

Article

Information Geometry and Asymptotic Theory for SMML Estimators

Enes Makalic ^{*}  and Daniel F. Schmidt 

Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia;
daniel.schmidt@monash.edu

^{*} Correspondence: enes.makalic@monash.edu

Abstract

Strict minimum message length (SMML) is an information-theoretic coding principle that represents a continuous statistical model by a finite set of assertions and a partition of the sample space. We show that the SMML objective decomposes into assertion entropy and conditional cross-entropy, balancing the cost of identifying an assertion against the cost of encoding data under the assigned model. For any fixed partition, the optimal codepoint for each cell is the model distribution that minimises Kullback–Leibler (KL) divergence from the data distribution restricted to that cell. Using the local Fisher–Rao geometry of regular parametric models, we show that, under a high-resolution LAN-scale regime, SMML partitions are asymptotically the pullback, through the maximum-likelihood estimator, of weighted Fisher–Rao Voronoi tessellations in parameter space, with assertion probabilities appearing as additive weights. For regular canonical exponential families, SMML codepoints satisfy a moment-matching condition and admit an interpretation as KL/Bregman centroids, while exact SMML cells are pullbacks of convex polyhedra in sufficient-statistic space. Together, these results show that SMML induces a natural information-geometric quantisation linking entropy-based coding, KL projection, and divergence-based Voronoi geometry.

Keywords: strict minimum message length; Kullback–Leibler divergence; Fisher–Rao geometry; Bregman divergence; exponential families; Voronoi tessellations

1. Introduction

Minimum message length (MML) inference [1–6] is an information-theoretic approach to statistical estimation in which data are encoded by a two-part message: an assertion specifying a model or parameter value, followed by an encoding the data conditional on that assertion. Strict minimum message length (SMML) [2,6] is the ideal finite-codebook version of this principle. It replaces a continuous parameter space by a finite set of assertions, assigns each data set to one assertion, and chooses both the assertions and the induced partition of the sample space by minimising expected codelength.

In this paper, we study SMML as an entropy–cross-entropy coding principle for finite-resolution statistical inference. The SMML objective balances assertion entropy against the conditional cross-entropy of encoding data under the assigned model, thereby quantising a continuous statistical model by a finite codebook. We show that, in regular parametric models and under a high-resolution local asymptotic normality (LAN) regime, this quantisation has an asymptotic Fisher–Rao geometry in which the codepoints are characterised by Kullback–Leibler (KL) projections [7] and the induced partitions are asymptotically governed



Academic Editor: Frank Nielsen

Received: 12 May 2026

Revised: 12 June 2026

Accepted: 15 June 2026

Published: 22 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

by weighted Fisher–Rao Voronoi tessellations. In exponential families, the same structure appears explicitly as KL/Bregman centroids satisfying moment-matching conditions.

Let $\mathbf{x} = (x_1, \dots, x_n) \in \mathcal{X}_n$ denote a data set of size n , where $\mathcal{X}_n$ is assumed to be countable, possibly infinite. We consider a parametric family

$$\{p_n(\mathbf{x} \mid \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p\},$$

with prior density $\pi(\boldsymbol{\theta})$. We assume that the data $\mathbf{x} \in \mathcal{X}_n$ form an independent and identically distributed sample, so that $p_n(\mathbf{x} \mid \boldsymbol{\theta}) = \prod_{i=1}^n f(x_i \mid \boldsymbol{\theta})$, where $f(\cdot \mid \boldsymbol{\theta})$ denotes the per-observation probability mass function. The marginal distribution of the data is

$$r_n(\mathbf{x}) = \int_{\Theta} p_n(\mathbf{x} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}. \quad (1)$$

Let $\mathcal{P} = \{P_1, \dots, P_k\}$ be a finite partition of $\mathcal{X}_n$ into non-empty cells, and define the assertion probabilities

$$q_j = \sum_{\mathbf{x} \in P_j} r_n(\mathbf{x}), \quad j = 1, \dots, k. \quad (2)$$

Throughout, we restrict attention to partitions with $q_j > 0$ for all j . Under SMML encoding [2,6,8], the expected two-part codelength associated with $\mathcal{P}$ is

$$\mathcal{I}(\mathcal{P}) = \underbrace{-\sum_{j=1}^k q_j \log q_j}_{H(Q): \text{assertion entropy}} - \underbrace{\sum_{j=1}^k \sum_{\mathbf{x} \in P_j} r_n(\mathbf{x}) \log p_n(\mathbf{x} \mid \boldsymbol{\theta}_j^*)}_{\text{expected detail codelength}} \quad (3)$$

$$= H(r_n) + \sum_{j=1}^k q_j D_{\text{KL}}(\bar{r}_{j,n} \parallel p_n(\cdot \mid \boldsymbol{\theta}_j^*)), \quad (4)$$

where $Q = (q_1, \dots, q_k)$ is the distribution of the assertion message, $H(r_n)$ is the entropy of the marginal distribution which is assumed to be finite, D_{KL} denotes Kullback–Leibler (KL) divergence and

$$\bar{r}_{j,n}(\mathbf{x}) = \frac{r_n(\mathbf{x})}{q_j} \mathbb{1}\{\mathbf{x} \in P_j\}$$

is the conditional marginal distribution in cell P_j . Unless otherwise stated, logarithms are natural, so codelengths are measured in nats. The first term in (3) is the Shannon entropy of the assertion distribution, while the second is the expected detail codelength incurred when data in cell P_j are encoded using $p_n(\cdot \mid \boldsymbol{\theta}_j^*)$.

For each cell P_j , the associated SMML codepoint is chosen to minimise the expected detail codelength within that cell:

$$\boldsymbol{\theta}_j^* = \arg \min_{\boldsymbol{\theta} \in \Theta} \sum_{\mathbf{x} \in P_j} r_n(\mathbf{x}) \{-\log p_n(\mathbf{x} \mid \boldsymbol{\theta})\} = \arg \max_{\boldsymbol{\theta} \in \Theta} \sum_{\mathbf{x} \in P_j} r_n(\mathbf{x}) \log p_n(\mathbf{x} \mid \boldsymbol{\theta}). \quad (5)$$

From (4), the optimal SMML codepoints satisfy

$$\boldsymbol{\theta}_j^* = \arg \min_{\boldsymbol{\theta} \in \Theta} D_{\text{KL}}(\bar{r}_{j,n} \parallel p_n(\cdot \mid \boldsymbol{\theta})), \quad j = 1, \dots, k. \quad (6)$$

Thus, for a fixed partition, each SMML codepoint is the KL, or information, projection of the normalised cellwise distribution onto the parametric model family in the sense of Csiszár [9]. The remaining optimisation problem is to choose the partition itself,

$$\mathcal{P}^* = \arg \min_{\mathcal{P} \in \Pi} \mathcal{I}(\mathcal{P}), \quad (7)$$

where Π is a prescribed class of admissible partitions of $\mathcal{X}_n$. The resulting partition and codepoints $\Theta^* = \{\theta_1^*, \dots, \theta_k^*\}$ determine the SMML estimator [2,6,8]. For the optimal partition, we write q_j^* , or $q_{j,n}^*$ in sample-size-dependent settings for the corresponding assertion probabilities.

