{\eta}})\nabla\check{F}(\boldsymbol{\eta})\ ,$$

where $I(p) = \operatorname{Cov}_p(\boldsymbol{T},\boldsymbol{T})$. As $p\mapsto\operatorname{Cov}_p(\boldsymbol{T},\boldsymbol{T})$ is the restriction to the simplex of a quadratic function, while $p\mapsto\boldsymbol{\eta}$ is the restriction to the exponential family $\mathcal{E}$ of a linear function, in some cases, we can naturally consider the extension of the gradient flow equation outside $M^\circ$. One notable case is when the function $F$ is a relaxation of a non-constant state space function $f\colon\Omega\to\mathbb{R}$, as it is defined in, e.g., [3].

**Proposition 4.** *Let $f\colon\Omega\to\mathbb{R}$, and let $F(p)=\mathrm{E}_p[f]$ be its relaxation on $p\in\mathcal{E}$. It follows:*

1. *$\nabla F(p)$ is the least square projection of $f$ onto $T_p\mathcal{E}$, that is:*

$$\nabla F(p) = I(p)^{-1}\operatorname{Cov}_p(f,\boldsymbol{T})\cdot(\boldsymbol{T}-\mathrm{E}_p[\boldsymbol{T}])\ .$$

2. *The expressions in the exponential parameters $\boldsymbol{\theta}$ are $\widetilde{\nabla}\hat{F}(\boldsymbol{\theta}) = (\operatorname{Hess}\psi(\boldsymbol{\theta}))^{-1}\operatorname{Cov}_{\boldsymbol{\theta}}(f,\boldsymbol{T})$, $\nabla\hat{F}(\boldsymbol{\theta}) = \operatorname{Cov}_{\boldsymbol{\theta}}(f,\boldsymbol{T})$, respectively.*

3. *The expressions in the mixture parameters $\boldsymbol{\eta}$ are $\widetilde{\nabla}\check{F}(\boldsymbol{\eta}) = \operatorname{Cov}_{\boldsymbol{\eta}}(f,\boldsymbol{T})$ and $\nabla\check{F}(\boldsymbol{\eta}) = \operatorname{Hess}\phi(\boldsymbol{\eta})\operatorname{Cov}_{\boldsymbol{\eta}}(f,\boldsymbol{T})$, respectively.*

**Proof.** On a generic curve through $p$ with velocity $V$, we have $\frac{d}{dt}\mathrm{E}_{p(t)}[f]\big|_{t=0} = \operatorname{Cov}_p(f,V) = \langle f,V\rangle_p$. If $V\in T_p\mathcal{E}$, we can orthogonally project $f$ to get $\langle\nabla F,V\rangle_p = \langle(I^{-1}(p)\operatorname{Cov}_p(f,\boldsymbol{T}))\cdot(\boldsymbol{T}-\mathrm{E}_p[\boldsymbol{T}]),V\rangle_p$. □

**Remark 2.** *Let us briefly recall the behavior of the gradient flow in the relaxation case. Let $\boldsymbol{\theta}_n$, $n=1,2,\dots$, be a minimizing sequence for $\hat{F}$, and let $\bar{p}$ be a limit point of the sequence $(p_{\boldsymbol{\theta}_n})_n$. It follows that $\bar{p}$ has a defective support, in particular $\bar{p}\notin\mathcal{E}$; see* [18,19]. *For a proof along lines*

*coherent with the present paper, see* [20] *(Theorem 1). It is found that the support* $\underline{F} \subset \Omega$ *is exposed, that is* $\boldsymbol{T}(\underline{F})$ *is a face of the marginal polytope* $M = \operatorname{con}\{\boldsymbol{T}(\boldsymbol{x}) | \boldsymbol{x} \in \Omega\}$. *In particular,* $\mathrm{E}_{\bar{p}}[\boldsymbol{T}] = \bar{\boldsymbol{\eta}}$ *belongs to a face of the marginal polytope* $M$. *If* $\boldsymbol{a}$ *is the (interior) orthogonal of the face, that is* $\boldsymbol{a} \cdot \boldsymbol{T}(\boldsymbol{x}) + b \geq 0$ *for all* $\boldsymbol{x} \in \Omega$ *and* $\boldsymbol{a} \cdot \boldsymbol{T}(\boldsymbol{x}) + b = 0$ *on the exposed set, then* $\boldsymbol{a} \cdot (\boldsymbol{T}(\boldsymbol{x}) - \bar{\boldsymbol{\eta}}) = 0$ *on the face, so that* $\boldsymbol{a} \cdot \mathrm{Cov}_{\bar{p}}(f, \boldsymbol{T}) = 0$. *If we extend the mapping* $\boldsymbol{\eta} \mapsto \mathrm{Cov}_{\boldsymbol{\eta}}(f, \boldsymbol{T})$ *on the closed marginal polytope* $M$ *to be the limit of the vector field of the gradient on the faces of the marginal polytope, we expect to see that such a vector field is tangent to the faces. This remark is further elaborated below in the binary case.*

### *2.4. The Saturated Model*

A case of special tutorial interest is obtained when the exponential family contains all probability densities, that is when $\mathcal{E} = \mathcal{P}_{>}$. This case has been treated by many authors; here, we use the presentation of [21].

It is convenient to recode the sample space as $\Omega = \{0, \ldots, d\}$, where $\boldsymbol{x} = 0$ is a distinguished point. If $X$ is the identity on $\Omega$, we define the sufficient statistics to be the indicator functions of points $T_j = (X = j)$, $j = 1, \ldots, d$. The saturated exponential family consists of all of the positive densities written as:

$$p(\boldsymbol{x}; \boldsymbol{\theta}) = \exp\left(\sum_{j=1}^{d} \theta_j (X = j) - \psi(\boldsymbol{\theta})\right) ,$$

where:

$$\psi(\boldsymbol{\theta}) = \log\left(1 + \sum_{j=1}^{d} \mathrm{e}^{\theta_j}\right) .$$

Note that, in this case, the expectation parameter $\eta_j = \mathrm{E}((X = j))$ is the probability of case $\boldsymbol{x} = j$ and the marginal polytope is the probability simplex $\Delta_d$.

The gradient mapping is:

$$\boldsymbol{\eta} = \nabla\psi(\boldsymbol{\theta}) = \left(\frac{\mathrm{e}^{\theta_j}}{1 + \sum_{i=1}^{d} \mathrm{e}^{\theta_i}} \middle| j = 1, \ldots, d\right) ,$$

the inverse gradient mapping is defined for $\boldsymbol{\eta} \in ]0, 1[^d$ by:

$$\boldsymbol{\theta} = (\nabla\psi)^{-1}(\boldsymbol{\eta}) = \nabla\phi(\boldsymbol{\eta}) = \left(\log\left(\frac{\eta_j}{1 - \sum_{i=1}^{d} \eta_i}\right) \middle| j = 1, \ldots, d\right) ,$$

the negative entropy (Legendre conjugate) is:

$$\phi(\boldsymbol{\eta}) = \boldsymbol{\eta} \cdot \nabla\phi(\boldsymbol{\eta}) - \psi \circ \nabla\phi(\boldsymbol{\eta}) = \sum_{j=1}^{d} \eta_j \log\left(\frac{\eta_j}{1 - \sum_{i=1}^{d} \eta_i}\right) + \log\left(1 - \sum_{i=1}^{d} \eta_i\right) ,$$

the $\boldsymbol{\eta}$ parameterization (1) of the probability is:

$$p_{\boldsymbol{\eta}} = \exp\left((\boldsymbol{T}-\boldsymbol{\eta})\cdot\nabla\phi(\boldsymbol{\eta})+\phi(\boldsymbol{\eta})\right) =$$
$$\exp\left(\sum_{j=1}^{d}((X=j)-\eta_j)\log\left(\frac{\eta_j}{1-\sum_{i=1}^{d}\eta_i}\right)+\sum_{j=1}^{d}\eta_j\log\left(\frac{\eta_j}{1-\sum_{i=1}^{d}\eta_i}\right)+\log\left(1-\sum_{i=1}^{d}\eta_i\right)\right) =$$
$$\exp\left(\sum_{j=1}^{d}(X=j)\log\left(\frac{\eta_j}{1-\sum_{i=1}^{d}\eta_i}\right)+\log\left(1-\sum_{i=1}^{d}\eta_i\right)\right) =$$
$$\prod_{j=1}^{d}\left(\frac{\eta_j}{1-\sum_{i=1}^{d}\eta_i}\right)^{(X=j)}\left(1-\sum_{i=1}^{d}\eta_i\right) = \left(1-\sum_{i=1}^{d}\eta_i\right)^{(X=0)}\prod_{j=1}^{d}\eta_j^{(X=j)}\,.$$

**Remark 3.** *The previous equation prompts three crucial remarks:*

1. *The expression of the probability in the $\boldsymbol{\eta}$ parameters is a normalized monomial in the parameters.*
2. *The expression continuously extends the exponential family to the probabilities in $\mathcal{P}_{\geq}$.*
3. *The expression actually is a polynomial parameterization of the signed densities $\mathcal{P}^1$.*

We proceed to approach the three issues above. The Hessian functions are:

$$\operatorname{Hess}\psi(\boldsymbol{\theta}) = \operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}\otimes\boldsymbol{p}, \quad \boldsymbol{p} = (1-\sum_{j=1}^{d}\mathrm{e}^{\theta_j})^{-1}\mathrm{e}^{\boldsymbol{p}}\,,$$
$$\operatorname{Hess}\phi(\boldsymbol{\eta}) = \operatorname{diag}(\boldsymbol{\eta})^{-1} - \eta_0^{-1}[1]_{i,j=1}^{d}, \quad \eta_0 = 1-\sum_{j=1}^{d}\eta_j\,.$$

The matrix $\operatorname{Hess}\psi(\boldsymbol{\theta})$ is the Fisher information matrix $I(p)$ of the exponential family at $\boldsymbol{p}=p_{\boldsymbol{\theta}}$, and the matrix $\operatorname{Hess}\phi(\boldsymbol{\eta})$ is the inverse Fisher information matrix $I^{-1}(p)$ at $\boldsymbol{p}=p_{\boldsymbol{\eta}}$. It follows that the natural gradient of a function $\boldsymbol{\eta}\mapsto h(\boldsymbol{\eta})$ will be:

$$\widetilde{\nabla}\,h(\boldsymbol{\eta}) = \operatorname{Hess}\phi(\boldsymbol{\eta})\nabla h(\boldsymbol{\eta})\,,$$

whose behavior depends on the following theorem; see [21] (Proposition 3).

**Proposition 5.**

1. *The inverse Fisher information matrix $I(p)^{-1}$ is zero on the vertexes of the simplex, only.*
2. *The determinant of the inverse Fisher information matrix $I(p)^{-1}$ is:*

$$\det\left(I(p)^{-1}\right) = \left(1-\sum_{i=1}^{n}p_i\right)\prod_{i=1}^{n}p_i.$$

3. *The determinant of the inverse Fisher information matrix $I(p)^{-1}$ is zero on the borders of the simplex, only.*

4. *On the interior of each facet, the rank of the inverse Fisher information matrix $I(p)^{-1}$ is $(n-1)$, and the $(n-1)$ linear independent column vectors generate the subspace parallel to the facet itself.*

A generic statistical model can be seen as a submanifold of the saturated model, so that the form of the gradient in the submanifold is derived according to the general results in differential geometry. We do not do that here, and we switch to some very specific examples.

## 3. Toric Models: A Tutorial Example

Exponential families whose sample space is an integer lattice, such as finite subsets of $\mathbb{Z}^2$ or $\{+1,-1\}^d$, have special algebro-combinatorial features that fall under the name of algebraic statistics. Seminal papers have been [22,23]. Monographs on the topic are [24–26]. The book [27] covers both information geometry and algebraic statistics.

We do not assume the reader has detailed information about algebraic statistics. In this section, we work on a toy example intended to show both the basic mechanism of algebraic statistics and how the algebraic concepts are applied to the gradient flow problem as it was described in the previous section.

First, we give a general definition of the object on which we focus. A toric model is an exponential family, such that the orthogonal space of the space generated by the sufficient statistics and the constant has a vector basis of integer-valued random variables. We consider this example:

| $\Omega$ | $T_1$ | $T_2$ | $T_3$ |
|---|---|---|---|
| 1 | 0 | 0 | $-2$ |
| 2 | 0 | 1 | 1 |
| 3 | 1 | 0 | 2 |
| 4 | 2 | 1 | $-1$ |

, (12)

which corresponds to a variation of the classical independence model, where the design corresponds to the vertices of a square. It this example we moved the point $\{4\}$ from $(1,1)$ to $(2,1)$.

In Equation (12), $T_1$ and $T_2$ are the sufficient statistics of the exponential family:

$$p_{\boldsymbol{\theta}} = \exp\left(\theta_1 T_1 + \theta_2 T_2 - \psi(\boldsymbol{\theta})\right), \quad \psi(\boldsymbol{\theta}) = \log\left(1 + \mathrm{e}^{\theta_2} + \mathrm{e}^{\theta_1} + \mathrm{e}^{2\theta_1+\theta_2}\right), \tag{13}$$

$T_3$ is an integer-valued vector basis of the orthogonal space $\mathrm{Span}\,(\mathbf{1}, T_1, T_2)^{\perp}$.

For the purpose of the generalization to less trivial examples, it should be noted that $T_3 = T_3^+ - T_3^-$, that is $(-2,1,2,-1) = (0,1,2,0) - (2,0,0,1)$. The couple $(T_3^+, T_3^-)$ connects the lattice defined by:

$$\mathcal{L} = \left\{(Y,Z) \in \mathbb{Z}^4_{\geq} \times \mathbb{Z}^4_{\geq} \middle| B^T y = B^T Z\right\}, \quad B = \begin{bmatrix} \mathbf{1} & T_1 & T_2 \end{bmatrix}.$$

Such a set of generators is called a Markov basis of the lattice; see [22]. Algorithms are available to compute such a set of generators and are implemented, for instance, in the software suite 4ti2; see [28].

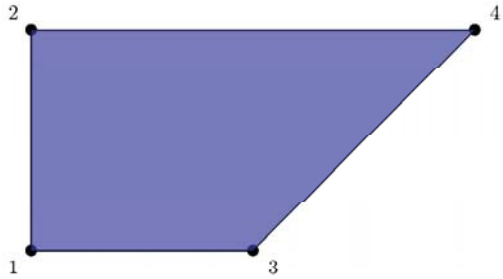



**Figure 1.** Marginal polytope of the exponential family in Equations (12) and (13). The coordinates of the vertices are given by $(T_1, T_2)$.

The sample space can be identified with the value of the sufficient statistics, hence with a finite subset of $\mathbb{Q}^2 \supset \Omega$, $\Omega = \{(0,0),(0,1),(1,0),(2,1)\}$; see Figure 1. Given a finite subset of $\mathbb{R}^d$, it is a general algebraic fact that there exists a filtering set of monomial functions that is a vector basis of all real functions on the subset itself; see an exposition and the applications to statistics in [24] or [27]. In our case, the monomial basis is $1, T_1, T_2, T_1T_2$, and we define the matrix of the saturated model to be:

$$A = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} \mathbf{1} & T_1 & T_2 & T_1T_2 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 2 & 1 & 2 \end{bmatrix} \end{array}, \quad A^{-1} = \frac{1}{2}\begin{bmatrix} 2 & 0 & 0 & 0 \\ -2 & 0 & 2 & 0 \\ -2 & 2 & 0 & 0 \\ 2 & -1 & -2 & 1 \end{bmatrix}. \tag{14}$$

The matrix $A$ one-to-one maps probabilities into expected values,

$$\begin{bmatrix} 1 & \eta_1 & \eta_2 & \eta_{12} \end{bmatrix} = \begin{bmatrix} 1 & \mathbb{E}[T_1] & \mathbb{E}[T_2] & \mathbb{E}[T_1T_2] \end{bmatrix} = \begin{bmatrix} p_1 & p_2 & p_3 & p_4 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 2 & 1 & 2 \end{bmatrix}, \tag{15}$$

and *vice versa*,

$$\begin{bmatrix} p_1 & p_2 & p_3 & p_4 \end{bmatrix} = \begin{bmatrix} 1 & \eta_1 & \eta_2 & \eta_{12} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 1 & 0 & 0 \\ 1 & -\frac{1}{2} & -1 & \frac{1}{2} \end{bmatrix}. \tag{16}$$

On Model (13), the (positive) probabilities are constrained by the model:

$$\begin{array}{c|ll} \Omega & p_{\boldsymbol{\theta}} & \exp\left(\theta_1 T_1 + \theta_2 T_2 - \log\left(1+\mathrm{e}^{\theta_2}+\mathrm{e}^{\theta_1}+\mathrm{e}^{2\theta_1+\theta_2}\right)\right) \\ \hline 1 & p(1;\boldsymbol{\theta}) & \exp\left(-\log\left(1+\mathrm{e}^{\theta_2}+\mathrm{e}^{\theta_1}+\mathrm{e}^{2\theta_1+\theta_2}\right)\right) \\ 2 & p(2;\boldsymbol{\theta}) & \exp\left(\theta_2-\log\left(1+\mathrm{e}^{\theta_2}+\mathrm{e}^{\theta_1}+\mathrm{e}^{2\theta_1+\theta_2}\right)\right) \\ 3 & p(3;\boldsymbol{\theta}) & \exp\left(\theta_1-\log\left(1+\mathrm{e}^{\theta_2}+\mathrm{e}^{\theta_1}+\mathrm{e}^{2\theta_1+\theta_2}\right)\right) \\ 4 & p(4;\boldsymbol{\theta}) & \exp\left(2\theta_1+\theta_2-\log\left(1+\mathrm{e}^{\theta_2}+\mathrm{e}^{\theta_1}+\mathrm{e}^{2\theta_1+\theta_2}\right)\right) \end{array} . \tag{17}$$

If we introduce the parameters $\zeta_1 = \exp(\theta_1)$, $\zeta_2 = \exp(\theta_2)$, the model is shown to be a (piece of an) algebraic variety, that is a set described by the rational parametric equations:

$$\begin{array}{c|ll} \Omega & p_{\boldsymbol{\zeta}} & \zeta^{T_1}\zeta^{T_2}/(1+\zeta_2+\zeta_1+\zeta_1^2\zeta_2) \\ \hline 1 & p(1;\boldsymbol{\zeta}) & 1/(1+\zeta_2+\zeta_1+\zeta_1^2\zeta_2) \\ 2 & p(2;\boldsymbol{\zeta}) & \zeta_2/(1+\zeta_2+\zeta_1+\zeta_1^2\zeta_2) \\ 3 & p(3;\boldsymbol{\zeta}) & \zeta_1/(1+\zeta_2+\zeta_1+\zeta_1^2\zeta_2) \\ 4 & p(4;\boldsymbol{\zeta}) & \zeta_1^2\zeta_2/(1+\zeta_2+\zeta_1+\zeta_1^2\zeta_2) \end{array} . \tag{18}$$

The peculiar structure of the toric model is best seen by considering the unnormalized probabilities:

$$\begin{array}{c|ll} \Omega & q_{\boldsymbol{\zeta}} & \zeta^{T_1}\zeta^{T_2} \\ \hline 1 & q(1;\boldsymbol{\zeta}) & 1 \\ 2 & q(2;\boldsymbol{\zeta}) & \zeta_2 \\ 3 & q(3;\boldsymbol{\zeta}) & \zeta_1 \\ 4 & q(4;\boldsymbol{\zeta}) & \zeta_1^2\zeta_2 \end{array} , \qquad p(\boldsymbol{x};\boldsymbol{\zeta}) = \frac{q(\boldsymbol{x};\boldsymbol{\zeta})}{1+\zeta_2+\zeta_1+\zeta_1^2\zeta_2}. \tag{19}$$

In algebraic terms, the homogeneous coordinates $[q_1 : q_2 : q_3 : q_4]$ belong to the projective space $\boldsymbol{P}^3$. Precisely, the (real) projective space $\boldsymbol{P}^3$ is the set of all non-zero points of $\mathbb{R}^4$ together with the equivalence relation $[q_1 : q_2 : q_3 : q_4] = [\bar{q}_1 : \bar{q}_2 : \bar{q}_3 : \bar{q}_4]$ if, and only if, $[q_1, q_2, q_3, q_4] = k[\bar{q}_1, \bar{q}_2, \bar{q}_3, \bar{q}_4]$, $k \neq 0$. The domain of unnormalized signed probabilities as projective points is the open subset $\mathbb{P}^3_*$ of $\mathbb{P}^3$ where $q_1+q_2+q_3+q_4 \neq 0$. On this set, we can compute the normalization:

$$\mathbb{P}^3_* \ni [q_1 : q_2 : q_3 : q_4] \mapsto [q_1, q_2, q_3, q_4]/(q_1+q_2+q_3+q_4) \in {}^*\mathcal{E} ,$$

where ${}^*\mathcal{E}$ is the affine space generated by the simplex $\Delta_3$. Notice that this embedding produces a number of natural geometrical structures on ${}^*\mathcal{E}$.

Because of the form of (13), a positive density $p$ belongs to that family if, and only if, $\log p \in \mathrm{Span}(1, T_1, T_2)$, which, in turn, is equivalent to $\log p \perp T_3$. We can rewrite the orthogonality as:

$$\begin{aligned} 0 = \sum_{\boldsymbol{x}\in\Omega} \log p(\boldsymbol{x}) T_3(\boldsymbol{x}) &= \sum_{\boldsymbol{x}:\, T_3(\boldsymbol{x})>0} \log p(\boldsymbol{x}) T_3^+(\boldsymbol{x}) - \sum_{\boldsymbol{x}:\, T_3(\boldsymbol{x})<0} \log p(\boldsymbol{x}) T_3^-(\boldsymbol{x}) \\ &= \log\left(\prod_{\boldsymbol{x}:\, T_3(\boldsymbol{x})>0} p(\boldsymbol{x})^{T_3^+(\boldsymbol{x})}\right) - \log\left(\prod_{\boldsymbol{x}:\, T_3(\boldsymbol{x})<0} p(\boldsymbol{x})^{T_3^-(\boldsymbol{x})}\right) . \end{aligned}$$

Dropping the $\log$ function in the last expression, we observe that the positive probabilities described by either Equation (17) with $\theta_1, \theta_2 \in \mathbb{R}$ or Equation (18) with $\zeta_1, \zeta_2 \in \mathbb{R}_>$ are equivalently described by the equations:

$$p_1 + p_2 + p_3 + p_4 - 1 = 0 \,, \tag{20}$$

$$p_1^2 p_4 - p_2 p_3^2 = 0 \,. \tag{21}$$

Equation (21) identifies a surface within the probability simplex $\Delta_3$, which is represented in Figure 2 by the triangularization of a grid of points that satisfy the invariant.

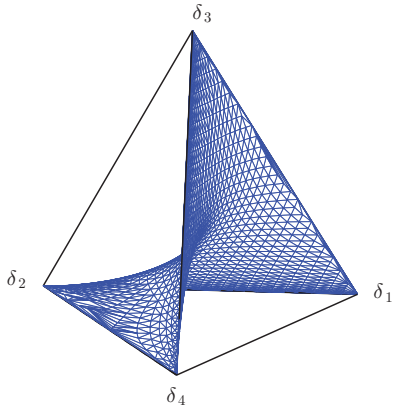



**Figure 2.** Representation of the exponential family in Equations (12) and (13) as a surface that intersects the probability simplex $\Delta_3$. The surface is obtained by the triangularization of a grid of points that satisfy the invariant in Equation (21).

By choosing a basis for the space orthogonal to $\mathrm{Span}\,(\mathbf{1}, T_1, T_2)^\perp$, we can embed the marginal polytope of Figure 1 into the associated full marginal polytope. By expressing probabilities as a function of the expectation parameters, Equation (21) identifies a relationship between $\eta_1, \eta_2$ and the expected values of the chosen basis for the orthogonal space. This corresponds to an equivalent invariant in the expectation parameters, which, in turn, identifies a surface in the full marginal polytope.

For instance, consider the full marginal polytope parametrized by $\boldsymbol{\eta} = (\eta_1, \eta_2, \eta_3)$, with $\eta_3 = \mathbb{E}[T_3]$, which corresponds to the choice of $T_3$ as a basis for the space orthogonal to the span of the sufficient statistics of the model, together with the constant $\mathbf{1}$, as in Equation (12). We introduce the following matrix:

$$B = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} \mathbf{1} & T_1 & T_2 & T_3 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & -2 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 2 \\ 1 & 2 & 1 & -1 \end{bmatrix} \end{array} , \tag{22}$$

and similarly to Equation (15), we use the $B$ matrix to one-to-one map probabilities into expected values, that is:

$$\begin{bmatrix} 1 \\ \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 2 \\ 0 & 1 & 0 & 1 \\ -2 & 1 & 2 & -1 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{bmatrix} , \tag{23}$$

and:

$$\begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{bmatrix} = \begin{bmatrix} \frac{3}{5} & -\frac{1}{5} & -\frac{2}{5} & -\frac{1}{5} \\ \frac{1}{5} & -\frac{2}{5} & \frac{7}{10} & \frac{1}{10} \\ \frac{2}{5} & \frac{1}{5} & -\frac{3}{5} & \frac{1}{5} \\ -\frac{1}{5} & \frac{2}{5} & \frac{3}{10} & -\frac{1}{10} \end{bmatrix} \begin{bmatrix} 1 \\ \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix} . \tag{24}$$

Then, by expressing probabilities as a function of the expectation parameters in Equation (21), we obtain the following invariant in $\boldsymbol{\eta}$ associated with the model:

$$(4\eta_1 + 3\eta_2 - \eta_3 - 2)(\eta_1 + 2\eta_2 + \eta_3 - 3)^2 + (4\eta_1 - 7\eta_2 - \eta_3 - 2)(\eta_1 - 3\eta_2 + \eta_3 + 2)^2 = 0 \, . \tag{25}$$

From the linear relationship between probabilities and expectation probabilities, we know that on the interior of the full marginal polytope, there exists a unique $\eta_3$ which can be computed as a function of the other expectation parameters. Solving Equation (25) for $\eta_3$ allows one to express explicitly the value of $\eta_3$ given $(\eta_1, \eta_2)$ and represent the surface associated with the invariant in the full marginal polytope. However, the cubic polynomial in Equation (25) in general admits three roots. The unique value of $\eta_3$ can be obtained from the roots of the cubic polynomial, by imposing that $\eta_3$ must be real and belong to the full marginal polytope given by $\mathrm{Conv}\left\{(T_1(\boldsymbol{x}), T_2(\boldsymbol{x}), T_3(\boldsymbol{x}))|\boldsymbol{x} \in \Omega\right\}$.

We remind that the determinant $\Delta$ associated with the cubic function in Equation (25) in the $\eta_3$ variable:

$$a\eta_3^3 + b\eta_3^2 + c\eta_3 + d = 0 \, , \tag{26}$$

with:

$$a = 1 \tag{27}$$

$$b = -2\eta_1 + \eta_2 + 1 \tag{28}$$

$$c = -(4\eta_1 + 3\eta_2 - 2)(\eta_1 + 2\eta_2 - 3) + \frac{1}{2}(\eta_1 + 2\eta_2 - 3)^2 - (4\eta_1 - 7\eta_2 - 2)(\eta_1 - 3\eta_2 + 2) + \\ + \frac{1}{2}(\eta_1 - 3\eta_2 + 2)^2 \tag{29}$$

$$d = -\frac{1}{2}(4\eta_1 + 3\eta_2 - 2)(\eta_1 + 2\eta_2 - 3)^2 - \frac{1}{2}(4\eta_1 - 7\eta_2 - 2)(\eta_1 - 3\eta_2 + 2)^2 \tag{30}$$

is given by:

$$\Delta = 18abcd - 4b^3d + b^2c^2 - 4ac^3 - 27a^2d^2 \, . \tag{31}$$

For $\Delta = 0$, the polynomial has a real root with multiplicity equal to three; for $\Delta < 0$, we have one real root and two complex conjugates roots, while for $\Delta > 0$, there exist three real roots. The three roots of the polynomial as a function of the coefficients are given by:

$$\eta_{3,k} = -\frac{1}{3}\left(b + u_k C + \frac{\Delta_0}{u_k C}\right) , \tag{32}$$

for $k \in \{1, 2, 3\}$, with:

$$u_1 = 1 , \tag{33}$$

$$u_2 = \frac{-1 + i\sqrt{3}}{2} , \tag{34}$$

$$u_3 = \frac{-1 - i\sqrt{3}}{2} , \tag{35}$$

and:

$$C = \sqrt[3]{\frac{\Delta_1 + \sqrt{(\Delta_1^2 - 4\Delta_0^3)}}{2}} , \tag{36}$$

$$\Delta_0 = b^2 - 3ac , \tag{37}$$

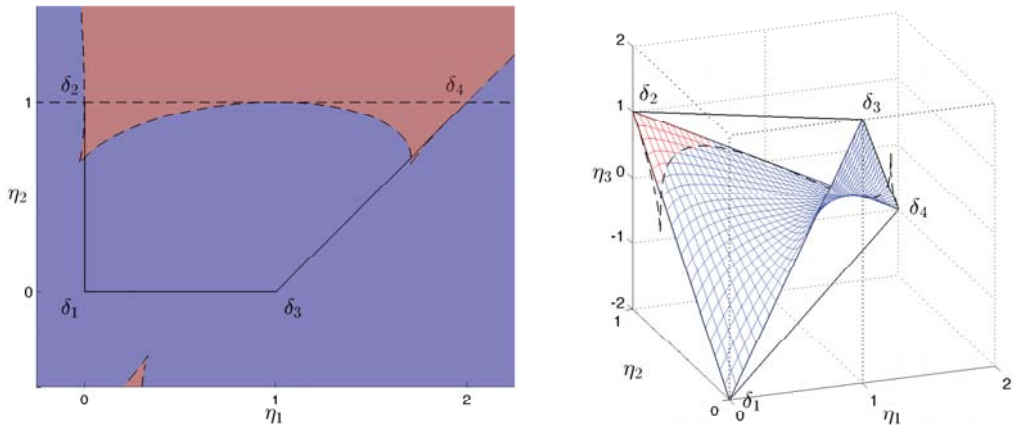
$$\Delta_1 = 2b^3 + 9abc + 27a^2 d . \tag{38}$$

For the cubic polynomial in $\eta_3$ of Equation (25), $\Delta < 0$ for $\eta_2 - 1 \neq 0$ and for:

$$4\eta_1^4 - 8\eta_1^3\eta_2 + 24\eta_1^2\eta_2^2 - 20\eta_1\eta_2^3 - 2\eta_2^4 - 8\eta_1^3 - 12\eta_1^2\eta_2 + 4\eta_2^3 + 8\eta_1^2 + 16\eta_1\eta_2 - \eta_2^2 - 4\eta_1 - 2\eta_2 + 1 > 0 . \tag{39}$$

In Figure 3(a), we represent in blue the region of the space $(\eta_1, \eta_2)$ where $\Delta < 0$, in red where $\Delta > 0$, and the points where $\Delta = 0$ with a dashed line. For $\Delta < 0$, the only real root is $\eta_{3,1}$, which identifies the blue surface in the full marginal polytope in Figure 3(b). For $\Delta > 0$, it is easy to verify that only $\eta_{3,2}$ belongs to the interior of the full marginal polytope parametrized by $(\eta_1, \eta_2, \eta_3)$, since it satisfies the inequalities given by the facets of the marginal polytope, and is represented in Figure 3(b) by the red surface. Finally, the three real roots coincide for $\Delta = 0$, that is, for $\eta_2 = 1$, and where:

$$4\eta_1^4 - 8\eta_1^3\eta_2 + 24\eta_1^2\eta_2^2 - 20\eta_1\eta_2^3 - 2\eta_2^4 - 8\eta_1^3 - 12\eta_1^2\eta_2 + 4\eta_2^3 + 8\eta_1^2 + 16\eta_1\eta_2 - \eta_2^2 - 4\eta_1 - 2\eta_2 + 1 = 0 . \tag{40}$$




(a) (b)

**Figure 3.** Marginal polytope of the exponential family in Equations (12) and (13) (**a**). The dashed lines correspond to the points where $\Delta = 0$, where $\Delta$ is the discriminant in Equation (31); over the red regions $\Delta > 0$ and over the blue regions $\Delta < 0$. Representation of the exponential family as a surface in the full marginal polytope parametrized by $(\eta_1, \eta_2, \eta_3)$ (**b**). The blue surface is given by the unique real root $\eta_{3,1}$ in Equation (32); the red surface corresponds to the unique real root $\eta_{3,2}$, which belongs to the full marginal polytope; over the dashed lines, which have been computed solving Equation (40) numerically, Equation (26) admits a real root with multiplicity equal to three.

In the polynomial ring $\mathbb{Q}[p_1, p_2, p_3, p_4]$, the model ideal:

$$\mathcal{I} = \left\langle p_1 + p_2 + p_3 + p_4 - 1, p_1^2 p_4 - p_2 p_3^2 \right\rangle \tag{41}$$

consists of all the polynomials of the form:

$$A\left(p_1 + p_2 + p_3 + p_4 - 1\right) + B\left(p_1^2 p_4 - p_2 p_3^2\right), \quad \forall A, B \in \mathbb{Q}[p_1, p_2, p_3, p_4] \, .$$

The algebraic variety of $\mathcal{I}$ uniquely extends the exponential family outside the positive octant. In the language of commutative algebra, it is the real Zariski closure of the exponential family model, *cf.* [29]. It is a notable example of toric variety. The general theory is in the monograph [30], and the applications to statistical models were first discussed in [31,32].

Let us discuss in some detail the parameterization of the toric variety as the submanifold of $\mathbb{R}^4$ defined by Equations (20) and (21). The Jacobian matrix is:

$$J = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2p_1 p_4 & -p_3^2 & -2p_2 p_3 & p_1^2 \end{bmatrix} .$$

It has rank one, that is, there is a singularity, if, and only if,

$$2p_1 p_4 = -p_3^2 = -2p_2 p_3 = p_1^2 \, .$$

This is equivalent to $p_1^2 = p_3^2 = 0$, which is a subspace of dimension two, whose intersection with Equation (20), is a line $\mathcal{C}$ in the affine space ${}^*\mathcal{E} = \{\boldsymbol{p} \in \mathbb{R}^4 | p_1 + p_2 + p_3 + p_4 = 1\}$. This (double) critical line intersects the simplex along the edge $\delta_2 \leftrightarrow \delta_4$. Outside $\mathcal{C}$, that is in the open complement set, the equations of the toric variety are locally solvable in two among the $p_i$'s under the condition that the corresponding minor is not zero. To have a picture of what this critical set looks like, let us intersect our surface with the plane $p_3 = 0$. On the affine space $p_1 + p_2 + p_4 = 1$, we have $p_1^2 p_4 = 0$, that is the union of the double line $p_1^2 = 0$ with the line $p_4 = 0$.

In the following, we derive a parameterization based on an algebraic argument, the Bézout theorem. In fact, it is remarkable that the cubic surface defined by Equations (20) and (21) is a well known example of ruled surface, see Exercise 5.8.15 in [33]. In fact, the singular line is a double line, so that the intersection of the cubic surface with any plane through the singular line is of degree $1 = 3 - 2$, by the Bézout theorem, and thus, it is a line.

The line $\mathcal{C}$ is said to be double because the polynomial $p_1^2 p_4 - p_2 p_3^2$ belongs to the ideal generated by $p_1^2$ and $p_3^2$. Let us consider the sheaf of planes through the singular line defined for each $[\alpha : \beta] \in \boldsymbol{P}^1$ by the equations:

$$\mathcal{P}[\alpha : \beta] = \{p_1 + p_2 + p_3 + p_4 - 1 = 0, \alpha p_1 + \beta p_3 = 0\} \ .$$

Let us intersect each plane $\mathcal{P}[\alpha : \beta]$ of the sheaf with the model variety $\mathcal{M}$ by solving the system of equations:

$$\begin{cases} p_1 + p_2 + p_3 + p_4 & = 1 \\ p_1^2 p_4 - p_2 p_3^2 & = 0 \\ \alpha p_1 + \beta p_3 & = 0 \end{cases} . \tag{42}$$

On the critical line $\mathcal{C}$, a generic point is parameterized as $\boldsymbol{p}(\tau, 0) = (0, \tau, 0, 1 - \tau)$, which satisfies Equation (42) for $\tau \in \mathbb{R}$. If $0 \leq \tau \leq 1$, then $\boldsymbol{p}(\tau, 0)$ belongs to the edge $\delta_2 \leftrightarrow \delta_4$.

As the critical line is double and the intersection of the model variety with the plane of the sheaf is a cubic curve, we expect the remaining part to be of degree $3 - 2 = 1$, that is to be a line. Assume first $\alpha, \beta \neq 0$. Outside the critical line, as $p_1, p_3$ are not both zero and $\alpha p_1 + \beta p_3 = 0$, then $\alpha p_1 = -\beta p_3 \neq 0$. It follows $(\alpha p_1)^2 = (\beta p_3)^2 \neq 0$; hence:

$$p_1^2 p_4 - p_2 p_3^2 = 0 \Rightarrow \beta^2 (\alpha p_1)^2 p_4 - \alpha^2 p_2 (\beta p_3)^2 = 0 \Rightarrow \beta^2 p_4 - \alpha^2 p_2 = 0 \ .$$

We have found that for $\alpha, \beta \neq 0$, the intersection between the plane $\mathcal{P}[\alpha : \beta]$ and the model variety $\mathcal{M}$ is the union of the critical line $\mathcal{C}$ and the line of equations:

$$\begin{cases} p_1 + p_2 + p_3 + p_4 & = 1 \\ \alpha p_1 + \beta p_3 & = 0 \\ -\alpha^2 p_2 + \beta^2 p_4 & = 0 \end{cases} . \tag{43}$$

This line intersects the critical line where:

$$p_1 = p_3 = 0, p_2 + p_4 = 1, -\alpha^2 p_2 + \beta^2 p_4 = 0 \ ,$$

that is in the point:

$$\boldsymbol{p}([\alpha:\beta],0)) = \left(0, \frac{\beta^2}{\alpha^2+\beta^2}, 0, \frac{\alpha^2}{\alpha^2+\beta^2}\right) .$$

In parametric form, the line in Equations (43) is:

$$\boldsymbol{p}([\alpha:\beta],t) = \boldsymbol{p}([\alpha:\beta],0) + \boldsymbol{u}t ,$$

with $\boldsymbol{u} = \left(\beta, \frac{\beta^2(\alpha-\beta)}{\alpha^2+\beta^2}, -\alpha, \frac{\alpha^2(\alpha-\beta)}{\alpha^2+\beta^2}\right)$,

$$\begin{aligned} p_1([\alpha:\beta],t) &= \beta t \\ p_2([\alpha:\beta],t) &= \frac{\beta^2}{\alpha^2+\beta^2} + \frac{\beta^2(\alpha-\beta)}{\alpha^2+\beta^2}t \\ p_3([\alpha:\beta],t) &= -\alpha t \\ p_4([\alpha:\beta],t) &= \frac{\alpha^2}{\alpha^2+\beta^2} + \frac{\alpha^2(\alpha-\beta)}{\alpha^2+\beta^2}t \end{aligned} . \tag{44}$$

The same equations hold in the previously excluded case $\alpha\beta = 0$.

Positive values of components 1 and 3 of the probability are obtained in Equation (44) for $\alpha\beta < 0$ and $\beta t > 0$, say $\alpha < 0$, $\beta > 0$, $t > 0$. In this case, we have for component 2:

$$\frac{\beta^2}{\alpha^2+\beta^2} + \frac{\beta^2(\alpha-\beta)}{\alpha^2+\beta^2}t = \frac{\beta^2}{\alpha^2+\beta^2}\left(1-(\beta-\alpha)t\right) ,$$

which is positive if $t < (\beta-\alpha)^{-1}$. The same condition applies to component 4. As $[\alpha:\beta] = \left[\frac{\alpha}{\beta-\alpha} : \frac{\beta}{\beta-\alpha}\right]$, we can always assume $\beta > 0$ and $\beta - \alpha = 1$ that is, $\alpha = \beta - 1$; hence $\beta < 1$. The parameterization of the positive probabilities in the model becomes:

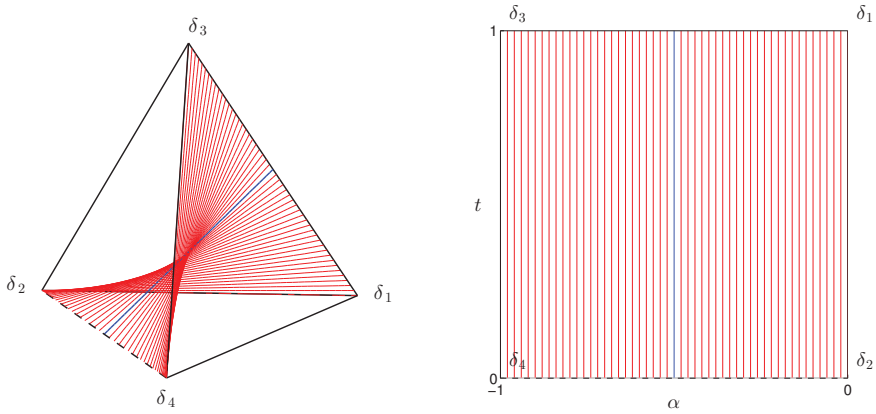
$$\begin{aligned} p_1(\alpha,t) &= (\alpha+1)t \\ p_2(\alpha,t) &= \frac{\alpha^2-(\alpha^2+2\alpha+1)t+2\alpha+1}{2\alpha^2+2\alpha+1} \\ p_3(\alpha,t) &= -\alpha t \\ p_4(\alpha,t) &= -\frac{\alpha^2 t-\alpha^2}{2\alpha^2+2\alpha+1} \end{aligned} , \qquad 0<t<1, -1<\alpha<0 . \tag{45}$$

For example, with $\alpha = -\frac{1}{2}$, we have:

$$\begin{aligned} p_1(\alpha,t) &= \frac{1}{2}t \\ p_2(\alpha,t) &= \frac{1}{2}(1-t) \\ p_3(\alpha,t) &= \frac{1}{2}t \\ p_4(\alpha,t) &= \frac{1}{2}(1-t) \end{aligned} , \qquad 0<t<1 .$$

In Figure 4(a), we represented the surface associated with the invariant of Equation (21) as a ruled surface in the probability simplex, according to Equations (45), where the blue line corresponds to

the case $\alpha = -\frac{1}{2}$. The ruled surface corresponds to the surface in Figure 2 that was approximated by the triangularization of a grid of points satisfying the invariant. In Figure 4(b), we represent the same lines of Figure 4(a) in the chart $(\alpha, t)$.




(a) (b)

**Figure 4.** Representation of the exponential family in Equations (12) and (13) as a ruled surface in the probability simplex (**a**) and in the parameter space $(\alpha, t)$ (**b**). The dashed line corresponds to the critical edge $\delta_2 \leftrightarrow \delta_4$ and the blue line to the case $\alpha = -\frac{1}{2}$.

From Equation (45), we can express the expectation parameters $\boldsymbol{\eta}$ as a function of $(\alpha, t)$, *i.e.*,

$$\eta_1 = \frac{2\alpha^2 - (2\alpha^3 + 4\alpha^2 + \alpha)t}{2\alpha^2 + 2\alpha + 1}\,, \tag{46}$$

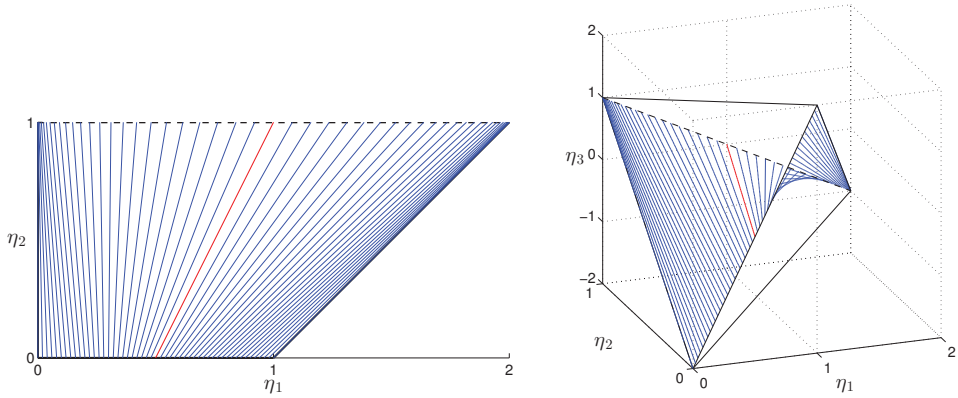
$$\eta_2 = -t + 1\,, \tag{47}$$

$$\eta_3 = -\frac{(8\alpha^3 + 12\alpha^2 + 10\alpha + 3)t - 2\alpha - 1}{2\alpha^2 + 2\alpha + 1}\,. \tag{48}$$

Notice that the dependence on $(\alpha, t)$ is rational. In Figure 5(a), the ruled surface has been represented in the full marginal polytope, while in Figure 5(a), the lines have been projected over the marginal polytope.

Let us invert Equation (45) to obtain the corresponding chart $\boldsymbol{p} \mapsto (\beta, t)$. From $p_1$ and $p_3$, we obtain $\beta = p_1/(p_1 + p_3)$. As $p_2 + p_4 = 1 - t$, we have the chart:

$$\beta = \frac{p_1}{p_1 + p_3}\,,$$
$$t = 1 - p_2 - p_4 = p_1 + p_3\,.$$




(a) (b)

**Figure 5.** Representation of the exponential family in Equations (12) and (13) as a ruled surface in the marginal polytope $(\eta_1, \eta_2)$ (**a**) and in the full marginal polytope parametrized by $(\eta_1, \eta_2, \eta_3)$ (**b**). The dashed line corresponds to the critical line $\delta_2 \leftrightarrow \delta_4$ and the red line to the case $\alpha = -\frac{1}{2}$.

It is remarkable that the model depends on the probability restricted to $\{1, 3\}$; similarly, the expectation parameters depend on $p_1$ and $p_3$ only.

From the theory of exponential families, we know that the gradient mapping:

$$(\theta_1, \theta_2) \mapsto \nabla\psi(\theta_1, \theta_2) = \left[ \begin{array}{cc} \frac{2\,e^{(2\,\theta_1+\theta_2)}+e^{\theta_1}}{e^{(2\,\theta_1+\theta_2)}+e^{\theta_1}+e^{\theta_2}+1} & \frac{e^{(2\,\theta_1+\theta_2)}+e^{\theta_2}}{e^{(2\,\theta_1+\theta_2)}+e^{\theta_1}+e^{\theta_2}+1} \end{array} \right]$$

is one-to-one from $\mathbb{R}^2$ onto the interior of the marginal polytope $M$; see Figure 3(b). The equations:

$$\eta_1 = \frac{\zeta_1 + 2\zeta_1^2\zeta_2}{1 + \zeta_2 + \zeta_1 + \zeta_1^2\zeta_2} \,,$$
$$\eta_2 = \frac{\zeta_2 + \zeta_1^2\zeta_2}{1 + \zeta_2 + \zeta_1 + \zeta_1^2\zeta_2} \,,$$

are uniquely solvable for $(\eta_1, \eta_2) \in M^\circ$. We study the local solvability in $\zeta_1, \zeta_2$ of:

$$(1 + \zeta_2 + \zeta_1 + \zeta_1^2\zeta_2)\eta_1 = \zeta_1 + 2\zeta_1^2\zeta_2 \,,$$
$$(1 + \zeta_2 + \zeta_1 + \zeta_1^2\zeta_2)\eta_2 = \zeta_2 + \zeta_1^2\zeta_2 \,,$$

that is,

$$0 = \eta_1 + (\eta_1 - 1)\zeta_1 + \eta_1\zeta_2 + (\eta_1 - 2)\zeta_1^2\zeta_2 \,,$$
$$0 = \eta_2 + \eta_2\zeta_1 + (\eta_2 - 1)\zeta_2 + (\eta_2 - 1)\zeta_1^2\zeta_2 \,.$$

The Jacobian is:

$$\begin{bmatrix} (\eta_1 - 1) + 2(\eta_1 - 2)\zeta_1\zeta_2 & \eta_1 + (\eta_1 - 2)\zeta_1^2 \\ \eta_2 + 2(\eta_2 - 1)\zeta_1\zeta_2 & (\eta_2 - 1) + (\eta_2 - 1)\zeta_1^2 \end{bmatrix} .$$

If we introduce the extra variable $\eta_{12}$, from Equations (15) and (18) we have the system:

$$\begin{aligned}(1+\zeta_2+\zeta_1+\zeta_1^2\zeta_2)\eta_1 &= \zeta_1+2\zeta_1^2\zeta_2\ ,\\ (1+\zeta_2+\zeta_1+\zeta_1^2\zeta_2)\eta_2 &= \zeta_2+\zeta_1^2\zeta_2\ ,\\ (1+\zeta_2+\zeta_1+\zeta_1^2\zeta_2)\eta_{12} &= 2\zeta_1^2\zeta_2\ ,\end{aligned}$$

Instead, if we use the variable $\eta_3$, from Equations (16) and (41), it is possible to derive the equation of the model variety in the $\eta_1, \eta_2, \eta_3$ parameters. From Equation (18), we have:

$$\begin{aligned}\eta_1 = \mathrm{E}_{\boldsymbol{\zeta}}[T_1] &= \frac{\zeta_1+2\zeta_1^2\zeta_2}{1+\zeta_2+\zeta_1+\zeta_1^2\zeta_2}\ ,\\ \eta_2 = \mathrm{E}_{\boldsymbol{\zeta}}[T_2] &= \frac{\zeta_2+\zeta_1^2\zeta_2}{1+\zeta_2+\zeta_1+\zeta_1^2\zeta_2}\ ,\\ \eta_3 = \mathrm{E}_{\boldsymbol{\zeta}}[T_3] &= \frac{-2+\zeta_2+2\zeta_1-\zeta_1^2\zeta_2}{1+\zeta_2+\zeta_1+\zeta_1^2\zeta_2}\ .\end{aligned}$$

Let us solve for the $\boldsymbol{\zeta}$, that is:

$$\begin{aligned}(1+\zeta_2+\zeta_1+\zeta_1^2\zeta_2)\eta_1 &= \zeta_1+2\zeta_1^2\zeta_2\ ,\\ (1+\zeta_2+\zeta_1+\zeta_1^2\zeta_2)\eta_2 &= \zeta_2+\zeta_1^2\zeta_2\ ,\\ (1+\zeta_2+\zeta_1+\zeta_1^2\zeta_2)\eta_3 &= -2+\zeta_2+2\zeta_1-\zeta_1^2\zeta_2\ .\end{aligned}$$

There is another way to derive the model constraint in the $\boldsymbol{\eta}$. In the example, the sample space has four points; the monomials $1, T_1, T_2, T_1T_2$ are a vector basis of the linear space of the columns of the matrix $A$, in particular $T_3$ is a linear combination:

| $\Omega$ | $\mathbf{1}$ | $T_1$ | $T_2$ | $T_1T_2$ | $T_3$ |
|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 | $-2$ |
| 2 | 1 | 0 | 1 | 0 | 1 |
| 3 | 1 | 1 | 0 | 0 | 2 |
| 4 | 1 | 2 | 1 | 2 | $-1$ |
| | $-2$ | 4 | 3 | $-5$ | $=$ |

.

It follows that:

$$\begin{aligned}\eta_3 = \mathrm{E}_{\boldsymbol{\theta}}[T_3] &= \mathrm{E}_{\boldsymbol{\theta}}[-2+4T_1+3T_2-5T_1T_2]\\ &= -2+4\mathrm{E}_{\boldsymbol{\theta}}[T_1]+3\mathrm{E}_{\boldsymbol{\theta}}[T_2]+3\,\mathrm{Cov}_{\boldsymbol{\theta}}(T_1,T_2)+3\mathrm{E}_{\boldsymbol{\theta}}[T_1]\,\mathrm{E}_{\boldsymbol{\theta}}[T_2]\\ &= -2+4\partial_1\psi(\boldsymbol{\theta})+3\partial_2\psi(\boldsymbol{\theta})-5\partial_1\partial_2\psi(\boldsymbol{\theta})-5\partial_1\psi(\boldsymbol{\theta})\partial_2\psi(\boldsymbol{\theta})\\ &= -2+4\eta_1+3\eta_2-5\partial_1\partial_2\psi(\boldsymbol{\theta})-5\eta_1\eta_2\ .\end{aligned}$$

*3.1. Border*

Let us consider the points in the model variety that are probabilities, that is,

$$p_1 + p_2 + p_3 + p_4 = 1, \quad p_1^2 p_4 = p_2 p_3^2, \quad p_1, p_2, p_3, p_4 \geq 0\ . \tag{49}$$

From the equation above, we see that single zeros are not allowed, that is to say there are no intersections between the model in Equation (49) and the open facets of the probability simplex. We now consider the full marginal polytope obtained by adding the sufficient statistics $T_1T_2$, and parametrized by $(\eta_1, \eta_2, \eta_{12})$. By Equation (16), the marginal polytope is represented by the inequalities:

$$\begin{aligned} p_1 &= 1 - \eta_1 - \eta_2 + \eta_{12} \geq 0\ , \\ p_2 &= \eta_2 - \frac{1}{2}\eta_{12} \geq 0\ , \\ p_3 &= \eta_1 - \eta_{12} \geq 0\ , \\ p_4 &= \frac{1}{2}\eta_3 \geq 0\ , \end{aligned}$$

which is a convex set with vertexes $(0,0,0)$, $(0,1,0)$, $(1,0,0)$, $(2,1,2)$, which corresponds to the full marginal polytope associated to the sufficient statistics $\{T_1, T_2, T_1T_2\}$. As the critical set is the edge $\delta_2 \leftrightarrow \delta_4$ in the $\boldsymbol{p}$ space, it is the edge $(0,1,0) \leftrightarrow (2,1,2)$ in the $\boldsymbol{\eta}$ space.

We have the following possible models on the border of the probability simplex and on the border of the full marginal polytope, where the values for $\eta_1$ and $\eta_2$ are obtained from Equation (15).

| $p_1$ | $p_2$ | $p_3$ | $p_4$ | $\eta_1$ | $\eta_2$ |
|---|---|---|---|---|---|
| 0 | 0 | + | + | $p_3 + 2p_4$ | $p_4$ |
| 0 | + | 0 | + | $2p_4$ | $p_2 + p_4$ |
| + | 0 | + | 0 | $p_3$ | 0 |
| + | + | 0 | 0 | 0 | $p_2$ |

| $p_1$ | $p_2$ | $p_3$ | $p_4$ | $\eta_1$ | $\eta_2$ |
|---|---|---|---|---|---|
| + | 0 | 0 | 0 | 0 | 0 |
| 0 | + | 0 | 0 | 0 | 1 |
| 0 | 0 | + | 0 | 1 | 0 |
| 0 | 0 | 0 | + | 2 | 1 |

That is, the domains that can be support of probabilities in the algebraic model are the faces of the marginal polytope. This is general; see [20,34].

*3.2. Fisher Information*

Let us consider the covariance matrix of the sufficient statistics. Let us denote by $A_{|12}$ the block of the two central columns in $A$ in Equation (14) and by $\boldsymbol{p}$ the row vector of probabilities. Then, the variance matrix is:

$$A_{|12}^T \operatorname{diag}(\boldsymbol{p})\, A_{|12} - (\boldsymbol{p}A_{|12})^T \boldsymbol{p}A_{|12} = A_{|12}^T \operatorname{diag}(\boldsymbol{p})\, A_{|12} - A_{|12}^T \boldsymbol{p}^T \boldsymbol{p} A_{|12} = A_{|12}^T \left(\operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}^T\boldsymbol{p}\right) A_{|12}\ .$$

On each of the cases of probabilities supported by a single point, the matrix $\boldsymbol{p} - \boldsymbol{p}^T\boldsymbol{p}$ is zero; hence, the covariance matrix is zero. In each of the cases where the probability is supported by a

facet, say $\{1,2\}$, the matrix $\boldsymbol{p}-\boldsymbol{p}^T\boldsymbol{p}$ reduces to the corresponding block, and the covariance matrix is:

$$\begin{aligned}
&\begin{bmatrix}0&0&1&1\\0&1&0&1\end{bmatrix}\begin{bmatrix}p_1-p_1^2 & -p_1p_2 & 0 & 0\\ -p_1p_2 & p_2-p_2^2 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}\begin{bmatrix}0&0\\0&1\\1&0\\2&1\end{bmatrix}\\
=\;&\begin{bmatrix}0&0\\0&1\end{bmatrix}\begin{bmatrix}p_1-p_1^2 & -p_1p_2\\ -p_1p_2 & p_2-p_2^2\end{bmatrix}\begin{bmatrix}0&0\\0&1\end{bmatrix}\\
=\;&\begin{bmatrix}0 & 0\\0 & p_2-p_2^2\end{bmatrix}.
\end{aligned}$$

The space generated by the covariance matrix is $\mathbb{Q}(0,1)$, that is the affine space that contains the facets itself. Analogous results hold for each facet, and this result is general.

We note that the determinant of the covariance matrix is a polynomial of degree six in the indeterminates $p_1, p_2, p_3$. This polynomial is zero on each facet.

The $\boldsymbol{\eta}$ parameters can be given as a function of either $\boldsymbol{\theta}$ or $\boldsymbol{\zeta}$. We have:

| $\boldsymbol{\eta}$ | $A^T[p_{\boldsymbol{\zeta}}]$ |
|---|---|
| $\eta_1$ | $(\zeta_1+2\zeta_1^2\zeta_2)/(1+\zeta_2+\zeta_1+\zeta_1^2\zeta_2)$ |
| $\eta_2$ | $(\zeta_2+\zeta_1^2\zeta_2)/(1+\zeta_2+\zeta_1+\zeta_1^2\zeta_2)$ |
| $\eta_3$ | $(-2+\zeta_2+2\zeta_1-\zeta_1^2\zeta_2)/(1+\zeta_2+\zeta_1+\zeta_1^2\zeta_2)$ |

(50)

We know from the theory of exponential families that the mapping:

$$]0,\infty[\times]0,\infty[\ni(\zeta_1,\zeta_2)\mapsto(\eta_1,\eta_2)\in\mathrm{Conv}\,\{(T_1(x),T_2(x))|x\in\Omega\}^\circ$$

is one-to-one. We look for an algebraic inversion of the equations:

$$\begin{aligned}
(1+\zeta_2+\zeta_1+\zeta_1^2\zeta_2)\eta_1&=\zeta_1+2\zeta_1^2\zeta_2\,,\\
(1+\zeta_2+\zeta_1+\zeta_1^2\zeta_2)\eta_2&=\zeta_2+\zeta_1^2\zeta_2\,.
\end{aligned}$$

If we rewrite Equations (50) as polynomials in $\zeta_1$, $\zeta_2$, we obtain:

$$\eta_1+(\eta_1-1)\zeta_1+\eta_1\zeta_2+(\eta_1-2)\zeta_1^2\zeta_2=0\,, \tag{51}$$

$$\eta_2+\eta_2\zeta_1+(\eta_2-1)\zeta_2+(\eta_2-1)\zeta_1^2\zeta_2=0\,, \tag{52}$$

$$-\eta_3+(\eta_3-2)\zeta_1+(\eta_3-1)\zeta_2+(\eta_3+1)\zeta_1^2\zeta_2=0\,. \tag{53}$$

Gauss elimination produces a linear system in $\zeta_1,\zeta_2$ with coefficients that are polynomials in $\eta_1,\eta_2,\eta_3$ to be considered with the implicit equation derived from $p_1^2p_4-p_2p_3^2=0$. The system is:

$$\begin{aligned}
-2\eta_2\eta_3-2\eta_1+2\eta_2&=(-2\eta_2\eta_3-2\eta_1+2)\zeta_1+(-2\eta_2\eta_3+2\eta_2+2\eta_3-2)\zeta_2\,,\\
\eta_2&=\eta_2\zeta 1+(\eta_2-1)\zeta_2\,.
\end{aligned}$$

*3.3. Extension of the Model*

In this subsection, we study an extension to signed probabilities of the exponential family in Equations (12) and (13) based on the representation of the statistical model as a ruled surface in the probability simplex. Our motivation for such an analysis is the study of the stability of the critical points of a gradient field in the $\boldsymbol{\eta}$ parameters, in particular when the critical points belong to the boundary of the model. Indeed, by extending the gradient field outside the marginal polytope, we can identify open neighborhoods for critical points on the boundary of the polytope, which allow one to study the convergence of the differential equations associated with the gradient flows, for instance by means of Lyapunov stability.

In the following, we describe more in detail how the extension can be obtained. Let $\boldsymbol{a}$ be a point along the edge $\delta_2 \leftrightarrow \delta_4$ of the full marginal polytope parametrized by $(\eta_1, \eta_2, \eta_3)$ and $\boldsymbol{b}$ the coordinates of the corresponding point over $\delta_1 \leftrightarrow \delta_3$ obtained by intersecting the line of the ruled surface through $\boldsymbol{a}$ with the edge $\delta_1 \leftrightarrow \delta_3$. The values of the $\eta_2$ coordinate for $\boldsymbol{a}$ and $\boldsymbol{b}$ are one and zero, respectively. The other coordinates of $\boldsymbol{b}$ depend on those of $\boldsymbol{a}$ though $\alpha$. First, we obtain the values of the $\eta_3$ coordinates as a function of the $\eta_1$ coordinate. For $\boldsymbol{a}$, we find the equation of the line to which $\delta_2 \leftrightarrow \delta_4$ belongs, given by:

$$\begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} + u \begin{pmatrix} 2 \\ 0 \\ -2 \end{pmatrix} = \begin{pmatrix} 2u \\ 1 \\ 1 - 2u \end{pmatrix} , \tag{54}$$

from which we obtain $\eta_3 = 1 - \eta_1$. Similarly, for the $\eta_3$ coordinate of $\boldsymbol{b}$, we consider the line through $\delta_1 \leftrightarrow \delta_3$, that is:

$$\begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ -2 \end{pmatrix} + u \begin{pmatrix} 1 \\ 0 \\ 4 \end{pmatrix} = \begin{pmatrix} u \\ 0 \\ 4u - 2 \end{pmatrix} , \tag{55}$$

which gives us $\eta_3 = 4\eta_1 - 2$. Finally, for the $\eta_1$ coordinate, we use Equations (44). In $\boldsymbol{a}$, since $t = 0$ and $p_1 = p_3 = 0$, then $p_2 = \frac{\beta^2}{\alpha^2+\beta^2}$ and $p_4 = \frac{\alpha^2}{\alpha^2+\beta^2}$. From Equation (24), it follows that:

$$\eta_1 = \frac{2\alpha^2}{2\alpha^2 + 2\alpha + 1} . \tag{56}$$

Similarly, for $\boldsymbol{b}$, we have $p_2 = p_4 = 0$ and $t = 1$, so that $p_1 = \alpha + 1$ and $p_3 = -\alpha$. From Equation (24), it follows that:

$$\eta_1 = -\alpha . \tag{57}$$

As a result, the coordinates of $\boldsymbol{a}$ and $\boldsymbol{b}$ both depend on $\alpha$ as follows,

$$\boldsymbol{a} = \left( \frac{2\alpha^2}{2\alpha^2 + 2\alpha + 1}, 1, \frac{2\alpha + 1}{2\alpha^2 + 2\alpha + 1} \right) \tag{58}$$

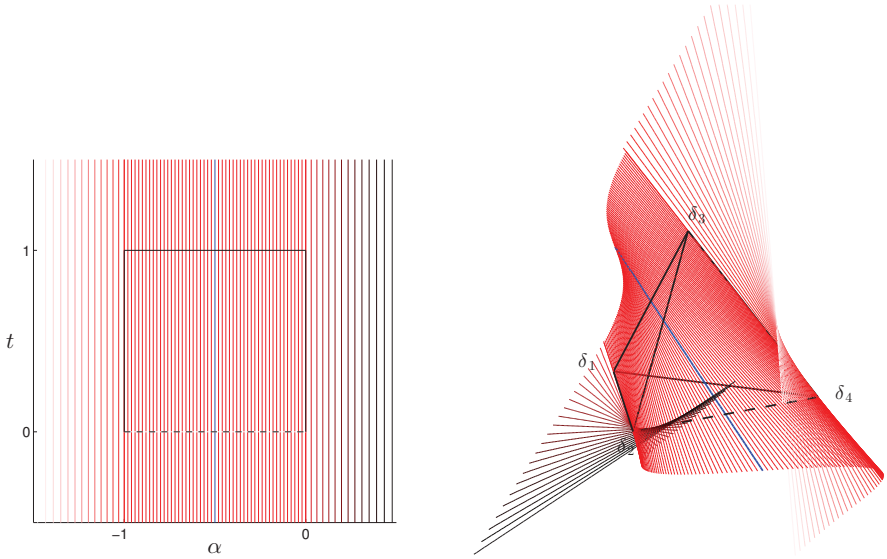
$$\boldsymbol{b} = (-\alpha, 0, -4\alpha - 2) \tag{59}$$

The ruled surface in the full marginal polytope is given by the lines through $\boldsymbol{a}$ and $\boldsymbol{b}$ described by the following implicit representation, for $-1 < \alpha < 1$ and $0 < t < 1$,

$$\begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix} = \begin{bmatrix} -\alpha \\ 0 \\ -4\alpha - 2 \end{bmatrix} + t \begin{bmatrix} \frac{2\alpha^3+4\alpha^2+\alpha}{2\alpha^2+2\alpha+1} \\ 1 \\ \frac{8\alpha^3+12\alpha^2+10\alpha+3}{2\alpha^2+2\alpha+1} \end{bmatrix}. \tag{60}$$

The ruled surface can be extended outside the marginal polytope by taking values of $\alpha, t \in \mathbb{R}$ and considering the set of lines through $\boldsymbol{a}$ and $\boldsymbol{b}$ for different values of $\alpha$. For $\alpha \to \pm\infty$, the $\eta_1$ coordinate of $\boldsymbol{b}$ tends to $\mp\infty$, while the $\eta_1$ of $\boldsymbol{a}$ tends to one. For $\alpha \to \pm\infty$, the ruled surface admits the same limit given by the line parallel to $\delta_1 \leftrightarrow \delta_3$ passing through $(1, 1, 0)$. The surface intersects the interior of the marginal polytope for $t \in (0, 1)$ and $\alpha \in (-1, 0)$. Moreover, the surface intersects the critical line twice, for $t = 0, \alpha \in [-1, 0]$ and for $t = 0, \alpha \notin [-1, 0]$.

In Figures 6 and 7, we represent the extension of the ruled surface outside the probability simplex and in the $(\alpha, t)$ chart, while in Figures 8 and 9, the extended surface has been represented in the full marginal polytope parametrized by $(\eta_1, \eta_2, \eta_3)$ and in the marginal polytope parametrized by $(\eta_1, \eta_2)$.




(a) (b)

**Figure 6.** The segments that form the ruled surface in Figure 4 have been extended, for $-0.5 < t < 1.5$. New lines described by Equations (60) have been represented for $0 < \alpha < \exp(0.7)$ (shading from red to black for increasing values of $\alpha$) and for $\exp(0.7) - 1 < \alpha < -1$ (shading from red to white for decreasing values of $\alpha$). The simplex in (**b**) has been rotated with respect to Figure 4(a) to better visualize the intersection of the lines with the critical edge $\delta_2 \leftrightarrow \delta_4$.

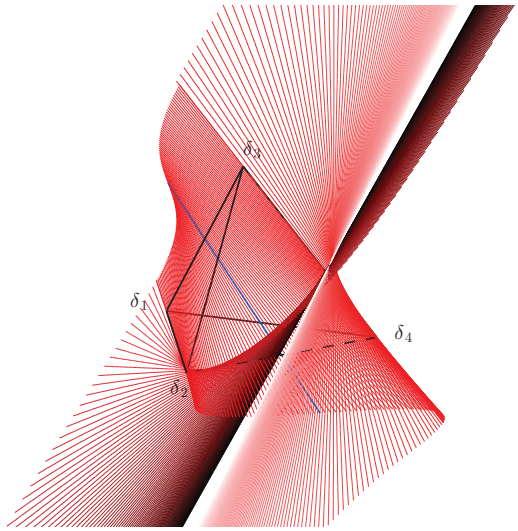
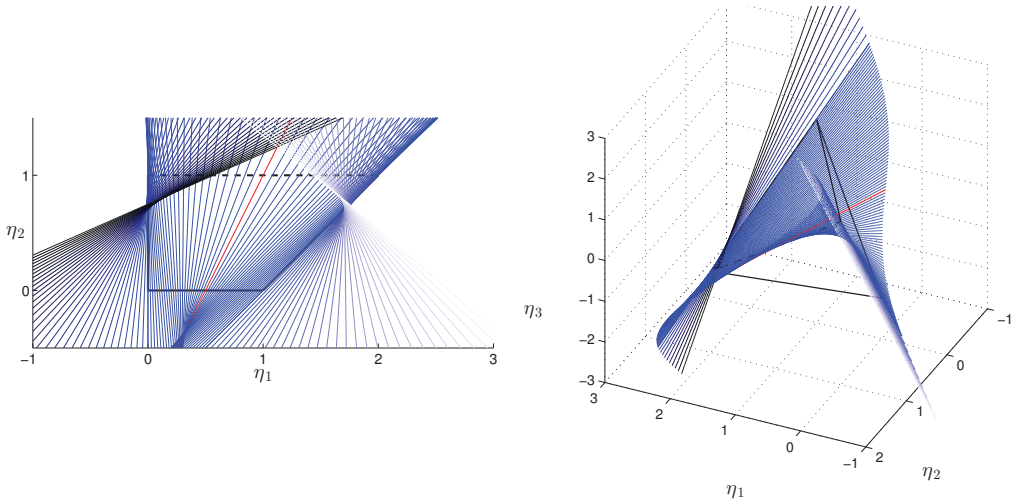



**Figure 7.** Extension of the ruled surface associated with the exponential family in Equations (12) and (13) as in Figure 6(b), for $\exp(3.5) - 1 < \alpha < \exp(3.5)$ and $-0.5 < t < 1.5$; for $\alpha \to \pm\infty$, the lines of the extended surface admit the same limit.




(a) (b)

**Figure 8.** The segments that form the ruled surface in Figure 5 have been extended, for $-0.5 < t < 1.5$. New lines described by Equations (60) have been represented for $0 < \alpha < \exp(1)$ (shading from blue to black for increasing values of $\alpha$) and $\exp(1) - 1 < \alpha < -1$ (shading from blue to white for decreasing values of $\alpha$). The full marginal polytope in (**b**) has been rotated with respect to Figure 5(b) to better visualize the intersection of the lines with the critical edge $\delta_2 \leftrightarrow \delta_4$.

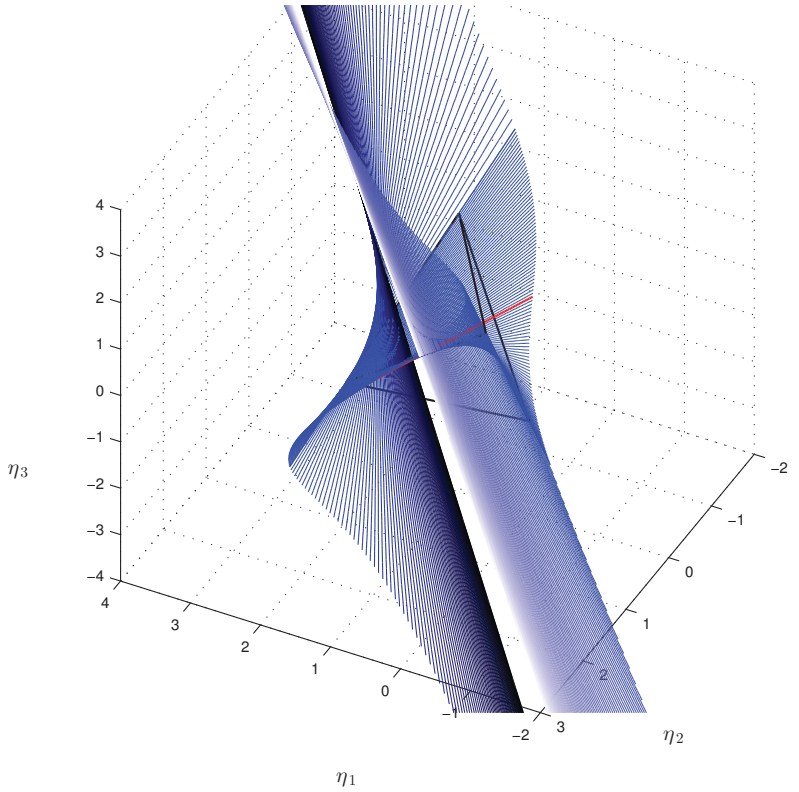



**Figure 9.** Extension of the ruled surface associated with the exponential family in Equations (12) and (13) as in Figure 8(b), for $\exp(3) - 1 < \alpha < \exp(3)$ and $-0.5 < t < 1.5$; notice that for $\alpha \to \pm\infty$, the lines of the extended surface admit the same limit.

### *3.4. Optimization and Natural Gradient Flows*

We are interested in the study of natural gradient flows of functions defined over statistical models. Our motivation is the study of the optimization of the stochastic relaxation of a function, *i.e.*, the optimization of the expected value of the function itself with respect to a distribution $p$ in a statistical model. Natural gradient flows associated with the stochastic relaxation converge to the boundary of the model, where the probability mass is concentrated on some instances of the search space. To study the convergence over the boundary, we proposed to extend the natural gradient field outside the marginal polytope and the probability simplex, by employing a parameterization that describes the model as a ruled surface, as we described in the tutorial example of this section.

In the following, we focus on the optimization of a function $f : \Omega \to \mathbb{R}$, and we consider its stochastic relaxation with respect to a probability distribution in the exponential family in Equations (12) and (13). First, we compute a basis for all real-valued functions defined over $\Omega$ using algebraic arguments. Consider the zero-dimensional ideal $I$ associated with the set of points

in $\Omega$, and let $R$ be the polynomial ring with the field of real coefficients; a vector space basis for the quotient ring $R/I$ defines a basis for all functions defined over $\Omega$. In CoCoA [36], this can be computed with the command `QuotientBasis`.

Coming back to our example, with $\Omega = \{1, 2, 3, 4\}$, by fixing the graded reverse lexicographical monomial order, which is the default one in CoCoA [36], we obtain a basis given by $\{1, x_1, x_2, x_1x_2\}$, so that any $f : \Omega \to \mathbb{R}$ can be written as:

$$f = c_0 + c_1x_1 + c_2x_2 + c_{12}x_1x_2 \, . \tag{61}$$

We are interested in the study of the natural gradient field of $F(p) = \mathbb{E}_p[f]$. Recall that $T_3 = 4x_1 + 3x_2 - 5x_1x_2 - 2$ and $\eta_3 = \mathbb{E}[T_3]$, so that:

$$\mathbb{E}[x_1x_2] = \frac{1}{5}(4\eta_1 + 3\eta_2 - \eta_3 - 2) \, , \tag{62}$$

which implies:

$$F_\eta(\boldsymbol{\eta}) = c_0 - \frac{2}{5}c_{12} + \left(c_1 + \frac{4}{5}c_{12}\right)\eta_1 + \left(c_2 + \frac{3}{5}c_{12}\right)\eta_2 - \frac{1}{5}c_{12}\eta_3 \, . \tag{63}$$

In order to study the gradient field of $F_\eta(\boldsymbol{\eta})$ over the marginal polytope parameterized by $(\eta_1, \eta_2)$, we need to express $\eta_3$ as a function of $\eta_1$ and $\eta_2$. In order to do that, we parametrize the exponential family as a ruled surface by means of the $(\alpha, t)$ parameters. Moreover, this parametrization has a natural extension outside the marginal polytope, which allows one to study the stability of the critical points on the boundary of the marginal polytope. We start by evaluating the gradient field of $F_{\alpha,t}(\alpha, t)$ in the $(\alpha, t)$ parametrization, then we map it to the marginal polytope in the $\boldsymbol{\eta}$ parameterization.

