information

Article

Principle of Information Increase: An Operational Perspective on Information Gain in the Foundations of Quantum Theory

Yang Yu * and Philip Goyal *

Department of Physics, University at Albany (SUNY), Albany, NY 12222, USA
* Correspondence: yyu9@albany.edu (Y.Y.); pgoyal@albany.edu (P.G.)

Abstract: A measurement performed on a quantum system is an act of gaining information about its state. However, in the foundations of quantum theory, the concept of information is multiply defined, particularly in the area of quantum reconstruction, and its conceptual foundations remain surprisingly under-explored. In this paper, we investigate the gain of information in quantum measurements from an operational viewpoint in the special case of a two-outcome probabilistic source. We show that the continuous extension of the Shannon entropy naturally admits *two* distinct measures of information gain, *differential information gain* and *relative information gain*, and that these have radically different characteristics. In particular, while differential information gain can increase or decrease as additional data are acquired, relative information gain consistently grows and, moreover, exhibits asymptotic indifference to the data or choice of Bayesian prior. In order to make a principled choice between these measures, we articulate a *Principle of Information Increase*, which incorporates a proposal due to Summhammer that more data from measurements leads to more knowledge about the system, and also takes into consideration black swan events. This principle favours differential information gain as the more relevant metric and guides the selection of priors for these information measures. Finally, we show that, of the symmetric beta distribution priors, the Jeffreys binomial prior is the prior that ensures maximal *robustness* of information gain for the particular data sequence obtained in a run of experiments.

Keywords: information gain; Kullback–Leibler divergence; Bayesian analysis; Jeffreys prior; foundation of quantum theory

check for updates

Citation: Yu, Y.; Goyal, P. Principle of Information Increase: An Operational Perspective on Information Gain in the Foundations of Quantum Theory. *Information* **2024**, *15*, 287. https://doi.org/10.3390/info15050287

Academic Editor: Gordana Dodig-Crnkovic

Received: 7 April 2024
Revised: 8 May 2024
Accepted: 8 May 2024
Published: 17 May 2024

1. Introduction

A measurement performed on a quantum system is an act of acquiring information about its state. This informational perspective on quantum measurement is widely embraced in practical applications such as quantum tomography [1–4], Bayesian experimental design [5], and informational analysis of experimental data [6,7]. It is also embraced in foundational research.

In particular, information assumes a central role in the quantum reconstruction program [8], which seeks to elucidate the fundamental physical origins of quantum theory by deriving its formalism from information-inspired postulates [9–17]. Nonetheless, in the foundational exploration of quantum theory, the concept of information is articulated and formalized in many different ways, which raises the question of whether there exists a more systematic basis for choosing how to formalize the concept of information within this domain.

In this paper, we scrutinize the notion of information from an operational standpoint and propose a physically intuitive postulate to determine the appropriate information gained from measurements.

In both tomographic applications and reconstruction of quantum theory, the focus often lies on probability distributions of physical parameters or quantities, which are updated based on the measurement results. In these contexts, the outcomes of a measurement performed on a quantum system are modelled as the interrogation of an n-outcome probabilistic source characterised by a set of parameters. For example, a given measurement on a given system can be described by a probability distribution $\Pr(x|D)$ of a quantity x, which is updated from a prior probability distribution given the results D obtained from a series of measurements performed on identical copies of a system. It is natural to consider using Shannon entropy to quantify the information gained from this updated distribution. However, Shannon entropy is limited to discrete distributions, whereas physical quantities and their associated probability distributions can be continuous.

The question thus arises: What is a suitable measure for quantifying the information obtained from real data, especially for quantities associated with continuous probability distributions?

One potential solution is to employ Kullback–Leibler (KL) divergence, also known as the relative entropy, $H(x|D) = \int \Pr(x|D) \ln \frac{\Pr(x|D)}{\Pr(x|I)} dx$, where $\Pr(x|I)$ represents the prior distribution of x, and $\Pr(x|D)$ represents the posterior distribution of x updated with the data D. This quantity is commonly referred to as the information gain from the prior distribution to the posterior distribution, and is widely used.