We make four main contributions. First, we characterise SMML codepoints as KL projections of conditional cell distributions onto the model family, making explicit the entropy–cross-entropy structure of the SMML objective. Second, under a high-resolution local asymptotic normality (LAN) regime, we show that SMML partitions are asymptotically the pullback, through the maximum-likelihood estimator (MLE), of weighted Fisher–Rao Voronoi tessellations in parameter space. Third, in the same regime, we derive an asymptotic weighted-average representation of the SMML codepoints and use it to establish consistency and the usual $n^{-1/2}$ parametric rate. Fourth, for exponential families, we show that SMML codepoints satisfy moment matching and that the induced partition is polyhedral in sufficient-statistic space, connecting SMML to KL/Bregman centroids and dually flat geometry.

Our exponential-family results are closely related to Dowty’s work on SMML estimators with continuous sufficient statistics [10]. Dowty showed that the regions of such estimators are convex polytopes described in terms of assertions and coding probabilities. We obtain the corresponding polyhedral structure in the countable sample-space setting and connect it to KL projection, Bregman centroids, and the asymptotic Fisher–Rao quantisation developed here. The covering interpretation used below is also related to the distinguishable-distribution perspective of Balasubramanian and of Myung et al. [11,12], where Fisher–Rao volume is interpreted as a count of statistically distinguishable probability distributions. Whereas that work concerns model selection and geometric model complexity, the present paper studies how SMML induces a finite entropy-coded codebook and partition within a fixed parametric model.

The remainder of the paper is organised as follows. Section 2 introduces the regularity conditions and the asymptotic quantisation framework. Section 3 establishes the connection between SMML partitions and weighted Fisher–Rao Voronoi tessellations, and derives the weighted-average representation of SMML codepoints. Section 4 gives the statistical consequences, including the consistency and the $n^{-1/2}$ convergence rate. Section 5 specialises the theory to exponential families, where codepoints become KL/Bregman centroids and the partition is polyhedral in sufficient-statistic space. Section 6 discusses connections with entropy-based coding, information-theoretic quantisation, and possible extensions.

2. Asymptotic Quantisation Framework and Regularity Conditions

The SMML objective in (3) defines a finite-resolution information-theoretic approximation to a continuous statistical model. Each assertion corresponds to a codepoint in parameter space, and each data set is assigned to one of these codepoints through the induced partition of the data space $\mathcal{X}_n$. Thus, an SMML code simultaneously induces a partition of the sample space and a finite quantisation of the parameter space Θ .

For a sample of size n , let

$$\Theta_n^* = \{\theta_{1,n}^*, \dots, \theta_{k,n}^*\}$$

denote the SMML codebook, where k_n is the number of assertions. The asymptotic regime considered here is a high-resolution regime in which k_n may grow with n and the local mesh of the induced quantisation shrinks as the data become more informative. This is the regime in which a finite SMML codebook can approximate the underlying continuous model.

The relevant local geometry is governed by Kullback–Leibler divergence [7]. For nearby parameter values,

$$D_{\text{KL}}(f(\cdot | \boldsymbol{\theta}) \| f(\cdot | \boldsymbol{\theta} + d\boldsymbol{\theta})) = \frac{1}{2} d\boldsymbol{\theta}' \mathbf{J}_1(\boldsymbol{\theta}) d\boldsymbol{\theta} + o(\|d\boldsymbol{\theta}\|^2),$$

where $\mathbf{J}_1(\boldsymbol{\theta})$ is the per-observation Fisher information matrix. Hence, the natural local metric for analysing SMML codepoints is the per-observation Fisher–Rao metric. Since the full-sample Fisher information is $n\mathbf{J}_1(\boldsymbol{\theta})$, displacements of order $n^{-1/2}$ in this metric correspond to constant-order changes in full-sample likelihood geometry.

This $n^{-1/2}$ scale is the statistical scale of distinguishability. Balasubramanian [11] and Myung et al. [12] show that Fisher–Rao volume may be interpreted as counting statistically distinguishable probability distributions. For a finite sample size, nearby distributions inside a Fisher information ellipsoid are difficult to distinguish, and these indistinguishability neighbourhoods shrink at the parametric rate as n grows. The same scale appears in local asymptotic normality (LAN). Under the local reparameterisation $\boldsymbol{\theta} = \boldsymbol{\theta}_0 + \mathbf{h}/\sqrt{n}$, the log-likelihood ratio admits a non-degenerate quadratic Gaussian limit, with curvature given by the Fisher information.

In the SMML setting, the assertion codebook may therefore be interpreted as a finite entropy-coded covering of the model at a chosen Fisher–Rao resolution. On a compact set $K \subset \Theta$, a Fisher–Rao ball of radius δ_n has volume of order δ_n^p . Thus, up to curvature and boundary effects, the number of assertions required to cover K at uniform resolution δ_n is heuristically

$$N_K(\delta_n) \asymp \frac{\text{Vol}_F(K)}{\delta_n^p}, \quad \text{Vol}_F(K) = \int_K |\mathbf{J}_1(\boldsymbol{\theta})|^{1/2} d\boldsymbol{\theta}.$$

More generally, for a non-uniform mesh with local radius $\delta_n(\boldsymbol{\theta})$,

$$k_n(K) \asymp \int_K \frac{|\mathbf{J}_1(\boldsymbol{\theta})|^{1/2}}{\delta_n(\boldsymbol{\theta})^p} d\boldsymbol{\theta}.$$

The natural benchmark resolution is the LAN-scale $\delta_n(\boldsymbol{\theta}) \asymp n^{-1/2}$, which yields $k_n(K) \asymp n^{p/2} \text{Vol}_F(K)$. Thus, $n^{p/2}$ is the benchmark order for a LAN-scale codebook on compact subsets.

The results below are established under the following regularity conditions. Assumptions (A1)–(A6) are standard likelihood regularity conditions; see, e.g., [13]. Assumption (A7) specifies the local high-resolution structure of the SMML quantisation by placing the selected SMML codepoint within an $O(n^{-1/2})$ Fisher–Rao neighbourhood of the MLE and requiring local quadratic Fisher–Rao loss inside the corresponding cells.

- (A1) The parameter space Θ is an open subset of $\mathbb{R}^p$.
- (A2) The true parameter $\boldsymbol{\theta}_0$ belongs to $\text{int}(\Theta)$.
- (A3) The support of $p_n(\mathbf{x}|\boldsymbol{\theta})$ does not depend on $\boldsymbol{\theta} \in \Theta$. The prior $\pi(\boldsymbol{\theta})$ is a proper density on Θ , continuous and strictly positive on a neighbourhood of the true parameter $\boldsymbol{\theta}_0$. For each n and each $\mathbf{x} \in \mathcal{X}_n$, the marginal distribution

$$r_n(\mathbf{x}) = \int_{\Theta} p_n(\mathbf{x} | \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

is finite.

(A4) For each $\mathbf{x} \in \mathcal{X}_n$, the map $\boldsymbol{\theta} \mapsto \log p_n(\mathbf{x}|\boldsymbol{\theta})$ is three times continuously differentiable on Θ . Moreover, for all data sets $\mathbf{x}$ in the SMML cells under consideration, the maximum likelihood estimator $\hat{\boldsymbol{\theta}}(\mathbf{x})$ exists and belongs to Θ . The model is identifiable and satisfies standard regularity conditions under which the maximum likelihood estimator is consistent at the usual parametric rate.