By expressing $(\eta_1, \eta_2)$ as a function of $(\alpha, t)$, we obtain:

$$F_{\alpha,t}(\alpha, t) = \frac{2\alpha^2(c_1 + c_{12}) + (2\alpha^2 + 2\alpha + 1)(c_0 + c_2) - (2\alpha^2(c_1 + c_{12}) + (2\alpha^2 + 2\alpha + 1)(c_1\alpha + c_2))t}{2\alpha^2 + 2\alpha + 1} \, . \tag{64}$$

If we take partial derivatives of Equation (64) with respect to $\alpha$ and $t$, we have:

$$\partial_\alpha F_{\alpha,t}(\alpha, t) = \frac{4(\alpha^2 + \alpha)(c_1 + c_{12}) - ((4\alpha^4 + 8\alpha^3 + 12\alpha^2 + 8\alpha + 1)c_1 + 4(\alpha^2 + \alpha)c_{12})t}{4\alpha^4 + 8\alpha^3 + 8\alpha^2 + 4\alpha + 1} \, , \tag{65}$$

$$\partial_t F_{\alpha,t}(\alpha, t) = -\frac{2\alpha^2 c_{12} + (2\alpha^3 + 4\alpha^2 + \alpha)c_1 + (2\alpha^2 + 2\alpha + 1)c_2}{2\alpha^2 + 2\alpha + 1} \, . \tag{66}$$

In the $(\alpha, t)$ parameterization, the Fisher information matrix reads:

$$I_{\alpha,t}(\alpha, t) = \mathbb{E}_{\alpha,t}[-\partial^2 \log p(x; \alpha, t)] = \begin{bmatrix} \frac{4\alpha^2 - (4\alpha^4 + 8\alpha^3 + 12\alpha^2 + 8\alpha + 1)t + 4\alpha}{4\alpha^6 + 12\alpha^5 + 16\alpha^4 + 12\alpha^3 + 5\alpha^2 + \alpha} & 0 \\ 0 & -(t^2 - t)^{-1} \end{bmatrix} \, . \tag{67}$$

Finally, the natural gradient becomes:

$$\widetilde{\nabla} F_{\alpha,t}(\alpha, t) = I_{\alpha,t}(\alpha, t)^{-1} \nabla F_{\alpha,t}(\alpha, t) \tag{68}$$

$$= \begin{bmatrix} \frac{(4\alpha^6 + 12\alpha^5 + 16\alpha^4 + 12\alpha^3 + 5\alpha^2 + \alpha)(4(\alpha^2 + \alpha)c_1 + 4(\alpha^2 + \alpha)c_{12} - ((4\alpha^4 + 8\alpha^3 + 12\alpha^2 + 8\alpha + 1)c_1 + 4(\alpha^2 + \alpha)c_{12})t)}{(4\alpha^4 + 8\alpha^3 + 8\alpha^2 + 4\alpha + 1)(4\alpha^2 - (4\alpha^4 + 8\alpha^3 + 12\alpha^2 + 8\alpha + 1)t + 4\alpha)} \\ \frac{(2\alpha^2 c_{12} + (2\alpha^3 + 4\alpha^2 + \alpha)c_1 + (2\alpha^2 + 2\alpha + 1)c_2)(t^2 - t)}{2\alpha^2 + 2\alpha + 1} \end{bmatrix}$$

We obtained a rational formula for the natural gradient in the $(\alpha, t)$ parameterization, which can be easily extended outside the marginal polytope. However, notice that the inverse Fisher information matrix and the natural gradient are not defined for:

$$t = \frac{4(\alpha^2 + \alpha)}{4\alpha^4 + 8\alpha^3 + 12\alpha^2 + 8\alpha + 1} \,. \tag{69}$$

We also remark that over the boundary of the model, for $t \in \{0, 1\}$ and $\alpha \in \{-1, 0\}$, the determinant of the inverse Fisher information vanishes, so that the matrix is not full rank. It follows that the trajectories associated with natural gradient flows with initial conditions in the interior of the marginal polytope remain in the marginal polytope.

In order to study the natural gradient field over the marginal polytope, we apply a reparameterization of a tangent vector from the $(\alpha, t)$ parameterization to the $(\eta_1, \eta_2)$ parameterization. Indeed, by the chain rule and the inverse function theorem, we have:

$$\nabla F_\eta(\alpha, t) = \nabla F_{\alpha,t}(\alpha, t)^T J(\alpha, t)^{-1} \tag{70}$$

The Jacobian of the map $(\alpha, t) \mapsto (\eta_1, \eta_2)$ is:

$$J(\alpha, t) = \begin{bmatrix} -\frac{(6\alpha^2+8\alpha+1)t-4\alpha}{2\alpha^2+2\alpha+1} & -\frac{2(2\alpha^2-(2\alpha^3+4\alpha^2+\alpha)t)(2\alpha+1)}{(2\alpha^2+2\alpha+1)^2} & -\frac{2\alpha^3+4\alpha^2+\alpha}{2\alpha^2+2\alpha+1} \\ 0 & & -1 \end{bmatrix} , \tag{71}$$

with inverse:

$$J(\alpha, t)^{-1} = \begin{bmatrix} \frac{4\alpha^4+8\alpha^3+8\alpha^2+4\alpha+1}{4\alpha^2-(4\alpha^4+8\alpha^3+12\alpha^2+8\alpha+1)t+4\alpha} & -\frac{4\alpha^5+12\alpha^4+12\alpha^3+6\alpha^2+\alpha}{4\alpha^2-(4\alpha^4+8\alpha^3+12\alpha^2+8\alpha+1)t+4\alpha} \\ 0 & -1 \end{bmatrix} . \tag{72}$$

It follows that:

$$\nabla F_\eta(\alpha, t) = \begin{bmatrix} \frac{4(\alpha^2+\alpha)c1+4(\alpha^2+\alpha)c12-((4\alpha^4+8\alpha^3+12\alpha^2+8\alpha+1)c1+4(\alpha^2+\alpha)c12)t}{4\alpha^2-(4\alpha^4+8\alpha^3+12\alpha^2+8\alpha+1)t+4\alpha} \\ -\frac{4(\alpha^3+\alpha^2)c12-4(\alpha^2+\alpha)c2+(2(2\alpha^4-\alpha^2)c12+(4\alpha^4+8\alpha^3+12\alpha^2+8\alpha+1)c2)t}{4\alpha^2-(4\alpha^4+8\alpha^3+12\alpha^2+8\alpha+1)t+4\alpha} \end{bmatrix} . \tag{73}$$

Notice that, as for the inverse Fisher information matrix, the inverse Jacobian $J(\alpha, t)^{-1}$ is not defined for $t$ which satisfies Equation (69).

We compute the inverse Fisher information matrix by evaluating the covariance between the sufficient statistics of the exponential family. Since over $\Omega$, we have $x_1^2 = x_1 + x_1 x_2$ and $x_1^2 = x_1$, it follows that:

$$I_\eta(\eta)^{-1} = \begin{bmatrix} \frac{1}{5}(9\eta_1 + 3\eta_2 - \eta_3 - 2) - \eta_1^2 & \frac{1}{5}(4\eta_1 + 3\eta_2 - \eta_3 - 2) - \eta_1\eta_2 \\ \frac{1}{5}(4\eta_1 + 3\eta_2 - \eta_3 - 2) - \eta_1\eta_2 & \eta_2 - \eta_2^2 \end{bmatrix} . \tag{74}$$

By parameterizing $I_\eta^{-1}$ with $(\alpha, t)$, we have:

$$I_\eta(\alpha,t)^{-1} \tag{75}$$
$$= \begin{bmatrix} \frac{4\alpha^4+8\alpha^3-(4\alpha^6+16\alpha^5+20\alpha^4+8\alpha^3+\alpha^2)t^2+4\alpha^2+(4\alpha^5-12\alpha^3-8\alpha^2-\alpha)t}{4\alpha^4+8\alpha^3+8\alpha^2+4\alpha+1} & -\frac{(2\alpha^3+4\alpha^2+\alpha)t^2-(2\alpha^3+4\alpha^2+\alpha)t}{2\alpha^2+2\alpha+1} \\ -\frac{(2\alpha^3+4\alpha^2+\alpha)t^2-(2\alpha^3+4\alpha^2+\alpha)t}{2\alpha^2+2\alpha+1} & -t^2+t \end{bmatrix}.$$

Finally, we derive the following rational formula for the natural gradient over the marginal polytope parametrized as a ruled surface by $(\alpha, t)$:

$$\widetilde{\nabla} F_\eta(\alpha,t) = I_\eta(\alpha,t)^{-1}\nabla F_\eta(\alpha,t) \tag{76}$$
$$= \begin{bmatrix} \frac{((4\alpha^6+16\alpha^5+20\alpha^4+8\alpha^3+\alpha^2)c_1+2(2\alpha^5+4\alpha^4+\alpha^3)c_{12}+(4\alpha^5+12\alpha^4+12\alpha^3+6\alpha^2+\alpha)c_2)t^2-4(\alpha^4+2\alpha^3+\alpha^2)c_1+}{-4(\alpha^4+2\alpha^3+\alpha^2)c_{12}-((4\alpha^5-12\alpha^3-8\alpha^2-\alpha)c_1+2(2\alpha^5+2\alpha^4-3\alpha^3-2\alpha^2)c_{12}+(4\alpha^5+12\alpha^4+12\alpha^3+6\alpha^2+\alpha)c_2)t} \\ \frac{4\alpha^4+8\alpha^3+8\alpha^2+4\alpha+1}{-\frac{(2\alpha^2c_{12}+(2\alpha^3+4\alpha^2+\alpha)c_1+(2\alpha^2+2\alpha+1)c_2)t^2-(2\alpha^2c_{12}+(2\alpha^3+4\alpha^2+\alpha)c_1+(2\alpha^2+2\alpha+1)c_2)t}{2\alpha^2+2\alpha+1}} \end{bmatrix}.$$

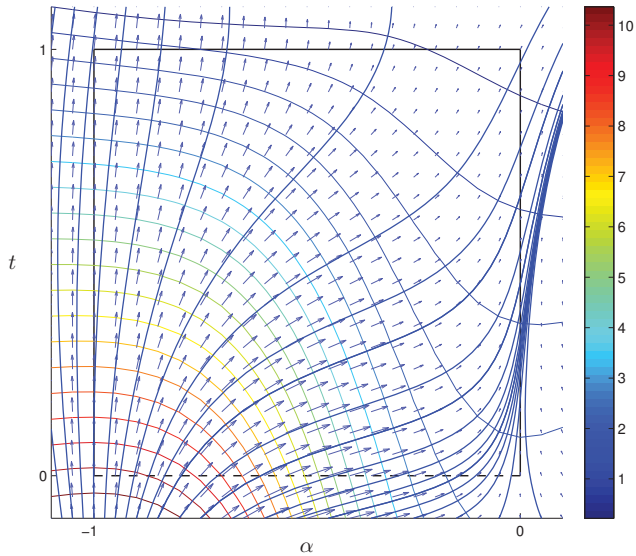
### *3.5. Examples with Global and Local Optima*

We conclude this section with two examples of natural gradient flows associated with two different $f$ functions. First, consider the case where $c_0 = 0, c_1 = 1, c_2 = 2, c_3 = 3$, so that:

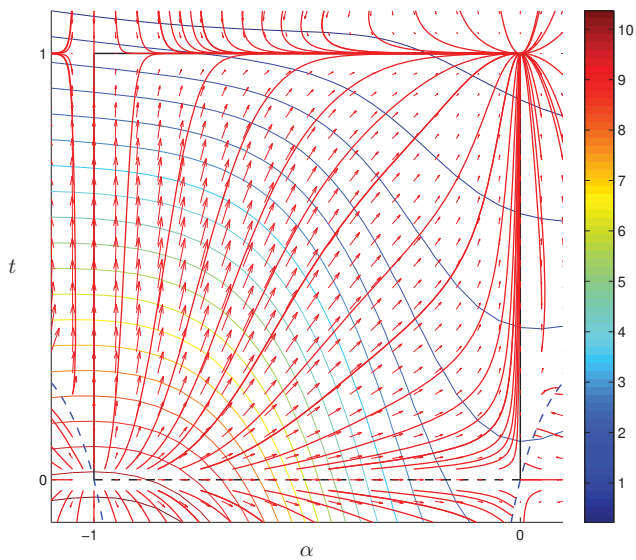
| $\Omega$ | $x_1$ | $x_2$ | $f_1$ |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 2 | 0 | 1 | 2 |
| 3 | 1 | 0 | 1 |
| 4 | 2 | 1 | 10 |

. (77)

The function admits a minimum on $\{1\}$. In Figure 10, we plotted the vector fields associated with the vanilla and natural gradient, together with some gradient flows for different initial conditions, in the $(\alpha, t)$ parameterization. In Figure 11, we represent the vanilla and natural gradient field over the marginal polytope in the $(\eta_1, \eta_2)$ parameterization. Notice that, as expected, differently from the vanilla gradient, the natural gradient flows converge to the unique global optima, which corresponds to the vertex where all of the probability is concentrated over $\{1\}$. In the $(\alpha, t)$ parameterization, the flows have been extended outside the statistical model by prolonging the lines of the ruled surface, and as we can see, they remain compatible with the flows on the interior of the model, in the sense that the nature of the critical point is the same for trajectories with initial conditions on the interior and on the exterior of the model. In other words, the global optima is an attractor from both the interior and the exterior of the model and similarly for the other critical points on the vertices, both for saddle points and the unstable points, where the natural gradient vanishes.

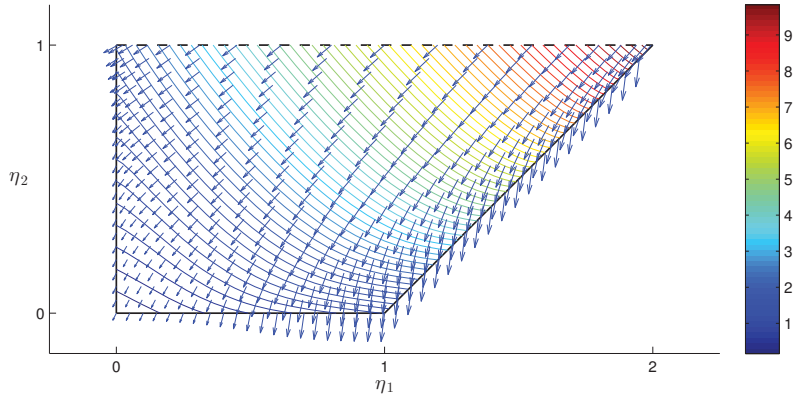



(a)

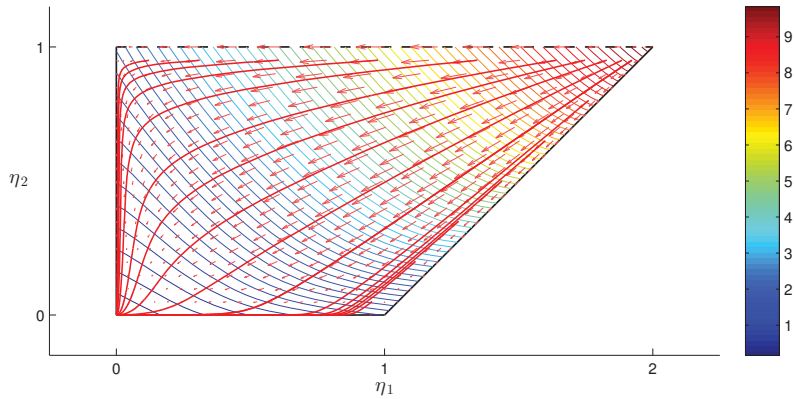



(b)

**Figure 10.** Vanilla gradient field and flows in blue (**a**) and natural gradient field and flows in red (**b**), together with level lines associated with $F_{\alpha,t}(\alpha, t)$ in the $(\alpha, t)$ parameterization, for $c_0 = 0, c_1 = 1, c_2 = 2$ and $c_3 = 3$; the dashed blue lines in (**b**) represent the points where $\widetilde{\nabla} F_{\alpha,t}(\alpha, t)$ is not defined; see Equation (68).




(a)




(b)

**Figure 11.** Vanilla gradient field in blue (**a**) and natural gradient field and flows in red (**b**), together with level lines associated with $F_\eta(\alpha, t)$ over the marginal polytope, for $c_0 = 0$, $c_1 = 1, c_2 = 2$ and $c_3 = 3$.

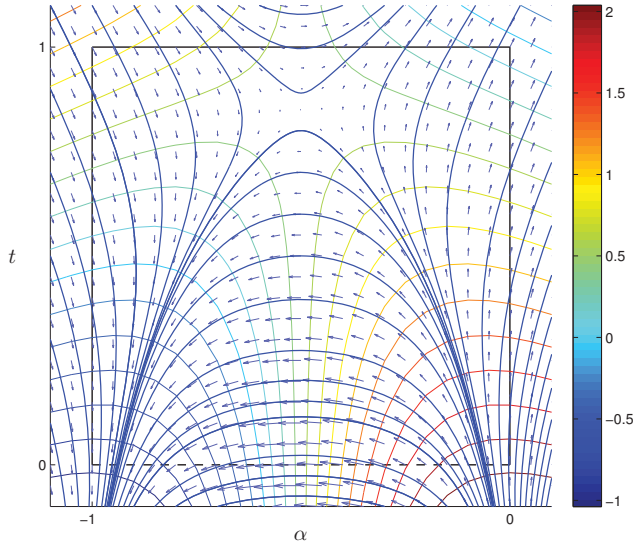
In the second example, we set $c_0 = 0, c_1 = 1, c_2 = 2, c_3 = -5/2$, and we have:

| $\Omega$ | $x_1$ | $x_2$ | $f_2$ |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 2 | 0 | 1 | 2 |
| 3 | 1 | 0 | 1 |
| 4 | 2 | 1 | $-1$ |

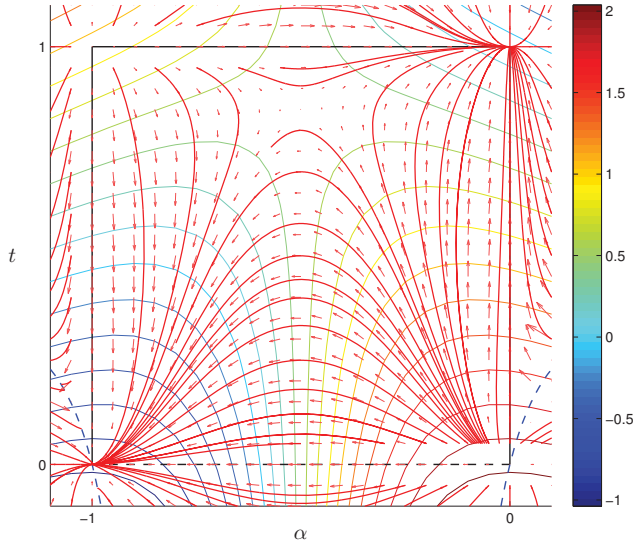
(78)

so that $f_2$ admits a minimum on $\{4\}$. In Figures 12 and 13, we plotted the vector fields associated with the vanilla and natural gradient, together with some gradient flows for different initial conditions, in the $(\alpha, t)$ and $(\eta_1, \eta_2)$ parameterization, respectively. As in the previous example, natural gradient flows converge to the vertices of the model; however, in this case, we have one local optima in $\{1\}$

and one global optima in $\{4\}$, together with a saddle point in the interior of the model. Similarly to the previous example, in the $(\alpha, t)$ parameterization, the flows have been extended outside the statistical model, and the nature of the critical points is the same for trajectories with initial conditions in the statistical model and in the extension of the statistical model.

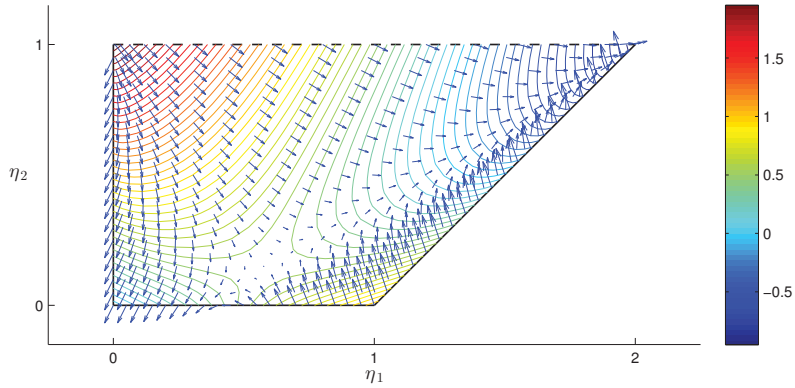



(a)

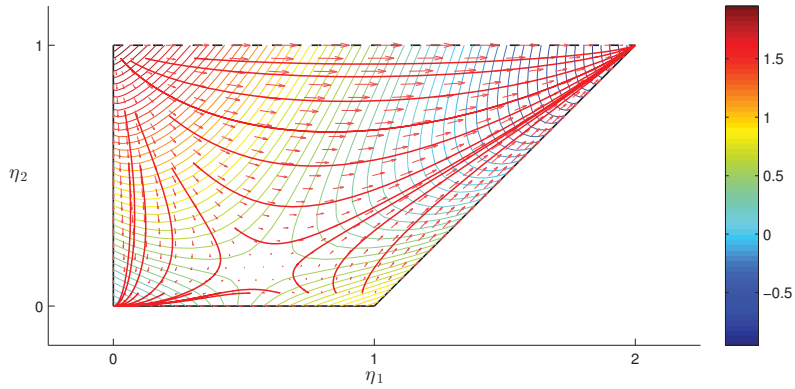



(b)

**Figure 12.** Vanilla gradient field and flows in blue (**a**) and natural gradient field and flows in red (**b**) as in Figure 10, for $c_0 = 0, c_1 = 1, c_2 = 2$ and $c_3 = -\frac{5}{2}$.




(a)




(b)

**Figure 13.** Vanilla gradient field in blue (**a**) and natural gradient field and flows in red (**b**) as in Figure 11, for $c_0 = 0, c_1 = 1, c_2 = 2$ and $c_3 = -\frac{5}{2}$.

We conclude the section by noticing that in both examples, for certain values of $t$ in Equation (69), the natural gradient flows are not defined on the extension of the statistical model. As represented in the figures, once a trajectory encounters the dashed blue line in the $(\alpha, t)$ parameterization, the flow stops at that point.

## 4. Pseudo-Boolean Functions

We turn to discuss a case of considerable practical interest to see which of the results obtained in the example of the previous section we are able to extend.

For binary variables, we use the coding $\pm 1$, that is $\boldsymbol{x} = (x_1, \dots, x_n) \in \{+1, -1\}^n = \Omega$. For any function $f \colon \Omega \mapsto \mathbb{R}$, with multi-index notation, $f(\boldsymbol{x}) = \sum_{\boldsymbol{\alpha} \in L} a_{\boldsymbol{\alpha}} \boldsymbol{x}^{\boldsymbol{\alpha}}$, with $L = \{0, 1\}^n$ and $\boldsymbol{x}^{\boldsymbol{\alpha}} = \prod_{i=1}^n x_i^{\alpha_i}$, $0^0 = 1$. If $M \subset L^* = L \setminus \{\mathbf{0}\}$, the model where $p \in \mathcal{E}$ if:

$$p \propto \exp\left(\sum_{\boldsymbol{\alpha}\in M} \theta_{\boldsymbol{\alpha}} \boldsymbol{X}^{\boldsymbol{\alpha}}\right) = \prod_{\boldsymbol{\alpha}\in M} \left(\mathrm{e}^{\theta_{\boldsymbol{\alpha}}}\right)^{\boldsymbol{X}^{\boldsymbol{\alpha}}}$$

has been considered in a number of papers on combinatorial optimization; see [3–5]. The following statements are results in algebraic statistics; *cf.* [20,35]. Let $\mathcal{P}^1 = \left\{ f \in \mathbb{R}^{\Omega} \middle| \sum_{\boldsymbol{x}\in\Omega} p(\boldsymbol{x}) = 1 \right\}$.

**Proposition 6** (Implicitization of the exponential family)**.** *Given a function $p\colon \Omega \to \mathbb{R}$, then $p \in \mathcal{E}$ if, and only if, the following conditions all hold:*

1. $p(\boldsymbol{x}) > 0$, $\boldsymbol{x} \in \Omega$*;*
2. $\sum_{\boldsymbol{x}\in\Omega} p(\boldsymbol{x}) = 1$*;*
3. $\prod_{\boldsymbol{x}\colon \boldsymbol{x}^{\boldsymbol{\beta}}=1} p(\boldsymbol{x}) = \prod_{\boldsymbol{x}\colon \boldsymbol{x}^{\boldsymbol{\beta}}=-1} p(\boldsymbol{x})$ *for all* $\boldsymbol{\beta} \in L^* \setminus M$ .

**Proof.** ($\Rightarrow$) If $p \in \mathcal{E}$, then $p(\boldsymbol{x}) > 0$, $\boldsymbol{x} \in \Omega$ (Item 1) and $\sum_{\boldsymbol{x}\in\Omega} p(\boldsymbol{x}) = 1$ (Item 2). Moreover, $\log p(\boldsymbol{x}) = \sum_{\boldsymbol{\alpha}\in M} \theta_{\alpha} \boldsymbol{x}^{\boldsymbol{\alpha}} - \psi(\boldsymbol{\theta})$. The function $\log p$ is orthogonal to each $\boldsymbol{X}^{\beta}$, $\beta \in L^* \setminus M$. Hence:

$$0 = \sum_{\boldsymbol{x}\in\Omega} \log p(\boldsymbol{x}) \boldsymbol{x}^{\boldsymbol{\beta}} = \sum_{\boldsymbol{x}\colon \boldsymbol{x}^{\boldsymbol{\beta}}=1} \log p(\boldsymbol{x}) - \sum_{\boldsymbol{x}\colon \boldsymbol{x}^{\boldsymbol{\beta}}=-1} \log p(\boldsymbol{x}) = \log \prod_{\boldsymbol{x}\colon \boldsymbol{x}^{\boldsymbol{\beta}}=1} p(\boldsymbol{x}) - \log \prod_{\boldsymbol{x}\colon \boldsymbol{x}^{\boldsymbol{\beta}}=-1} p(\boldsymbol{x}) \,, \quad (79)$$

which is equivalent to Item 3.

($\Leftarrow$) Oppositely, the computation in Equation (79) implies that $\log p$ is orthogonal to each $\boldsymbol{X}^{\boldsymbol{\beta}}$; hence, there exists $\boldsymbol{\theta}$, such that $\log p = \sum_{\boldsymbol{\alpha}\in M} \theta_{\boldsymbol{\alpha}} \boldsymbol{X}^{\boldsymbol{\alpha}} + C$. Now, Item 2 implies $C = -\psi(\boldsymbol{\theta})$. □

Let $\mathbb{R}[\Omega]$ denote the ring of polynomials in the indeterminates $\{p(\boldsymbol{x}) | \boldsymbol{x} \in \Omega\}$. Given a binary model $M$, the set of polynomials:

$$\left\{ \prod_{\boldsymbol{x}\colon \boldsymbol{x}^{\boldsymbol{\beta}}=1} p(\boldsymbol{x}) - \prod_{\boldsymbol{x}\colon \boldsymbol{x}^{\boldsymbol{\beta}}=-1} p(\boldsymbol{x}) \middle| \boldsymbol{\beta} \in L^* \setminus M \right\},$$

generates an ideal $\mathcal{J}(M)$, which is called the toric ideal of the model $M$. Its variety $\mathcal{V}(M)$ is called the exponential variety of $M$.

**Proposition 7.**

1. *The exponential variety of $M$ is the Zariski closure of the exponential model $\mathcal{E}$.*
2. *The closure $\overline{\mathcal{E}}$ of $\mathcal{E}$ in $\mathcal{P}_{\geq}$ is characterized by $p(\boldsymbol{x}) \geq 0$, $\boldsymbol{x} \in \Omega$, together with Items 2 and 3 of Proposition 6.*

3. *The algebraic variety of the ring* $\mathbb{R}[p(\boldsymbol{x})\colon \boldsymbol{x} \in \Omega]$*, which is generated by the polynomials* $\sum_{\boldsymbol{x}\in\Omega} p(\boldsymbol{x}) - 1$, $\prod_{\boldsymbol{x}\colon \boldsymbol{x}^{\boldsymbol{\beta}}=1} p(\boldsymbol{x}) - \prod_{\boldsymbol{x}\colon \boldsymbol{x}^{\boldsymbol{\beta}}=-1} p(\boldsymbol{x})$*,* $\boldsymbol{\beta} \in L^* \setminus M$*, is an extension* $\mathcal{E}^1$ *of* $\mathcal{E}$ *to* $\mathcal{P}^1$*.*

4. *Define the moments* $\eta_{\boldsymbol{\alpha}} = \sum_{\boldsymbol{x}\in\Omega} \boldsymbol{x}^{\boldsymbol{\alpha}} p(\boldsymbol{x})$*,* $\boldsymbol{\alpha} \in L$*,* i.e.*, the discrete Fourier transform of* $p$*, with inverse* $p(\boldsymbol{x}) = 2^{-n} \sum_{\boldsymbol{\alpha}\in L} \boldsymbol{x}^{\boldsymbol{\alpha}} \eta_{\boldsymbol{\alpha}}$*. There exists an algebraic extension of the moment function* $\mathcal{E} \ni p \mapsto \boldsymbol{\eta}(p) \in M^\circ$ *to a mapping defined on* $\mathcal{E}^1$*.*

**Proof.** 1. According to the implicitization Proposition 6, the exponential family is characterized by the positivity condition together with the algebraic binomial conditions.

2. This follows from the implicit form, and it is proven, for example, in [20].

3. By definition.

4. As the mapping from the probabilities to the moments is affine and one-to-one, such a transformation extends to a one-to-one mapping from the extended model to the affine space of the marginal polytope.
□

We conclude this section by introducing the so-called *no three-way interaction* example. On $\Omega = \{0,1\}^3$, the full model in the statistics $0 \mapsto 1$, $1 \mapsto -1$, that is $t = (-1)^x = 1 - 2x$, is described by the matrix:

$$D_3 = \begin{array}{c} \\ 000 \\ 001 \\ 010 \\ 011 \\ 100 \\ 101 \\ 110 \\ 111 \end{array} \begin{array}{c} \begin{array}{cccccccc} 1 & T_3 & T_2 & T_2T_3 & T_1 & T_1T_3 & T_1T_2 & T_1T_2T_3 \end{array} \\ \left[\begin{array}{cccccccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{array}\right] \end{array}. \tag{80}$$

Note the lexicographic order of both the sample points and the statistics' exponents.

The exponential family without the interaction term $T_1T_2T_3$ is the same model as the toric model without the three-way interaction, which is based on the matrix:

$$B = \begin{array}{c} \\ 000 \\ 001 \\ 010 \\ 011 \\ 100 \\ 101 \\ 110 \\ 111 \end{array} \begin{array}{c} \begin{array}{ccccccc} C & \zeta_1 & \zeta_2 & \zeta_3 & \zeta_4 & \zeta_5 & \zeta_6 \end{array} \\ \left[\begin{array}{ccccccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 \end{array}\right] \end{array}. \tag{81}$$

that is the probabilities as a function of the $\zeta$'s are:

$$\begin{cases} p_1 = & c \\ p_2 = & c\zeta_1\zeta_3\zeta_5 \\ p_3 = & c\zeta_2\zeta_3\zeta_6 \\ p_4 = & c\zeta_1\zeta_2\zeta_5\zeta_6 \\ p_5 = & c\zeta_4\zeta_5\zeta_6 \\ p_6 = & c\zeta_1\zeta_3\zeta_4\zeta_6 \\ p_7 = & c\zeta_2\zeta_3\zeta_4\zeta_5 \\ p_8 = & c\zeta_1\zeta_2\zeta_4 \end{cases} . \tag{82}$$

The toric ideal of the toric model in Equation (82) is generated by the polynomial:

$$p_2p_3p_5p_8 - p_1p_4p_6p_7 = 0 \, , \tag{83}$$

this means that the closure of the exponential family is given by the solution of the equations:

$$\begin{cases} p_1 + p_2 + p_3 + p_4 + p_5 + p_6 + p_7 + p_8 = 1 \\ p_2p_3p_5p_8 - p_1p_4p_6p_7 = 0 \end{cases} . \tag{84}$$

The $\eta$ parameters are the expected values of the sufficient statistics of the full model,

$$\begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \\ \eta_5 \\ \eta_6 \\ \eta_7 \end{bmatrix} = \begin{array}{c} \\ 001 \\ 010 \\ 011 \\ 100 \\ 101 \\ 110 \\ 111 \end{array} \begin{array}{c} \begin{array}{cccccccc} 000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \end{array} \\ \begin{bmatrix} 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{bmatrix} \end{array} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \\ p_6 \\ p_7 \\ p_8 \end{bmatrix} . \tag{85}$$

In the ring:

$$R = \mathbb{Q}[p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8, \eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6, \eta_7] \tag{86}$$

we can consider the ideal $\mathcal{I}$ generated by the Equations (84) together with Equations (85). The elimination ideal:

$$\mathcal{J} = \mathcal{I} \cap \mathbb{Q}[\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6, \eta_7] \tag{87}$$

will express the model as a dependence between the $\eta$'s.

Computation with CoCoA [36] gives the following polynomial:

$$
\begin{aligned}
&f(\eta_1,\eta_2,\eta_3,\eta_4,\eta_5,\eta_6;\eta_7) = \\
&\eta_1^2\eta_3\eta_4+\eta_2^2\eta_3\eta_4-\eta_3^3\eta_4-\eta_3\eta_4^3+\eta_1^2\eta_2\eta_5-\eta_2^3\eta_5+\eta_2\eta_3^2\eta_5+\eta_2\eta_4^2\eta_5+\eta_3\eta_4\eta_5^2-\eta_2\eta_5^3-\eta_1^3\eta_6+\eta_1\eta_2^2\eta_6+\eta_1\eta_3^2\eta_6 \\
&+\eta_1\eta_4^2\eta_6+\eta_1\eta_5^2\eta_6+\eta_3\eta_4\eta_6^2+\eta_2\eta_5\eta_6^2-\eta_1\eta_6^3-2\eta_1\eta_2\eta_4-2\eta_1\eta_3\eta_5-2\eta_2\eta_3\eta_6-2\eta_4\eta_5\eta_6+\eta_3\eta_4+\eta_2\eta_5+\eta_1\eta_6 \\
&\quad+\left(-2\eta_1\eta_2\eta_3-2\eta_1\eta_4\eta_5-2\eta_2\eta_4\eta_6-2\eta_3\eta_5\eta_6+\eta_1^2+\eta_2^2+\eta_3^2+\eta_4^2+\eta_5^2+\eta_6^2-1\right)\eta_7 \\
&\qquad+\left(\eta_3\eta_4+\eta_2\eta_5+\eta_1\eta_6\right)\eta_7^2+(-1)\eta_7^3\,. \qquad (88)
\end{aligned}
$$

The equation:

$$f(\eta_1,\eta_2,\eta_3,\eta_4,\eta_5,\eta_6;\eta_7)=0 \qquad (89)$$

is an expression of the model in the expectation parameters, and this expression is a polynomial equation. We know unique solvability in $\eta_7$ if $(\eta_1,\eta_2,\eta_3,\eta_4,\eta_5,\eta_6)$ is in the interior of the marginal polytope. As in the example of the previous section, it is possible to intersect the polynomial invariant in Equation (83) with one or more sheaves of hyperplanes around some faces of the simplex, in order to lower the degree of the invariant and thus decompose the model as the convex hull of probabilities on the boundary of the model. We do not describe the details here, and we postpone the discussion of this example to a paper which is in preparation.

## 5. Conclusions

Geometry and algebra play a fundamental role in the study of statistical models, and in particular in the exponential family. In the fist part of the paper, starting from the definition of the natural gradient over an exponential family, we described the relationship between its expression in the basis of the sufficient statistics and in the conjugate basis. From this perspective, the terms natural gradient and vanilla gradient, to denote gradients evaluated with respect to the Fisher and the Euclidean geometry, together with their duality in the natural and expectation parameters, assume a new meaning, since these definitions depend on the choice of the basis for the tangent space.

In order to study natural gradient flows for a generic discrete exponential model and, in particular, their convergence, it is convenient to move to the mixture geometry of the expectation parameters and to study trajectories over the marginal polytope. However, in order to obtain explicit equations for the flows, it is necessary to determine the dependence between the moments associated with the sufficient statistics of the model, which are constrained to belong to the marginal polytope, and the remaining moments, which on the other side are not free. Such a relationship, which for finite search spaces is given by a system of polynomial invariants, cannot be easily solved explicitly in general. In the second part of the paper, by using algebraic tools, we proposed a novel parameterization based on ruled surfaces for an exponential family, which does not require to solve the polynomial invariant explicitly. We applied our approach to a simple example, and we showed that the surface associated with the model in the full marginal polytope is a ruled surface. We claim that these results are not peculiar to the example we described, and we are working towards an extension of this approach in a more general case.

**Acknowledgments**

The authors would like to thank Gianfranco Casnati from Politecnico di Torino for the useful discussions on the geometry of ruled surfaces. Giovanni Pistone is supported by de Castro Statistics of Collegio Carlo Alberto at Moncalieri and is a member of INdAM/GNAMPA.

**Author Contributions**

Both authors contributed to the design of the research. The research was carried out by all of the authors. The manuscript was written by Luigi Malagò and Giovanni Pistone. Both authors read and approved the final manuscript.

**Conflicts of Interest**

The authors declare no conflict of interest.

**References**

1. Pistone, G. In Proceedings of the First International Conference (GSI 2013), Paris, France, 28–30 August 2013; Nonparametric information geometry. In *Geometric Science of Information*; Nielsen, F., Barbaresco, F., Eds.; Lecture Notes in Computer Science, Volume 8085; Springer: Heidelberg, Germany, 2013; pp. 5–36.
2. Malagò, L.; Matteucci, M.; Pistone, G. Stochastic Relaxation as a Unifying Approach in 0/1 Programming, 2009. In Proceedings of the NIPS 2009 Workshop on Discrete Optimization in Machine Learning: Submodularity, Sparsity & Polyhedra (DISCML), Whistler Resort & Spa, BC, Canada, 11–12 December 2009.
3. Malagò, L.; Matteucci, M.; Pistone, G. Towards the geometry of estimation of distribution algorithms based on the exponential family. In Proceedings of the 11th Workshop on Foundations of Genetic Algorithms (FOGA '11), Schwarzenberg, Austria, 5–8 January 2011; ACM: New York, NY, USA, 2011; pp. 230–242.
4. Malagò, L.; Matteucci, M.; Pistone, G. Stochastic Natural Gradient Descent by estimation of empirical covariances. In Proceedings of the 2011 IEEE Congress on Evolutionary Computation (CEC), New Orleans, LA, USA, 5–8 June 2011; pp. 949–956.
5. Malagò, L.; Matteucci, M.; Pistone, G. Natural gradient, fitness modelling and model selection: A unifying perspective. In Proceedings of the 2013 IEEE Congress on Evolutionary Computation (CEC), Cancun, Mexico, 20–23 June 2013; pp. 486–493.
6. Wierstra, D.; Schaul, T.; Peters, J.; Schmidhuber, J. Natural evolution strategies. In Proceedings of the 2008 IEEE Congress on Evolutionary Computation, Hong Kong, China, 1–6 June 2008; pp. 3381–3387.
7. Ollivier, Y.; Arnold, L.; Auger, A.; Hansen, N. Information-Geometric Optimization Algorithms: A Unifying Picture via Invariance Principles. **2011**, arXiv:1106.3708.

8. Malagò, L.; Pistone, G. Combinatorial Optimization with Information Geometry: Newton method. *Entropy* **2014**, *16*, 4260–4289.
9. Amari, S.; Nagaoka, H. *Methods of Information Geometry*; American Mathematical Society: Providence, RI, USA, 2000; Translated from the 1993 Japanese original by Daishi Harada.
10. Bourbaki, N. *Variétés differentielles et analytiques. Fascicule de résultats / Paragraphes 1 à 7*; Number XXXIII in Éléments de mathématiques; Hermann: Paris, France, 1971.
11. Pistone, G.; Sempi, C. An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *Ann. Stat.* **1995**, *23*, 1543–1561.
12. Malagò, L.; Pistone, G. Gradient Flow of the Stochastic Relaxation on a Generic Exponential Family. In Proceedings of Conference of Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MaxEnt 2014), Clos Lucé, Amboise, France, 21–26 September 2014; Mohammad-Djafari, A., Barbaresco, F., Eds.; pp. 353–360.
13. Brown, L.D. *Fundamentals of Statistical Exponential Families With Applications in Statistical Decision Theory*; Number 9 in IMS Lecture Notes, Monograph Series; Institute of Mathematical Statistics: Hayward, CA, USA, 1986;
14. Rockafellar, R.T. *Convex Analysis*; Princeton Mathematical Series, No. 28; Princeton University Press: Princeton, NJ, USA, 1970.
15. Do Carmo, M.P. *Riemannian Geometry*; Mathematics: Theory & Applications, Birkhäuser Boston Inc.: Boston, MA, USA, 1992; Translated from the second Portuguese edition by Francis Flaherty.
16. Amari, S.I. Natural gradient works efficiently in learning. *Neur. Comput.* **1998**, *10*, 251–276.
17. Shima, H. *The Geometry of Hessian Structures*; World Scientific Publishing Co. Pte. Ltd.: Hackensack, NJ, USA, 2007.
18. Rinaldo, A.; Fienberg, S.E.; Zhou, Y. On the geometry of discrete exponential families with application to exponential random graph models. *Electron. J. Stat.* **2009**, *3*, 446–484.
19. Rauh, J.; Kahle, T.; Ay, N. Support Sets in Exponential Families and Oriented Matroid Theory. *Int. J. Approx. Reas.* **2011**, *52*, 613–626.
20. Malagò, L.; Pistone, G. A note on the border of an exponential family. **2010**, arXiv:1012.0637v1.
21. Pistone, G.; Rogantin, M. The gradient flow of the polarization measure. With an appendix. **2015**, doi:arXiv:1502.06718.
22. Diaconis, P.; Sturmfels, B. Algebraic algorithms for sampling from conditional distributions. *Ann. Stat.* **1998**, *26*, 363–397.
23. Pistone, G.; Wynn, H.P. Generalised confounding with Gröbner bases. *Biometrika* **1996**, *83*, 653–666.
24. Pistone, G.; Riccomagno, E.; Wynn, H.P. *Algebraic Statistics: Computational Commutative Algebra in Statistics*; Volume 89, Monographs on Statistics and Applied Probability, Chapman & Hall/CRC: Boca Raton, FL, USA, 2001.

25. Drton, M.; Sturmfels, B.; Sullivant, S. *Lectures on Algebraic Statistics*; Volume 39, Oberwolfach Seminars; Birkhäuser Verlag: Basel, Germany, 2009.
26. Pachter, L., Sturmfels, B., Eds. *Algebraic Statistics for Computational Biology*; Cambridge University Press: Cambridge, UK, 2005.
27. Gibilisco, P., Riccomagno, E., Rogantin, M.P., Wynn, H.P., Eds. *Algebraic and Geometric Methods in Statistics*; Cambridge University Press: Cambridge, UK, 2010.
28. 4ti2 team. 4ti2—A software package for algebraic, geometric and combinatorial problems on linear spaces. Available online: http://www.4ti2.de (accessed on 2 June 2015).
29. Michałek, M.; Sturmfels, B.; Uhler, C.; Zwiernik, P. Exponential Varieties. **2014**, arXiv:1412.6185.
30. Sturmfels, B. *Gröbner Bases and Convex Polytopes*; American Mathematical Society: Providence, RI, USA, 1996.
31. Geiger, D.; Meek, C.; Sturmfels, B. On the toric algebra of graphical models. *Ann. Stat.* **2006**, *34*, 1463–1492.
32. Rapallo, F. Toric statistical models: Parametric and binomial representations. *Ann. Inst. Stat. Math.* **2007**, *59*, 727–740.
33. Beltrametti, M.; Carletti, E.; Gallarati, D.; Monti Bragadin, G. *Lectures on Curves, Surfaces and Projective Varieties: A Classical View of Algebraic Geometry*; EMS textbooks in mathematics; European Mathematical Society: Zürich, Switzerland, 2009.
34. Rinaldo, A.; Fienberg, S.E.; Zhou, Y. On the geometry of discrete exponential families with application to exponential random graph models. *Electron. J. Stat.* **2009**, *3*, 446–484.
35. Pistone, G. Algebraic varieties *vs.* differentiable manifolds in statistical models. In *Algebraic and Geometric Methods in Statistics*; Gibilisco, P., Riccomagno, E., Rogantin, M., Wynn, H.P., Eds.; Cambridge University Press: Cambridge, UK, 2009; Chapter 21, pp. 339–363.
36. Abbott, J.; Bigatti, A.; Lagorio, G. CoCoA-5: A system for doing Computations in Commutative Algebra. Available online: http://cocoa.dima.unige.it (accessed on 2 June 2015).

# Distributed Consensus for Metamorphic Systems Using a Gossip Algorithm for $CAT(0)$ Metric Spaces

**Anass Bellachehab**

**Abstract:** We present an application of distributed consensus algorithms to metamorphic systems. A metamorphic system is a set of identical units that can self-assemble to form a rigid structure. For instance, one can think of a robotic arm composed of multiple links connected by joints. The system can change its shape in order to adapt to different environments via reconfiguration of its constituting units. We assume in this work that several metamorphic systems form a network: two systems are connected whenever they are able to communicate with each other. The aim of this paper is to propose a distributed algorithm that synchronizes all of the systems in the network. Synchronizing means that all of the systems should end up having the same configuration. This aim is achieved in two steps: (i) we cast the problem as a consensus problem on a metric space; and (ii) we use a recent distributed consensus algorithm that only makes use of metrical notions.

Reprinted from *Entropy*. Cite as: Bellachehab, A. Distributed Consensus for Metamorphic Systems Using a Gossip Algorithm for $\boldsymbol{CAT(0)}$ Metric Spaces. *Entropy* **2015**, *17*, 1165–1180.

## 1. Introduction

Many problems in robotics, computer science and biology involve systems that can be described as reconfigurable or metamorphic. These systems change their state through a set of local rules, and in order to move from a given State A to another given State B, the system has to determine the sequence of local moves that it should perform. Examples of such systems are: metamorphic manufacturing systems [1], phylogenetic trees [2] and metamorphic robots [3].

The examples we will deal with in this paper will be about metamorphic robots, which consists of a collection of individual modules that can connect/disconnect from each other and form a rigid structure, referred to as a configuration or a state. Individual modules can change their position relative to their neighbors, as long as the whole system remains connected and according to a set of local rules. This allows the system to dynamically change its configuration and position. We assume here that we have many identical metamorphic robots that are able to communicate with each other and share information about their respective states. These robots form a communication network that we will assume to be decentralized (*i.e.*, without a central fusion node). We are interested in the problem of distributed consensus among metamorphic systems: each system is in its own initial state, and we would like all of them to end up in a common configuration.

Distributed consensus algorithms have been thoroughly studied, mainly in the cases of vector data [4–6] or ordered data (e.g., [7]). The algorithm in [5], for example, is a distributed algorithm that uses pairwise arithmetic averages of the data. However, without specific assumptions, configurations cannot be averaged. The algorithm presented in [7] relies on pairwise maximum computations. In the setting of metamorphic robots, however, configurations cannot be easily ordered, hence the need to find an appropriate framework.

In [8,9], the authors introduce a mathematical framework for analyzing metamorphic systems based on embedding the state space—the space of all possible states of a given metamorphic system—into a continuous space. This embedding, called the state complex is well suited to a recent consensus algorithm [10] that relies on pairwise midpoint computations. Indeed, the state complex can be equipped with an adequate metric, so as to yield a $CAT(0)$ metric space. Hence, the main contribution of this paper is to propose a distributed consensus algorithm that provably converges and is well adapted to the state space of metamorphic systems.

The paper is organized as follows. The first section details the mathematical background underpinning the metamorphic systems state space embedding, as well as gives a formal description of the consensus problem. The second section exposes the random pairwise midpoint algorithm. Section 3 explains how to compute the midpoint of any two points in a cubical complex. Finally, Section 4 provides numerical simulations for the proposed consensus algorithm.

## 2. Framework

### *2.1. Metamorphic Systems: Definition and Examples*

There are many examples of metamorphic/reconfigurable systems, such as reconfigurable manufacturing systems [1] and phylogenetic trees [2]. Another example, one in which we will be more interested in this paper, is that of metamorphic robots, which were described by Østergaard *et al.* in [11] as robotic systems:

(1) That consist of several identical and physically independent unit modules;
(2) For which its modules can be connected to each other in many possible ways in order to form rigid structures;
(3) For which its modules can disconnect and reconnect while the system is active;
(4) For which it can change the way its modules are connected, *i.e.*, it is fully automatic.

Some of these robots are lattice-based, meaning that the robotic modules occupy a discrete set of possible positions, this set of possible positions forming a lattice. The nature of the lattice depends on the geometry of the modules: it can be hexagonal, squared, dodecahedral, *etc*. These metamorphic systems can be mathematically described as a collection of states on a graph (see Appendix A).

Representing the various states of a metamorphic system by their lattice configuration will prove to be insufficient for finding a simple consensus protocol. Indeed, the lattice representation does not provide an ordering of states. Following [8], we represent a system configuration as a point in a cubical complex $\mathcal{S}$, called the state complex (see, also, Appendix B for a definition). The zero-dimensional skeleton of this complex is the set of states, and two vertices are linked by an edge if their corresponding states differ by a single action of a generator (see Appendix A). A $k$-cube of the complex represents $k$ commutative movements, *i.e.*, movements that are non-overlapping whatever their order (see [8] for a rigorous definition).

An example of a metamorphic system is the robotic arm, which consists of attached links, inside a grid with one of its extremities attached to the base point $(0, 0)$ of the grid (see Figure 1).

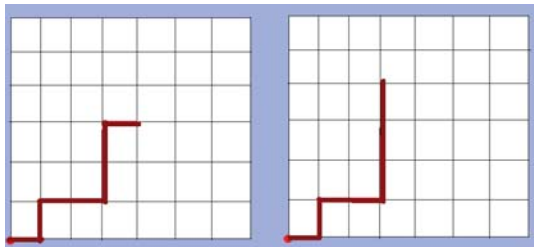


**Figure 1.** Example of a lattice-based reconfigurable system: the robotic arm. The edges in color indicate the presence of a unit module; a black edge indicates its absence. The arm is attached at its base point $(0, 0)$. Here, an elementary movement has been performed by the module at the end of the arm, which changes the overall form of the system.

We assume in this work that the metamorphic systems form a network; two systems are connected whenever they are able to communicate with each other. The aim is to synchronize all of the systems that compose the network, *i.e.*, all of the systems should have the same configuration, as shown in the example of Figure 2.

Having described the model chosen for metamorphic systems, we next review the mathematical framework of the consensus problem.

### *2.2. Framework of the Consensus Problem*

#### 2.2.1. Network

Following the approach of [5], we model the network of metamorphic systems by a connected graph $G = (V, E)$ with vertices $V$ representing the metamorphic systems (agents) and whose edges $E$ represent the communication links between these agents. We assume the graph to be undirected, which means that if an agent can communicate with another agent, then the converse is also true; this hypothesis is not too unrealistic if we suppose that all of the agents are identical and that the movement speed of the agents is very small compared to their communication speed. When a communication link exists between two agents, we say that the two agents are neighbors. We denote by $\mathcal{N}(v)$ the set of all neighbors of the agent $v \in V$.

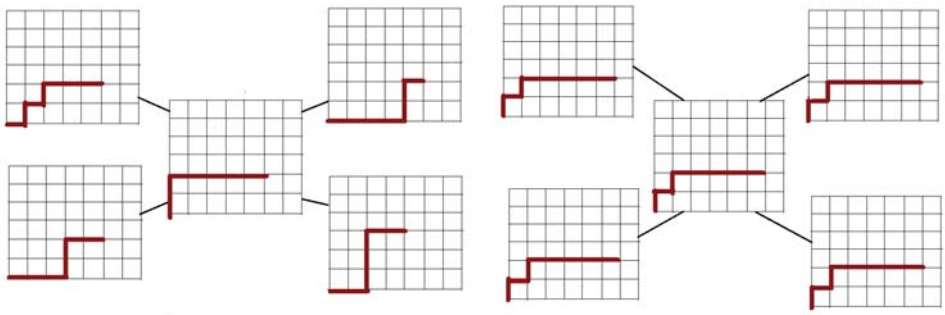


**Figure 2.** In this example, we are given five robotic arms. In the leftmost figure (describing the initial state), each arm has its own configuration. The rightmost figure represents a consensus state, in which all of the arms share a common configuration.

As in [5], we assume that the time model is asynchronous, *i.e.*, that each agent has its own Poisson clock that ticks with a common intensity $\lambda$ (the clocks are identically made), and moreover, each clock is independent of the other clocks. When an agent clock ticks, the agent is able to perform some computations and wake up some neighboring agents. This time model has the same probability distribution as a global single clock ticking with intensity $N\lambda$ and selecting uniformly randomly a single agent at each tick. This equivalence is described, e.g., in [5]. Notice also that link $e = \{v, w\}$ is not necessarily used by agents $v$ and $w$ at a given time: $v$ or $w$ might not be awakened.

### *2.3. Communication*

At a given time $k$, we denote by $V_k$ the agent whose clock ticked and by $W_k$ the neighbor that was in turn awaken. Therefore, at time $k$, the only communicating agents in the whole network are $V_k$ and $W_k$. A single link is then active at each time; hence, at a given time, most links are not used. We assume that $(V_k, W_k)$ are independent and identically distributed and that the distribution of $V_k$ is uniform over the network, while the distribution of $W_k$ is uniform in the neighborhood of $V_k$. More precisely, the probability distribution of $(V_k, W_k)$ is given by:

$$\mathbb{P}[V_k = v, W_k = w] = \begin{cases} \frac{1}{N}\frac{1}{\deg(v)} & \text{if } v \sim w \\ 0 & \text{otherwise} \end{cases}$$

Notice that this probability is not symmetric in $(v, w)$. It is going to turn out to be convenient to also consider directly the link $\{V_k, W_k\}$, forgetting which node was the first to wake up and which node was second. In this case, $\mathbb{P}[\{V_k, W_k\} = \{v, w\}]$ is of course symmetric in $(v, w)$. One has:

$$\mathbb{P}[\{V_k, W_k\} = \{v, w\}] = \begin{cases} \frac{1}{N}\left(\frac{1}{\deg(v)} + \frac{1}{\deg(w)}\right) & \text{if } v \sim w \\ 0 & \text{otherwise} \end{cases}$$

The communication framework considered here is standard [5].

2.3.1. Data Space

Each node $v \in V$ can store a value $x_v \in \mathcal{S}$ that lies in a cubical complex $\mathcal{S}$ (see Appendix B).

We assume that the cubical complex is equipped with the metric $d$ induced from the Euclidean metric on each cube. The skeleton of the complex is a metric graph called the transition graph and is denoted.

Initially, each node $v$ has an initial value $x_v(0)$, and $X_0 = (x_1(0), \ldots, x_N(0))$ is the tuple of initial values. A consensus state has the form $X_\infty = (x_\infty, \ldots, x_\infty)$ with: $x_\infty \in \mathcal{S}$. We denote by $x_v(k)$ the value stored by the agent $v \in V$ at time $k$ and $X_k = (x_1(k), \ldots, x_N(k))$ the global state of the system at that instant.

We want to apply a distributed algorithm in order to achieve the consensus state. $(\mathcal{S}, d)$ being a metric space, a candidate algorithm could be the average pairwise midpoint [10] algorithm. This algorithm requires $(\mathcal{S}, d)$ to be a $CAT(0)$ metric space.

**Definition 1.** *Let $(M, d)$ be a metric space. A geodesic curve $c$ in $X$ is a map $c : [0, l] \to X$ from a closed interval $I = [0, l]$ to $X$, such that, for all $t.t' \in I$:*

$$d(c(t), c(t')) = |t - t'|$$

*A metric space is said to be geodesic if and only if, for any two points $x, y \in X$, there exists a geodesic $c : [0, l] \to X$, such that $c(0) = x$ and $c(l) = y$.*

**Definition 2** ($CAT(0)$ inequality)**.** *Assume $(X, d)$ is a geodesic metric space (a metric space in which any two points can be related by a geodesic) and $\Delta = (c_0, c_1, c_2)$ is a geodesic triangle with vertices $p = c_0(0)$, $q = c_1(0)$ and $r = c_2(0)$. Let $\bar{\Delta} = (\bar{p}, \bar{q}, \bar{r})$ denote a comparison triangle (a triangle with the same edge lengths as $\Delta$) in the Euclidean space $\mathbb{R}^2$. $\Delta$ is said to satisfy the $CAT(0)$ inequality if, for any $x = c_0(t)$ and $y = c_2(t')$, one has:*

$$d(x, y) \leq \bar{d}(\bar{x}, \bar{y})$$

*where $\bar{x}$ is the unique point of $[\bar{p}, \bar{q}]$, such that $d(p, x) = \bar{d}(\bar{p}, \bar{x})$, and $\bar{y}$ on $[\bar{p}, \bar{r}]$, such that $d(p, y) = \bar{d}(\bar{p}, \bar{y})$.*

*A geodesic metric space is said to be locally $CAT(0)$ if any geodesic triangle of sufficiently small perimeter verifies the $CAT(0)$ inequality. It is said to be globally $CAT(0)$ if any geodesic triangle verifies the $CAT(0)$ inequality.*

For a thorough introduction to the subject, see [12,13].

Any state complex can be shown to be a locally $CAT(0)$ space [14]. The global $CAT(0)$ proprietyrequires an additional constraint on the state complex. In [14], a combinatorial criterion based on the notion of posets with inconsistent pairs is provided to verify whether a state complex is globally $CAT(0)$.

**Assumption 1.** *We shall make the fundamental assumption that the state complex of any metamorphic system involved in this paper is globally $CAT(0)$. This assumption, while restrictive, still covers many interesting examples, like the robotic arm* [3]*, phylogenetic trees* [2]*, the hexagonal system* [8]*, etc.*

The following proposition links the existence and uniqueness of geodesics and midpoints with the $CAT(0)$ property:

**Proposition 1.** *If $x$ and $y$ are two points in a globally $CAT(0)$ space $X$, then there is but one geodesic $\gamma : [0,1] \to X$, such that $\gamma(0) = x$ and $\gamma(1) = y$, which we will denote from now on as $[x, y]$. The midpoint of $x$ and $y$ is defined as $\left\langle \frac{x+y}{2} \right\rangle := [x, y](\frac{1}{2})$ and is always well defined and unique.*

Now that the working framework is set (connected and undirected communications graph, Poisson clocks, data space, *etc*.), we will propose in the next section a consensus algorithm based on the gossip protocol [5] adapted to $CAT(0)$ metric spaces.

## 3. Algorithm

### *3.1. Description*

In this section, we expose a distributed algorithm that relies on distributed midpoint computation in $\mathcal{S}$ to drive a system of identical metamorphic systems into a consensus configuration—the random pairwise midpoint algorithm [10]—which works as follows: At each count of the virtual global clock, one node $v$ is selected uniformly randomly from the set of agents $V$. The node $v$ then randomly selects a node $w$ from the neighbors of $v$ in the communications graph. Both node $v$ and $w$ then compute and update their value to $\left\langle \frac{x_v + x_w}{2} \right\rangle$.

This algorithm is well defined, since for any couple of points $x, y \in \mathcal{S}$, their midpoint exists and is unique. This is due to the fact that $\mathcal{S}$ is a globally $CAT(0)$ space.

**Algorithm 1** Random Pairwise Midpoint.

**Input**: a graph $G = (V, E)$ and the initial nodes configuration $X_v(0), v \in V$
**for all** $k > 0$ **do**
At instant $k$, uniformly randomly choose a node $V_k$ from $V$ and a node $W_k$ uniformly randomly from $\mathcal{N}(V_k)$.
Update:
$X_{V_k}(k) = \left\langle \frac{X_{V_k}(k-1) + X_{W_k}(k-1)}{2} \right\rangle$
$X_{W_k}(k) = \left\langle \frac{X_{V_k}(k-1) + X_{W_k}(k-1)}{2} \right\rangle$
$X_v(k) = X_v(k-1)$ for $v \notin \{V_k, W_k\}$
**end for**

This algorithm belongs to the class of consensus protocols, which has been shown to achieve convergence towards a consensus state in [15]; the convergence rate is established as linear in [10]. This assumes, however, that it is possible to compute the midpoint of any two points with reasonable complexity. While there is no closed expression for the midpoint of any two points in $\mathcal{S}$, there is a procedure described in the next section that permits the computation of the midpoint.

## 4. Computing the Midpoint

In [16], an algorithm is given for computing means and medians in general $CAT(0)$ metric spaces; we are here interested in finding the midpoint $\left\langle \frac{x+y}{2} \right\rangle$ of any two given points $x$ and $y$ of a $CAT(0)$ cubical complex $\mathcal{S}$, so we first determine the geodesic $[x, y]$ between $x$ and $y$.

Let $x = (x_1 \dots x_N)$ and $y = (y_1 \dots y_N)$ be the coordinates of $x$ and $y$ in the current standard embedding of $\mathcal{S}$ in $\mathbb{Z}^N$ (see Appendix C for an exposition of the standard embedding).

Then, we define $v = (v_1, \dots, v_N)$ and $w = (w_1, \dots, w_N)$, such that for $i \in \{1, \dots, N\}$:

$$\begin{cases} v_i = 0 & \text{if } x_i \leq 0.5 \\ v_i = 1 & \text{if } x_i > 0.5 \end{cases}$$

Additionally:

$$\begin{cases} w_i = 0 & \text{if } y_i \leq 0.5 \\ w_i = 1 & \text{if } y_i > 0.5 \end{cases}$$

$v$ and $w$ are, respectively, the closest vertices of $\mathcal{S}$ to $x$ and $y$ (see Figure 3).

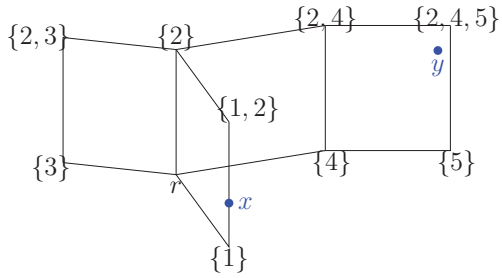



**Figure 3.** Example of a cubical complex with an initial root vertex $r$. We want the geodesic between the points $x = (1, 0.25, 0, 0, 0)$ and $y = (0, 0.75, 0, 1, 0.75)$.

Next, we re-root the cubical complex at $v$ and change the labeling of its vertices consequently, as well as the coordinates of $x$ and $y$. To obtain the new coordinates of any point $a = (a_1, \dots, a_N)$ in the new standard embedding, we update: $a_{i,new} = 1 - a_{i,old}$ if $v_i = 1$ and $a_{i,new} = a_{i,old}$ if $v_i = 0$ (see Figures 3 and 4).

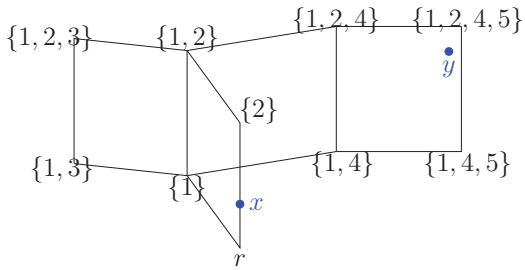



**Figure 4.** We re-root the cubical complex at $v$, and the new coordinates are: $x = (0, 0.25, 0, 0, 0)$ and $y = (1, 0.75, 0, 1, 0.75)$. The coordinates of $w$ are $(0, 1, 0, 1, 1)$.

We denote by $I_w = \{i \in \{1, \dots N\} | w_i = 1\}$, the order ideal associated with the vertex $w$. Now that the complex is re-rooted, we need to find the cubical sequence that contains the geodesic between $x$ and $y$. According to [14], a valid sequence $(C_k)_{1 \leq k \leq n} = (I_k, M_k)_{1 \leq k \leq n}$ of cubes containing the geodesic, consists of a sequence of order ideals $I_1 \subset I_2 \subset \cdots \subset I_n = I_w$ and maximal antichains $M_k \subset I_k$, $1 \leq k \leq n$ (see Appendix B for the representation of individual cubes of a complex). To find this sequence, we take the subset of minimal elements of $I_w$ and $I_1$, and set $M_1 = I_1$. Then, in order to form $C_k = (I_k, M_k)$ from $C_{k-1} = (I_{k-1}, M_{k-1})$, we take the subset of minimal elements $m_k$ of $Q - I_{k-1}$, put $I_k = I_{k-1} \cup m_k$ and let $M_k$ be the maximal antichain of $I_k$. Using this procedure until $I_k = P$, we obtain a valid cube sequence, which contains the geodesic. In the example of Figures 3 and 4, we have: $I_1 = \{2, 3\}$, $M_1 = \{2, 3\}$ and $I_2 = \{2, 3, 4\}$, $M_2 = \{2\}$.

After the cube sequence has been determined, we then have to find the 'breakpoints' from which the geodesic passes. The points $\{p_1, \dots, p_{n-1}\}$ are such that: $\|x - p_1\| + \|p_1 - p_2\| + \cdots + \|p_{n-1} - p_n\| + \|p_n - y\|$ is minimal and that for each $k \in \{1, \dots, n-1\}$ $p_k \in F_k$ with $F_k$ being the common frontier of $(I_k, M_k)$ and $(I_{k+1}, M_{k+1})$, which is the cubical cell $(I_k, M_k \cap M_{k+1})$. This problem can be cast as a touring problem with $n$ polyhedral regions and $2n$ facets [14].

$$\min t_0 + t_1 + \cdots + t_n$$

$$\forall k \in \{0 \dots n\} : \; t_k \geq \|p_k - p_{k+1}\|$$

$$p_k \in F_k, \;\; k = 1 \dots n$$

$$p_0 = x, p_{n+1} = y$$

The touring problem is a second order cone optimization problem, for which numerical solvers exist. We solve the touring problem and obtain a series of points $p_1 \dots p_n$ and the distances $t_1 \dots t_{n-1}$ between $p_k$ and $p_{k+1}$ (see Figure 5).

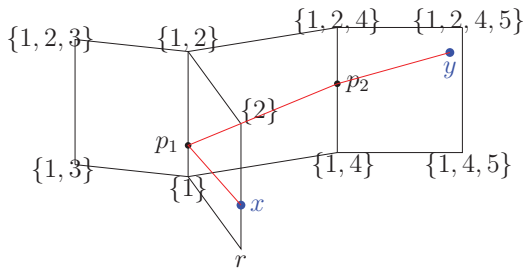



**Figure 5.** In this example, the cube complex and the points $x$ and $y$ are such that we have a non-co-linear sequence of points $\{p_1, p_2\}$ that the geodesic between $x$ and $y$ (drawn in red) passes through. The cube sequence $\{C_1, C_2, C_3\}$ contains this geodesic.

Knowing the sequence of points $(p_i)_{1\leq i\leq n}$ and distances $(t_i)_{0\leq i\leq n}$, we want to determine the midpoint of $x$ and $y$; we have two cases:

Case 1: $n > 0$.

First, we determine the cubical cell of $(C_k)_{1\leq k\leq n+1}$ that contains the midpoint. Let $i_0 = \min\left\{1 \leq i \leq n | d(x, p_i) \geq \frac{d(x,y)}{2}\right\}$; then, we have: $\left\langle \frac{x+y}{2} \right\rangle \in C_{i_0}$ and $\left\langle \frac{x+y}{2} \right\rangle \in [p_{i_0}, p_{i_0+1}]$. Using the convention $p_0 = x$ and $p_{n+1} = y$, the analytic expression of $\left\langle \frac{x+y}{2} \right\rangle$ is:

$$\left\langle \frac{x+y}{2} \right\rangle = p_{i_0} + \frac{D}{t_{i_0}}(p_{i_0+1} - p_{i_0})$$

with:

- $D = \frac{1}{2}\left|\sum_{k=0}^{i_0-1} t_k - \sum_{k=i_0+1}^{n+1} t_k - t_{i_0}\right|$: If $2 \leq i_0 \leq n-1$.
- $D = \frac{1}{2} t_{i_0}$: If $i_0 = 1$.
- $D = \frac{1}{2}\left|t_{i_0} - \sum_{k=0}^{i_0-1} t_k\right|$: If $i_0 = n$.

Case 2: $n = 0$.

In this case, $x$ and $y$ belong to the same cubical cell, and we have: $\left\langle \frac{x+y}{2} \right\rangle = \frac{x+y}{2}$.

## 5. Numerical Results

In this section, we apply the previously described algorithm to two examples of a lattice-based metamorphic systems, the robotic arm, and the hexagonal lattice robot.

To understand why the robotic arm is $CAT(0)$, one must construct a poset with inconsistent pairs associated with the state complex of the arm (see Appendix B.2).

**Definition 3.** *Define for any robotic arm with $n$ articulations $R_n$ the set $P_n = \{(x, y) | y \geq 0, y \leq x, x \leq n-1\}$, and define the partial order relation $\leq$, such that $(x_1, y_1) \leq (x_2, y_2)$ if and only if $x_1 \leq x_2$ and $y_1 \leq y_2$.*

Using the partially ordered set $P_n$, we can show that the state complex of $R_n$ is $CAT(0)$ through the following proposition:

**Proposition 2.** [3] *Let $\mathcal{S}_n$ be the cubical complex of $R_n$ rooted at the state where the arm is completely horizontal.*

*Then, there is a bijection between the possible states of $R_n$ and the order ideals of $P_n$.*

We analyze the convergence of the pairwise midpoint algorithm using as a criterion the variance function:

**Definition 4.** *Given a configuration $x = (x_1, \ldots, x_N) \in \mathcal{S}^N$, the variance function is defined as:*

$$\sigma^2(x) = \frac{1}{N} \sum_{\{v,w\} \in \mathcal{P}_2(V)} d^2(x_v, x_w)$$

*where $d$ is the distance between two points in $\mathcal{S}$*

If $\sigma^2(x) = 0$, then $x_1 = x_2 \cdots = x_N$, and we have achieved a consensus state. In [10], it is proven that for a sequence of points $X(k) = (x_1(k), \ldots, x_N(k))$ generated according to the pairwise midpoint algorithm, the function $\sigma^2$ converges to zero at a linear rate.

**Theorem 1.** *Let $X_k = (x_1(k), ..., x_N(k))$ denote the sequence of random variables generated by Algorithm 1; then, there exists $L < 0$, such that,*

$$\limsup_{k \to \infty} \frac{\log \mathbb{E}[\sigma(X_k)]}{k} \leq L$$

In order to randomly sample a set of initial coordinates, we begin by assigning to each metamorphic system a vector $(x_1, \ldots, x_N)$, where each $x_i$ is sampled randomly and uniformly from the interval $[0, 1]$. Then, we do the following: we check if there exists $i, j$, such that $i \prec j$ and $x_j \neq 0$; if so, then set $x_i = 1$ (the $\prec$ relation being the partial order of the poset associated with the cubical complex $\mathcal{S}$ in its initial rooting).

### *5.1. Results for the Robotic Arm*

We plot the log-variance $\log \sigma^2$ as a function of the number of iterations $k$, for a complete graph of $10$ robotic arms of $n = 7$ joints (see Figure 6), then for a path graph of $10$ robotic arms of the same type. In both cases, we observe in Figure 7 a linear curve with a negative slope in accordance with the results of [10]. In the case of the path graph, however, the slope is smaller than for the complete graph.

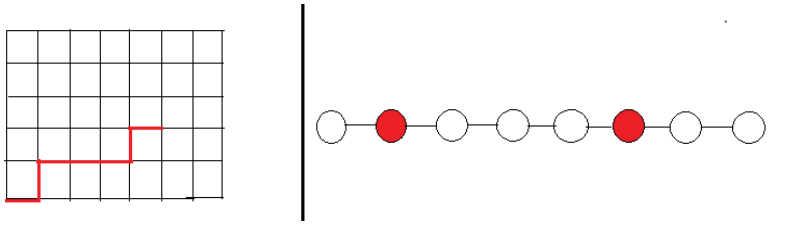
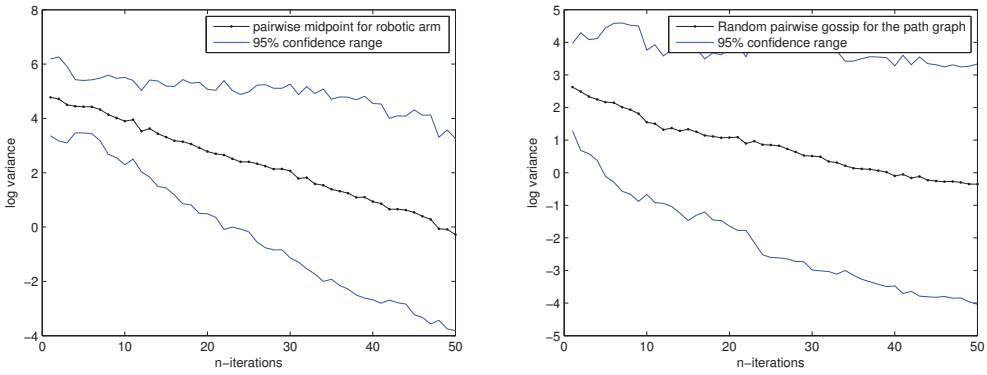


**Figure 6.** On the right, a robotic arm with $n = 7$ articulations and the associated state graph.




(**a**) (**b**)

**Figure 7.** Plot of $n \to \log \sigma^2$ for a network of $k = 10$ robotic arms with $n = 7$ articulations. (**a**) The underlying communications graph is the complete graph, while on (**b**), it is the path graph. Because of the stochastic nature of the algorithm, the procedure is averaged over 30 runs. The resulting curve is a line of a negative slope of bigger magnitude for the complete graph than for the path graph.

### *5.2. Results for the Planar Hexagonal Lattice*

The same analysis is applied to the hexagonal lattice system [8], which is a connected aggregate of hexagonally-shaped modules that occupy a planar lattice. Its graph representation and associated cubical complex can be seen in Figure 8.

We observe similar patterns as for the robotic arms experiment; in Figure 9 the function $\log \sigma^2$ decreases at a linear rate for both the complete graph and the path graph cases.

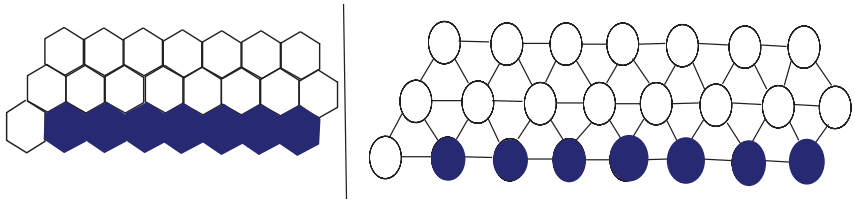
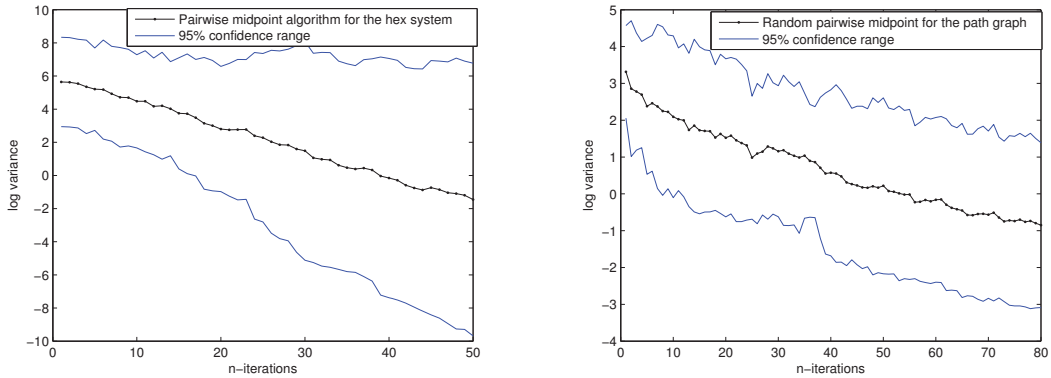


**Figure 8.** On the right, a hexagonal lattice system and its associated state graph.




(**a**) (**b**)

**Figure 9.** Plot of $n \to \log \sigma^2$ for a network of $k = 10$ hexagonal systems. (**a**) The underlying communications graph is the complete graph, while on (**b**), it is the path graph. Because of the stochastic nature of the algorithm, the procedure is averaged over 30 runs. The resulting curve is a line of a negative slope of bigger magnitude for the complete graph than for the path graph.

## 6. Conclusions

The random pairwise midpoint algorithm can be successfully applied to discrete combinatorial systems, like the case of metamorphic systems. This is a non-trivial example of a purely metric application of this algorithm, whose exponential convergence towards a consensus state has been confirmed through numerical experiments on two examples of metamorphic systems. The same method could be applied to the space of phylogenetic trees. Another application could be that of distributed optimization, where $N$ identical metamorphic systems, each with its own utility function $f_i : \mathcal{S} \to \mathbb{R}$: $i \in \{1, \ldots, N\}$, try to minimize a collective objective function $\frac{1}{N}\sum_{i=1}^{N} f_i$; the underlying communication network has no fusion node.

## Appendix

### A. Modeling Metamorphic Systems

For a rigorous definition of metamorphic systems, we follow the approach of [8]. We represent the lattice by its dual graph $\mathcal{G} = (\mathcal{V}(\mathcal{G}), \mathcal{E}(\mathcal{G}))$, whose vertices represent the individual cells of the lattice; two vertices are joined by an edge if and only if their corresponding cells are adjacent in the lattice. We associate with this graph a set of labels $\mathcal{A}$ on the set of vertices to indicate whether the corresponding cell is occupied or not and, in the former case, whether it is occupied by an obstacle or by a module and of which type. A state of the system is any map $U : \mathcal{G} \to \mathcal{A}$. The metamorphic system dynamically changes its state through a set of elementary movements

that satisfy the following rules: (i) units cannot overlap during reconfiguration; and (ii) overall connectivity should always be maintained. This is done via generators, which are defined below:

**Definition 5.** [9] *Let $\mathcal{G} = (\mathcal{V}(\mathcal{G}), \mathcal{E}(\mathcal{G}))$ be a graph and $\mathcal{A}$ a set of labels. A generator $\phi$ is a collection of three objects:*

- *A support $\mathrm{SUP}(\phi) \subset \mathcal{G}$, which is a subgraph of $\mathcal{G}$;*
- *A trace, $\mathrm{TR}(\phi) \subset \mathrm{SUP}(\phi)$, which is a subgraph of $\mathrm{SUP}(\phi)$;*
- *An non-ordered pair of states $u_0, u_1 : \mathcal{V}(\mathrm{SUP}(\phi)) \to \mathcal{A}$, verifying:*

$$u_0|_{\mathcal{V}(\mathrm{SUP}(\phi))-\mathcal{V}(\mathrm{TR}(\phi))} = u_1|_{\mathcal{V}(\mathrm{SUP}(\phi))-\mathcal{V}(\mathrm{TR}(\phi))}$$

*where $\mathcal{V}(\mathrm{SUP}(\phi))$ are the vertices of $\mathrm{SUP}(\phi)$.*

The support of the generator corresponds to the region of the graph that contains the necessary information to verify whether the movement is feasible (*i.e.*, the absence of obstacles, no modules overlapping and the connectivity of the system is maintained). The trace of the generator is the region of the graph where the movement actually takes place.