Since the KL divergence is non-negative and invariant under changes of coordinates, it appears to be a reasonable generalization of the Shannon entropy for continuous probability distributions. However, there are situations where *information gain* defined in terms of the KL divergence does not have a unique representation. Consider a scenario where one has acquired a series of data D, and one proceeds to take *additional* measurements, obtaining additional data D'. What is the additional information gain pertaining to D'? Using the KL divergence, there are two distinct ways to express the information related to this additional data. The first, to which we refer henceforth as the *differential information gain*, is simply the difference between the information gain from the combined dataset $\{D, D'\}$ and the information gain from D alone (see Figure 1). The second, which we refer to as the *relative information gain*, is given by the KL divergence of the posterior distribution after obtaining the complete dataset $\{D, D'\}$ compared to the posterior distribution after receiving data D alone (see Figure 2). These two measures of information gain exhibit notably different characteristics. For instance, whether the differential information gain increases or decreases when data D' is acquired depends on the choice of the prior distribution over the parameter, while the relative information gain consistently increases regardless of the choice of prior.

As we shall discuss in Section 2, both of these measures can be viewed as arising as a consequence of seeking to generalize the Shannon entropy to continuous probability distributions. In order to determine which of these options is most appropriate for our purposes, we seek a physically intuitive informational postulate to guide our selection. The first criterion comes from the intuitive notion proposed by Summhammer [18,19] that *more data from measurements leads to more knowledge about the system*. This idea has its origin in the observation that, as we conduct more measurements to determine the value of a physical quantity, the measurement uncertainty tends to decrease. In the following, we employ information theory to formalize and explore the plausibility of this idea. We find that relative information gain is consistently non-negative, whereas the positivity of differential information gain hinges on the choice of the prior distribution.

Contrary to Summhammer’s criterion, we argue that under certain circumstances, negative information gain due to acquisition of additional data D' is also meaningful. Take, for instance, the occurrence of a *black swan event*: an event so rare and unexpected that it significantly increases one’s uncertainty about the colour of swans. If the gain of information is considered to result from a reduction in the degree of uncertainty, the information gain associated with the observation of a black swan should indeed be negative. By combining this observation with Summhammer’s criterion, we are led to the *Principle of Information Increase:* the information gain from additional data should be positive asymptotically and negative in extreme cases. On the basis of the Principle of Information Increase, in the case of a two-outcome probabilistic source, we show that differential information gain is the more appropriate measure.

In addition, we formulate a new criterion, *the robustness of information gain*, for selecting priors to use with the differential information gain. The essential idea behind this criterion is as follows. If the result of the additional data D' is fixed, then the information gain due to D' will vary for different D. Robustness quantifies this difference in information gain across all possible data D. We show that for a two-outcome probabilistic source amongst the symmetric beta distributions, the Jeffreys binomial prior exhibits the highest level of robustness.

The quantification of knowledge gained from additional data is a topic that has received limited attention in the literature. In the realm of foundational research on quantum theory, this issue has been acknowledged but not extensively explored. Summhammer initially proposed the notion that "more data from measurements lead to more knowledge about the system" but did not employ information theory to address this problem, instead using changes in measurement uncertainty to quantify knowledge obtained in the asymptotic limit. This approach limits the applicability of the idea, as it excludes considerations pertaining to prior probability distributions and does not readily apply to finite data.

Wootters demonstrated the significance of the Jeffreys prior in the context of quantum systems from a different information-theoretical perspective [20]. In the domain of communication through quantum systems, the Jeffreys prior can maximize the information gained from measurements. Wootters approaches the issue from a more systematic perspective, utilizing mutual information to measure the information obtained from measurements. However, mutual information quantifies the *average information gain over all possible data sequences*, which is not suitable for addressing the specific scenario we discussed earlier, for which the focus is on the information gain from a fixed data sequence.

More broadly, the question of how much information is gained with the acquisition of additional data has been a relatively under-explored topic in both practical applications and foundational research on quantum theory. Commonly, mutual information is employed as a utility function. However, as noted above, mutual information essentially represents the expected information gain averaged over all possible data sequences. Consequently, it does not address the specific question of how much information is gained when a particular additional data point is obtained. From our perspective, this averaging process obscures essential edge effects, including black swan events, which, as we will discuss, serve as valuable guides for selecting appropriate information measures.

While our investigation primarily focuses on information gain in quantum systems, we conjecture that the principles and conclusions we draw can be extended to general probabilistic systems. Based on our analysis, we recommend quantification using differential information gain and the utilization of the Jeffreys multinomial prior. If one seeks to calculate the *expected* information gain in the next step, both the expected differential information gain and the expected relative information gain can be employed since, as we demonstrate for the two-outcome probabilistic case, they yield the same result.