(A5) Let

$$\ell_n(\mathbf{x}|\boldsymbol{\theta}) = \log p_n(\mathbf{x}|\boldsymbol{\theta})$$

denote the log-likelihood for an i.i.d. data set of size n , and let

$$\mathbf{J}_1(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[\nabla_{\boldsymbol{\theta}} \log f(X_1|\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \log f(X_1|\boldsymbol{\theta})'] \quad (8)$$

denote the per-observation Fisher information matrix. Thus, the full-sample Fisher information is

$$\mathbf{J}_n(\boldsymbol{\theta}) = n \mathbf{J}_1(\boldsymbol{\theta}).$$

Assume that $\mathbf{J}_1(\boldsymbol{\theta})$ exists and is locally uniformly positive definite on Θ : for every compact set $K \subset \Theta$, there exist constants $0 < c_1(K) \leq c_2(K) < \infty$ such that

$$c_1(K) \mathbf{I}_p \preceq \mathbf{J}_1(\boldsymbol{\theta}) \preceq c_2(K) \mathbf{I}_p, \quad \boldsymbol{\theta} \in K, \quad (9)$$

where $\mathbf{I}_p$ is the $p \times p$ identity matrix and $\preceq$ denotes the Loewner order.

(A6) Differentiation under the integral sign is valid whenever it is used below.

(A7) Fix a compact set $K \subset \Theta$ on which (A5) holds. The conditional distributions of $\sqrt{n} d_F(\hat{\boldsymbol{\theta}}(\mathbf{X}_n), \boldsymbol{\theta}_{j,n}^*)$ given $\mathbf{X}_n \in P_{j,n}^*$ are uniformly tight over SMML cells with $\boldsymbol{\theta}_{j,n}^* \in K$: for every $\epsilon > 0$, there exists $M < \infty$ such that, for all sufficiently large n ,

$$\sup_{j: \boldsymbol{\theta}_{j,n}^* \in K} \mathbb{P}\left(\sqrt{n} d_F(\hat{\boldsymbol{\theta}}(\mathbf{X}_n), \boldsymbol{\theta}_{j,n}^*) > M \mid \mathbf{X}_n \in P_{j,n}^*\right) < \epsilon.$$

Moreover,

$$\ell_n(\mathbf{X}_n \mid \boldsymbol{\theta}_{j,n}^*) = \ell_n(\mathbf{X}_n \mid \hat{\boldsymbol{\theta}}(\mathbf{X}_n)) - \frac{n}{2} d_F^2(\hat{\boldsymbol{\theta}}(\mathbf{X}_n), \boldsymbol{\theta}_{j,n}^*) + o_p(1),$$

conditional on $\mathbf{X}_n \in P_{j,n}^*$, uniformly over such cells.

Assumption (A7) should be understood as a conditional high-resolution regularity condition. It places the SMML codebook in an LAN-scale regime in which the selected codepoint is $O(n^{-1/2})$ from the MLE in Fisher–Rao distance, and likelihood differences inside the corresponding cells are governed by local quadratic Fisher–Rao loss. Consequently, the asymptotic results in Sections 3 and 4 are conditional on this high-resolution regime. We do not claim here that this regime follows automatically from global SMML optimality for every regular model.

Remark 1 (Rate–distortion interpretation of the high-resolution regime). *The SMML objective decomposes into an assertion-entropy term and a conditional Kullback–Leibler distortion term, up to partition-independent constants. The assertion entropy plays the role of rate, while the conditional KL term plays the role of distortion. In local coordinates, the KL distortion is quadratically approximated by Fisher information, so the high-resolution SMML problem is naturally compared with high-rate quantisation under quadratic Fisher–Rao distortion. From this viewpoint, (A7) places the SMML estimator in the regime where the entropy–cross-entropy trade-off is resolved at the LAN scale.*

A local split-optimality heuristic suggests why this scale is natural. Splitting a cell of assertion mass q_j incurs an entropy cost of order q_j , whereas, under local quadratic Fisher–Rao loss, the potential reduction in detail codelength is of order $q_j n r_j^2$ for a cell of effective Fisher–Rao radius r_j . Balancing these terms heuristically gives $r_j^2 \asymp n^{-1}$. A rigorous derivation of this LAN-scale localisation directly from global SMML optimality, including boundary and heterogeneity effects, is left for future work.

3. Fisher–Rao Geometry of SMML

We now derive the local geometric form of the SMML decision rule under the high-resolution regime described in (A7). In this regime, each data set $\mathbf{x}$ is first mapped to parameter space by its maximum-likelihood estimate (MLE) $\hat{\theta}(\mathbf{x})$. The selected SMML assertion is then the nearby codepoint that gives the shortest asymptotic two-part message. Since the loss from replacing $\hat{\theta}(\mathbf{x})$ by a nearby codepoint is governed locally by Fisher–Rao quadratic loss, the assignment rule takes the form of a weighted Fisher–Rao Voronoi decision rule, with the assertion probabilities appearing as additive weights.

Throughout this section, $d_F(\theta_1, \theta_2)$ denotes the per-observation Fisher–Rao distance induced by the metric $J_1(\theta)$. For nearby parameter values,

$$d_F^2(\theta_1, \theta_2) = (\theta_1 - \theta_2)' J_1(\bar{\theta})(\theta_1 - \theta_2) + o(\|\theta_1 - \theta_2\|^2), \quad (10)$$

where $\bar{\theta}$ lies on the line segment joining θ_1 and θ_2 . The use of the per-observation metric separates the intrinsic geometry of the model from the sample-size scaling of the full likelihood, whose curvature is $nJ_1(\theta)$.

Theorem 1 (SMML partitions and weighted Fisher–Rao Voronoi cells). *Let*

$$\mathcal{P}_n^* = \{P_{j,n}^*\}_{j=1}^{k_n}, \quad \Theta_n^* = \{\theta_{j,n}^*\}_{j=1}^{k_n}$$

denote an optimal SMML partition and its associated codepoints, with assertion probabilities

$$q_{j,n}^* = \sum_{\mathbf{x} \in P_{j,n}^*} r_n(\mathbf{x}) > 0.$$

Define the weighted Fisher–Rao Voronoi cells

$$V_{j,n} = \left\{ \theta \in \Theta : d_F^2(\theta, \theta_{j,n}^*) + \omega_{j,n} \leq d_F^2(\theta, \theta_{\ell,n}^*) + \omega_{\ell,n}, \forall \ell \right\}, \quad (11)$$

where

$$\omega_{j,n} = -\frac{2}{n} \log q_{j,n}^*.$$

Under (A1)–(A7), and assuming that the local quadratic approximation in (A7) holds uniformly for all assertions that are locally competitive with the selected assertion, the optimal SMML partition is, up to tie-breaking on a set of probability $o(1)$, the pullback of this weighted Fisher–Rao Voronoi tessellation under the MLE map

$$\mathbb{P}(\mathbf{X}_n \in P_{j,n}^* \iff \hat{\theta}(\mathbf{X}_n) \in V_{j,n}) \rightarrow 1, \quad (12)$$

for each $j = 1, \dots, k_n$. Here, an assertion is called locally competitive for x if its two-part codelength differs from the minimum two-part codelength by $O(1)$. Moreover, the pairwise boundary between two locally competing cells, j and ℓ , satisfies

$$\frac{n}{2} \left\{ d_F^2(\boldsymbol{\theta}, \boldsymbol{\theta}_{j,n}^*) - d_F^2(\boldsymbol{\theta}, \boldsymbol{\theta}_{\ell,n}^*) \right\} = \log \frac{q_{j,n}^*}{q_{\ell,n}^*} + o(1), \quad (13)$$

uniformly on compact subsets where the local quadratic approximation is valid. The corresponding data-space boundary is obtained by pulling this boundary back under $\mathbf{x} \mapsto \hat{\boldsymbol{\theta}}(\mathbf{x})$.