**Definition 6.** [9] *A generator is said to be admissible at a state $U$ if: $U|_{\mathcal{V}(\mathrm{SUP}(\phi))} = \hat{u}_0$. The action of $\phi$, denoted $\Phi$, maps a state $u : \mathcal{V}(\mathcal{G}) \to \mathcal{A}$ to a new one $\Phi[u] : \mathcal{V}(\mathcal{G}) \to \mathcal{A}$, given by:*

$$\Phi[U] := \begin{cases} U & : \text{on } \mathcal{V}(\mathcal{G}) - \mathcal{V}(\mathrm{TR}(\phi)) \\ \hat{u}_1 & : \text{on } \mathcal{V}(\mathrm{TR}(\phi)) \end{cases}$$

Thus, a generator $\phi$ acts on a state $U$ by modifying its restriction on the vertices of $\mathrm{TR}(\phi)$ from $u_0$ to $u_1$. This corresponds to an elementary movement of unit modules. However, a metamorphic system should be able to perform many such elementary movements simultaneously; in order for that to be feasible, the movements have to be compatible, *i.e.*, their simultaneous execution does not lead to module overlap or loss of system connectivity. For that, we introduce the following definition:

**Definition 7.** [9] *In a metamorphic system, a collection of generators $\{\phi_i\}$ is said to be commutative if:*

$$i \neq j \Rightarrow \mathrm{TR}(\phi_i) \cap \mathrm{SUP}(\phi_j) = \emptyset,$$

## B. The State Complex

### *B.1. Definition*

A cubical complex can be seen as a collection of cubes glued together using isometries.

**Definition 8.** [17] (p.112) *Let $\Gamma \subset \mathbb{N}$ and $(C_i)_{i\in\Gamma}$ be a collection of Euclidean unit cubes of various dimensions. Additionally, let $X = \coprod_{i\in\Gamma} C_i$ be a disjoint union of these cubes. A cubical complex $\mathcal{S}$ is the quotient of $X$ by an equivalence relation $\sim$, such that if $p : X \to \mathcal{S}$ is the natural projection, then:*

- *For every $i \in \Gamma$, the restriction $p_i$ of $p$ to the cube $C_i$ is injective.*
- *If $p_i(C_i) \cap p_j(C_j) \neq \emptyset$, then there is an isometry $h_{i,j}$ from a face $T_i \subset C_i$ onto a face $T_j \subset C_j$, such that $p_i(x) = p_j(x')$ if and only if $x' = h_{i,j}(x)$.*

**Definition 9.** [9] *The state complex $\mathcal{S}$ of a metamorphic system is a cubical complex. Each $k$-cube $e^{(k)}$ of $\mathcal{S}$ is an equivalence class $[u, (\phi_i)_{i=1}^k]$, where:*

- *$(\phi_i)_{i=1}^k$ is a $k$-upletof commuting generators.*
- *$u$ is a state for which the generators $(\phi_i)_{i=1}^k$ are admissible.*
- *$[u_0, (\phi_i)_{i=1}^k] = [u_1; (\phi'_i)_{i=1}^k]$ if and only if $\exists \sigma \in \mathfrak{t}_k$, such that: $\forall i \in \{1, \ldots, k\}$; we have: $\phi_i = \phi'_{\sigma(i)}$; and $u_0 = u_1$ on the subgraph: $\mathcal{G} - \bigcup \mathrm{TR}(\phi_{\alpha_i})$.*

*The boundary of each $k$-cube is a collection of $2k$ faces:*

$$\partial[u; (\phi_{\alpha_i})_{i=1}^k] = \bigcup_{i=1}^{k} ([u; (\phi_{\alpha_j})_{j \neq i}] \cup [\phi_{\alpha_i}[u]; (\phi_{\alpha_j})_{j \neq i}])$$

One advantage of using the cubical complex representation over that of the transition graph is that it shows which elementary movements can be performed simultaneously; and, thus, contains more information than the transition graph. It is not generally computationally feasible to construct the cubical complex associated with a given reconfigurable system, but there are some interesting cases where it is possible to do so, like the case of the robotic arm [3].

*B.2. Encoding as a Partially-Ordered Set*

Given a cubical complex $\mathcal{S}$ and a vertex $v$ (called the root vertex) of $\mathcal{S}$, one can introduce a partial order relation in the set of vertices of $\mathcal{S}$ by stating that a $u \prec w$ if and only if there is an edge path geodesic (*i.e.*, a geodesic on the metric graph associated with $\mathcal{S}$) from the root $v$ to $w$ that passes through $u$. The complex $\mathcal{S}$ can thus be seen as a partially-ordered set. In [14], it is shown that if one can choose the special node $v$, such that the set of vertices of $\mathcal{S}$ equipped with the partial order relation $\prec$ is a poset with inconsistent pairs, then $\mathcal{S}$ is a globally $CAT(0)$ cubical complex. This last propriety guarantees the existence and uniqueness of the midpoint of any given two points and allows us to apply the random midpoint algorithm on a network of agents whose data are encoded in a cubical complex. We assume in this paper that all of the studied complexes are $CAT(0)$.

Reciprocally, given a poset with inconsistent pairs $(P, \prec)$, one can build a lattice whose summits are the order ideals of $(P, \prec)$ that do not contain any inconsistent pairs (such order ideals are said to be consistent). An edge is drawn between two summits if their corresponding ideals differ by exactly one element. The vertices are ordered according to their corresponding ideals, $u < w$ if and only if the order ideal $I_u$ corresponding to the vertex $u$ is a subset of $I_w$, the order ideal associated with the vertex $w$ ($u < w \Leftrightarrow I_u \subset I_w$), as shown in Figure B1.

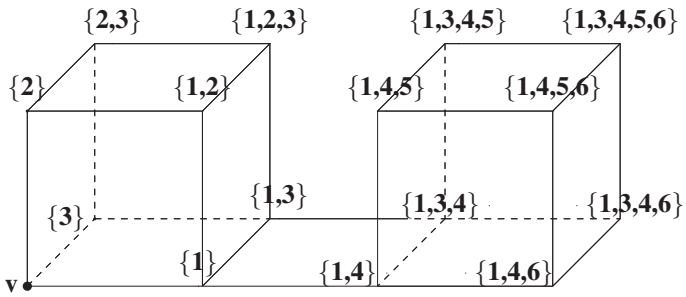



**Figure B1.** The cubical complex associated with the poset $(P = \{1, 2, 3, 4, 5, 6\}, \prec)$, such that: $1 \prec 4$, $1 \prec 5$ and $4 \prec 6$. Each vertex is labeled by a consistent order ideal of $P$.

Individual cubes are represented by a pair $(I, M)$, with $I$ a consistent order ideal and $M$ a subset of the maximal element of $I$; such a cube is of dimension $|M|$, and its vertices are found by removing from $I$ all of the possible subsets of $M$ (Figure B2 gives an example).

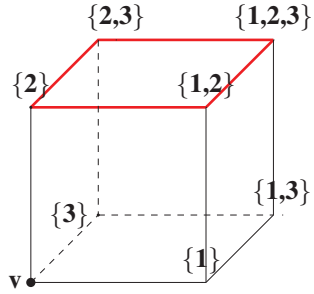



**Figure B2.** In this example, the 2-dimensional cube highlighted in color is represented by the pair $(I, M)$ with $I = \{1, 2, 3\}$ and $M = \{1, 3\}$.

## C. The Standard Embedding

One way to embed the cubical complex associated with a poset with inconsistent pairs of cardinal $N$ is the so-called standard embedding [14]. An element $u \in \mathcal{S}$ is represented by an $N$-tuple $u = (u_1, \ldots, u_N)$, where $\forall i \; u_i \in [0, 1]$, and if for some $(i, j) \in \{1, \ldots, N\}^2$, we have $u_j \neq 0$ and $u_j > u_i$, then we must have $u_i = 1$. Furthermore, if $x_i$ and $x_j$ are inconsistent, then: $u_i u_j = 0$.

Figure 3 shows an example of a cubical complex where the standard embedding coordinates of the points $x$ and $y$ are: $x = (1; 0.25; 0.25; 0)$ and $y = (1; 0.75; 1; 0.75)$.

## Conflicts of Interest

The authors declare no conflict of interest.

## References


1. Koren, Y.; Heisel, U.; Jovane, F.; Moriwaki, T.; Pritschow, G.; Ulsoy, G.; van Brussel, H. Reconfigurable Manufacturing Systems. In *Manufacturing Technologies for Machines of the Future*; Dashchenko, A., Ed.; Springer: Berlin/Heidelberg, Germany, 2003; Part IV, Chapter 19, pp. 627–665.
2. Billera, L.J.; Holmes, S.P.; Vogtmann, K. Geometry of the space of phylogenetic trees. *Adv. Appl. Math.* **2001**, *27*, 733–767.
3. Ardila, F.; Baker, T.; Yatchak, R. Moving Robots Efficiently Using the Combinatorics of CAT(0) Cubical Complexes. **2012**, arXiv preprint arXiv:1211.1442.
4. Tsitsiklis, J. Problems in Decentralized Decision Making and Computation. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 1984.
5. Boyd, S.; Ghosh, A.; Prabhakar, B.; Shah, D. Randomized gossip algorithms. *IEEE Trans. Inf. Theory* **2006**, *52*, 2508–2530.
6. DeGroot, M.H. Reaching a consensus. *J. Am. Stat. Assoc.* **1974**, *69*, 118–121.
7. Nejad, B.M.; Attia, S.A.; Raisch, J. Max-Consensus in a Max-Plus Algebraic Setting: The Case of Fixed Communication Topologies. In Proceedings of the IEEE XXII International Symposium on Information, Communication and Automation Technologies (ICAT 2009), Sarajevo, Bosnia and Herzegovina, 29–31 October 2009; pp. 1–7.
8. Abrams, A.; Ghrist, R. State complexes for metamorphic robots. *Int. J. Robot. Res.* **2004**, *23*, 811–826.
9. Ghrist, R.; Peterson, V. The geometry and topology of reconfiguration. *Adv. Appl. Math.* **2007**, *38*, 302–323.
10. Bellachehab, A.; Jakubowicz, J. Random Pairwise Gossip on $CAT(\kappa)$ Metric Spaces. **2014**, arXiv preprint arXiv:1405.4190.
11. Østergaard, E.H.; Kassow, K.; Beck, R.; Lund, H.H. Design of the ATRON lattice-based self-reconfigurable robot. *Auton. Robots* **2006**, *21*, 165–183.
12. Bacak, M. *Convex Analysis and Optimization in Hadamard Spaces*; De Gruyter Series in Nolinear Analysis and Applications; De Gruyter: Genthiner Strasse 13, Berlin, Germany, 2014; Volume 22.
13. Jost, J. *Nonpositive Curvature: Geometric and Analytic Aspects*; Lectures in Mathematics. ETH Zurich; Springer: Basel, Switzerland, 1997.
14. Ardila, F.; Owen, M.; Sullivant, S. Geodesics in CAT (0) cubical complexes. *Adv. Appl. Math.* **2012**, *48*, 142–163.
15. Grohs, P. *Wolfowitz's Theorem and Convergence of Consensus Algorithms in Hadamard Spaces*; Technical Report SAM Report 2012-27; ETH: Zurich, Switzerland, 2012.
16. Bacak, M. Computing medians and means in Hadamard spaces. *SIAM J. Optim.* **2014**, *24*, 1542–1566.
17. Bridson, M.; Haefliger, A. *Metric Spaces of Non-Positive Curvature*; Springer: Berlin/Heidelberg, Germany, 1999.

# Geometric Shrinkage Priors for Kählerian Signal Filters

**Jaehyung Choi and Andrew P. Mullhaupt**

**Abstract:** We construct geometric shrinkage priors for Kählerian signal filters. Based on the characteristics of Kähler manifolds, an efficient and robust algorithm for finding superharmonic priors which outperform the Jeffreys prior is introduced. Several ansätze for the Bayesian predictive priors are also suggested. In particular, the ansätze related to Kähler potential are geometrically intrinsic priors to the information manifold of which the geometry is derived from the potential. The implication of the algorithm to time series models is also provided.

Reprinted from *Entropy*. Cite as: Choi, J.; Mullhaupt, A.P. Geometric Shrinkage Priors for Kählerian Signal Filters. *Entropy* **2015**, *17*, 1347–1357.

## 1. Introduction

In information geometry, signal processing is one of the most important applications. In particular, an information geometric approach to various linear time series models has been also well-known [1–7]. The geometric description of the linear systems is not confined to the pursuit of mathematical beauty. Komaki's work [8] is in the line of developing practical tools for Bayesian inference. Using the Kullback–Leibler divergence as a risk function for estimation, he found that superharmonic shrinkage priors outperform the Jeffreys prior in the viewpoint of information theory. Better prediction in the Bayesian framework is attainable by the Komaki priors.

However, a difficult part of Komaki's idea in practice is verifying whether or not a prior function is superharmonic. In particular, when high-dimensional statistical manifolds are considered, it is technically tricky to test the superharmonicity of prior functions because Laplace–Beltrami operators on the manifolds are non-trivial. Although some superharmonic priors for the autoregressive (AR) models were found not only in the two-dimensional cases [5,7] but also in arbitrary dimensions [6], there is no clue about the Bayesian shrinkage priors of more complicated models such as the autoregressive moving average (ARMA) models, the fractionally integrated ARMA (ARFIMA) models, and any arbitrary signal filters. Additionally, generic algorithms for systematically obtaining the information shrinkage priors are not known yet.

The connection between Kähler manifolds and information geometry has been reported [4,9–12] and the mathematical correspondence between a Kähler manifold and the information geometry of a linear system is recently revealed. It is found that the information geometry of a signal filter with a finite complex cepstrum norm is a Kähler manifold [7]. In particular, the Hermitian condition on the Kählerian information manifolds is clearly seen under conditions on the transfer function of the linear system. Moreover, many practical aspects of introducing Kähler manifolds to information geometry for signal processing were also reported in the same literature [7]. One of the benefits in the Kählerian information geometry is that the simpler form of the Laplace–Beltrami operator on the Kähler manifold is beneficial to finding the Komaki priors.

In this paper, we construct Komaki-style shrinkage priors for Kählerian signal filters. By introducing an algorithm which is based on the characteristics of Kähler manifolds, the Bayesian predictive priors outperforming the Jeffreys prior can be obtained in a more efficient and more robust way. Several prior ansätze are also suggested. Among the ansätze, the geometric shrinkage priors related to Kähler potential are intrinsic priors on the information manifold because the geometry is given by the Kähler potential. We also provide the geometric priors for the ARFIMA models where the Komaki priors have not been reported. The structure of this paper is as follows. In next section, theoretical backgrounds of Kählerian information geometry and superharmonic priors are introduced. In Section 3, an algorithm and ansätze for the geometric shrinkage priors are suggested. The implication of the algorithm to the ARFIMA models is given in Section 4. We conclude the paper in the last section.

## 2. Theoretical Backgrounds

### *2.1. Kählerian Filters*

A linear filter with $n$-dimensional complex parameters $\boldsymbol{\xi}$ is characterized by a transfer function $h(w;\boldsymbol{\xi})$ in the frequency domain $w$ with

$$y(w) = h(w;\boldsymbol{\xi})x(w)$$

where $y$ and $x$ are complex output and input signals, respectively. A spectral density function $S(w;\boldsymbol{\xi})$ is defined as the absolute square of the transfer function

$$S(w;\boldsymbol{\xi}) = |h(w;\boldsymbol{\xi})|^2$$

and it is a real-valued measurable quantity.

In information geometry, it is well-known by Amari and Nagaoka [1] that the geometry of a linear system is determined by the spectral density function $S(w;\boldsymbol{\xi})$ under the stability condition, minimum phase, and

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} |\log S(w;\boldsymbol{\xi})|^2 dw < \infty.$$

The last condition is also known as the finite unweighted norm of the power cepstrum of a filter [13,14]. For a linear system with the spectral density function satisfying the above conditions, the metric tensor of the information geometry is given by

$$g_{\mu\nu}(\boldsymbol{\xi}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} (\partial_\mu \log S)(\partial_\nu \log S) dw$$

where the partial derivatives are taken with respect to the model parameters $\boldsymbol{\xi}$.

The metric tensor can be expressed in a complexified coordinate system and the Z-transformed transfer function. With the Z-transformation, the holomorphic transfer function can be written in the form of series expansion of $z$

$$h(z;\boldsymbol{\xi}) = \sum_{r=0}^{\infty} h_r(\boldsymbol{\xi}) z^{-r} \tag{1}$$

where $h_r$ is an impulse response function. The Z-transformed power spectrum is also defined in the similar way. In this case, the conditions on the transfer function for constructing information geometry are identical to the spectral density function representation except for

$$\frac{1}{2\pi i}\oint_{|z|=1} |\log h(z;\boldsymbol{\xi})|^2 \frac{dz}{z} < \infty$$

and it is a necessary condition for the finite power cepstrum norm. The condition indicates that the Hardy norm of the logarithmic transfer function, also known as the unweighted complex cepstrum norm [14,15], is finite. The metric tensor of the geometry is given by the transfer function,

$$g_{ij}(\boldsymbol{\xi}) = \frac{1}{2\pi i}\oint_{|z|=1} \partial_i \log h(z;\boldsymbol{\xi}) \partial_j \log h(z;\boldsymbol{\xi}) \frac{dz}{z} \tag{2}$$

$$g_{i\bar{j}}(\boldsymbol{\xi}) = \frac{1}{2\pi i}\oint_{|z|=1} \partial_i \log h(z;\boldsymbol{\xi}) \partial_{\bar{j}} \log \bar{h}(\bar{z};\bar{\boldsymbol{\xi}}) \frac{dz}{z} \tag{3}$$

where $i, j$ run from 1 to $n$ and $g_{\bar{i}\bar{j}}, g_{\bar{i}j}$ are the complex conjugates of $g_{ij}$ and $g_{i\bar{j}}$, respectively.

After plugging the Z-transformed transfer function, Equation (1), into the metric tensor expressions, Equations (2) and (3), the metric tensor is expressed with the series expansion coefficients in $z$ of the logarithmic transfer function by

$$g_{ij} = \partial_i \eta_0 \partial_j \eta_0$$

$$g_{i\bar{j}} = \partial_i \eta_0 \partial_{\bar{j}} \bar{\eta}_0 + \sum_{r=1}^{\infty} \partial_i \eta_r \partial_{\bar{j}} \bar{\eta}_r$$

where $\eta_r$ is the coefficient of $z^{-r}$ in the series expansion of the logarithmic transfer function, also known as a complex cepstrum coefficient [15]. It is obvious that $\eta_0 = \log h_0$.

Recently, it is found by Choi and Mullhaupt [7] that the information geometry of a linear system with a finite Hardy norm of a logarithmic transfer function (or the complex cepstrum norm) is the Kähler manifold that is the Hermitian manifold with the closed Kähler two-form: $g_{ij} = g_{\bar{i}\bar{j}} = 0$ for the Hermitian manifold and $\partial_i g_{j\bar{k}} = \partial_j g_{i\bar{k}}$, $\partial_{\bar{i}} g_{k\bar{j}} = \partial_{\bar{j}} g_{k\bar{i}}$ for the closed Kähler two-form. Additionally, the Hermitian structure can be explicitly seen in the metric tensor if and only if the impulse response function with the highest degree in $z$, *i.e.*, $h_0$ in the unilateral transfer function case, is a constant in model parameters $\boldsymbol{\xi}$. In this paper, for simplicity, we only consider unilateral transfer functions with non-zero $h_0$ and the Kähler manifolds with the explicit Hermitian conditions on the metric tensors because complex manifolds are always Hermitian manifolds [16]. In this case, the necessary and sufficient condition for being a Kähler manifold is that $h_0(\boldsymbol{\xi})$ is a constant in $\boldsymbol{\xi}$ [7].

According to Choi and Mullhaupt [7], the benefits of the Kählerian description are the followings. First of all, geometric objects are straightforwardly computed on a Kähler manifold. The non-trivial metric tensor component is simply derived from the following formula

$$g_{i\bar{j}} = \partial_i \partial_{\bar{j}} \mathcal{K} \tag{4}$$

where $\mathcal{K}$ is the Kähler potential of the geometry. The Kähler potential in the information geometry of a linear filter is the square of the Hardy norm (or $H^2$-norm) of the logarithmic transfer function (or the square of the complex cepstrum norm) on the unit disk $\mathbb{D}$

$$\mathcal{K} = \frac{1}{2\pi i}\oint_{|z|=1} |\log h(z;\boldsymbol{\xi})|^2 \frac{dz}{z} = ||\log h(z;\boldsymbol{\xi})||^2_{H^2} \tag{5}$$

and the details of the derivation are given in the literature [7]. The non-trivial components of the Levi–Civita connection are expressed as

$$\Gamma_{ij,\bar{k}} = \partial_i g_{j\bar{k}} = \partial_i \partial_j \partial_{\bar{k}} \mathcal{K} \tag{6}$$

and the other connection components are all vanishing. Notice that it is much simpler than the connection components on a non-Kähler manifold given by

$$\Gamma_{ij,k} = \frac{1}{2}(\partial_i g_{jk} + \partial_j g_{ik} - \partial_k g_{ij})$$

and it is obvious that the number of calculation steps is significantly reduced in the Kähler case. The Riemann curvature tensor of the linear system geometry is also represented in the simpler form which is given in Choi and Mullhaupt [7]. The Ricci tensor on the Kähler manifold is obtained as

$$R_{i\bar{j}} = -\partial_i \partial_{\bar{j}} \log \mathcal{G} \tag{7}$$

where $\mathcal{G}$ is the determinant of the metric tensor. It is evident that we can skip the calculation of the Riemann curvature tensor in order to compute the Ricci tensor on a Kähler manifold.

Additionally, the $\alpha$-generalization of the geometric objects is linear in $\alpha$ on Kähler manifolds. Since the Riemann curvature tensor on a Kähler manifold is linear in the $\alpha$-connection which is $\alpha$-linear, the Riemann tensor also exhibits the $\alpha$-linearity which leads to the $\alpha$-linear Ricci tensor and scalar curvature.

In addition to these advantages, any submanifolds of a Kähler manifold are also Kähler manifolds. If the information geometry of a given statistical model is a Kähler manifold, its submodels also have Kähler manifolds as the information geometry and all the properties of the ambient manifold are also equipped with the submanifolds.

Lastly, the Kählerian information geometry is also useful to find superharmonic priors because of the simpler Laplace–Beltrami operators on the manifolds. We will cover the details of the superharmonic priors soon.

### *2.2. Superharmonic Priors*

For further discussions, we need to introduce the superharmonic priors suggested by Komaki [8]. When we want to find the true probability distribution $p(y|\boldsymbol{\xi})$ based on given samples $x$ of size $N$, one of the best approaches is using Bayesian predictive density $p_\pi(y|x^{(N)})$ with a prior $\pi(\boldsymbol{\xi})$:

$$p_\pi(y|x^{(N)}) = \frac{\int p(y|\boldsymbol{\xi}) p(x^{(N)}|\boldsymbol{\xi}) \pi(\boldsymbol{\xi}) d\boldsymbol{\xi}}{\int p(x^{(N)}|\boldsymbol{\xi}) \pi(\boldsymbol{\xi}) d\boldsymbol{\xi}}.$$

The superharmonic priors $\pi_I$ are derived from the difference between two risk functions with respect to the true probability density, one from the Jeffreys prior and another from the superharmonic prior:

$$\begin{aligned}&\mathbb{E}[D_{KL}(p(y|\boldsymbol{\xi})||p_{\pi_J}(y|x^{(N)}))|\boldsymbol{\xi}] - \mathbb{E}[D_{KL}(p(y|\boldsymbol{\xi})||p_{\pi_I}(y|x^{(N)}))|\boldsymbol{\xi}] \\ &= \frac{1}{2N^2} g^{ij}\partial_i \log\Big(\frac{\pi_I}{\pi_J}\Big)\partial_j \log\Big(\frac{\pi_I}{\pi_J}\Big) - \frac{1}{N^2}\frac{\pi_J}{\pi_I}\Delta\Big(\frac{\pi_I}{\pi_J}\Big) + o(N^{-2})\end{aligned}$$

where $D_{KL}$ is the Kullback–Leibler divergence and $\pi_J$ is the Jeffreys prior which is the volume form of the statistical manifold. Each risk function indicates how far a given Bayesian predictive density is from the true distribution in the Kullback–Leibler divergence in average. Sine better priors are obtained from smaller risk functions, the priors outperforming the Jeffreys prior make the above expression greater than zero. Since the first term on the right-hand side is non-negative, the risk function of the Komaki prior is decreased with respect to the risk function of the Jeffreys prior if a prior function $\psi = \pi_I/\pi_J$ is superharmonic. If a superharmonic prior function $\psi$ can be found, it is possible to do better Bayesian prediction in the viewpoint of information theory. In the same paper, Komaki also pointed out that shrinkage priors are information-theoretically more improved in prediction than the Jeffreys prior if and only if the square root of a prior function is superharmonic.

Since Komaki's paper [8], several superharmonic priors for the AR models have been found [5–7]. The Komaki prior for the AR(2) model in the pole coordinates [5] is given by

$$\psi = 1 - \xi^1\xi^2$$

where $\xi^i$ is a pole of the transfer function. Tanaka [6] generalized the two-dimensional case to superharmonic priors for the AR model in an arbitrary dimension $p$. The shrinkage prior function for the AR($p$) model is in the form of

$$\psi = \prod_{i<j}^{p}(1 - \xi^i\xi^j)$$

where $\xi^i$ is a pole of the AR transfer function.

As mentioned before, one of the advantages in the Kählerian description is that finding the Komaki prior functions becomes more efficient than those in non-Kähler description because the Laplace–Beltrami operators on Kähler manifolds are in the simpler forms. For a differentiable function $\psi$, the Laplace–Beltrami operator in the Kähler geometry is represented with

$$\Delta\psi = 2g^{i\bar{j}}\partial_i\partial_{\bar{j}}\psi.$$

Meanwhile, the Laplace–Beltrami operator on a non-Kähler manifold is expressed as

$$\begin{aligned}\Delta\psi &= \frac{1}{\sqrt{\mathcal{G}}}\partial_i\Big(\sqrt{\mathcal{G}}g^{ij}\partial_j\psi\Big) \\ &= g^{ij}\partial_i\partial_j\psi + \frac{1}{2}g^{ij}\partial_i \log\mathcal{G}\partial_j\psi + \partial_i g^{ij}\partial_j\psi\end{aligned}$$

where $\mathcal{G}$ is the determinant of the metric tensor. It is obvious that additional calculations for the latter two terms in the right-hand side are indispensable in the non-Kähler cases.

With the computational benefits on the Kählerian information manifolds, the superharmonic prior function for the Kähler-AR(2) model [7] is found

$$\psi = (1-|\xi^1|^2)(1-\xi^1\bar{\xi}^2)(1-\xi^2\bar{\xi}^1)(1-|\xi^2|^2)$$

where $\xi^i$ is the $i$-th pole of the transfer function and $\bar{\xi}^i$ is the complex conjugate of $\xi^i$. However, its generalization to any arbitrary dimensions has been unknown. Moreover, the Komaki priors for the ARMA models and the ARFIMA models are not reported yet.

## 3. Geometric Shrinkage Priors

As shown in the previous section, Kähler manifolds in information geometry are useful in order to obtain the superharmonic priors. In this section, we introduce an algorithm to find the geometric shrinkage priors by using the properties of Kähler geometry. Moreover, several ansätze for the priors are suggested.

For further discussions, let us set $\tau = u^* - \kappa(\boldsymbol{\xi}, \bar{\boldsymbol{\xi}})$ where $u^*$ is a constant in $\boldsymbol{\xi} = (\xi^1, \xi^2, \cdots, \xi^n)$ and its complex conjugate $\bar{\boldsymbol{\xi}}$. The following lemma is worthwhile when the algorithm for the prior functions is constructed.

**Lemma 1.** *On a Kähler manifold, a function $\psi(\boldsymbol{\xi}, \bar{\boldsymbol{\xi}})$ is superharmonic if $\psi(\boldsymbol{\xi}, \bar{\boldsymbol{\xi}})$ is in the form of $\psi(\boldsymbol{\xi}, \bar{\boldsymbol{\xi}}) = \Psi(u^* - \kappa(\boldsymbol{\xi}, \bar{\boldsymbol{\xi}}))$ such that $\kappa$ is subharmonic (or harmonic) and $\Psi'(\tau) > 0, \Psi''(\tau) \leq 0$ (or $\Psi'(\tau) > 0, \Psi''(\tau) < 0$).*

**Proof.** The Laplace–Beltrami operator on $\psi$ is given by

$$\begin{aligned}\Delta\psi &= 2g^{i\bar{j}}\partial_i\partial_{\bar{j}}\psi = 2g^{i\bar{j}}\partial_i\Big(\big(-\partial_{\bar{j}}\kappa\big)\Psi'\Big) \\ &= 2\Psi'' g^{i\bar{j}}\partial_i\kappa\partial_{\bar{j}}\kappa - 2\Psi' g^{i\bar{j}}\partial_i\partial_{\bar{j}}\kappa \\ &= 2\Psi''||\partial\kappa||_g^2 - \Psi'\Delta\kappa\end{aligned}$$

where the derivatives on $\Psi$ are taken with respect to $\tau$. It is obvious that if $\kappa$ is subharmonic (or harmonic) and if $\Psi'(\tau) > 0, \Psi''(\tau) \leq 0$ (or $\Psi'(\tau) > 0, \Psi''(\tau) < 0$), then the right-hand side is negative, *i.e.*, $\psi$ is a superharmonic function. □

According to Lemma 1, superharmonic functions are easily obtained from subharmonic or harmonic functions by simply plugging the (sub-)harmonic functions as $\kappa$ into Lemma 1.

By considering that a prior function should be positive, it is able to utilize Lemma 1 for obtaining the superharmonic prior functions. Let us confine the function $\psi$ in Lemma 1 to be positive.

**Theorem 1.** *On a Kähler manifold, a positive function $\psi = \Psi(u^* - \kappa)$ is a superharmonic prior function if $\kappa$ is subharmonic (or harmonic) and $\Psi'(\tau) > 0$, $\Psi''(\tau) \leq 0$ (or $\Psi'(\tau) > 0, \Psi''(\tau) < 0$).*

**Proof.** Since this is a special case of Lemma 1, the proof is obvious. □

Although any (sub-)harmonic function $\kappa$ can be used for constructing superharmonic priors, restriction on $\kappa$ makes finding the ansätze of the geometric priors easier. From now on, upper-bounded functions are only our concerns. Additionally, we assume that $\kappa$ and $u^*$ are real. With these assumptions, it is possible to set $u^*$ as a constant greater than the upper bound of $\kappa$ in order for $\tau$ to be positive.

Ansätze for $\Psi$ can be found in the following example.

**Example 1.** *Given subharmonic (or harmonic) $\kappa$ and positive $\tau$,* i.e.*, upper-bounded $\kappa$, the following functions are candidates for $\Psi$*

$$\Psi_1(\tau) = \tau^a$$
$$\Psi_2(\tau) = \log(1+\tau^a)$$

*where $0 < a \leq 1$ (or $0 < a < 1$).*

**Proof.** We only cover a subharmonic case for $\kappa$ here and it is also straightforward for the harmonic case. First of all, $\Psi_1$ and $\Psi_2$ are all positive. For $\Psi_1$, it is easy to verify the followings:

$$\Psi_1'(\tau) = a\tau^{a-1} > 0$$
$$\Psi_1''(\tau) = a(a-1)\tau^{a-2} \leq 0$$

for $0 < a \leq 1$. The similar calculation is repeated for $\Psi_2$:

$$\Psi_2'(\tau) = \frac{a\tau^{a-1}}{(1+\tau^a)} > 0$$
$$\Psi_2''(\tau) = \frac{a\tau^{a-2}(a-(1+\tau^a))}{(1+\tau^a)^2} \leq 0$$

for $0 < a \leq 1$.

Both functions $\Psi_1$ and $\Psi_2$ satisfy the conditions for $\Psi$ in Lemma 1. □

It is also possible to find ansätze for upper-bounded subharmonic $\kappa$. The following functions are candidates for upper-bounded and subharmonic $\kappa$.

**Example 2.** *For positive real numbers $a_r$ and $b_i$, the following subharmonic functions are candidates for $\kappa$ in the cases that those are upper-bounded:*

$$\kappa_1 = \mathcal{K}$$
$$\kappa_2 = \sum_{r=0}^{\infty} a_r |h_r(\boldsymbol{\xi})|^2$$
$$\kappa_3 = \sum_{i=1}^{n} b_i |\xi^i|^2.$$

**Proof.** Let us assume that the ansätze are upper-bounded in given domains. For $\kappa_1$, it is easy to show that the Kähler potential $\mathcal{K}$ is subharmonic:

$$\begin{aligned}\Delta\kappa_1 = \Delta\mathcal{K} &= 2g^{i\bar{j}}\partial_i\partial_{\bar{j}}\mathcal{K} \\ &= 2g^{i\bar{j}}g_{i\bar{j}} = 2n > 0.\end{aligned}$$

The proof for subharmonicity of $\kappa_2$ is as follows:

$$\begin{aligned}\Delta\kappa_2 = \Delta\left(\sum_{r=0}^{\infty} a_r|h_r(\boldsymbol{\xi})|^2\right) &= 2g^{i\bar{j}}\partial_i\partial_{\bar{j}}\left(\sum_{r=0}^{\infty} a_r|h_r(\boldsymbol{\xi})|^2\right) \\ = \sum_{r=0}^{\infty} 2a_r g^{i\bar{j}}\partial_i h_r\partial_{\bar{j}}\bar{h}_r &= \sum_{r=0}^{\infty} 2a_r||\partial h_r||_g^2 > 0.\end{aligned}$$

The subharmonicity of $\kappa_3$ is tested by

$$\Delta\kappa_3 = \Delta\left(\sum_{i=1}^{n} b_i|\xi^i|^2\right) = 2g^{i\bar{j}}\partial_i\partial_{\bar{j}}\left(\sum_{i=1}^{n} b_i|\xi^i|^2\right) = \sum_{i=1}^{n} 2b_i g^{i\bar{i}} > 0.$$

If the upper-boundedness is satisfied, the above subharmonic functions are ansätze for $\kappa$. □

Superharmonic prior functions on the Kähler manifolds are efficiently constructed from the following algorithm which exploits Theorem 1 and the ansätze for $\Psi$ and $\kappa$. When we find positive and superharmonic functions, it is automatically the Komaki-style prior functions as usual. If positive, upper-bounded, and (sub-)harmonic functions are found, those functions are plugged into Theorem 1 in order to obtain superharmonic prior functions. Multiplying the Jeffreys prior by the superharmonic prior functions, we finally acquire the geometric shrinkage priors. Additionally, since the ansätze are already given, there is no extra cost to find the Komaki prior functions except for verifying whether or not the information geometry is a Kähler manifold. Comparing with the literature on the Komaki priors of the time series models [5–7], obtaining the geometric priors on the Kähler manifolds becomes more efficient and more robust.

## 4. Example: ARFIMA Models

The ARFIMA model is the generalization of the ARMA model with a fractional differencing parameter in order to model the long memory process. The transfer function of the ARFIMA$(p, d, q)$ model with parameters $\boldsymbol{\xi} = (\xi^{-1}, \xi^0, \xi^1, \cdots, \xi^{p+q}) = (\sigma, d, \lambda_1, \cdots, \lambda_p, \mu_1, \cdots, \mu_q)$ is given by

$$h(z;\boldsymbol{\xi}) = \frac{\sigma^2}{2\pi}\frac{(1-\mu_1 z^{-1})(1-\mu_2 z^{-1})\cdots(1-\mu_q z^{-1})}{(1-\lambda_1 z^{-1})(1-\lambda_2 z^{-1})\cdots(1-\lambda_p z^{-1})}(1-z^{-1})^d$$

where $d$ is the differencing parameter and $\mu_i, \lambda_i, \sigma$ are a pole, a root, and a gain in the ARMA model, respectively. It is noteworthy that the transfer function of the ARFIMA model is decomposed into the ARMA model part and the fractionally integration part. Additionally, every poles and roots of the linear system are located inside the unit disk, *i.e.*, $|\lambda_i| < 1$ for $i = 1, \cdots, p$ and $|\mu_i| < 1$ for $i = 1, \cdots, q$.

Similar to the ARMA case [7], the full geometry of the ARFIMA model is a Kähler manifold and the submanifold of a constant gain $\sigma$ is also Kähler geometry. This submanifold also exhibits the explicit Hermitian condition on the metric tensor. It is easy to cross-check the Hermitian structure by fixing $h_0 = 1$ up to the gain of the signal filter. We will work on this submanifold.

Since the information geometry of the ARFIMA model is a Kähler manifold, the Kähler potential of the ARFIMA geometry is obtained from the square of the Hardy norm of the logarithmic transfer function (or the square of the complex cepstrum norm), Equation (5), represented with

$$\mathcal{K} = \sum_{r=1}^{\infty} \left| \frac{d + (\mu_1^r + \cdots + \mu_q^r) - (\lambda_1^r + \cdots + \lambda_p^r)}{r} \right|^2. \tag{8}$$

It is obvious that the Kähler potential for the ARFIMA model, Equation (8), is reducible to the Kähler potential of the ARMA geometry by setting $d = 0$. It is easy to verify that the Kähler potential of the ARFIMA geometry is upper-bounded by $(d + p + q)^2 \frac{\pi^2}{6}$.

By using Equation (4), the metric tensor of the Kähler geometry is simply derived from the Kähler potential. The metric tensor of the Kähler-ARFIMA geometry is given by

$$g_{i\bar{j}} = \begin{pmatrix} \frac{\pi^2}{6} & \frac{1}{\bar{\lambda}_j} \log\left(1 - \bar{\lambda}_j\right) & -\frac{1}{\bar{\mu}_j} \log\left(1 - \bar{\mu}_j\right) \\ \frac{1}{\lambda_i} \log\left(1 - \lambda_i\right) & \frac{1}{1 - \lambda_i \bar{\lambda}_j} & -\frac{1}{1 - \lambda_i \bar{\mu}_j} \\ -\frac{1}{\mu_i} \log\left(1 - \mu_i\right) & -\frac{1}{1 - \mu_i \bar{\lambda}_j} & \frac{1}{1 - \mu_i \bar{\mu}_j} \end{pmatrix}$$

and it is easy to show that the metric tensor contains the pure ARMA metric. The metric tensor is also in the similar form to the ARFIMA geometry in non-complexified coordinates [3]. The metric tensor indicates that the ARMA geometry is embedded in the ARFIMA geometry and corresponds to the submanifold of the ARFIMA manifold. The ARMA part of the metric tensor is the same metric with the Kähler-ARMA geometry in Choi and Mullhaupt [7]. In addition to that, we can cross-check the fact that the ARMA geometry is also a Kähler manifold based on a property of a Kähler manifold that a submanifold of the Kähler geometry is Kähler.

Other geometric objects can be derived from the metric tensor. For example, the non-trivial components of the 0-connection are given by Equation (6). It is noteworthy that any connection components with the $d$-coordinate in the first two indices of the connection are trivially zero and the others might not be vanishing. Similar to the 0-connection, the Ricci tensor components along the fractionally integrated direction are also zero because there is no dependence on $d$ in the metric tensor. Considering the Schur complement, the non-vanishing Ricci tensor components are decomposed into the Ricci tensor from the pure ARMA part and the term from the mixing between the ARMA part and the fractionally integrated (FI) part:

$$R_{i\bar{j}} = R_{i\bar{j}}^{ARMA} + R_{i\bar{j}}^{ARMA-FI}$$

where $i$ and $j$ are not along the $d$-coordinate.

It is the time to be back to the geometric shrinkage priors. Since the Kähler potential of a given ARFIMA model is upper-bounded by a constant $u^* = (d + p + q)^2 \frac{\pi^2}{6}$, the intrinsic priors on the Kähler manifold can be found as it is proven in the previous section. By using the algorithm and the

ansätze related to the Kähler potential, some geometric shrinkage prior functions for the ARFIMA model are constructed as

$$\psi_1 = (u^* - \mathcal{K})^a$$
$$\psi_2 = \log\left(1 + (u^* - \mathcal{K})^a\right)$$

where $0 < a \leq 1$. It is also noteworthy that when $d = 0$ in the Kähler potential, superharmonic priors of the ARMA (or AR/MA) models are obtained and finding the priors becomes much simpler than the literature on the Komaki priors of the AR models [5–7]. Similarly, $\kappa_2$ and $\kappa_3$ are also utilized for the superharmonic prior function ansätze in the ARFIMA models because the both functions are upper-bounded on the ARFIMA manifold. Moreover, if we set $d = 0$ for $\kappa_2$ or $b_0 = 0$ for $\kappa_3$, the ansätze for the ARFIMA models are reducible to the Komaki priors of the ARMA models.

## 5. Conclusion

In this paper, we build up an algorithm and ansätze for the geometric shrinkage priors of Kählerian signal filters. By using the properties of Kähler manifolds, an algorithm to find the Komaki priors is constructed and ansätze for the prior functions are suggested. Additionally, some ansätze associated with the Kähler potential are geometrically intrinsic to Kählerian information manifolds because the geometry is derived from the Kähler potential which is the square of the complex cepstrum norm of a linear system.

Comparing with the literature on the Komaki priors of the time series models, verification of the geometric priors is much easier on the Kähler manifold and it is also possible to acquire the geometric shrinkage priors for highly complicated models in the more efficient and robust way. For example, Bayesian predictive priors for the ARFIMA model are obtained from the algorithm and ansätze for the prior functions. The shrinkage priors of the ARMA cases are simply found from the geometric shrinkage priors of the ARFIMA models by using the property of submanifolds in the Kähler geometry.

## Acknowledgments

We are thankful to Michael Tiano for useful discussions.

## Author Contributions

Both authors contributed equally to the main idea. The research was conducted out by both authors. Jaehyung Choi wrote the paper. Both authors have read and approved the final manuscript.

## Conflicts of Interest

The authors declare no conflict of interest.

## References


1. Amari, S.; Nagaoka, H. *Methods of information geometry*; Oxford University Press: Oxford, UK, 2000.
2. Ravishanker, N.; Melnick, E. L.; Tsai, C. Differential geometry of ARMA models. *J. Time Ser. Anal.* **1990**, *11*, 259–274.
3. Ravishanker, N. Differential geometry of ARFIMA processes. *Commun. Stat. Theory Methods* **2001**, *30*, 1889–1902.
4. Barbaresco, F. Information intrinsic geometric flows. *AIP Conf. Proc.* **2006**, *872*, 211–218.
5. Tanaka, F.; Komaki, F. A superharmonic prior for the autoregressive process of the second order. *J. Time Ser. Anal.* **2008**, *29*, 444–452.
6. Tanaka, F. *Superharmonic priors for autoregressive models*; Mathematical Engineering Technical Reports; University of Tokyo: Tokyo, Japan, 2009.
7. Choi, J.; Mullhaupt, A. P. Kählerian information geometry for signal processing. arXiv:1404.2006.
8. Komaki, F. Shrinkage priors for Bayesian prediction. *Ann. Stat.* **2006**, *34*, 808–819.
9. Barndorff-Nielsen, O. E.; Jupp, P. E. Statistics, yokes and symplectic geometry. *Annales de la faculté des sciences de Toulouse 6 série* **1997**, *6*, 389–427.
10. Barbaresco, F. Information geometry of covariance matrix: Cartan-Siegel homogeneous bounded domains, Mostow/Berger fibration and Fréchet median. In *Matrix Information Geometry*; Bhatia, R., Nielsen, F., Eds.; Springer: Berlin and Heidelberg, Germany, 2012; pp. 199–256.
11. Zhang, J.; Li, F. Symplectic and Kähler structures on statistical manifolds induced from divergence functions. *Geom. Sci. Inf.* **2013**, *8085*, 595–603.
12. Barbaresco, F. Koszul information geometry and Souriau geometric temperature/capacity of Lie group thermodynamics. *Entropy* **2014**, *16*, 4521–4565.
13. Bogert, B.; Healy, M.; Tukey, J. The quefrency alanysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. In Proceedings of the Symposium on Time Series Analysis, Brown University, Providence, RI, USA, 11–14 June 1963; pp. 209–243.
14. Martin, R.J. A metric for ARMA processes. *IEEE Trans. Signal Process.* **2000**, *48*, 1164–1170.
15. Oppenheim, A. V. Superposition in a class of nonlinear systems. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 1965.
16. Nakahara, M. *Geometry, Topology and Physics*; Institute of Physics Publishing: Bristol, UK and Philadelphia, PA, USA, 2003.

# Kählerian Information Geometry for Signal Processing

**Jaehyung Choi and Andrew P. Mullhaupt**

**Abstract:** We prove the correspondence between the information geometry of a signal filter and a Kähler manifold. The information geometry of a minimum-phase linear system with a finite complex cepstrum norm is a Kähler manifold. The square of the complex cepstrum norm of the signal filter corresponds to the Kähler potential. The Hermitian structure of the Kähler manifold is explicitly emergent if and only if the impulse response function of the highest degree in $z$ is constant in model parameters. The Kählerian information geometry takes advantage of more efficient calculation steps for the metric tensor and the Ricci tensor. Moreover, $\alpha$-generalization on the geometric tensors is linear in $\alpha$. It is also robust to find Bayesian predictive priors, such as superharmonic priors, because Laplace–Beltrami operators on Kähler manifolds are in much simpler forms than those of the non-Kähler manifolds. Several time series models are studied in the Kählerian information geometry.

Reprinted from *Entropy*. Cite as: Choi, J.; Mullhaupt, A.P. Kählerian Information Geometry for Signal Processing. *Entropy* **2015**, *17*, 1581–1605.

## 1. Introduction

Since the introduction of Riemannian geometry to statistics [1,2], information geometry has been developed along various directions. The statistical curvature as the differential-geometric analogue of information loss and sufficiency was proposed by Efron [3]. The $\alpha$-duality of information geometry was found by Amari [4]. Not being limited to statistical inference, information geometry has become popular in many different fields, such as information-theoretic generalization of the expectation-maximization algorithm [5], hidden Markov models [6], interest rate modeling [7], phase transition [8,9] and string theory [10]. More applications can be found in the literature [11] and the references therein.

In particular, time series analysis and signal processing are well-known applications of information geometry. Ravishanker *et al*. [12] found the information geometry of autoregressive moving average (ARMA) models in the coordinate system of poles and zeros. It was also extended to fractionally-integrated ARMA (ARFIMA) models [13]. The information geometry of autoregressive (AR) models in the reflection coefficient coordinates was also reported by Barbaresco [14]. In the information-theoretic framework, Bayesian predictive priors outperforming the Jeffreys prior were derived for the AR models by Komaki [15].

Kähler manifolds are interesting topics in differential geometry. On a Kähler manifold, the metric tensor and the Levi–Civita connection are straightforwardly calculated from the Kähler potential, and the Ricci tensor is obtained from the determinant of the metric tensor. Moreover, its holonomy group is related to the unitary group. Because of these properties, many implications of Kähler manifolds are found in mathematics and theoretical physics. In addition to these fields, information geometry is one of those fields where the Kähler manifolds are intriguing. After the symplectic structure in

information geometry and its connection to statistics were discovered [16], Barbaresco [14] notably introduced Kähler manifolds to information geometry for time series models and also generalized the differential-geometric approach with mathematical structures, such as Koszul geometry [17,18]. Additionally, Zhang and Li [19] found symplectic and Kähler structures in divergence functions.

In this paper, we prove that the information geometry of a signal filter with a finite complex cepstrum norm is a Kähler manifold. The Kähler potential of the geometry is the square of the Hardy norm of the logarithmic transfer function of a linear system. The Hermitian structure of the manifold is explicitly seen in the metric tensor under certain conditions on the transfer functions of linear models and filters. The calculation of geometric objects and the search for Bayesian predictive priors are simplified by exploiting the properties of Kähler geometry. Additionally, $\alpha$-correction terms on the geometric objects exhibit $\alpha$-linearity. This paper is structured as follows. In the next section, we shortly review information geometry for signal processing and derive basic lemmas in terms of the spectral density function and transfer function. In Section 3, main theorems for Kählerian information manifolds are proven and the consequences of the theorems are provided. The implications of Kähler geometry to time series models are reported in Section 4. We conclude the paper in Section 5.

## 2. Information Geometry for Signal Processing

### *2.1. Spectral Density Representation in the Frequency Domain*

We model an output signal $y(w)$ as a linear system with a transfer function $h(w;\boldsymbol{\xi})$ of model parameters $\boldsymbol{\xi} = (\xi^1, \xi^2, \cdots, \xi^n)$:

$$y(w) = h(w;\boldsymbol{\xi})x(w)$$

where $x(w)$ is an input signal in frequency domain $w$. Complex inputs, outputs and model parameters are considered in this paper. The properties of a given signal filter are characterized by the transfer function $h(w;\boldsymbol{\xi})$ and the model parameters $\boldsymbol{\xi}$.

In signal processing, one of the most important quantities is the spectral density function. The spectral density function $S(w;\boldsymbol{\xi})$ is defined as the absolute square of the transfer function:

$$S(w;\boldsymbol{\xi}) = |h(w;\boldsymbol{\xi})|^2. \tag{1}$$

The spectral density function describes the way that energy in the frequency domain is distributed by a given signal filter. In terms of signal amplitude, the spectral density function encodes an amplitude response to a monochromatic input $e^{iw}$. For example, the spectral density function of the all-pass filter is constant in the frequency domain, because the filter passes all inputs to outputs up to the phase difference regardless of frequency. The high-pass filters only allow the signals in the high-frequency domain. Meanwhile, the low-pass filters only permit low-frequency inputs. The properties of other well-known filters are also described by their specific spectral density functions.

The spectral density function is also important in information geometry, because the information-geometric objects of the signal processing geometry are derived from the spectral

density function [20,21]. Among the geometric objects, the length and distance concepts are most fundamental in geometry. One of the most important distance measures in information geometry is the $\alpha$-divergence, also known as Chernoff's $\alpha$-divergence, that is the only divergence which is both an $f$-divergence and a Bregman divergence [22]. The $\alpha$-divergence between two spectral density functions $S_1$ and $S_2$ is defined as

$$D^{(\alpha)}(S_1||S_2) = \begin{cases} \frac{1}{2\pi\alpha^2}\int_{-\pi}^{\pi}\left\{\left(\frac{S_2}{S_1}\right)^{\alpha} - 1 - \alpha\log\frac{S_2}{S_1}\right\}dw & (\alpha \neq 0) \\ \frac{1}{4\pi}\int_{-\pi}^{\pi}\left(\log S_2 - \log S_1\right)^2 dw & (\alpha = 0) \end{cases}$$

and the divergence conventionally measures the distance from $S_1$ to $S_2$. The $\alpha$-divergence, except for $\alpha = 0$, is a pseudo-distance, because it is not symmetric under exchange between $S_1$ and $S_2$. In spite of the asymmetry, the $\alpha$-divergence is frequently used for measuring differences between two linear models or two filters. Some $\alpha$-divergences are more popular than others, because those divergences have been already known in information theory and statistics. For example, the $(-1)$-divergence is the Kullback–Leibler divergence. The 0-divergence is well known as the square of the Hellinger distance in statistics. The Hellinger distance is locally asymptotically equivalent to the information distance and globally tightly bounded by the information distance [23].

The metric tensor of a statistical manifold, also known as the Fisher information matrix, is derived from the $\alpha$-divergence. In order to define the information geometry of a linear system, the conditions on a signal filter are found in Amari and Nagaoka [21]: stability, minimum phase and

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} |\log S(w;\boldsymbol{\xi})|^2 dw < \infty$$

which imposes that the unweighted power cepstrum norm [24,25] is finite. According to the literature [20,21], the metric tensor of the linear system geometry is given by

$$g_{\mu\nu}(\boldsymbol{\xi}) = \frac{1}{2\pi}\int_{-\pi}^{\pi}(\partial_\mu \log S)(\partial_\nu \log S)dw \tag{2}$$

where the partial derivatives are taken with respect to the model parameters $\boldsymbol{\xi}$, *i.e.*, $\partial_\mu = \frac{\partial}{\partial\xi^\mu}$. Since the dimension of the manifold is $n$, the metric tensor is an $n \times n$ matrix.

Other information-geometric objects are also determined by the spectral density function. The $\alpha$-connection, which encodes the change of a vector being parallel-transported along a curve, is expressed with

$$\Gamma^{(\alpha)}_{\mu\nu,\rho}(\boldsymbol{\xi}) = \frac{1}{2\pi}\int_{-\pi}^{\pi}(\partial_\mu\partial_\nu \log S - \alpha(\partial_\mu \log S)(\partial_\nu \log S))(\partial_\rho \log S)dw \tag{3}$$

where $\alpha$ is a real number. Notice that the $\alpha$-connection is not a tensor. The $\alpha$-connection is related to the Levi–Civita connection, $\Gamma_{\mu\nu,\rho}(\boldsymbol{\xi})$, also known as the metric connection. The relation is given by the following equations:

$$\Gamma^{(\alpha)}_{\mu\nu,\rho}(\boldsymbol{\xi}) = \Gamma_{\mu\nu,\rho}(\boldsymbol{\xi}) - \frac{\alpha}{2}T_{\mu\nu,\rho}(\boldsymbol{\xi}) \tag{4}$$

$$T_{\mu\nu,\rho}(\boldsymbol{\xi}) = \frac{1}{\pi}\int_{-\pi}^{\pi}(\partial_\mu \log S)(\partial_\nu \log S)(\partial_\rho \log S)dw \tag{5}$$

where the tensor $T$ is symmetric under the exchange of the indices. The Levi–Civita connection corresponds to the $\alpha = 0$ case.

These information-geometric objects have interesting properties with the reciprocality of spectral density functions. The spectral density function of an inverse system is the reciprocal spectral density function of the original system. The geometric properties of the inverse system are described by the $\alpha$-dual description. The following lemma shows the correspondence between the reciprocality of the spectral density function and the $\alpha$-duality.

**Lemma 1.** *The information geometry of an inverse system is the $\alpha$-dual geometry to the information geometry of the original system.*

**Proof.** The metric tensor is invariant under the reciprocality of spectral density functions, *i.e.*, plugging $S^{-1}$ into Equation (2) provides the identical metric tensor.

Meanwhile, the $\alpha$-connection is not invariant under the reciprocality and exhibits a more interesting property. The $\alpha$-connection from the reciprocal spectral density function is given by

$$\begin{aligned}\Gamma^{(\alpha)}_{\mu\nu,\rho}(S^{-1};\boldsymbol{\xi}) &= \frac{1}{2\pi}\int_{-\pi}^{\pi}(\partial_\mu\partial_\nu \log S + \alpha(\partial_\mu \log S)(\partial_\nu \log S))(\partial_\rho \log S)dw \\ &= \Gamma^{(-\alpha)}_{\mu\nu,\rho}(S;\boldsymbol{\xi})\end{aligned}$$

and the above equation shows that the $\alpha$-connection induced by the reciprocal spectral density function corresponds to the $(-\alpha)$-connection of the original geometry.

Similar to the $\alpha$-connection, the $\alpha$-divergence is equipped with the same property. The $\alpha$-divergence between two reciprocal spectral density functions is straightforwardly found from the definition of the $\alpha$-divergence, and it is represented by the $(-\alpha)$-divergence between the two spectral density functions:

$$D^{(\alpha)}(S_1^{-1}||S_2^{-1}) = D^{(-\alpha)}(S_1||S_2).$$

Using the inverse systems, we can construct the $\alpha$-dual description of signal processing models in information geometry. The multiplicative inverse of a spectral density function corresponds to the $\alpha$-duality of the geometry. □

Lemma 1 indicates that given a linear system geometry, there is no way to discern whether the metric tensor is derived from the filters with $S$ or $S^{-1}$. Additionally, the model $S^{-1}$ is $(-\alpha)$-flat if and only if $S$ is $\alpha$-flat. The 0-connection is self-dual under the reciprocality. A consequence of Lemma 1 is the following multiplication rule:

$$\begin{aligned}D^{(\alpha)}(S_1||S_2^{-1}) &= \frac{1}{2\pi\alpha^2}\int_{-\pi}^{\pi}\left\{(S_1S_2)^{-\alpha} - 1 + \alpha\log(S_1S_2)\right\}dw \\ &= D^{(-\alpha)}(S_0||S_1S_2) = D^{(\alpha)}(S_1S_2||S_0)\end{aligned}$$

where $S_0$ is the unit spectral density function of the all-pass filter. Plugging $S_1 = S_0$ and $S_2 = S$, we have $D^{(0)}(S_0||S^{-1}) = D^{(0)}(S_0||S) = D^{(0)}(S||S_0)$.

We observe that the bilateral transfer functions $\log|h(e^{iw};\boldsymbol{\xi})|^2 \in L^2(\mathbb{T})$ are isomorphically embedded in the space $\mathbb{R} \oplus zH^2(\mathbb{D})$.

**Lemma 2.** *Let* $\log\left|h(e^{iw};\boldsymbol{\xi})\right|^2 \in L^2(\mathbb{T})$. *Then, there is an analytic function* $f \in \exp\left(H^2(\mathbb{D})\right)$, *such that*

$$\left|h(e^{iw};\boldsymbol{\xi})\right|^2 = \left|f(e^{iw};\boldsymbol{\xi})\right|^2$$

*and*

$$\left\|\log\left|h(e^{iw};\boldsymbol{\xi})\right|^2 - \log\left|h(1;\boldsymbol{\xi})\right|^2\right\|_{L^2(\mathbb{T})} = \left\|\log\left|f(e^{iw};\boldsymbol{\xi})\right|^2 - \log\left|f(1;\boldsymbol{\xi})\right|^2\right\|_{H^2(\mathbb{D})}.$$

*This has the interpretation that the information manifold of* $\log\left|h(e^{iw};\boldsymbol{\xi})\right|^2 \in L^2$ *is isometric to the Hardy–Hilbert space.*

**Proof.** $\log h(e^{iw};\boldsymbol{\xi})$ is represented by the Fourier series:

$$\log\left|h(e^{iw};\boldsymbol{\xi})\right|^2 = \sum_{r=-\infty}^{\infty} a_r e^{irw}$$

and since $\log\left|h(e^{iw};\boldsymbol{\xi})\right|^2$ is real, we have $a_{-r} = \bar{a}_r$, and in particular, $a_0$ is real. We define the conjugate series by the coefficients $\tilde{a}_r$, so that $a_r + i\tilde{a}_r = 0$ for $r < 0$ and $\tilde{a}_r$ for $r > 0$; so that $\tilde{a}\left(e^{i\theta}\right)$ is real valued, we choose $\tilde{a}_0 = 0$. This implies

$$\tilde{a}_r = \begin{cases} -\frac{1}{i}a_r & (r<0) \\ \frac{1}{i}a_r & (r>0) \end{cases}$$

and if $\{a_r\} \in l^p$ for $1 \le p \le \infty$, then $\{\tilde{a}_r\} \in l^p$, in particular,

$$\sum_{r\neq 0} |a_r|^2 = \sum_{r\neq 0} |\tilde{a}_r|^2. \tag{6}$$

The analytic function $f(z) = \exp\left(a_0 + a(z) + i\tilde{a}(z)\right)$ has

$$\log\left|h(e^{iw};\boldsymbol{\xi})\right|^2 = \log\left|f(e^{iw};\boldsymbol{\xi})\right|^2$$

and

$$\|\log f(z;\boldsymbol{\xi}) - \log f(1;\boldsymbol{\xi})\|_{H^2}^2 = \left\|\log\left|h(e^{iw};\boldsymbol{\xi})\right|^2 - \log\left|h(1;\boldsymbol{\xi})\right|^2\right\|_{L^2(\mathbb{T})}^2 < \infty$$

and because $f \in \exp\left(zH^2(\mathbb{D})\right)$, $f$ (and $f^{-1}$) is outer, we may write

$$h(e^{iw};\boldsymbol{\xi}) = u(e^{iw};\boldsymbol{\xi}) f(e^{iw};\boldsymbol{\xi})$$

where $\log u(e^{iw};\boldsymbol{\xi}) \in L^2$ is pure imaginary, that is, $|u(e^{iw};\boldsymbol{\xi})| = 1$. □

This has the interpretation that $h$ has a well-defined outer factor, and the information geometry of $h$ depends only on $h$. In the case that the power series coefficients $a_k(\boldsymbol{\xi})$ are continuous, smooth, analytic, *etc*., then the embedding is likewise smooth.

*2.2. Transfer Function Representation in the $z$ Domain*

By using transfer functions, it is also possible to reproduce all of the previous results with the spectral density function. With Fourier transformation and $Z$-transformation, $z = e^{iw}$, a transfer function $h(z;\boldsymbol{\xi})$ is expressed with a series expansion of $z$,

$$h(z;\boldsymbol{\xi}) = \sum_{r=-\infty}^{\infty} h_r(\boldsymbol{\xi}) z^{-r} \tag{7}$$

where $h_r(\boldsymbol{\xi})$ is an impulse response function. It is a bilateral (or two-sided) transfer function expression, which has both positive and negative degrees in $z$, including the zero-th degree. In the causal response case that $h_r(\boldsymbol{\xi}) = 0$ for all negative $r$, the transfer function is unilateral. In many applications, the main concern is the causality of linear filters, which is represented by unilateral transfer functions. In this paper, we start with bilateral transfer functions as generalization and then will focus on causal filters.

In the complex $z$-domain, all formulae for the information-geometric objects are identical to the expressions in the frequency domain, except for the change of the integral measure:

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} G(e^{iw};\boldsymbol{\xi})dw \to \frac{1}{2\pi i}\oint_{|z|=1} G(z;\boldsymbol{\xi})\frac{dz}{z}$$

for an arbitrary integrand $G$. Since the evaluation of the integration is obtained from the line integral along the unit circle on the complex plane, it is easy to calculate the above integration with the aid of the residue theorem. According to the residue theorem, the poles only inside the unit circle contribute to the value of the integration. If $G(z;\boldsymbol{\xi})$ is analytic on the unit disk, the constant term in $z$ of $G(z;\boldsymbol{\xi})$ is the value of the integration. For more details, see Cima *et al.* [26] and the references therein.

One advantage of using $Z$-transformation is that a transfer function can be understood in the framework of functional analysis. A transfer function defined on the complex plane is expanded by the orthonormal basis $z^{-r}$ for integers $r$ with impulse response functions as the coefficients. In functional analysis, it is possible to define the inner product between two complex functions $F$ and $G$ in the Hilbert space:

$$\langle F, G\rangle = \frac{1}{2\pi i}\oint_{|z|=1} F(z)\overline{G(z)}\frac{dz}{z}.$$

By using this inner product, the condition for the stationarity, $\sum_{r=0}^{\infty}|h_r|^2 < \infty$, is written as the Hardy norm ($H^2$-norm) in complex functional analysis,

$$\|h(z;\boldsymbol{\xi})\|_{H^2}^2 = \langle h(z;\boldsymbol{\xi}), h(z;\boldsymbol{\xi})\rangle = \sum_{r=0}^{\infty}|h_r|^2 < \infty.$$

Since the functional space with a finite Hardy norm is called the Hardy–Hilbert space $H^2$, the unilateral transfer functions satisfying the stationarity condition live on the $H^2$-space. A transfer function of a stationary system is a function in the $L^2$-space if the transfer function is in the bilateral form.

The conditions on the transfer function of a signal filter are also necessary for defining the information geometry of a linear system in terms of the transfer function. Similar to the spectral density representation, the conditions on the linear filters are stability and minimum phase. In addition to these conditions, we also need the following condition on the finite $H^2$-norm of the logarithmic transfer function,

$$\frac{1}{2\pi i}\oint_{|z|=1} |\log h(z;\boldsymbol{\xi})|^2 \frac{dz}{z} < \infty$$

and for the above condition, it is also known that the unweighted complex cepstrum norm [25,27] is finite. From now on, signal filters in this paper are the linear systems satisfying the above norm conditions. This is a necessary condition for a finite power cepstrum norm.

It is natural to complexify the coordinate system as being used in the complex differential geometry. In holomorphic and anti-holomorphic coordinates, the metric tensor of a linear system geometry is represented by

$$g_{\mu\nu} = \frac{1}{2\pi i}\oint_{|z|=1} \partial_\mu\big(\log h(z;\boldsymbol{\xi}) + \log \bar{h}(\bar{z};\bar{\boldsymbol{\xi}})\big)\partial_\nu\big(\log h(z;\boldsymbol{\xi}) + \log \bar{h}(\bar{z};\bar{\boldsymbol{\xi}})\big)\frac{dz}{z}$$

where both $\mu$ and $\nu$ run over all holomorphic and anti-holomorphic coordinates, *i.e.*, $\mu, \nu = 1, 2, \cdots, n, \bar{1}, \bar{2}, \cdots, \bar{n}$.

The components of the metric tensor are categorized into two classes: one with pure indices from holomorphic coordinates and anti-holomorphic coordinates, and another with the mixed indices. The metric tensor components in these categories are given by

$$g_{ij}(\boldsymbol{\xi}) = \frac{1}{2\pi i}\oint_{|z|=1} \partial_i \log h(z;\boldsymbol{\xi})\partial_j \log h(z;\boldsymbol{\xi})\frac{dz}{z} \tag{8}$$

$$g_{i\bar{j}}(\boldsymbol{\xi}) = \frac{1}{2\pi i}\oint_{|z|=1} \partial_i \log h(z;\boldsymbol{\xi})\partial_{\bar{j}} \log \bar{h}(\bar{z};\bar{\boldsymbol{\xi}})\frac{dz}{z} \tag{9}$$

where $g_{\bar{i}\bar{j}} = (g_{ij})^*$ and $g_{\bar{i}j} = (g_{i\bar{j}})^*$, and the indices $i$ and $j$ run from one to $n$. It is also possible to express the $\alpha$-connection and the $\alpha$-divergence in terms of the transfer function by using Equation (1), the relation between the transfer function and the spectral density function.

It is noteworthy that the information geometry of a linear system is invariant under the multiplicative factor of $z$ in the transfer function, because the metric tensor is not changed by the factorization. The invariance is also true for the geometry induced by the spectral density function.

**Lemma 3.** *The information geometry of a signal filter is invariant under the multiplicative factor of $z$.*

**Proof.** Any transfer function can be factored $z^R$ out in the form of

$$h(z;\boldsymbol{\xi}) = z^R \tilde{h}(z;\boldsymbol{\xi})$$

where $R$ is an integer and $\tilde{h}$ is the factored-out transfer function. In the spectral density function representation, the contribution of the factorization is $|z|^{2R}$, and it is a unity in the line integration.

It imposes that the metric tensor, the $\alpha$-connection and the $\alpha$-divergence are independent of the factorization.

When a transfer function is considered, the same conclusion is obtained. Since the contribution from the factorization parts, $\log z^R$, is canceled by the partial derivatives in the metric tensor and the $\alpha$-connection expression, the geometry is invariant under the factorization. It is also easy to show that $\alpha$-divergence is also not changed by the factorization. Another explanation is that the terms of $\partial_i h/h$ in the metric tensor and the $\alpha$-connection are invariant under $z^R$-scaling. □

Based on Lemma 3, it is possible to obtain the unilateral transfer function from a transfer function with a finite upper bound in degrees of $z$. In particular, this factorization invariance of the geometry is useful in the case that the transfer function has a finite number of terms in the non-causal direction of the bilateral transfer function. If the highest degree in $z$ of the transfer function is finite, the transfer function is factored out as

$$\begin{aligned} h(z;\boldsymbol{\xi}) &= z^R(h_{-R} + h_{-(R-1)}z^{-1} + \cdots) \\ &= z^R\tilde{h}(z;\boldsymbol{\xi}) \end{aligned}$$

where $R$ is the maximum degree in $z$ of the transfer function and $\tilde{h}$ is a unilateral transfer function.

A bilateral transfer function can be expressed with the multiplication of a unilateral transfer function $f(z;\boldsymbol{\xi})$ and an analytic function $a(z;\boldsymbol{\xi})$ on the disk:

$$\begin{aligned} h(z;\boldsymbol{\xi}) &= f(z;\boldsymbol{\xi})a(z;\boldsymbol{\xi}) \\ &= (f_0 + f_1 z^{-1} + f_2 z^{-2} + \cdots)(a_0 + a_1 z^1 + a_2 z^2 + \cdots) \end{aligned}$$

where $f_r$ and $a_r$ are functions of $\boldsymbol{\xi}$. For a causal filter, all $a_i$'s are zero, except for $a_0$. This decomposition also includes the case of Lemma 3 by setting $a_i = 0$ for $i < R$ and $a_R = 1$. However, it is natural to take $f_0$ and $a_0$ as non-zero functions of $\boldsymbol{\xi}$. This is because powers of $z$ could be factored out for non-zero coefficient terms with the maximum degree in $f(z;\boldsymbol{\xi})$ and the minimum degree in $a(z;\boldsymbol{\xi})$, and the transfer function is reducible to

$$h(z;\boldsymbol{\xi}) = z^R\tilde{h}(z;\boldsymbol{\xi})$$

where $\tilde{h}(z;\boldsymbol{\xi})$ has non-zero $\tilde{f}_0$ and $\tilde{a}_0$ and $R$ is an integer, which is the sum of the degrees in $z$ with the first non-zero coefficient terms from $f(z;\boldsymbol{\xi})$ and $a(z;\boldsymbol{\xi})$, respectively. By Lemma 3, the information geometry of the linear system with the transfer function $h(z;\boldsymbol{\xi})$ is the same as the geometry induced by the factored-out transfer function $\tilde{h}(z;\boldsymbol{\xi})$.

The relation between $f(z;\boldsymbol{\xi}), a(z;\boldsymbol{\xi})$ and $h(z;\boldsymbol{\xi})$ is described by the following Toeplitz system:

$$\begin{pmatrix} h_0 & h_1 & h_2 & \dots \\ h_{-1} & h_0 & h_1 & \dots \\ h_{-2} & h_{-1} & h_0 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} = \begin{pmatrix} f_0 & f_1 & f_2 & \dots \\ 0 & f_0 & f_1 & \dots \\ 0 & 0 & f_0 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} a_0 & 0 & 0 & \dots \\ a_1 & a_0 & 0 & \dots \\ a_2 & a_1 & a_0 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

For a given $h(z;\boldsymbol{\xi})$, $f_r$ is determined by the coefficients of $a(z;\boldsymbol{\xi})$, *i.e.*, if we choose $a(z;\boldsymbol{\xi})$, $f(z;\boldsymbol{\xi})$ is conformable to the choice under the above Toeplitz system. The following lemma is noteworthy for further discussions. It is the generalization of Lemma 3.

**Lemma 4.** *The information geometry of a signal filter is invariant under the choice of* $a(z;\boldsymbol{\xi})$.

**Proof.** It is obvious that the information geometry of a linear system is only decided by the transfer function $h(z;\boldsymbol{\xi})$. Whatever $a(z;\boldsymbol{\xi})$ is chosen, the transfer function is the same, because $f(z;\boldsymbol{\xi})$ is conformable to the Toeplitz system. □

For further generalization, the transfer function is extended to the consideration of the Blaschke product $b(z)$, which corresponds to the all-pass filter in signal processing. The transfer function can be decomposed into the following form:

$$h(z;\boldsymbol{\xi}) = f(z;\boldsymbol{\xi})a(z;\boldsymbol{\xi})b(z)$$

where the Blaschke product $b(z)$ is given by

$$b(z) = \prod_s b(z, z_s) = \prod_s \frac{|z_s|}{z_s}\frac{z_s - z}{1 - \bar{z}_s z}$$

and every $z_s$ is on the unit disk. Although the Blaschke product can be written in $z^{-1}$ instead of $z$, our conclusion is not changed, and we choose $z$ for our convention. When $z_s = 0$, the Blaschke product is given by $b(z, z_s) = z$. Regardless of $z_s$, the Blaschke product is analytic on the unit disk. Since the Taylor expansion of the Blaschke product provides positive order terms in $z$, it is also possible to incorporate the Blaschke product into $a(z;\boldsymbol{\xi})$. However, the Blaschke product is separately considered in the paper.

The logarithmic transfer function of a linear system is represented in terms of $f, a$ and $b$:

$$\begin{aligned}\log h(z;\boldsymbol{\xi}) &= \log(f_0 a_0) + \log\Big(1 + \sum_{r=1}^{\infty}\frac{f_r}{f_0}z^{-r}\Big) + \log\Big(1 + \sum_{r=1}^{\infty}\frac{a_r}{a_0}z^{r}\Big) + \log b(z)\\ &= \phi_0 + \sum_s \log|z_s| + \sum_{r=1}^{\infty}\phi_r(\boldsymbol{\xi})z^{-r} + \sum_{r=1}^{\infty}\alpha_r(\boldsymbol{\xi})z^{r} + \sum_{r=1}^{\infty}\beta_r z^{r}\end{aligned}$$

where $\phi_0 = \log(f_0 a_0)$ and $\phi_r, \alpha_r$ are the $r$-th coefficients of the logarithmic expansions. $\phi_r$ and $\alpha_r$ are functions of $\boldsymbol{\xi}$ unless all $f_r/f_0$ and $a_r/a_0$ are constant. Meanwhile, $\beta_r = \frac{1}{r}\sum_s \frac{|z_s|^{2r}-1}{z_s^r}$ is a constant in $\boldsymbol{\xi}$.

It is also straightforward to show that the information geometry is independent of the Blaschke product.

**Lemma 5.** *The information geometry of a signal filter is independent of the Blaschke product.*

**Proof.** It is obvious that the Blaschke product is independent of the coordinate system $\boldsymbol{\xi}$. Plugging the above series into the expression of the metric tensor in complex coordinates, Equations (8) and (9), the metric tensor components are expressed in terms of $\phi_r$ and $\alpha_r$:

$$g_{ij} = \partial_i\phi_0\partial_j\phi_0 + \sum_{r=1}^{\infty}\partial_i\phi_r\partial_j\alpha_r + \sum_{r=1}^{\infty}\partial_i\alpha_r\partial_j\phi_r$$
$$g_{i\bar{j}} = \sum_{r=0}^{\infty}\partial_i\phi_r\partial_{\bar{j}}\bar{\phi}_r + \sum_{r=1}^{\infty}\partial_i\alpha_r\partial_{\bar{j}}\bar{\alpha}_r$$

and it is noteworthy that the metric tensor components are independent of the $\beta_r$ terms, which are related to the Blaschke product, because those are not functions of $\boldsymbol{\xi}$. This is why the $z$-convention for the Blaschke product is not important. It is straightforward to repeat the same calculation for the $\alpha$-connection. Based on these, the information geometry of a linear system is independent of the Blaschke product. □

According to Lemma 4, the geometry is invariant under the degree of freedom in choosing $a(z;\boldsymbol{\xi})$. By using the invariance of the geometry, it is possible to fix the degree of freedom as $a_r/a_0$ constant. With the choice of the degree of freedom, the metric tensor components of the information manifold are given by

$$g_{ij} = \partial_i\phi_0\partial_j\phi_0 \tag{10}$$
$$g_{i\bar{j}} = \sum_{r=0}^{\infty}\partial_i\phi_r\partial_{\bar{j}}\bar{\phi}_r \tag{11}$$

and it is easy to verify that the metric tensor components are only dependent on $\phi_r$ and $\bar{\phi}_r$. In other words, the metric tensor is dependent only on the unilateral part of the transfer function and a constant term in $z$ of the analytic part.

By Lemma 3, any transfer function with the upper-bounded degree in $z$ is reducible to a unilateral transfer function with a constant term. For this class of transfer functions, a similar expression for the metric tensor can be obtained. First of all, the logarithmic transfer function is given in the series expansion:

$$\begin{aligned}\log h(z;\boldsymbol{\xi}) &= \log z^R + \log h_{-R} + \log\left(1 + \sum_{r=1}^{\infty}\frac{h_{-R+r}}{h_{-R}}z^{-r}\right)\\ &= \log z^R + \sum_{r=0}^{\infty}\eta_r z^{-r}\end{aligned}$$

where $R$ is the highest degree in $z$. The coefficients $\eta_r$ are also known as the complex cepstrum [27], and $\eta_0 = \log h_{-R}$. After the series expansion of this logarithmic transfer function is plugged into the formulae of the metric tensor components, Equations (8) and (9), the metric tensor components are obtained as

$$g_{ij} = \partial_i \eta_0 \partial_j \eta_0 \tag{12}$$

$$g_{i\bar{j}} = \sum_{r=0}^{\infty} \partial_i \eta_r \partial_{\bar{j}} \bar{\eta}_r \tag{13}$$

and these expressions for the metric tensor components are similar to Equations (10) and (11) with the exchange of $\phi_r \leftrightarrow \eta_r$.

As an extension of Lemma 5, it is possible to generalize it to the inner-outer factorization of the $H^2$-functions. A function in the $H^2$-space can be expressed as the product of outer and inner functions by the Beurling factorization [28]. The generalization with the Beurling factorization is given by the following lemma.