The paper is organized as follows. In Section 2, we detail the two information gain measures, both of which have their origins in the generalization of Shannon entropy to continuous probability distributions. We will also examine Jaynes' approach to continuous entropy, which serves as the foundation for understanding these two information gain measures. Sections 3 and 4 focus on the numerical and asymptotic analysis of differential information gain and relative information gain for two-outcome probabilistic sources. Our primary emphasis is on how these measures behave under different prior distributions. We will explore black swan events, where the additional data D' are highly improbable given D. In this unique context, we will assess the physical meaningfulness of the two information gain measures. In Section 5, we will discuss expected information gain under the assumption that data D' from additional measurements have not yet been received. Despite the general differences between the two measures, it is intriguing to note that the two expected information gain measures are equal. Section 6 presents a comparison of the two information gain measures and the expected information gain. It is within this section that we propose the *Principle of Information Increase*, which crystallises the results of our analysis of the two measures of information gain. Finally, Section 7 explores the relationships between our work and other research in the field.

2. Continuous Entropy and Bayesian Information Gain

2.1. Entropy of Continuous Distribution

The Shannon entropy serves as a measure of uncertainty concerning a random variable before we have knowledge of its value. If we regard information as the absence of uncertainty, the Shannon entropy can also be used as a measure of information gained about a variable after acquiring knowledge about its value. However, it is important to note that Shannon entropy is applicable only to discrete random variables. To extend the concept of entropy to continuous variables, Shannon introduced the idea of differential entropy. Unlike Shannon entropy, differential entropy was not derived on an axiomatic basis. Moreover, it has a number of limitations.

First, the differential entropy can yield negative values, as exemplified by the differential entropy of a uniform distribution over the interval $[0, \frac{1}{2}]$, which equals $-\log 2$. Negative entropy, indicating a negative degree of uncertainty, lacks meaningful interpretation. Second, the differential entropy is coordinate-dependent [21], so that its value is not conserved under a change of variables. This implies that viewing the same data through different coordinate systems may result in the assignment of different degrees of uncertainty. Since the choice of coordinate systems is usually considered arbitrary, this coordinate-dependence also lacks a meaningful interpretation.

In an attempt to address the challenges associated with continuous entropy, Jaynes introduced a solution known as the limiting density of discrete points (LDDP) approach in his work [22]. In this approach, the probability density $p(x)$ of a random variable X is initially defined on a set of discrete points $x \in x_1, x_2, \cdots, x_n$. Jaynes proposed an invariant measure $m(x)$ such that, as the collection of points x_i becomes increasingly numerous, in the limit as $n \to \infty$,

$$\lim_{n\to\infty} \frac{1}{n} \,(\text{number of points in } a < x < b) = \int_a^b m(x)dx \tag{1}$$

With the help of $m(x)$, the entropy of X can then be represented as

$$H(X) = \lim_{n\to\infty} \log n - \int p(x) \log \frac{p(x)}{m(x)} dx \tag{2}$$

In this manner, the weaknesses associated with differential entropy appear to be resolved. This quantity remains invariant under changes of variables and is always non-negative. A similar approach is also discussed in [21]. However, two new issues arise. In Equation (2), $H(X)$ contains an infinite term, and the measure function $m(x)$ is unknown.

Regarding the infinite term, two potential solutions exist. The first option is to retain this infinite term and to reserve interpretation to the *difference* in the continuous entropy of two continuous distributions. The second solution is more straightforward: simply to omit the problematic $\log n$ term.

1. *Entropy of continuous distribution as a difference:*
   For example, when variable X is updated to X' due to certain actions, the *decrease* in entropy can be expressed as:

$$\Delta H(X \to X') \equiv H(X') - H(X) = \int p'(x) \log \frac{p'(x)}{m(x)} dx - \int p(x) \log \frac{p(x)}{m(x)} dx \tag{3}$$

   where $p'(x)$ represents the probability distribution of X'. Here, the two infinite terms cancel. The quantity ΔH quantifies the reduction in uncertainty about variable X resulting from these actions. This reduction in uncertainty can also be interpreted as an increase in information.

2. *Straightforward solution:*
   Jaynes directly discards the infinite term in Equation (2). For the sake of convenience, the minus sign is also dropped. This leads to the definition of Shannon–Jaynes information:
   $$H_{Jaynes}(X) = \int p(x) \log \frac{p(x)}{m(x)} dx \tag{4}$$
   This term quantifies the amount of information we possess regarding the outcome of X rather than the degree of uncertainty about X. H_{Jaynes} is equivalent to the KL divergence between the distributions $p(x)$ and $m(x)$.

In short, there are two ways to represent the entropy of a continuous distribution, with no obvious criterion to choose between them. In a special case where the variable X initially follows a distribution identical to the measure function, i.e., $p(x) = m(x)$, and X undergoes evolution to X' with distribution $p'(x)$, then we find that $\Delta H(X \to X') = H_{Jaynes}(X')$.