Proof. Fix the optimal codebook $\{(\boldsymbol{\theta}_{j,n}^*, q_{j,n}^*)\}_{j=1}^{k_n}$. Conditional on this codebook, assigning $\mathbf{x}$ to assertion j gives the codelength

$$\Lambda_{j,n}(\mathbf{x}) = -\log q_{j,n}^* - \log p_n(\mathbf{x} \mid \boldsymbol{\theta}_{j,n}^*).$$

Thus, the induced partition assigns $\mathbf{x}$ to an index minimising $\Lambda_{j,n}(\mathbf{x})$, up to ties. Let $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{x})$. By (A7),

$$d_F(\hat{\boldsymbol{\theta}}(\mathbf{X}_n), \boldsymbol{\theta}_{j,n}^*) = O_p(n^{-1/2}) \quad \text{conditional on } \mathbf{X}_n \in P_{j,n}^*.$$

Moreover,

$$-\log p_n(\mathbf{X}_n \mid \boldsymbol{\theta}_{j,n}^*) = -\log p_n(\mathbf{X}_n \mid \hat{\boldsymbol{\theta}}(\mathbf{X}_n)) + \frac{n}{2} d_F^2(\hat{\boldsymbol{\theta}}(\mathbf{X}_n), \boldsymbol{\theta}_{j,n}^*) + o_p(1),$$

conditional on $\mathbf{X}_n \in P_{j,n}^*$, uniformly over such cells. By the additional uniform local-competition assumption in the statement of the theorem, the same expansion holds for all assertions whose two-part codelengths are within $O(1)$ of the minimum, and hence for all assertions that can affect the asymptotic decision boundary. Therefore,

$$\Lambda_{j,n}(\mathbf{X}_n) = -\log q_{j,n}^* - \log p_n(\mathbf{X}_n \mid \hat{\boldsymbol{\theta}}) + \frac{n}{2} d_F^2(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_{j,n}^*) + o_p(1).$$

Since the middle term does not depend on j , the selected assertion asymptotically minimises

$$-\log q_{j,n}^* + \frac{n}{2} d_F^2(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_{j,n}^*),$$

or equivalently

$$d_F^2(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_{j,n}^*) - \frac{2}{n} \log q_{j,n}^*,$$

which is the weighted Fisher–Rao Voronoi rule. Equating the asymptotic codelengths for two locally competing assertions j and ℓ yields (13). Pulling this boundary back through the MLE map gives the corresponding boundary in data space. $\square$

Remark 2 (Uniform assertion probabilities). If assertion probabilities are uniformly asymptotically equal, in the sense that

$$\sup_{1 \leq j, \ell \leq k_n} \left| \log \frac{q_{j,n}^*}{q_{\ell,n}^*} \right| \rightarrow 0,$$

then

$$\omega_{j,n} - \omega_{\ell,n} = -\frac{2}{n} \log \frac{q_{j,n}^*}{q_{\ell,n}^*} = o(n^{-1}).$$

Hence, the weighted Fisher–Rao Voronoi tessellation reduces asymptotically to the ordinary Fisher–Rao Voronoi tessellation. In this case, local SMML assignment is governed primarily by Fisher–Rao distance.

Theorem 1 shows that SMML partitions are governed locally by Fisher–Rao geometry. Figure 1 depicts the SMML assignment mechanism, showing how the partition of data space arises due to the pullback, through the MLE map, of weighted Fisher–Rao Voronoi cells in parameter space. The next result identifies the corresponding location of the codepoints in the same high-resolution regime. It shows that, within each SMML cell, the associated codepoint is asymptotically the marginal-predictive weighted average of the MLEs assigned to that cell.

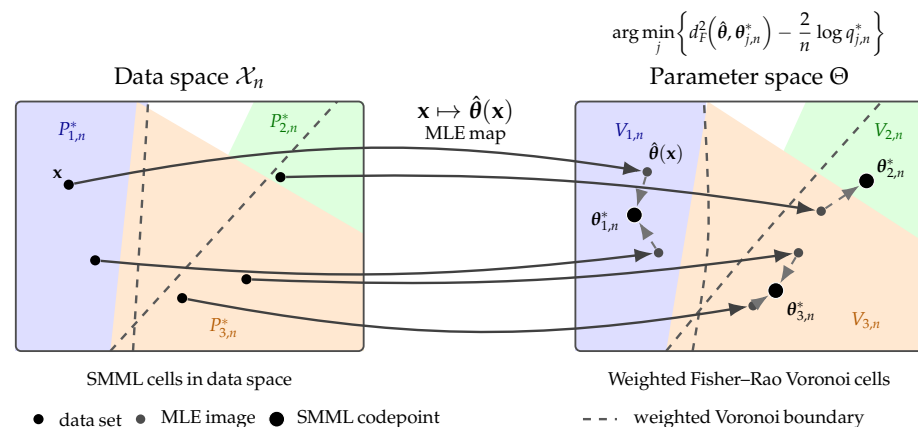


Figure 1. SMML as the pullback of a weighted Fisher–Rao Voronoi tessellation. The (left) panel shows the induced SMML partition of data space $\mathcal{X}_n$, with cells $P_{j,n}^*$. Data sets are mapped to parameter space by the MLE map $\mathbf{x} \mapsto \hat{\boldsymbol{\theta}}(\mathbf{x})$. The (right) panel shows the corresponding weighted Fisher–Rao Voronoi cells $V_{j,n}$ around the active SMML codepoints $\boldsymbol{\theta}_{j,n}^*$. Dashed curves indicate pairwise weighted Voronoi boundaries. The SMML cells in data space are the pullbacks of the parameter-space cells under the MLE map.

Theorem 2 (SMML codepoints as weighted averages of MLEs). *With the notation of Theorem 1, define*

$$w_{j,n}(\mathbf{x}) = \frac{r_n(\mathbf{x})}{q_{j,n}^*}, \quad \mathbf{x} \in P_{j,n}^*.$$

Under (A1)–(A7), fix a compact set $K \subset \Theta$ on which (A5) holds, and consider any sequence of cells such that $\boldsymbol{\theta}_{j,n}^ \in K$. Suppose, in addition, that the following score linearisation holds uniformly for $\mathbf{x} \in P_{j,n}^*$:*

$$\nabla_{\boldsymbol{\theta}} \ell_n(\mathbf{x} \mid \boldsymbol{\theta}_{j,n}^*) = -n \mathbf{J}_1(\boldsymbol{\theta}_{j,n}^*) \{\boldsymbol{\theta}_{j,n}^* - \hat{\boldsymbol{\theta}}(\mathbf{x})\} + \mathbf{r}_{j,n}(\mathbf{x}),$$

with

$$\sup_{\mathbf{x} \in P_{j,n}^*} \|\mathbf{r}_{j,n}(\mathbf{x})\| = o(\sqrt{n}).$$

Then

$$\boldsymbol{\theta}_{j,n}^* = \sum_{\mathbf{x} \in P_{j,n}^*} w_{j,n}(\mathbf{x}) \hat{\boldsymbol{\theta}}(\mathbf{x}) + \boldsymbol{\varepsilon}_{j,n}, \quad \|\boldsymbol{\varepsilon}_{j,n}\| = o(n^{-1/2}).$$

Equivalently,

$$\boldsymbol{\theta}_{j,n}^* = \mathbb{E}_{r_n} [\hat{\boldsymbol{\theta}}(\mathbf{X}_n) \mid \mathbf{X}_n \in P_{j,n}^*] + o(n^{-1/2}).$$