**Lemma 6.** *The information geometry of a signal filter is independent of the inner function.*

**Proof.** A transfer function $h(z;\boldsymbol{\xi})$ in the $H^2$-space can be decomposed by the inner-outer factorization:

$$h(z;\boldsymbol{\xi}) = \mathcal{O}(z;\boldsymbol{\xi})\mathcal{I}(z;\boldsymbol{\xi})$$

where $\mathcal{O}(z;\boldsymbol{\xi})$ is an outer function and $\mathcal{I}(z;\boldsymbol{\xi})$ is an inner function. The $\alpha$-divergence is expressed with $S(z;\boldsymbol{\xi}) = |h(z;\boldsymbol{\xi})|^2 = |\mathcal{O}(z;\boldsymbol{\xi})\mathcal{I}(z;\boldsymbol{\xi})|^2 = |\mathcal{O}(z;\boldsymbol{\xi})|^2$ on the unit circle, because the inner function has the unit modulus on the unit circle. Since the $\alpha$-divergence is represented only with the outer function, other geometric objects, such as the metric tensor and the $\alpha$-connection, are also independent of the inner function. □

## 3. Kähler Manifold for Signal Processing

An advantage of the transfer function representation in the complex $z$-domain is that it is easy to test whether or not the information geometry of a given signal processing filter is a Kähler manifold. As mentioned before, choosing the coefficients in $a(z;\boldsymbol{\xi})$ is considered as fixing the degrees of freedom in calculation without changing any geometry. By setting $a(z;\boldsymbol{\xi})/a_0(\boldsymbol{\xi})$ a constant function in $\boldsymbol{\xi}$, the description of a statistical model becomes much simpler, and the emergence of Kähler manifolds can be easily verified. Since causal filters are our main concerns in practice, we concentrate on unilateral transfer functions. Although we will work with causal filters, the results in this section are also valid for the cases of bilateral transfer functions.

**Theorem 1.** *For a signal filter with a finite complex cepstrum norm, the information geometry of the signal filter is a Kähler manifold.*

**Proof.** The information manifold of a signal filter is described by the metric tensor $g$ with the components of the expressions, Equation (10) and Equation (11). Any complex manifold admits a Hermitian manifold by introducing a new metric tensor $\hat{g}$ [29]:

$$\hat{g}_p(X,Y) = \frac{1}{2}\big(g_p(X,Y) + g_p(J_pX, J_pY)\big)$$

where $X, Y$ are tangent vectors at point $p$ on the manifold and $J$ is the almost complex structure, such that

$$J_p \frac{\partial}{\partial \xi^i} = i \frac{\partial}{\partial \xi^i}, \; J_p \frac{\partial}{\partial \bar{\xi}^i} = -i \frac{\partial}{\partial \bar{\xi}^i}.$$

With the new metric tensor $\hat{g}$, it is straightforward to verify that the information manifold is equipped with the Hermitian structure:

$$\begin{aligned} \hat{g}_{ij} &= \hat{g}(\partial_i, \partial_j) = 0 \\ \hat{g}_{i\bar{j}} &= \hat{g}(\partial_i, \partial_{\bar{j}}) = g_{i\bar{j}}. \end{aligned}$$

Based on the above metric tensor expressions, it is obvious that the information geometry of a linear system is a Hermitian manifold.

The Kähler two-form $\Omega$ of the manifold is given by

$$\Omega = i \hat{g}_{i\bar{j}} d\xi^i \wedge d\bar{\xi}^j$$

where $\wedge$ is the wedge product. By plugging Equation (11) into $\Omega$, it is easy to check that the Kähler two-form is closed by satisfying $\partial_k \hat{g}_{i\bar{j}} = \partial_i \hat{g}_{k\bar{j}}$ and $\partial_{\bar{k}} \hat{g}_{i\bar{j}} = \partial_{\bar{j}} \hat{g}_{i\bar{k}}$.

Since Kähler manifolds are defined as the Hermitian manifolds with the closed Kähler two-forms, the information geometry of a signal filter is a Kähler manifold. □

An information manifold for a linear system with purely real parameters is a submanifold of a Kählerian information manifold where the metric tensor has the isometry of exchanging holomorphic- and anti-holomorphic coordinates. In addition to that, a given linear system can be described by two manifolds: one is Kähler, and another is non-Kähler. Although the dimension is doubled, working with Kähler manifolds has many advantages, which will be reiterated later.

In Theorem 1, the Hermitian condition is clearly seen after introducing the new metric tensor $\hat{g}$. It is also possible to find a condition for which the metric tensor $g$ shows the explicit Hermitian structure. To impose the explicit Hermitian condition, the following theorem is worthwhile to mention.

**Theorem 2.** *In the Kählerian information geometry of a signal filter, the Hermitian structure is explicit in the metric tensor if and only if $\phi_0$ (or $f_0 a_0$) is a constant in $\boldsymbol{\xi}$. Similarly, for the transfer function of which the highest degree in $z$ is finite, the Hermitian condition is directly found if and only if the coefficient of the highest degree in $z$ of the logarithmic transfer function is a constant in $\boldsymbol{\xi}$.*

**Proof.** Let us prove the first statement.

($\Rightarrow$) If the geometry is Kähler, it should be the Hermitian manifold satisfying

$$g_{ij} = \partial_i \phi_0 \partial_j \phi_0 = 0$$

for all $i$ and $j$. This equation exhibits that $f_0 a_0$ is a constant in $\boldsymbol{\xi}$, because $\phi_0 = \log(f_0 a_0)$.

($\Leftarrow$) If $\phi_0$ (or $f_0 a_0$) is a constant in $\boldsymbol{\xi}$, the metric tensor is found from Equations (10) and (11),

$$\begin{aligned} g_{ij} &= 0 \\ g_{i\bar{j}} &= \sum_{r=0}^{\infty} \partial_i \phi_r \partial_{\bar{j}} \bar{\phi}_r \end{aligned} \tag{14}$$

and these metric tensor conditions impose that the geometry is the Hermitian manifold. It is noteworthy that the non-vanishing metric tensor components are expressed only with $\phi_r$ and $\bar{\phi}_r$, which are functions of the impulse response functions $f_r$ in $f(z;\boldsymbol{\xi})$, the unilateral part of the transfer function. For the manifold to be a Kähler manifold, the Kähler two-form $\Omega$ needs to be a closed two-form. The condition for the closed Kähler two-form $\Omega$ is that $\partial_k g_{i\bar{j}} = \partial_i g_{k\bar{j}}$ and $\partial_{\bar{k}} g_{i\bar{j}} = \partial_{\bar{j}} g_{i\bar{k}}$. It is easy to verify that the metric tensor components, Equation (14), satisfy the conditions for the closed Kähler two-form. The Hermitian manifold with the closed Kähler two-form is a Kähler manifold.

The proof for the second statement is straightforward, because it is similar to the proof of the first one by exchanging $\phi_r \leftrightarrow \eta_r$. Let us assume that the highest degree in $z$ is $R$. According to Lemma 3, it is possible to reduce a bilateral transfer function with finite terms along the non-causal direction to the unilateral transfer function by using the factorization of $z^R$. After that, we need to replace $\eta_0$ with $\phi_0$ in the proof. The two theorems are equivalent. □

Theorem 2 can be applied to submanifolds of the information manifolds. For example, a submanifold of a linear system is a Kähler manifold if and only if $\phi_0$ (or $f_0 a_0$) is constant on the submanifold, *i.e.*, $\phi_0$ is a function of the coordinates orthogonal to the submanifold.

On a Kähler manifold, the metric tensor is derived from the following equation:

$$g_{i\bar{j}} = \partial_i \partial_{\bar{j}} \mathcal{K} \tag{15}$$

where $\mathcal{K}$ is the Kähler potential. There exists the degree of freedom in Kähler potential up to the holomorphic and anti-holomorphic function: $\mathcal{K}(\xi,\bar{\xi}) = \mathcal{K}'(\zeta,\bar{\zeta}) + \phi(\zeta) + \psi(\bar{\zeta})$. However, geometry is derived from the same relation: $g_{i\bar{j}} = \partial_i \partial_{\bar{j}} \mathcal{K}$. By using Equation (15), the information on the geometry can be extracted from the Kähler potential. It is necessary to find the Kähler potential for the signal processing geometry. The following corollary shows how to get the Kähler potential for the Kählerian information manifold.

**Corollary 1.** *For a given Kählerian information geometry, the Kähler potential of the geometry is the square of the Hardy norm of the logarithmic transfer function. In other words, the Kähler potential is the square of the complex cepstrum norm of a signal filer.*

**Proof.** Given a transfer function $h(z;\boldsymbol{\xi})$, the non-trivial components of the metric tensor for a signal processing model are given by Equation (9). By using integration by parts, the metric tensor component is represented by

$$g_{i\bar{j}} = \frac{1}{2\pi i} \oint_{|z|=1} \left\{ \partial_i \Big( \log h(z;\boldsymbol{\xi}) \partial_{\bar{j}} \log \bar{h}(\bar{\xi};\bar{\boldsymbol{\xi}}) \Big) - \log h(z;\boldsymbol{\xi}) \partial_i \partial_{\bar{j}} \log \bar{h}(\bar{\xi};\bar{\boldsymbol{\xi}}) \right\} \frac{dz}{z}$$

where the latter term goes to zero by holomorphicity. When we integrate by parts with respect to the anti-holomorphic derivative once again, the metric tensor is expressed with

$$g_{i\bar{j}} = \frac{1}{2\pi i}\oint_{|z|=1}\left\{\partial_i\partial_{\bar{j}}\Big(\log h(z;\boldsymbol{\xi})\log\bar{h}(\bar{\xi};\bar{\boldsymbol{\xi}})\Big) - \partial_i\Big(\partial_{\bar{j}}\log h(z;\boldsymbol{\xi})\log\bar{h}(\bar{\xi};\bar{\boldsymbol{\xi}})\Big)\right\}\frac{dz}{z}$$

and the latter term is also zero, because $h(z;\boldsymbol{\xi})$ is a holomorphic function.

Finally, the metric tensor is obtained as

$$g_{i\bar{j}} = \partial_i\partial_{\bar{j}}\left(\frac{1}{2\pi i}\oint_{|z|=1}\big(\log h(z;\boldsymbol{\xi})\big)\big(\log h(z;\boldsymbol{\xi})\big)^*\frac{dz}{z}\right)$$

and by the definition of the Kähler potential, Equation (15), the Kähler potential of the linear system geometry is given by

$$\mathcal{K} = \frac{1}{2\pi i}\oint_{|z|=1}\big(\log h(z;\boldsymbol{\xi})\big)\big(\log h(z;\boldsymbol{\xi})\big)^*\frac{dz}{z}$$

up to a holomorphic function and an anti-holomorphic function. The right-handed side of the above equation is known as the square of the Hardy norm for the logarithmic transfer function. It is straightforward to derive the relation between the Kähler potential and the square of the Hardy norm of the logarithmic transfer function:

$$\mathcal{K} = \frac{1}{2\pi i}\oint_{|z|=1}\big(\log h(z;\boldsymbol{\xi})\big)\big(\log h(z;\boldsymbol{\xi})\big)^*\frac{dz}{z} = \|\log h(z;\boldsymbol{\xi})\|_{H^2}^2. \tag{16}$$

Additionally, the Hardy norm of the logarithmic transfer function is also known as the complex cepstrum norm of a linear system [25,27]. □

For a given linear system, the Kähler potential of the geometry is given by $\phi_r$, $\alpha_r$ and the complex conjugates of $\phi_r$, $\alpha_r$:

$$\mathcal{K} = \sum_{r=0}^{\infty}(\phi_r\bar{\phi}_r + \alpha_r\bar{\alpha}_r).$$

However, the geometry is not dependent on $\alpha$ and $\bar{\alpha}$, because those are not the functions of the model parameters $\boldsymbol{\xi}$ under fixing the degree of the freedom. By using Equation (14), the Kähler potential is expressed with

$$\mathcal{K} = \sum_{r=0}^{\infty}\phi_r\bar{\phi}_r$$

and it is noticeable that the Kähler potential only depends on $\phi_r$ and $\bar{\phi}_r$, which come from the unilateral part of the transfer function decomposition. It is possible to obtain a similar expression for the finite highest upper-degree case by changing $\phi_r$ to $\eta_r$.

Since we assume that the complex cepstrum norm is finite, a transfer function $h(z;\boldsymbol{\xi})$ in the $H^2$-space also lives in the Hardy space of

$$\mathcal{K} = \|\log h(z;\boldsymbol{\xi})\|_{H^2}^2 < \infty.$$

This implies that the transfer function lives not only in $H^2$, but also in $\exp(H^2)$, equivalently $\log h$ in the $H^2$-space.

From Equation (15), the metric tensor is derived from the Kähler potential. Additionally, the metric tensor is also calculated from the $\alpha$-divergence. These facts indicate that there exists a connection between the Kähler potential and the $\alpha$-divergence.

**Corollary 2.** *The Kähler potential is a constant term in $\alpha$, up to purely holomorphic or purely anti-holomorphic functions, of the $\alpha$-divergence between a signal processing filter and the all-pass filter of a unit transfer function.*

**Proof.** After replacing the spectral density function with the transfer function, the 0-divergence between a signal filter and the all-pass filter with a unit transfer function is given by

$$\begin{aligned} D^{(0)}(1||h) &= \frac{1}{2\pi i}\oint_{|z|=1}\frac{1}{2}(\log h+\log\bar{h})^2\frac{dz}{z} \\ &= \mathcal{K}+\frac{1}{2\pi i}\oint_{|z|=1}\frac{1}{2}\big((\log h)^2+(\log\bar{h})^2\big)\frac{dz}{z} \\ &= \mathcal{K}+F(\boldsymbol{\xi})+\bar{F}(\bar{\boldsymbol{\xi}}) \end{aligned}$$

where $F(\boldsymbol{\xi})=\frac{1}{2}\phi_0^2=\frac{1}{2}(\log(f_0a_0))^2$. For a bilateral transfer function, $F(\boldsymbol{\xi})=\frac{1}{2}(\phi_0+\sum\log|z_s|)^2+\sum_{r=1}\phi_r(\alpha_r+\beta_r)$.

For non-zero $\alpha$, the $\alpha$-divergence between a signal and the white noise is also obtained as

$$\begin{aligned} D^{(\alpha)}(1||h) &= \frac{1}{2\pi i\alpha^2}\oint_{|z|=1}\big\{h^\alpha-1-\alpha(\log h+\log\bar{h})\big\}\frac{dz}{z} \\ &= \frac{1}{2\pi i}\oint_{|z|=1}\Big(\frac{1}{2}(\log h+\log\bar{h})^2+\sum_{n=1}^{\infty}\frac{1}{(n+2)!}\alpha^n(\log h+\log\bar{h})^{n+2}\Big)\frac{dz}{z} \\ &= D^{(0)}(1||h)+\mathcal{O}(\alpha) \\ &= \mathcal{K}+F(\boldsymbol{\xi})+\bar{F}(\bar{\boldsymbol{\xi}})+\mathcal{O}(\alpha). \end{aligned}$$

When $f_0a_0$ is unity, a constant term in $\alpha$ of the $\alpha$-divergence is the Kähler potential. This shows the relation between the $\alpha$-divergence and the Kähler potential. □

The $\alpha$-connection on a Kähler manifold is expressed with the transfer function by using Equation (1) and Equation (3). It is also cross-checked from the $\alpha$-divergence in the transfer function representation.

**Corollary 3.** *The $\alpha$-connection components of the Kählerian information geometry are found as*

$$\begin{aligned} \Gamma^{(\alpha)}_{ij,\bar{k}} &= \frac{1}{2\pi i}\oint_{|z|=1}\big(\partial_i\partial_j\log h(z;\boldsymbol{\xi})-\alpha\partial_i\log h(z;\boldsymbol{\xi})\partial_j\log h(z;\boldsymbol{\xi})\big)\big(\partial_k\log h(z;\boldsymbol{\xi})\big)^*\frac{dz}{z} \\ \Gamma^{(\alpha)}_{ij,k} &= \frac{1}{2\pi i}\oint_{|z|=1}\big(\partial_i\partial_j\log h(z;\boldsymbol{\xi})-\alpha\partial_i\log h(z;\boldsymbol{\xi})\partial_j\log h(z;\boldsymbol{\xi})\big)\big(\partial_k\log h(z;\boldsymbol{\xi})\big)\frac{dz}{z} \\ \Gamma^{(\alpha)}_{i\bar{j},k} &= \frac{1}{2\pi i}\oint_{|z|=1}-\alpha\big(\partial_i\log h(z;\boldsymbol{\xi})\big)\big(\partial_j\log h(z;\boldsymbol{\xi})\big)^*\big(\partial_k\log h(z;\boldsymbol{\xi})\big)\frac{dz}{z} \\ \Gamma^{(\alpha)}_{i\bar{j},\bar{k}} &= \frac{1}{2\pi i}\oint_{|z|=1}-\alpha\big(\partial_i\log h(z;\boldsymbol{\xi})\big)\big(\partial_{\bar{j}}\log h(z;\boldsymbol{\xi})\big)^*\big(\partial_k\log h(z;\boldsymbol{\xi})\big)^*\frac{dz}{z} \end{aligned}$$

*and the non-trivial components of the symmetric tensor $T$ are given by*

$$T_{ij,\bar{k}} = \frac{1}{\pi i}\oint_{|z|=1} \big(\partial_i \log h(z;\boldsymbol{\xi})\big)\big(\partial_j \log h(z;\boldsymbol{\xi})\big)\big(\partial_k \log h(z;\boldsymbol{\xi})\big)^* \frac{dz}{z} \tag{17}$$
$$T_{ij,k} = \frac{1}{\pi i}\oint_{|z|=1} \big(\partial_i \log h(z;\boldsymbol{\xi})\big)\big(\partial_j \log h(z;\boldsymbol{\xi})\big)\big(\partial_k \log h(z;\boldsymbol{\xi})\big) \frac{dz}{z}.$$

*In particular, the non-vanishing 0-connection components are expressed with*

$$\Gamma^{(0)}_{ij,\bar{k}} = (\Gamma^{(0)}_{\bar{i}\bar{j},k})^* = \frac{1}{2\pi i}\oint_{|z|=1} \big(\partial_i\partial_j \log h(z;\boldsymbol{\xi})\big)\big(\partial_k \log h(z;\boldsymbol{\xi})\big)^* \frac{dz}{z}$$

*and the 0-connection is directly derived from the Kähler potential:*

$$\Gamma^{(0)}_{ij,\bar{k}} = \partial_i\partial_j\partial_{\bar{k}}\mathcal{K}. \tag{18}$$

*Additionally, the $\alpha$-connection and the $(-\alpha)$-connection are dual to each other.*

**Proof.** After plugging Equation (1) into Equation (3), the derivation of the $\alpha$-connection is straightforward by considering holomorphic and anti-holomorphic derivatives in the expression. The same procedure is applied to the derivation of the symmetric tensor $T$.

The 0-connection is also directly derived from the Kähler potential. The proof is as follows:

$$\begin{aligned}
\Gamma^{(0)}_{ij,\bar{k}} &= \frac{1}{2\pi i}\oint_{|z|=1} \big(\partial_i\partial_j \log h(z;\boldsymbol{\xi})\big)\big(\partial_k \log h(z;\boldsymbol{\xi})\big)^* \frac{dz}{z} \\
&= \partial_i\partial_j\partial_{\bar{k}}\Big(\frac{1}{2\pi i}\oint_{|z|=1} \big(\log h(z;\boldsymbol{\xi})\big)\big(\log h(z;\boldsymbol{\xi})\big)^* \frac{dz}{z}\Big) \\
&= \partial_i\partial_j\partial_{\bar{k}}\big(||\log h(z;\boldsymbol{\xi})||^2_{H^2}\big) \\
&= \partial_i\partial_j\partial_{\bar{k}}\mathcal{K}.
\end{aligned}$$

To prove the $\alpha$-duality, we need to test the following relation:

$$\partial_\mu g_{\nu\rho} = \Gamma^{(\alpha)}_{\mu\nu,\rho} + \Gamma^{(-\alpha)}_{\mu\rho,\nu}$$

where the Greek letters run from $1,\cdots,n,\bar{1},\cdots,\bar{n}$. After tedious calculation, it is obvious that the above equation is satisfied regardless of combinations of the indices. Therefore, the $\alpha$-duality also exists on the Kählerian information manifolds. □

The 0-connection and the symmetric tensor $T$ are expressed in terms of $\phi_r$ and $\bar{\phi}_r$,

$$\begin{aligned}
\Gamma^{(0)}_{ij,\bar{k}} &= \sum_{r=0}^{\infty} \partial_i\partial_j\phi_r\partial_{\bar{k}}\bar{\phi}_r \\
\Gamma^{(0)}_{ij,k} &= \partial_i\partial_j\phi_0\partial_k\phi_0 \\
T_{ij,\bar{k}} &= 2\sum_{r,s=0}^{\infty} \partial_i\phi_r\partial_j\phi_s\partial_{\bar{k}}\bar{\phi}_{r+s} \\
T_{ij,k} &= 2\partial_i\phi_0\partial_j\phi_0\partial_k\phi_0.
\end{aligned}$$

With the degree of freedom that $\phi_0$ is a constant in the model parameters $\boldsymbol{\xi}$, the non-trivial components of the 0-connection and the symmetric tensor $T$ are $\Gamma^{(0)}_{ij,\bar{k}}$ and $T_{ij,\bar{k}}$, respectively. In this degree of freedom, the Hermitian condition on the metric tensor is obviously emergent, and it is also beneficial to check the $\alpha$-duality condition for non-vanishing components:

$$\begin{aligned}\partial_k g_{i\bar{j}} &= \Gamma^{(\alpha)}_{ki,\bar{j}} + \Gamma^{(-\alpha)}_{k\bar{j},i} \\ \partial_{\bar{k}} g_{i\bar{j}} &= \Gamma^{(\alpha)}_{\bar{k}i,\bar{j}} + \Gamma^{(-\alpha)}_{\bar{k}\bar{j},i}.\end{aligned}$$

We can cross-check these formulae for the geometric objects of the linear system geometry with the well-known results on a Kähler manifold. First of all, the fact that the 0-connection is the Levi–Civita connection can be verified as follows:

$$\Gamma^{(0)\,k}_{\ ij} = g^{k\bar{m}}\Gamma^{(0)}_{ij,\bar{m}} = g^{k\bar{m}}\partial_i\partial_j\partial_{\bar{m}}\mathcal{K} = g^{k\bar{m}}\partial_i g_{j\bar{m}} = \partial_i(\log g_{m\bar{n}})^k{}_j = \Gamma^k_{ij}$$

where the last equality comes from the expression for the Levi–Civita connection on a Kähler manifold. This is well-matched to the Levi–Civita connection on a Kähler manifold.

In Riemannian geometry, the Riemann curvature tensor, corresponding to the 0-curvature tensor, is given by

$$R^{\rho}{}_{\sigma\mu\nu} = \partial_\mu\Gamma^\rho_{\nu\sigma} - \partial_\nu\Gamma^\rho_{\mu\sigma} + \Gamma^\rho_{\mu\lambda}\Gamma^\lambda_{\nu\sigma} - \Gamma^\rho_{\nu\lambda}\Gamma^\lambda_{\mu\sigma}$$

where the Greek letters can be any holomorphic and anti-holomorphic indices. Similar to a Hermitian manifold, the non-vanishing components of the 0-curvature tensor on a Kähler manifold are $R^\rho_{\sigma\bar{\mu}j}$ and its complex conjugate, *i.e.*, the components with three holomorphic indices and one anti-holomorphic index (and the complex conjugate component). The non-trivial components of the Riemann curvature tensor are represented by

$$\begin{aligned}R^{(0)l}{}_{k\bar{i}j} &= \partial_{\bar{i}}\Gamma^l_{jk} - \partial_j\Gamma^l_{\bar{i}k} + \Gamma^l_{\bar{i}m}\Gamma^m_{jk} - \Gamma^l_{jm}\Gamma^m_{\bar{i}k} \\ &= \partial_{\bar{i}}\Gamma^l_{jk} = \partial_{\bar{i}}(g^{l\bar{m}}\partial_j\partial_l\partial_{\bar{m}}\mathcal{K}) = \left(R^{(0)\bar{l}}{}_{\bar{k}i\bar{j}}\right)^*\end{aligned}$$

because the 0-connection components with the mixed indices are vanishing.

Taking index contraction on holomorphic upper and lower indices in the Riemann curvature tensor, the 0-Ricci tensor is found as

$$\begin{aligned}R^{(0)}_{i\bar{j}} &= R^{(0)k}{}_{ki\bar{j}} = -R^{(0)k}{}_{k\bar{j}i} \\ &= -\partial_{\bar{j}}\partial_i(\log g_{m\bar{n}})^k{}_k = -\partial_{\bar{j}}\partial_i tr(\log g_{m\bar{n}}) \\ &= -\partial_{\bar{j}}\partial_i\log\mathcal{G}\end{aligned} \tag{19}$$

where $\mathcal{G}$ is the determinant of the metric tensor. This result is consistent with the expression of the Ricci tensor on a Kähler manifold. It is also straightforward to obtain the 0-scalar curvature by contracting the indices in the 0-Ricci tensor:

$$R^{(0)} = g^{i\bar{j}}R^{(0)}_{i\bar{j}} = -\frac{1}{2}\Delta\log\mathcal{G}$$

where $\Delta$ is the Laplace–Beltrami operator on the Kähler manifold.

The $\alpha$-generalization of the curvature tensor, the Ricci tensor and the scalar curvature is based on the $\alpha$-connection, Equation (4). The $\alpha$-curvature tensor is given by

$$\begin{aligned} R^{(\alpha)l}{}_{k\bar{i}ij} &= \partial_{\bar{i}}\Gamma^{(\alpha)l}{}_{jk} = \partial_{\bar{i}}\Big(\Gamma^{(0)l}{}_{jk} - \frac{\alpha}{2}g^{l\bar{m}}T_{jk,\bar{m}}\Big) \\ &= R^{(0)l}{}_{k\bar{i}ij} - \frac{\alpha}{2}\partial_{\bar{i}}\Big(g^{l\bar{m}}T_{jk,\bar{m}}\Big). \end{aligned}$$

The $\alpha$-Ricci tensor and the $\alpha$-scalar curvature are obtained as

$$\begin{aligned} R^{(\alpha)}_{i\bar{j}} &= R^{(\alpha)k}{}_{ki\bar{j}} = -R^{(\alpha)k}{}_{k\bar{j}i} \\ &= -\partial_{\bar{j}}\Big(\Gamma^{(0)k}{}_{ik} - \frac{\alpha}{2}g^{k\bar{l}}T_{ik,\bar{l}}\Big) \\ &= R^{(0)}_{i\bar{j}} + \frac{\alpha}{2}\partial_{\bar{j}}T^{k}_{ik} \\ R^{(\alpha)} &= R^{(0)} + \frac{\alpha}{2}g^{i\bar{j}}\partial_{\bar{j}}T^{\rho}_{i\rho}. \end{aligned}$$

It is noteworthy that the $\alpha$-curvature tensor, the $\alpha$-Ricci tensor and the $\alpha$-scalar curvature on a Kähler manifold have the linear corrections in $\alpha$ comparing with the quadratic corrections in $\alpha$ on non-Kähler manifolds. A submanifold of a Kähler manifold is also a Kähler manifold. When a submanifold of dimension $m$ exists, the transfer function of a linear system can be decomposed into two parts:

$$h(z;\boldsymbol{\xi}) = h_{\parallel}(z;\xi^1,\cdots,\xi^m)h_{\perp}(z;\xi^{m+1},\cdots,\xi^n)$$

where $h_{\parallel}$ is the transfer function on the submanifold and $h_{\perp}$ is the transfer function orthogonal to the submanifold. When it is plugged into Equation (16), the Kähler potential of the geometry is decomposed into three terms as follows:

$$\begin{aligned} \mathcal{K} &= \frac{1}{2\pi i}\oint_{|z|=1}(\log h_{\parallel} + \log h_{\perp})(\log h_{\parallel} + \log h_{\perp})^*\frac{dz}{z} \\ &= \frac{1}{2\pi i}\oint_{|z|=1}\log h_{\parallel}\log\bar{h}_{\parallel}\frac{dz}{z} + \frac{1}{2\pi i}\oint_{|z|=1}\log h_{\perp}\log\bar{h}_{\perp}\frac{dz}{z} + \frac{1}{2\pi i}\oint_{|z|=1}\log h_{\parallel}\log\bar{h}_{\perp}\frac{dz}{z} + (c.c.) \\ &= \mathcal{K}_{\parallel} + \mathcal{K}_{\perp} + \mathcal{K}_{\times} \end{aligned}$$

where $\mathcal{K}_{\parallel}$ contains the coordinates from the submanifold, $\mathcal{K}_{\times}$ is for the cross-terms and $\mathcal{K}_{\perp}$ is orthogonal to the submanifold.

It is obvious that each part in the decomposition of the Kähler potential provides the metric tensors for submanifolds,

$$\begin{aligned} g_{M\bar{N}} &= \partial_M\partial_{\bar{N}}\mathcal{K}_{\parallel} \\ g_{M\bar{n}} &= \partial_M\partial_{\bar{n}}\mathcal{K}_{\times} \\ g_{m\bar{n}} &= \partial_m\partial_{\bar{n}}\mathcal{K}_{\perp} \end{aligned}$$

where an uppercase index is for the coordinates on the submanifold and a lowercase index is for the coordinates orthogonal to the submanifold. As we already know, the induced metric tensor for the submanifold is derived from $\mathcal{K}_{\parallel}$, the Kähler potential of the submanifold. Based on this

decomposition, it is also possible to use $\mathcal{K}$ as the Kähler potential of the submanifold, because it endows the same metric with $\mathcal{K}_{\|}$. However, the Riemann curvature tensor and the Ricci tensors include the mixing terms from embedding in the ambient manifold, because the inverse metric tensor contains the orthogonal coordinates by the Schur complement. In statistical inference, connections, tensors and scalar curvature play important roles. If those corrections are negligible, dimensional reduction to the submanifolds is meaningful from the viewpoints not only of Kähler geometry, but also of statistical inference.

The benefits of introducing a Kähler manifold as an information manifold are as follows. First of all, on a Käher manifold, the calculation of geometric objects, such as the metric tensor, the $\alpha$-connection and the Ricci tensor, is simplified by using the Kähler potential. For example, the 0-connection on a non-Kähler manifold is given by

$$\Gamma^{(0)}_{ij,k} = \frac{1}{2}(\partial_i g_{kj} + \partial_j g_{ik} - \partial_k g_{ij})$$

demanding three-times more calculation steps than the Kähler case, Equation (18). Additionally, the Ricci tensor on a Kähler manifold is directly derived from the determinant of the metric tensor. Meanwhile, the Ricci tensor on a non-Kähler manifold needs more procedures. In the beginning, the connection should be calculated from the metric tensor. Additionally, then, the Riemann curvature is obtained after taking the derivatives on the connection and considering quadratic terms of the connection. Finally, the Ricci tensor on the non-Kähler manifold is found by the index contraction on the curvature tensor indices.

Secondly, $\alpha$-corrections on the Riemann curvature tensor, the Ricci tensor and the scalar curvature on the Kähler manifold are linear in $\alpha$. Meanwhile, there exist the quadratic $\alpha$-corrections in non-Kähler cases. The $\alpha$-linearity makes it much easier to understand the properties of $\alpha$-family.

Moreover, submanifolds in Kähler geometry are also Kähler manifolds. When a statistical model is reducible to its lower-dimensional models, the information geometry of the reduced statistical model is a submanifold of the geometry. If the ambient manifold is Kähler, the dimensional reduction also provides a Kähler manifold as the information geometry of the reduced model, and the submanifold is equipped with all of the properties of the Kähler manifold.

Lastly, finding the superharmonic priors suggested by Komaki [15] is more straightforward in the Kähler setup, because the Laplace–Beltrami operator on a Kähler manifold is of the more simplified form compared to that in non-Kähler cases. For a differentiable function $\psi$, the Laplace–Beltrami operator on a Kähler manifold is given by

$$\Delta\psi = 2g^{i\bar{j}}\partial_i\partial_{\bar{j}}\psi \tag{20}$$

comparing with the Laplace–Beltrami operator on a non-Kähler manifold:

$$\Delta\psi = \frac{1}{\sqrt{\mathcal{G}}}\partial_i\big(\sqrt{\mathcal{G}}g^{ij}\partial_j\psi\big) \tag{21}$$

where $\mathcal{G}$ is the determinant of the metric tensor. On a Kähler manifold, the partial derivatives only act on the superharmonic prior functions. Meanwhile, the contributions from the derivatives acting

on $\mathcal{G}$ and $g^{ij}$ should be considered in the non-Kähler cases. This computational redundancy is not on the Kähler manifold.

## 4. Example: AR, MA and ARMA Models

In the previous section, we show that the information geometry of a signal filter is a Kähler manifold. From the viewpoint of signal processing, time series models can be interpreted as a signal filter that transforms a randomized input $x(z)$ to an output $y(z)$. The geometry of a time series model can also be found by using the results in the previous section. In particular, we cover the AR, the MA and the ARMA models as examples.

First of all, the transfer functions of these time series models need to be identified. The transfer functions of the AR, the MA and the ARMA models with model parameters $\boldsymbol{\xi} = (\sigma, \xi^1, \cdots, \xi^n)$ are represented by

$$h(z;\boldsymbol{\xi}) = \frac{\sigma^2}{2\pi}\prod_{i=1}^{n}(1-\xi^i z^{-1})^{c_i}$$

where $c_i = -1$ if $\xi^i$ is an AR pole and $c_i = 1$ if $\xi^i$ is an MA root.

The ARMA models can be considered as the fraction of two AR models or two MA models. By Lemma 1, the correspondence between the $\alpha$-duality and the reciprocality of transfer functions is also valid for the ARMA$(p,q)$ models. For example, the ARMA$(p,q)$ model with $\alpha$-connection is $\alpha$-dual to the ARMA$(q,p)$ model with the $(-\alpha)$-connection under the reciprocality of the transfer function. Simply speaking, the AR model and the MA model are exchangeable by Lemma 1. The correspondence is given as follows:

$$\begin{aligned}
\text{ARMA}(p,q) &\leftrightarrow \text{ARMA}(q,p)\\
\text{poles} &\leftrightarrow \text{zeros}\\
\text{zeros} &\leftrightarrow \text{poles}\\
\sigma/\sqrt{2\pi} &\leftrightarrow \sqrt{2\pi}/\sigma\\
\alpha &\leftrightarrow -\alpha\\
\Gamma^{(\alpha)} &\leftrightarrow \Gamma^{(-\alpha)}\\
D^{(\alpha)}(h^{(0)}||h) &\leftrightarrow D^{(-\alpha)}(h^{(0)}||h)
\end{aligned}$$

where $h^{(0)}$ is the unit transfer function of an all-pass filter.

### *4.1. Kählerian Information Geometry of ARMA$(p,q)$ Models*

The ARMA$(p,q)$ model is the ($p$+$q$+1)-dimensional model with $\boldsymbol{\xi} = (\sigma, \xi^1, \cdots, \xi^{p+q})$, and the time series model is characterized by its transfer function:

$$h(z;\boldsymbol{\xi}) = \frac{\sigma^2}{2\pi}\frac{(1-\xi^{p+1}z^{-1})(1-\xi^{p+2}z^{-1})\cdots(1-\xi^{p+q}z^{-1})}{(1-\xi^1 z^{-1})(1-\xi^2 z^{-1})\cdots(1-\xi^p z^{-1})}$$

where $\sigma$ is the gain and $\xi^i$ is a pole with the condition of $|\xi^i| < 1$. The logarithmic transfer function of the ARMA$(p, q)$ model is given by

$$\log h(z; \boldsymbol{\xi}) = \log \frac{\sigma^2}{2\pi} + \sum_{i=1}^{p+q} c_i \log\left(1 - \xi^i z^{-1}\right)$$

and it is easy to verify that $f_0 a_0 = \sigma^2/2\pi$.

According to Theorem 1, the information geometry of the ARMA model is a Kähler manifold because of stability, minimum phase and the finite complex cepstrum norm of the ARMA filter. By using Theorem 2, the Hermitian condition on the metric tensor is explicitly checked on the submanifold of the ARMA model, where $\sigma$ is a constant. In addition to that, this submanifold is also a Kähler manifold, because a submanifold of a Kähler manifold is also Kähler. Since it is possible to gauge $\sigma$ by normalizing the amplitude of an input signal, the $\sigma$-coordinate can be considered as the denormalization coordinate [21]. Similar to the non-complexified ARMA models [12], $g_{0i}$ for all non-zero $i$ vanish by direct calculation using Equation (2). Considering these facts, we work only with the submanifolds of a constant gain.

As mentioned, the Käher potential is crucial for the Kähler manifolds and defined as the square of the Hardy norm of the logarithmic transfer function, equivalently the square of the complex cepstrum norm, Equation (16). For the ARMA$(p, q)$ model, the Kähler potential is given by

$$\mathcal{K} = \sum_{r=1}^{\infty} \frac{1}{r^2} \Big| \sum_{i=1}^{p+q} c_i (\xi^i)^r \Big|^2$$

Since the metric tensor is simply derived from taking the partial derivatives on the Kähler potential, Equation (15), the metric tensor of the ARMA$(p, q)$ model is represented as

$$g_{i\bar{j}} = \frac{c_i c_j}{1 - \xi^i \bar{\xi}^j}.$$

where other fully holomorphic- and fully anti-holomorphic-indexed components are all zero. It is easily verified that if $c_i$ and $c_j$ are both from the AR or the MA models, $c_i$ and $c_j$ exhibit the same signature, which imposes that the AR$(p)$- and the MA$(q)$-submanifolds of the ARMA$(p, q)$ model have the same metric tensors with the AR$(p)$ and the MA$(q)$ models, respectively. If two indices are from the different models, there exists only the sign difference in the metric tensor. The metric tensor of the geometry is of a similar form as the metric tensor in Ravishanker's work on the ARMA geometry [12].

By considering the Schur complement, the inverse metric tensor can be deduced from the inverse metric tensor of the AR($p$+$q$) model. The inverse metric tensor of the geometry is represented by

$$g^{i\bar{j}} = c_i c_j \frac{(1 - \xi^i \bar{\xi}^j) \prod_{k \neq i} (1 - \xi^k \bar{\xi}^j) \prod_{k \neq j} (1 - \xi^i \bar{\xi}^k)}{\prod_{k \neq i} (\xi^k - \xi^i) \prod_{k \neq j} (\bar{\xi}^k - \bar{\xi}^j)}$$

and the only difference with the AR case is the signature $c_i c_j$ in the AR-MA mixed components. With the sign difference in the metric tensor components with the AR-MA mixed indices, the determinant

of the metric tensor can be calculated with the aid of the Schur complement. The determinant of the metric tensor is found as

$$\mathcal{G} = \det g_{i\bar{j}} = \frac{\prod_{1\le j<k\le n} |\xi^k - \xi^j|^2}{\prod_{j,k}(1-\xi^j\bar{\xi}^k)}.$$

The 0-connection and the symmetric tensor $T$ for the Kähler-ARMA model can be found from the results in the previous section. The non-trivial 0-connection components are calculated from Equation (18):

$$\Gamma^{(0)}_{ij,\bar{k}} = \frac{c_j c_k \delta_{ij} \bar{\xi}^k}{(1-\xi^j\bar{\xi}^k)^2}$$

and the non-zero components of the symmetric tensor $T$ are given by Equation (17):

$$T_{ij,\bar{k}} = \frac{2c_i c_j c_k \bar{\xi}^k}{(1-\xi^i\bar{\xi}^k)(1-\xi^j\bar{\xi}^k)}.$$

Based on the above expressions, the $\alpha$-connection is easily obtained from Equation (4).

The 0-Ricci tensor of the ARMA geometry is represented by Equation (19):

$$R^{(0)}_{i\bar{j}} = -\frac{1}{(1-\xi^i\bar{\xi}^j)^2}$$

and it is noteworthy that the Ricci tensor is not dependent on $c_i$. The 0-scalar curvature is calculated from the 0-Ricci tensor by index contraction:

$$R^{(0)} = -\sum_{i,j} \frac{c_i c_j \prod_{k\neq i}(1-\xi^k\bar{\xi}^j)\prod_{k\neq j}(1-\xi^i\bar{\xi}^k)}{(1-\xi^i\bar{\xi}^j)\prod_{k\neq i}(\xi^k-\xi^i)\prod_{k\neq j}(\bar{\xi}^k-\bar{\xi}^j)}$$

where $c_i, c_j$ are from the inverse metric tensor of the ARMA model.

It is straightforward to derive the $\alpha$-generalization of the Riemann curvature tensor, the Ricci tensor and the scalar curvature by using the results in Section 3.

### *4.2. Superharmonic Priors for Kähler-ARMA($p, q$) Models*

As mentioned before, the Laplace–Beltrami operator on a Kähler manifold is of a much simpler form than that on a non-Kähler manifold. The simplified Laplace–Beltrami operator of the geometry makes finding superharmonic priors easier. Although it is also valid in any arbitrary dimension, let us confine ourselves to the ARMA(1,1) model as a simplification. For the ARMA$(1,1)$ model, the metric tensor is expressed with

$$g_{i\bar{j}} = \begin{pmatrix} \frac{1}{1-|\xi^1|^2} & -\frac{1}{1-\xi^1\bar{\xi}^2} \\ -\frac{1}{1-\xi^2\bar{\xi}^1} & \frac{1}{1-|\xi^2|^2} \end{pmatrix}.$$

It is trivial to show that $\psi_1 = (1-|\xi^1|^2)+(1-|\xi^2|^2)$ and $\psi_2 = (1-|\xi^1|^2)(1-|\xi^2|^2)$ are superharmonic prior functions.

In order to compare with the literature on superharmonic priors for the non-Kählerian AR models [30,31], let us consider the Kähler-AR($p$) models. For $p = 2$, the metric tensor is given by

$$g_{i\bar{j}} = \begin{pmatrix} \frac{1}{1-|\xi^1|^2} & \frac{1}{1-\xi^1\bar{\xi}^2} \\ \frac{1}{1-\xi^2\bar{\xi}^1} & \frac{1}{1-|\xi^2|^2} \end{pmatrix}.$$

With the Laplace–Beltrami operator on a Kähler manifold, it is obvious that $(1 - |\xi^k|^2)$ for $k = 1, \cdots, p$ is a superharmonic function in arbitrary $p$-dimensional AR geometry. The proof for superharmonicity is as follows:

$$\begin{aligned} \Delta(1 - |\xi^k|^2) &= 2g^{i\bar{j}}\partial_i\partial_{\bar{j}}(1 - |\xi^k|^2) \\ &= -2g^{i\bar{j}}\delta_{i,k}\delta_{j,k} = -2g^{k\bar{k}} < 0 \end{aligned}$$

because the diagonal components of the inverse metric tensor are all positive. By additivity, the sum of these prior functions, $\sum_{k=1}^{n}(1 - |\xi^k|^2)$, are also superharmonic. Obviously, $\psi_1 = (1 - |\xi^1|^2) + (1 - |\xi^2|^2)$ is a superharmonic prior function in the two-dimensional case.

Another superharmonic prior function for the AR(2) model is $\psi_2 = (1 - |\xi^1|^2)(1 - |\xi^2|^2)$. The Laplace–Beltrami operator acting on $\psi_2$ is represented by

$$\left(\frac{\Delta\psi_2}{\psi_2}\right) = -\frac{2(2 - \xi^1\bar{\xi}^2 - \xi^2\bar{\xi}^1)}{|\xi^1 - \xi^2|^2}$$

and it is simply verified that $\left(\frac{\Delta\psi_2}{\psi_2}\right) < 0$, because $2 - \xi^1\bar{\xi}^2 - \xi^2\bar{\xi}^1 > 0$. In addition to that, since $\psi_2$ is positive, $\psi_2 = (1 - |\xi^1|^2)(1 - |\xi^2|^2)$ is a superharmonic prior function.

Additionally, it is found that $\psi_3 = (1-\xi^1\bar{\xi}^2)(1-\xi^2\bar{\xi}^1)(1-|\xi^1|^2)(1-|\xi^2|^2)$ is also a superharmonic prior function. The Laplace–Beltrami operator acting on this prior function gives

$$(\frac{\Delta\psi_3}{\psi_3}) = -\frac{6}{\mathcal{G}}\frac{|\xi^1 - \xi^2|^2}{(1 - \xi^1\bar{\xi}^2)(1 - \xi^2\bar{\xi}^1)(1 - |\xi^1|^2)(1 - |\xi^2|^2)} = -6$$

and it is straightforward that $\psi_3$ is superharmonic, because $\psi_3$ is positive. This prior function is similar to the prior function found in the literature [30,31]. If the prior function is represented in the complexified coordinates, the prior function is $(1 - |\xi^1|^2)$, because the two coordinates in his paper are complex conjugate to each other.

To obtain superharmonic priors, the superharmonic prior functions found above are multiplied by the Jeffreys prior, which is the volume form of the information manifold. After that, the superharmonic priors outperform the Jeffreys prior [15].

## 5. Conclusion

In this paper, we prove that the information geometry of a signal filter with a finite complex cepstrum norm is a Kähler manifold. The conditions on the transfer function of the filter make the Hermitian structure explicit. The first condition on the transfer function for the Kählerian information

manifold is whether or not multiplication between the zero-th degree terms in $z$ of the unilateral part and the analytic part in the transfer function decomposition is a constant. The second condition is whether or not the coefficient of the highest degree in $z$ is a constant in the model parameters. These two conditions are equivalent to each other for some transfer functions.

It is also found that the square of the Hardy norm of a logarithmic transfer function is the Kähler potential of the information geometry. It is also known as the unweighted complex cepstrum norm of a linear system. Using the Kähler potential, it is easy to derive the geometric objects, such as the metric tensor, the $\alpha$-connection and the Ricci tensor. Additionally, the Kähler potential is a constant term in $\alpha$ of the $\alpha$-divergence, *i.e.*, it is related to the 0-divergence.

The Kählerian information geometry for signal processing is not only mathematically interesting, but also computationally practical. Contrary to non-Kähler manifolds where tedious and lengthy calculation is needed in order to obtain the tensors, it is relatively easier to calculate the metric tensor, the connection and the Ricci tensor on a Kähler manifold. Taking derivatives on the Kähler potential provides the metric tensor and the connection on a Kähler manifold. The Ricci tensor is obtained from the determinant of the metric tensor. Moreover, $\alpha$-generalization on the curvature tensor, the Ricci tensor and the scalar curvature is linear in $\alpha$. Meanwhile, there exist the non-linear corrections in the non-Kähler cases. Additionally, since the Laplace–Beltrami operator in Kähler geometry is of the simpler form, it is more straightforward to find superharmonic priors.

The information geometries of the AR, the MA and the ARMA models, the most well-known time series models, are the Kähler manifolds. The metric tensors, the connections and the divergences of the linear system geometries are derived from the the Kähler potentials with simplified calculation. In addition to that, the superharmonic priors for those models are found with much less computational efforts.

**Acknowledgments**

We are thankful to Robert J. Frey and Michael Tiano for useful discussions.

**Author Contributions**

Both authors contributed equally to the main idea. The research was conducted out by both authors. Both authors wrote the paper. Both authors have read and approved the final manuscript.

**Conflicts of Interest**

The authors declare no conflict of interest.

**References**

1. Rao, C.R. Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **1945**, *37*, 81–89.

2. Jeffreys, H. An invariant form for the prior probability in estimation problems. *Proc. R. Soc. London Ser. A* **1946**, *196*, 453–461.
3. Efron, B. Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Stat.* **1975** *3*, 1189–1217.
4. Amari, S. *Differential-Geometrical Methods in Statistics*; Springer: Berlin and Heidelberg, Germany, 1990.
5. Matsuyama, Y. The $\alpha$-EM algorithm: Surrogate likelihood maximization using $\alpha$-Logarithmic information measures. *IEEE Transactions on Information Theory* **2003**, *49*, 692–706.
6. Matsuyama, Y. Hidden Markov model estimation based on alpha-EM algorithm: Discrete and continuous alpha-HMMs. In Proceedings of International Conference on Neural Networks, San Jose, CA, USA, 31 July–5 August 2011; pp. 808–816.
7. Brody, D.C.; Hughston, L.P. Interest rates and information geometry. *Proc. R. Soc. Lond. A* **2001**, *457*, 1343–1363.
8. Janke, W.; Johnston, D.A.; Kenna, R. Information geometry and phase transitions. *Physica A* **2004**, *336*, 181–186.
9. Zanardi, P.; Giorda, P.; Cozzini, M. Information-theoretic differential geometry of quantum phase transitions. *Phys. Rev. Lett.* **2007**, *99*, 100603.
10. Heckman, J.J. Statistical inference and string theory. arXiv:1305.3621.
11. Arwini, K.; Dodson, C.T.J. *Information Geometry: Near Randomness and Near Independence*; Springer: Berlin and Heidelberg, Germany, 2008.
12. Ravishanker, N.; Melnick, E.L.; Tsai, C. Differential geometry of ARMA models. *J. Time Ser. Anal.* **1990**, *11*, 259–274.
13. Ravishanker, N. Differential geometry of ARFIMA processes. *Commun. Stati. Theory Methods* **2001**, *30*, 1889–1902.
14. Barbaresco, F. Information intrinsic geometric flows. *AIP Conf. Proc.* **2006**, *872*, 211–218.
15. Komaki, F. Shrinkage priors for Bayesian prediction. *Ann. Stat.* **2006**, *34*, 808–819.
16. Barndorff-Nielsen, O.E.; Jupp, P.E. Statistics, yokes and symplectic geometry. *Annales de la faculté des sciences de Toulouse 6 série* **1997**, *6*, 389–427.
17. Barbaresco, F. Information geometry of covariance matrix: Cartan-Siegel homogeneous bounded domains, Mostow/Berger fibration and Fréchet median. In *Matrix Information Geometry*; Bhatia, R., Nielsen, F., Eds.; Springer: Berlin and Heidelberg, Germany, 2012; pp. 199–256.
18. Barbaresco, F. Koszul information geometry and Souriau geometric temperature/capacity of Lie group thermodynamics. *Entropy* **2014**, *16*, 4521–4565.
19. Zhang, J.; Li, F. Symplectic and Kähler structures on statistical manifolds induced from divergence functions. *Geom. Sci. Inf.* **2013**, *8085*, 595–603.
20. Amari, S. Differential geometry of a parametric family of invertible linear systems—Riemannian metric, dual affine connections and divergence. *Math. Syst. Theory* **1987**, 20, 53–82.

21. Amari, S.; Nagaoka, H. *Methods of information geometry*; Oxford University Press: Oxford, UK, 2000.
22. Amari, S. $\alpha$-divergence is unique, belonging to both $f$-divergence and Bregman divergence classes. *IEEE Trans. Inf. Theory* **2009**, 55, 4925–4931.
23. Zhang, K.; Mullhaupt, A.P. Hellinger distance and information distance. **2015**, in preparation.
24. Bogert, B.; Healy, M.; Tukey, J. The quefrency alanysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. In Proceedings of the Symposium on Time Series Analysis, Brown University, Providence, RI, USA, 11–14 June 1963; pp. 209–243.
25. Martin, R. J. A metric for ARMA processes. *IEEE Trans. Signal Process.* **2000**, *48*, 1164–1170.
26. Cima, J.A.; Matheson, A.L.; Ross, W.T. *The Cauchy Transform*; American Mathematical Society: Providence, RI, USA, 2006.
27. Oppenheim, A.V. Superposition in a class of nonlinear systems. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 1965.
28. Beurling, A. On two problems concerning linear transformations in Hilbert space. *Acta Math.* **1949**, *81*, 239–255.
29. Nakahara, M. *Geometry, Topology and Physics*; Institute of Physics Publishing: Bristol, UK and Philadelphia, PA, USA, 2003.
30. Tanaka, F.; Komaki, F. A superharmonic prior for the autoregressive process of the second order. *J. Time Ser. Anal.* **2008**, *29*, 444–452.
31. Tanaka, F. *Superharmonic Priors for Autoregressive Models*; Mathematical Engineering Technical Reports; University of Tokyo: Tokyo, Japan, 2009.

# Most Likely Maximum Entropy for Population Analysis with Region-Censored Data

**Youssef Bennani, Luc Pronzato and Maria João Rendas**

**Abstract:** The paper proposes a new non-parametric density estimator from region-censored observations with application in the context of population studies, where standard maximum likelihood is affected by over-fitting and non-uniqueness problems. It is a maximum entropy estimator that satisfies a set of constraints imposing a close fit to the empirical distributions associated with the set of censoring regions. The degree of relaxation of the data-fit constraints is chosen, such that the likelihood of the inferred model is maximal. In this manner, the estimator is able to overcome the singularity of the non-parametric maximum likelihood estimator and, at the same time, maintains a good fit to the observations. The behavior of the estimator is studied in a simulation, demonstrating its superior performance with respect to the non-parametric maximum likelihood and the importance of carefully choosing the degree of relaxation of the data-fit constraints. In particular, the predictive performance of the resulting estimator is better, which is important when the population analysis is done in the context of risk assessment. We also apply the estimator to real data in the context of the prevention of hyperbaric decompression sickness, where the available observations are formally equivalent to region-censored versions of the variables of interest, confirming that it is a superior alternative to non-parametric maximum likelihood in realistic situations.

Reprinted from *Entropy*. Cite as: Bennani, Y.; Pronzato, L.; Rendas, M.J. Most Likely Maximum Entropy for Population Analysis with Region-Censored Data. *Entropy* **2015**, *17*, 3963–3988.

## 1. Introduction

### *1.1. Motivation*

The paper presents a new density estimator motivated by problems of population modeling, where the interest is in estimating the probability distribution $\pi_\theta, \theta \in \Theta$, of the parameters of a mathematical model $M(\cdot|\theta)$ characterizing the response $y(t|\theta)$ of individuals to applied stimuli $x(t)$. The ultimate goal is in general to be able to predict the dispersion of the response of the population to an arbitrary future stimulus $x(t)$, rather than to make a "tomography" of the population itself. These types of problems are frequent in domains like biomedical engineering, insurance studies or environmental management.

If the parameter $\theta$ can be estimated from each observation $y(t|\theta)$ and each individual's parameter is chosen independently from $\pi_\theta$, the problem of estimating $\pi_\theta$ from a collection of responses $\{(y_i(t|\theta_i), x_i(t))\}_{i=1}^N$ is formally equivalent to the usual density estimation problem from a set of independent and identically distributed samples $\{\theta_i\}_{i=1}^N \sim \pi_\theta$ and can be solved using standard parametric or non-parametric methods; see the abundant literature on non-linear mixed-effects models. The situation considered in this paper is more complex, in that the response $y(\cdot|\theta)$ of the model is not observable, and we only have access to the result of the classification of its assignment

to a finite number $(L+1)$ of possible labels by a known classifier $C(\cdot)$. Figure 1 illustrates the structural modeling/observation framework that we consider.

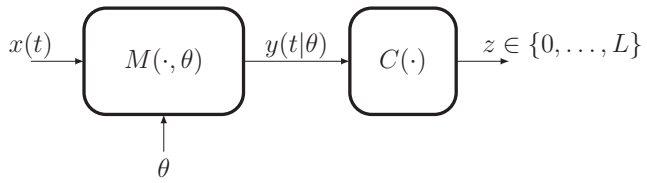



**Figure 1.** Partial response observation: $z$ is the classification of the response to stimulus $x(t)$ in a finite set.

In this setup, each observation can no longer be related to a single point $\theta \in \Theta$, the same label $z$ being assigned, for the same stimulus $x(t)$, to all responses inside a subset $R \subset \Theta$. The set $R$ is completely determined by the pair $(z, x(t))$ together with knowledge of the model $M(\cdot|\theta)$ and of the classifier rule $C(\cdot)$. This situation, when a single observation does not give information with respect to the individual value $\theta$, but only the indication that it belongs to a set, is commonly known in the statistical literature as "censored observations". While in general studies of the density estimation under censored observations have assumed that the censoring sets $R$ are intervals, the geometry of our censoring regions is determined by the structure of the (possibly highly non-linear) operators $M(\cdot|\theta)$ and $C(\cdot)$ and can have an arbitrary morphology, requiring modification of the existing methods.

In Section 4, we detail a particular instance of the problem formally presented above, relevant in the context of the prevention of decompression sickness in hyperbaric diving. Readers may want to read the material in Section 4.1 to have a concrete instantiation of the generic stimuli and operators used in the presentation above.

### *1.2. Notation and Problem Formulation*

Consider the notation introduced in Section 1.1 (see also Figure 1), and let $\{(z_n, x_n(\cdot))\}_{n=1}^{N}$ denote the available set of observations, where label $z_n \in \{0,\ldots,L\}$ has been observed for input $X^{(n)} = \{x_n(t), t \in T_n\}$, where $T_n$ is the duration of the stimulus. Denote by $R_n \subset \Theta$ the set of all individual parameters whose response to $X^{(n)}$ receives label $z_n$:

$$R_n = \left\{\theta \in \Theta \,:\, C(M(X^{(n)}|\theta)) = z_n\right\}$$

We assume that for all possible stimuli $X^{(n)}$, the composition $C\left(M(X^{(n)}|\cdot)\right)$ (of the model and the classifier) is a measurable function from $\Theta$ to $\{0,\ldots,L\}$ with respect to the restriction of the Lebesgue measure to the set $\Theta$. Under this assumption, the probability of the sets $M^{-1}_{X^{(n)}}\left(C^{-1}(\ell)\right)$ is well defined for all $0 \le \ell \le L$ and all stimuli for any distribution absolutely continuous with respect to the Lebesgue measure.

Usually, in population studies, the same stimulus is applied to several individuals. We assume here that stimuli $X^{(j)}$ are chosen in a finite set $\mathcal{X} = \left\{X^{(1)}, \ldots, X^{(J)}\right\}$. Each possible input function $X^{(j)}$ in $\mathcal{X}$ determines a partition of $\Theta$ in $L+1$ sets, that we denote by $\mathcal{Q}^{(j)} = \{R_0^{(j)}, \ldots, R_L^{(j)}\}$:

$$R_\ell^{(j)} = \left\{\theta \in \Theta \,:\, C(M(t|X^{(j)}, \theta)) = \ell\right\}, \qquad \Theta = \cup_{\ell=0}^{L} R_\ell^{(j)}, \qquad \ell_1 \neq \ell_2 \Rightarrow R_{\ell_1}^{(j)} \cap R_{\ell_2}^{(j)} = \emptyset$$

The top row of Figure 2 illustrates schematically partitions that correspond to classification in two ($L_1 = 1$) and three ($L_2 = 2$) classes of the response to two distinct stimuli.

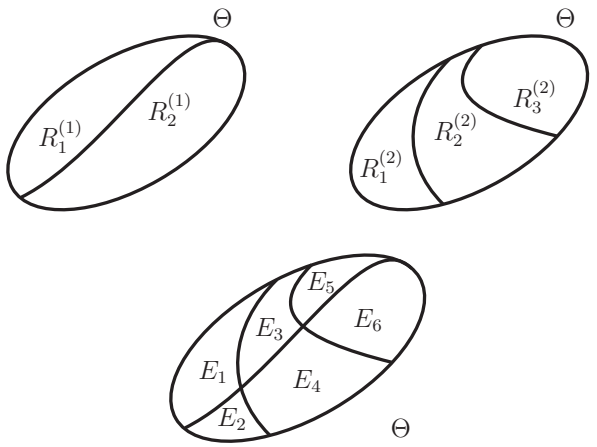



**Figure 2.** Two partitions associated with distinct stimuli, $\mathcal{Q}^{(1)} = \{R_1^1, R_2^1\}$ (top left) and $\mathcal{Q}^{(2)} = \{R_1^2, R_2^2, R_3^2\}$ (top right) and the resulting partition $\mathcal{Q}$ (bottom); see Definition 1.

Let $n_j$ be the number of times that stimulus $X^{(j)}$ has been used in the $N$ observations and $n_\ell^{(j)}$ the number of times label $\ell$ occurred in these $n_j$ experiences. The observed dataset determines $J$ empirical laws $\tilde{f}^{(j)}$, each one associated with a distinct partition $\mathcal{Q}^{(j)}$:

$$\tilde{f}_\ell^{(j)} = \frac{n_\ell^{(j)}}{n_j}, \qquad \ell = 0, \ldots, L, j = 1, \ldots, J, \qquad \sum_{\ell=0}^{L} n_\ell^{(j)} = n_j, \sum_{j=1}^{J} n_j = N \tag{1}$$

When we want to emphasize the number of observations on which these empirical laws are based, we will call $\tilde{f}^{(j)}$ an $n_j$-type. With the notation defined above, we can finally state the problem addressed in this paper with full generality.

**Problem 1. (Density estimation from region-censored data)**
*Find the non-parametric estimate of $\pi_\theta$ from the set of $J$ $n_j$-types $\tilde{f}^{(j)}, j = 1, \ldots, J$ (see Equation (1)) of the discrete random variables associated with the known partitions $\{\mathcal{Q}^{(j)}\}_{j=1}^{J}$ (see Equation (4)).*

Before initiating the study of this estimation problem, we show below how a set of constraints can be related to the observations (1) leading to an alternative problem formulation.

Let $\mathbf{1}_A(\theta)$ be the indicator function of set $A \subset \Theta$ and $\tilde{\pi}_\theta^{(n_j)}$ the (non-observed) empirical distribution:

$$\tilde{\pi}_\theta^{(n_j)}(\theta) = \frac{1}{n_j} \sum_{i=1}^{n_j} \delta\left(\theta - \theta_i^{(j)}\right), \qquad j = 1, \ldots, J$$

where $\theta_i^{(j)}$, $i = 1, \ldots, n_j$, is the parameter of the $i$-th individual to whom stimulus $X^{(j)}$ has been applied. It is immediate that $\tilde{f}_\ell^{(j)}$ in Equation (1) can be written as the statistical expectation of the indicator function of $R_\ell^{(j)}$ with respect to $\tilde{\pi}_\theta^{(n_j)}$:

$$\tilde{f}_\ell^{(j)} = \mathrm{E}_{\tilde{\pi}_\theta^{(n_j)}}\left[\mathbf{1}_{R_\ell^{(j)}}(\theta)\right], \qquad \ell = 0, \ldots, L, \;\; j = 1, \ldots, J \tag{2}$$

We stress that in our context, the (virtual) datasets $\theta^{(j)} = \left\{\theta_i^{(j)}\right\}_{i=1}^{n_j}$ are distinct for different values of $j \in \{1, \ldots, J\}$, since they correspond to statistically-independent samples from $\pi_\theta$.

The remarks above allow us to relate Problem 1 to two alternative problems: Problem 2 formulated below and Problem 3 presented in the next subsection.

**Problem 2. (Density estimation under moment constraints)**
*Consider a set of partitions $\mathcal{R}^{(j)}, j \in \{0 \ldots L\}$ all of size $L+1$, and let $\{g_m(\cdot)\}_{m=1}^M$, with $M = (L+1)J$, be the set of indicator functions $\{\mathbf{1}_{R_\ell^{(j)}}(\cdot)\}_{j=1,\ell=0}^{J,L}$. Denote by $\tilde{g}_m, m = 1, \ldots, M$, the corresponding empirical moments as in (2). Find the non-parametric estimate of $\pi_\theta$ that satisfies the set of constraints:*

$$E_{\pi_\theta}\left[g_m(\theta)\right] = \tilde{g}_m, \qquad m = 1, \ldots, M$$

Note that the existence and unicity of the solution to this problem is not guaranteed: depending on the set of partitions and empirical moments, the problem may have no solution or admit a solution (possibly non-unique).

The next subsection summarizes the present background on the two problems formulated above. Prior to that, we present three definitions that will be useful in the sequel.

**Definition 1.** *Let $\mathcal{Q}$ be the smallest partition of $\Theta$ whose generated $\sigma$-algebra, $\sigma(\mathcal{Q})$, contains all partitions $\{\mathcal{Q}^{(j)}\}_{j=1}^J$ (elements of $\mathcal{Q}$ are the minimal elements of the closure of the union of all partitions $\mathcal{Q}^{(j)}$ with respect to set intersection). The size $Q = |\mathcal{Q}|$ is necessarily finite. We denote by $E_m, m \in \{1, \ldots, Q\}$ a generic element of $\mathcal{Q}$.*

The bottom row of Figure 2 shows the partition $\mathcal{Q}$ generated by the two partitions in the top.

**Definition 2.** $\mathbf{E}_\ell^{(j)}$ *is the set of elements of $\mathcal{Q}$ that intersect $R_\ell^{(j)}$, such that:*

$$R_\ell^{(j)} = \bigcup_{E_m \in \mathbf{E}_\ell^{(j)}} E_m, \qquad \ell = 0, \ldots, L, j = 1, \ldots, J \tag{3}$$

**Definition 3.** *Let $\pi_\theta$ be a probability distribution over $\Theta$ and $\mathcal{Q}$ a finite partition of $\Theta$. We denote by $\pi_{\theta,\mathcal{Q}}$ the probability law induced by $\pi_\theta$ over the elements of $\mathcal{Q}$:*

$$\pi_{\theta,\mathcal{Q}}(E_m) = \pi_\theta\left(E_m\right), \qquad \forall E_m \in \mathcal{Q} \tag{4}$$

*1.3. Background*

1.3.1. Density Estimation from Region-Censored Data

Determination of $\hat{\pi}_\theta$, the NPMLE (non-parametric maximum likelihood estimate) of $\pi_\theta$ from censored observations, *i.e.*, the solution of Problem 1, has been studied by many authors, starting with the pioneering formulation of the Kaplan–Meier product-limit estimator [1]. Several types of censoring (one-sided, interval, *etc*.) have been considered since, first for scalar and more recently for multivariate distributions.

The problem assessed here departs from previous studies in that our (multi-dimensional) censoring regions $R_\ell^{(j)} \subset \Theta$ can have arbitrary geometry. To emphasize this, we speak of "region-censoring", instead of the more common term "interval-censoring." Another important difference concerns the fact that our regions are elements of a known set of partitions, being in general observed several times, while in general, no relation between the censoring intervals is assumed in the literature, each one being usually applied once.

Several facts are known about the NPMLE for censored observations.

**Proposition 1.**

*(i) The support of $\hat{\pi}_\theta$, $\mathcal{S}_{\text{NPMLE}} = \{\theta, :\ \hat{\pi}_\theta(\theta) > 0\}$ is confined to a finite number $K \leq Q$ of elements of $\mathcal{Q}$, the so-called "elementary regions":*

$$\mathcal{S}_{\text{NPMLE}} = \cup_{k=1}^{K} E_k, E_k \in \mathcal{Q} \tag{5}$$

*This set necessarily has a non-empty intersection with all observed lists $\mathbf{E}_\ell^{(j)}$, i.e.,*

$$n_\ell^{(j)} \neq 0 \Rightarrow \mathbf{E}_\ell^{(j)} \cap \mathcal{S}_{\text{NPMLE}} \neq \emptyset$$

*(ii) all distributions that put the same probability mass $w_k = \{\pi_\theta(E_k)\}, k = 1, \ldots, K$ in the elementary regions have the same likelihood;*

*(iii) there is in general no unique assignment of probabilities $\{\hat{w}_k\}_{k=1}^K$ that maximizes the likelihood.*

Turnbull [2] has first demonstrated (i) giving an algorithm to find the pairs $\{(E_k, w_k)\}_{k=1}^K$ for the scalar case. Gentleman and Vandal [3] addressed the multivariate interval-censored case, showing that the $E_k$'s are the intersections of the elements of the maximal cliques of the intersection graph of the set of observed intervals; see Figure 3a for a bi-dimensional example. We have shown elsewhere [4] that (i) also holds when the censoring sets have arbitrary geometry, but that some elementary regions are now associated with non-maximal cliques of the intersection graph, as shown in Figure 3b, requiring a slightly more complex identification of the sets $E_k$, which we do not detail here.

Facts (i) and (ii) together imply that the NPMLE problem can be studied in the $K$-dimensional probability simplex $\mathbb{S}^K$, since $\hat{\pi}_\theta(\cdot)$ is determined only up to the probability vector $\hat{\mathbf{w}} = \{\hat{w}_1, \ldots, \hat{w}_K\}$. The two types of "non-uniqueness" of the NPMLE, (ii) and (iii), have been first

pointed out by Turnbull [2]. More recently, they were studied in detail for the multi-variate case in [3], where the authors coined the terms representational (ii) and mixture (iii) non-uniqueness, further showing that the set of probability laws $\hat{\pi}_\theta$ defining NPMLEs is a polytope.

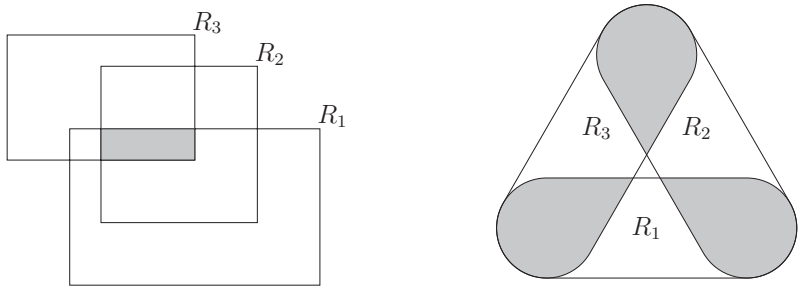



**Figure 3.** Definition of elementary regions from the cliques of the intersection graph. (**a**) Three intervals: maximal clique and corresponding elementary region $E_k$ (shaded region); (**b**) three regions with empty intersection resulting in three disjoint elementary regions $E_k$ (the shaded regions).

The NPMLE under censored observations retains the typical consistency properties of the maximum likelihood estimates, in particular $\hat{\pi}_\theta(\mathcal{R}_\ell^{(j)})$ tends to $\pi_{\theta,\mathcal{R}^{(j)}}(\ell)$ (see Equation (4)) when $n_j \to \infty$. It is not possible to guarantee the consistency of the estimate of the distribution of $\pi_{\theta,\mathcal{Q}}$ over the finer partition $\mathcal{Q}$. However, the simulations studies presented in Section 3 show that as the number of partitions $J$ tends to infinity and this $\sigma$-algebra gets finer, while keeping fixed each $n_j$ (and thus, $n \to \infty$ with $J$), the distance between the true and estimated probability laws decreases to zero.

Facts (i)–(iii) seriously hinder application of NPMLEs in many domains, in particular when, as is the case in our study, they provide a model of the diversity of the population under analysis that will be used for subsequent risk assessment. Besides being affected by some degree of arbitrariness (Facts (ii) and (iii)), the concentration of the probability mass in a small number of bounded regions reveals a tendency to underestimate population diversity, which may result in strong biases when estimating risk under unobserved stresses. The simulation studies that will be presented in Section 4 illustrates to what extent a lack of identifiability and a tendency to concentrate its support compromise the ability to predict the empirical laws corresponding to stimuli that were not used in the available dataset.

### 1.3.2. Density Estimation under Moment Constraints

Eventual non-unicity problems in density estimation under constraints on moments, like Problem 2, have been most often solved by relying on the maximum entropy (MaxEnt) principle [5] to select the most un-informative density that can match the observed moments $\{\tilde{g}_m\}_{m=1}^M$. Several information entropies have been considered in this context, the original Shannon entropy $H_1(\cdot)$ remaining the most commonly used due to its simple interpretation in terms of coding theory and

its intimate link to fundamental results in estimation theory, while amongst generalized entropies, the Rényi entropy $H_\alpha(\cdot)$, coinciding with Shannon when $\alpha \to 1$, is often chosen due to its appealing numerical and analytical tractability for $\alpha = 2$:

$$H_1(\pi) = \mathrm{E}_\pi\left[-\log(\pi)\right], \qquad H_\alpha(\pi) = \frac{1}{1-\alpha}\log\left(\mathrm{E}_\pi\left[\pi^{\alpha-1}\right]\right)$$

**Problem 3. ($H$-MaxEnt density estimator)**
*Let $H(\cdot)$ be a generalized entropy. The $H$-MaxEnt estimate $\hat{\pi}_\theta^H$ of Problem 2 is the solution of:*

$$\hat{\pi}_\theta^H = \arg\max_{\pi_\theta \in \mathcal{G}} H(\pi_\theta) \ , \qquad \mathcal{G} = \{\pi_\theta \ \textit{s.t.}\ \boldsymbol{E}_{\pi_\theta}\left[g_m(\theta)\right] = \tilde{g}_m, \qquad m = 1, \ldots, Q\}$$

When $\mathcal{G}$ is non-empty (*i.e.*, the constraints are compatible) the MaxEnt density can be analytically determined for some choices of $H(\cdot)$.

**Proposition 2. (Equivalence to ML estimation in the exponential family)**
*Assume that the constraints $\{\tilde{g}_m\}_{m=1}^M$ of Problem 2 are statistical averages with respect to the empirical distribution of a common dataset $\theta^{(N)} = \{\theta_n\}_{n=1}^N$, i.e., $\tilde{\pi}_\theta^{(n_j)} = \tilde{\pi}_\theta^{(n)}$ in Equation (2), such that:*

$$\tilde{g}_m = \frac{1}{N}\sum_{n=1}^{N} g_m(\theta_n), \qquad m = 1, \ldots, M$$

*Then:*

*(1) (Boltzmann theorem [6]) the $H_1$-MaxEnt estimate $\hat{\pi}_\theta^{H_1}$ maximizes the likelihood of the observations in the exponential family,*

$$\hat{\pi}_\theta^{H_1}(\theta) = \frac{1}{Z_\lambda}\prod_{m=1}^{M}\exp\left(\lambda_m g_m(\theta)\right) \tag{6}$$

*where $Z_\lambda$ is a normalizing constant (the partition function), and the $\{\lambda_m\}_{m=1}^M$ are determined such that the $M$ constraints are satisfied.*
*In short, the MaxEnt (non-parametric) estimate coincides with the maximum likelihood parametric estimate inside the exponential distributions.*

*(2) the $H_2$-MaxEnt estimate [7] $\hat{\pi}_\theta^{H_2}$ is:*

$$\hat{\pi}_\theta^{H_2}(\theta) = \left[-\frac{1}{2}\sum_{m=1}^{M}\lambda_m g_m(\theta)\right]_+$$

*where $[\cdot]_+ = \max(\cdot, 0)$ and the $\{\lambda_m\}_{m=1}^M$ are such that the $M$ constraints are satisfied.*

Note that the $H_1$-MaxEnt/ML equivalence is lost when the empirical averages $\tilde{g}_m$ are not all obtained from the same dataset, as is the case in our problem, where (see Equation (2)) constraints associated with distinct stimuli are being derived from distinct empirical distributions.

When the constraints are not compatible, *i.e.*, $\mathcal{G} = \emptyset$ and Problem 2 has no solution, $\hat{\pi}_\theta^H$ is not defined, and only a relaxed version of the original problem can be solved.

**Problem 4. (Relaxed $H$–MaxEnt density estimator)**
*Let $H$ be a generalized entropy, and $\boldsymbol{\epsilon} \in \mathbb{R}^{+^M}$. The $\boldsymbol{\epsilon}$-relaxed $H$-MaxEnt density estimate $\hat{\pi}_\theta^{ME,\boldsymbol{\epsilon}}$ is the solution of:*

$$\hat{\pi}_\theta^{ME,\boldsymbol{\epsilon}} = \underset{\pi_\theta \in \mathcal{G}^{(\boldsymbol{\epsilon})}}{\arg\max}\, H(\pi_\theta) \ , \qquad \mathcal{G}^{(\boldsymbol{\epsilon})} = \left\{ \pi_\theta \ s.t. \ \ \left\| \mathbf{g} - \tilde{\mathbf{g}} \right\|_{\pi_\theta} \leq \boldsymbol{\epsilon} \right\}$$

*where $\mathbf{g}$ is the $M$-dimensional vector function with $m$-th component $g_m(\cdot)$, $\tilde{\mathbf{g}}$ is the $M$-dimensional vector of empirical expectations of $\mathbf{g}$, $\|\cdot\|_\pi$ is a vector of norms depending on $\pi$ and inequality is understood component-wise.*

This estimator has been studied in detail in [8,9] for the Shannon entropy and moment constraints derived from a single empirical distribution, where the authors fully exploit the equivalence between regularized MaxEnt as formulated above and $\ell_1$-penalized maximum likelihood in the exponential family, showing that Proposition 2 holds in a more generic sense.

**Proposition 3. (Equivalence of $\ell_1$-regularized $H_1$-MaxEnt and penalized log-likelihood [9])**
*Problem 4 with $H = H_1$ (Shannon entropy) and $\|\cdot\|_\pi$ the $\ell_1$ norm for the expected value:*

$$\left[ \left\| \mathbf{g} - \tilde{\mathbf{g}} \right\|_{\pi_\theta} \right]_m = \left| E_{\pi_\theta}[g_m(\cdot)] - \tilde{g_m} \right|$$

*where the constraints $\tilde{\mathbf{g}}$ are empirical averages computed using a dataset $\Theta$, is equivalent to the maximization of the sum of the log-likelihood of $\Theta$ for the exponential family (6) penalized by the term $\sum_m \epsilon_m |\lambda_m|$, where $\epsilon_m$ is the $m$-th element of $\boldsymbol{\epsilon}$.*

By linking the relaxation level (the parameter $\boldsymbol{\epsilon}$ in Problem 4) to the expected level of accuracy of the empirical averages $\tilde{g}_m$, in [8,9], the authors are able to establish performance guarantees for the resulting density estimate, in terms of log-likelihood loss.

As before, this regularized-MaxEnt/penalized-ML equivalence only holds when all constraints are on the empirical moments with respect to the same underlying empirical distribution. This is not true in population analysis, where an individual is observed only through one of the partitions, and we cannot invoke the properties of maximum likelihood estimators to characterize the properties of regularized MaxEnt estimators, as is done in [8].