The remaining challenge lies in the selection of the measure function $m(x)$. When applying this concept of continuous entropy to the relationship between information theory and classical statistical physics, Jaynes opted for a uniform measure over phase space [22]. However, there is no established criterion for the choice of the measure function in any given application. We note that this measure function is analogous to the prior distribution in the context of Bayesian probability, with which it is often identified, which then leads to the well-known challenge of prior selection in Bayesian data analysis.

2.2. Bayesian Information Gain

In a coin-tossing model, let p denote the probability of getting a head in a single toss, and let N be the total number of tosses. After N tosses, the outcomes of these N tosses can be represented by an N-tuple, denoted as $T_N = (t_1, t_2, \cdots, t_N)$, where each t_i represents the result of the ith toss, with t_i taking values in the set $\{\text{Head}, \text{Tail}\}$. Applying the Bayes rule, the posterior probability for the probability of getting a head is given by:

$$\Pr(p|N, T_N, I) = \frac{\Pr(T_N|N, p, I)\Pr(p|I)}{\int \Pr(T_N|N, p, I)\Pr(p|I)dp} \tag{5}$$

where $\Pr(p|I)$ represents the prior. The information gain after N tosses would be the KL divergence from the prior distribution to the posterior distribution:

$$I(N) = D_{\text{KL}}(\Pr(p|N, T_N, I)||\Pr(p|I)) = \int_0^1 \Pr(p|N, T_N, I) \ln \frac{\Pr(p|N, T_N, I)}{\Pr(p|I)} dp \tag{6}$$

Based on the earlier discussion on continuous entropy, this quantity can be interpreted in two ways, either as the difference between the information gain after N tosses and the information gain without any tosses or as the KL divergence from the posterior distribution to the prior distribution.

When considering the information gain of additional tosses based on the results of the previous N tosses, we may observe two different approaches to represent this quantity.

Let t_{N+1} represent the outcome of the $(N+1)$th toss, and let $T_{N+1} = (t_1, t_2, \ldots, t_N, t_{N+1})$ denote the combined outcomes of the first N tosses and the $(N+1)$th toss. The posterior distribution after these $N+1$ tosses is given by:

$$\Pr(p|N+1, T_{N+1}, I) = \frac{\Pr(T_{N+1}|N+1, p, I)\Pr(p|I)}{\int \Pr(T_{N+1}|N+1, p, I)\Pr(p|I)dp} \tag{7}$$

When considering information gain as a difference between two quantities, the first form of information gain for this single toss t_{N+1} can be expressed as:

$$I_{\text{diff}} = D_{\text{KL}}(\Pr(p|N+1, T_{N+1}, I)||\Pr(p|I)) - D_{\text{KL}}(\Pr(p|N, T_N, I)||\Pr(p|I)) \tag{8}$$

In this expression, the first term $H(\Pr(p|N+1, t_{N+1}, I)||\Pr(p|I))$ represents the information gain from 0 tosses to $N+1$ tosses, while the second term $H(\Pr(p|N, T_N, I)||\Pr(p|I))$ represents the information gain from 0 tosses to N tosses. The difference between these terms quantifies the information gain in the single $(N+1)$th toss (see Figure 1). In this context, we can refer to I_{diff} as the *differential information gain in a single toss.*

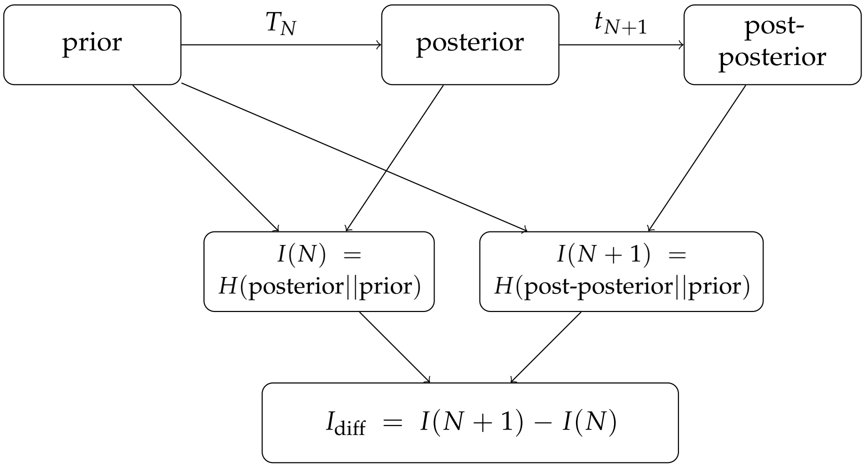

prior
T_N
posterior
t_{N+1}
post-posterior
$I(N) = H(\text{posterior}||\text{prior})$
$I(N+1) = H(\text{post-posterior}||\text{prior})$
$I_{\text{diff}} = I(N+1) - I(N)$