Proof. For a fixed cell j , define the cellwise expected log-likelihood

$$L_{j,n}(\boldsymbol{\theta}) = \sum_{\mathbf{x} \in P_{j,n}^*} r_n(\mathbf{x}) \ell_n(\mathbf{x} \mid \boldsymbol{\theta}). \quad (14)$$

By definition, $\theta_{j,n}^*$ maximises $L_{j,n}(\theta)$. Since Θ is open (A1) and $\theta_{j,n}^* \in K \subset \Theta$ by hypothesis, the first-order condition holds:

$$0 = \nabla_{\theta} L_{j,n}(\theta) \big|_{\theta=\theta_{j,n}^*} = \sum_{\mathbf{x} \in P_{j,n}^*} r_n(\mathbf{x}) \nabla_{\theta} \ell_n(\mathbf{x} | \theta_{j,n}^*). \quad (15)$$

For each $\mathbf{x} \in P_{j,n}^*$, the score linearisation gives

$$\nabla_{\theta} \ell_n(\mathbf{x} | \theta_{j,n}^*) = -n \mathbf{J}_1(\theta_{j,n}^*) \{ \theta_{j,n}^* - \hat{\theta}(\mathbf{x}) \} + \mathbf{r}_{j,n}(\mathbf{x}),$$

where

$$\sup_{\mathbf{x} \in P_{j,n}^*} \|\mathbf{r}_{j,n}(\mathbf{x})\| = o(\sqrt{n}).$$

Substituting this into the first-order condition

$$0 = \sum_{\mathbf{x} \in P_{j,n}^*} r_n(\mathbf{x}) \nabla_{\theta} \ell_n(\mathbf{x} | \theta_{j,n}^*)$$

and dividing by $nq_{j,n}^*$ yields

$$\mathbf{J}_1(\theta_{j,n}^*) \left(\theta_{j,n}^* - \sum_{\mathbf{x} \in P_{j,n}^*} w_{j,n}(\mathbf{x}) \hat{\theta}(\mathbf{x}) \right) = o(n^{-1/2}),$$

where $w_{j,n}(\mathbf{x}) = r_n(\mathbf{x})/q_{j,n}^*$. Since $\mathbf{J}_1(\theta_{j,n}^*)$ is uniformly non-singular on compact subsets, we obtain

$$\theta_{j,n}^* = \sum_{\mathbf{x} \in P_{j,n}^*} w_{j,n}(\mathbf{x}) \hat{\theta}(\mathbf{x}) + o(n^{-1/2}),$$

as claimed. $\square$

4. Consistency

The preceding section shows that, in the high-resolution regime of (A7), SMML behaves locally as an entropy-weighted Fisher–Rao quantisation of the statistical model. The selected SMML codepoint remains within the same local $n^{-1/2}$ -scale neighbourhood as the MLE. Consequently, replacing the continuous MLE by the finite SMML codepoint perturbs estimation only at the usual parametric scale.

Theorem 3 (Consistency and rate of convergence of the SMML estimator). *Let*

$$\mathcal{P}_n^* = \{P_{j,n}^*\}_{j=1}^{k_n}, \quad \Theta_n^* = \{\theta_{j,n}^*\}_{j=1}^{k_n}$$

denote an optimal SMML partition and its associated codepoints. Define the SMML estimator by

$$\hat{\theta}_n^{\text{SMML}}(\mathbf{x}) = \sum_{j=1}^{k_n} \theta_{j,n}^* \mathbb{1}_{\{\mathbf{x} \in P_{j,n}^*\}}.$$

Let

$$\hat{\theta}_n = \hat{\theta}(\mathbf{X}_n)$$

denote the MLE, and suppose that the true parameter is $\theta_0 \in \text{int}(\Theta)$. Under (A1)–(A7), fix compact sets $K_0 \Subset K_1 \Subset \Theta$, such that $\theta_0 \in \text{int}(K_0)$ and (A5) holds for K_1 . Assume, in addition, that the

selected SMML codepoint lies in K_1 with a probability tending to one whenever the MLE lies in K_0 , that is,

$$P(\hat{\theta}_n \in K_0, \hat{\theta}_n^{\text{SMML}} \notin K_1) \rightarrow 0.$$

Then

$$\hat{\theta}_n^{\text{SMML}} \xrightarrow{p} \theta_0, \quad n \rightarrow \infty.$$

Moreover,

$$\|\hat{\theta}_n^{\text{SMML}} - \theta_0\| = O_p(n^{-1/2}). \quad (16)$$

Proof. Write

$$\hat{\theta}_n^{\text{SMML}} := \hat{\theta}_n^{\text{SMML}}(\mathbf{X}_n), \quad \hat{\theta}_n := \hat{\theta}(\mathbf{X}_n).$$

By standard likelihood theory under (A1)–(A6),

$$\hat{\theta}_n \xrightarrow{p} \theta_0, \quad \|\hat{\theta}_n - \theta_0\| = O_p(n^{-1/2}).$$

Since $\theta_0 \in \text{int}(K_0)$, it follows that

$$P(\hat{\theta}_n \in K_0) \rightarrow 1.$$

By the additional compact-localisation assumption,

$$P(\hat{\theta}_n \in K_0, \hat{\theta}_n^{\text{SMML}} \in K_1) \rightarrow 1.$$

On this event, (A7) gives

$$d_F(\hat{\theta}_n, \hat{\theta}_n^{\text{SMML}}) = O_p(n^{-1/2}).$$

By (A5), the Fisher–Rao and Euclidean metrics are locally equivalent on K_1 . Therefore,

$$\|\hat{\theta}_n^{\text{SMML}} - \hat{\theta}_n\| = O_p(n^{-1/2}).$$

Hence,

$$\|\hat{\theta}_n^{\text{SMML}} - \theta_0\| \leq \|\hat{\theta}_n^{\text{SMML}} - \hat{\theta}_n\| + \|\hat{\theta}_n - \theta_0\| = O_p(n^{-1/2}),$$

which proves (16). Since $O_p(n^{-1/2}) = o_p(1)$, consistency follows. $\square$

Remark 3 (Interpretation). *Theorem 3 shows that, under the high-resolution regime of (A7), replacing the continuous MLE by the selected SMML codepoint introduces only an $O_p(n^{-1/2})$ perturbation.*

5. SMML and Exponential Families

We now specialise the general SMML framework to regular exponential family models. In this setting, the entropy–cross-entropy structure of the SMML objective has an explicit information-geometric form. Exponential families are dually flat statistical manifolds. The Fisher information defines the Riemannian metric, natural and expectation parameters provide dual affine coordinates, and KL divergence is the Bregman divergence generated by the log-partition function [14]. Consequently, SMML codepoints and cells can be described in terms of KL projections, Bregman centroids, and affine Voronoi-type partitions.

The closest SMML-specific prior work is Dowty [10], who studied exponential families with continuous sufficient statistics and showed that SMML regions are convex polytopes determined by the assertions and coding probabilities. The results below give the corre-

sponding countable-data formulation and connect it to the asymptotic Fisher–Rao and KL-projection framework developed in the preceding sections.

We consider models whose joint likelihood for a data set $\mathbf{x}$ of size n has the form

$$p_n(\mathbf{x} \mid \boldsymbol{\theta}) = h(\mathbf{x}) \exp\{\eta(\boldsymbol{\theta})'T(\mathbf{x}) - A_n(\eta(\boldsymbol{\theta}))\}, \quad (17)$$

where $T(\mathbf{x}) \in \mathbb{R}^d$ is a sufficient statistic, $\eta(\boldsymbol{\theta}) \in \mathbb{R}^d$ is the natural parameter, and A_n is the log-partition function for the full sample. The model is canonical when $\eta(\boldsymbol{\theta}) = \boldsymbol{\theta}$.