We remark that the regularized MaxEnt estimates are unique for strictly concave entropy functionals and always exist for sufficiently large $\boldsymbol{\epsilon}$. They do not suffer from neither representational non-unicity, the optimal continuous distribution being constant inside each element of $\mathcal{Q}$, nor from mixture non-uniqueness, being the solution of a concave criterion under linear inequality constraints.

### *1.4. Contributions*

As largely documented in the literature, the NPMLE using censored data frequently exhibits a singular behavior. By concentrating probability mass in a subset of $\Theta$ of a small Lebesgue measure, they favor "over-homogenous" population models that may lead to dangerous biases in the context of risk assessment, by masking the existence of individuals for which risk can be large. As shown above,

the problem of density estimation from censored observations addressed in the paper can be recast as the problem of density estimation under a set of constraints derived from the censored observations, each constraint being associated with one of the censoring regions.

While MaxEnt has been frequently used for density estimation from the joint observation of empirical moments of a set of features, its use for region-censored data arising from strongly quantified data, as we consider in this paper, violates the conditions under which previous equivalence to maximum likelihood estimation in the Gibbs family can be established. In these circumstances, guarantees on the likelihood of the original data can be no longer given.

We propose a novel estimator that explicitly relies on the two criteria, the most likely maximum entropy estimator (MLME), where the degree $\epsilon$ of regularization of a MaxEnt estimate (*i.e.*, of the solutions to Problem 4) is chosen such that the resulting estimate has maximum likelihood. The duality of the two criteria is exploited to allow suppression of singularities that are due to inconsistent or small datasets, and the resulting solution converges to the non-parametric maximum likelihood solution as the size of the datasets associated with each constraint (censoring region) grows. By using the Rényi entropy of order two instead of the Shannon entropy, we are led to a quadratic optimization problem with linear inequality constraints that has an efficient numerical implementation.

While no theoretical performance guarantees are given, the paper presents numerical studies of the performance of the proposed MLME estimator in real and simulated data, comparing it to the NPMLE and to the best fitting MaxEnt solutions. The results of cross-validation on a real dataset show that our novel estimator is better than the NPMLE or the minimally-regularized MaxEnt estimator, leading to better predictions of the population risk under unobserved stress conditions.

The paper is organized as follows. Section 2 illustrates the poor behavior of the NPMLE using simulated data. We show (Section 2.4) that even the most uncertain of the NPMLEs still presents singularities that are unlikely to occur in a natural population. The section starts by presenting the likelihood function and defining the polytope of NPMLE solutions. It also addresses the numerical determination of the NPMLE, and two optimization algorithms are presented.

Section 3 presents the main contribution of the paper, introducing the most likely Rényi MaxEnt estimator (MLME; see Definition 4). We compare our estimator to the NPMLE, demonstrating using simulated datasets that it performs better. We also present numerical studies of its asymptotic behavior as the number $J$ of different stimuli becomes large, revealing a remarkably better behavior.

In Section 4, the proposed estimator is applied to the real problem that motivated this study, in the context of the prevention of decompression sickness in hyperbaric deep sea diving. The new estimator is compared to classical maximum likelihood and maximum entropy estimators on real and simulated data, illustrating the superior performance of the new estimator in a realistic situation.

## 2. The NPMLE

### *2.1. Simulated Data Generation Mechanism*

Before presenting the NPMLE and discussing its determination, we present the data generation mechanism that will be used to illustrate the different estimators presented in this and the next sections. The simulated population distribution $\pi_\theta$ is the restriction of the joint distribution of two independent and identically distributed normal variables of mean $\mu = 0.5$ and variance $\sigma^2 = 0.2$ to the unit square $\Theta = [0, 1]^2$.

Partitions $\mathcal{R}^{(j)}$ are randomly generated by considering random unions of the elements of the Voronoi tessellation of $S = 50$ points uniformly drawn in $\Theta$; see Figure 4. The partition $\mathcal{Q}$ induced by 10 random binary splits of $\Theta$ is shown in Figure 5a, and Figure 5b is a color-coded representation of the probability law $p_{\pi,\mathcal{Q}}$, where the fine partition $\mathcal{Q}$ is easily recognizable (black delimited polygonal regions). We remark that the size of the elements of the partitions generated by our simulation mechanism tends have low dispersion, following approximately a gamma distribution with both parameters equal to $(7/2)\lambda^{-2}$, where $\lambda$ is the intensity of the homogenous Poisson process [10] ($\lambda = 1/50$ in our simulations). In Section 4, we will see that this may not be the case in practical applications.

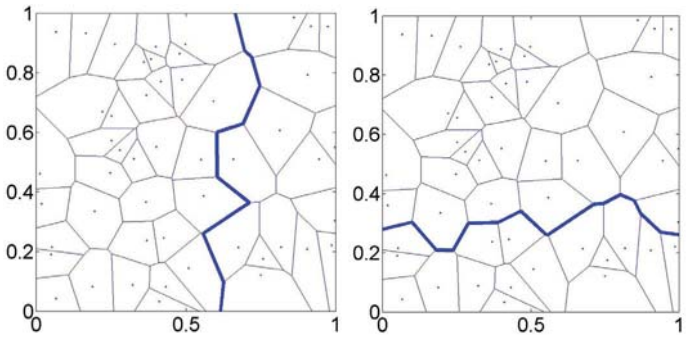
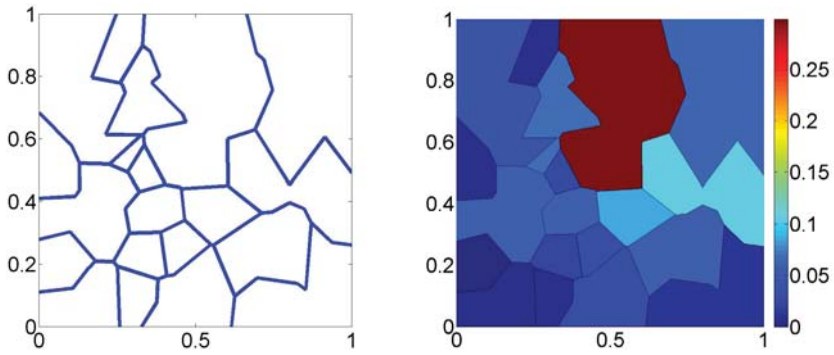



**Figure 4.** Two randomly-generated partitions of the unit square.

Observations are then generated by independently sampling $n_j$ times from each of the probability laws associated with the individual partitions $\mathcal{R}^{(j)}$, for $j = 1, \ldots, J$. In the numerical studies presented in this section, $J = 10$. To simulate the situation when some stimuli are seldom applied (for instance, if they may have compromised the safety of the individual to which they are applied), the partitions are divided into two groups, representing "safe" and "dangerous" stimuli, of sizes seven and three, respectively. The probability that a dangerous partition is chosen is $10^{-3}$, and inside each group, partitions are chosen uniformly. Except when indicated otherwise, we will consider a total of $N = 10^4$ observations.




**(a)** **(b)**

**Figure 5.** (**a**) Partition $\mathcal{Q}$ determined by $J = 10$ random binary partitions of $\Theta$. (**b**) Probability law $\pi_{\theta,\mathcal{Q}}$ induced over the elements of the partition $\mathcal{Q}$.

### *2.2. Likelihood Function*

The log-likelihood function for Problem 1 is:

$$\mathcal{L}(\pi_\theta; \{\tilde{\mathbf{f}}^{(j)}, \mathcal{Q}^{(j)}\}) = \frac{1}{N} \sum_{j=1}^{J} n_j \sum_{\ell=0}^{L} \tilde{f}_\ell^{(j)} \log p_{\pi_\theta, \mathcal{Q}^{(j)}}(\ell) \tag{7}$$

Consider the partition $\Theta = S_{\text{NPMLE}} \cup \overline{S_{\text{NPMLE}}}$, with $\overline{A}$ the complement of set $A$, and where $S_{\text{NPMLE}}$ is the union of the elementary regions $\{E_k\}_{k=1}^K$ in Proposition 1 (i), such that $p_{\pi_\theta,\mathcal{Q}^{(j)}}(\ell) = \pi_\theta(R_\ell^{(j)} \cap S_{\text{NPMLE}}) + \pi_\theta(R_\ell^{(j)} \cap \overline{S_{\text{NPMLE}}})$.

Note that since the elementary regions $\{E_k\}_{k=1}^K$ are elements of $\mathcal{Q}$, notation $\mathbf{E}_i^{(j)}$ introduced in Definition 2 is well defined.

From Proposition 1 (i) $\hat{\pi}_\theta(R_\ell^{(j)} \cap \overline{S_{\text{NPMLE}}}) = 0$ and, thus, using (3):

$$p_{\pi_\theta,\mathcal{Q}^{(j)}}(\ell) = \sum_{E_m \in \mathbf{E}_\ell^{(j)}} \pi_\theta(E_m) \quad = \mathbf{B}_{\ell.}^{(j)} \mathbf{w} \tag{8}$$

where $\mathbf{B}_{\ell.}^{(j)}$ is the $\ell$-th row of $\mathbf{B}^{(j)}$, the $(L+1) \times K$ binary matrix, with $\mathbf{B}_{\ell k}^{(j)} = 1 \Leftrightarrow E_k \in \mathbf{E}_\ell^{(j)}$, and $\mathbf{w} \in \mathbb{S}^K$ is the vector of probabilities of the elementary regions $E_k$: $w_k = \pi_\theta(E_k), k = 1, \ldots, K$, with $\mathbb{S}^K$ the $K$-dimensional probability simplex:

$$\mathbb{S}^K = \{\mathbf{w} \in \mathbb{R}^K : w_k \geq 0, \sum_{k=1}^{K} w_k = 1\}$$

Equations (7) and (8) show that (Proposition 1 (iii)) all $\pi_\theta$ leading to the same $\mathbf{w}$ have the same likelihood.

**Proposition 4.** *There is in general no single* $\mathbf{w}$ *maximizing (7) and all elements of:*

$$\mathcal{P} = \left\{\mathbf{w} \in \mathbb{S}^K, \ s.t. \ \forall j, \ \mathbf{B}^{(j)}\mathbf{w} = \mathbf{B}^{(j)}\hat{\mathbf{w}}\right\} \tag{9}$$

*where* $\hat{\mathbf{w}}$ *is a NPMLE are also NPMLEs. We call* $\mathcal{P}$ *the NPMLE polytope.*

Note that the non-unicity statement above concerns $\mathbf{w}$ the probabilities of the elementary regions $E_k$, but that the probability of the censoring regions $R_\ell^{(j)}$ is uniquely estimated, all $\mathbf{w} \in \mathcal{P}$ assigning the same probabilities to the elements of the partitions $\mathcal{Q}^{(j)}$. It is obvious that the estimator is consistent for these, but no stronger statement seems to be possible.

### *2.3. Optimizing the Likelihood*

Several algorithms have been proposed to maximize (7); see e.g., [11]. Gentleman and Vandal [3] discussed several methods and summarized them in two categories: those based on convex optimization and those based on finite mixture estimation. Two algorithms are compared in [11]: the iterative convex minorant (ICM), initially presented by Groeneboom and Wellner [12], and the vertex exchange method [13].

We show below that a multiplicative algorithm, known as the Richardson–Lucy algorithm [14] in the framework of image deconvolution, can be used to maximize $\mathcal{L}$. This follows from the fact that maximization of $\mathcal{L}$ is equivalent to an optimal design problem, enabling application of a vast collection of efficient algorithms originating from optimal design theory. As far as we know, this link of NPMLE estimation using censored observations to a $D$-optimal design problem has not been remarked on before.

Consider the $n_j(L+1) \times K$ matrices $\mathbf{B}^{(j)'}$, obtained from the $((L+1) \times K)$ matrix $\mathbf{B}^{(j)}$ by repeating $n_\ell^{(j)}$ times line $\ell$, the $N \times K$ matrix $\mathbf{B}'$ that stacks all $\mathbf{B}^{(j)'}, j = 1, \ldots, J$, the $N \times N$ diagonal matrix $\mathbf{H}_k = \text{diag}(\mathbf{B}'_{\cdot k})$ and the matrix $\mathbf{M}(\mathbf{w}) = \sum_{k=1}^K w_k \mathbf{H}_k$. Then, it is easy to show that $\mathcal{L}$ can be written as:

$$\mathcal{L}(\mathbf{w}; \{\tilde{\mathbf{f}}^{(j)}, \mathcal{Q}^{(j)}\}) = \log\det \mathbf{M}(\mathbf{w})$$

demonstrating that the determination of $\hat{\mathbf{w}}$ maximizing $\mathcal{L}(\mathbf{w}; \{\tilde{\mathbf{f}}^{(j)}, \mathcal{Q}^{(j)}\})$ with respect to $\mathbf{w} \in \mathbb{S}^K$ corresponds to a $D$-optimal design problem for the matrix $\mathbf{M}(\mathbf{w})$, with $\mathbf{w}$ considered as a design measure allocating weight $\mathbf{w}_k$ to the elementary design matrix $\mathbf{H}_k$ (see, e.g., [15]). A number of important properties follow from this equivalence with a $D$-optimal design problem. In particular, see [16,17], the iterations:

$$w_k^{(t+1)} = \frac{1}{N}\left(\sum_{j=1}^{J}\sum_{\ell=0}^{L} n_\ell^{(j)} \frac{\mathbf{B}^{(j)}_{(\ell+1)k}}{\mathbf{B}^{(j)}_{(\ell+1)\cdot}\mathbf{w}^{(t)}}\right) w_k^{(t)} \tag{10}$$

initialized at some strictly positive $\mathbf{w}^{(0)}$ converge to a maximizer of (7). This multiplicative algorithm is easy to implement, but the following vertex exchange method (VEM) [13] ensures a faster convergence to the optimum. The VEM updating rule is:

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \alpha w_{k^\star}(\mathbf{e}_{k^\star} - \mathbf{e}_{k_\star}) \tag{11}$$

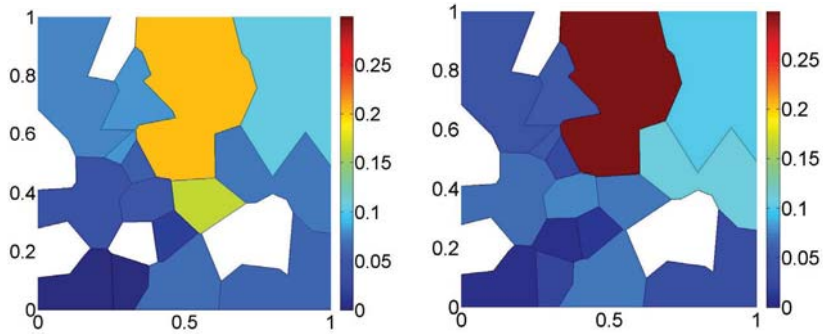
where:

$$\mathbf{w}^{(0)} \in \mathbb{S}^K,\ k^\star = \underset{k\in\{1,\ldots,K\}}{\arg\max}\, d(\mathbf{w}^{(t)}, k),\ k_\star = \underset{k\in\{1,\ldots,K\}, w_k^{(t)}>0}{\arg\min}\, d(\mathbf{w}^{(t)}, k),\ d(\mathbf{w}, k) = \text{trace}\left[\mathbf{M}^{-1}(\mathbf{w})\mathbf{H}_k\right]$$

and $\alpha$ is chosen to maximize a quadratic approximation of the log-likelihood evaluated at $\mathbf{w}^{(t+1)}$. In the multiplicative and VEM algorithms, we use the stopping condition $\max_{k\in\{1,\ldots,K\}} \frac{d(\mathbf{w},k)}{N} - 1 < \delta \ll 1$.

Our numerical studies show that the VEM Algorithm (11) is faster than the multiplicative Algorithm (10) requiring on average three-times less iterations to converge. We stress that these optimization algorithms can be applied to all $Q$ elements of the complete partition $\mathcal{Q}$ and automatically sets to zero the entries of $\mathbf{w}$ that do not correspond to the elementary sets $\{E_k\}_{k=1}^K$ (in particular, when using the result in [18] to detect the entries of $\mathbf{w}$ that can be set to zero), so that the computationally-expensive analysis of the intersection graph presented in [4] is not required.

Figure 6a shows one of the NPMLE estimates (*i.e.*, one element of the NPMLE polytope) for a simulated dataset produced, as we explained in the beginning of the section. This example clearly displays the NPMLE singularities that have been mentioned before: while $\pi_\theta$ is strictly positive inside the complete unit square, significant regions of $\Theta$ are assigned zero probability mass (the white regions in the figure), and the support of $\hat{\pi}_\theta$ is strictly contained in $\Theta$.




**(a)** **(b)**

**Figure 6.** (**a**) $\hat{\pi}_\theta$, one non-parametric maximum likelihood estimate (NPMLE) solution found by (10). (**b**) $\hat{\pi}_\theta^{\mathcal{L}}$, the Rényi-MaxEnt NPMLE. The white regions have zero probability mass.

### *2.4. Least Informative NPMLE*

As stated in Proposition 1 (iii), the NPMLE is not unique, and we have seen (Proposition (4)) that the set of solutions is the polytope $\mathcal{P}$ defined in Equation (9), associated with the matrix $\mathbf{B}$,

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}^{(1)} \\ \vdots \\ \mathbf{B}^{(J)} \end{bmatrix}$$

Motivated by the ultimate goal of capturing the largest possible diversity of the underlying population, we select from the NPMLE polytope $\mathcal{P}$ the distribution that is least informative, *i.e.*, that has maximum entropy.

Let $\hat{\mathbf{w}}$ be an NPMLE, and define $\hat{\mathbf{f}}^{(j)} = \mathbf{B}^{(j)}\hat{\mathbf{w}}$, with $\hat{\mathbf{w}}$ the vector of probabilities $\hat{\pi}_\theta(E_k), k = 1, \ldots, K$. We denote by $\hat{\pi}_\theta^{\mathcal{L}}$ the distribution in $\mathcal{P}$ maximizing the entropy $H$; it satisfies:

$$\hat{\pi}_\theta^{\mathcal{L}}(R_\ell^{(j)}) = \hat{f}_\ell^{(j)} = \hat{\pi}_\theta(R_\ell^{(j)}) \ , \qquad \ell = 0, \ldots, L; \ j = 1, \ldots, J \tag{12}$$

Determining $\hat{\pi}_\theta^{\mathcal{L}}$ for the Shannon entropy, *i.e.*, for $H = H_1$, is a non-trivial non-linear constrained optimization problem. However, for $H = H_2$, the Rényi-MaxEnt NPMLE probability vector $\tilde{\mathbf{w}}$ is the solution to the following quadratic program with linear equality constraints:

$$\tilde{\mathbf{w}} = \underset{\mathbf{w} \in \mathbb{S}^{\mathbf{K}}}{\operatorname{argmin}} \qquad \sum_{k=1}^{K} \frac{1}{\nu(E_k)} w_k^2, \qquad \text{s.t.} \quad \mathbf{B}\mathbf{w} = \mathbf{B}\hat{\mathbf{w}}$$

with $\nu(E_k)$ the volume of $E_k$, for which efficient solutions exist.

The Rényi-MaxEnt NPMLE for the same dataset that leads to the NPMLE in Figure 6a is displayed in Figure 6b. We can see that the restriction to the NPMLE polytope still forces the density to be concentrated in a strict subset of $\Theta$, with areas of zero measure (white zones in Figures 6a,b). This is inherent to the likelihood criterion, which favors the most concentrated densities that are able to explain the observed data.

Indeed, it is easy to see that the support of a NPMLE density may importantly shrink when a stimulus that is applied only once is added to the dataset, confirming the ill-conditioning of the NPMLE for small datasets. Suppose a new stimulus $X^{(J+1)}$ applied only once with resulting label $\ell^\star$ is added to a dataset already containing $J$ stimuli:

$$n_{\ell^\star}^{(J+1)} = n_{J+1} = 1, \qquad n_\ell^{(J+1)} = 0, \ell \neq \ell^\star$$

Let $\mathcal{Q}$ be the partition of $\Theta$ corresponding to "old" stimuli $j \leq J$ and $\mathcal{Q}'$ the new partition, which also integrates $(X^{(J+1)}, n^{(J+1)})$. If $R_{\ell^\star}^{(J+1)}$ intersects an elementary set $E_k \in \mathcal{Q}$, such that:

$$E_k' = E_k \cap R_{\ell^\star}^{(J+1)} \in \mathcal{Q}'$$

then $E_k \setminus E_k'$ will no longer be an elementary set, showing that the support of the NPMLE will shrink. Note that we may have $\nu(E_k') \ll \nu(E_k)$, with $\nu(\cdot)$ the Lebesgue measure.

### 3. Most Likely Rényi-MaxEnt

To avoid the singular behavior of the NPMLE, we must estimate $\pi_\theta$ with a criterion other than maximum likelihood. Relying on the link of our problem with density estimation under constraints, we propose to estimate $\pi_\theta$ through the maximum entropy principle.

If there exists a $\pi$ that can satisfy all constraints, *i.e.*, if there exists a solution to Problem 2, the corresponding $\mathbf{w}$ belongs to the NPMLE polytope $\mathcal{P}$. However, being derived from $J$ distinct empirical distributions, the $J$ constraints are in general inconsistent, and as in [9], we consider entropy maximization under relaxed constraints, *i.e.*, Problem 4. For reasons of numerical efficiency, we consider the Rényi entropy $H_2$.

**Problem 5. (Relaxed ME estimator)**
*For $\epsilon \in \mathbb{R}^+$, define the $\epsilon$-relaxed MaxEnt estimator as:*

$$\hat{\pi}_\theta^{H_2,\epsilon} = \underset{\pi}{\operatorname{argmax}} \; H_2(\pi)$$
$$s.t. \quad \left\| \Sigma^{(j)^{-1/2}} \left( E_\pi[\mathbf{f}_+^{(j)}] - \tilde{\mathbf{f}}_+^{(j)} \right) \right\|_\infty \leq \epsilon, \;\; \forall j = 1, \ldots, J \tag{13}$$

*where $\Sigma^{(j)}$ is the covariance of the empirical estimate $\tilde{\mathbf{f}}_+^{(j)}$ and $\mathbf{f}_+^{(j)}$ is obtained from $\mathbf{f}^{(j)}$ by retaining all but one of its non-zero elements.*

We remark that the constraints in Problem 5, the relaxed MaxEnt problem that we solve, take into account the correlation between the observed frequencies, contrary to what is done in [9], where the degrees of relaxation of each constraint are fixed independently, as in Problem 4. As we will verify in Section 4 (see also the discussion around Figure 8), use of an inappropriate metric in the constraints directs the estimator towards sets of solutions that have have lower likelihood, resulting in a poor ability to reproduce the observed empirical moments.

Denote by $\epsilon^\star \geq 0$ the smallest value of $\epsilon$ for which there exists a solution to Problem 5. Since in (13) we use the $\ell_\infty$ metric to evaluate the deviation of a model $\pi$ with respect to the empirical moments and $\ell_\infty$ is not equivalent to the (Riemannian) metric induced by maximum likelihood in the simplex $\mathbb{S}^K$, we cannot guarantee that likelihood is monotonically decreasing with the degree of relaxation, *i.e.*, that $\mathcal{L}(\hat{\pi}_\theta^{H_2,\epsilon}) < \mathcal{L}(\hat{\pi}_\theta^{H_2,\epsilon^\star})$, for $\epsilon > \epsilon^\star$. In fact, as the plot of the log-likelihood of $\hat{\pi}_\theta^{H_2,\epsilon}$ as a function of $\epsilon/\epsilon^\star$ in Figure 7 shows, this is not necessarily true for values of $\epsilon$ close to $\epsilon^\star$. More importantly, this figure shows that a suitable choice of the relaxation term can lead to a likelihood loss with respect to the NPMLE that is minimal, improving the fit to the data. These remarks motivate the definition of the new estimator proposed in this paper.

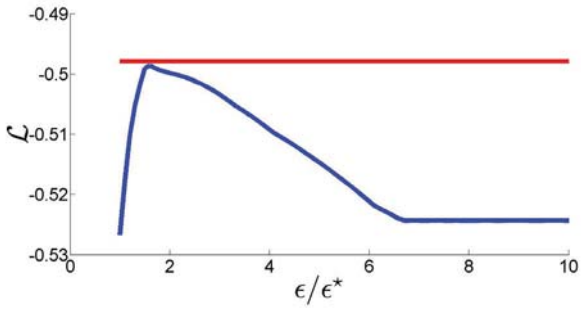



**Figure 7.** Log likelihood variation of $\hat{\pi}_\theta^{H_2,\epsilon}$ as a function of $\epsilon/\epsilon^\star$. Red line: $\mathcal{L}(\hat{\pi}_\theta)$.

**Definition 4.** ***(MLME: the most likely MaxEnt estimator)***
*Let $\hat{\pi}_\theta^{H_2,\epsilon}$ denote the solution of Problem 5 for a generic $\epsilon \geq \epsilon^\star$. The most likely Rényi-MaxEnt estimator is:*

$$\hat{\pi}_\theta^{H_2,ml} = \underset{\hat{\pi}_\theta^{H_2,\epsilon},\; \epsilon \geq \epsilon^\star}{\operatorname{argmax}} \; \mathcal{L}(\hat{\pi}_\theta^{H_2,\epsilon}; \{\tilde{\mathbf{f}}^{(j)}, \mathcal{Q}^{(j)}\}) \tag{14}$$

**Proposition 5.** *($\epsilon^\star = 0$)*
*If $\epsilon^\star = 0$, then the feasible set of the constrained optimization Problem 5 coincides with the NPMLE polytope. Since the likelihood of all solutions with $\epsilon > 0$ will be smaller, the MLME estimate coincides in this case with the MaxEnt NPMLE: $\epsilon^\star = 0 \Rightarrow \hat{\pi}_\theta^{H_2,ml} = \hat{\pi}_\theta^{H_2,\epsilon^\star} = \hat{\pi}_\theta^{\mathcal{L}}$.*

Since the probability that $\epsilon^\star = 0$ is small for finite datasets, the solution space of our constrained optimization problem is in general larger than the NPMLE polytope $\mathcal{P}$. We illustrate now the geometry of the NLME $\hat{\pi}_\theta^{H_2,ml}$ using the following simple example for which $L = 1$, $J = 2$, $K = 3$ and:

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}^{(1)} \\ \\ \mathbf{B}^{(2)} \end{bmatrix} = \left[ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right]$$

This choice allows us to represent graphically the elements of $\mathbb{S}^3$; see Figure 8. The empirical moments $(\tilde{f}_0^{(1)}, \tilde{f}_1^{(1)}, \tilde{f}_0^{(2)}, \tilde{f}_1^{(2)})$ have been chosen such that the constraints are incompatible, avoiding the trivial case where $\hat{\pi}_\theta^{\mathcal{L}}$, $\hat{\pi}_\theta^{H_2,\epsilon^\star}$ and $\hat{\pi}_\theta^{H_2,ml}$ all coincide.

Figure 8 illustrates in $\mathbb{S}^3$ the geometry behind the MLME. Black lines $w_1 = \tilde{f}_0^{(1)}$ and $w_3 = \tilde{f}_1^{(2)}$ correspond to the constraints, which do not intersect since they are incompatible. For this example, the NPMLE (orange dot on the boundary of $\mathbb{S}^3$, its second component being zero) is unique. All distributions that satisfy the minimally-relaxed constraints (*i.e.*, with $\epsilon = \epsilon^\star$) belong to the two gray areas, their intersection defining $\hat{\pi}_\theta^{H_2,\epsilon^\star}$ (the green dot, also on the boundary of $\mathbb{S}^3$). The dashed green line is the curve defined by $\hat{\pi}_\theta^{H_2,\epsilon}$ in $\mathbb{S}^3$ for $\epsilon \geq \epsilon^\star$, which has an accumulation point in the uniform distribution $w_1 = w_2 = w_3 = \frac{1}{3}$ as $\epsilon$ becomes sufficiently large for the uniform distribution to satisfy the constraints. Our estimator MLME is the point in this green curve at which the value of the likelihood is the largest, that is the highest level set of the likelihood function over $\mathbb{S}^3$ whose intersection with the green curve is a single point. The orange curve shows this level set, the contact point (red dot) being the MLME $\hat{\pi}_\theta^{H_2,ml}$.

The MLME estimator $\hat{\pi}_\theta^{H_2,ml}$ corresponds in general to an $\epsilon > \epsilon^\star$ in the constraints (13). In terms of vector $\mathbf{w}$ of probabilities of the elementary regions $E_k$, this set is a polytope $\mathcal{P}_\epsilon$, defined by its linear boundaries, which characterizes all solutions compatible with the data. One may notice that although the determination of its vertices is a difficult task, approximation of $\mathcal{P}_\epsilon$ by the maximum-volume interior ellipsoid is feasible at a reasonable computational cost [19], providing directly a lower bound on the volume of $\mathcal{P}_\epsilon$.

Figure 9 shows the proposed estimator $\hat{\pi}_\theta^{H_2,ml}$ for the same dataset as in Figure 6. Note that the distribution of the probability mass is much smoother than in Figure 6 and that the support of $\hat{\pi}_\theta^{H_2,ml}$ is now the entire $\Theta$. This example shows that the new estimator $\hat{\pi}_\theta^{H_2,ml}$ is able to exploit the dual characteristics of the ML and MaxEnt criteria to produce an estimate that is not too informative while still fitting the observed data reasonably well.

Two common measures of the difference between two distributions are the Kolmogorov and the total variation distances. The Kolmogorov distance $d_{\mathrm{K}}$ is the maximum value of the absolute

difference between the two cumulative distributions, while the total variation distance $d_{\mathrm{TV}}$ is the sum of all absolute differences [20]. Figure 10 addresses the performance of the estimation of the true probability law over $\mathcal{Q}$, showing box-plots of the Kolmogorov–Smirnov (left) and total variation (right) distances between $\pi_{\theta,\mathcal{Q}}$ and the NPMLE and the MLME estimates observed in 200 simulations, each for $N = 10^3$ observations. In each plot, the box in the left corresponds to the MaxEnt-NPMLE estimator $\hat{\pi}_\theta^{\mathcal{L}}$ and the one on the right to the the proposed estimator $\hat{\pi}_\theta^{H_2,ml}$. This clearly demonstrates the superiority of the estimator proposed in the paper. Note that the difference is more pronounced for the total variation, which is the criterion that best indicates the predictive power of the identified population model.

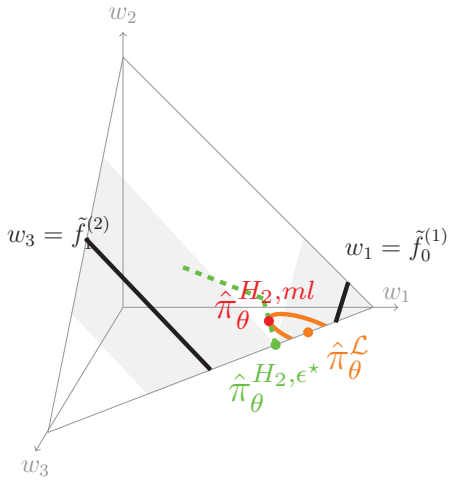



**Figure 8.** Illustration of the three proposed estimators in a simple example.

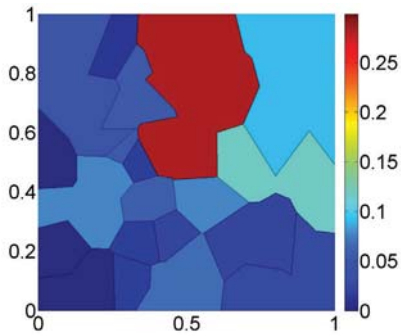



**Figure 9.** $\hat{\pi}_\theta^{H_2,ml}$.

Finally, Figure 11 shows the behavior under an increasing number of randomly-generated binary partitions. The total number of observations grows with $J$: $N = 100J$. The plots show the empirical average of the two Kullback–Leibler divergences $D(\cdot\|\pi_\theta)$ (Figure 11a) and $D(\pi_\theta\|\cdot)$ (Figure 11b)

over 100 randomly-generated datasets for each value of $J$, with $J$ varying from 10 to 100 in steps of 10. Here, the probability of "dangerous" partitions has been increased to $10^{-2}$, to guarantee a sufficient number of samples censored by them. Figure 11a suggests that $\hat{\pi}_\theta^{H_2,ml}$ may be consistent, which is strongly contradicted by the behavior observed for the NPMLE. The divergence $D(\pi_\theta||\hat{\pi}_\theta^{\mathcal{L}})$ was infinite in all simulations (due to $\hat{\pi}_\theta^{\mathcal{L}}(E_k) = 0$ for some $E_k \in \mathcal{Q}$) and, thus, cannot be presented in Figure 11b.

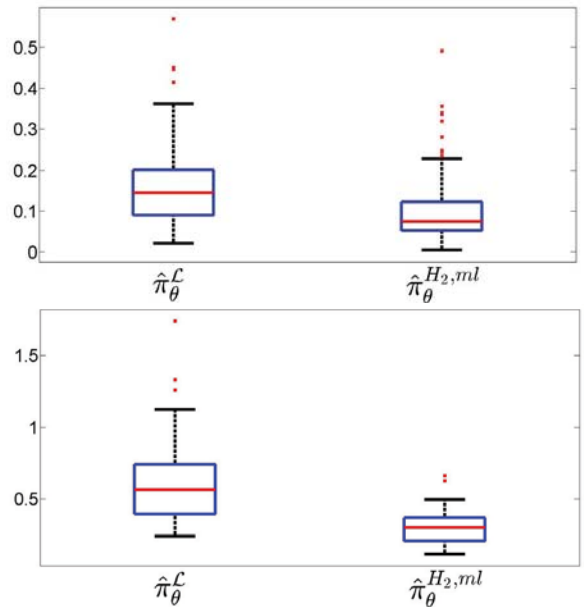
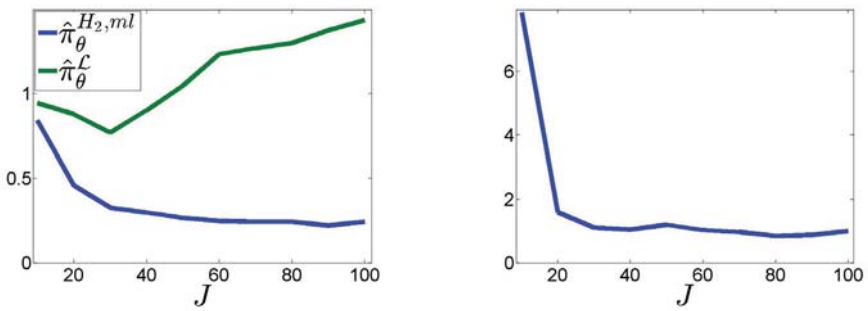



**Figure 10.** Box-plots of the Kolmogorov–Smirnov (Top) and total variation (Bottom) distances between $\pi_{\theta,\mathcal{Q}}$ and estimates $\hat{\pi}_\theta^{\mathcal{L}}$ and $\hat{\pi}_\theta^{H_2,ml}$ observed in 200 simulations.




**(a)** **(b)**

**Figure 11.** Kullback–Leibler divergence for an increasing number $J$ of partitions. **(a)** Empirical average of $D(\hat{\pi}_\theta^{H_2,ml}\|\pi_\theta)$ and $D(\hat{\pi}_\theta^{\mathcal{L}}\|\pi_\theta)$; **(b)** empirical average of $D(\pi_\theta\|\hat{\pi}_\theta^{H_2,ml})$.

## 4. Numerical Results

### *4.1. Application to a Real Problem: Modeling Decompression Sickness*

The density estimation problem studied in this paper has been motivated by a problem of population analysis in the context of the prevention of decompression sickness (DCS) in deep sea diving, which is known to be highly correlated with the presence of gas bubbles in the diver's blood. The ability to correctly predict the probability that this volume becomes exceedingly high can thus be exploited to establish safe diving rules, avoiding diving profiles (duration/depth) that may be dangerous for a non-negligible part of the population.

More precisely, we are interested in estimating the distribution $\pi_\theta$ of the biophysical parameters $\theta$ of a mathematical model [21] for the instantaneous volume $B(t)$ of micro-bubbles flowing through the right ventricle of a diver's heart when executing a decompression profile $P(t)$ (see Figure 12a):

$$(\theta, \{P(u)\}_{u\leq t}) \rightarrow B(t|\theta, \{P(u)\}_{u\leq t})) \tag{15}$$

Gas presence in the diver's circulatory system is only observed through "bubble grades", which are strongly quantified samples of $B(t)$ (the red horizontal lines in Figure 12b indicate the quantification levels $\tau = \{\tau_\ell\}_{\ell=1}^{L}$ applied to $B(t)$, represented by the blue curve). In our case $L = 4$, as shown in Figure 12b, and thresholds $\tau_0 = 0 < \tau_1 < \cdots < \tau_L < \tau_{L+1} = \infty$ are assumed known. Since it is usually accepted that DCS is related to the maximum observed grade, only the grade corresponding to the peak volume:

$$b(\theta, P) = \max_t B(t|\theta, \{P(u)\}_{u\leq t})$$

is retained, such that for each executed dive $D_n$ where (the known) profile $P_n$ has been followed by a diver with (unknown) bio-physical parameter $\theta_n$, a single grade measure $G_n$ is recorded:

$$G_n = \ell \Leftrightarrow b(\theta_n, P_n) \in [\tau_\ell, \tau_{\ell+1}[\,, \;\; \ell \in \{0, \ldots, L\} \tag{16}$$

In Figure 12, a simplified model with $\theta \in \Theta \subset \mathbb{R}^2$, with $\Theta$ the rectangular colored region in Figure 12c, has been used, all other parameters of model (15) being held fixed. Note that all biophysical parameters $\theta$ in region $R_n$:

$$R_n = R_n^{P_n} \equiv \{\theta \in \Theta \;:\; b(\theta, P_n) \in [\tau_{G_n}, \tau_{G_n+1}[\} \tag{17}$$

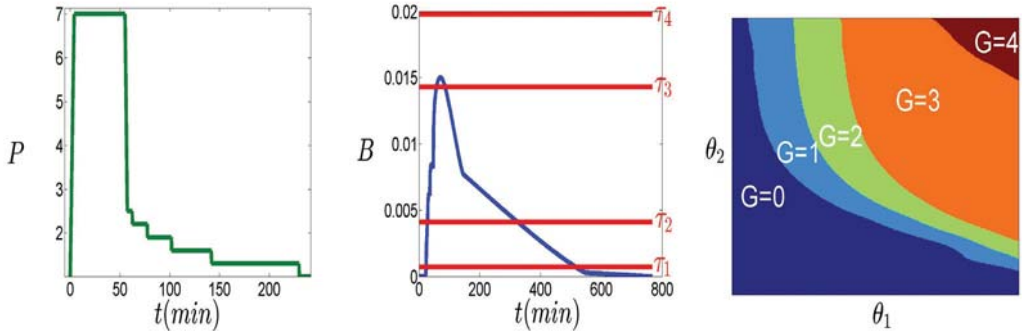
yield the same grade $G_n$ for all dives that use profile $P_n$. Each diving profile $P$ induces in this manner a partition $\mathcal{Q}^{(P)}$ of $\Theta$:

$$\Theta = \cup_\ell R_\ell^P, \qquad R_{\ell_1}^P \cap R_{\ell_2}^P = \emptyset, \ell_1 \neq \ell_2$$

Figure 12c displays the regions corresponding to the $L + 1 = 5$ possible grade values for the profile $P$ in Figure 12a. In this example, observation of a grade $G = 3$ indicates that the diver's biophysical parameters $\theta$ belong to the orange region.

The dataset available for this study contains records of the bubble grades observed over a total of $J = 19$ distinct profiles, repeated a number $n_j$ of times ranging from 12 to 41 (see Table 1; the most

dangerous profiles have been executed less often) and leads to the partition $\mathcal{Q}$ shown in Figure 13. We remark on the strong dispersion of the sizes of the elements of $\mathcal{Q}$ in this case, in particular the presence of very narrow regions that are contained in the elements of several partitions. The elements of $\mathcal{Q}$ have in this case strongly elongated shapes, markedly different from the partitions built from Voronoi cells used in the simulations of the previous sections.




**(a)** **(b)** **(c)**

**Figure 12.** Definition of bubble grades $G$ and regions $R_\ell^P$. (**a**) Diving profile $P(t)$; (**b**) blue: gas volume $B$; red: thresholds $\tau_\ell$; (**c**) regions corresponding to the $L+1=5$ bubble grades $G$.

**Table 1.** Number of experiments by profile in a real dataset.

| Profile | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n_j$ | 31 | 41 | 24 | 31 | 28 | 12 | 18 | 14 | 14 | 17 | 16 | 26 | 14 | 16 | 18 | 30 | 12 | 41 | 30 |

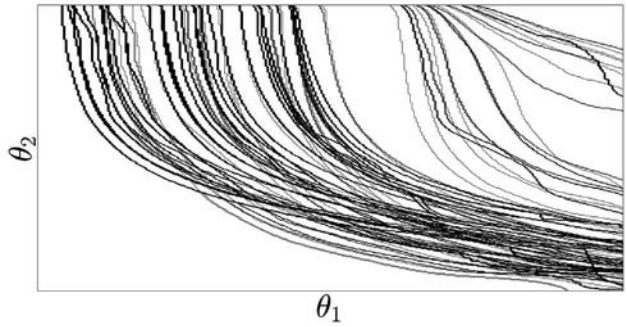
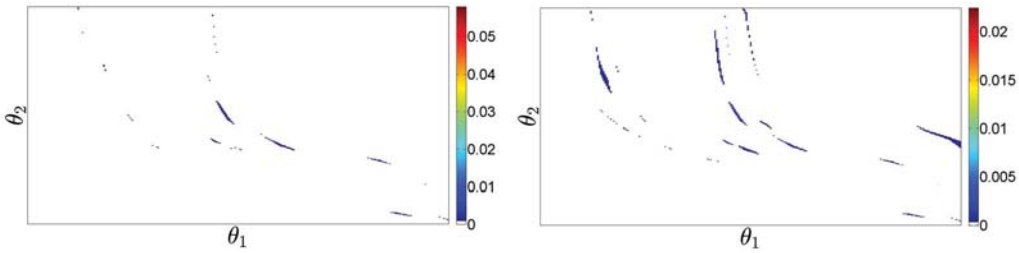



**Figure 13.** Partition induced by the 19 profiles in the real diving dataset.

4.1.1. Simulated Data

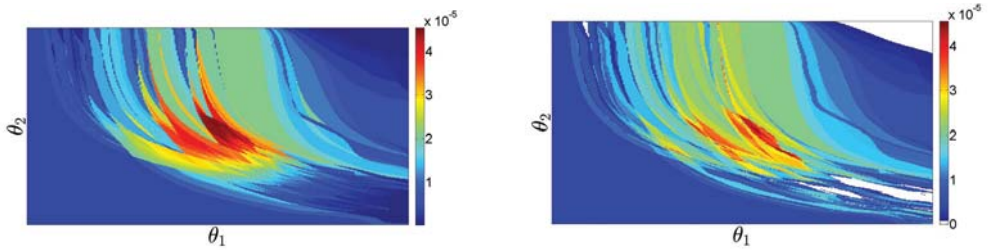
Before showing the results obtained in the real dataset of grade measures for the set of profiles available, we study the performance of the method proposed on the set of partitions corresponding to the set of observed profiles using simulated data. We considered the simulation of two normally and independent random variables restricted to a (biologically motivated) rectangular domain $\Theta$. We kept the same $n_j$ as shown in Table 1 and, thus, the same total $N = 433$.

Figures 14 and 15 show the results obtained with a total of $N = 10^4$ observations. The singularity of both $\hat{\pi}_\theta^{\mathcal{L}}$ and $\hat{\pi}_\theta^{H_2,\epsilon^\star}$, represented in Figure 14, is very strong in this case, the probability mass being concentrated in a subset of $\Theta$ of small Lebesgue measure. On the contrary, even for a partition of complex geometry like this one, the proposed MLME estimator (see Figure 15b) is able to overcome the shortcomings of the maximum-likelihood based estimates, producing an estimate that resembles the simulated law (in Figure 15a). The resulting population model has limited complexity while still retaining a superior predictive power, as is obvious from these plots.




**(a)** **(b)**

**Figure 14.** (**a**) Rényi-MaxEnt NPMLE $\hat{\pi}_\theta^{\mathcal{L}}$. (**b**) Rényi-MaxEnt $\hat{\pi}_\theta^{H_2,\epsilon^\star}$. White regions have zero probability mass.




**(a)** **(b)**

**Figure 15.** (**a**) Simulated distribution $\pi_{\theta,\mathcal{Q}}$. (**b**) MLME estimate $\hat{\pi}_\theta^{H_2,ml}$. White regions have zero probability mass.

Figure 16 shows the evolution of the likelihood along the curve $\hat{\pi}_\theta^{H_2,\epsilon}, \epsilon > \epsilon^\star$. We can see that the likelihood loss is larger than for the random partitions and that $\hat{\pi}_\theta^{H_2,ml}$ is obtained for $\epsilon \simeq \epsilon^\star$.

The larger likelihood loss can be explained by a smaller number of observations ($N = 433$ here, while for the previous simulation study $N = 10^4$) and also by the more irregular geometry of the partition $\mathcal{Q}$, with a large number of small elongated sets, which can produce over-optimistic values of the likelihood by concentrating mass over those sets.

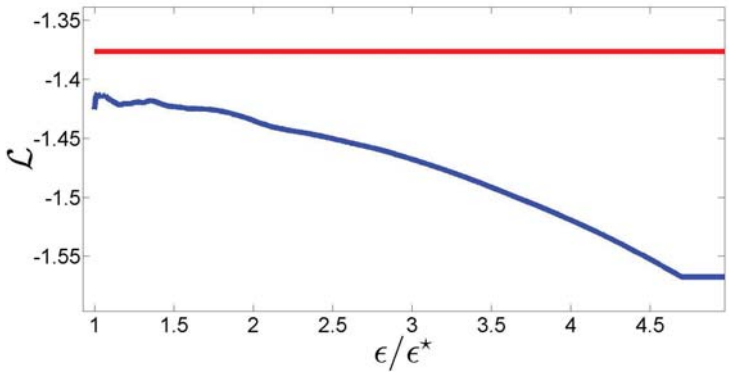



**Figure 16.** Variation of $\mathcal{L}(\hat{\pi}_\theta^{H_2,\epsilon})$ with $\epsilon/\epsilon^\star$. Red line: $\mathcal{L}(\hat{\pi}_\theta^{\mathcal{L}})$.

4.1.2. Real Data (Part of the Material in this Section Has Been Previously Presented in [22])

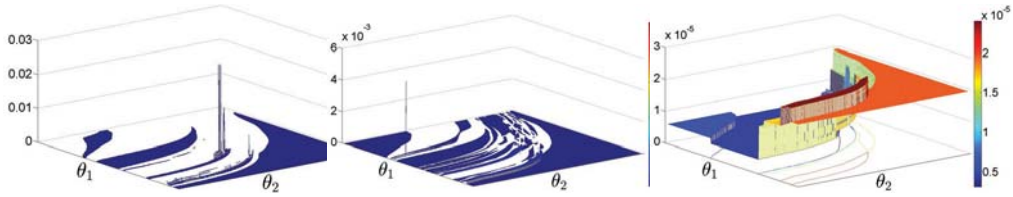
Figure 17 shows the densities obtained for the real dataset by the three different estimators discussed in the previous sections: the least informative NPMLE $\hat{\pi}_\theta^{\mathcal{L}}$, the minimally-regularized MaxEnt estimate $\hat{\pi}_\theta^{H_2,\epsilon^\star}$ and the new most likely MaxEnt estimate $\hat{\pi}_\theta^{H_2,ml}$. Analysis of this figure reveals the marked singularity of the first two estimates, which are highly concentrated in regions of small Lebesgue measure. The estimator proposed in the paper, $\hat{\pi}_\theta^{H_2,ml}$, leads to a much smoother solution, resembling $\pi_{\theta,\mathcal{Q}}$ and covering nearly all of the domain, which seems to provide a more natural model of a biological population than the solution found by the two other estimators.

For the dataset sizes of our study with $Q = 665$, we observed very fast convergence of (11) for the complete $\mathcal{Q}$ (35 iterations for $\delta = 10^{-4}$), confirming the applicability of the proposed algorithm.

We now assess the likelihood loss of our solution. Figure 18 shows the variation of $\mathcal{L}(\hat{\pi}_\theta^{H_2,\epsilon})$ with $\epsilon/\epsilon^\star$ for this real dataset. Compared to what we observed with random partitions in Figure 7, there is now a significant likelihood loss, the blue curve staying well below the maximum likelihood value for all values of the regularization parameter. This is natural, being an expected consequence of eventual misfits of the biophysical/classification model, which induce errors in the definition of the partitions $\mathcal{Q}^{(j)}$ associated with the distinct profiles $P^{(j)}$ and, thus, compromise the ability to closely fit the data.

Finally, we show, for this real dataset, the importance of accounting for the correlation of the empirical distributions in the constrained optimization problem. Figure 19 shows the estimates $\hat{\pi}_\theta^{H_2,\epsilon^\star}$ (top) and $\hat{\pi}_\theta^{H_2,ml}$ (bottom) obtained using the entire matrix $\Sigma$ (left) or just its diagonal elements

(right). We can see that simple independent relaxation of the empirical laws is not able to prevent the estimates from becoming highly concentrated, indicating an over-homogeneous population.




**(a)** **(b)** **(c)**

**Figure 17.** Estimates of $\pi_\theta$ with real dataset. (**a**) Least informative NPMLE $\hat{\pi}_\theta^{\mathcal{L}}$. (**b**) Rényi-MaxEnt $\hat{\pi}_\theta^{H_2,\epsilon^\star}$. (**c**) MLME $\hat{\pi}_\theta^{H_2,ml}$. White regions have zero probability mass.

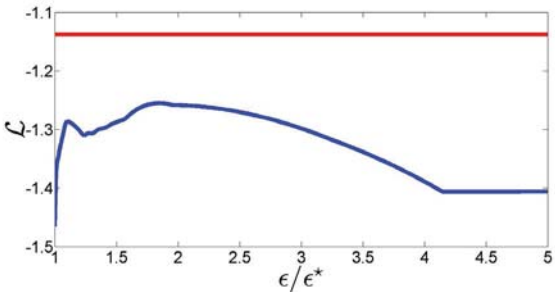
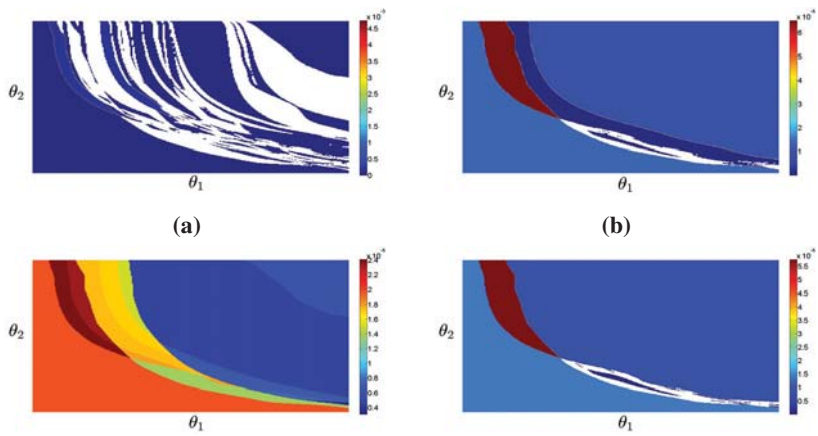



**Figure 18.** Variation of $\mathcal{L}(\hat{\pi}_\theta^{H_2,\epsilon})$ with $\epsilon/\epsilon^\star$. Red line: $\mathcal{L}(\hat{\pi}_\theta^{\mathcal{L}})$.




**(a)** **(b)**

**(c)** **(d)**

**Figure 19.** (**a**) $\hat{\pi}_\theta^{H_2,\epsilon^\star}$, full $\Sigma^{(j)^{-1/2}}$; (**b**) $\hat{\pi}_\theta^{H_2,\epsilon^\star}$, $diag(\Sigma^{(j)^{-1/2}})$; (**c**) $\hat{\pi}_\theta^{H_2,ml}$, full $\Sigma^{(j)^{-1/2}}$; (**d**) $\hat{\pi}_\theta^{H_2,ml}$, $diag(\Sigma^{(j)^{-1/2}})$.

*4.2. Assessing Predictive Power*

To assess the predictive power of the three estimators $\hat{\pi}_\theta^{\mathcal{L}}$, $\hat{\pi}_\theta^{H_2,\epsilon^\star}$ and $\hat{\pi}_\theta^{H_2,ml}$, we performed leave-one-out cross-validation, removing at each time all observations relative to one profile $P^{(j)}$ and computing the three estimators using the data for the remaining 18 profiles. We then compare the estimated and observed grades' frequencies $\tilde{f}^{(j)}$ for the retained profile.

Figure 20 represents boxplots of the total variation distance for each of the 19 profiles in the dataset, confirming the superiority of the estimator proposed for prediction purposes. In the sense of the total variation distance, $\hat{\pi}_\theta^{H_2,ml}$ is on average the closest distribution to the empirical frequencies in the dataset.

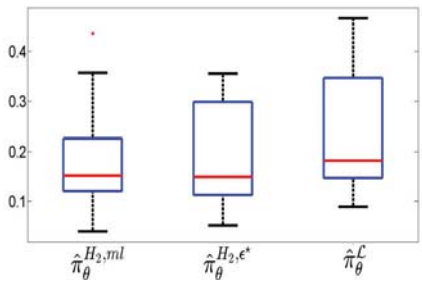



**Figure 20.** Boxplots of the total variation distance $d_{\mathrm{TV}}$ for the 19 datasets in the leave-on-out cross-validation study.

## 5. Conclusions

The paper studied the estimation of a probability density from region-censored observations, with application to the prevention of decompression sickness during hyperbaric diving. We show that the NPMLE is intrinsically ill-posed, leading to unstable solutions that are biologically implausible. Expressing counts of the censored observations as empirical means of a set of features, we derive the MaxEnt solution that best approximates the empirical distributions. The degree of fitting to the observed frequencies is chosen by selecting the MaxEnt solution that has the largest likelihood. The tests conducted show that the proposed most likely Rényi-MaxEnt estimator has superior behavior compared to the minimally-relaxed MaxEnt estimator, being able to approximate the observed dataset while at the same time being plausible as a description of a natural population. In particular, our numerical experiments show that our construction leads to a distribution estimate with good generalization properties, being able to predict grade probabilities for unseen profiles well, and can thus be used to detect profiles with a high risk of decompression sickness.

## Acknowledgments

This work has been partially funded by contract SAFE-DIVE (Rapid, DGA/DGCIS) 122906109 EJThe authors acknowledge the support of Julien Hugon and Axel Barbaud (BF Systèmes, France)

on the biophysical modeling of hyperbaric diving, as well as providing the dataset used for the study. They also thank the anonymous reviewers for their careful reading and constructive comments, which helped us to improve the quality of the manuscript.

**Author Contributions**

The work presented in this paper is carried out as part of the Ph.D. Thesis of Youssef Bennani, supervised by Luc Pronzato and Maria João Rendas. All authors have read and approved the final manuscript.

**Conflicts of Interest**

The authors declare no conflict of interest.

**References**

1. Kaplan, E.L.; Meier, P. Nonparametric estimation from incomplete observations. *J. Am. Stat. Assos.* **1958**, *53*, 457–481.
2. Turnbull, B.W. The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. R. Stat. Soc. Ser. B (Methodol.)* **1976**, *38*, 290–295.
3. Gentleman, R.; Vandal, A.C. Computational algorithms for censored-data problems using intersection graphs. *J. Comput. Graph. Stat.* **2001**, *10*, 403–421.
4. Bennani, Y. *Intersection Graph for Region-Censored Data*; Rapport de recherche I3S; I3S: Sophia-Antipolis, France, 2013.
5. Jaynes, E.T. Information theory and statistical mechanics. *Phys. Rev.* **1957**, *106*, 620–630.
6. Della Pietra, S.; Della Pietra, V.; Lafferty, J. Inducing features of random fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **1997**, *19*, 380–393 .
7. Grechuk, B.; Molyboha, A.; Zabarankin, M. Maximum entropy principle with general deviation measures. *Math. Oper. Res.* **2009**, *34*, 445–467.
8. Dudik, M. Maximum Entropy Density Estimation and Modeling of Geographic Distributions of Species. Ph.D. Thesis, Princeton University, Princeton, NJ, USA, 2007.
9. Dudik, M.; Phillips, M.; Schapire, R. Performance guarantees for regularized maximum entropy density estimation. In Proceedings of the 17th Annual Conference on Computational Learning Theory, Banff, AL, Canada, 1–4 July 2004.
10. Járai-Szabó, F.; Néda, Z. On the size-distribution of Poisson Voronoi cells. *Physica A* **2007**, *385*, 518–526.
11. Liu, X. Nonparametric Estimation With Censored Data: A Discrete Approach. Ph.D. Thesis, McGill University, Montreal, QC, Canada, 2005.
12. Groeneboom, P.; Wellner, J.A. *Information Bounds and Nonparametric Maximum Likelihood Estimation*; Birkhauser Verlag: Basel, Switzerland, 1992.
13. Böhning, D.; Schlattmann, P.; Dietz, E. Interval censored data: A note on the nonparametric maximum likelihood estimator of the distribution function. *Biometrika* **1996**, *83*, 462–466, .

14. Fish, D.; Brinicombe, A.; Pike, E.; Walker, J. Blind deconvolution by means of the Richardson–Lucy algorithm. *JOSA A* **1995**, *12*, 58–65.
15. Fedorov, V.V. *Theory of Optimal Experiments*; Academic Press: New York, NY, USA, 1972.
16. Silvey, S.D.; Titterington, D.H.; Torsney, B. An algorithm for optimal designs on a finite design space. *Commun. Stat.-Theor. M.* **1978**, *7*, 1379–1389.
17. Torsney, B. A moment inequality and monotonicity of an algorithm. In *Semi-Infinite Programming and Applications*; Springer: Berlin, Germany, 1983; pp. 249–260.
18. Harman, R.; Pronzato, L. Improvements on removing nonoptimal support points in D-optimum design algorithms. *Stat. Probab. Lett.* **2007**, *77*, 90–94.
19. Khachiyan, L.G.; Todd, M.J. On the complexity of approximating the maximal inscribed ellipsoid for a polytope. *Math. Program.* **1993**, *61*, 137–159.
20. Strassen, V. The existence of probability measures with given marginals. *Ann. Math. Stat.* **1965**, *36*, 423–439.
21. Hugon, J. Vers Une Modélisation Biophysique De La Décompression. Ph.D. Thesis, Université Aix Marseille, Aix-en-Provence, France, 22 November 2010.
22. Bennani, Y.; Pronzato, L.; Rendas, M.J. Nonparametric density estimation with region-censored data. In Proceedings of the 22nd European Signal Processing Conference (EUSIPCO), Lisbon, Portugal, 1–5 September 2014; pp. 1098–1102.

# General Hyperplane Prior Distributions Based on Geometric Invariances for Bayesian Multivariate Linear Regression

**Udo von Toussaint**

**Abstract:** Based on geometric invariance properties, we derive an explicit prior distribution for the parameters of multivariate linear regression problems in the absence of further prior information. The problem is formulated as a rotationally-invariant distribution of $L$-dimensional hyperplanes in $N$ dimensions, and the associated system of partial differential equations is solved. The derived prior distribution generalizes the already known special cases, e.g., 2D plane in three dimensions.

Reprinted from *Entropy*. Cite as: von Toussaint, U. General Hyperplane Prior Distributions Based on Geometric Invariances for Bayesian Multivariate Linear Regression. *Entropy* **2015**, *17*, 3898–3912.

## 1. Introduction

In the context of Bayesian probability theory, a proper assignment of prior probabilities is crucial. Depending on the domain, quite different prior information can be available. It may be in the form of point estimates provided by domain experts (see, e.g., [1] for prior distribution elicitation) or in the form of invariances (of the prior knowledge) of the system of interest, which should be reflected in the prior probability density [2]. However, especially for the ubiquitous case of the estimation of parameters of linear equation systems (like a straight line or hyperplane fitting), the latter requirement is often violated. Consider, for concreteness, the simple case of $y = ax$, a straight line through the origin, with $a$ the parameter of interest. Here, the commonly-applied prior is constant, $p\,(a \mid I) = \mathrm{const.}$, often accompanied by statements like "Since we do not have specific prior information, we chose a uniform prior on $a$...". In Figure 1 on the left-hand side, 15 random samples generated from this prior distribution with $a \in [0, 50]$ are displayed. Confronted with this result, the typical response is (at least in the experience of the author) that instead, a more "uniform" prior distribution of the slopes was intended, which is often depicted like in Figure 1 on the right-hand side. This plot was generated from a prior distribution that has an equal probability density for the angle of the line to the abscissa, corresponding to

$$p\,(a \mid I) \sim \frac{1}{\left(1 + a^2\right)^{3/2}}. \tag{1}$$

Additionally, in fact, in practice, the units of the axes are commonly chosen in such a way that extreme values of the slopes are not *a priori* overrepresented. If we generalize this requirement to more than one independent or dependent variable, then the desired prior probability should be invariant under arbitrary rotations in this parameter space. Some important special cases have been given already in [3], e.g., for a 1D line in two dimensions or a 2D plane in three dimensions. There also, the governing transformation invariance equation underlying invariant priors is derived. These special cases have since then been generalized to invariant priors for $(N-1)$-dimensional hyperplanes in $N$-dimensional space; see, e.g., [4]. These hyperplane priors proved to be valuable for Bayesian neural networks [5], where the specific properties of the prior density favored node-pruning instead

of simple edge pruning of standard (quadratic) weight regularizers. This is especially helpful for a Bayesian approach to fully-connected deep convolutional networks; see e.g., [6,7].

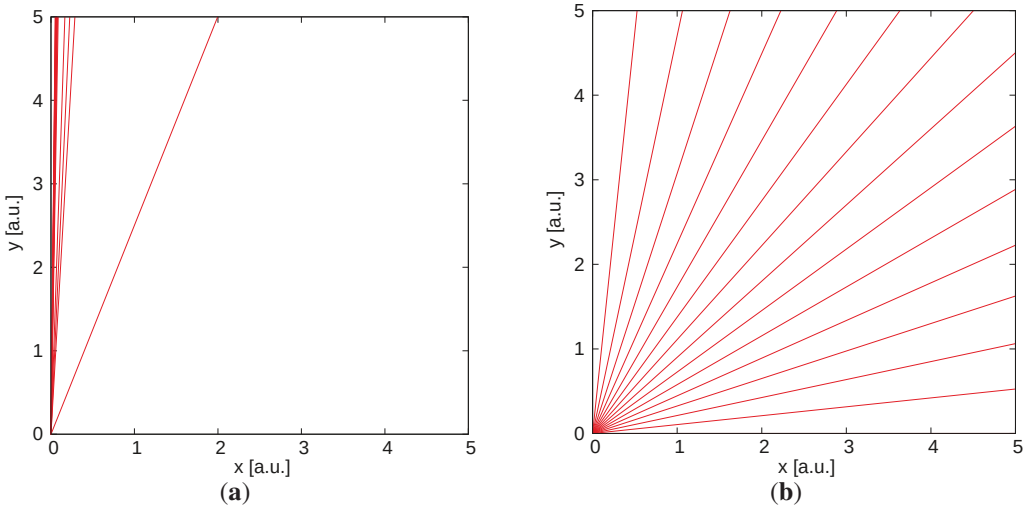



**Figure 1.** Comparison of two different priors. (**a**) 15 random samples drawn from $p\,(a \mid I) = 1/50$, *i.e.*, a uniform distribution in the slope with $0 \leq a \leq 50$. (**b**) the density $p\,(a \mid I) \sim (1 + a^2)^{-3/2}$, corresponding to a distribution uniform in the angle, is visualized by 15 samples.

Nevertheless, the general case of prior probability densities for $L$-dimensional hyperplanes in $N$-dimensions ($N > L$) in a suitable parameterization has not been available so far. It has even been conjectured that it is impossible to derive a general solution [8]. Luckily, this conjecture has been too pessimistic, and an explicit formula for the prior density, which can directly be applied to linear regression problems, is derived below.

It should be pointed out that multivariate regression is of course a longstanding topic in Bayesian inference. with classical contributions, e.g., by Box and Tiao [9], Zellner [10] or West [11]. However, the standard approach is based on the use of conjugate priors (instead of invariance priors), mostly for computational convenience [12]. In contrast, the subsequently derived prior distribution is determined by the basic desideratum of consistency if the available prior information is invariant under the considered transformations (*i.e.*, rotations). Whether this invariance holds depends on the considered problem and must not be assumed without further consideration (similar to the case of flat priors for the coefficients). For example, the assumption of rotation invariance may not be suitable for covariates with different underlying units (e.g., $\mathrm{m}^2$, kg).

## 2. Problem Statement

In standard notation, a multivariate regression model is notated as follows:

$$\mathbf{y}_i = \mathbf{A}\mathbf{x}_i + \mathbf{t}, \qquad x_i \in \mathbb{R}^L, \mathbf{A} \in \mathbb{R}^{M\times L}, \mathbf{t} \in \mathbb{R}^M \text{ and } y_i \in \mathbb{R}^M, \tag{2}$$

with:

$$\mathbf{z}_i = \mathbf{y}_i + \epsilon_i, \epsilon_i \in \mathbb{R}^M, \tag{3}$$

where $\mathbf{z}_i$ is the response vector, $\mathbf{y}_i$ the model value vector, $\mathbf{x}_i$ the vector of the $L$ covariates for observation $i$, $\mathbf{t}$ the intercept vector and $\mathbf{A}$ the $M \times L$-dimensional matrix of adjacent regression coefficients. The observation noise $\epsilon_i$ of each data point is often considered as Gaussian distributed, $\epsilon_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. This regression model can also be considered as estimating the "best" $L$-dimensional hyperplane in an $N$-dimensional space, because in an $N$-dimensional space, an $L$-dimensional hyperplane is given by:

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1L}x_L + t_1 \\ y_2 &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2L}x_L + t_2 \\ y_3 &= a_{31}x_1 + a_{32}x_2 + \cdots + a_{3L}x_L + t_3 \\ &\vdots \\ y_M &= a_{M1}x_1 + a_{M2}x_2 + \cdots + a_{ML}x_L + t_M \end{aligned} \tag{4}$$

with $M = N - L$.

The quantity of interest is the prior probability density $F(\mathbf{A}) = F(a_{11}, \cdots, a_{ML}, t_1, \cdots, t_M|I)$ for the coefficients $a_{11}, \cdots, a_{ML}, t_1, \cdots, t_M$, which remains invariant under translations and rotations of the coordinate system.

## 3. Derivation

Using the transformation invariance equation derived in [3]:

$$\sum_{i=1}^{N} \frac{\partial}{\partial z_i} \left(F(z_1, \cdots, z_N)\, g_i(z_1, \cdots, z_N)\right) = 0 \tag{5}$$

for infinitesimal transformations of the form $z_i^{'} = z_i + \epsilon g_i(z_1, \cdots, z_N)$, we can establish a partial differential equation system for $F$.

### *3.1. Invariance under Translations*

Let us first consider a translation with respect to $y_i : y_i^{'} = y_i + \epsilon$, *i.e.*, $g_i = 1, g_{j,j\neq i} = 0$. Then, the equation in the primed variables reads:

$$y_i^{'} = y_i + \epsilon = a_{i1}^{'}x_1 + a_{i2}^{'}x_2 + \cdots + a_{iL}^{'}x_L + t_i^{'} \tag{6}$$

Collecting the coefficients yields $t_i^{'} = t_i + \epsilon$, and therefore, Equation (5) results in:

$$0 + \cdots + 0 + \frac{\partial}{\partial t_i} \left( F \left( \mathbf{A}, \mathbf{t} \right) \cdot 1 \right) + 0 + \cdots + 0 = 0, \tag{7}$$

which holds for any $i$. Therefore, $F\left(\mathbf{A}, \mathbf{t}\right)$ can be a function of $\vec{a}$ only. Since $F\left(\mathbf{A} \mid I\right)$ does not depend on $\vec{t}$, the prior distribution is improper (not normalizable in $\mathbf{t}$) as long as there are no limits on the magnitude of $\mathbf{t}$.

The translation with respect to $x_i$ results in the same conclusion.

### *3.2. Invariance under Rotations*

The general rotation in $n$-dimensional space may be expressed as a sequence of rotations around rotation axes, which are perpendicular to the planes spanned by appropriately-chosen pairs of coordinate system basis vectors [13]. This is based on the fact that any orthogonal matrix, *i.e.*, rotation matrices, can be written uniquely as a product of $2 \times 2$ rotations. To avoid convoluted language, we denote in the following the rotation around the rotation axis that is perpendicular to the plane spanned by the linear combination of the basis vectors $e_i$ and $e_j$ simply as rotation in the $x_i x_j$-plane.

#### 3.2.1. Rotation in the $x_i x_j$-Plane

Now, we perform one such infinitesimal $2 \times 2$-rotation for independent parameters around an arbitrary rotation axis perpendicular to the plane spanned by $e_i$ and $e_j$, preserving all other coordinates: $x_k^{'} = x_k \quad \forall \quad k \neq (j, i)$ and

$$x_i^{'} = x_i - \epsilon x_j, \tag{8}$$

$$x_j^{'} = \epsilon x_i + x_j. \tag{9}$$

Substituting the primed coordinates into Equation (5) yields the implied transformations:

$$a_{ki}^{'} = a_{ki} - a_{kj}\epsilon, \tag{10}$$

$$a_{kj}^{'} = a_{kj} + a_{ki}\epsilon, \tag{11}$$

$$t_k^{'} = t_k \tag{12}$$

and, therefore, the partial differential equation:

$$\sum_{k=1}^{M} \frac{\partial}{\partial a_{ki}} \left( F \left( \mathbf{A} \right) \cdot \left( -a_{kj} \right) \right) + \sum_{k=1}^{M} \frac{\partial}{\partial a_{kj}} \left( F \left( \mathbf{A} \right) \cdot \left( a_{ki} \right) \right) = 0. \tag{13}$$

#### 3.2.2. Rotation in the $y_i y_j$-Plane

Now, we perform one such rotation in the plane of two dependent parameters $e_i$ and $e_j$; thus $y_k^{'} = y_k \quad \forall \quad k \neq (j, i)$ and:

$$y_i^{'} = y_i - \epsilon y_j, \tag{14}$$
$$y_j^{'} = \epsilon y_i + y_j. \tag{15}$$

Substituting the primed coordinates into Equation (5) yields the implied transformations:

$$a_{ik}^{'} = a_{ik} - a_{jk}\epsilon, \tag{16}$$
$$a_{jk}^{'} = a_{jk} + a_{ik}\epsilon, \tag{17}$$
$$t_i^{'} = t_i - t_j\epsilon, \tag{18}$$
$$t_j^{'} = t_j + t_i\epsilon, \tag{19}$$
$$t_k^{'} = t_k \tag{20}$$

and, therefore, the partial differential equation:

$$\sum_{k=1}^{M} \frac{\partial}{\partial a_{ik}} \left(F\left(\mathbf{A}\right) \cdot \left(-a_{jk}\right)\right) + \sum_{k=1}^{M} \frac{\partial}{\partial a_{jk}} \left(F\left(\mathbf{A}\right) \cdot \left(a_{ik}\right)\right) = 0. \tag{21}$$

3.2.3. Rotation in a Plane Spanned by $x_i y_j$-Axes

Performing a rotation in the xy-plane, we obtain:

$$x_i^{'} = x_i - \epsilon y_j, \tag{22}$$
$$y_j^{'} = \epsilon x_i + y_j. \tag{23}$$

which yields (see the Appendix):

$$a_{ji}^{'} = a_{ji} + \left(1 + a_{ji}^2\right)\epsilon, \tag{24}$$
$$a_{kl}^{'} = a_{kl} + \left(a_{jl}a_{ki}\right)\epsilon, \tag{25}$$
$$t_k^{'} = t_k + \left(a_{ki}t_j\right)\epsilon \tag{26}$$

and therefore:

$$\sum_{k=1}^{M}\sum_{l=1}^{L} \frac{\partial}{\partial a_{kl}} \left(F \cdot \left(a_{jl}a_{ki}\right)\right) + \frac{\partial}{\partial a_{ji}} F + F \cdot a_{ji} = 0. \tag{27}$$

## 4. The PDE System

The translation invariance of Equation (5) excludes a dependence of $F$ on $t_1, \cdots, t_M$, so $F$ is of the form $F\left(a_{11}, \cdots, a_{ML} | I\right)$. Rotation invariance with respect to the y-axis requires $F$ to fulfill the homogeneous, linear system of first order partial differential equations $\left(i, j \in [1, M], i \neq j\right)$ (*i.e.*, Equation (21)):

$$\sum_{k=1}^{L} \frac{\partial}{\partial a_{jk}} \left(F \cdot a_{ik}\right) - \sum_{k=1}^{L} \frac{\partial}{\partial a_{ik}} \left(F \cdot a_{jk}\right) = 0 \tag{28}$$

and similar for rotations around the x-axis $\left(i, j \in [1, L], i \neq j\right)$ (Equation (13)):

$$\sum_{k=1}^{M} \frac{\partial}{\partial a_{kj}} \left(F \cdot a_{ki}\right) - \sum_{k=1}^{M} \frac{\partial}{\partial a_{ki}} \left(F \cdot a_{kj}\right) = 0. \tag{29}$$

Rotations around an axis perpendicular to a plane given by an x,y-pair require the probability distribution to obey also ($i \in [1, L], j \in [1, M]$):

$$\sum_{k=1}^{M}\sum_{l=1}^{L}\frac{\partial}{\partial a_{kl}}\left(F\cdot\left(a_{jl}a_{ki}\right)\right)+\frac{\partial}{\partial a_{ji}}F+F\cdot a_{ji}=0. \tag{30}$$

Using the product rule, the double sum can be rewritten as

$$\sum_{k=1}^{M}\sum_{l=1}^{L}\frac{\partial}{\partial a_{kl}}\left(F\cdot\left(a_{jl}a_{ki}\right)\right)=\sum_{k=1}^{M}\sum_{l=1}^{L}a_{jl}a_{ki}\frac{\partial}{\partial a_{kl}}F+F\cdot\sum_{k=1}^{M}\sum_{l=1}^{L}\frac{\partial}{\partial a_{kl}}\left(a_{jl}a_{ki}\right) \tag{31}$$

and the last term of the previous equation can be split into three parts and simplified:

$$\begin{aligned}&F\cdot\sum_{k=1}^{M}\sum_{l=1}^{L}\frac{\partial}{\partial a_{kl}}\left(a_{jl}a_{ki}\right)=\\&F\cdot\sum_{k=1,k\neq j}^{M}\frac{\partial}{\partial a_{ki}}\left(a_{ji}a_{ki}\right)+F\cdot\sum_{l=1,l\neq i}^{L}\frac{\partial}{\partial a_{jl}}\left(a_{jl}a_{ji}\right)+F\cdot\frac{\partial}{\partial a_{ji}}a_{ji}^{2}=\\&\left(M-1\right)a_{ji}F+\left(L-1\right)a_{ji}F+2a_{ji}F=\left(M+L\right)a_{ji}F.\end{aligned} \tag{32}$$

Using this, Equation (30) can be written as:

$$\sum_{k=1}^{M}\sum_{l=1}^{L}a_{jl}a_{ki}\frac{\partial}{\partial a_{kl}}F+\frac{\partial}{\partial a_{ji}}F+\left(M+L+1\right)a_{ji}F=0. \tag{33}$$

## 5. Solution

This system of PDEs (Equations (28), (29) and (33)) can be tackled with the theory of Lie groups, which provides a systematic, though algebraically-intensive solution strategy, which is implemented in contemporary computer algebra systems. The solutions of several test cases computed by the Maple computer algebra system (http://www.maplesoft.com/) (it proved to be superior to MATHEMATICA (www.http://www.wolfram.com/mathematica/) for the present PDE-systems) led to the conjecture that a general solution to this PDE system is given by the sum of the squares of all possible minors of the coefficient matrix:

$$\begin{aligned}&F\left(a_{11},\cdots,a_{ML}|I\right)\\&=\left[1+\sum_{k=1}^{\binom{M}{P}\binom{L}{P}}\left(\det\left(A^{P,k}\right)\right)^{2}+\sum_{k=1}^{\binom{M}{P-1}\binom{L}{P-1}}\left(\det\left(A^{P-1,k}\right)\right)^{2}+\cdots+\sum_{k=1}^{\binom{M}{1}\binom{L}{1}}\left(\det\left(A^{1,k}\right)\right)^{2}\right]^{-\frac{M+L+1}{2}}\end{aligned} \tag{34}$$

where $A^n$ denotes a submatrix (minor) of size $n\times n$ (this notation is used at various places throughout the paper and should not be confused with the power of a matrix, which does not occur in this paper) and $P=\mathrm{Min}\left(M,L\right)$. Equation (34) does not appear unreasonable from the onset as prior density, because it preserves the underlying symmetry of the problem (permutation invariance of the parameters) and it is non-negative.

An explicit example for the case $N = 4, L = 2$ is:

$$F\left(a_{11}, a_{12}, a_{21}, a_{22}|I\right) = \left[1 + a_{11}^2 + a_{12}^2 + a_{21}^2 + a_{22}^2 + \left(a_{11} \cdot a_{22} - a_{12} \cdot a_{21}\right)^2\right]^{-5/2}. \tag{35}$$

A two-dimensional slice of this probability density is given in Figure 2. The high symmetry of the prior distribution with respect to parameter permutations results in similar, "Cauchy"-like shapes if slices along other parameter axis are displayed.

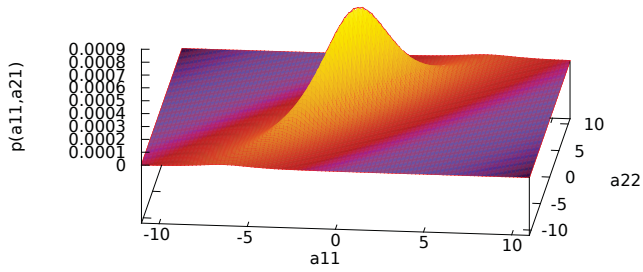



**Figure 2.** Probability density of $p\left(a_{11}, a_{21} \mid a_{12}, a_{22}, I\right)$ for $a_{12} = 3$ and $a_{22} = 5$ for the case $N = 4, L = 2$. The probability density exhibits the typical "Cauchy-"like shape with heavy tails compared to a binormal distribution. Due to the symmetry of the prior distribution, slices with respect to the other parameters display the same basic features.

For the case $N = 6, L = 3$, the solution is given by:

$$\begin{aligned}
& F\left(a_{11}, \cdots, a_{33}|I\right) \\
= \; & \left(1 + a_{11}^2 + a_{12}^2 + a_{13}^2 + a_{21}^2 + a_{22}^2 + a_{23}^2 + a_{31}^2 + a_{32}^2 + a_{33}^2 + \right. \\
& \left(a_{22}a_{33} - a_{23}a_{32}\right)^2 + \left(a_{21}a_{33} - a_{23}a_{31}\right)^2 + \left(a_{21}a_{32} - a_{22}a_{31}\right)^2 + \\
& \left(a_{12}a_{33} - a_{13}a_{32}\right)^2 + \left(a_{11}a_{33} - a_{13}a_{31}\right)^2 + \left(a_{11}a_{32} - a_{12}a_{31}\right)^2 + \\
& \left(a_{12}a_{23} - a_{13}a_{22}\right)^2 + \left(a_{11}a_{23} - a_{13}a_{21}\right)^2 + \left(a_{11}a_{22} - a_{12}a_{21}\right)^2 + \\
& \left.\left(a_{11} \cdot \left(a_{22}a_{33} - a_{23}a_{32}\right) - a_{12} \cdot \left(a_{21}a_{33} - a_{23}a_{31}\right) + a_{13} \cdot \left(a_{21}a_{32} - a_{22}a_{31}\right)\right)^2\right)^{-7/2}.
\end{aligned}$$

## 6. Proof

### *6.1. Preliminaries*

To prove that Equation (34) fulfills the equation system given by Equations (28), (29) and (33), we verify directly that Equation (34) solves the PDEs.

We will make repeated use of the Laplace expansion of determinants:

$$\det\left(A^n\right) = \sum_{j=1}^{n} a_{ij} \left(-1\right)^{i+j} \det\left(M_{ij}^{n-1}\right) \tag{36}$$

where the minor $M_{ij}^{n-1}$ is the $(n-1)\times(n-1)$-matrix derived from the $n\times n$-matrix $A^n$ by deletion of the $i$-th row and $j$-th column (by definition $M^0 := 1$). The cofactor matrix $A_{ij}^{n-1}$ is defined to be:

$$A_{ij}^{n-1} = (-1)^{i+j} M_{ij}^{n-1} \tag{37}$$

and satisfies the following $n$-equations $(i,j,k = 1,2,\cdots,N)$:

$$\sum_{j=1}^{n} a_{ij}\det\left(A_{kj}^{n-1}\right) = \delta_{ik}\det\left(A^n\right), \quad \sum_{i=1}^{n} a_{ij}\det\left(A_{ik}^{n-1}\right) = \delta_{jk}\det\left(A^n\right). \tag{38}$$

Further useful is the following form of the Laplace expansion, taking into account index shifts of a previous deletion of row k and column i of an $(n+1)$-matrix $A^{n+1}$, resulting in the minor $M_{ki}^n$:

$$\det\left(M_{ki}^n\right) = \sum_{l=1,l\neq i}^{n+1} a_{jl}\,(-1)^{(l'+j')}\det\left(M_{(jk)(li)}^{n-1}\right) \tag{39}$$

where $M_{(jk)(li)}^{n-1}$ is the minor given by deletion of the $j$-th and $k$-th row and the $l$-th and $i$-th column. $l'$ and $j'$ are defined as:

$$\begin{aligned} l' = l \quad \forall \quad (l<i) \quad \text{and} \quad l' = l-1 \quad \forall \quad (l>i) \\ j' = j \quad \forall \quad (j<k) \quad \text{and} \quad j' = j-1 \quad \forall \quad (j>k). \end{aligned} \tag{40}$$

In the following, we face the problem of possibly too heavy of a nomenclature, because we need summation indices, while we also need to keep track of the original indices underlying the entries in the minors, where some rows and columns have been deleted, although the relative order is preserved. The mapping could be expressed, e.g., as $a_{i(i')j(j')}$ with $i', j' \in [1,m]$ and $i\,(:) \in [1,M]$ and $j\,(:) \in [1,L]$. To avoid this cumbersome notation, we implicitly assume from now on (up to the Conclusion Section) this mapping for all summations that are indexed by either $k$ or $l$. Therefore:

$$\sum_{k=1}^{m} a_{ki}\det\left(A_{kj}^{m-1}\right) \quad \text{has to be read as} \quad \sum_{k'=1}^{m} a_{k(k')i}\det\left(A_{k(k')j}^{m-1}\right). \tag{41}$$

### *6.2. $x_ix_j$- and $y_iy_j$-Rotations*

We now verify that Equation (34) solves Equation (29). It is obvious that only those determinants of Equation (34) that contain column $i$ or column $j$ have the potential to provide non-zero contributions in Equation (29): if column $j$ is missing, the derivative in the first term is zero. If, instead, column $i$ is missing, then the derivative in the second term of Equation (29) yields zero. To proceed, we introduce $H(\mathbf{A})$ via:

$$F(\mathbf{A}) = H(\mathbf{A})^{-\frac{L+M+1}{2}}. \tag{42}$$

It is noteworthy that $H(\mathbf{A})$ has a very simple form: it is given by a sum of positive terms. This almost decouples the problem, and we can largely proceed on a term-by-term basis. Using the equality:

$$\frac{\partial}{\partial a_{pq}}\left(\det\left(A^m\right)\right)^2 = 2\det\left(A^m\right)\det\left(A_{pq}^{m-1}\right) \tag{43}$$

the left-hand side of Equation (29) transforms to $(i, j \in [1, L], i \neq j)$:

$$-\left(M+L+1\right) H\left(a\right)^{-\frac{L+M+3}{2}} \det\left(A^m\right) \cdot \left(\sum_{k=1}^{m} a_{ki} \det\left(A_{kj}^{m-1}\right) - \sum_{k=1}^{m} a_{kj} \det\left(A_{ki}^{m-1}\right)\right) \tag{44}$$

and using Equation (38), we obtain:

$$-\left(M+L+1\right) H\left(a\right)^{-\frac{L+M+3}{2}} \det\left(A^m\right) \cdot \left(\delta_{ij} \det\left(A^m\right) - \delta_{ij} \det\left(A^m\right)\right) = 0 \tag{45}$$

and, therefore, Equation (34) solves Equation (29). The calculation is similar for Equation (28) and yields the result that Equation (34) solves also the system Equation (28).

### *6.3.* $(x_i y_j)$*-Rotations*

The verification of the successful solution of Equation (33) by Equation (34) requires some more steps. As before, Equation (33) can be written as:

$$-\frac{M+L+1}{2}\left(\sum_{k=1}^{M}\sum_{l=1}^{L} a_{jl}a_{ki} H\left(\mathbf{A}\right)^{-\frac{L+M+3}{2}} \frac{\partial H\left(\mathbf{A}\right)}{\partial a_{kl}} + H\left(\mathbf{A}\right)^{-\frac{L+M+3}{2}} \frac{\partial H\left(\mathbf{A}\right)}{\partial a_{ji}} - 2a_{ji} H\left(\mathbf{A}\right)^{-\frac{L+M+1}{2}}\right) = 0 \tag{46}$$

and after multiplication with $H\left(\mathbf{A}\right)^{\frac{L+M+3}{2}}$ as:

$$\begin{aligned} & -\left(M+L+1\right)\cdot \\ & \left(\sum_{m=1}^{P}\sum_{r=1}^{\binom{M}{m}\binom{L}{m}} \left(\sum_{k=1}^{m}\sum_{l=1}^{m} a_{jl}a_{ki} \det\left(A^{m,r}\right) \det\left(A_{kl}^{m-1,r}\right) + \det\left(A^{m,r}\right) \det\left(A_{ji}^{m-1,r}\right)\right) - a_{ji} H(\mathbf{A})\right) \\ = \; & 0 \end{aligned} \tag{47}$$

#### 6.3.1. Matrices with Either Row j or Column i

The inner double sum can be simplified for all matrices containing either row $j$ or column $i$ (*i.e.*, all matrices of size $P \times P$ and all matrices $A^{m,r}$ of size $m \times m, m \in (1, 2, \cdots, P-1)$ with label $r = 1, 2, \cdots, \binom{M}{m}\binom{L}{m} - \binom{M-1}{m}\binom{L-1}{m}$) using the Laplace expansion (here, the expansion with respect to row $j$ is shown):

$$\begin{aligned} & \sum_{k=1}^{m}\sum_{l=1}^{m} a_{jl}a_{ki} \det\left(A^{m,r}\right) \det\left(A_{kl}^{m-1,r}\right) \\ = \; & \det\left(A^{m,r}\right) \sum_{k=1}^{m} a_{ki} \sum_{l=1}^{m} a_{jl} \det\left(A_{kl}^{m-1,r}\right) \\ = \; & \det\left(A^{m,r}\right) \sum_{k=1}^{m} a_{ki}\delta_{jk} \det\left(A^{m,r}\right) = a_{ji} \left(\det\left(A^{m,r}\right)\right)^2 \end{aligned} \tag{48}$$

which cancels the corresponding determinant of $H\left(\mathbf{A}\right)$ in the last term of Equation (48).

6.3.2. Matrices with Neither Row $j$ nor Column $i$

The basic idea is to show that $\binom{M-1}{m}\binom{L-1}{m}$-matrices with neither row $j$ nor column $i$, $m \in (1, 2, \cdots, P-1)$, cancel with the contributions of the corresponding matrices including row $j$ and column $i$ of size $(m+1) \times (m+1)$ of the second term.

Please note that there is a one-to-one correspondence of minors of size $m \times m$ without the $j$-th row and $i$-th column and the matrices of size $(m+1) \times (m+1)$ with row $j$ and column $i$ in the second term, therefore allowing one to label both with the same index $r$. After division by $-(M+L+1)$, the remaining terms of Equation (48) are (taking into account that the labeling of the rows and columns of the matrices of size $(m+1) \times (m+1)$ and $(m) \times (m)$ must be consistent):

$$\sum_{k=1,k\neq j}^{m+1} \sum_{l=1,l\neq i}^{m+1} a_{jl}a_{ki}\det\left(A_{ji}^{m,r}\right) \det\left(A_{(jk)(il)}^{m-1,r}\right) + \det\left(A^{m+1,r}\right) \det\left(A_{ji}^{m,r}\right) - a_{ji}H_{ji}(\mathbf{A}) = 0 \quad (49)$$

with $H_{ji}$ now only containing determinants with neither row $j$ nor column $i$. If we now only consider the relevant term of $H_{ji}$, we can write:

$$\sum_{k=1,k\neq j}^{m+1} \sum_{l=1,l\neq i}^{m+1} a_{jl}a_{ki}\det\left(A_{ji}^{m,r}\right) \det\left(A_{(jk)(il)}^{m-1,r}\right) + \det\left(A^{m+1,r}\right) \det\left(A_{ji}^{m,r}\right) - a_{ji}\det\left(A_{ji}^{m,r}\right)^2 = 0. \quad (50)$$

The equation is trivially true if $\det\left(A_{ji}^{m,r}\right) = 0$; otherwise, we can divide by $\det\left(A_{ji}^{m,r}\right)$ and obtain:

$$\sum_{k=1,k\neq j}^{m+1} \sum_{l=1,l\neq i}^{m+1} a_{jl}a_{ki}\det\left(A_{(jk)(il)}^{m-1,r}\right) + \det\left(A^{m+1,r}\right) - a_{ji}\det\left(A_{ji}^{m,r}\right) = 0. \quad (51)$$

Replacing the various cofactors by the corresponding minors (*cf.* Equations (36) and (37)) yields:

$$\sum_{k=1,k\neq j}^{m+1} \sum_{l=1,l\neq i}^{m+1} a_{jl}a_{ki}\,(-1)^{(i+j+k'+l')} \det\left(M_{(jk)(il)}^{m-1,r}\right) + \det\left(A^{m+1,r}\right) - a_{ji}\,(-1)^{(i+j)} \det\left(M_{ji}^{m,r}\right) = 0 \quad (52)$$

and after replacing $\det\left(A^{m+1,r}\right)$ by its Laplace expansion together with multiplication by $(-1)^{(i+j)}$, the equation reads:

$$\sum_{k=1,k\neq j}^{m+1} \sum_{l=1,l\neq i}^{m+1} a_{jl}a_{ki}\,(-1)^{(k'+l')} \det\left(M_{(jk)(il)}^{m-1,r}\right) + (-1)^{(i+j)} \sum_{k=1} a_{ki}\,(-1)^{(i+k)} \det\left(M_{ki}^{m,r}\right) - a_{ji}\det\left(M_{ji}^{m,r}\right) = 0 \quad (53)$$

and can be simplified to:

$$\sum_{k=1,k\neq j}^{m+1} \sum_{l=1,l\neq i}^{m+1} a_{jl}a_{ki}\,(-1)^{(k'+l')} \det\left(M_{(jk)(il)}^{m-1,r}\right) + \sum_{k=1} a_{ki}\,(-1)^{(j+k)} \det\left(M_{ki}^{m,r}\right) - a_{ji}\det\left(M_{ji}^{m,r}\right) = 0 \quad (54)$$

because $(-1)^{2i}$ equals one in the second term. Therefore, the third term cancels with the second term for $k = j$, and the remaining equation is given by:

$$\sum_{k=1,k\neq j}^{m+1} a_{ki}\,(-1)^{k'} \sum_{l=1,l\neq i}^{m+1} a_{jl}\,(-1)^{l'} \det\left(M^{m-1,r}_{(jk)(il)}\right) + \sum_{k=1,k\neq j} a_{ki}\,(-1)^{(j+k)} \det\left(M^{m,r}_{ki}\right) = 0. \quad (55)$$

Using Equation (39) together with the definition Equation (40), the inner sum of the first term of Equation (55) can be rewritten as:

$$\sum_{l=1,l\neq i}^{m+1} a_{jl}\,(-1)^{l'} \det\left(M^{m-1,r}_{(jk)(il)}\right) = (-1)^{j}\, M^{m,r}_{ki} \quad \forall\ (j<k) \quad (56)$$

and:

$$\sum_{l=1,l\neq i}^{m+1} a_{jl}\,(-1)^{l'} \det\left(M^{m-1,r}_{(jk)(il)}\right) = (-1)^{j-1}\, M^{m,r}_{ki} \quad \forall\ (j>k)\,. \quad (57)$$

Splitting the summation over k into two parts $(k<j)$ and $(k>j)$ and inserting the definition for $k'$, we obtain:

$$\begin{aligned} \sum_{k=1,k<j} a_{ki}\,(-1)^{(j+k)} \det\left(M^{m,r}_{ki}\right) + \sum_{k>j} a_{ki}\,(-1)^{(j+k)} \det\left(M^{m,r}_{ki}\right) \quad + \\ \sum_{k=1,k<j} a_{ki}\,(-1)^{((j-1)+k)} \det\left(M^{m,r}_{ki}\right) + \sum_{k>j} a_{ki}\,(-1)^{(j+(k-1))} \det\left(M^{m,r}_{ki}\right) \quad = \quad 0 \end{aligned} \quad (58)$$

where the first two terms cancel the last two terms.

Summarizing the previous approach, we have shown that for an arbitrary $n \times n$-determinant, the first and third term of Equation (48) almost cancel. Only determinants not containing the $j$-th row and the $i$-th column remain. These remaining contributions are canceled by the $(n+1)$-order determinant (required to contain the matrix element $a_{ji}$) of the second term in Equation (48). This schema can be repeated down to $n = 1$, and the last step $(n = 0)$ is easily explicitly calculated. This finishes our derivation.

## 7. Relation to Previously-Derived Special Cases

The underlying equation systems of the special case of an $(n-1)$-dimensional hyperplane in an $n$-dim space used in [8] and in this paper differ slightly due to a different parameterization, and therefore, the derived priors appear on first glance to be different, although they are identical, as will be shown below.

For probability density functions in different coordinate systems, the following equation holds:

$$p\left(\vec{a}\right) d\vec{a} = p\left(\vec{b}\left(\vec{a}\right)\right) \left|\frac{\partial\left(\vec{b}\right)}{\partial\left(\vec{a}\right)}\right| d\vec{a}, \quad (59)$$

where $|\cdots|$ denotes the absolute value of the Jacobi determinant:

$$\left|\frac{\partial\left(\vec{b}\right)}{\partial\left(\vec{a}\right)}\right| = \left|\det\begin{pmatrix} \frac{\partial b_1}{\partial a_1} & \frac{\partial b_1}{\partial a_2} & \cdots & \frac{\partial b_1}{\partial a_n} \\ \vdots & & & \vdots \\ \frac{\partial b_n}{\partial a_1} & \frac{\partial b_n}{\partial a_2} & \cdots & \frac{\partial b_n}{\partial a_n} \end{pmatrix}\right|. \quad (60)$$

The equation describing the (n-1)-dim hyperplane in an n-dim space in this paper is given by:

$$y_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1(n-1)}x_{n-1} + t_1 \tag{61}$$

and results in the following prior:

$$p\left(a_{11}, a_{12}, \cdots, a_{1(n-1)}, t_1\right) = \left(1 + \sum_{i=1}^{n-1} a_{1i}^2\right)^{-\frac{n+1}{2}}. \tag{62}$$

In [8], the corresponding hyperplane equation reads:

$$0 = b_1x_1 + b_2x_2 + \cdots + b_nx_n + 1 \tag{63}$$

with prior distribution:

$$p\left(b_1, b_2, \cdots, b_n\right) = \left(\sum_{i=1}^{n} b_i^2\right)^{-\frac{n+1}{2}}, \text{with} \quad \sum_{i=1}^{n} b_i^2 > R_0^2. \tag{64}$$

The latter constraints yield a proper (normalizable) prior. The relation of the two different parameterizations is given by:

$$b_i = \frac{a_{1i}}{t_1} \quad \forall i \neq n \qquad \text{and} \qquad b_n = -\frac{1}{t_1} \tag{65}$$

which yields the Jacobian:

$$\left|\frac{\partial\left(\vec{b}\right)}{\partial\left(\vec{a}\right)}\right| = \left|\det\begin{pmatrix} \frac{1}{t_1} & 0 & 0 & \cdots & 0 & -\frac{a_{11}}{t_1} \\ 0 & \frac{1}{t_1} & 0 & \cdots & 0 & -\frac{a_{12}}{t_1} \\ \vdots & & \ddots & \cdots & & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{t_1} & -\frac{a_{1(n-1)}}{t_1} \\ 0 & 0 & 0 & \cdots & 0 & -\frac{1}{t_1^2} \end{pmatrix}\right| = \frac{1}{t_1^{n+1}} \tag{66}$$

Using this result and Equation (65), we can write:

$$p\left(\vec{b}\left(\vec{a}\right)\right)\left|\frac{\partial\left(\vec{b}\right)}{\partial\left(\vec{a}\right)}\right| d\vec{a} = \frac{1}{\left(\sum_{i=1}^{n-1}\left(\frac{a_{1i}}{t_1}\right)^2 + \frac{1}{t_1^2}\right)^{\frac{n+1}{2}} t_1^{n+1}} d\vec{a} = \frac{1}{\left(1 + \sum_{i=1}^{n-1} a_{1i}^2\right)^{\frac{n+1}{2}}} d\vec{a} \tag{67}$$

which shows the equivalence of the two priors (Equations (62) and (64)). The requirement of $\sum b_i^2 > R_0^2$ leads to:

$$R_0^2 \leq \sum_{i=1}^{n} b_i^2 = \sum_{i=1}^{n-1}\left(\frac{a_{1i}}{t_1}\right)^2 + \frac{1}{t_1^2} = \frac{1}{t_1^2}\left(1 + \sum_{i=1}^{n-1} a_{1i}^2\right). \tag{68}$$

In the case of all $a_{1i} = 0$, we obtain:

$$t_1^2 \leq \frac{1}{R_0^2} \tag{69}$$

which means that the lower limit $R_0^2$ corresponds to an upper limit of $t_1^2$.

## 8. Practical Hints

In the worst case, the hyperplane prior has an exponentially-increasing number of determinants with increasing dimension. The total number of individual determinants for an $N$-dimensional plane in a $2N$-dimensional space is given by:

$$\sum_{k=0}^{N} \binom{N}{k}^2 = \binom{2N}{N} \tag{70}$$

which is already $70$ for a 4D hyperplane in an 8D space. Therefore, it is advantageous to compute the determinants using iteratively the Laplace expansion, starting from small determinants, storing the determinants of the previous step. This requires the storage of at most $\binom{N}{N/2}^2$ terms. As a proposal density for Markov chain Monte Carlo (MCMC) sampling methods (e.g., rejection sampling), the dominating multivariate Cauchy distribution is a good candidate. Source code for the set up of the PDE system and for the solution, together with a Maple script for the verification of the solution, can be obtained from the author.

## 9. Conclusions

This paper has derived a prior density for $L$-dimensional hyperplanes in $N$-dimensional space, based on geometric invariances. It is suited, e.g., to parameter estimation of multilinear regression problems in the absence of further prior knowledge or Bayesian model estimation for neural networks. In the latter case, the prior has to be made proper by suitable restriction of the range of the offset parameters, which depends on domain knowledge. The obtained prior density avoids the too strong weight of "large" values of the regression coefficients typically assigned by uniform priors. Being a rational function, its influence on the parameter estimates on standard problems with Gaussian uncertainties (resulting in an exponential likelihood) on the data will be limited. However, this can be different for robust estimation approaches with heavy-tailed likelihood distributions.

## Appendix

In this section, the relation between the primed coefficient $a'_{nm}$ and the unprimed coefficient $a_{nm}$ is derived. A rotation perpendicular to the $x_i y_j$-plane relates $x_i, y_j$ with $x'_i, y'_j$ by:

$$x'_i = x_i - \epsilon y_j, \tag{A1}$$
$$y'_j = \epsilon x_i + y_j. \tag{A2}$$

and $x'_k = x_k, k = 1, \cdots, L; k \neq i$ and $y'_k = y_k, k = 1, \cdots, M; k \neq j$. Using this, the system Equation (5) in the transformed coordinate system reads $(n = 1, \cdots, M; n \neq j)$:

$$\begin{aligned} y_n &= a'_{n1}x_1 + a'_{n2}x_2 + \cdots + a'_{ni}(x_i - \epsilon y_j) + a'_{n(i+1)}x_{(i+1)} + \cdots + a'_{nL}x_L + t'_n \\ y_j + \epsilon x_i &= a'_{j1}x_1 + a'_{j2}x_2 + \cdots + a'_{ji}(x_i - \epsilon y_j) + a'_{j(i+1)}x_{(i+1)} + \cdots + a'_{jL}x_L + t'_j. \end{aligned} \tag{A3}$$

Solving for $y_j$, we obtain:

$$y_j = \frac{1}{1 + a_{ji}^{'}\epsilon}\left(t_j^{'} - x_i\epsilon + \sum_{k=1}^{L} a_{jk}^{'} x_k\right) \tag{A4}$$

and subsequently:

$$y_n = \left(t_n^{'} + \sum_{k=1}^{L} a_{nk}^{'} x_k\right) - a_{ni}^{'}\epsilon\frac{1}{1 + a_{ji}^{'}\epsilon}\left(t_j^{'} - x_i\epsilon + \sum_{k=1}^{L} a_{jk}^{'} x_k\right). \tag{A5}$$

Using the Taylor expansion $1/\left(1 + a_{ji}^{'}\epsilon\right) = 1 - a_{ji}^{'}\epsilon + O\left(\epsilon^2\right)$ up to first order and collecting the coefficients, the previous equations yield:

$$\begin{aligned} a_{ji} &= a_{ji}^{'} - a_{ji}^{'\,2}\epsilon - \epsilon \quad ; \quad t_j = t_j^{'} - t_j^{'} a_{ji}^{'}\epsilon \\ a_{nk} &= a_{nk}^{'} - a_{ni}^{'} a_{jk}^{'}\epsilon \quad ; \quad t_n = t_n^{'} - t_j^{'} a_{ni}^{'}\epsilon. \end{aligned} \tag{A6}$$

First, we solve for $a_{ji}^{'}$:

$$a_{ji}^{'\,2}\epsilon - a_{ji}^{'} + \epsilon + a_{ji} = 0 \rightarrow a_{ji}^{'} = \frac{1 - \sqrt{1 - 4\epsilon\left(\epsilon + a_{ji}\right)}}{2\epsilon} = \underline{a_{ji} + \left(1 + a_{ji}^{\,2}\right)\epsilon} + O\left(\epsilon^2\right) \tag{A7}$$

and next for $a_{jk}^{'}$:

$$\begin{aligned} a_{jk} = a_{jk}^{'}\left(1 - a_{ji}^{'}\epsilon\right) \rightarrow \\ a_{jk}^{'} &= \frac{a_{jk}}{1 - a_{ji}^{'}\epsilon} = a_{jk}\left(1 + a_{ji}^{'}\epsilon\right) + O\left(\epsilon^2\right) \\ &= \underline{a_{jk}\left(1 + a_{ji}\epsilon\right)} + O\left(\epsilon^2\right). \end{aligned} \tag{A8}$$

A similar calculation for $a_{ni}^{'}$ yields:

$$a_{ni}^{'} = \underline{a_{ni}\left(1 + a_{ji}\epsilon\right)} + O\left(\epsilon^2\right) \tag{A9}$$

which then allows one to compute $a_{nk}^{'}$ for index pairs with $\{nk\} \neq \{ji\}$:

$$a_{nk}^{'} = a_{nk} + \left(a_{ni} + a_{ni}a_{ji}\epsilon\right)\left(a_{jk} + a_{jk}a_{ji}\epsilon\right)\epsilon = \underline{a_{nk} + a_{ni}a_{jk}\epsilon} + O\left(\epsilon^2\right). \tag{A10}$$

The offset variable $t_j$ is given by:

$$\begin{aligned} t_j = t_j^{'}\left(1 - a_{ji}^{'}\epsilon\right) \rightarrow \\ t_j^{'} &= \frac{t_j}{1 - a_{ji}^{'}\epsilon} = t_j\left(1 + a_{ji}^{'}\epsilon\right) + O\left(\epsilon^2\right) \\ &= \underline{t_j\left(1 + a_{ji}\epsilon\right)} + O\left(\epsilon^2\right). \end{aligned} \tag{A11}$$

and the other offset variables $t_n$ by:

$$t_n^{'} = t_n + \left(a_{ni} + a_{ni}a_{ji}\epsilon\right)\left(t_j + t_j a_{ji}\epsilon\right)\epsilon = \underline{t_n + a_{ni}t_j\epsilon} + O\left(\epsilon^2\right) \tag{A12}$$

which concludes the derivation of Equations (24)–(26).

## Conflicts of Interest

The authors declare no conflict of interest.

## References


1. Gosling, J.P.; Oakley, J.E.; O'Hagan, A. Nonparametric elicitation for heavy-tailed prior distributions. *Bayesian Anal.* **2007**, *2*, 693–718.
2. Jaynes, E.T. Prior Probabilities. *IEEE Trans. Syst. Sci. Cybern.* **1968**, *SSC4*, 227–241.
3. Kendall, M.; Moran, P. *Geometrical Probability*; Griffin: London, UK, 1963.
4. Von der Linden, W.; Dose, V.; von Toussaint, U. *Bayesian Probability Theory: Application to the Physical Sciences*, 1st ed.; Cambridge University Press: Cambridge, UK, 2014.
5. Von Toussaint, U.; Gori, S.; Dose, V. Bayesian Neural-Networks-Based Evaluation of Binary Speckle Data. *Appl. Opt.* **2004**, *43*, 5356–5363.
6. Hinton, G.; Salakhutdinov, R. Reducing the Dimensionality of Data with Neural Networks. *Science* **2006**, *313*, 504–507.
7. Minh, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; *et al.* Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533.
8. Dose, V. Hyperplane Priors. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*; Williams, C.J., Ed.; American Institute of Physics: Melville, NY, USA, 2003; Volume AIP Conference Proceedings 659, pp. 350–357.
9. Box, G.E.P.; Tiao, G.C. *Bayesian Inference in Statistical Analysis*; Wiley: New York, NY, USA, 1992; Reprint from 1973.
10. Zellner, A. *An Introduction to Bayesian Inference in Econometrics*; Wiley: New York, NY, USA, 1971.
11. West, M. Outlier Models and Prior Distributions in Bayesian Linear Regression. *J. R. Stat. Soc. B* **1984**, *46*, 431–439.
12. O'Hagan, A. *Kendall's Advanced Theory of Statistics, Bayesian Inference*, 1st ed.; Arnold Publishers: New York, NY, USA, 1994; Volume 2B.
13. Landau, L.; Lifschitz, E. *Lehrbuch der Theoretischen Physik I*, 1st ed.; Akademie Verlag: Berlin, Germany, 1962.

# A New Robust Regression Method Based on Minimization of Geodesic Distances on a Probabilistic Manifold: Application to Power Laws

**Geert Verdoolaege**

**Abstract:** In regression analysis for deriving scaling laws that occur in various scientific disciplines, usually standard regression methods have been applied, of which ordinary least squares (OLS) is the most popular. In many situations, the assumptions underlying OLS are not fulfilled, and several other approaches have been proposed. However, most techniques address only part of the shortcomings of OLS. We here discuss a new and more general regression method, which we call geodesic least squares regression (GLS). The method is based on minimization of the Rao geodesic distance on a probabilistic manifold. For the case of a power law, we demonstrate the robustness of the method on synthetic data in the presence of significant uncertainty on both the data and the regression model. We then show good performance of the method in an application to a scaling law in magnetic confinement fusion.

Reprinted from *Entropy*. Cite as: Verdoolaege, G. A New Robust Regression Method Based on Minimization of Geodesic Distances on a Probabilistic Manifold: Application to Power Laws. *Entropy* **2015**, *17*, 4602–4626.

## 1. Introduction

Regression analysis is an essential instrument for data analysis in numerous branches of science. It is used for investigating deterministic relations between variables, for model building and for prediction by extrapolation to a previously unseen range of the involved variables. In this paper, we focus on regression analysis applied to the estimation of scaling laws. In various scientific disciplines, such as astronomy, biology, ecology and geology, scaling laws are used to characterize the underlying mechanisms at work in the respective complex systems under study. In general, a scaling law describes how a quantity of interest $y$ scales when changing other quantities $x_1, x_2, \ldots, x_P$, on which it depends. Scaling laws are often expressed in terms of a power law:

$$y = \beta_0 x_1^{\beta_1} x_2^{\beta_2} \ldots x_P^{\beta_P}. \tag{1}$$

A crucial property of such a power law is scale-invariance, *i.e.*, when multiplying any of the variables $x_i$ by a constant $a$, the power law in Equation (1) essentially remains the same, being multiplied only by a constant $a^{\beta_i}$.

In nuclear fusion experiments based on magnetic confinement of a hot hydrogen plasma, scaling laws are crucial for predicting the performance of future fusion reactors, which will have a larger size, magnetic field, plasma density, *etc*., compared to present-day experimental devices [1]. These scaling laws can be estimated on the basis of datasets from multiple fusion devices, spanning a significant part of the parameter space. Ordinary least squares regression (OLS) combined with

frequentist theory is the statistical workhorse that is employed for this purpose in the vast majority of cases. However, often, there is considerable uncertainty in the experimental data, including the predictor variables, and in the model equations (regression model). However, OLS only deals with uncertainty on the response variables and does not cover additional complications, including atypical observations (outliers), heteroscedasticity, correlations, non-Gaussian distributions, *etc*. As such, OLS regression is often unsuitable for deriving scaling laws [2,3], and many scientific fields could benefit greatly from a unified regression methodology that is flexible and robust and yet relatively simple to implement.

In order to be able to handle the complications mentioned above, we have developed a new regression method, called geodesic least squares regression (GLS). It is based on minimization of the Rao geodesic distance between probability distributions on a manifold equipped with the Fisher metric. In this paper, we briefly introduce the method by means of a simple example involving a power law and Gaussian noise. We show the good performance of the method on synthetic data, introducing outliers in the first case and studying the effect of a logarithmic transformation of the data in the second case. Finally, we present an application to the important scaling concerning the power threshold for the transition into the high confinement regime (H-mode) in nuclear fusion experiments based on magnetic plasma confinement. The details of the quantities involved in this scaling, their experimental determination and the underlying physics are not important for the purpose of this paper. Rather, we here aim at showing the performance of GLS on a challenging and heterogeneous real-life dataset.

The paper is structured as follows. The method of geodesic least squares regression is described in Section 2, including a short discussion on calculating geodesic distances on a Gaussian probabilistic manifold, within the framework of information geometry. The next section, Section 3, briefly introduces the database that is used in the subsequent regression experiments, in relation to the scaling law for the H-mode power threshold in fusion plasmas. The experiments involving synthetic data are described in Section 4, while the real power threshold scaling is derived in Section 5. Section 6 concludes the paper and contains an outlook towards future work related to the methodology.

## 2. Geodesic Least Squares Regression

We start by describing the GLS methodology, which was already introduced in [4,5], but here, we go into some more detail. We describe a specific form of GLS regression, and it should be stressed that various aspects can be generalized, as will be noted accordingly. Furthermore, several elements on which GLS is based are also found in other regression techniques. The strength of GLS regression is that it integrates many of these aspects in an elegant way, resulting in a method that is very general, flexible and robust. From one point of view, GLS is similar to a class of parameter estimation methods that are collectively referred to in the statistics community as minimum distance estimation, in that GLS minimizes a distance between a parametric model distribution of the data and an empirical distribution [6]. We use the Rao geodesic distance (GD) as a similarity measure, which

has the advantage that it offers an intuitive geometric interpretation. In addition, there are similarities between GLS and the generalized linear model [7].

We will consider the case of regression with multiple predictor variables (regressors) and a single response variable. For this case, we will show that GLS regression can be regarded as a generalization of OLS. However, GLS takes place on a probabilistic manifold, whereas classic OLS operates in a flat Euclidean space. Indeed, OLS is based on minimizing the difference, *i.e.*, the Euclidean distance, between the predicted and measured values of the response variable. Likewise, GLS is based on minimizing the GD between distributions on the probabilistic manifold. Therefore, we start by briefly introducing some concepts from information geometry related to distance calculation.

### *2.1. Distance in Information Geometry*

In information geometry, a parametric family of probability densities is interpreted as a Riemannian differentiable manifold [8]. Each point on the manifold corresponds to a specific probability density function (pdf) within the family, and the family parameters represent a coordinate system on the manifold. The Fisher information (covariance of the score) provides a unique metric tensor. For a probability model $p(\{x_m\}|\{\theta^k\})$ [9] describing a set $\{x_m\}$ of $M$ variables ($m = 1, \ldots, M$), parameterized by a set $\{\theta^k\}$ of $P$ parameters ($k = 1, \ldots, P$), the entries $g_{ij}$ of the Fisher information matrix are given by (no summation):

$$\begin{aligned} g_{ij}\big(\{\theta^k\}\big) &= \mathbb{E}\left[\frac{\partial}{\partial\theta^i}\ln p\big(\{x_m\}|\{\theta^k\}\big)\frac{\partial}{\partial\theta^j}\ln p\big(\{x_m\}|\{\theta^k\}\big)\right], \qquad i,j,k = 1,\ldots,P. \\ &= -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^i\partial\theta^j}\ln p\big(\{x_m\}|\{\theta^k\}\big)\right] \end{aligned}$$

The metric provides the basis for distance measurement between pdfs. Specifically, a geodesic curve locally minimizes the distance between two points on the manifold equipped with that metric. Through calculus of variations, it can be shown that a geodesic is the solution of the following system of nonlinear second-order ordinary differential equations, known in the language of variational analysis as Euler–Lagrange equations [10] and in the present context as geodesic equations:

$$\ddot{\theta}^r(t) + \sum_{i,j=1}^{P} \Gamma^r{}_{ij}\,\dot{\theta}^i(t)\dot{\theta}^j(t) = 0, \qquad r = 1,\ldots,P. \tag{2}$$

Here, the $\theta^i$ have been parameterized along the geodesic by $t$ and $\Gamma^r{}_{ij}$ are the Christoffel symbols of the second kind, defined through the metric as:

$$\Gamma^k{}_{ij} = \frac{1}{2}\sum_r g^{kr}\left(\frac{\partial g_{jr}}{\partial\theta^i} + \frac{\partial g_{ir}}{\partial\theta^j} - \frac{\partial g_{ij}}{\partial\theta^r}\right),$$

where $g^{ij}$ denotes the components of the inverse metric. The boundary value problem Equation (2) needs to be solved assuming the known values of the coordinates at the boundary points of the geodesic.

From the metric and the solution of the geodesic equations, the length $L_g$ of the geodesic curve between two distributions with parameter sets $\{\theta_1^i\}$ and $\{\theta_2^i\}$, *i.e.*, the geodesic distance between these distributions, may be locally calculated as follows (assuming $t$ runs from zero to one):

$$L_g = \int_{\{\theta_1^i\}}^{\{\theta_2^i\}} \mathrm{d}s = \int_0^1 \left( \sum_{i,j} g_{ij} \dot{\theta}^i \dot{\theta}^j \right)^{1/2} \mathrm{d}t, \tag{3}$$

where $s$ represents the arc length. In the framework of information geometry, the geodesic distance based on the Fisher metric is often referred to as the Rao geodesic distance (GD).

Coming back to Equations (2) and (3), it should be noted that closed-form expressions for the GD are rarely available. On the other hand, provided the Fisher metric can be calculated relatively easily, the framework of information geometry is very useful, since straightforward approximations of the geodesic curves can be found in a geometrically intuitive way [11]. This intuitive approach by means of geometry is an important and attractive aspect of the theory, as it provides enhanced insight into various concepts and algorithms in probability theory and statistics [12]. This is also the case for GLS, as will be demonstrated below. Furthermore, as far as the GD is concerned, visualization of geodesics may guide controlled approximations to the geodesic paths and geodesic distances [11].

Besides the attractive feature of providing intuitive geometrical insight into problems involving similarity measurement between probability distributions, the GD has several more advantages over other similarity measures for distributions. First, it is a distance measure (a metric) in the strict sense of the word. As a result, it is symmetric in its arguments, a desirable property for measuring the similarity between two given states of information in terms of probability distributions. In addition, it obeys the triangle inequality, yielding various practical advantages, for instance in the field of data retrieval from large databases [13]. Furthermore, closed-form expressions may be available for the GD, or its approximation, for various families of distributions where no such analytic form has been found in the case of, for instance, the Kullback–Leibler divergence (KLD) [11]. Finally, there is considerable experimental evidence suggesting that the GD in general is a more effective similarity measure between distributions than the KLD (see [11] and the references therein). We note that for distributions that lie infinitesimally close on the probabilistic manifold, it can be proven that the Kullback–Leibler divergence equals half of the squared geodesic distance between the distributions (see, e.g., [14]). Hence, in such a case, the KLD and GD yield similar results, but in general, they are quite different measures of similarity between distributions.

### *2.2. Geodesics for the Univariate Normal Distribution*

In this paper, we discuss applications that are based on a univariate normal distribution $\mathcal{N}(\mu, \sigma^2)$, parameterized by its mean $\mu$ and standard deviation $\sigma$. In this case, an analytic expression for the Fisher–Rao metric is available. It turns out to be the familiar Poincaré metric, which, when applied to a half-plane, is a well-known model for hyperbolic geometry that has constant negative scalar curvature. The line-element is given by [15,16]:

$$\mathrm{d}s^2 = \frac{\mathrm{d}\mu^2}{\sigma^2} + 2\frac{\mathrm{d}\sigma^2}{\sigma^2}. \tag{4}$$

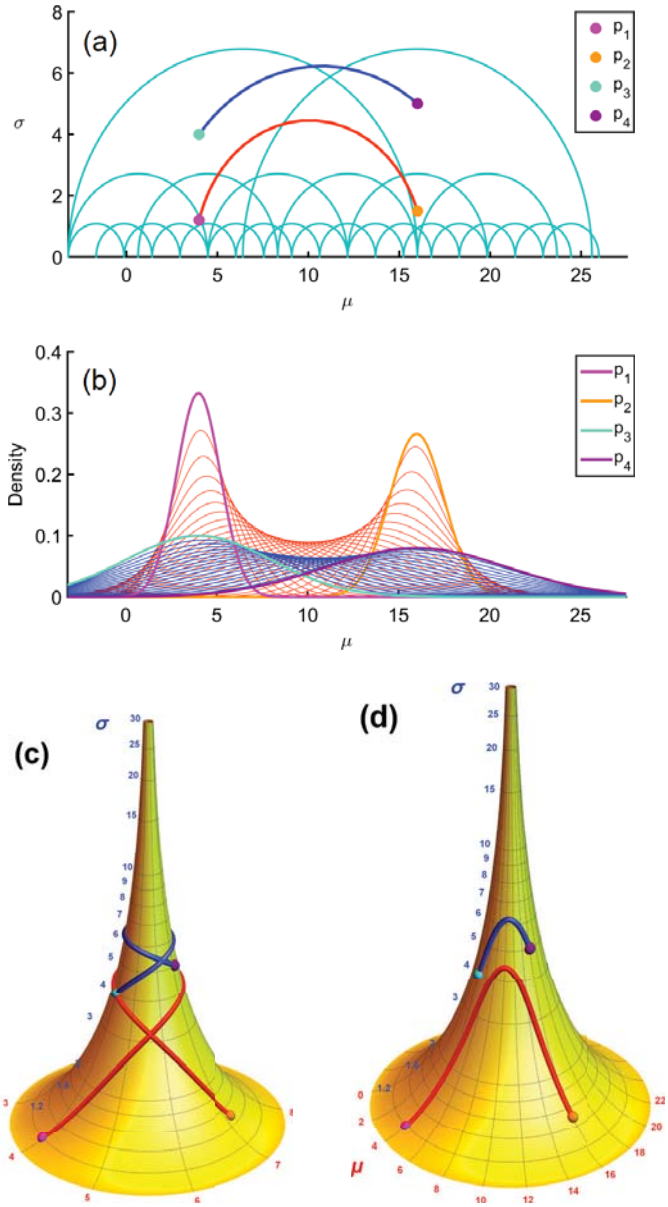



**Figure 1.** (**a**) Illustration of the Poincaré half-plane with several half-circle geodesics in the background, together with the geodesic between the points $p_1$ and $p_2$ and between $p_3$ and $p_4$, defined in the main text. (**b**) Probability densities corresponding to the points $p_1$, $p_2$, $p_3$ and $p_4$ indicated in (a). The densities associated with some intermediate points on the geodesics between $p_1$ and $p_2$ and between $p_3$ and $p_4$ are also drawn. (**c**) Rendering of one blade of the tractroid, again with the two geodesics superimposed. The parallels of the tractroid are lines of constant standard deviation σ, while the meridians (the tractrices) are lines of constant mean μ. This representation of the normal manifold is periodic in the μ-direction, and a rescaled version (longer period along μ) is shown in (**d**).

As an illustration, the Poincaré half-plane is pictured in Figure 1a, together with two geodesics between, on the one hand, the points $p_1 = \mathcal{N}(4, 1.2^2)$ and $p_2 = \mathcal{N}(16, 1.5^2)$ and, on the other hand, $p_3 = \mathcal{N}(4, 4.0^2)$ and $p_4 = \mathcal{N}(16, 5.0^2)$. The corresponding normal density functions are drawn in Figure 1b, as well as a number of densities associated with some intermediate points on each geodesic. As a further illustration, Figure 1c shows one blade of a particular pseudosphere, namely the tractroid, which is locally isometric to the Poincaré half-plane and the univariate normal manifold for $\sigma > 1$, with periodicity in $\mu$. In order to better visualize the geodesics, a rescaled version of the tractroid is shown in Figure 1d. This surface has a longer period in the $\mu$-direction. However, it should be kept in mind that only the visualization in Figure 1c can be used to measure absolute distances on the surface, for in Figure 1d, the pictured geodesics are no longer the shortest curves between the points in question. It is clear that the geodesics on the Gaussian manifold are different from straight lines in the Euclidean space, wherein the manifold has been immersed. The shape of the geodesics can be made intuitively clear by noting that they always pass through a region of increased standard deviation relative to that of the boundary points. This provides the shortest route, as can be seen from the line element Equation (4). Interestingly, similar arguments will be shown to enable a deeper insight into the operation of GLS regression. We further note that various alternative models exist to visualize hyperbolic geometry; see, e.g., [17].

A closed-form expression is available for the GD on the normal manifold, permitting fast evaluation. Indeed, for two univariate normal distributions $p_1(x|\mu_1, \sigma_1^2)$ and $p_2(x|\mu_2, \sigma_2^2)$, the GD is given by [16]:

$$\mathrm{GD}(p_1, p_2) = \sqrt{2} \ln \frac{1+\delta}{1-\delta} = 2\sqrt{2} \tanh^{-1} \delta, \qquad \delta \equiv \left[ \frac{(\mu_1 - \mu_2)^2 + 2(\sigma_1 - \sigma_2)^2}{(\mu_1 - \mu_2)^2 + 2(\sigma_1 + \sigma_2)^2} \right]^{1/2}. \quad (5)$$

Furthermore, since the injectivity radius of the hyperbolic plane is infinite, the geodesics are globally length-minimizing [10].

### *2.3. Geodesic Least Squares Methodology*

GLS starts from the premise that the probability distribution underlying experimental measurements is the fundamental object resulting from the measurement. As such, GLS does not perform regression based on data points in a Euclidean space, but rather operates on probability distributions lying on a probabilistic manifold. This introduces additional flexibility that renders the method robust in the presence of large uncertainties, as will be demonstrated in the experiments.

Briefly, the idea is to consider two different proposals for the distribution of the response (dependent) variable $y$, conditional on the predictor variables. On the one hand, there is the distribution that one would expect if all assumptions were correct regarding the deterministic component of the regression model (regression function) and the stochastic component. We call this the modeled distribution. On the other hand, we try to capture the conditional distribution of $y$ by relying less on the model assumptions, but directly on the measurements of $y$. For this, we will use the term observed distribution. It is in this sense that GLS is similar to minimum distance estimation (MDE), where the Hellinger distance is a popular similarity measure [18]. This was first applied to

regression in [19], but there are several differences with GLS. First and foremost, GLS calculates the geodesic distance between each individual pair of modeled and observed distributions of the response variable, corresponding to an individual measurement point. As such, each individual data point acquires the status of a probability distribution in its own right. Consequently, GLS performs regression between probability distributions on a probabilistic manifold. In contrast, MDE usually considers a distance between a kernel density estimate of the distribution of residuals, on the one hand, and the parametric model, on the other hand, based on the entire data sample. Secondly, we explicitly model all parameters of the modeled distribution, which is similar to the ideas behind the link function in the generalized linear model (GLM) [7]. In the present work, this will be accomplished by explicitly modeling both the mean and standard deviation of the Gaussian modeled distribution. Additionally, a final difference is that we use the Rao geodesic distance as a similarity measure.

As a simple example that we will use also in the experiments, consider a linear relation $\eta = \beta\xi$ between a single predictor variable $\xi$, and a response variable $\eta$, with $\beta$ a constant. In accordance with the discussion above, we explicitly wish to allow for the challenging case of uncertainty on the predictor variable $\xi$. Therefore, we assume that, in reality, $N$ samples of a stochastic (noisy) variable $x$ are observed, together with $N$ samples of a stochastic response variable $y$. We take the simple case of normally distributed (Gaussian) noise:

$$\begin{aligned} y &= \eta + \epsilon_y = \beta\xi + \epsilon_y, & \epsilon_y &\sim \mathcal{N}\left(0, \sigma_y^2\right), \\ x &= \xi + \epsilon_x, & \epsilon_x &\sim \mathcal{N}\left(0, \sigma_x^2\right). \end{aligned} \tag{6}$$

The observations $x_n$ ($n = 1, \ldots, N$) are taken as mutually independent and so are the $y_n$. $\sigma_x$ and $\sigma_y$ are assumed to be known, and in this example, they are taken constant for all measurements, *i.e.*, we have homoscedasticity. However, we will also consider heteroscedasticity later on. According to the regression model, conditionally on $x_n$, each measurement $y_n$ is drawn from a normal distribution:

$$p_{\mathrm{mod}}(y|x_n) = \mathcal{N}\left(\beta x_n, \sigma_{\mathrm{mod}}^2\right), \quad \text{where} \quad \sigma_{\mathrm{mod}}^2 \equiv \sigma_y^2 + \beta^2\sigma_x^2, \tag{7}$$

with the subscript "mod" referring to the modeled distribution. In our simple example, Equation (7) follows from standard Gaussian error propagation rules. However, for nonlinear regression laws, the conditional distribution for $y$ has to be obtained by marginalizing the unknown true values $\xi_n$. Nevertheless, the Gaussian error propagation laws may be used in the nonlinear case as well, to approximate the conditional distribution $p(y|x_n)$ by a normal distribution, as will be shown in the experiments.

We next choose a specific form of the observed distribution corresponding to each realization of the variable $y$, conditional on the observations, *i.e.*, $p_{\mathrm{obs}}(y|y_n)$. In this example, we take again the normal distribution, but centered on each data point: $\mathcal{N}\left(y_n, \sigma_{\mathrm{obs}}^2\right)$, where $\sigma_{\mathrm{obs}}$ is to be estimated from the data. In the context of the GLM, this is known as the saturated model. The extra parameter $\sigma_{\mathrm{obs}}$ gives the method added flexibility, since it is not *a priori* required to equal $\sigma_{\mathrm{mod}}$. As a result, GLS is less sensitive to incorrect model assumptions. Note that in this example, we have chosen the observed distribution from the same model (Gaussian) as the modeled distribution. Furthermore,

$\sigma_{\mathrm{mod}}$ is taken as a fixed value for all measurements and so is $\sigma_{\mathrm{obs}}$. These assumptions can of course be relaxed, leading to a more general method. However, the transition from OLS to GLS is best explained by means of a Gaussian observed distribution, which, in addition, offers computational advantages, since the expression for the GD has a closed form; see Equation (5).

GLS now proceeds by minimizing the total GD between, on the one hand, the joint observed distribution of the $N$ realizations of the variable $y$ and, on the other hand, the joint modeled distribution. Thanks to the independence assumption in this example, we can write this in terms of products of the corresponding marginal distributions:

$$\begin{aligned}\hat{\beta} &= \underset{\beta\in\mathbb{R}}{\operatorname{argmin}}\, \mathrm{GD}\left[\prod_{n=1}^{N} p_{\mathrm{obs}}(y|y_n), \prod_{n=1}^{N} p_{\mathrm{mod}}(y|x_n)\right] \\ &= \underset{\beta\in\mathbb{R}}{\operatorname{argmin}} \sum_{n=1}^{N} \mathrm{GD}^2\big[p_{\mathrm{obs}}(y|y_n), p_{\mathrm{mod}}(y|x_n)\big]. \end{aligned} \tag{8}$$

The last equality entails a considerable simplification, owing to the property that the squared GD between products of distributions can be written as the sum of squared GDs between the corresponding factors [16]. Hence, the optimization procedure involves matching not only $y_n$ with $bx_n$, but also $\sigma^2_{\mathrm{obs}}$ with $\sigma_y^2 + \beta^2\sigma_x^2$. Note that the parameter $\beta$ occurs both in the mean and the variance of the modeled distribution. Incidentally, forcing $\sigma^2_{\mathrm{obs}} \equiv \sigma_y^2 + \beta^2\sigma_x^2$ would take us back to standard maximum likelihood estimation, for the Rao GD between the two Gaussians $p_{\mathrm{obs}}$ and $p_{\mathrm{mod}}$ with means $y_n$ and $bx_n$, respectively, but with identical standard deviations (fixed along the geodesic path), is precisely the Mahalanobis distance [20]:

$$\mathrm{GD}(p_{\mathrm{obs}}, p_{\mathrm{mod}}) = \frac{|y_n - bx_n|}{\sigma_y^2 + \beta^2\sigma_x^2}, \qquad \text{if } \sigma^2_{\mathrm{obs}} \equiv \sigma_y^2 + \beta^2\sigma_x^2.$$

We note that the GLS scheme addresses many of the difficulties with classic OLS regression. First, GLS explicitly allows uncertainty on the predictor variables, and it is not restricted to normal or symmetric noise distributions, nor does it necessarily assume homoscedasticity. In addition, correlations among variables and among observations can be built into the stochastic component of the regression model. Furthermore, GLS can operate with any (nonlinear) regression function. Moreover, it will be shown in the experiments that GLS is relatively insensitive to uncertainties in both the stochastic and deterministic components of the regression model. The same quality renders the method also robust against outliers.

In the experiments below, we employed a classic active-set algorithm to carry out the optimization [21]. Furthermore, presently, the GLS method does not directly offer confidence (or credible) intervals on the estimated quantities. Future work will address this issue in more detail, but for now, error estimates were derived by Monte Carlo sampling in the case of the numerical simulations (Section 4) and by bootstrapping in the case of the real data (Section 5) [22]. The bootstrapping involved creating, from the measured dataset, a large number of artificial datasets of the same size, by resampling with replacement. The regression analysis was then carried out on each of the datasets, and the mean and standard deviation, over all datasets, of each estimated regression

parameter and of the predicted quantities were used as estimates of the parameter or prediction value and its error bar, respectively. This scheme typically results in rather conservative error bars, which could possibly be narrowed down using more sophisticated methods.

## 3. The L-H Power Threshold and Database

We now provide some background information regarding the main regression application that will be treated in the experiments with synthetic and real data. It concerns one of the most important scaling relations in fusion science based on magnetic plasma confinement, related to the threshold $P_{\mathrm{thr}}$ for the heating power that is required for the plasma to make the transition into a desired regime of high energy confinement (H-mode) in the next-step fusion device ITER (International Thermonuclear Experimental Reactor) [1,23,24]. To a good approximation, this so-called L-H (or H-mode) power threshold depends on the electron density in the plasma $\bar{n}_{\mathrm{e}}$ (in $10^{20}\ \mathrm{m}^{-3}$), the main magnetic field $B_{\mathrm{t}}$ (in tesla (T)) and the total surface area $S$ of the confined plasma (in $\mathrm{m}^2$). This is usually expressed by means of the following scaling relation:

$$P_{\mathrm{thr}} = \beta_0\, \bar{n}_{\mathrm{e}}^{\beta_1} B_{\mathrm{t}}^{\beta_2} S^{\beta_3}. \tag{9}$$

To estimate the coefficients in this relation, we employed data from eight fusion devices of the tokamak type (ASDEX, AUG, CMOD, DIII-D, JET, JFT-2M, JT-60U, PBXM) in the International Tokamak Physics Activity (ITPA) multi-machine database for the L-H power threshold (subset IAEA02 [23,25–27]). This yields $616$ measurement sets of power, density, magnetic field and surface area, each set obtained during a brief time of plasma operation under stationary conditions in one of the eight devices involved in the study [28].

The ITPA database contains some information regarding the error bars on the measurements, specifically relative errors expressed as percentages. This is important for our purposes, because we need the error bars to calculate $\sigma_{\mathrm{mod}}$. Unfortunately, the error estimates are not available in some cases, and if they are, the precise definition of the error bars is not always clear. Usually, an error bar in the database represents an estimate by the experimentalist of the typical range in which the "true" quantity can be expected to lie, where the uncertainty is assumed to be caused by both stochastic and systematic effects. Moreover, it is difficult to assess the probability that is covered by the stochastic component of the errors mentioned in the database. Since a detailed investigation of the uncertainty of the threshold data is beyond the scope of the present paper, we will assume that the error bars pertain to a stochastic uncertainty corresponding to a single standard deviation of a Gaussian distribution. For some derived quantities, the error bars had to be calculated from the uncertainty on more fundamental measurements. In those cases, we employed Gaussian error propagation rules to estimate the standard deviation on the derived quantities. For the case of the global H-mode confinement database, this strategy has been shown to provide reasonable information on the actual error bars [29].

It is important to mention that the main source of uncertainty in the data used for power threshold scaling, when compared to the predictions of a simple power law regression model, is not expected

to be due to the measurement uncertainty on the individual variables. There are far larger sources owing to the variability of the experiments that produced the data. To estimate the variability of each of the physical quantities with respect to the scaling law, we performed the following calculation. First, the nonlinear scaling law was estimated using OLS, as explained in Section 5.2. Then, for a specific variable $z$ (one of the predictor variables or the dependent variable) and for each data point, the relative difference was computed between the $z$-value of the data point itself and the $z$-value of the projection of the data point on the hypersurface given by the scaling law, keeping the values of the other variables fixed. This difference can be interpreted as the deviation of the point from the theoretical scaling law, assuming the deviation is solely due to the variability of the variable $z$. Finally, the standard deviation of these relative differences was taken, and the procedure was repeated for every predictor variable and the dependent variable. The resulting standard deviations can be interpreted as upper bounds of the relative variability of each of the variables around their 'theoretical' values given by the scaling law. This way, for $\bar{n}_{\mathrm{e}}$, $B_{\mathrm{t}}$, $S$ and $P_{\mathrm{thr}}$, we obtained $39\%$, $31\%$, $28\%$ and $38\%$, respectively. These levels are clearly much larger than the relative uncertainties due to measurement error alone. Indeed, the typical measurement error bars quoted in the ITPA database, on average, over all devices, are estimated at $4\%$ for $\bar{n}_{\mathrm{e}}$, $1\%$ for $B_{\mathrm{t}}$, $3\%$ for $S$ and $15\%$ for $P_{\mathrm{thr}}$ [25,26].

## 4. Numerical Simulations

We next demonstrate some of the potential of the GLS regression scheme by means of a number of experiments with synthetically-generated data. We treat two particular cases of deviation from the model according to which the data were created and show that, in comparison with a number of standard regression techniques, GLS yields the most accurate results across all experiments. The first case concerns the effect of outliers, while in the second case, the influence of a logarithmic transformation is studied. In each case, we start with a very simple experiment that is easily reproduced, using a single predictor variable, providing some intuitive feeling regarding the performance of the method. We then proceed to a more elaborate test, still based on partly synthetic data, but using a regression challenge similar to that used in the real-world experiment for scaling of the L-H power threshold in fusion plasmas, presented in Section 5.

### *4.1. Effect of Outliers*

The robustness of minimum distance estimators to outliers in the data was noted in the classic literature of minimum distance estimation [18]. We now show that this is a quality also enjoyed by GLS regression.

4.1.1. Single Predictor Variable

We first concentrate on estimating the slope of a regression line with a single predictor variable. To this end, a dataset was generated consisting of ten points labeled by coordinates $\xi_n$ and $\eta_n$ ($n = 1, \ldots, 10$), with the $\xi_n$ chosen unevenly between zero and 50 and $\eta_n = \beta\xi_n$, taking $\beta = 3$. Then, Gaussian noise was added to all coordinates according to Equation (6), with $\sigma_y = 2.0$ and $\sigma_x = 0.5$, resulting in values $x_n$ for the predictor variable and $y_n$ for the response variable. Finally, one outlier was created by multiplying the value of $y_k$ by a factor distributed uniformly between $1.5$ and $2.5$, with $k$ chosen uniformly among the indices 8, 9 and 10.

We next estimated $\beta$ by means of GLS and compared the estimates with those obtained by OLS, maximum *a posteriori* (MAP) using the model in Equation (7) for the likelihood and an uninformative prior [30], total least squares (TLS), which is a typical errors-in-variables technique [31], and a robust method (ROB) based on iteratively reweighted least squares (bisquare weighting) [32], included in the MATLAB Statistics toolbox [33]. It should be noted that MAP takes into account the error bars on the predictor variables. In all cases, we assumed knowledge of the values of $\sigma_x$ and $\sigma_y$. In order to get an idea of the variability of the estimates, Monte Carlo sampling of the data-generating distributions was performed, and the estimation was carried out $100$ times.

The results are given in Table 1, mentioning the sample average and standard deviation of the estimates $\hat{\beta}$ over the $100$ runs for each of the methods. GLS is seen to perform very well and similar to the robust method ROB, but the other techniques yield considerably worse results. The average estimate of $\sigma_{\text{obs}}$ was $5.43$ with a standard deviation of $0.24$. On the other hand, the modeled value of the standard deviation in the conditional distribution for $y$ was $\sigma_{\text{mod}} = \sqrt{\sigma_y^2 + 9\sigma_x^2} = 2.5$. Hence, GLS succeeds in ignoring the outlier by increasing the estimated variability of the data. Put differently, the effect of the outlier is, in a sense, to increase the overall variability of the data, which GLS takes into account by increasing the observed standard deviation of the data ($\sigma_{\text{obs}}$) with respect to the standard deviation predicted by the model ($\sigma_{\text{mod}}$).

**Table 1.** Monte Carlo estimates of the mean and standard deviation for the slope parameter in linear regression with errors on both variables and one outlier. GLS, geodesic least squares regression; TLS, total least squares; ROB, robust method.

| Original | GLS | OLS | MAP | TLS | ROB |
|---|---|---|---|---|---|
| $\beta = 3.00$ | **$3.031 \pm 0.035$** | $3.68 \pm 0.29$ | $3.83 \pm 0.36$ | $4.6 \pm 1.0$ | $2.992 \pm 0.041$ |

As mentioned before, this result can also be understood in terms of the pseudosphere as a geometrical model for the normal distribution. To see this, we refer to Figure 2, where several sets of points (distributions) are drawn on a portion of the surface of the pseudosphere for one particular dataset generated as described above. First, the modeled distributions are plotted with their means $\hat{\beta}x_n$ (see Equation (7)) and standard deviations $\sigma_{\text{mod}} = 2.5$, using the average estimate $\hat{\beta} = 3.031$ obtained by GLS. These are the green points on the surface, and they lie on a parallel, since they all correspond to Gaussians with the same standard deviation $\sigma_{\text{mod}}$. In this particular dataset, the index

of the outlier was $k = 10$, so the point $\hat{\beta}x_{10}$ is indicated individually. Obviously, according to the model, no outlier is expected, so the modeled distribution corresponding to $k = 10$, which is the green point just mentioned, lies close to the other predicted points (distributions). Next, we plot the observed distributions with their means $y_n$ and standard deviations $\sigma_{\mathrm{obs}}$ (for this dataset estimated at $\hat{\sigma}_{\mathrm{obs}} = 5.43$). These are the blue points, lying at a constant standard deviation $\sigma_{\mathrm{obs}}$, which is higher than $\sigma_{\mathrm{mod}}$ ($5.43 > 2.5$). The outlier $y_{10}$ can clearly be observed, and being an outlier, it lies relatively far away from the rest of the blue points (observed distributions). Now suppose that, like MAP, GLS would not be able to increase $\sigma_{\mathrm{obs}}$ relative to $\sigma_{\mathrm{mod}}$ in order to accommodate the outlier. Then, the observed distributions would have the same observed means (the measured values $y_n$), but they would have the standard deviation predicted by the model. Hence, they would lie on the parallel corresponding to $\sigma_{\mathrm{mod}}$, just like the green points. We have plotted these fictitious distributions as the red points at the level of $\sigma_{\mathrm{mod}}$, and they are labeled $\tilde{y}$. Again, the outlier (labeled $\tilde{y}_{10}$) can be seen, but it seems to lie further away from the other red points (the points $\tilde{y}_n$) compared to the actually observed situation, *i.e.*, the distance from $y_{10}$ to the other $y_n$ (blue points). At least this is the case when using (visually) the Euclidean distance in the embedding Euclidean space. We can verify that this is indeed so by using the proper geodesic distance on the surface: overall, the blue points lie closer together (including the outlier) than the red points. Now, in fact, GLS aims at minimizing the distance between each green point (modeled distribution) and its corresponding blue point (observed distribution), so as far as the outlier is concerned, we should really be looking at the geodesic between the point $(\hat{\beta}x_{10}, \sigma_{\mathrm{mod}})$ and the point $(y_{10}, \sigma_{\mathrm{obs}})$. The geodesic (labeled "$\mathrm{Geo}_1$") between these points is also drawn on the surface, and again, we compare this to the fictitious situation, represented by the geodesic (labeled "$\mathrm{Geo}_2$") between $(\hat{\beta}x_{10}, \sigma_{\mathrm{mod}})$ and $(\tilde{y}_{10}, \sigma_{\mathrm{mod}})$. Indeed, again, we see that the geodesic $\mathrm{Geo}_1$ is shorter than $\mathrm{Geo}_2$. Therefore, by increasing $\sigma_{\mathrm{obs}}$ relative to $\sigma_{\mathrm{mod}}$, the outlier is not so much an outlier anymore, as measured on the pseudosphere! When calculating the GD, one finds $2.4$ for $\mathrm{Geo}_1$ and $2.8$ for $\mathrm{Geo}_2$. Therefore, GLS obtains a lower value of the objective function (sum of squared geodesic distances) if it increases $\sigma_{\mathrm{obs}}$ with respect to $\sigma_{\mathrm{mod}}$. Of course, there is a limit to this: GLS cannot continue raising $\sigma_{\mathrm{obs}}$ indefinitely, trying to mitigate the distorting effect of the outlier, for then, the other points would get a too high observed standard deviation, which is not supported by the data. The image that we see in Figure 2 is the best compromise that GLS could find. In fact, we note that, in the case we suspect that $y_{10}$ could be an outlier, it may very well be worthwhile to introduce two parameters to describe the observed standard deviation: one for the nine points that seem to follow the model and one to take care of the outlier. This would be a very straightforward extension of the method, and we explore this to some extent when using data from the ITPA database below. There, we assign a separate parameter to describe the observed standard deviation of all data coming from a specific tokamak, hence defining an individual parameter for each machine.

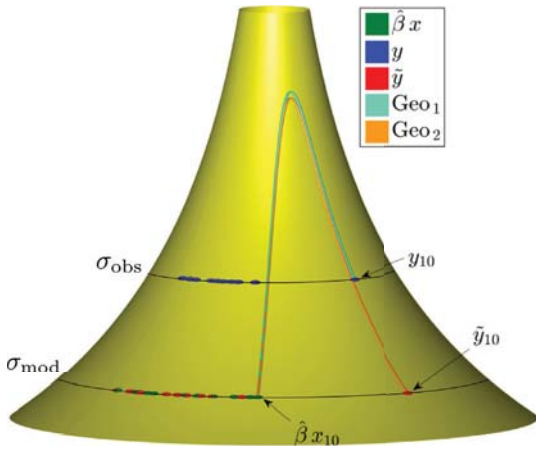



**Figure 2.** A portion of the pseudosphere together with the regression results on synthetic data with an outlier, as described in the main text.

4.1.2. Multiple Predictor Variables

In this experiment, a regression problem with multiple predictor variables and a power law is studied. The deterministic part of the regression model is based on the real-world problem for the L-H power threshold in fusion plasmas, which we are going to consider in Section 5. Furthermore, the values of the predictor variables are taken from the same international power threshold database, and values of the response variable are synthetically generated from this.

More specifically, the dataset for this experiment was created as follows. First, an artificial linear regression law was put forward for a variable $\eta$, depending on the predictor variables $\bar{n}_{\mathrm{e}}$, $B_{\mathrm{t}}$ and $S$, which were introduced in the context of the power threshold scaling law in Section 3 [34]. In particular, we generated a number of realizations of the variable $\eta$ from the following prescription:

$$\eta = \beta_0 + \beta_1 \bar{n}_{\mathrm{e}} + \beta_2 B_{\mathrm{t}} + \beta_3 S. \tag{10}$$

This was considered as the "true" relation between the predictor and response variables, where, as mentioned above, the values of the predictor variables were chosen to be exactly those from the ITPA database, which are normally used in the real power threshold scaling law. A whole range of datasets was created using the following values of the coefficients $\beta_0$, $\beta_1$, $\beta_2$ and $\beta_3$:

$$\begin{aligned} \beta_0 &= 1, 1.1, \ldots, 20, \\ \beta_1, \beta_2, \beta_3 &= 0.1, 0.2, \ldots, 2. \end{aligned} \tag{11}$$

Thus, for each combination of values of $\beta_0$, $\beta_1$, $\beta_2$ and $\beta_3$, all $616$ values of $\eta$ were calculated according to Equation (10), based on the values of $\bar{n}_{\mathrm{e}}$, $B_{\mathrm{t}}$ and $S$ from the ITPA database. The range of coefficient values in Equation (11) was chosen to be representative for the values that are typically obtained from a regression analysis on the true scaling law (see Section 5). The exception is $\beta_0$, for

which the range was chosen of roughly the same order as $\eta - \beta_0$ (much smaller values of $\beta_0$ would not be estimable in comparison with $\eta - \beta_0$).

Next, Gaussian noise was added to both the predictor and response variables. The noise level was chosen according to the typical relative measurement errors in the ITPA database, *i.e.*, $4\%$ for $\bar{n}_{\mathrm{e}}$, resulting in a variable $x_1$, $1\%$ for $B_{\mathrm{t}}$ (variable $x_2$), $3\%$ for $S$ (variable $x_3$) and $15\%$ for the dependent variable (variable $y$, which is $P_{\mathrm{thr}}$ in the real-world regression problem). It should be stressed that, in the light of our comments in Section 3 regarding the variability of the predictor variables, these are rather low noise levels. We further note that fixed relative noise levels lead to a different standard deviation for each measurement (heteroscedasticity).

Furthermore, in this experiment studying the effect of atypical observations, $10$ outliers were created in each of the datasets. In particular, from the total of $616$ points in each dataset, $10$ points were randomly chosen, and the associated value of $y$ was multiplied with a factor $F$, where $F$ was distributed uniformly between $1.5$ and $2.5$. For each combination of coefficient values $\beta_i$ ($i = 0, \ldots, 3$) taken from Equation (11), $10$ datasets were realized, each time performing the sampling of noise and outliers.

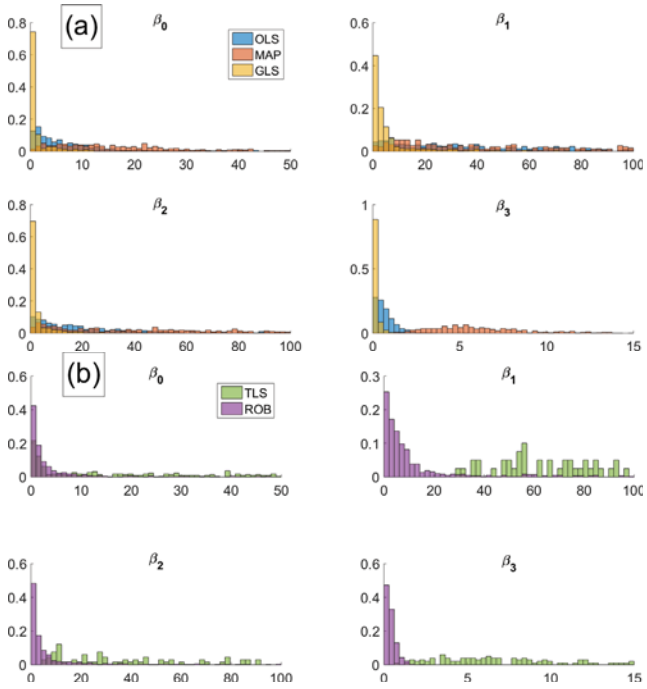



**Figure 3.** (**a**) Histograms of the relative error in estimating the regression coefficients $\beta_i$ by means of OLS, MAP and GLS for a linear regression problem with outliers. Horizontal axes represent the error in percent and vertical axes probability, normalized to one. (**b**) Similar, for TLS and ROB.

Finally, the regression analysis was carried out for every dataset, and for each choice of the $\beta_i$, the obtained estimates $\hat{\beta}_i$ were defined as the average over the $10$ data realizations. Next, histograms were created based on these averages for the estimated coefficients, specifically the normalized histograms of the relative difference $(\beta_i - \hat{\beta}_i)/\beta_i$ $(i = 0, \ldots, 3)$, expressed as a percentage, between the true value $\beta_i$ and the estimated value $\hat{\beta}_i$ of each regression parameter. The histograms of these percentage errors are shown in Figure 3. In order to avoid a cluttered figure, the results of OLS, MAP and GLS are plotted in one panel and those of TLS and ROB in another.

From these histograms, it is clear that, for each parameter, GLS performs much better than OLS, MAP and TLS, with the latter failing completely. In case of GLS, the vast majority of relative errors is of the order of a few percent and certainly smaller than $20\%$. Overall, the most difficult to estimate parameter turns out to be $\beta_1$, which is associated with $\bar{n}_{\mathrm{e}}$. The robust estimation technique in MATLAB also delivers good results (in fact, not much worse than GLS), as it is designed to cope with outliers. However, we will see that in the next experiment ROB does not perform well at all.

### *4.2. Effect of Logarithmic Transformation*

We next tested the effect of a logarithmic transformation, which is often used to transform a power-law regression model into a linear form. However, the logarithm alters the data distribution, which may lead to misguided inferences from OLS [2,3]. Therefore, the flexibility offered by GLS is expected to be beneficial in this case, as it allows the observed distribution to deviate from the modeled distribution.

#### 4.2.1. Single Predictor Variable

Again, we first performed a simple regression experiment involving a single predictor variable, with a power law deterministic model and additive Gaussian noise on all variables. In accordance with the typical situation of fitting fusion scaling laws to multi-machine data, the noise standard deviation was taken proportional to the simulated measurements, corresponding to a given set of relative error bars. As a result, in the logarithmic space the distributions were only approximately Gaussian, with the standard deviation given by the constant relative error on the original measurement (homoscedasticity). Ten points were chosen with predictor values $\xi_n$ unevenly spread between zero and $60$. A power law was proposed to relate the unobserved $\xi_n$ and $\eta_n$:

$$\eta_n = \beta_0 \, \xi_n^{\beta_1}, \quad n = 1, \ldots, 10.$$

Then, Gaussian noise was added to the $\xi_n$ and $\eta_n$, corresponding to a substantial relative error of $40\%$. We finally took the natural logarithm of all observed values $x_n$ and $y_n$, enabling application of the same linear regression methods that were used in the previous experiment. In this particular experiment, we chose $\beta_0 = 0.8$ and $\beta_1 = 1.4$, but we found that other values yield similar conclusions. Again, $100$ data replications were generated, allowing calculation of Monte Carlo averages.

The averages and standard deviations over all 100 runs are given in Table 2. Again, the results show that GLS is robust against the flawed model assumptions, now performing similar to TLS.

**Table 2.** Monte Carlo estimates of the mean and standard deviation for the parameters in a log-linear regression experiment with proportional additive noise on both variables.

| Parameter | Original | GLS | OLS | MAP | TLS | ROB |
|---|---|---|---|---|---|---|
| $\beta_0$ | 0.80 | **0.94 ± 0.47** | 2.2 ± 2.3 | 3.0 ± 1.7 | 0.99 ± 0.70 | 2.72 ± 0.77 |
| $\beta_1$ | 1.40 | **1.39 ± 0.11** | 1.19 ± 0.16 | 1.08 ± 0.26 | 1.41 ± 0.14 | 1.17 ± 0.11 |

4.2.2. Multiple Predictor Variables

In the last experiment with synthetic data, we studied the effect of a logarithmic transformation in a similar problem as the one described in Section 4.1.2, but in the case of a power law. In particular, the variable $\eta$ was calculated for the same range of values of the parameters $\beta_i$ as given in Equation (11), but now according to a power law:

$$\eta = \beta_0 \bar{n}_e^{\beta_1} B_t^{\beta_2} S^{\beta_3}.$$

Then, Gaussian noise was added to all variables. However, when applying the relatively low noise levels used in Section 4.1.2, no significant differences were observed in the performance of GLS and MAP. Therefore, the noise levels for the predictor variables were augmented to 20% for $\bar{n}_e$ (variable $x_1$), 5% for $B_t$ (variable $x_2$) and 15% for $S$ (variable $x_3$). The level for $P_{thr}$ was kept at 15%, as before. This is still well within the maximum variability range that can be expected for the predictor variables in the ITPA database, as discussed in Section 3.

After adding the noise, all data were transformed to the logarithmic domain, and 10 datasets were generated for each combination of regression coefficients. Subsequently, linear regression analysis was applied to each of the log-transformed datasets. The coefficient estimates, defined as the average over the 10 replications, were then compared among the various regression methods, as shown in Figure 4. Again, the normalized histograms of the relative error on the estimated parameters are displayed, showing the consistently better performance of GLS over all other methods tested, including TLS and ROB. For GLS, the errors on $\beta_0$ and $\beta_1$ are the largest, compared to those on $\beta_2$ and $\beta_3$, but the majority is still below 20%. As for $\beta_0$, the slightly inferior performance of GLS relative to the results with outliers in Section 4.1.2 is simply due to the fact that $\log \beta_0$ for the lowest values of $\beta_0$ is negligibly small compared to $\log \eta - \log \beta_0$.