5.1. SMML Codepoints as KL/Bregman Centroids

For a fixed SMML cell P_j , the associated codepoint satisfies

$$\boldsymbol{\theta}_j^* = \arg \max_{\boldsymbol{\theta} \in \Theta} \sum_{\mathbf{x} \in P_j} r_n(\mathbf{x}) \log p_n(\mathbf{x} \mid \boldsymbol{\theta}). \quad (18)$$

Substituting (17) and dropping the term involving $\log h(\mathbf{x})$, which does not depend on $\boldsymbol{\theta}$, gives

$$\boldsymbol{\theta}_j^* = \arg \max_{\boldsymbol{\theta} \in \Theta} \{\eta(\boldsymbol{\theta})'S_j - q_j A_n(\eta(\boldsymbol{\theta}))\}, \quad (19)$$

where

$$S_j = \sum_{\mathbf{x} \in P_j} r_n(\mathbf{x})T(\mathbf{x}), \quad q_j = \sum_{\mathbf{x} \in P_j} r_n(\mathbf{x}) > 0.$$

The first-order condition is

$$[D\eta(\boldsymbol{\theta}_j^*)]' \{S_j - q_j \nabla_{\eta} A_n(\eta(\boldsymbol{\theta}_j^*))\} = 0. \quad (20)$$

In the canonical case, this reduces to

$$\nabla A_n(\boldsymbol{\theta}_j^*) = \frac{S_j}{q_j}. \quad (21)$$

Since $\nabla A_n(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[T(\mathbf{X})]$, this is a moment-matching condition. The expectation parameter of the SMML codepoint equals the r_n -weighted average of the sufficient statistic over the cell.

Proposition 1 (SMML codepoints in canonical exponential families). *Suppose that (17) is a regular canonical exponential family, and suppose that the cellwise mean S_j/q_j lies in the interior of the convex support of the sufficient statistic T , so that the maximiser in (18) exists, is unique, and lies in Θ . Then, for each SMML cell P_j ,*

$$\mathbb{E}_{\boldsymbol{\theta}_j^*}[T(\mathbf{X})] = \sum_{\mathbf{x} \in P_j} w_j(\mathbf{x})T(\mathbf{x}), \quad w_j(\mathbf{x}) = \frac{r_n(\mathbf{x})}{q_j}. \quad (22)$$

Thus, the expectation parameter associated with the SMML codepoint is the cellwise r_n -weighted average of the sufficient statistic.

Proof. In the canonical case, $\eta(\boldsymbol{\theta}) = \boldsymbol{\theta}$, so the first-order condition in (20) reduces to

$$\nabla A_n(\boldsymbol{\theta}_j^*) = \frac{S_j}{q_j}.$$

Since

$$S_j = \sum_{\mathbf{x} \in P_j} r_n(\mathbf{x}) T(\mathbf{x}), \quad q_j = \sum_{\mathbf{x} \in P_j} r_n(\mathbf{x}),$$

we obtain

$$\frac{S_j}{q_j} = \sum_{\mathbf{x} \in P_j} \frac{r_n(\mathbf{x})}{q_j} T(\mathbf{x}) = \sum_{\mathbf{x} \in P_j} w_j(\mathbf{x}) T(\mathbf{x}).$$

Finally, for a regular canonical exponential family,

$$\nabla A_n(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[T(\mathbf{X})],$$

which yields (22). $\square$

The moment-matching condition in (21) is the standard characterisation of the KL (or I-) projection onto a regular exponential family [9]. It is also the countable-data analogue of Dowty's centroid condition for SMML estimators with continuous sufficient statistics [10]. Since KL divergence in a canonical exponential family is the Bregman divergence generated by A_n , Proposition 1 identifies SMML codepoints as KL/Bregman centroids in expectation coordinates, connecting the result to Bregman clustering theory [15]. Unlike standard Bregman clustering, however, the cells and weights here arise from the SMML coding criterion. In general, the natural parameter itself need not be an affine average of the corresponding MLEs. Such an interpretation holds only in special cases where the inverse mean map is affine and the MLE depends affinely on the sufficient statistic.

5.2. Polyhedral Structure of SMML Cells

The Fisher–Rao Voronoi geometry of Section 3 is an asymptotic local result for regular parametric models. In exponential families, an exact finite-sample geometry is available because the log-likelihood is affine in the sufficient statistic. The next result shows that exact SMML cells are pullbacks of convex polyhedra in sufficient-statistic space.

Theorem 4 (Polyhedral SMML cells in sufficient-statistic space). *Let $\mathcal{P}^* = \{P_1^*, \dots, P_k^*\}$ be an optimal SMML partition for the exponential family (17), with associated codepoints $\boldsymbol{\Theta}^* = \{\boldsymbol{\theta}_1^*, \dots, \boldsymbol{\theta}_k^*\}$ and assertion probabilities $q_1^*, \dots, q_k^*$. For each j , define*

$$\mathcal{V}_j = \bigcap_{\ell \neq j} \left\{ \mathbf{t} \in \mathbb{R}^d : (\boldsymbol{\eta}(\boldsymbol{\theta}_j^*) - \boldsymbol{\eta}(\boldsymbol{\theta}_\ell^*))' \mathbf{t} \geq \log q_\ell^* - \log q_j^* + A_n(\boldsymbol{\eta}(\boldsymbol{\theta}_j^*)) - A_n(\boldsymbol{\eta}(\boldsymbol{\theta}_\ell^*)) \right\}. \quad (23)$$

Then each $\mathcal{V}_j$ is a convex polyhedron in $\mathbb{R}^d$, and, up to ties on boundaries,

$$P_j^* = \{\mathbf{x} \in \mathcal{X}_n : T(\mathbf{x}) \in \mathcal{V}_j\}.$$

Proof. For a fixed codebook and assertion probabilities, assigning $\mathbf{x}$ to assertion j incurs the codelength

$$\Lambda_j(\mathbf{x}) = -\log q_j^* - \log p_n(\mathbf{x} | \boldsymbol{\theta}_j^*).$$

The optimal partition assigns $\mathbf{x}$ to an index minimising $\Lambda_j(\mathbf{x})$, up to ties. Thus, $\mathbf{x} \in P_j^*$ precisely when

$$\Lambda_j(\mathbf{x}) \leq \Lambda_\ell(\mathbf{x}) \quad \text{for all } \ell \neq j.$$

Substituting the exponential-family representation and cancelling the common base-measure term $\log h(\mathbf{x})$, this inequality becomes

$$(\boldsymbol{\eta}(\boldsymbol{\theta}_j^*) - \boldsymbol{\eta}(\boldsymbol{\theta}_\ell^*))' T(\mathbf{x}) \geq \log q_\ell^* - \log q_j^* + A_n(\boldsymbol{\eta}(\boldsymbol{\theta}_j^*)) - A_n(\boldsymbol{\eta}(\boldsymbol{\theta}_\ell^*)).$$

For fixed j and ℓ , this is a closed half-space in sufficient-statistic space. Intersecting these half-spaces over all $\ell \neq j$ gives $\mathcal{V}_j$. Hence, $\mathcal{V}_j$ is a convex polyhedron. Since the membership condition depends on $\mathbf{x}$ only through $T(\mathbf{x})$, the SMML cell is the pullback

$$P_j^* = \{\mathbf{x} \in \mathcal{X}_n : T(\mathbf{x}) \in \mathcal{V}_j\},$$

up to boundary ties. $\square$

Unlike Theorems 1–3, Theorem 4 is an exact finite-sample result for exponential families. It follows directly from affine comparisons of two-part codelengths in sufficient-statistic space and does not rely on the high-resolution Assumption (A7).