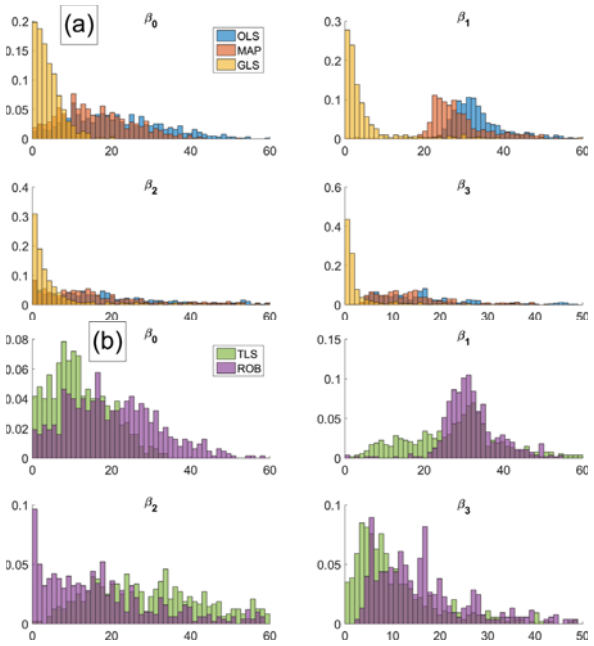



**Figure 4.** Histograms of the relative error in estimating the regression coefficients $\beta_i$ by means of OLS, MAP and GLS for a power-law regression problem after a logarithmic transformation. Horizontal axes represent the error in percent and vertical axes probability, normalized to one. (**b**) Similar, for TLS and ROB.

## 5. Power Threshold Scaling

We finally come to the application of power threshold scaling using real-world data from the ITPA database for all variables, including the response variable $P_{\mathrm{thr}}$. We start with log-linear regression and then apply nonlinear regression analysis. Next, we perform a simple analysis of the influence of the error bars on the estimation results, and we finally provide a discussion of the results in this section.

### *5.1. Linear Scaling*

We first followed the standard practice of transforming to the logarithmic scale to estimate the coefficients $\beta_0$, $\beta_1$, $\beta_2$ and $\beta_3$ in Equation (9) via linear regression. In the GLS method, we introduced additional parameters $\sigma_{\mathrm{obs},\alpha}$ ($\alpha = 1, \ldots, N_{\mathrm{t}}$), one for each of the $N_{\mathrm{t}} = 8$ tokamaks contributing data to the scaling. That is, if a certain data point with index $n$ originated from tokamak $\alpha$, then in term $n$ of the objective function in Equation (8), an observed distribution was used, parameterized by means of the $\sigma_{\mathrm{obs},\alpha}$ corresponding to that machine. The $\sigma_{\mathrm{obs},\alpha}$ serve a similar purpose as the parameter $\sigma_{\mathrm{obs}}$ defined above, except that they describe the observed standard deviations of the logarithmic power threshold. This, of course, corresponds to the relative errors on

the power threshold itself. To calculate $\sigma_{\mathrm{mod}}$ for each data point, we used the relative measurement error bars quoted in the database (typically $4\%$ for $\bar{n}_{\mathrm{e}}$, $1\%$ for $B_{\mathrm{t}}$, $3\%$ for $S$ and $15\%$ for $P_{\mathrm{thr}}$). Considering the discussion in Section 3 regarding other sources of uncertainty, it is clear that the $\sigma_{\mathrm{obs},\alpha}$ will need to take into account other, "unexpected" uncertainty sources, hence increasing the flexibility of the method.

In this analysis, we compared GLS only with OLS and the powerful MAP method. The results on the IAEA02 data are given in Table 3. The predictions for ITER are also shown, for two typical densities ($0.5$ and $1.0 \times 10^{20}$ $\mathrm{m}^{-3}$). All estimates are accompanied by their 95% credible intervals obtained from 100 bootstrap samples (artificial datasets). We stress that this notion of a credible interval corresponds to the standard Bayesian definition of an interval wherein the true value of a stochastic variable is assumed to lie with a certain probability (e.g., $0.95$).

**Table 3.** Estimates of regression parameters and predictions for ITER in log-transformed linear scaling of the H-mode threshold power using the IAEA02 dataset. The bootstrap averages are given, as well as the 95% credible intervals (CI).

| **Method** | | $\hat{\beta}_0$ | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_3$ | $\hat{P}_{\mathrm{thr},0.5}$ (MW) | $\hat{P}_{\mathrm{thr},1.0}$ (MW) |
|---|---|---|---|---|---|---|---|
| OLS | Average | 0.0507 | 0.485 | 0.873 | 0.843 | 38.0 | 53.2 |
| | CI | $\pm$0.0060 | $\pm$0.073 | $\pm$0.061 | $\pm$0.041 | $\pm$4.4 | $\pm$8.0 |
| MAP | Average | 0.0449 | 0.567 | 0.867 | 0.901 | 45.6 | 67.6 |
| | CI | $\pm$0.0051 | $\pm$0.078 | $\pm$0.069 | $\pm$0.039 | $\pm$5.0 | $\pm$9.6 |
| GLS | Average | 0.0426 | 0.660 | 0.795 | 0.946 | 48.3 | 76.4 |
| | CI | $\pm$0.0042 | $\pm$0.069 | $\pm$0.059 | $\pm$0.034 | $\pm$4.7 | $\pm$9.8 |

**Table 4.** Estimates of the observed standard deviations $\sigma_{\mathrm{obs},\alpha}$ of the logarithmic power threshold, expressed as percentage errors on $P_{\mathrm{thr}}$ itself, for the tokamaks contributing to the IAEA02 dataset, obtained using log-transformed linear scaling. The bootstrap averages are given, as well as the 95% credible intervals (CI).

| | **ASDEX** | **AUG** | **CMOD** | **DIII-D** | **JET** | **JFT-2M** | **JT-60U** | **PBXM** |
|---|---|---|---|---|---|---|---|---|
| Average (%) | 41.8 | 23.0 | 22.0 | 15.7 | 24.6 | 15.9 | 22.8 | 27.6 |
| CI (%) | $\pm$5.3 | $\pm$1.4 | $\pm$1.1 | $\pm$1.8 | $\pm$2.0 | $\pm$1.2 | $\pm$2.3 | $\pm$2.9 |

The estimates by GLS of the parameters $\sigma_{\mathrm{obs},\alpha}$ (observed standard deviation on $\log P_{\mathrm{thr}}$), for each of the devices contributing to the IAEA02 data, were expressed as a relative error on the bootstrap-averaged $P_{\mathrm{thr}}$. These relative errors and their credible intervals are given in Table 4. The relative error on the power threshold lies around 15% to 30% for the various machines, except for ASDEX, where the uncertainty reaches a higher level of about 40%. On average, this yields an estimated error of $24.2\%$ for $P_{\mathrm{thr}}$, which is quite somewhat higher than the average of $15\%$

mentioned in the database, although still considerably lower than the upper bound of $38\%$, as calculated in Section 3. Again, this is an indication of additional sources of uncertainty, on top of mere measurement error, causing the data points to deviate from the proposed regression model, as discussed already in Section 3. That extra uncertainty is captured by the GLS method.

### *5.2. Nonlinear Scaling*

Next, we show the results of nonlinear regression in the original data space, *i.e.*, without logarithmic transformation. Whereas this prevents an analytic solution using OLS, the advantage is that the distribution of the data is left undistorted [2,3], while the implementation of OLS, MAP and GLS is not significantly more complex. Indeed, the distribution of the right-hand side in Equation (9) can be approximated by a Gaussian with mean $\mu_{\rm mod} = \beta_0\, \bar{n}_{\rm e}^{\beta_1} B_{\rm t}^{\beta_2} S^{\beta_3}$ and standard deviation $\sigma_{\rm mod}$, given by:

$$\sigma_{\rm mod}^2 = \sigma_{P_{\rm thr}}^2 + \mu_{\rm mod}^2 \left[ \beta_1^2 \left( \frac{\sigma_{\bar{n}_{\rm e}}}{\bar{n}_{\rm e}} \right)^2 + \beta_2^2 \left( \frac{\sigma_{B_{\rm t}}}{B_{\rm t}} \right)^2 + \beta_3^2 \left( \frac{\sigma_S}{S} \right)^2 \right].$$

Hence, the modeled standard deviations depend on the measurements (heteroscedasticity). Nevertheless, in defining the observed standard deviations $\sigma_{{\rm obs},\alpha}$, we introduced an approximation assuming constant error bars for all measurements from a single machine. This assumption may be relaxed in the future.

The results of the scalings and predictions are presented in Tables 5 and 6. We compared GLS with OLS and MAP using uniform priors. It may be possible to derive even less informative priors for MAP, as was done in the log-linear case in Section 5.1 (and see [30,35]), but this was not pursued here. Moreover, even in the log-linear analysis, we observed only a marginal difference between the results under various choices of priors.

**Table 5.** Estimates of regression parameters and predictions for ITER in power-law scaling on the original scale of the H-mode threshold power using the IAEA02 dataset. The bootstrap averages are given, as well as the 95% credible intervals (CI).

| **Method** | | $\hat{\beta}_0$ | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_3$ | $\hat{P}_{\rm thr,0.5}$ **(MW)** | $\hat{P}_{\rm thr,1.0}$ **(MW)** |
|---|---|---|---|---|---|---|---|
| OLS | Average | 0.0274 | 0.773 | 0.96 | 1.038 | 69 | 118 |
| | CI | $\pm$0.0083 | $\pm$0.090 | $\pm$0.10 | $\pm$0.071 | $\pm$15 | $\pm$32 |
| MAP | Average | 0.0425 | 0.643 | 0.788 | 0.933 | 44.2 | 69.1 |
| | CI | $\pm$0.0041 | $\pm$0.074 | $\pm$0.079 | $\pm$0.034 | $\pm$3.8 | $\pm$8.2 |
| GLS | Average | 0.0397 | 0.715 | 0.751 | 0.984 | 51.6 | 84.7 |
| | CI | $\pm$0.0036 | $\pm$0.071 | $\pm$0.081 | $\pm$0.031 | $\pm$4.0 | $\pm$8.8 |

**Table 6.** Estimates of the observed standard deviations $\sigma_{\mathrm{obs},\alpha}$ of the power threshold $P_{\mathrm{thr}}$, expressed as percentage errors, for the machines contributing to the IAEA02 dataset, obtained using power-law scaling. The bootstrap averages are given, as well as the 95% credible intervals (CI).

| | **ASDEX** | **AUG** | **CMOD** | **DIII-D** | **JET** | **JFT-2M** | **JT-60U** | **PBXM** |
|---|---|---|---|---|---|---|---|---|
| Average (%) | 35.8 | 21.2 | 20.4 | 15.9 | 22.4 | 15.7 | 22.3 | 27.7 |
| CI (%) | ±9.1 | ±4.3 | ±3.4 | ±2.4 | ±3.8 | ±2.2 | ±4.6 | ±8.1 |

It should also be mentioned that, in obtaining Table 6, we again calculated relative errors from the observed standard deviations estimated by GLS. However, this time, the relative errors are not the same for all measurements coming from a single machine, so we calculated an average for each machine (and similar for the credible interval). The resulting errors on $P_{\mathrm{thr}}$ are relatively similar to those using log-linear scaling, with an average over all devices of 22.7%, which is again higher than the 15% expected from measurement error only.

### *5.3. Influence of Error Bars*

In the last couple of experiments, we intended to assess the sensitivity of the regression analysis on the accuracy of the error bars on the ITPA data. A systematic study of this influence is outside the scope of this paper, and as a first simple test, we doubled the error bars on all root ITPA variables (basically the electron density and the magnetic field together with various geometrical plasma parameters and sources of input power), which were used for calculation of the variables involved in the power threshold scaling law. On average, over all machines, this resulted in the following derived error bars: 9% on $\bar{n}_{\mathrm{e}}$, 2% on $B_{\mathrm{t}}$, 5% on $S$ and 32% on $P_{\mathrm{thr}}$. Again, these are all below the maxima quoted in Section 3.

We then performed power-law regression with MAP and GLS on the ITPA data using these larger error bars; the results are given in Table 7 [36]. It is observed that, based on MAP, the predictions for ITER are lowered relative to the analysis with the original error bars in Section 5.2. In contrast, the predictions by GLS remain about the same as before. On the other hand, the GLS estimates of the observed standard deviations, listed in Table 8, are increased for all devices. This is how GLS accommodates the increased error bars on the data.

**Table 7.** Estimates of regression parameters and predictions for ITER in power-law scaling on the original scale of the H-mode threshold power using the IAEA02 dataset with all error bars (on the root quantities) doubled.

| **Method** | $\hat{\boldsymbol{\beta}}_{\mathbf{0}}$ | $\hat{\boldsymbol{\beta}}_{\mathbf{1}}$ | $\hat{\boldsymbol{\beta}}_{\mathbf{2}}$ | $\hat{\boldsymbol{\beta}}_{\mathbf{3}}$ | $\hat{\mathbf{P}}_{\mathbf{thr,0.5}}$ **(MW)** | $\hat{\mathbf{P}}_{\mathbf{thr,1.0}}$ **(MW)** |
|---|---|---|---|---|---|---|
| MAP | 0.0436 | 0.581 | 0.828 | 0.900 | 41.0 | 61.3 |
| GLS | 0.0393 | 0.725 | 0.742 | 0.990 | 52.1 | 86.2 |

**Table 8.** Estimates of the observed standard deviations $\sigma_{\mathrm{obs},\alpha}$ of the power threshold $P_{\mathrm{thr}}$, expressed as percentage errors, for the machines contributing to the IAEA02 dataset with all error bars doubled, obtained using power-law scaling.

| ASDEX | AUG | CMOD | DIII-D | JET | JFT-2M | JT-60U | PBXM |
|---|---|---|---|---|---|---|---|
| 49.5 | 35.9 | 31.7 | 24.9 | 32.9 | 27.6 | 38.9 | 47.7 |

In another simple test, we changed the error bars on $\bar{n}_{\mathrm{e}}$, $B_{\mathrm{t}}$, $S$ and $P_{\mathrm{thr}}$ to values computed from the average percentages mentioned earlier in Section 3: $4\%$ for $\bar{n}_{\mathrm{e}}$, $1\%$ for $B_{\mathrm{t}}$, $3\%$ for $S$ and $15\%$ for $P_{\mathrm{thr}}$. These are averages over all machines, rendering the final absolute error bars (standard deviations), computed from the relative errors, less precise. The estimation results using power-law regression with MAP and GLS are shown in Table 9. The results of both methods are clearly affected by the averaging step, but again, MAP is seen to be more sensitive to the change in the error bars compared to GLS, which maintains estimates in a similar range as those given in Tables 3 and 5. The estimates of the observed standard deviations, given in Table 10, are adjusted accordingly by GLS.

**Table 9.** Estimates of regression parameters and predictions for ITER in power-law scaling on the original scale of the H-mode threshold power using the IAEA02 dataset with averaged error bars.

| Method | $\hat{\beta}_0$ | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_3$ | $\hat{P}_{\mathrm{thr},0.5}$ (MW) | $\hat{P}_{\mathrm{thr},1.0}$ (MW) |
|---|---|---|---|---|---|---|
| MAP | 0.0488 | 0.552 | 0.807 | 0.862 | 35.1 | 51.5 |
| GLS | 0.0429 | 0.647 | 0.780 | 0.938 | 45.7 | 71.5 |

**Table 10.** Estimates of the observed standard deviations $\sigma_{\mathrm{obs},\alpha}$ of the power threshold $P_{\mathrm{thr}}$, expressed as percentage errors, for the machines contributing to the IAEA02 dataset with averaged error bars, obtained using power-law scaling.

| ASDEX | AUG | CMOD | DIII-D | JET | JFT-2M | JT-60U | PBXM |
|---|---|---|---|---|---|---|---|
| 29.6 | 19.1 | 20.5 | 19.5 | 22.5 | 18.1 | 18.7 | 20.4 |

*5.4. Discussion*

Several interesting observations can be made from the experiments regarding the power threshold scaling in this section. First, considering Tables 3 and 5, it should be noted that there are several instances where the regression parameters estimated by OLS differ significantly from those obtained by GLS. For log-linear regression, this is particularly the case for the dependence of the power threshold on density and surface area and for the predicted power thresholds for ITER, as shown by the non-overlapping credible intervals. For power-law regression, the difference is rather situated

in the dependence on the magnetic field. In this case, the power thresholds predicted by OLS are also quite different from the results given by GLS, but this time, the credible intervals on the OLS estimates are so wide, that they overlap with those obtained from GLS. Apart from this discrepancy, the three methods provide comparable absolute error bars on their estimates.

Furthermore, we see that the correspondence between GLS and MAP is significantly better, although the remaining differences become particularly clear for the predicted power at higher density in ITER. The estimate by GLS is higher than that provided by MAP, especially for power-law scaling.

In addition, and quite remarkably, when comparing the coefficient estimates and predictions obtained by GLS between the linear and nonlinear case, relatively consistent results can be noted. The same goes for the MAP estimates. In contrast, OLS provides widely different results, depending on whether a linear (log-transformed) or nonlinear (power-law) model is used. The relatively good consistency of the GLS estimates across regression models is a solid argument in favor of the method.

Another noteworthy point comes from the results of the two additional tests with increased and averaged error bars. They indicate that for MAP (and maximum likelihood) regression, reliable estimates of the variability of the measurements is important. However, as discussed in Section 3, the standard error bars that were used in the analysis in Sections 5.1 and 5.2 are small compared to the actual variability of the data around the theoretical scaling law. Hence, one could speculate whether the results of the MAP analysis are in fact trustworthy, given its sensitivity to the error bars on the data. Therefore, at least for MAP, it would be better to encode the available information on the error bars in sufficiently wide prior distributions (which, incidentally, would be possible for GLS, too).

A related comment is that GLS is clearly less vulnerable to inaccurate error specification compared to MAP. The mechanism behind this behavior is similar to the one that makes GLS less sensitive to outliers, *i.e.*, the observed standard deviation is able to capture deviations from the expected data variability with respect to the model. In the simple implementation of the GLS method used in the present paper, the distinction that is made between the modeled and observed standard deviation is the main difference between GLS and MAP.

## 6. Conclusions

Regression and scaling laws represent crucial tools in science in general and in the analysis of complex physical systems in particular. We have presented geodesic least squares regression (GLS) as a method that is able to handle large uncertainties on the data and on the regression model, and we have demonstrated its application to power-law regression. Operating on a manifold of probability distributions, GLS has the advantage that its results can be easily visualized in the case of the univariate Gaussian distribution. However, GLS is sufficiently flexible to allow tackling much more general regression problems within the same framework.

We have shown two examples of the enhanced robustness of the method using synthetic data. GLS showed a better stability in the presence of outliers and under a logarithmic transformation of a power-law, compared to established techniques. In addition, we have addressed the scaling of the L-H

power threshold in magnetically-confined fusion plasmas. On the basis of data from a multi-machine database, it was observed that geodesic least squares provides estimates of regression parameters and predictions that are consistent across different regression models, in contrast to ordinary least squares. Furthermore, because GLS allows the data uncertainty predicted by the model to be different from the empirically observed uncertainty, whereas with maximum *a posteriori* they are, by design, the same, GLS is more flexible and robust at the same time. As a consequence, the degrees of freedom provided by the parameters of the regression model better serve their actual purpose: to parameterize a model that best describes a trend in the data, with minimal distraction by the data "noise".

In future work, we intend to present a more general formulation of geodesic least squares, targeted at a wider class of regression problems. In addition, various theoretical performance issues need to be addressed, including uniqueness and convergence properties of the optimization problem, asymptotic behavior of the parameter estimates, *etc*. On the practical side, we aim at establishing a broader basis for the performance of the GLS method on simulated data. This should increase the confidence over a wider range of regression problems, as well as deviations from the regression model.

Finally, although we have noted that GLS performs regression on a probabilistic manifold, we have actually made little use of the geometrical structure of the manifold, save for calculating geodesic distances. Nowadays, there are various schemes, more sophisticated than a least-squares approach, to perform regression on manifold-valued data. From that point of view, one can expect advantages of a method performing regression between probability distributions, each of them containing more information than structureless data points in a Euclidean space. One possibility that we will explore in future work is a Bayesian regression method on a probabilistic manifold, by describing the distribution corresponding to the regression model intrinsically on the manifold [37]. At the same time, this will provide uncertainty estimates on the parameters through the posterior distribution.

**Acknowledgments**

The author wishes to acknowledge the ITPA Topical Groups on Transport and Confinement and on Pedestal and Edge Physics for maintaining and kindly providing the data in the H-mode threshold databases.

**Conflicts of Interest**

The author declares no conflict of interest.

**References**

1. Doyle, E.J.; Houlberg, W.A.; Kamada, Y.; Mukhovatov, V.; Osborne, T.H.; Polevoi, A.; Bateman, G.; Connor, J.W.; Cordey, J.G.; Fujita,T.; *et al.* Chapter 2: Plasma confinement and transport. *Nucl. Fusion* **2007**, *47*, S18–S127.
2. Xiao, X.; White, E.P.; Hooten, M.B.; Durham, S.L. On the use of log-transformations *vs.* nonlinear regression for analyzing biological power laws. *Ecology* **2011**, *92*, 1887–1894.

3. McDonald, D.; Meakins, A.J.; Svensson, J.; Kirk, A.; Cordey, J.G.; ITPA H-mode Threshold Database WG. The impact of statistical models on scalings derived from multi-machine H-mode threshold experiments. *Plasma Phys. Control. Fusion* **2006**, *48*, A439–A447.
4. Verdoolaege, G. Geodesic least squares regression on information manifolds. *AIP Conf. Proc.* **2013**, *1636*, 43–48.
5. Verdoolaege, G. Geodesic least squares regression for scaling studies in magnetic confinement fusion. *AIP Conf. Proc.* **2014**, *1641*, 564–571.
6. Basu, A.; Shioya, H.; Park, C. *Statistical Inference: The Minimum Distance Approach*; Chapman & Hall/CRC: Boca Raton, FL, USA, 2011; Volume 120.
7. McCullagh, P.; Nelder, J. *Generalized Linear Models*, 2nd ed.; Chapman & Hall/CRC: Boca Raton, FL, USA, 1989; Volume 37.
8. Amari, S.; Nagaoka, H. *Methods of Information Geometry*; American Mathematical Society: New York, NY, USA, 2000.
9. We follow standard notational practice from differential geometry with respect to index placement in the following definitions for the metric, Christoffel symbols and geodesic distance. However, in the remainder of the paper we will revert to subscript indices only, in order to avoid other notational problems.
10. Oprea, J. *Differential Geometry and Its Applications*, 2nd ed.; The Mathematical Association of America: Washington, DC, USA, 2007.
11. Verdoolaege, G.; Scheunders, P. On the geometry of multivariate generalized Gaussian models. *J. Math. Imaging Vis.* **2011**, *43*, 180–193.
12. Kass, R.; Vos, P. *Geometrical Foundations of Asymptotic Inference*; Wiley: New York, NY, USA, 1997.
13. Verdoolaege, G.; Scheunders, P. Geodesics on the manifold of multivariate generalized Gaussian distributions with an application to multicomponent texture discrimination. *Int. J. Comput. Vis.* **2011**, *95*, 265–286.
14. Kullback, S. *Information Theory and Statistics*; Dover Publications: New York, NY, USA, 1968.
15. Atkinson, C.; Mitchell, A. Rao's distance measure. *Indian J. Stat.* **1981**, *48*, 345–365.
16. Burbea, J.; Rao, C. Entropy differential metric, distance and divergence measures in probability spaces: A unified approach. *J. Multivar. Anal.* **1982**, *12*, 575–596.
17. Nielsen, F.; Nock, R. Visualizing hyperbolic Voronoi diagrams. In Proceedings of the 30th Annual Symposium on Computational Geometry (SOCG'14), Kyoto, Japan, 8–1 June 2014; p. 90.
18. Beran, R. Minimum Hellinger distance estimates for parametric models. *Ann. Stat.* **1977**, *5*, 445–463.
19. Pak, R. Minimum Hellinger distance estimation in simple regression models; distribution and efficiency. *Stat. Probab. Lett.* **1996**, *26*, 263–269.
20. Rao, C. Differential metrics in probability spaces. In *Differential Geometry in Statistical Inference*; Institute of Mathematical Statistics: Hayward, CA, USA, 1987.

21. Gill, P.; Murray, W.; Wright, M. *Numerical Linear Algebra and Optimization*; Addison Wesley: Boston, MA, USA, 1991; Volume 1.
22. Casella, G.; Berger, R. *Statistical Inference*, 2nd ed.; Cengage Learning: Hampshire, UK, 2002.
23. Snipes, J.A.; Greenwald, M.; Ryter, F.; Kardaun, O.J.W.F.; Stober, J.; Valovic, M.; Valovic, S.J.; Sykes, A.; Dnestrovskij, A.; Walsh, M.; *et al.* Multi-Machine global confinement and H-mode threshold analysis. In Proceedings of the 19th IAEA Fusion Energy Conference, Lyon, France, 14–19 October 2002.
24. Martin, Y.R.; Takizuka, T.; The ITPA CDBM H-mode Threshold Database Working Group. Power requirements for accessing the H-mode in ITER. *J. Phys. Conf. Ser.* **2008**, *123*, 012033.
25. Ryter, F.; The H-Mode Database Working Group. H Mode power threshold database for ITER. *Nucl. Fusion* **1996**, *36*, 1217–1264.
26. Ryter, F.; The H-Mode Threshold Database Group. Progress of the international H-Mode power threshold database activity. *Plasma Phys. Control. Fusion* **2002**, *44*, A415–A421.
27. ITPA—Threshold database. Available online: http://efdasql.ipp.mpg.de/threshold (accessed on 30 June 2015).
28. Whereas the most recent update of the database dates from 2008 [24], we used the earlier version from 2002, because it allows a better illustration of the advantages of GLS with respect to other methods. The reason is that the data in the most recent version is significantly better conditioned, in which case even a simple regression technique such as OLS turns out to be able to provide acceptable estimates of the regression parameters. This point is not relevant for the present discussion, as here our aim is to demonstrate the advantages of GLS in cases where the data are not in the best shape.
29. Verdoolaege, G.; Karagounis, G.; Tendler, M.; van Oost, G. Pattern recognition in probability spaces for visualization and identification of plasma confinement regimes and confinement time scaling. *Plasma Phys. Control. Fusion* **2012**, *54*, 124006.
30. Preuss, R.; Dose, V. Errors in all variables. *AIP Conf. Proc.* **2005**, *803*, 448–455.
31. Markovsky, I.; van Huffel, S. Overview of total least-squares methods. *Signal Process.* **2007**, *87*, 2283–2302.
32. Maronna, R.; Martin, D.; Yohai, V. *Robust Statistics: Theory and Methods*; Wiley: New York, NY, USA, 2006.
33. *MATLAB and Statistics Toolbox Release 2015a*; The Mathworks Inc: Natick, MA, USA, 2015.
34. We use the notation $\eta$ for the response variable instead of $P_{\mathrm{thr}}$ because in this experiment $\eta$ is generated artificially and therefore it is not necessarily related to the actual power threshold in fusion devices.
35. Von Toussaint, U.; Frey, M.; Gori, S. Fitting of functions with uncertainties in dependent and independent variables. *AIP Conf. Proc.* **2009**, *1193*, 302–310.
36. OLS is not repeated here because it does not depend on the error bars.
37. Pennec, X. Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. *J. Math. Imaging Vis.* **2006**, *25*, 127–154.

# On Monotone Embedding in Information Geometry

**Jun Zhang**

**Abstract:** A paper was published (Harsha and Subrahamanian Moosath, 2014) in which the authors claimed to have discovered an extension to Amari's $\alpha$-geometry through a general monotone embedding function. It will be pointed out here that this so-called $(F, G)$-geometry (which includes $F$-geometry as a special case) is identical to Zhang's (2004) extension to the $\alpha$-geometry, where the name of the pair of monotone embedding functions $\rho$ and $\tau$ were used instead of $F$ and $H$ used in Harsha and Subrahamanian Moosath (2014). Their weighting function $G$ for the Riemannian metric appears cosmetically due to a rewrite of the score function in log-representation as opposed to $(\rho, \tau)$-representation in Zhang (2004). It is further shown here that the resulting metric and $\alpha$-connections obtained by Zhang (2004) through arbitrary monotone embeddings is a unique extension of the $\alpha$-geometric structure. As a special case, Naudts' (2004) $\phi$-logarithm embedding (using the so-called $\log_\phi$ function) is recovered with the identification $\rho = \phi, \tau = \log_\phi$, with $\phi$-exponential $\exp_\phi$ given by the associated convex function linking the two representations.

Reprinted from *Entropy*. Cite as: Zhang, J. On Monotone Embedding in Information Geometry. *Entropy* **2015**, *17*, 4485–4499.

In a recent paper that appeared in *Entropy* (Harsha and Subrahamanian Moosath, 2014) [1], the authors proposed an extension to Amari's $\alpha$-geometry, which they call $F$- or $(F, G)$-geometry, where $F$ is a monotone embedding function and $G$ is the weighting function for taking the expectation of random variables in calculating the Riemannian metric ($G = 1$ reduces to $F$-geometry, with the standard Fisher–Rao metric). This paper serves the purpose of pointing out that $(F, G)$-geometry as proposed is the same as what Zhang (2004) [2] has obtained for extending the $\alpha$-geometry and captured in his subsequent work [4–8]. The metric and affine connections proposed by [1] are identical to [2] apart from the notations: the embedding functions $F$ and $H$ in [1] were denoted as $\rho$ and $\tau$ in [2], and weighting function $G$ in [1] is a trivial rewriting of the convex function $f$ used by [2].

This paper will start in Section 1 with a review of Amari's $\alpha$-geometry and $\alpha$-embedding, a review of Zhang's (2004) [2] extension to $\rho$-embedding with an arbitrary monotone function and a summary of Harsha and Subrahamanian Moosath (2014) [1]. Then, the equivalence of [1] to [2] is shown. In Section 2, after analyzing the group of monotone embedding functions, a stronger statement is made: the construction of [2] is a unique dualistic extension of Amari's $\alpha$-geometry through arbitrary monotone embedding in place of $\alpha$-embedding. As an important special case, we illustrate how the deformed logarithm $\log_\phi$ associated with an arbitrary strictly increasing function $\phi$ as investigated by Naudts (2004) [3] arises naturally from identifying $\phi$ with $\rho$ and with a proper choice of the auxiliary function $f$ as a part of Zhang's theory.

## 1. Equivalence of $(F, G)$-Geometry to Zhang's (2004) [2] $(\rho, \tau)$-Geometry

### *1.1. Amari's $\alpha$-Geometry and $\alpha$-Embedding*

The now standard differential geometric characterization of the manifold $\mathcal{M}_\Theta = \{\, p(\cdot \mid \theta), \theta \in \Theta \subseteq \mathbb{R}^n\}$ of parametric probability functions $p$ (probability density or probability distributions) is through the Fisher–Rao metric $g_{ij}$ as its Riemannian metric:

$$g_{ij}(\theta) = \mathrm{E}_\mu \left\{ p(\zeta|\theta)\, \frac{\partial \log p(\zeta|\theta)}{\partial \theta^i} \frac{\partial \log p(\zeta|\theta)}{\partial \theta^j} \right\} \tag{1}$$

and a family of $\alpha$-connections (given by Amari [9,10]) with coefficients $\Gamma^{(\alpha)}$ ($\alpha \in \mathbb{R}$):

$$\Gamma^{(\alpha)}_{ij,k}(\theta) = \mathrm{E}_\mu \left\{ \left( \frac{1-\alpha}{2}\, \frac{\partial \log p(\zeta|\theta)}{\partial \theta^i}\, \frac{\partial \log p(\zeta|\theta)}{\partial \theta^j} + \frac{\partial^2 \log p(\zeta|\theta)}{\partial \theta^i \partial \theta^j} \right) \frac{\partial p(\zeta|\theta)}{\partial \theta^k} \right\} . \tag{2}$$

Here, $\mathrm{E}_\mu$ denotes the expectation with respect to a background measure $\mu$ of the random variable denoted by $\zeta$:

$$\mathrm{E}_\mu\{\cdot\} = \int (\cdot) d\mu(\zeta). \tag{3}$$

The $\alpha$-connection is constructed as a convex combination of a pair of conjugate connections $\Gamma, \Gamma^*$

$$\Gamma^{(\alpha)}_{ij,k}(\theta) = \frac{1+\alpha}{2} \Gamma_{ij,k}(\theta) + \frac{1-\alpha}{2} \Gamma^*_{ij,k}(\theta), \tag{4}$$

where $\Gamma \equiv \Gamma^{(1)}$ is frequently called $e$-connection ($\alpha = 1$) and $\Gamma^* \equiv \Gamma^{(-1)}$ called $m$-connection ($\alpha = -1$). A Riemannian manifold $\mathcal{M}_\mu$ with its metric $g$ and the family of $\alpha$-connections $\Gamma^{(\alpha)}$ in the form of (1) and (2) has been called $\alpha$-geometry. Amari's $\alpha$-geometry can be specified in terms of a symmetric $(0,2)$-tensor $g_{ij}$ (the Fisher–Rao metric) and a totally symmetric $(0,3)$-tensor $T_{ijk}$ (sometimes called the Amari–Chentsov tensor), which is linked to the $\alpha$-connections via:

$$\Gamma^{(\alpha)}_{ij,k} = \Gamma^{LC}_{ij,k}(\theta) - \frac{\alpha}{2}\, T_{ijk}(\theta) \ , \tag{5}$$

where $\Gamma^{LC}_{ij,k}$ is the Levi–Civita connection corresponding to the Riemannian metric $g$.

As an extension of the logarithmic embedding $l(p) = \log p$ of probability density function $p$, an $\alpha$-embedding function [10] is defined through $l^{(\alpha)} : \mathbb{R}^+ \to \mathbb{R}$:

$$l^{(\alpha)}(t) = \begin{cases} \log t & \alpha = 1 \\ \frac{2}{1-\alpha}\, t^{(1-\alpha)/2} & \alpha \neq 1 \end{cases} . \tag{6}$$

It is an interesting observation (e.g., p. 46 in [11]) that the $\alpha$-geometry can be recovered under such $\alpha$-representation (scaling) of the probability function, that is the Fisher–Rao metric turns out to be $\alpha$-independent (*i.e.*, embedding independent) and the $\pm 1$-connections precisely the $\alpha$-connections:

$$g_{ij}(\theta) \;=\; \mathrm{E}_\mu \left\{ \frac{\partial l^{(\alpha)}(p(\zeta|\theta))}{\partial \theta^i}\, \frac{\partial l^{(-\alpha)}(p(\zeta|\theta))}{\partial \theta^j} \right\} , \tag{7}$$

$$\Gamma^{(\alpha)}_{ij,k}(\theta) \;=\; \mathrm{E}_\mu \left\{ \frac{\partial^2 l^{(\alpha)}(p(\zeta|\theta))}{\partial \theta^i \partial \theta^j}\, \frac{\partial l^{(-\alpha)}(p(\zeta|\theta))}{\partial \theta^k} \right\} . \tag{8}$$

A variance of $\alpha$-embedding of a probability function plays an important role in Tsallis statistics; see [12–14]. On the geometric side, [15,16] illuminated that the $\alpha$-scaling of the probability functions leads to a conformal transformation.

*1.2. Zhang (2004) [2] Extension: $\rho$-Embedding and $(\rho, \tau)$-Geometry*

Zhang [2,4,6] obtained generalizations of the $\alpha$-geometry for a pair of monotone embeddings, called $\rho$- and $\tau$-embeddings generalizing $\alpha$-embedding. Given any smooth strictly convex function $f : \mathbb{R} \to \mathbb{R}$, with convex conjugate $f^*$ given by:

$$f^*(t) = t\,(f')^{-1}(t) - f((f')^{-1}(t)) \ , \tag{9}$$

Zhang (2004) defines a pair of conjugate representations [2] (Section 3.2) using two strictly increasing functions $\rho, \tau$ from $\mathbb{R} \to \mathbb{R}$:

(1) we call $\rho$-representation of a probability function $p$ the mapping $p \mapsto \rho(p)$;

(2) we say $\tau$-representation of the probability function $p \mapsto \tau(p)$ is conjugate to $\rho$-representation with respect to a smooth and strictly convex function $f$, or simply $\tau$ is $f$-conjugate to $\rho$, if:

$$\tau(p) = f'(\rho(p)) = ((f^*)')^{-1}(\rho(p)) \ , \tag{10}$$

which can be equivalently written as:

$$\rho(p) = (f')^{-1}(\tau(p)) = (f^*)'(\tau(p)) \ . \tag{11}$$

These equalities in (10) and (11) hold, and they are equivalent, because $f'$ and $(f^*)'$ are both strictly increasing (due to their strict convexity) and that $(f^*)^* = f, \ (f^*)' = (f')^{-1}$. Sometimes, we write $f' = \sigma, (f^*)^{-1} = \sigma^{-1}$ for convenience, so $\sigma(\rho) = \tau, \sigma^{-1}(\tau) = \rho$, for a strictly increasing function $\tau$.

As a first example, we may set $\rho(t) = t, \tau(t) = \log t$. Then, we can derive that $f^*(t) = \exp(t)$ and $f(t) = t \log t - t + 1$. That $\rho(p)$ and $\tau(p)$ are just the $p$ and $\log p$ representation reflects the conventional dual embeddings that have later been extended to $\phi$- and $\log_\phi$-embedding in ([3]). In Section 2.2, it will be shown that Naudts' $\phi$-logarithm formulation is recovered as a special case of the $(\rho, \tau)$-embedding.

As another example, we may set $\rho(p) = l^{(\beta)}(p)$ to be the $\beta$-representation given by Equation (6); this would have been traditionally called "alpha-embedding", except we use the symbol $\beta$, so that the $\alpha$-parameter will be reserved for indexing $\alpha$-connections. In this case, the conjugate representation is the $(-\beta)$-representation $\tau(p) = l^{(-\beta)}(p)$:

$$\rho(p) = l^{(\beta)}(p) \longleftrightarrow \tau(p) = l^{(-\beta)}(p) \ . \tag{12}$$

In this case, $\rho$ and $\tau$ are conjugate with respect to $f$, where $f$ is given by:

$$f(t) = \frac{2}{1+\beta}\left(\left(\frac{1-\beta}{2}\right)t\right)^{\frac{2}{1-\beta}} \ , \quad f^*(t) = \frac{2}{1-\beta}\left(\left(\frac{1+\beta}{2}\right)t\right)^{\frac{2}{1+\beta}} . \tag{13}$$

Based on divergence functions constructed under monotone embedding, Zhang ([2]) showed:

**Proposition 1.** ([2], Proposition 7) *Using an arbitrary monotone embedding function $\rho$ and an arbitrary smooth strictly convex function $f$, a generalization of $\alpha$-geometry is obtained, with metric and $\alpha$-connections taking the form:*

$$g_{ij}(\theta) = E_\mu \left\{ f''(\rho(p(\zeta|\theta)) \frac{\partial \rho(p(\zeta|\theta))}{\partial \theta^i} \frac{\partial \rho(p(\zeta|\theta))}{\partial \theta^j} \right\} \tag{14}$$

$$\Gamma^{(\alpha)}_{ij,k}(\theta) = E_\mu \left\{ \frac{1-\alpha}{2} f'''(\rho(p(\zeta|\theta))) A_{ijk} + f''(\rho(p(\zeta|\theta))) B_{ijk} \right\} , \tag{15}$$

*where:*

$$A_{ijk}(\zeta,\theta) = \frac{\partial \rho(p(\zeta|\theta))}{\partial \theta^i} \frac{\partial \rho(p(\zeta|\theta))}{\partial \theta^j} \frac{\partial \rho(p(\zeta|\theta))}{\partial \theta^k} , \quad B_{ijk}(\zeta,\theta) = \frac{\partial^2 \rho(p(\zeta|\theta))}{\partial \theta^i \partial \theta^j} \frac{\partial \rho(p(\zeta|\theta))}{\partial \theta^k} . \tag{16}$$

As special cases,

$$\Gamma_{ij,k}(\theta) = \mathrm{E}_\mu \left\{ f''(\rho(p)) \frac{\partial^2 \rho(p)}{\partial \theta^i \partial \theta^j} \frac{\partial \rho(p)}{\partial \theta^k} \right\} , \tag{17}$$

$$\Gamma^*_{ij,k}(\theta) = \mathrm{E}_\mu \left\{ \frac{\partial \rho(p)}{\partial \theta^k} \left( f'''(\rho(p)) \frac{\partial \rho(p)}{\partial \theta^i} \frac{\partial \rho(p)}{\partial \theta^j} + f''(\rho(p)) \frac{\partial^2 \rho(p)}{\partial \theta^i \partial \theta^j} \right) \right\} . \tag{18}$$

Furthermore, taking a pair of monotone representations, the metric tensor and affine connections stated in Proposition 1 have dualistic expressions:

**Corollary 1.** ([2], Proposition 8) *Using two arbitrary monotone embedding functions $\rho$ and $\tau$, the metric and $\alpha$-connections of (14)–(16) are:*

$$g_{ij}(\theta) = E_\mu \left\{ \frac{\partial \rho(p(\zeta|\theta))}{\partial \theta^i} \frac{\partial \tau(p(\zeta|\theta))}{\partial \theta^j} \right\} , \tag{19}$$

$$\Gamma^{(\alpha)}_{ij,k}(\theta) = E_\mu \left\{ \frac{1-\alpha}{2} \frac{\partial^2 \tau(p(\zeta,\theta))}{\partial \theta^i \partial \theta^j} \frac{\partial \rho(p(\zeta|\theta))}{\partial \theta^k} + \frac{1+\alpha}{2} \frac{\partial^2 \rho(p(\zeta|\theta))}{\partial \theta^i \partial \theta^j} \frac{\partial \tau(p(\zeta|\theta))}{\partial \theta^k} \right\} . \tag{20}$$

As a special case, when $\rho, \tau$ take the familiar alpha-embeddings (12) (using $\beta$ as the parameter), the $\alpha$-connections becomes $(\alpha\beta)$-connections:

$$\Gamma^{(\alpha)}_{ij,k}(\theta) = \mathrm{E}_\mu \left\{ \left( \frac{1-\alpha\beta}{2} \frac{\partial \log p(\zeta|\theta)}{\partial \theta^i} \frac{\partial \log p(\zeta|\theta)}{\partial \theta^j} + \frac{\partial^2 \log p(\zeta|\theta)}{\partial \theta^i \partial \theta^j} \right) \frac{\partial p(\zeta|\theta)}{\partial \theta^k} \right\} , \tag{21}$$

with the product $\alpha \cdot \beta$ playing the role of the alpha-parameter indexing the family of connections.

### *1.3. Harsha and Subrahamanian Moosath's (2014) Work [1]*

Using a monotone embedding function denoted as $F$ and a weighting function denoted as $G$ ($G = 1$ is a special case to reduce to what they called $F$-geometry), these authors [1] proposed $(F, G)$-metric as (their Equation (33) in [1]):

$$g^{F,G}_{ij} = \mathrm{E}_\mu \left\{ pG(p) \frac{\partial \log p}{\partial \theta^i} \frac{\partial \log p}{\partial \theta^j} \right\} \tag{22}$$

with affine connection given as (their Equation (34)):

$$\Gamma^{F,G}_{ijk} = \mathrm{E}_\mu \left\{ p\, G(p) \frac{\partial \log p}{\partial \theta^k} \left( \frac{\partial^2 \log p}{\partial \theta^i \partial \theta^j} + \left( 1 + \frac{pF''(p)}{F'(p)} \right) \frac{\partial \log p}{\partial \theta^i} \frac{\partial \log p}{\partial \theta^j} \right) \right\}. \tag{23}$$

Note $\mathrm{E}_p\{(\cdot)\} = \mathrm{E}_\mu\{(\cdot)p\}$. (23) is the expression for the $e$-connection ($\alpha = 1$), $\Gamma^{F,G}_{ijk}$. To express the conjugate connection ($m$-connection, $\alpha = -1$), $\Gamma^{H,G}_{ijk}$, a dual embedding function $H$ is introduced, which is shown ([1], Theorem 3.2) to be related to $F$ and $G$ via (their Equation (36)):

$$H'(p) = \frac{G(p)}{pF'(p)}. \tag{24}$$

In such a case, the conjugate connection $\Gamma^{H,G}_{ijk}$ (*sic*, more accurately $(\Gamma^{F,G})^*_{ijk}$) is expressed as (their Equation (37)):

$$\Gamma^{H,G}_{ijk} = \mathrm{E}_\mu \left\{ pG(p) \frac{\partial \log p}{\partial \theta^k} \left( \frac{\partial^2 \log p}{\partial \theta^i \partial \theta^j} + \left( \frac{pG'(p)}{G(p)} - \frac{pF''(p)}{F'(p)} \right) \frac{\partial \log p}{\partial \theta^i} \frac{\partial \log p}{\partial \theta^j} \right) \right\}. \tag{25}$$

We now show the equivalence of the three expressions (14), (17), (18) from the work [2] with the three corresponding expressions (22), (23), (25) from the work [1].

**Statement 1.** *Equations (14) and (22) give the same Riemannian metric; Equations (17) and (23) give the same affine connection; and Equations (18) and (25) give the same conjugate connection, as long as:*

$$F(p) = \rho(p)\ , \quad G(p) = (\rho')^2\, p\, f''(\rho(p)). \tag{26}$$

**Proof.** Re-writing (14), and keeping in mind:

$$\frac{\partial \rho(p)}{\partial \theta^i} = \rho'(p)\, \frac{\partial p}{\partial \theta^i} = p\, \rho'(p)\, \frac{\partial \log p}{\partial \theta^i}, \tag{27}$$

so:

$$g_{ij}(\theta) = \mathrm{E}_\mu \left\{ f''(\rho(p)) (p\, \rho'(p))^2\, \frac{\partial \log p}{\partial \theta^i}\, \frac{\partial \log p}{\partial \theta^j} \right\}. \tag{28}$$

Comparing the above with (22), obviously, $F$ is just $\rho$, and $G$ is linked to $f$ and $\rho$:

$$G(p) = (\rho')^2\, p\, f''(\rho(p)) = p\, \rho'(p) \tau'(p) \tag{29}$$

where we have used (10).

Next, differentiate (27); we obtain:

$$\begin{aligned}
\frac{\partial^2 \rho(p)}{\partial \theta^i \partial \theta^j} &= \frac{\partial p}{\partial \theta^j} \rho'(p) \frac{\partial \log p}{\partial \theta^i} + p\, \rho''(p) \frac{\partial p}{\partial \theta^j} \frac{\partial \log p}{\partial \theta^i} + p\, \rho'(p) \frac{\partial^2 \log p}{\partial \theta^i \partial \theta^j} && (30) \\
&= p\, \rho'(p) \left( \frac{\partial \log p}{\partial \theta^i} \frac{\partial \log p}{\partial \theta^j} + \frac{\partial^2 \log p}{\partial \theta^i \partial \theta^j} + \frac{p\rho''(p)}{\rho'(p)} \frac{\partial \log p}{\partial \theta^i} \frac{\partial \log p}{\partial \theta^j} \right) && (31) \\
&= p\, \rho'(p) \left( \frac{\partial^2 \log p}{\partial \theta^i \partial \theta^j} + \left( 1 + \frac{p\rho''(p)}{\rho'(p)} \right) \frac{\partial \log p}{\partial \theta^i} \frac{\partial \log p}{\partial \theta^j} \right). && (32)
\end{aligned}$$

Identifying $F = \rho$ and making use of (29), we see that (17) is precisely (23).

Finally, differentiate (29),

$$G'(p) = (\rho')^2 f''(\rho(p)) + (\rho')^3 \, p \, f'''(\rho(p)) + 2\rho'(p)\rho''(p) \, p \, f''(\rho(p)). \tag{33}$$

Therefore,

$$\frac{pG'(p)}{G(p)} - \frac{pF''(p)}{F'(p)} = 1 + \frac{p\rho'(p)f'''(\rho(p))}{f''(\rho(p))} + \frac{p\rho''(p)}{\rho'(p)}. \tag{34}$$

After substituting (34) and (29) into (25) and making use of (31), the expression (18) results. □

**Statement 2.** *The conjugate embedding function $H$ is the same as $\tau$. The conjugate connection (25), when expressed using $H$, has the same form as (23) for $\Gamma_{ij,k}^{G,F}$ using $F$.*

**Proof.** Applying Definition (24) immediately yields $H' = \tau'$. Therefore, (apart from constant) $H(p) = \tau(p)$. Next, we will express (25) explicitly using the conjugate embedding function $H$ (rather than $F$) and the weighting function $G$. That is to say, we will simplify the terms in the middle parenthesis of (25):

$$\frac{pG'(p)}{G(p)} - \frac{pF''(p)}{F'(p)} = p\left(\log\frac{G(p)}{F'(p)}\right)' = p\left(\log(pH'(p)\right)' = p\left(\log p + \log H'(p)\right)' \tag{35}$$

$$= p\left(\frac{1}{p} + \frac{H''(p)}{H'}\right) = 1 + \frac{pH''}{H'(p)} \, . \tag{36}$$

Hence, (25) has the same expression as (23) showing the duality between the embedding function $H$ and the embedding function $F$. □

By Statement 1, starting from $F$ (that is, $\rho$) and $G$ and imposing conjugacy requirement on the pair of affine connections, one is guaranteed to derive $H$ (that is, $\tau$) as the conjugate embedding function.

From Statements 1 and 2, we conclude that, Harsha and Moosath's $F$-embedding [1] replicates the $\rho$-embedding of Zhang (2004) [2]; the conjugate $H$-embedding turns out to be identical to $\tau$-embedding of [2]. Contrary to the authors' claim (Remark 3.7 of [1], p. 2480), $(F, G)$-geometry is identical to Zhang's $(\rho, \tau)$ geometry [2]. In particular, their $F$-geometry is recovered by simply choosing $f$ to satisfy $f''(t) = 1/(\rho^{-1}(t)\,(\rho'(\rho^{-1}(t)))^2)$, for a given $\rho$. The subsequent development in their paper [1], e.g., the definition of the $F$-affine manifold (their Equation (50)), replicates the definition of $\rho$-affine manifold in [2] (Section 3.4).

During the review of their manuscript [1] and in subsequent personal communications, these authors argued that they used a different approach: $(F, G)$-geometry is derived by embedding the manifold into the space of random variables and suitably defining the inner product through using the $F$-expectation (their Equation (15)) and $(F, G)$-expectation (their Equation (32)) as a general weighted expectation of a random variable, while Zhang (2004) [2] derived the geometry through constructing a divergence function. This difference, however, is entirely superficial, because the relationship between divergence functions and geometric structure (metric and affine connection) is well-established by Eguchi's work [17,18] and known to information geometers. Therefore, neither the approach nor the results of Harsha and Moosath's proposed $(F, H, G)$ extension to Amari's

$\alpha$-geometry differs from Zhang's proposed $(\rho, \tau, f)$ extension, with the following correspondence in different symbols by the two papers:

$$F \Longleftrightarrow \rho \ , \quad H \Longleftrightarrow \tau \ , \tag{37}$$

$$G(t) \Longleftrightarrow t\rho'(t)\tau'(t) = tf''(\rho(t))(\rho'(t))^2 = t(f^*)''(\tau(t))(\tau'(t))^2 \ ; \tag{38}$$

the difference in the representation of score function as $\log$-representation in [1] or under $\rho$ or $\tau$-representation in [2] is cosmetic.

## 2. Uniqueness of $(\rho, \tau)$-Geometry and Representation Duality

### *2.1. Monotone Embedding as a Transformation Group*

Monotone representations of any given probability function form a transformation group, with functional composition as group composition operation and the functional inverse as the group inverse operation. This was pointed out by Zhang [6] (Section 2.2.2). We state it as a lemma here.

**Lemma 1.** *Denote $\Omega$ as the set of strictly increasing functions from $\mathbb{R} \to \mathbb{R}$. Then, $(\Omega, \circ)$ forms a group, with $\circ$ denoting functional composition.*

**Proof.** We easily verify that:

(1) closure for $\circ$: for any $\rho_1, \ \rho_2 \in \Omega$, $\rho_2 \circ \rho_1$, defined as $\rho_2(\rho_1(\cdot))$, is strictly increasing, and hence, $\rho_2 \circ \rho_1 \in \Omega$;

(2) existence of unique identity element: the identity function $\iota$, which satisfies $\rho \circ \iota = \iota \circ \rho = \rho$, is strictly increasing, and hence, $\iota \in \Omega$ and is unique;

(3) existence of inverse: for any $\rho \in \Omega$, its functional inverse $\rho^{-1}$, which satisfies $\rho^{-1} \circ \rho = \rho^{-1} \circ \rho = \iota$, is also strictly increasing, and hence, $\rho^{-1} \in \Omega$;

(4) associativity of $\circ$: for any three $\rho_1, \ \rho_2, \ \rho_3 \in \Omega$, then $(\rho_1 \circ \rho_2) \circ \rho_3 = \rho_1 \circ (\rho_2 \circ \rho_3)$.

□

Recall that the derivative of smooth strictly convex functions are strictly increasing functions. From this perspective, $f' = \tau \circ \rho^{-1} = \tau(\rho^{-1}(\cdot))$, $(f^*)' = \rho \circ \tau^{-1} = \rho(\tau^{-1}(\cdot))$, encountered above, are themselves two mutually inverse strictly increasing functions. This is the rationale behind Zhang's ([2]) choice of $f$ (and $f^*$) as the auxiliary function to capture conjugate embedding, rather than using $G$ as in [1]. The following identities are useful; they are obtained by differentiating (10) and (11):

$$f''(\rho(t))\, \rho'(t) = \tau'(t) \ , \quad (f^*)''(\tau(t))\, \tau'(t) = \rho'(t) \ ; \tag{39}$$

therefore:

$$f''(\rho(t))\, (\rho'(t))^2 = (f^*)''(\tau(t))\, (\tau'(t))^2 \ , \tag{40}$$

and:

$$f''(\rho(t))\,(f^*)''(\tau(t)) = 1. \tag{41}$$

With respect to (41), taking $\log$ on both sides yields:

$$\log f''(\rho(t)) + \log(f^*)''(\tau(t)) = 0\ . \tag{42}$$

Move and differentiate:

$$\frac{f'''(\rho(t))\,\rho'(t)}{f''(\rho(t))} = -\frac{(f^*)'''(\tau(t))\,\tau'(t)}{(f^*)''(\tau(t))}\ . \tag{43}$$

Making use of (40) yields:

$$f'''(\rho(t))\,(\rho'(t))^3 = -(f^*)'''(\tau(t))\,(\tau'(t))^3\ . \tag{44}$$

Note the coupling between $f$ and $\rho, \tau$ given by (10), (11), (40) and (44). They allow us to cast (14) and (15) in terms of $f^*$ and $\tau$.

Among the triple $(f, \rho, \tau)$, given any two, the third is specified. In particular, if we arbitrary choose two strictly increasing functions $\rho$ and $\tau$ as embedding functions and require them to be conjugate embeddings, then $f$ is specified by $f'(t) = \tau(\rho^{-1}(t))$. In terms of conjugate function $f^*$, the relation is $(f^*)'(t) = \rho(\tau^{-1}(t))$. The function $f$ (or $f^*$) is important in constructing the general class of divergence function.

### *2.2. Naudts' $\phi$-Logarithm as a Special Case*

In his 2004 publication [3], Naudts considered the "deformed" logarithm function as an extension to the exponential family of densities that is log-linear. Given a strictly increasing and strictly positive function $\phi : \mathbb{R}_+ \to \mathbb{R}_+$, the $\phi$-logarithm is defined as:

$$\log_\phi(t) = \int_1^t \frac{1}{\phi(s)}\,ds\,, \quad (t > 0). \tag{45}$$

The deformed exponential denoted $\exp_\psi$, is defined by:

$$\exp_\psi(t) = 1 + \int_0^t \psi(s)\,ds. \tag{46}$$

(Naudts (2004) used the notation $\exp_\phi$, so our current rendition has a subtle difference shown as (48) and (49) below.) It can be shown that the deformed functions $\log_\phi$ and $\exp_\psi$ are in fact inverse functions of each other if:

$$\psi(\log_\phi(t)) = \phi(t), \qquad \psi(t) = \phi(\exp_\psi(t)). \tag{47}$$

Stated alternatively, the deformed logarithmic function $h(t) = \log_\phi(t)$ can be viewed as the solution to the following integral and its equivalent differential equation:

$$h(t) = \int_1^t \frac{1}{\psi(h(s))} ds \iff \frac{dh}{dt} = \frac{1}{\psi(h(t))}, \tag{48}$$

whereas the deformed exponential function $h(t) = \exp_\psi(t)$ can be viewed as the solution to the following integral and its equivalent differential equation:

$$h(t) = 1 + \int_0^t \phi(h(s))ds \iff \frac{dh}{dt} = \phi(h(t)). \tag{49}$$

We now show that the above formulation can be re-written as $(\rho, \tau)$-embeddings with a particular choice of $f$ (or equivalently, $f^*$) function. Set $\phi(t) = \rho(t)$ and $f^*(t) = \exp_\psi(t)$, so that $(f^*)'(t) = \psi(t)$ from (46). Therefore, we derive:

$$\log_\phi(t) = \psi^{-1}(\phi(t)) = ((f^*)')^{-1}(\rho(t)) = f'(\rho(t)) = \tau(t).$$

That is, when $\phi$ is chosen as $\rho$-representation, the deformed logarithm $\log_\phi$ turns out to be the $\tau$-representation, while the deformed exponential is nothing but $f^*$. The relationship (47) is identical to (10) and (11).

In the $\phi$-logarithm approach, once $\phi$ (that is, $\rho$) is specified, then $\log_\phi$ (that is, $\tau$) is specified, through the integral relation (45). Viewing $\tau(\cdot) = f'(\rho(\cdot))$, the relation (45) essentially specifies a strictly convex function $f$, through its derivative $f'$, which operates on $\rho$.

**Proposition 2.** *Denote $\rho \equiv \phi$. The deformed logarithmic transformation $\phi \to \log_\phi$ given by (45) can be viewed as the function composition $f' : \rho \to f'(\rho)$, where $f$ is given by:*

$$f(\rho(t)) = \rho(t) f'(\rho(t)) - t. \tag{50}$$

*Equivalently, using conjugate function $f^*$ given by (9),*

$$\rho = (f^*)' \circ (f^*)^{-1}, \tag{51}$$

*or*

$$\rho = \frac{1}{((f^*)^{-1})'}. \tag{52}$$

**Proof.** From (45), we write:

$$f'(\rho(t)) = \int_1^t \frac{1}{\rho(s)}\, ds, \tag{53}$$

with unknown $f$. Multiply both sides by $\rho'(t)$ and then integrate from one to $x$; the left-hand side of (53) is:

$$\int_1^x f'(\rho(t))\, \rho'(t)\, dt = \int_1^x f'(\rho(t)) d(\rho(t)) = f(\rho(x)) - f(\rho(1)).$$

The right-hand side of (53), after the same operation, is:

$$\int_1^x \rho'(t)\, dt \int_1^t \frac{1}{\rho(s)}\, ds = \int_1^x \frac{1}{\rho(s)}\, ds \int_s^x \rho'(t)\, dt = \int_1^x \frac{\rho(x) - \rho(s)}{\rho(s)}\, ds$$
$$= \int_1^x \left(\frac{\rho(x)}{\rho(s)} - 1\right) ds = \rho(x) \left(\int_1^x \frac{1}{\rho(s)}\, ds\right) - \int_1^x ds = \rho(x)\, f'(\rho(x)) - (x - 1).$$

Clearly, $f'(\rho(1)) = 0$ by (53). We set $f(\rho(1)) = -1$. Comparing expressions from the left- and right-hand side, we obtain (50).

Applying (9), we obtain the equivalent expression:

$$f^*(f'(\rho(t))) = t.$$

That is, $f$ is chosen, such that $f^* \circ f'$ is the inverse function of $\rho$, or:

$$\rho = (f^* \circ f')^{-1} = (f')^{-1} \circ (f^*)^{-1} = (f^*)' \circ (f^*)^{-1}.$$

Hence, (51) holds.

Finally, differentiate the identity:

$$f^*((f^*)^{-1}(t)) = t,$$

we obtain:

$$1 = (f^*)'((f^*)^{-1}(t)) \cdot (f^*)^{-1}(t) = \rho(t) \cdot (f^*)^{-1}(t)$$

upon substituting (51). Hence, (52) holds. □

The expression (51) in Proposition 2 shows that for any $\rho$, if one can find a decomposition: $\rho = g' \circ g^{-1}$ in terms of $g$, then $g$ would be the $\rho$-exponential, $g^{-1}$ the $\rho$-logarithm and $g'$ the linking function. In the case of $\phi \mapsto \log_\phi$ transformation, $g = f^*(t)$.

Naudts' ([3]) deformed logarithm/exponential embedding approach and Zhang's ([2]) $(\rho, \tau)$-embedding approach can be seen as playing complementary roles in information geometry: the former makes it easy to generalize the exponentiation and logarithm as inverse operations obeying desired differential/integral equations, while the latter makes it apparent how conjugate $(\rho, \tau)$-embeddings lead to bidualistic expressions for the underlying geometric structures (metric and conjugate connections).

### *2.3. Uniqueness of $(\rho, \tau)$-Geometry*

It is known [19,20] that the Fisher–Rao metric and $\alpha$-connections (equivalently, Amari–Chentsov tensor $T$) are the only invariants of sufficient statistics under the Markov morphism of a random variable. In [22,23], the Fisher–Rao metric has been extended to allow a weighting function. In [2,6], general weighting functions for affine connections were made compatible with the generalized (*i.e.*, weighted) Fisher–Rao metric, since they result from divergence functions that are allowed to have the freedom of monotone embedding. The recent reinvention [1] constructed weighted connections that turned out to be identical to the expressions given by [2]. A natural question is, then, whether Zhang's $(\rho, \tau)$ geometry is the unique construction given the freedom of arbitrary monotone embedding. Below, arguments will be provided, along with a proof, for a positive answer to this question.

First, when a probability function $p(\zeta|\theta)$ (as a function of a random variable indexed by $\zeta$ and a background measure of $\mu$) is embedded into the parametric manifold $\mathcal{M}_\Theta$, there are several traditional choices for tangent vectors: $\partial_i p$, $\partial_i \log p$, $\partial_i \sqrt{p}$, *etc*. Each of these are linked with a weighting function (expectation operator), so that the tangent vectors are zero-mean random variables:

$$0 = \mathrm{E}_\mu\{\partial_i p\} = \mathrm{E}_\mu\{(p)\, \partial_i \log p\} = \mathrm{E}_\mu\{(\sqrt{p})\, \partial_i (\sqrt{p})\} = \cdots \tag{54}$$

where the weighting functions are, respectively, one, $p$, $\sqrt{p}$:

$$0 = \mathrm{E}_\mu\{\partial_i p\} = \mathrm{E}_p\{\partial_i \log p\} = \mathrm{E}_{\sqrt{p}}\{\partial_i(\sqrt{p})\} = \cdots$$

For these various choices, the direction of the tangent vectors are all the same. We can consider the above as special cases of $\rho$-embedding, with $\rho(t) = t, \log t, \sqrt{t}$, respectively. Because $\partial_i(\rho(p)) = \rho'(p)\partial_i p$, so a tangent vector retains its direction with any choice of monotone embedding function.

To investigate the weighting function for general monotone $\rho$-embedding, let us consider the $f$-normalization (foliation) condition, *cf.* [21],

$$\mathrm{E}_\mu\{f(\rho(p)\} = 1, \tag{55}$$

where $f$ is a given convex function. Differentiate the above; we get:

$$0 = \mathrm{E}_\mu\left\{ f'(\rho(p)) \frac{\partial \rho'(p)}{\partial \theta^i} \right\} = \mathrm{E}_\mu\{\tau(p)\, \partial_i \rho\}. \tag{56}$$

Therefore, we can see that $\tau(p) = f'(\rho(p))$, what we have called the $f$-conjugate of $\rho$, is precisely the weighting function to make $\partial_i \rho$ a zero-mean random function at any point of $\mathcal{M}_\Theta$ (*i.e.*, for any value of $\theta \in \Theta$).

Next, consider the Fisher–Rao metric (1), which can be written as $\mathrm{E}_\mu\{\partial_i p\, \partial_j \log p\} = \mathrm{E}_\mu\{\partial_i \log p\, \partial_j p\}$, the pairing of a random function with a random functional under two embeddings $p$ and $\log p$. A natural generalization (see [6]) is to use two (independently chosen) monotone embeddings $\rho, \tau$:

$$g_{ij}(\theta) = \mathrm{E}_\mu\{\partial_i \rho\, \partial_j \tau\} = \mathrm{E}_\mu\{\partial_j \rho\, \partial_i \tau\} = \mathrm{E}_\mu\{\rho'(p)\, \tau'(p)\, \partial_i p\, \partial_j p\}\ . \tag{57}$$

This is precisely (14), with the weighting function for the Riemannian metric as $f''(\rho(p))(\rho'(p))^2 = \tau'(p)\rho'(p)$, when tangent vectors are expressed as $\partial_i p$ (identity representation). When $\rho$-representation or $\tau$-representation is adopted, the weighting function is simply $f''(\rho(p))$ or $(f^*)''(\tau(p))$, respectively.

Third, given $\rho, \tau$ embedding, we can construct two affine connections on the manifold as follows. Differentiate (57),

$$\frac{\partial g_{ij}(\theta)}{\partial \theta^k} = \mathrm{E}_\mu\left\{ \frac{\partial^2 \rho(p)}{\partial \theta^k \partial \theta^i} \frac{\partial \tau}{\partial \theta^j} + \frac{\partial^2 \tau(p)}{\partial \theta^k \partial \theta^j} \frac{\partial \rho(p)}{\partial \theta^i} \right\}\ , \tag{58}$$

and compare with the relation that defines conjugate connections:

$$\frac{\partial g_{ij}(\theta)}{\partial \theta^k} = \Gamma_{ki,j}(\theta) + \Gamma^*_{kj,i}(\theta)\ ; \tag{59}$$

we can identify:

$$\mathrm{E}_\mu\left\{ \frac{\partial^2 \rho(p)}{\partial \theta^k \partial \theta^i} \frac{\partial \tau(p)}{\partial \theta^j} \right\} \tag{60}$$

with $\Gamma_{ki,j}$ and:

$$\mathrm{E}_\mu\left\{ \frac{\partial^2 \tau(p)}{\partial \theta^k \partial \theta^i} \frac{\partial \rho(p)}{\partial \theta^j} \right\} \tag{61}$$

with $\Gamma^*_{kj,i}$, respectively. Their difference is, by definition, the Amari–Chentsov (0,3)-tensor $T$:

$$T_{ijk}(\theta) \equiv \mathrm{E}_\mu \left\{ \frac{\partial^2 \tau(p)}{\partial\theta^i \partial\theta^j} \frac{\partial \rho(p)}{\partial \theta^k} - \frac{\partial^2 \rho(p)}{\partial\theta^i \partial\theta^j} \frac{\partial \tau(p)}{\partial \theta^k} \right\} . \tag{62}$$

**Proposition 3.** *$T$ as given by (62) is a totally symmetric (0,3)-tensor.*

**Proof.** First, we prove that $T(\theta)$ is totally symmetric:

$$T_{ijk} = T_{jik} = T_{ikj} = T_{jki} = T_{kij} = T_{kji}. \tag{63}$$

Since (62) clearly implies $T_{ijk} = T_{jik}$, we only need to establish $T_{ijk} = T_{ikj}$. Applying the chain-rule of differentiation,

$$\frac{\partial}{\partial\theta^i}\left(\frac{\partial \tau(p)}{\partial\theta^j}\frac{\partial \rho(p)}{\partial\theta^k}\right) = \frac{\partial^2 \tau(p)}{\partial\theta^i\partial\theta^j}\frac{\partial \rho(p)}{\partial\theta^k} + \frac{\partial^2 \rho(p)}{\partial\theta^i\partial\theta^k}\frac{\partial \tau(p)}{\partial\theta^j} , \tag{64}$$

$$\frac{\partial}{\partial\theta^i}\left(\frac{\partial \rho(p)}{\partial\theta^j}\frac{\partial \tau(p)}{\partial\theta^k}\right) = \frac{\partial^2 \rho(p)}{\partial\theta^i\partial\theta^j}\frac{\partial \tau(p)}{\partial\theta^k} + \frac{\partial^2 \tau(p)}{\partial\theta^i\partial\theta^k}\frac{\partial \rho(p)}{\partial\theta^j} , \tag{65}$$

and taking into account:

$$\frac{\partial \tau(p)}{\partial\theta^j}\frac{\partial \rho(p)}{\partial\theta^k} = \frac{\partial \tau(p)}{\partial\theta^k}\frac{\partial \rho(p)}{\partial\theta^j} = \tau'(p)\rho'(p)\frac{\partial p}{\partial\theta^j}\frac{\partial p}{\partial\theta^k} , \tag{66}$$

(62) becomes:

$$T_{ijk}(\theta) = \mathrm{E}_\mu \left\{ -\left( \frac{\partial^2 \rho(p)}{\partial\theta^i\partial\theta^k}\frac{\partial \tau(p)}{\partial\theta^j} - \frac{\partial^2 \tau(p)}{\partial\theta^i\partial\theta^k}\frac{\partial \rho(p)}{\partial\theta^j} \right) \right\} = T_{ikj}(\theta) . \tag{67}$$

Next, we prove that $T_{ijk}$ is indeed a (0,3)-tensor. This is done through examining the behavior of $T$ under a coordinate transform $\theta \mapsto \bar{\theta}$, with the (inverse) Jacobian matrix $\frac{\partial\theta^k}{\partial\bar{\theta}^l}$, which affects:

$$\frac{\partial \rho(p)}{\partial\bar{\theta}^i} = \sum_l \frac{\partial \rho(p)}{\partial\theta^l}\frac{\partial\theta^l}{\partial\bar{\theta}^i} , \qquad \frac{\partial \tau(p)}{\partial\bar{\theta}^i} = \sum_l \frac{\partial \tau(p)}{\partial\theta^l}\frac{\partial\theta^l}{\partial\bar{\theta}^i} , \tag{68}$$

and:

$$\frac{\partial^2 \rho(p)}{\partial\bar{\theta}^i\partial\bar{\theta}^j} = \sum_{l,m} \frac{\partial^2 \rho(p)}{\partial\theta^l\partial\theta^m}\frac{\partial\theta^l}{\partial\bar{\theta}^i}\frac{\partial\theta^m}{\partial\bar{\theta}^j} + \sum_l \frac{\partial \rho(p)}{\partial\bar{\theta}^l}\frac{\partial^2\theta^l}{\partial\bar{\theta}^i\partial\bar{\theta}^j} , \tag{69}$$

$$\frac{\partial^2 \tau(p)}{\partial\bar{\theta}^i\partial\bar{\theta}^j} = \sum_{l,m} \frac{\partial^2 \tau(p)}{\partial\theta^l\partial\theta^m}\frac{\partial\theta^l}{\partial\bar{\theta}^i}\frac{\partial\theta^m}{\partial\bar{\theta}^j} + \sum_l \frac{\partial \tau(p)}{\partial\bar{\theta}^l}\frac{\partial^2\theta^l}{\partial\bar{\theta}^i\partial\bar{\theta}^j} . \tag{70}$$

Therefore:

$$\bar{T}_{ijk}(\bar{\theta}) \equiv \mathrm{E}_\mu \left\{ \frac{\partial^2 \tau(p)}{\partial\bar{\theta}^i\partial\bar{\theta}^j}\frac{\partial \rho(p)}{\partial\bar{\theta}^k} - \frac{\partial^2 \rho(p)}{\partial\bar{\theta}^i\partial\bar{\theta}^j}\frac{\partial \tau(p)}{\partial\bar{\theta}^k} \right\} = \sum_{lmn} \frac{\partial\theta^i}{\partial\bar{\theta}^l}\frac{\partial\theta^j}{\partial\bar{\theta}^m}\frac{\partial\theta^k}{\partial\bar{\theta}^n} T_{lmn}(\theta) . \tag{71}$$

after substituting (69), (70) and (62). $T$ indeed transforms to $\bar{T}$ in a manner that defines a $(0,3)$-tensor. Therefore, the proposition is proven. □

We now cast the Amari–Chentsov tensor $T$ in an alternative form that gives an explicit form of weighting function. Given $\rho, \tau$, because of Lemma 1, there exists another monotone embedding $\sigma$, such that $\sigma(\rho) = \tau$. Differentiating,

$$\frac{\partial \sigma(\rho(p))}{\partial \theta^i} = \sigma'(\rho(p)) \, \frac{\partial \rho(p)}{\partial \theta^i} \ . \tag{72}$$

Differentiate again, we obtain:

$$\frac{\partial^2 \sigma(\rho(p))}{\partial \theta^i \partial \theta^j} = \sigma''(\rho(p)) \frac{\partial \rho(p)}{\partial \theta^j} \frac{\partial \rho(p)}{\partial \theta^i} + \sigma'(\rho(p)) \frac{\partial^2 \rho(p)}{\partial \theta^i \partial \theta^j} \ . \tag{73}$$

Substituting the above into (62), we obtain an expression of $T$ in terms of $\rho$ (which plays the role of embedding function) and $\sigma$ (which plays the role of weighting function):

$$T_{ijk}(\theta) = \mathrm{E}_\mu \left\{ \sigma''(\rho(p)) \frac{\partial \rho(p)}{\partial \theta^i} \frac{\partial \rho(p)}{\partial \theta^j} \frac{\partial \rho(p)}{\partial \theta^k} \right\} \ . \tag{74}$$

Similarly, we can obtain:

$$T_{ijk}(\theta) = -\mathrm{E}_\mu \left\{ (\sigma^{-1})''(\tau(p)) \frac{\partial \tau(p)}{\partial \theta^i} \frac{\partial \tau(p)}{\partial \theta^j} \frac{\partial \tau(p)}{\partial \theta^k} \right\} \ . \tag{75}$$

Therefore, under $\tau$-representation, $\sigma^{-1}$ (the inverse function of $\sigma$) serves as the weighting function. Note that $\sigma = f', \sigma^{-1} = (f^*)'$ when $\rho$ and $\tau$ are said to be conjugate. Furthermore, note the negative sign in (75) compared with (74); this precisely reflects "representation duality" with a $\rho \longleftrightarrow \tau$ exchange.

To summarize, because $\alpha$-geometry $\{\mathcal{M}, g, T\}$ is uniquely specified given a Riemannian metric $g$ and the Amari–Chentsov tensor $T$, the above derivations show that they both enjoy the freedom of two monotone/convex functions, with the freedom in specifying $g$ coupled to the freedom in specifying $T$ in the same way that the metric and connections are coupled via Codazzi relation for statistical manifolds. That the weighting functions used to construct linear, symmetric bilinear and totally symmetric trilinear functionals (on random functions) turns out to be $f'(\rho(\cdot)), f''(\rho(\cdot)), f'''(\rho(\cdot))$, respectively, is noteworthy. See [6] for more discussions.

### *2.4. Representation Duality versus Reference Duality*

Going beyond extending $\alpha$-embedding to dual monotonic embeddings, Reference [2] illuminated two different senses of duality in the $\alpha$-geometry. Prior to [2], there have been several different usages of $\alpha$-parameter in Amari's theory of information geometry [10,11]:

(1) parameterizing the divergence functions ($\alpha$-divergences);

(2) parameterizing monotone embedding of probability functions ($\alpha$-embedding);

(3) parameterizing the convex mixture of connections ($\alpha$-connections).

Zhang (2004) [2] showed that (1) and (2) reflect two different types of duality in information geometry, with (1) concerning the reference/comparison status of a pair of points (functions) expressed in the divergence function ("reference duality") and (2) concerning their representation under arbitrary monotone scaling ("representation duality"). Both can lead to (3), the family of $\alpha$-connections. Therefore, care has to be taken in carefully delineating these two kinds of duality; for instance, the $\alpha\beta$-connection we derived in (21) reflects how reference duality and representation duality interacts in the alpha-connections.

The present analysis elaborated representation duality in information geometry by working out the freedom in allowing two (independently chosen) embedding functions $\rho, \tau$ or, equivalently, one embedding function $\rho$ along with a weighting function $f$, while the $(\rho, f)$ pair can be dually chosen to be the $(\tau, f^*)$ pair. Naudts' (2004) [3] $\phi$-logarithm is but a special case of the $(\rho, \tau)$ duality, in which $f'$ plays the role of the "integral-of-the-reciprocal" operation, that is taking the $\log$ of a function. This linkage then leads to $f^*$ and $\tau$ as inverse functions. The phenomena of biduality emerges when exchanging $\rho \longleftrightarrow \tau$ or $(\rho, f) \longleftrightarrow (\tau, f^*)$ leads to invariance of the Riemannian metric, but switches the two connections (the latter half of the statement is equivalent to changing signs of the Amari–Chentsov tensor). Therefore, the present paper, while elaborating the theory developed in [2], re-asserts the distinction between two distinct kinds of duality that was originally confounded in Amari's theory of $\alpha$-geometry, one through the freedom of selecting monotone embedding functions ("representation duality") and the other through the freedom of assigning referential status to points for pair comparison ("reference duality").

Finally, it is noted that the (bi)dualistic structure of the $(\rho, \tau)$-geometry (generalizing $\alpha$-geometry) is preserved in the non-parametric (infinite-dimensional) setting, as well [4,6], with the $\alpha$-connection structure cast in a more general way. Theorem 1 of [4] gives non-parametric expressions of the metric and connections under monotone embedding, mirroring the forms (14) and (15) in the parametric case.

## 3. Conclusion

The Riemannian metric with the pair of conjugate connections derived by Harsha and Moosath [1] are identical to the $(\rho, \tau)$-geometry obtained by Zhang in [2]. The $(\rho, \tau)$-embedding also recovers Naudts' deformed logarithm/exponential formulation. It is further shown in this paper that such $(\rho, \tau)$-geometry obtained is, when $\alpha$-embedding is relaxed to arbitrary monotone embeddings, the unique extension of Amari's $\alpha$-geometry in terms of its representational freedom.

## Acknowledgments

The writing of this paper was supported by research grant ARO W911NF-12-1-0163 awarded to Jun Zhang.

## Conflicts of Interest

The author declares no conflict of interest.

## References


1. Harsha, K.V; Subrahamanian Moosath, K.S. $F$-geometry and Amari's $\alpha$-geometry on a statistical manifold. *Entropy* **2014**, *16*, 2472–2487.
2. Zhang, J. Divergence function, duality, and convex analysis. *Neural Comput.* **2004**, *16*, 159–195.
3. Naudts, J. Estimators, escort probabilities, and $\phi$-exponential families in statistical physics. *J. Inequal. Pure Appl. Math.* **2004**, *5*, 102.
4. Zhang, J. Referential Duality and Representational Duality on Statistical Manifolds. In Proceedings of the Second International Symposium on Information Geometry and Its Applications, Tokyo, Japan, 12–16 December 2005; pp. 58–67.
5. Zhang, J. Referential duality and representational duality in the scaling of multi-dimensional and infinite-dimensional stimulus space. In *Measurement and Representation of Sensations: Recent Progress in Psychological Theory*; Dzhafarov, E., Colonius, H., Eds.; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 2006.
6. Zhang, J. Nonparametric information geometry: From divergence function to referential-representational biduality on statistical manifolds. *Entropy* **2013**, *15*, 5384–5418.
7. Zhang, J. Divergence functions and geometric structures they induce on a manifold. In *Geometric Theory of Information*; Nielsen, F., Ed.; Springer: Cham, Switzerland, 2014; pp. 1–30.
8. Zhang, J. Reference duality and representation duality in information geometry. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MaxEnt2014)*, Proceedings of 34th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Amboise, France, 21–26 September 2014; Volume 1641; pp. 130–146.
9. Amari, S. Differential geometry of curved exponential families—curvatures and information loss. *Ann. Stat.* **1982**, *10*, 357–385.
10. Amari, S. *Differential Geometric Methods in Statistics*; Lecture Notes in Statistics, Volume 28; Springer: New York, NY, USA, 1985.
11. Amari, S.; Nagaoka, H. *Method of Information Geometry*; Oxford University Press: Oxford, UK, 2000.
12. Ohara, A. Geometry of distributions associated with Tsallis statistics and properties of relative entropy minimization. *Phys. Lett. A* **2007**, *370*, 184–193.
13. Naudts, J. Generalised exponential families and associated entropy functions. *Entropy* **2008**, *10*, 131–149.
14. Ohara, A.; Matsuzoe, H.; Amari, S. A dually flat structure on the space of escort distributions. *J. Phys. Conf. Ser.* **2010**, *201*, 012012.

15. Amari. S.; Ohara, A. Geometry of q-exponential family of probability distributions. *Entropy* **2011**, *13*, 1170–1185.
16. Amari, S.; Ohara, A.; Matsuzoe, H. Geometry of deformed exponential families: Invariant, dually-flat and conformal geometry. *Physica A* **2012**, *391*, 4308–4319.
17. Eguchi, S. Second order efficiency of minimum contrast estimators in a curved exponential family. *Ann. Stat.* **1983**, *11*, 793–803.
18. Eguchi, S. A differential geometric approach to statistical inference on the basis of contrast functionals. *Hiroshima Math. J.* **1985**, *15*, 341–391.
19. Chentsov, N.N. *Statistical Decision Rules and Optimal Inference*; American Mathematics Society: Providence, RI, USA, 1982.
20. Ay, N.; Jost, J.; Le, H.V.; Schwachhöfer, L. Information geometry and sufficient statistics. *Probab. Theory Relat. Fields* **2014**, doi: 10.1007/s00440-014-0574-8.
21. Zhang, J.; Hasto, P. Statistical manifold as an affine space: A functional equation approach. *J. Math. Psychol.* **2006**, *50*, 60–65.
22. Burbea, J.; Rao, C.R. Entropy differential metric, distance and divergence measures in probability spaces: A unified approach. *J. Multivar. Anal.* **1982**, *12*, 575–596.
23. Burbea, J.; Rao, C.R. Differential metrics in probability spaces. *Probab. Math. Stat.* **1984**, *3*, 241–258.

# Binary Classification with a Pseudo Exponential Model and Its Application for Multi-Task Learning

**Takashi Takenouchi, Osamu Komori and Shinto Eguchi**

**Abstract:** In this paper, we investigate the basic properties of binary classification with a pseudo model based on the Itakura–Saito distance and reveal that the Itakura–Saito distance is a unique appropriate measure for estimation with the pseudo model in the framework of general Bregman divergence. Furthermore, we propose a novel multi-task learning algorithm based on the pseudo model in the framework of the ensemble learning method. We focus on a specific setting of the multi-task learning for binary classification problems. The set of features is assumed to be common among all tasks, which are our targets of performance improvement. We consider a situation where the shared structures among the dataset are represented by divergence between underlying distributions associated with multiple tasks. We discuss statistical properties of the proposed method and investigate the validity of the proposed method with numerical experiments.

Reprinted from *Entropy*. Cite as: Takenouchi, T.; Komori, O.; Eguchi, S. Binary Classification with a Pseudo Exponential Model and Its Application for Multi-Task Learning. *Entropy* **2015**, *17*, 4892–4910.

## 1. Introduction

In the framework of multi-task learning problems, we assume that there are multiple related tasks (datasets) sharing a common structure and can utilize the shared structure to improve the generalization performance of classifiers for multiple tasks [1,2]. This framework has been successfully employed in various kind of applications, such as medical diagnosis. Most methods utilize the similarity among tasks to improve the performance of classifiers by representing the shared structure as a regularization term [3,4]. We tackle this problem using a boosting method, which makes it possible to adaptively learn complicated problems with low computational cost. The boosting methods are notable implementations of the ensemble learning and try to construct a better classifier by combining weak classifiers. AdaBoost is the most popular boosting method, and many variations, including TrAdaBoost for the multi-task learning [5], have been developed. In face recognition [6], as well as web search ranking [7], the computational efficiency of boosting is paid attention to in the framework of multi-task learning.

In this paper, we firstly reveal that AdaBoost can be derived by a sequential minimization of the Itakura–Saito (IS) distance between an empirical distribution and a pseudo measure model associated with a classifier. The IS distance is a special case of the Bregman divergence [8] between two positive measures and is frequently used for non-negative matrix factorization (NMF) in the region of signal processing [9,10]. Secondly, we propose a novel boosting algorithm for the multi-task learning based on the IS distance. We utilize the IS distance as a discrepancy measure between pseudo models associated with tasks and incorporate the IS distance as a regularizer into AdaBoost. The proposed

method can capture the shared structure, *i.e.*, the relationship between underlying distributions by considering the IS distance between pseudo models constructed by classifiers. We discuss the statistical properties of the proposed method and investigate the validity of the regularization by the IS distance with small experiments using synthetic datasets and a real dataset.

This paper is organized as follows. In Section 2, basic settings are described, and a divergence measure is introduced. In Section 3, we briefly introduce the IS distance, which is a special case of the Bregman divergence, and investigate the relationship between a well-known ensemble algorithm, AdaBoost and estimation with a pseudo model using the Itakura–Saito distance. In Section 4, we propose a method for multi-task learning, which is derived from a minimization of the weighted sum of divergence, and the performance of the proposed methods is examined in Section 5 using a synthetic dataset and a real dataset (a short version of this article has been presented as a conference paper [11]; some theoretical results and numerical experiments are added to the current version).

## 2. Settings

In this study, we focus on binary classification problems. Let $\boldsymbol{x}$ be an input and $y \in \mathcal{Y} = \{\pm 1\}$ be a class label. Let us assume that $J$ datasets $\mathcal{D}_j = \{\boldsymbol{x}_i^{(j)}, y_i^{(j)}\}_{i=1}^{n_j}$ $(j = 1, \ldots, J)$ are given, and let $p_j(y|\boldsymbol{x})r_j(\boldsymbol{x})$ and $\tilde{p}_j(y|\boldsymbol{x})\tilde{r}_j(\boldsymbol{x})$ be an underlying distribution and an empirical distribution associated with the dataset $\mathcal{D}_j$, respectively. Here, we assume that each conditional distribution of $y$ given $\boldsymbol{x}$ is written as:

$$p_k(y|\boldsymbol{x}) = p_0(y|\boldsymbol{x}) + \delta_k(\boldsymbol{x})y \tag{1}$$

where $p_0(y|\boldsymbol{x})$ is a common conditional distribution for all datasets and $\delta_k(x)$ is a term that is specific to the dataset $\mathcal{D}_k$. Note that $\sum_{y\in\mathcal{Y}} \delta_k(\boldsymbol{x})y = 0$ holds, because $p_k(y|\boldsymbol{x})$ is a probability distribution. While a discriminant function $F_k$ is usually constructed using only the dataset $\mathcal{D}_k$, the multi-task learning aims to improve the performance of the discriminant function for each dataset $\mathcal{D}_k$ with the help of datasets $\mathcal{D}_j$ $(j \neq k)$. For this purpose, we consider a risk minimization problem defined with a pseudo model and the Itakura–Saito (IS) distance, which is a discrepancy measure frequently used in a region of signal processing.

Let $\mathcal{M} = \left\{ m(y) \;\middle|\; 0 \leq \sum_{y\in\mathcal{Y}} m(y) < \infty \right\}$ be a space of all positive finite measures over $\mathcal{Y}$. The Itakura–Saito distance between $p, q \in \mathcal{M}$ is defined as:

$$\mathrm{IS}(p, q; r) = \int r(\boldsymbol{x}) \sum_{y\in\mathcal{Y}} \left\{ \log \frac{q(y|\boldsymbol{x})}{p(y|\boldsymbol{x})} - 1 + \frac{p(y|\boldsymbol{x})}{q(y|\boldsymbol{x})} \right\} d\boldsymbol{x} \tag{2}$$

where $r(\boldsymbol{x})$ is a marginal distribution of $\boldsymbol{x}$ shared by $p$, $q \in \mathcal{M}$. Note that the IS distance is a kind of statistical version of the Bregman divergence [12], which makes it possible to directly plug-in the empirical distribution. We observe that $\mathrm{IS}(p, q; r) \geq 0$ and $\mathrm{IS}(p, q; r) = 0$ if and only if $p = q$. Banerjee *et al.* [13] showed that there exists a unique Bregman divergence corresponding to every regular exponential family, and the Itakura–Saito distance is associated with the exponential distribution.

## 3. Itakura–Saito Distance and Pseudo Model

### *3.1. Parameter Estimation with the Pseudo Model*

Let $q_F(y|\boldsymbol{x})$ be an (un-normalized) pseudo model associated with a function $F(\boldsymbol{x})$,

$$q_F(y|\boldsymbol{x}) = \exp(F(\boldsymbol{x})y). \tag{3}$$

Note that $q_F(y|\boldsymbol{x})$ is not a probability function, *i.e.*, $\sum_{y\in\mathcal{Y}} q_F(y|\boldsymbol{x}) \neq 1$ in general. If $q_F(y|\boldsymbol{x})$ is normalized, the model reduces to the classical logistic model as:

$$\bar{q}_F(y|\boldsymbol{x}) = \frac{\exp(F(\boldsymbol{x})y)}{\exp(F(\boldsymbol{x})) + \exp(-F(\boldsymbol{x}))}. \tag{4}$$

When the function $F$ is parameterized by $\boldsymbol{\theta}$, the maximum likelihood estimation (MLE) $\mathrm{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log \bar{q}_F(y_i|\boldsymbol{x}_i)$ or equivalently minimization of the (extended) Kullback–Leibler (KL) divergence is a powerful tool for the estimation of $\boldsymbol{\theta}$, and the MLE has properties such as asymptotic consistency and efficiency under some regularity conditions. Here, we consider parameter estimation with the pseudo model Equation (3) rather than the normalized model Equation (4).

**Proposition 1.** *Let $p(y|\boldsymbol{x}) = \bar{q}_{F_0}(y|\boldsymbol{x})$ be the underlying distribution. Then, we observe:*

$$\operatorname*{argmin}_F \mathrm{IS}(p, q_F; r) = F_0, \tag{5}$$

$$\operatorname*{argmin}_F \mathrm{IS}(q_F, p; r) = F_0. \tag{6}$$

**Proof.** See Appendix A □

On the other hand, when we consider an estimation based on the extended KL divergence, *i.e.*, $\mathrm{argmin}_F \mathrm{KL}(p, q_F; r)$ where:

$$\mathrm{KL}(p, q; r) = \int r(\boldsymbol{x}) \sum_{y\in\mathcal{Y}} \{p(y|\boldsymbol{x}) \log \frac{p(y|\boldsymbol{x})}{q(y|\boldsymbol{x})} - p(y|\boldsymbol{x}) + q(y|\boldsymbol{x})\} d\boldsymbol{x}, \tag{7}$$

we observe the following.

**Proposition 2.** *Let $F_0$ be a function $F_0(\neq 0)$ and $p(y|\boldsymbol{x}) = \bar{q}_{F_0}(y|\boldsymbol{x})$ be the underlying distribution. Then, we observe:*

$$F_{\mathrm{KL},1} = \operatorname*{argmin}_F \mathrm{KL}(p, q_F; r) \neq F_0, \tag{8}$$

$$F_{\mathrm{KL},2} = \operatorname*{argmin}_F \mathrm{KL}(q_F, p; r) \neq F_0. \tag{9}$$

**Proof.** See Appendix B. □

**Remark 1.** *Let $p(y|\boldsymbol{x}) = \bar{q}_{F_0}(y|\boldsymbol{x})$ be the underlying distribution. Then, minimizer Equation* (8) *or* (9) *of the extended KL divergence attains the Bayes rule, i.e.,*

$$\mathrm{sgn}(F_{\mathrm{KL},1}(\boldsymbol{x})) = \mathrm{sgn}(F_{\mathrm{KL},2}(\boldsymbol{x})) = \mathrm{sgn}\left(\frac{1}{2} \log \frac{p(+1|\boldsymbol{x})}{p(-1|\boldsymbol{x})}\right). \tag{10}$$

The proposition and the remark show that the extended KL divergence is not completely appropriate for estimation with the pseudo model.

### *3.2. Characterization of the Itakura–Saito Distance*

In this section, we investigate the characterization of the Itakura–Saito distance for estimation with the pseudo model, in the framework of the Bregman $U$-divergence. Firstly, we briefly introduce the statistical version of Bregman $U$-divergence [12]. The statistical version of Bregman $U$-divergence is a discrepancy measure between positive measures in $\mathcal{M}$ defined by a generating function $U$ and enables us to directly plug-in the empirical distribution for estimation. [12] proposed a general boosting-type algorithm for classification using the Bregman $U$-divergence and discussed properties of the method from the viewpoint of information geometry [14]. By changing the generating function $U$, the Bregman $U$-divergence can have a useful property as robustness against noise. For example, the β-divergence is a special case of the Bregman $U$-divergence and is frequently used for robust estimation in the context of unsupervised learning, such as clustering or component analysis [15,16]. Another example of the Bregman $U$-divergence is the η-divergence, which is employed to robustify the classification algorithm and is closely related to probability models of mislabeling [17,18].

Let $U$ be a monotonically-increasing convex function and $\xi$ be an inverse function of $U'$, the derivative of $U$. From the convexity of the function $U$, the function $\xi$ is a monotonically-increasing function. The statistical version of Bregman $U$-divergence between two measures $p, q \in \mathcal{M}$ is defined as follows.

$$D_U(p,q;r) = \int r(\boldsymbol{x}) \sum_{y\in\mathcal{Y}} \left\{ U(\xi(q(y|\boldsymbol{x}))) - U(\xi(p(y|\boldsymbol{x}))) - p(y|\boldsymbol{x}) \left(\xi(q(y|\boldsymbol{x})) - \xi(p(y|\boldsymbol{x}))\right) \right\} d\boldsymbol{x}. \tag{11}$$

Note that the function $\xi$ should be defined at least on $z > 0$.

**Remark 2.** *The KL divergence and the Itakura–Saito distance are special cases of the Bregman $U$-divergence Equation* (11) *with generating functions $U(z) = \exp(z)$ and $U(z) = -\log(c - z) + c_1$ $(z < c)$, where $c$ and $c_1$ are constants, respectively.*

Here, we introduce the concept of reflection-symmetric for characterization of the IS distance.

**Definition 3.** *A function $f(z)$ is reflection-symmetric if:*

$$f(z) = f\left(z^{-1}\right) \tag{12}$$

*holds for all $z \neq 0$.*

If the function $f$ is reflection-symmetric, we observe that:

$$\lim_{z\to 0} f(z) = \lim_{z\to\infty} f(z). \tag{13}$$

Because of this property, the reflection-symmetric function often has a singular point at $z = 0$, and to investigate the behavior of the function, we can employ the Laurent series as:

$$f(z) = c + \sum_{k=1}^{\infty} \left( a_k z^k + b_k z^{-k} \right). \tag{14}$$

Note that if the function $f$ is holomorphic over $R$, $b_k = 0$ for all $k$, and the Laurent series is equivalent to the Taylor series.

**Remark 3.** *If the function $f$ is reflection-symmetric and holomorphic over $R$, $a_k = b_k = 0$ holds for all $k$, and then, $f$ is a constant function.*

For the Bregman $U$-divergence Equation (11), we observe the following Lemma.

**Lemma 4.** *Let $F_0$ be an arbitrary function, $p(y|\boldsymbol{x}) = \bar{q}_{F_0}(y|\boldsymbol{x})$ be the underlying distribution and $q_F(\boldsymbol{x})$ be the pseudo model Equation* (3)*. If the Bregman $U$-divergence associated with the function $U$ attains:*

$$F_0 = \operatorname*{argmin}_F D_U(p, q_F; r), \tag{15}$$

*a function $\xi'(z)z^2$ derived from $U$ is reflection-symmetric. In addition, if the Bregman $U$-divergence associated with the function $U$ attains:*

$$F_0 = \operatorname*{argmin}_F D_U(q_F, p; r), \tag{16}$$

*a function $z\left\{\xi(z) - \xi\left(\frac{z}{z+z^{-1}}\right)\right\}$ derived from $U$ is reflection-symmetric.*

**Proof.** See Appendix C. □

**Remark 4.** *Proposition 1 implies that the function $\xi$ associated with the IS distance satisfies Lemma 4.*

**Remark 5.** *Propositions imply that the function $U$, i.e., Bregman $U$-divergence, attains Equation* (15) *or* (16) *is not unique and there exists divergences satisfying Equation* (15) *or* (16)*, other than the Itakura–Saito distance. For example, a function:*

$$\xi(z) = -2z^{-\frac{2}{3}} - z^{-\frac{4}{3}} \tag{17}$$

*satisfies $\xi'(z)z^2 = \frac{4}{3}(z^{1/3} + z^{-1/3})$, and then, $\xi'(z)z^2$ is reflection-symmetric. The associated generating function $U$ is written as:*

$$U(z) = \int^z \xi^{-1}(z')dz' = -4\frac{-2 + \sqrt{1-z}}{\sqrt{-1 + \sqrt{1-z}}} + C_1 \tag{18}$$

*where $C_1$ is a constant.*

In the following theorem, we reveal the characterization of the Itakura–Saito distance for estimation with the pseudo model Equation (3) and the Bregman $U$-divergence.

**Theorem 5.** *Let $p(y|\boldsymbol{x}) = \bar{q}_{F_0}(y|\boldsymbol{x})$ be the underlying distribution and $q_F(\boldsymbol{x})$ be the pseudo model Equation* (3)*. If conditions:*

$$F_0 = \operatorname*{argmin}_F D_U(p, q_F; r), \tag{19}$$

$$F_0 = \operatorname*{argmin}_F D_U(q_F, p; r) \tag{20}$$

*simultaneously hold, then $U(z) = -\log(-z)$, i.e., $D_U(p, q; r)$ is the Itakura–Saito distance* $\mathrm{IS}(p, q; r)$.

**Proof.** See Appendix D. □

**Remark 6.** *If we assume that a function $\xi'(z)z^2$ derived from $U$ is reflection-symmetric and holomorphic over $R$, $\xi'(z)z^2$ is a constant function from Remark 3. Then, we obtain $\xi(z) = c + \frac{b_1}{z}$ where $c, b_1$ are constants, implying that the associated divergence is equivalent to the Itakura–Saito distance.*

### *3.3. Relationship with AdaBoost*

The IS distance between the underlying conditional distribution $p(y|\boldsymbol{x})$ and the pseudo model $q_F(y|\boldsymbol{x})$ is written as:

$$\begin{aligned}\mathrm{IS}(p, q_F; r) &= C + \int r(\boldsymbol{x}) \sum_{y\in\mathcal{Y}} \left\{ F(\boldsymbol{x})y + \frac{p(y|\boldsymbol{x})}{q_F(y|\boldsymbol{x})} \right\} d\boldsymbol{x} \\ &= C + \int r(\boldsymbol{x}) \sum_{y\in\mathcal{Y}} p(y|\boldsymbol{x}) e^{-F(\boldsymbol{x})y} d\boldsymbol{x}, \end{aligned} \tag{21}$$

where $C$ is a constant, and Equation (21) is equivalent to an expected loss of AdaBoost, except for the constant term. Then, sequential minimization of an empirical version of Equation (21) is equivalent to the algorithm of AdaBoost, which is the most popular boosting method for the binary classification. Furthermore, [12,19] discussed that a gradient-based boosting algorithm can be derived from the minimization of the KL divergence or the Bregman $U$-divergence between the underlying distribution and a pseudo model. An important difference between these frameworks and our framework Equation (21) is the employed pseudo model. The pseudo model employed by the previous frameworks assumes a condition called "consistent data assumption" and is defined with the empirical distribution, implying that the pseudo model varies depending on the dataset. On the other hand, the pseudo model Equation (3) employed in Equation (21) is fixed against the dataset as usual statistical models.

The IS distance between two pseudo models $q_F(y|\boldsymbol{x})$ and $q_{F'}(y|\boldsymbol{x})$ is written as,

$$\begin{aligned}\mathrm{IS}(q_F, q_{F'}; r) &= \int r(\boldsymbol{x}) \sum_{y\in\mathcal{Y}} \{F'(\boldsymbol{x})y - F(\boldsymbol{x})y - 1 + \exp(F(\boldsymbol{x})y - F'(\boldsymbol{x})y)\}\, d\boldsymbol{x} \\ &= 2 + \int r(\boldsymbol{x}) \{\exp(F(\boldsymbol{x}) - F'(\boldsymbol{x})) + \exp(F'(\boldsymbol{x}) - F(\boldsymbol{x}))\}\, d\boldsymbol{x}. \end{aligned} \tag{22}$$

Note that $\mathrm{IS}(q_{F'}, q_F; r) = \mathrm{IS}(q_F, q_{F'}; r)$ holds for arbitrary $q_F$ and $q_{F'}$, while the IS distance itself is not necessarily symmetric. Furthermore, note that the symmetric property does not hold for normalized models $\bar{q}_F$ and $\bar{q}_{F'}$.

## 4. Application for Multi-Task Learning

There are two main types of frameworks for multi-task learning [20,21].

Case 1 : There is a target dataset $\mathcal{D}_k$, and our interest is to construct a discriminant function $F_k$ utilizing remaining datasets $\mathcal{D}_j$ $(j \neq k)$ or *a priori* constructed discriminant functions $F_j$ $(j \neq k)$.

Case 2 : Our interest is to simultaneously construct better discriminant functions $F_1, \ldots, F_J$ using all $J$ datasets $\mathcal{D}_1, \ldots, \mathcal{D}_J$ by utilizing shared information among datasets.

### *4.1. Case 1*

In this section, we focus on the above first framework. Let us assume that discriminant functions $F_j(\boldsymbol{x})$ $(j \neq k)$ are given or are constructed by an arbitrary binary classification method. Then, let us consider a risk function:

$$\begin{aligned} L_k(F_k) &= \mathrm{IS}(p_k, q_{F_k}; r_k) + \sum_{j\neq k} \lambda_{k,j}\, \mathrm{IS}(q_{F_k}, q_{F_j}; r_k) \\ &= \int r_k(\boldsymbol{x}) \left\{ \sum_{y\in\mathcal{Y}} p_k(y|\boldsymbol{x}) e^{-F_k(\boldsymbol{x})y} + \sum_{j\neq k} \lambda_{k,j} \left\{ e^{F_k(\boldsymbol{x}) - F_j(\boldsymbol{x})} + e^{F_j(\boldsymbol{x}) - F_k(\boldsymbol{x})} \right\} \right\} d\boldsymbol{x}, \end{aligned} \tag{23}$$

where $\lambda_{k,j} \geq 0$ $(j \neq k)$ are regularization constants. Note that the risk function depends on functions $F_j$ $(j \neq k)$, and the second term becomes small when the target discriminant function $F_k$ is similar to functions $F_j (j \neq k)$ in the sense of the IS distance; and the second term corresponds to a regularizer incorporating the shared information among datasets into the target function $F_k$. Furthermore, note that the marginal distribution $r_k$ is shared in the second term for the ease of implementation and the simplicity of theoretical analysis.

An empirical version of Equation (23) is written as:

$$\bar{L}_k(F_k) = \frac{1}{n_k} \sum_{i=1}^{n_k} \left( e^{-F_k(\boldsymbol{x}_i^{(k)}) y_i^{(k)}} + \sum_{j\neq k} \lambda_{k,j} \left( e^{F_k(\boldsymbol{x}_i^{(k)}) - F_j(\boldsymbol{x}_i^{(k)})} + e^{F_j(\boldsymbol{x}_i^{(k)}) - F_k(\boldsymbol{x}_i^{(k)})} \right) \right). \tag{24}$$

An algorithm is derived by sequential minimization of Equation (24) by updating $F_k$ to $F_k + \alpha f$, *i.e.*, $(\alpha, f) = \mathrm{argmin}_{\alpha,f}\, \bar{L}_k(F_k + \alpha f)$, where $f$ is a weak classifier and $\alpha$ is a coefficient [22].

(1) Initialize the function to $F_k^0$, and define weights for the $i$-th example with a function $F$ as:

$$w_1(i;F) = \frac{e^{-F(\boldsymbol{x}_i^{(k)})y_i^{(k)}}}{Z_1(F)},$$
$$w_2(i;F) = \frac{\sum_{j\neq k}\lambda_{k,j}e^{f(\boldsymbol{x}_i^{(k)})(F(\boldsymbol{x}_i^{(k)})-F_j(\boldsymbol{x}_i^{(k)}))}}{Z_2(F)}$$

where:

$$Z_1(F) = \sum_{i=1}^{n_k} e^{-F(\boldsymbol{x}_i^{(k)})y_i^{(k)}},$$
$$Z_2(F) = \sum_{i=1}^{n_k}\sum_{j\neq k}\lambda_{k,j}\left(e^{F(\boldsymbol{x}_i^{(k)})-F_j(\boldsymbol{x}_i^{(k)})} + e^{-F(\boldsymbol{x}_i^{(k)})+F_j(\boldsymbol{x}_i^{(k)})}\right).$$

(2) For $t = 1, \ldots, T$

(a) Select a weak classifier $f_k^t \in \{\pm 1\}$, which minimizes the following quantity:

$$\varepsilon(f) = \frac{Z_1(F_k^{t-1})}{Z_1(F_k^{t-1}) + Z_2(F_k^{t-1})}\varepsilon_1(f) + \frac{Z_2(F_k^{t-1})}{Z_1(F_k^{t-1}) + Z_2(F_k^{t-1})}\varepsilon_2(f). \tag{25}$$

where $\varepsilon_1(f) = \sum_{i=1}^{n_k} w_1(i;F_k^{t-1})\,\mathrm{I}(f(\boldsymbol{x}_i^{(k)}) \neq y_i^{(k)})$ and $\varepsilon_2(f) = \sum_{i=1}^{n} w_2(i;F_k^{t-1})$.

(b) Calculate a coefficient of $f_k^t$ by $\alpha_k^t = \frac{1}{2}\log\frac{1-\varepsilon(f_k^t)}{\varepsilon(f_k^t)}$.

(c) Update the discriminant function as $F_k^t = F_k^{t-1} + \alpha_k^t f_k^t$.

(3) Output $F_k^T(\boldsymbol{x}) = F_k^0(\boldsymbol{x}) + \sum_{t=1}^{T}\alpha_k^t f_k^t(\boldsymbol{x})$.

In Step 1, $F_k^0$ is typically initialized as $F_k^0(\boldsymbol{x}) = 0$. The quantity Equation (25) is a mixture of two terms: $\varepsilon_1(f)$ is a weighted error rate of the classifier $f$, and $\varepsilon_2(f)$ is the sum of weights $w_2(f)$, which represents the degree of discrepancy between $f$ and $F - F_j$. $\varepsilon_2(f)$ becomes large when $F$ is updated by $f$ as departed from $F_j$. Note that if we set $\lambda_{k,j} = 0$ for all $j$, the risk function Equation (24) coincides with that of AdaBoost, and the above derived algorithm reduces to the usual AdaBoost.

Because the empirical risk function Equation (24) is convex with respect to $F$ or $F'$, we can consider another version of the risk function as:

$$\bar{L}_k(F_k) = \frac{1}{n_k}\sum_{i=1}^{n_k}\left(e^{-F_k(\boldsymbol{x}_i^{(k)})y_i^{(k)}} + \lambda_k\left(e^{F_k(\boldsymbol{x}_i^{(k)})-\bar{F}_k(\boldsymbol{x}_i^{(k)})} + e^{-F_k(\boldsymbol{x}_i^{(k)})+\bar{F}_k(\boldsymbol{x}_i^{(k)})}\right)\right) \tag{26}$$

where $\bar{F}_k(\boldsymbol{x}) = \sum_{j\neq k}\frac{\lambda_{k,j}}{\lambda_k}F_j(\boldsymbol{x})$. The risk function is upper bounded by the risk function Equation (24), implying that the effect of regularization by the shared information is weakened. The derived algorithm is almost the same as the one derived from Equation (24).

*4.2. Case 2*

In this section, we consider simultaneous construction of discriminant functions $F_1, \ldots, F_J$ by minimizing the following risk function:

$$L(F_1, \ldots, F_J) = \sum_{j=1}^{J} \pi_j L_j(F_j) \tag{27}$$

where $\pi_j (j = 1, \ldots, J)$ is a positive constant satisfying $\sum_{j=1}^{J} \pi_j = 1$ and $L_k$ is defined in Equation (23).

Though we can directly minimize the empirical version of Equation (27), a derived algorithm is complicated and is computationally heavy. Then, we derive a simplified algorithm utilizing the algorithm shown in Case 1 in which a target dataset is fixed.

(1) Initialize functions $F_1, \ldots, F_J$.
(2) For $t = 1, \ldots, T$:
    (a) Randomly choose a target index $k \in \{1, \ldots, J\}$.
    (b) Update the function $F_k$ using the algorithm in Case 1 by $S$ steps, with fixed functions $F_j$ $(j \neq k)$.
(3) Output learned functions $F_1, \ldots, F_J$.

Note that the empirical risk function cannot be monotonically decreased because the minimization of $L_k(F_k)$ is a trade-off of the first term and the second regularization term, and a decrease of $L_k(F_k)$ does not necessarily mean a decrease of the regularization term.

*4.3. Statistical Properties of the Proposed Methods*

In this section, we discuss the statistical properties of the proposed methods. Firstly, we focus on Case 1, and the minimizer $F_k^*$ of the risk function Equation (23) satisfies the following:

$$\left.\frac{\delta L_k(F_k)}{\delta F_k(\boldsymbol{x})}\right|_{F_k = F_k^*} \propto -p_k(+1|\boldsymbol{x}) e^{-F_k^*(\boldsymbol{x})} + p_k(-1|\boldsymbol{x}) e^{F_k^*(\boldsymbol{x})} + \sum_{j \neq k} \lambda_{k,j} \left\{ e^{F_k^*(\boldsymbol{x}) - F_j(\boldsymbol{x})} - e^{F_j(\boldsymbol{x}) - F_k^*(\boldsymbol{x})} \right\} = 0, \tag{28}$$

which implies:

$$F_k^*(\boldsymbol{x}) = \frac{1}{2} \log \frac{p_k(+1|\boldsymbol{x}) + \sum_{j \neq k} \lambda_{k,j} \exp(F_j(\boldsymbol{x}))}{p_k(-1|\boldsymbol{x}) + \sum_{j \neq k} \lambda_{k,j} \exp(-F_j(\boldsymbol{x}))}, \tag{29}$$

or equivalently:

$$p_k(y|\boldsymbol{x}) = p_{0,k}(y|\boldsymbol{x}) \left( 1 + \sum_{j \neq k} \lambda_{k,j} \exp(-F_j(\boldsymbol{x}) y) \right) - p_{0,k}(-y|\boldsymbol{x}) \sum_{j \neq k} \lambda_{k,j} \exp(F_j(\boldsymbol{x}) y), \tag{30}$$

where $p_{0,k}(y|\boldsymbol{x}) = \frac{\exp(F_k^*(\boldsymbol{x})y)}{\exp(F_k^*(\boldsymbol{x}))+\exp(-F_k^*(\boldsymbol{x}))}$. This can be interpreted as a probabilistic model of asymmetric mislabeling [17,18]. In Equation (29), the confidence of classification is discounted by the results of remaining discriminant functions when the classifier $\mathrm{sgn}(F_k^*(\boldsymbol{x}))$ makes a different decision from these of $\mathrm{sgn}(F_j(\boldsymbol{x}))$ $(j \neq k)$.

**Remark 7.** *$F_k^*(\boldsymbol{x}) \geq 0$ does not mean $p_k(+1|\boldsymbol{x}) \geq \frac{1}{2}$ unless $F_j(\boldsymbol{x}) = \frac{1}{2}\log\frac{p_k(+1|\boldsymbol{x})}{p_k(-1|\boldsymbol{x})}$ holds.*

**Proposition 6.** *Let us assume that $F_j(\boldsymbol{x})$ satisfies:*

$$\frac{\exp(F_j(\boldsymbol{x})y)}{\exp(F_j(\boldsymbol{x})) + \exp(-F_j(\boldsymbol{x}))} = p_0(y|\boldsymbol{x}) + \epsilon_j(\boldsymbol{x})y,\ ||\epsilon_j(\boldsymbol{x})|| \ll 1. \tag{31}$$

*Then, Equation* (29) *can be approximated as:*

$$F_k^*(\boldsymbol{x}) \simeq \frac{1}{2}\log\frac{p_0(+1|\boldsymbol{x})}{p_0(-1|\boldsymbol{x})} + \frac{1}{2P^2}\frac{P\delta_k(\boldsymbol{x}) + \sum_{j\neq k}\lambda_{k,j}\epsilon_j(\boldsymbol{x})}{P + \lambda_k} \tag{32}$$

*where $P = \sqrt{p_0(+1|\boldsymbol{x})p_0(-1|\boldsymbol{x})}$ and $\lambda_k = \sum_{j\neq k}\lambda_{k,j}$.*

**Proof.** We obtain Equation (32) by considering the Taylor expansion of Equation (29). □

We observe that a discrepancy derived by $\delta_k$ is moderated by the mixture of $\epsilon_j$ when perturbations $\epsilon_j$ are independently and identically distributed.

**Proposition 7.** *Let $\eta_j(\boldsymbol{x}) = F_j(\boldsymbol{x}) - F_k(\boldsymbol{x})$ be a difference between two functions. Then, $F_k^*$ can be approximated as:*

$$F_k^*(\boldsymbol{x}) \simeq \frac{1}{2}\log\frac{p_k(+1|\boldsymbol{x})}{p_k(-1|\boldsymbol{x})} + \frac{1}{P}\sum_{j\neq k}\lambda_{k,j}\eta_j(\boldsymbol{x}). \tag{33}$$

**Proof.** See Appendix E. □

**Proposition 8.** *Let $\bar{F}_k^*$ be a minimizer of the risk function Equation* (23) *with $\lambda_{k,j} = 0 (j \neq k)$. Then, we observe:*

$$\left(\bar{F}_k^*(\boldsymbol{x}) - \frac{1}{2}\log\frac{p_0(+1|\boldsymbol{x})}{p_0(-1|\boldsymbol{x})}\right)^2 \geq \left(F_k^*(\boldsymbol{x}) - \frac{1}{2}\log\frac{p_0(+1|\boldsymbol{x})}{p_0(-1|\boldsymbol{x})}\right)^2, \tag{34}$$

*i.e., the proposed method improves the performance in the sense of the squared error, when:*

$$|\delta_k(\boldsymbol{x})| \geq \frac{|\sum_{j\neq k}\lambda_{k,j}\epsilon_j(\boldsymbol{x})|}{\lambda_k} \tag{35}$$

*holds.*

**Proof.** See Appendix F. □

Secondly, we consider a property of the algorithm for Case 2.

**Proposition 9.** *Let $r(\boldsymbol{x}) = r_j(\boldsymbol{x})$ $(j = 1, \ldots, J)$ be a common marginal distribution shared by all tasks. Then, the minimizer of the risk function is written as:*

$$F_k(\boldsymbol{x}) = \frac{1}{2} \log \frac{p_k(+1|\boldsymbol{x}) + \sum_{j \neq k} \lambda_{jk} e^{F_j(\boldsymbol{x})}}{p_k(-1|\boldsymbol{x}) + \sum_{j \neq k} \lambda_{jk} e^{-F_j(\boldsymbol{x})}}, \tag{36}$$

*where* $\lambda_{jk} = \lambda_{k,j} + \frac{\pi_j}{\pi_k} \lambda_{k,j}$.

**Proof.** See Appendix G. □

The only difference from Equation (28) is that regularization is strengthened by $\frac{\pi_j}{\pi_k} \lambda_{j,k}$, and then, the same propositions in Section 4.1 hold for Equation (36).

### *4.4. Comparison of Regularization Terms*

The proposed method incorporates the regularization term defined by the IS distance into AdaBoost. In this section, we discuss a property of the regularization term.

**Proposition 10.** *Let $\epsilon(\boldsymbol{x})$ be a perturbation function satisfying $|\epsilon(\boldsymbol{x})| \ll 1$. Then, we observe:*

$$\mathrm{KL}(\bar{q}_F, \bar{q}_{F+\epsilon}; r) \simeq \int 2r(\boldsymbol{x}) \epsilon(\boldsymbol{x})^2 \bar{q}_F(+1|\boldsymbol{x}) \bar{q}_F(-1|\boldsymbol{x}) d\boldsymbol{x}, \tag{37}$$

$$\mathrm{KL}(q_F, q_{F+\epsilon}; r) \simeq \int \frac{r(\boldsymbol{x})}{2} \epsilon(\boldsymbol{x})^2 \frac{1}{\sqrt{\bar{q}_F(+1|\boldsymbol{x}) \bar{q}_F(-1|\boldsymbol{x})}} d\boldsymbol{x}, \tag{38}$$

$$\mathrm{IS}(\bar{q}_F, \bar{q}_{F+\epsilon}; r) \simeq \int 2r(\boldsymbol{x}) \epsilon(\boldsymbol{x})^2 \sum_{y \in \mathcal{Y}} \bar{q}_F(y|\boldsymbol{x})^2 d\boldsymbol{x}, \tag{39}$$

$$\mathrm{IS}(q_F, q_{F+\epsilon}; r) \simeq \int r(\boldsymbol{x}) \epsilon(\boldsymbol{x})^2 d\boldsymbol{x}. \tag{40}$$

**Proof.** We obtain these approximations by considering the Taylor expansion up to second order. □

Figure 1 shows values of divergences against a value of $\bar{q}_F(\boldsymbol{x})$. Those relations implies that the KL divergence Equation (37) emphasizes a region of input $\boldsymbol{x}$ whose conditional distribution $\bar{q}_F(\boldsymbol{x})$ is nearly equal to $\frac{1}{2}$, *i.e.*, the classification boundary, while the IS distance Equation (39) focuses on a region of $\boldsymbol{x}$ whose conditional distribution is nearly equal to zero or one. The IS distance between pseudo model Equation (40), *i.e.*, the proposed method, considers the intermediate of Equations (37) and (39). This implies that the regularization Equation (40) with the IS distance puts more focus on a region far from the classification boundary compared to Equation (37), while Equation (39) tends to relatively ignore the region near the classification boundary. Furthermore, note that the employment of Equation (40) makes it possible to derive the simple algorithm shown in Section 4.1.

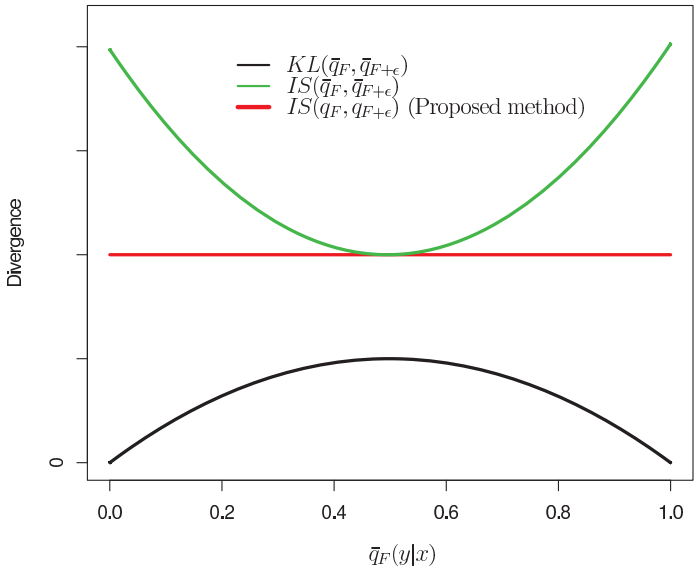



**Figure 1.** Values of divergences (regularization terms) against $\bar{q}_F$.

## 5. Experiments

In this section, we investigate the performance of the proposed multi-task algorithm with synthetic datasets and a real dataset.

### *5.1. Synthetic Dataset*

Firstly, we investigate the performance of the proposed method using two synthetic datasets within the situation described in Case 2. We compared the proposed method with AdaBoost trained with an individual dataset and AdaBoost trained with all datasets simultaneously. We employed the boosting stump (the boosting stump is a decision tree with only one node) as the weak classifier and fixed as $\pi_j = 1/J$. A boosting-type method has a hyper-parameter $T$, the step number of boosting, and the proposed method additionally has the hyper-parameter $\lambda_{k,j}$. In the experiment, we determined these parameters $T$ and $\lambda_{k,j}$ by the validation technique. Especially, we investigated two kinds of scenarios for the determination of $\lambda_{k,j}$.

1. We set that $\lambda_{k,j} = \lambda$ for all $j, k$ and determined $\lambda$.
2. We set that $\lambda_{k,j} = \frac{\lambda}{\mathrm{IS}\left(q_{\hat{F}_k}, q_{\hat{F}_j}; r_k\right)}$ where $\hat{F}_j$ is a discriminant function constructed by AdaBoost with the dataset $\mathcal{D}_j$ and determined $\lambda$.

Scenario 2 can incorporate more detailed information about the relationship between tasks, and the proposed method can ignore the information of tasks having less shared information. In summary, we compared the following four methods:

A : The proposed method with $\lambda_{k,j}$ determined by Scenario 1.

B : The proposed method with $\lambda_{k,j}$ determined by Scenario 2.
C : AdaBoost trained with an individual dataset.
D : AdaBoost trained with all datasets simultaneously.

We utilized $80\%$ of the training dataset for training of classifiers and the remaining $20\%$ for the validation. We repeated the above procedure $20$ times and observed the averaged performance of the methods.

5.1.1. Dataset 1

We set the number $J$ of tasks to three and assume that a marginal distribution of $\boldsymbol{x}$ is a uniform distribution on $[-1, 1]^2$, and a discriminant function $F_j$ $(j = 1, 2, 3)$ associated with each dataset is generated by $F_j(\boldsymbol{x}) = (1 + c_{j,2})(x_1 - c_{j,1}) - x_2$, where $c_{j,1} \sim \mathcal{N}(0, 0.2^2)$ and $c_{j,2} \sim \mathcal{N}(0, 0.1^2)$. In addition, we randomly added a contamination noise on label $y$. Under these settings, we generated a training dataset, including $400$ examples, and a test dataset, including $600$ examples. Generated datasets are shown in Figure 2. We observe that each discriminant function and noise structure are different from the other two.

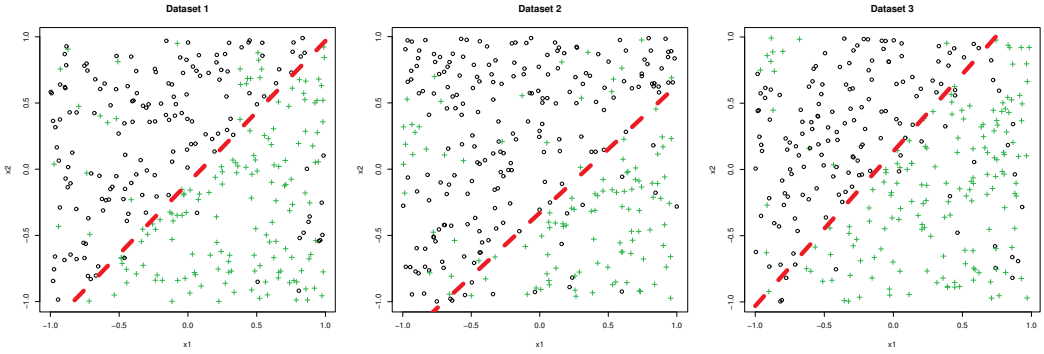



**Figure 2.** The three generated datasets and decision boundaries.

Figure 3 shows boxplots of the test errors of each method for datasets $\mathcal{D}_j$ $(j = 1, 2, 3)$. We observe that the proposed method consistently outperforms individually trained AdaBoost, and AdaBoost trained with all datasets simultaneously. The figure shows that the proposed method can incorporate shared information among datasets into classifiers.

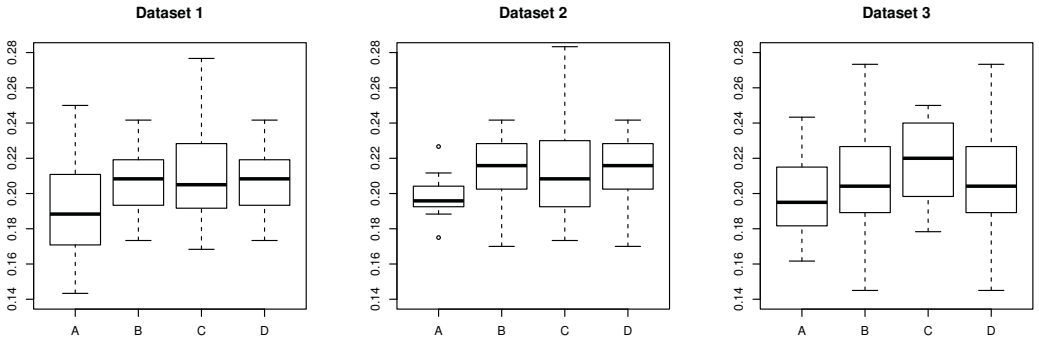



**Figure 3.** Boxplots of the test error of each method: A—proposed method with λ in Scenario 1; B—proposed method with λ in Scenario 2; C—AdaBoost trained with the individual dataset; D—AdaBoost trained with all datasets simultaneously; for three datasets, over the 20 simulation trials.

5.1.2. Dataset 2

We set the number $J$ of tasks to 6 and assume that a marginal distribution of $\boldsymbol{x}$ is a uniform distribution on $[-1, 1]^2$. Discriminant functions associated with each dataset are generated by:

$$F_j(\boldsymbol{x}) = \begin{cases} (1 + c_{j,2})(x_1 - c_{j,1}) - x_2, & j = 1, 2, 3, \\ -(1 + c_{j,2})(x_1 - c_{j,1}) + x_2, & j = 4, 5, 6, \end{cases}$$

where $c_{j,1} \sim \mathcal{N}(0, 0.1^2)$ and $c_{j,2} \sim \mathcal{N}(0, 0.1^2)$. In addition, we randomly added a contamination noise on label $y$. Under these settings, we generated training dataset, including 400 examples, and the test dataset, including 600 examples. Generated datasets are shown in Figure 4. We observe that Datasets 1, 2 and 3 share a structure, and Datasets 4, 5 and 6 share another structure.

Figure 5 shows boxplots of the test errors of each method for datasets $\mathcal{D}_j$ ($j = 1, \ldots, 6$). We omitted the result of AdaBoost trained with all datasets simultaneously (D) from the figure, because its performance is significantly worse than those of the other methods: the median of classification errors is around 0.5. This is because the structures of Datasets 1, 2, 3 and Datasets 4, 5, 6 are opposite, and the labeling of concatenated dataset seems to be random. We observe that the proposed method with Scenario 2 (B) improves performance against individually-trained AdaBoost (C) and the proposed method in Scenario 1 (A). This is because the structure shared among Datasets 1, 2 and 3 does not have information about Datasets 4, 5 and 6 (and *vice versa*), and Method (B) can ignore the influence of the irrelevant information by adjusting $\lambda_{k,j}$ responding to $\mathrm{IS}(q_{\hat{F}_j}, q_{\hat{F}_k}; r_k)$. Note that the performance of Method (A) is not so degraded, because the regularization parameter $\lambda_{k,j}$ was determined, so as to be zero, implying AdaBoost trained with the individual dataset.

Figure 6 shows examples of classification boundaries estimated by Methods A, B, C and D, for Dataset 6.

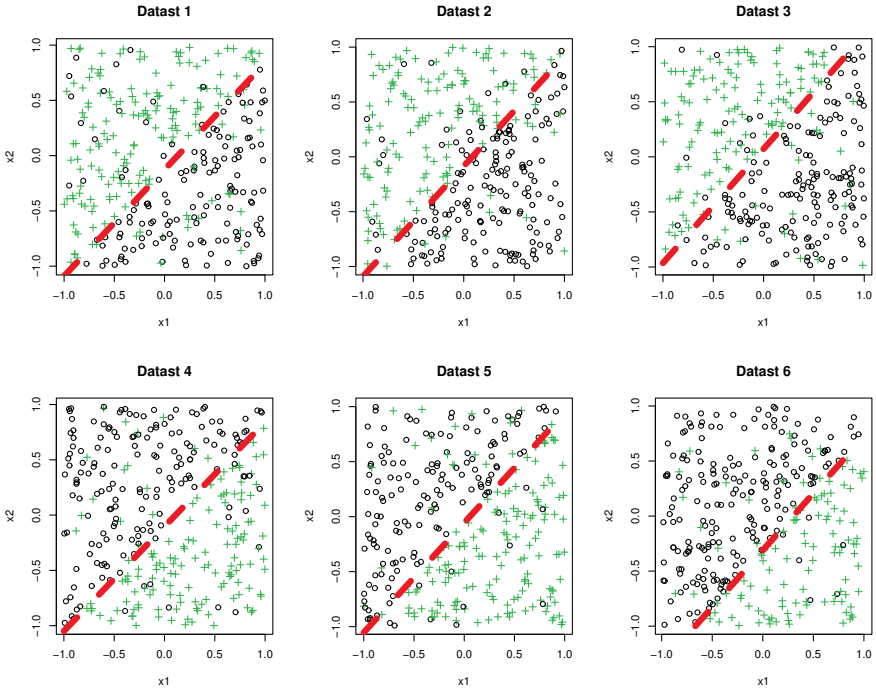



**Figure 4.** The six generated datasets and decision boundaries.

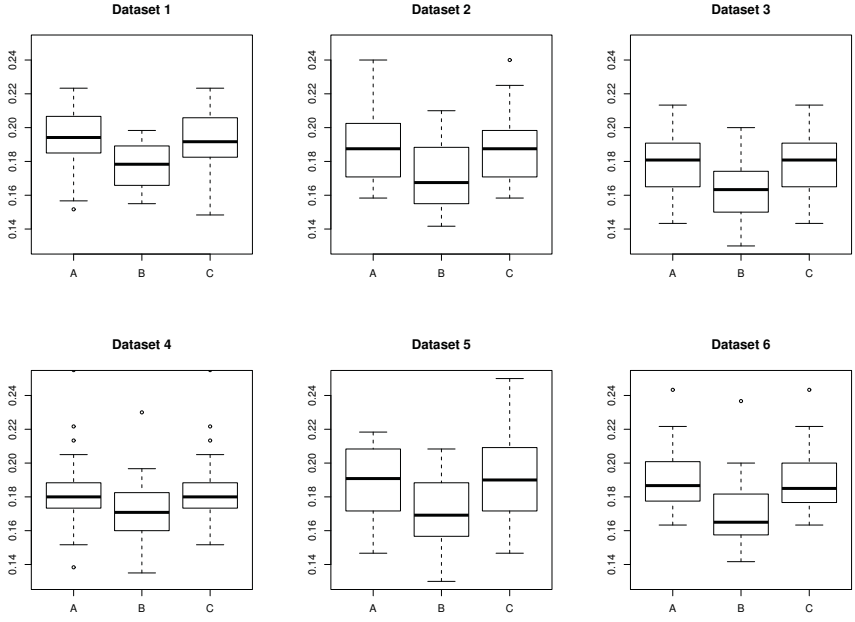



**Figure 5.** Boxplots of the test error of each method: A, Proposed method with $\lambda$ in Scenario 1; B, proposed method with $\lambda$ in Scenario 2; C, AdaBoost trained with the individual dataset ; for 6 datasets, over the 20 simulation trials.

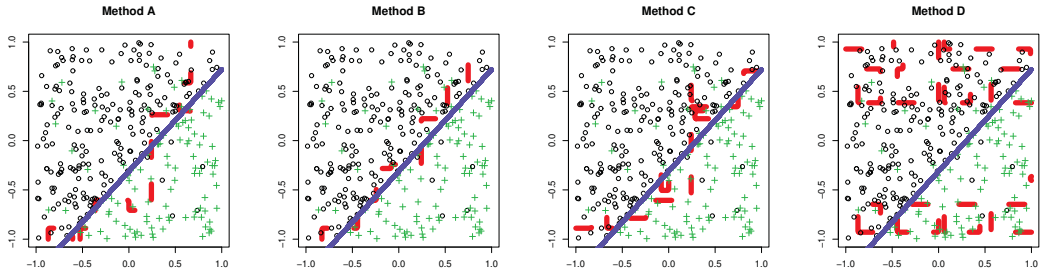



**Figure 6.** Classification boundaries by Methods A, B, C and D for Dataset 6. The blue line is the true classification boundary, and the red line represents the estimated classification boundary.

### *5.2. Real Dataset: School Dataset*

In this section, we compared the proposed method (Scenario 2) to the a binary decision tree-based ensemble method, called extremely randomized trees (ExtraTrees) [23], applying to a real dataset, "school data", reported from the Inner London Education Authority [24]. The dataset consists of examination records of 15,362 students from 139 secondary schools, *i.e.*, we had 139 tasks. The dimension of input $x$ is 27, in which original variables that are categorical were transformed into dummy variables. The original target variable $y_0$ represents score values in the range $[1, 70]$, and we transformed the target variable $y_0$ to a binary variable as:

$$y = \mathrm{sgn}(y_0 - 20).$$

We set the threshold to 20 to balance the ratio of classes ($-1 : +1 = 7930 : 7432$). We randomly divided the dataset of each tasks into $80\%$ of the training dataset and remaining $20\%$ test dataset. In addition, we used $20\%$ of the divided training dataset as a validation dataset to determine the hyper-parameter $\lambda$ and step number $T$. We repeated the above procedure 20 times and observed the average performance of the methods. Figure 7 shows the medians of error rates over 20 trials, by the proposed method and the ExtraTrees for 139 tasks. The horizontal axis indicates an index of a task, which is ranked in increasing order of the median error rate of the ExtraTrees. We observe that the proposed method is comparable to the ExtraTrees and especially has an advantage for tasks, in which the error rates of the ExtraTrees are large.

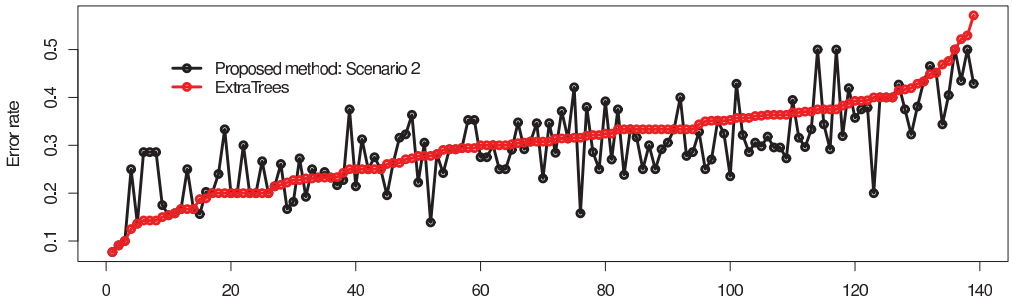



**Figure 7.** Medians of error rates by the proposed method and extremely randomized trees (ExtraTrees) for 139 tasks. The horizontal axis represents an index of a task, and the vertical axis indicates the median of error rates over 20 trials. Tasks are ranked in increasing order of the median error rate of the ExtraTrees.

## 6. Conclusions

In this paper, we investigate the properties of binary classification with the pseudo model and reveal that minimization of the Itakura–Saito distance between the empirical distribution and the pseudo model is equivalent to AdaBoost and provides suitable properties for the binary classification. In addition, we pointed out that the Itakura–Saito distance is a unique divergence, having a suitable property for estimation with the pseudo model in the framework of the Bregman divergence. Based on the framework, we proposed a novel binary classification method for the multi-task learning, which incorporates shared information among tasks into the targeted task. The risk function of the proposed method is defined by the mixture of IS distance. The IS distance between pseudo models can be interpreted as the regularization term, incorporating shared information among tasks into the binary classifier for the target task. We investigated statistical properties of the risk function and derived computationally-feasible boosting-based algorithms. Furthermore, we considered a mechanism for the adjustment of the degree of information sharing and numerically investigated the validity of the proposed methods.

## Acknowledgments

This study was partially supported by a Grant-in-Aid for Young Scientists (B), 25730018, from MEXT, Japan. Shinto Eguchi and Osamu Komori were supported by the Japan Science and Technology Agency (JST), Core Research for Evolutionary Science and Technology (CREST).

## Author Contributions

Takashi Takenouchi made major contributions to employing the Itakura–Saito divergence, and Shinto Eguchi gave a proof for the characterization associated with the divergence. Takashi Takenouchi and Osamu Komori contributed to the statistical discussion for the multi-task learning.

## Conflicts of Interest

The authors declare no conflict of interest.

## Appendix

### A. Proof of Proposition 1

By a variational calculation, a minimizer of Equation (5) satisfies:

$$\frac{\delta \,\mathrm{IS}(p, q_F; r)}{\delta F(\boldsymbol{x})} \propto \frac{e^{F_0(\boldsymbol{x})-F(\boldsymbol{x})} - e^{-F_0(\boldsymbol{x})+F(\boldsymbol{x})}}{e^{F_0(\boldsymbol{x})} + e^{-F_0(\boldsymbol{x})}} = 0, \tag{41}$$

and $F = F_0$ satisfies the above equation for an arbitrary $F_0$, which concludes Equation (5). Furthermore,

$$\frac{\delta \,\mathrm{IS}(q_F, p; r)}{\delta F(\boldsymbol{x})} \propto \left(e^{F_0(\boldsymbol{x})} + e^{-F_0(\boldsymbol{x})}\right)\left(e^{F(\boldsymbol{x})-F_0(\boldsymbol{x})} - e^{-F(\boldsymbol{x})+F_0(\boldsymbol{x})}\right) = 0, \tag{42}$$

and $F = F_0$ satisfies the above equation for an arbitrary $F_0$, concluding Equation (6).

### B. Proof of Proposition 2

By a straightforward variational calculation, we observe that a minimizer $F_{\mathrm{KL},1}$ of Equation (8) satisfies:

$$\begin{aligned}\frac{\delta \,\mathrm{KL}(p, q_F; r)}{\delta F(\boldsymbol{x})} &\propto -p(+1|\boldsymbol{x}) + p(-1|\boldsymbol{x}) + \exp(F_{\mathrm{KL},1}(\boldsymbol{x})) - \exp(-F_{\mathrm{KL},1}(\boldsymbol{x})) \\ &= \frac{-e^{F_0(\boldsymbol{x})} + e^{-F_0(\boldsymbol{x})}}{e^{F_0(\boldsymbol{x})} + e^{-F_0(\boldsymbol{x})}} + e^{F_{\mathrm{KL},1}(\boldsymbol{x})} - e^{-F_{\mathrm{KL},1}(\boldsymbol{x})} = 0,\end{aligned} \tag{43}$$

and $F_{\mathrm{KL},1} = F_0$ means $F_0(\boldsymbol{x}) = 0$ ($\forall \boldsymbol{x}$), which concludes Equation (8). Furthermore, for Equation (9), $F_{\mathrm{KL},2}$ satisfies:

$$\begin{aligned}&\frac{\delta \,\mathrm{KL}(q_F, p; r)}{\delta F(\boldsymbol{x})} \\ &\propto (F_{\mathrm{KL},2}(\boldsymbol{x}) - F_0(\boldsymbol{x}))(e^{F_{\mathrm{KL},2}(\boldsymbol{x})} + e^{-F_{\mathrm{KL},2}(\boldsymbol{x})}) + (e^{F_{\mathrm{KL},2}(\boldsymbol{x})} - e^{-F_{\mathrm{KL},2}(\boldsymbol{x})}) \log(e^{F_0(\boldsymbol{x})} + e^{-F_0(\boldsymbol{x})}) \\ &= 0,\end{aligned}$$

and $F_{\mathrm{KL},2} = F_0$ means $F_0(\boldsymbol{x}) = 0$ ($\forall \boldsymbol{x}$), concluding Equation (9).

### C. Proof of Lemma 4

If Equation (15) holds, $F_0$ satisfies:

$$\begin{aligned}\left.\frac{\delta D_U(p, q_F; r)}{\delta F(\boldsymbol{x})}\right|_{F=F_0} &= \left(1 - \frac{1}{\sum_{y\in\mathcal{Y}} q_{F_0}(y|\boldsymbol{x})}\right) \sum_{y\in\mathcal{Y}} y\xi'(q_{F_0}(y|\boldsymbol{x})) q_{F_0}(y|\boldsymbol{x})^2 \\ &\propto \xi'(e^{F_0(\boldsymbol{x})}) e^{2F_0(\boldsymbol{x})} - \xi'(e^{-2F_0(\boldsymbol{x})}) e^{-2F_0(\boldsymbol{x})} \\ &= 0.\end{aligned}$$

By setting $z = e^{F_0(\boldsymbol{x})}$, we have $z^2\xi'(z) = z^{-2}\xi'(z^{-1})$, and the function $\xi'(z)z^2$ is reflection-symmetric.

If Equation (16) holds, $F_0$ satisfies:

$$\begin{aligned}
&\left.\frac{\delta D_U(q_F, p; r)}{\delta F(\boldsymbol{x})}\right|_{F=F_0} \\
=&\sum_{y\in\mathcal{Y}} y q_{F_0}(y|\boldsymbol{x}) \left\{\xi(q_{F_0}(y|\boldsymbol{x})) - \xi(\bar{q}_{F_0}(y|\boldsymbol{x}))\right\} \\
=&e^{F_0(\boldsymbol{x})}\left\{\xi(e^{F_0(\boldsymbol{x})}) - \xi\left(\frac{e^{F_0(\boldsymbol{x})}}{e^{F_0(\boldsymbol{x})}+e^{-F_0(\boldsymbol{x})}}\right)\right\} - e^{-F_0(\boldsymbol{x})}\left\{\xi(e^{-F_0(\boldsymbol{x})}) - \xi\left(\frac{e^{-F_0(\boldsymbol{x})}}{e^{F_0(\boldsymbol{x})}+e^{-F_0(\boldsymbol{x})}}\right)\right\} \\
=&0,
\end{aligned}$$

implying that the function $z\left\{\xi(z) - \xi\left(\frac{z}{z+z^{-1}}\right)\right\}$ is reflection-symmetric.

### D. Proof of Theorem 5

For the proof of the theorem, we firstly prepare the following lemmas.

**Lemma 11.** *Let $f(z)$ be a reflection-symmetric and holomorphic function on $z \neq 0$. Then, $a_k = b_k$ holds for all $k \geq 1$.*

**Proof.** The function $f$ can be expressed as Equation (14), and let us assume that there exists an integer $k_0$, such that $a_{k_0} \neq b_{k_0}$. From the reflection-symmetric property, we have:

$$(a_{k_0} - b_{k_0})(z^{k_0} - z^{-k_0}) = 0 \tag{44}$$

for all $z > 0$, which contradicts $a_{k_0} \neq b_{k_0}$. □

**Lemma 12.** *Let $\xi(z)$ be a holomorphic function on $z \neq 0$. If two functions:*

$$\xi'(z)z^2, \text{ and } z\left\{\xi(z) - \xi\left(\frac{z}{z+z^{-1}}\right)\right\} \tag{45}$$

*are both reflection-symmetric, then $\xi(z) = \frac{c_1}{z} + c_0$.*

**Proof.** We can express the function $\xi(z)$ by a Laurent series as:

$$\xi(z) = c + \sum_{k=1}^{\infty}\left(a_k z^k + b_k z^{-k}\right). \tag{46}$$

Then, we have:

$$\begin{aligned}
\xi'(z)z^2 &= \sum_{k=1}^{\infty} k\left(a_k z^{k+1} - b_k z^{-k+1}\right) \\
&= -b_1 - 2b_2 z^{-1} + \sum_{k=1}^{\infty}\left(k a_k z^{k+1} - (k+2) b_{k+2} z^{-k-1}\right).
\end{aligned} \tag{47}$$

Because of the assumption of reflection-symmetry for $z^2\xi'(z)$ and Lemma 11, we have $b_2 = 0$ and $ka_k = -(k+2)b_{k+2}$ for all $k \geq 1$. Thus, we obtain:

$$\begin{aligned}\xi(z) &= \int -\frac{b_1}{z^2} + \sum_{k=1}^{\infty} a_k \left(kz^{k-1} + kz^{-k-3}\right) dz \\ &= c + b_1 z^{-1} + \sum_{k=1}^{\infty} a_k \left(z^k - \frac{k}{k+2} z^{-k-2}\right). \end{aligned} \tag{48}$$

Then, we have:

$$\begin{aligned}&z\left\{\xi(z) - \xi\left(\frac{z}{z+z^{-1}}\right)\right\} \\ =&b_1(1-(z+z^{-1})) + \sum_{k=1}^{\infty} a_k \left\{z^{k+1}(1-(z+z^{-1})^{-k}) - \frac{k}{k+2} z^{-k-1}(1-(z+z^{-1})^{k+2})\right\}. \end{aligned} \tag{49}$$

From Equation (48) and the assumption of the reflection-symmetry of the function $z\left\{\xi(z) - \xi\left(\frac{z}{z+z^{-1}}\right)\right\}$, we observe that for all $z$,

$$\begin{aligned} z\left\{\xi(z) - \xi\left(\frac{z}{z+z^{-1}}\right)\right\} - z^{-1}\left\{\xi(z^{-1}) - \xi\left(\frac{z^{-1}}{z+z^{-1}}\right)\right\} &= \sum_{k=1}^{\infty} a_k h_k(z) \\ &= 0 \end{aligned} \tag{50}$$

where:

$$h_k(z) = \left(z^{k+1} - z^{-k-1}\right)\left\{1 - (z+z^{-1})^{-k} + \frac{k}{k+2}\left\{1-(z+z^{-1})^{k+2}\right\}\right\}. \tag{51}$$

Since $\{h_k(z)\}_{k=1}^{\infty}$ is functionally independent, we conclude that $a_k = 0$ for all $k \geq 1$ or, equivalently, $\xi(z) = c + \frac{b_1}{z}$. $\square$

We now give a proof for Theorem 5 using Lemma 12.

**Proof.** If condition Equations (19) and (20) hold, functions $\xi'(z)z^2$ and $z\left\{\xi(z) - \xi\left(\frac{z}{z+z^{-1}}\right)\right\}$ are both reflection-symmetric from Lemma 4. From Lemma 12, the reflection-symmetric property of these two functions implies $\xi(z) = \frac{b_1}{z} + c$. Since the function $\xi$ should be defined on $z > 0$, the generating function $U$ derived from $\xi$ is written as:

$$U(z) = \int \xi^{-1}(z)dz = b_1 \log(c-z) + c_1 \ (z < c) \tag{52}$$

where $c_1$ is a constant and $b_1 < 0$ holds because of the convexity of function $U$. Then, we have $U(\xi(z))) = b_1 \log(-b_1) - b_1 \log z + c_1 (z > 0)$, and the associated divergence is equivalent to the IS distance, *i.e.*,

$$\begin{aligned} D_U(p,q;r) &= \int r(\boldsymbol{x}) \sum_{y\in\mathcal{Y}} \left\{-b_1 \log \frac{q(y|\boldsymbol{x})}{p(y|\boldsymbol{x})} - p(y|\boldsymbol{x})\left\{\frac{b_1}{q(y|\boldsymbol{x})} - \frac{b_1}{p(y|\boldsymbol{x})}\right\}\right\} d\boldsymbol{x} \\ &= -b_1 \int r(\boldsymbol{x}) \sum_{y\in\mathcal{Y}} \left\{\log \frac{q(y|\boldsymbol{x})}{p(y|\boldsymbol{x})} + \frac{p(y|\boldsymbol{x})}{q(y|\boldsymbol{x})} - 1\right\} d\boldsymbol{x} \\ &= -b_1 \,\mathrm{IS}(p,q;r), \end{aligned} \tag{53}$$

up to the constant $-b_1$. $\square$

## E. Proof of Proposition 7

From Equation (28), we observe:

$$
\begin{aligned}
&F_k^*(\boldsymbol{x}) \\
=&\log \frac{\sqrt{p_k(+1|\boldsymbol{x}) + \frac{1}{4p_k(-1|\boldsymbol{x})}(\sum_{j\neq k}\lambda_{k,j}\left\{e^{-\eta_j(\boldsymbol{x})} - e^{\eta_j(\boldsymbol{x})}\right\})^2} - \frac{1}{2\sqrt{p_k(-1|\boldsymbol{x})}}\sum_{j\neq k}\lambda_{k,j}\left\{e^{-\eta_j(\boldsymbol{x})} - e^{\eta_j(\boldsymbol{x})}\right\}}{\sqrt{p_k(-1|\boldsymbol{x})}} \\
\simeq&\frac{1}{2}\log\frac{p_k(+1|\boldsymbol{x})}{p_k(-1|\boldsymbol{x})} + \frac{1}{P}\sum_{j\neq k}\lambda_{k,j}\eta_j(\boldsymbol{x}).
\end{aligned}
$$

## F. Proof of Proposition 8

We observe that:

$$
\begin{aligned}
&\left(\bar{F}_k^*(\boldsymbol{x}) - \frac{1}{2}\log\frac{p_0(+1|\boldsymbol{x})}{p_0(-1|\boldsymbol{x})}\right)^2 - \left(F_k^*(\boldsymbol{x}) - \frac{1}{2}\log\frac{p_0(+1|\boldsymbol{x})}{p_0(-1|\boldsymbol{x})}\right)^2 \\
=&\frac{1}{4P^4(P+\lambda_k)^2}\left(\lambda_k\delta_k(\boldsymbol{x}) - \sum_{j\neq k}\lambda_{k,j}\epsilon_j(\boldsymbol{x})\right)\left((\lambda_k+2P)\delta_k(\boldsymbol{x}) + \sum_{j\neq k}\lambda_{k,j}\epsilon_j(\boldsymbol{x})\right),
\end{aligned}
$$

which implies the proposition.

## G. Proof of Proposition 9

The minimizer of the risk function Equation (27) satisfies:

$$
\begin{aligned}
\frac{\delta L(F_1,\ldots,F_J)}{\delta F_k} \propto& e^{F_k(\boldsymbol{x})}\left\{\pi_k p_k(-1|\boldsymbol{x}) + \sum_{j\neq k}(\pi_k\lambda_{k,j} + \pi_j\lambda_{j,k})e^{-F_j(\boldsymbol{x})}\right\} \\
&- e^{-F_k(\boldsymbol{x})}\left\{\pi_k p_k(+1|\boldsymbol{x}) + \sum_{j\neq k}(\pi_k\lambda_{k,j} + \pi_j\lambda_{j,k})e^{F_j(\boldsymbol{x})}\right\} \\
=&0,
\end{aligned}
$$

implying Equation (36).

## References


1. Caruana, R. Multitask learning. *Mach. Learn.* **1997**, *28*, 41–75.
2. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359.
3. Argyriou, A.; Pontil, M.; Ying, Y.; Micchelli, C.A. A spectral regularization framework for multi-task structure learning. In *Advances in Neural Information Processing Systems 19*; MIT Press: Cambridge, MA, USA, 2007.
4. Evgeniou, A.; Pontil, M. Multi-task feature learning. In *Advances in Neural Information Processing Systems 19*; MIT Press: Cambridge, MA, USA, 2007.

5. Dai, W.; Yang, Q.; Xue, G.R.; Yu, Y. Boosting for transfer learning. In Proceedings of the 24th International Conference on Machine Learning, Corvallis, OR, USA, 20–24 June 2007; pp. 193–200.
6. Wang, X.; Zhang, C.; Zhang, Z. Boosted multi-task learning for face verification with applications to web image and video search. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp.142–149.
7. Chapelle, O.; Shivaswamy, P.; Vadrevu, S.; Weiinberger, K.; Zhang, Y.; Tseng, B. Multi-task learning for boosting with application to web search ranking. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 25–28 July 2010; pp. 1189–1198
8. Cichocki, A.; Amari, S. Families of alpha- beta- and gamma-divergences: Flexible and robust measures of similarities. *Entropy* **2010**, *12*, 1532–1568.
9. Févotte, C.; Bertin, N.; Durrieu, J.L. Nonnegative matrix factorization with the Itakura–Saito divergence: With application to music analysis. *Neural Comput.* **2009**, *21*, 793–830.
10. Lefevre, A.; Bach, F.; Févotte, C. Itakura–Saito nonnegative matrix factorization with group sparsity. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech, 22–27 May 2011; pp. 21–24.
11. Takenouchi, T.; Komori, O.; Eguchi, S. A novel boosting algorithm for multi-task learning based on the Itakura–Saito divergence. In Proceedings of the Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MaxEnt 2014), Amboise, France, 21–26 September 2014; pp. 230–237.
12. Murata, N.; Takenouchi, T.; Kanamori, T.; Eguchi, S. Information geometry of $U$-boost and Bregman divergence. *Neural Comput.* **2004**, *16*, 1437–1481.
13. Banerjee, A.; Merugu, S.; Dhillon, I.S.; Ghosh, J. Clustering with Bregman divergences. *J. Mach. Learn. Res.* **2005**, *6*, 1705–1749.
14. Amari, S.; Nagaoka, H. *Methods of Information Geometry of Translations of Mathematical Monographs*; Oxford University Press: Providence, RI, USA, 2000; Volume 191.
15. Mihoko, M.; Eguchi, S. Robust blind source separation by beta divergence. *Neural Comput.* **2002**, *14*, 1859–1886.
16. Cichocki, A.; Cruces, S.; Amari, S.I. Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy* **2011**, *13*, 134–170.
17. Takenouchi, T.; Eguchi, S. Robustifying AdaBoost by adding the naive error rate. *Neural Comput.* **2004**, *16*, 767–787.
18. Takenouchi, T.; Eguchi, S.; Murata, T.; Kanamori, T. Robust boosting algorithm against mislabeling in multi-class problems. *Neural Comput.* **2008**, *20*, 1596–1630.
19. Lafferty, G.L.J. Boosting and maximum likelihood for exponential models. In *Advances in Neural Information Processing Systems 14*; MIT Press: Cambridge, MA, USA, 2002.
20. Evgeniou, T.; Pontil, M. Regularized multi-task learning. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA, 22–25 August 2004; pp. 109–117.

21. Xue, Y.; Liao, X.; Carin, L.; Krishnapuram, B. Multi-task learning for classification with Dirichlet process priors. *J. Mach. Learn. Res.* **2007**, *8*, 35–63.
22. Mason, L.; Baxter, J.; Bartlett, P.; Frean, M. Boosting algorithms as gradient decent in function space. In *Advances in Neural Information Processing Systems 11*; MIT Press: Cambridge, MA, USA, 1999.
23. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3–42.
24. Goldstein, H. Multilevel modelling of survey data. *J. R. Stat. Soc. Ser. D* **1991**, *40*, 235–244.

MDPI AG
Klybeckstrasse 64
4057 Basel, Switzerland
Tel. +41 61 683 77 34
Fax +41 61 302 89 18
http://www.mdpi.com/

*Entropy* Editorial Office
E-Mail: entropy@mdpi.com
http://www.mdpi.com/journal/entropy

MDPI • Basel • Beijing • Wuhan ÅBarcelona
ISBN 978-3-03842-104-7
www.mdpi.com