Figure 2 illustrates the exact exponential-family structure in Theorem 4 where the pairwise SMML codelength comparisons become affine boundaries in sufficient-statistic space, and the resulting data-space cells are pullbacks under the sufficient-statistic map. Theorem 4 is the countable-data analogue of Dowty’s polytope theorem for SMML estimators with continuous sufficient statistics [10]. In both settings, pairwise comparisons of two-part codelengths become affine inequalities in sufficient-statistic space; the assertion probabilities provide the offsets, while differences between natural parameters determine the normal vectors of the separating hyperplanes. This affine structure also agrees with the general geometry of Bregman Voronoi diagrams, where nearest-codepoint comparisons for Bregman divergences become affine inequalities in suitable dual coordinates [16].

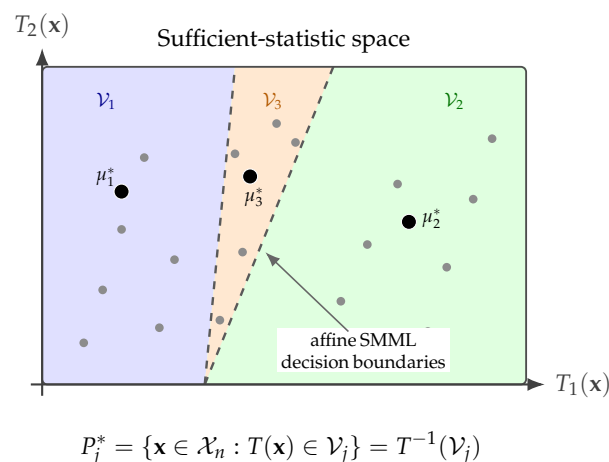


Figure 2. Exact SMML cells in sufficient-statistic space for an exponential family. Pairwise comparisons of two-part codelengths become affine inequalities in $T(\mathbf{x})$, so each region $\mathcal{V}_j$ is a convex polyhedron. The grey points represent attainable sufficient-statistic values, while the black points indicate the corresponding expectation-coordinate centroids μ_j^* . The SMML cell in data space is the pullback $P_j^* = T^{-1}(\mathcal{V}_j)$.

For multinomial models, the structure is especially transparent. In the binomial case, the sufficient statistic is one-dimensional, so each SMML cell is an interval of count values. For a K -category multinomial model, each pairwise SMML boundary is an affine hyperplane in the count vector, and each cell is obtained by intersecting the multinomial lattice with a finite collection of half-spaces. For a fixed finite codebook, this is a weighted analogue of the logarithmic Voronoi partitions of Alexandr and Heaton [17].

5.3. Example: Poisson Model

The Poisson model is a one-dimensional example of both the centroid condition in Proposition 1 and the exact affine cell structure in Theorem 4. Let $X_1, \dots, X_n$ be independent Poisson observations with common mean $\lambda > 0$, and write

$$T(\mathbf{x}) = \sum_{i=1}^n x_i$$

for the sufficient statistic of the sample $\mathbf{x} = (x_1, \dots, x_n)$. The joint likelihood is

$$p_n(\mathbf{x} \mid \lambda) = \left(\prod_{i=1}^n \frac{1}{x_i!} \right) \exp\{T(\mathbf{x}) \log \lambda - n\lambda\},$$

so the model is a regular canonical exponential family with natural parameter

$$\eta = \log \lambda,$$

sufficient statistic $T(\mathbf{x})$, and log-partition function

$$A_n(\eta) = ne^\eta = n\lambda.$$

For a fixed SMML cell P_j , Proposition 1 gives the moment-matching condition

$$E_{\lambda_j^*}[T(\mathbf{X})] = \sum_{\mathbf{x} \in P_j} w_j(\mathbf{x}) T(\mathbf{x}), \quad w_j(\mathbf{x}) = \frac{r_n(\mathbf{x})}{q_j}.$$

Since $E_\lambda[T(\mathbf{X})] = n\lambda$, this becomes

$$n\lambda_j^* = E_{r_n}[T(\mathbf{X}) \mid \mathbf{X} \in P_j],$$

and hence

$$\lambda_j^* = \frac{1}{n} E_{r_n}[T(\mathbf{X}) \mid \mathbf{X} \in P_j].$$

Thus, the Poisson SMML codepoint is the marginal-predictive conditional expectation of the sample mean over the cell, which is the Poisson instance of the KL/Bregman centroid condition in Proposition 1.

The exact SMML partition is equally explicit. For a fixed codebook $\{\lambda_1, \dots, \lambda_k\}$ and assertion probabilities $\{q_1, \dots, q_k\}$, the two-part codelength for assigning sample $\mathbf{x}$ to assertion j is

$$\Lambda_j(\mathbf{x}) = -\log q_j - \log p_n(\mathbf{x} \mid \lambda_j).$$

Comparing assertions j and ℓ , and cancelling the common base measure term, gives

$$(\log \lambda_j - \log \lambda_\ell) T(\mathbf{x}) \geq \log q_\ell - \log q_j + n(\lambda_j - \lambda_\ell).$$

Hence, each exact SMML cell is the pullback of an interval in the sufficient statistic $T(\mathbf{x})$, and the pairwise threshold is

$$t_{j\ell} = \frac{\log q_\ell - \log q_j + n(\lambda_j - \lambda_\ell)}{\log \lambda_j - \log \lambda_\ell}, \quad \lambda_j \neq \lambda_\ell,$$

where the inequality defining membership of cell j is $T(\mathbf{x}) \geq t_{j\ell}$ when $\lambda_j > \lambda_\ell$ and $T(\mathbf{x}) \leq t_{j\ell}$ when $\lambda_j < \lambda_\ell$. In particular, after ordering the assertions by their Poisson means, the exact SMML partition is obtained by intersecting the integer lattice of attainable

sufficient-statistic values with a finite collection of threshold intervals. This is the one-dimensional Poisson instance in Theorem 4.

The asymptotic Fisher–Rao geometry of Section 3 is also especially simple in this example. For one Poisson observation,

$$J_1(\lambda) = \frac{1}{\lambda},$$

so the per-observation Fisher–Rao metric is

$$ds^2 = \frac{d\lambda^2}{\lambda}.$$

The corresponding Fisher–Rao distance is [18]

$$d_F(\lambda_1, \lambda_2) = 2|\sqrt{\lambda_1} - \sqrt{\lambda_2}|.$$

Thus, in the high-resolution regime of Theorem 1, local SMML assignment compares nearby Poisson assertions using

$$4(\sqrt{\lambda} - \sqrt{\lambda_j})^2 - \frac{2}{n} \log q_j,$$

showing explicitly how assertion probabilities shift the Voronoi boundaries. The finite-sample threshold structure in the sufficient statistic $T(\mathbf{x})$ and the asymptotic weighted Voronoi structure in λ -space are therefore consistent descriptions of the same underlying SMML geometry at different levels of approximation.

6. Discussion

This paper has developed an information-theoretic and geometric perspective on strict minimum message length (SMML) estimation in regular parametric models. The central insight is that SMML is an entropy–cross-entropy optimisation principle that produces a finite-resolution approximation of a continuous statistical model. In the high-resolution regime, this approximation acquires a local geometry governed by Kullback–Leibler (KL) divergence and the Fisher information.

At the level of partitions, the SMML decision rule is asymptotically equivalent to a weighted Fisher–Rao Voronoi rule in parameter space, pulled back through the maximum likelihood estimator (MLE). The assertion probabilities appear as additive weights, reflecting the contribution of assertion entropy to the two-part codelength. At the level of codepoints, each SMML assertion is the KL projection of the normalised cellwise predictive distribution onto the model family. Thus, SMML simultaneously quantises the parameter space and performs local KL-optimal fitting within each quantisation cell. The SMML coding constraint preserves consistency and the usual $n^{-1/2}$ rate. It need not preserve the exact first-order limiting distribution unless the quantisation error is $o_p(n^{-1/2})$.

For regular exponential families, the geometry becomes exact. In canonical form, SMML codepoints satisfy a moment-matching condition whereby the expectation parameter of the assertion equals the marginal-predictive weighted average of the sufficient statistic in the cell. Equivalently, SMML codepoints are KL/Bregman centroids, connecting the codepoint condition to Bregman centroid and clustering theory [15]. Pairwise code-length comparisons also yield affine inequalities in sufficient-statistic space, so that the exact SMML partition is the pullback of a polyhedral partition. This is the countable-data analogue of Dowty’s polytope theorem for SMML estimators with continuous sufficient statistics [10], and is consistent with the affine structure of Bregman Voronoi diagrams [16].

In the SMML setting, the assertion probabilities provide the affine offsets in these Voronoi-type boundaries.

The computational difficulty of exact SMML optimisation also suggests a role for geometric approximations. Farr and Wallace [19] proved that SMML inference is NP-hard via a reduction from Exact Cover by 3-Sets (X3C), although special one-dimensional cases admit efficient algorithms. The Fisher–Rao and Bregman Voronoi structures identified here may therefore provide useful guidance for practical SMML codebook-construction procedures. Theorem 1 implies that, in the high-resolution regime, the assignment step is approximately a weighted Fisher–Rao Voronoi rule. Theorem 2 further suggests a corresponding update step, in which each codepoint is replaced by the weighted average of the MLEs assigned to its cell, or, more generally, by the fixed-cell KL projection of the conditional marginal distribution. This leads naturally to alternating schemes for approximate SMML optimisation. In exponential families, Proposition 1 and Theorem 4 show that assignment reduces to threshold or affine half-space comparisons in sufficient-statistic space, while codepoint updates reduce to moment matching. Such procedures do not remove the global NP-hardness of exact SMML optimisation, but they provide a practical way to exploit the induced Fisher–Rao or Bregman geometry when constructing approximate SMML codebooks.

A limitation of the present analysis is that the high-resolution LAN-scale regime is assumed rather than derived from the global SMML optimisation problem. The results of Sections 3 and 4 therefore describe the local geometry of SMML once the finite assertion codebook resolves the model at the LAN $n^{-1/2}$ scale. A local split-optimality heuristic supports this scale, since the entropy cost of splitting a cell of mass q_j is of order q_j , whereas the local quadratic reduction in detail codelength is of order $q_j n r_j^2$ for a cell of effective Fisher–Rao radius r_j . Establishing a rigorous global derivation of this regime for exact SMML optima remains an important direction for future work.

Overall, the results show that SMML provides an information-geometric bridge between statistical inference, quantisation, and compression. Starting from an entropy-based coding criterion, it induces KL projections, Bregman centroids, and Voronoi-type geometry in both asymptotic and exponential-family settings. In practical terms, the results suggest geometry-aware approximate SMML procedures for finite codebook design, model discretisation, and finite-resolution statistical compression in regular parametric families.

Author Contributions: Conceptualization, E.M.; methodology, E.M. and D.F.S.; formal analysis, E.M. and D.F.S.; investigation, E.M. and D.F.S.; writing—original draft preparation, E.M.; writing—review and editing, E.M. and D.F.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: Generative AI tools (Claude Opus 4.5; Microsoft Copilot GPT-5.5) were used in the preparation of this manuscript for generation and exploration of ideas as well as language improvement. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wallace, C.S.; Boulton, D.M. An information measure for classification. *Comput. J.* **1968**, *11*, 185–194. [[CrossRef](#)]
2. Wallace, C.S.; Boulton, D.M. An invariant Bayes method for point estimation. *Classif. Soc. Bull.* **1975**, *3*, 11–34. [[CrossRef](#)]
3. Wallace, C.S.; Freeman, P.R. Estimation and inference by compact coding. *J. R. Stat. Soc. (Ser. B)* **1987**, *49*, 240–252. [[CrossRef](#)]

4. Wallace, C.S. False oracles and SMML estimators. In *Proceedings of the International Conference on Information, Statistics and Induction in Science*; World Scientific: Singapore, 1996; pp. 304–316.
5. Wallace, C.S.; Dowe, D.L. Minimum Message Length and Kolmogorov Complexity. *Comput. J.* **1999**, *42*, 270–283. [[CrossRef](#)]
6. Wallace, C.S. *Statistical and Inductive Inference by Minimum Message Length*, 1st ed.; Information Science and Statistics; Springer: New York, NY, USA, 2005. [[CrossRef](#)]
7. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [[CrossRef](#)]
8. Wallace, C.S.; Dowe, D.L. Refinements of MDL and MML Coding. *Comput. J.* **1999**, *42*, 330–337. [[CrossRef](#)]
9. Csiszar, I. I-Divergence Geometry of Probability Distributions and Minimization Problems. *Ann. Probab.* **1975**, *3*, 146–158. [[CrossRef](#)]
10. Dowty, J.G. SMML estimators for exponential families with continuous sufficient statistics. *arXiv* **2013**, arXiv:1302.0581. [[CrossRef](#)]
11. Balasubramanian, V. Statistical Inference, Occam’s Razor, and Statistical Mechanics on the Space of Probability Distributions. *Neural Comput.* **1997**, *9*, 349–368. [[CrossRef](#)]
12. Myung, I.J.; Balasubramanian, V.; Pitt, M.A. Counting probability distributions: Differential geometry and model selection. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 11170–11175. [[CrossRef](#)] [[PubMed](#)]
13. Vaart, A.W.v.d. *Asymptotic Statistics*; Cambridge University Press: Cambridge, UK, 1998. [[CrossRef](#)]
14. Amari, S.i. *Information Geometry and Its Applications*; Springer: Tokyo, Japan, 2016. [[CrossRef](#)]
15. Banerjee, A.; Merugu, S.; Dhillon, I.S.; Ghosh, J. Clustering with Bregman Divergences. *J. Mach. Learn. Res.* **2005**, *6*, 1705–1749.
16. Boissonnat, J.D.; Nielsen, F.; Nock, R. Bregman Voronoi Diagrams. *Discret. Comput. Geom.* **2010**, *44*, 281–307. [[CrossRef](#)]
17. Alexandr, Y.; Heaton, A. Logarithmic Voronoi cells. *Algebr. Stat.* **2021**, *12*, 75–95. [[CrossRef](#)]
18. Miyamoto, H.K.; Meneghetti, F.C.C.; Pinele, J.; Costa, S.I.R. On Closed-Form Expressions for the Fisher–Rao Distance. *Inf. Geom.* **2024**, *7*, 311–354. [[CrossRef](#)]
19. Farr, G.E.; Wallace, C.S. The complexity of Strict Minimum Message Length inference. *Comput. J.* **2002**, *45*, 285–292. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.