Figure 1. *Differential information gain in a single toss.* Assuming we have data from the first N tosses, denoted as T_N. Using a specific prior distribution, we can calculate the information gain for these first N tosses, denoted as $I(N)$. If we now consider the $(N+1)$th toss and obtain the result t_{N+1}, we can repeat the same procedure to calculate the information gain for a total of $N+1$ tosses, denoted as $I(N+1)$. The information gain specific to the $(N+1)$th toss can be obtained as the difference between $I(N+1)$ and $I(N)$.

Alternatively, we directly calculate the information gain from the Nth toss to the $(N+1)$th toss. Hence, the second form of information gain is defined as follows:

$$I_{\text{rel}} = D_{\text{KL}}(\Pr(p|N+1, T_{N+1}, I)||\Pr(p|N, T_N, I)), \tag{9}$$

which is simply the KL divergence from the posterior distribution after N tosses to the posterior distribution after $N+1$ tosses (see Figure 2). We refer to I_{rel} as the *relative information gain in a single toss.*

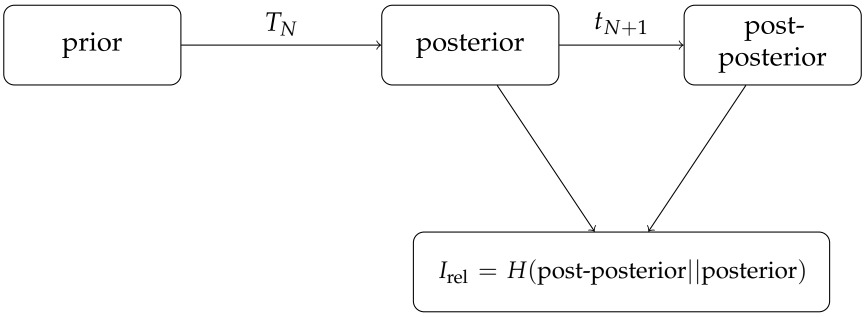

prior
T_N
posterior
t_{N+1}
post-posterior
$I_{\text{rel}} = H(\text{post-posterior}||\text{posterior})$

Figure 2. *Relative information gain in a single toss*: The posterior distribution calculated from the results of the first N tosses serves as the prior for the $(N+1)$th toss. The KL divergence between this posterior and the subsequent posterior represents the information gain in the $(N+1)$th toss.

In general, these two quantities, I_{diff} and I_{rel}, are not the same unless $N=0$, which implies that no measurements have been performed. I_{diff} could take on negative values, while I_{rel} is always non-negative due to the properties of the KL divergence. (This non-negativity is a consequence of Jensen's inequality applied to the convex logarithmic function, ensuring that the expected logarithmic difference between two probability distributions, which constitutes the KL divergence, cannot be negative.) Although KL divergence is not a proper distance metric between probability distributions (as it does not satisfy the triangle inequality), it is a valuable tool for illustrating the analogy of displacement and distance in a random walk model. (In a random walk, the change in total distance after $N+1$ steps compared to after N steps could be either positive or negative, analogous to how I_{diff} can have positive or negative values. On the other hand, the net displacement between the positions at step N and step $N+1$ represents the absolute change in position, which is analogous to I_{rel} always having a non-negative value.) This analogy helps elucidate the subtle difference between the two types of information gain.

Our goal is to determine which information gain measure is a more suitable choice. To do so, we use Summhammer's aforementioned postulate—"more measurements lead to more knowledge about the physical system" [18,19]—as our point of departure. If we quantify "knowledge" in terms of information gain from data, this notion suggests that the information gain from additional data should be positive if it indeed contributes to our understanding. This consideration makes relative information gain seem an appealing choice, as it is always non-negative. However, the derivation of differential information gain also carries significance. This leads to the question of whether Summhammer's intuitive idea is sufficient, and if not, what can replace it. In the following sections, we first will investigate differential information gain in both the finite N and asymptotic cases. We will explore the implications of negative values of differential information gain, particularly in extreme situations. We will then conduct numerical and asymptotic analyses of relative information gain. After analysing both measures of information gain, we will be better equipped to compare and establish connections between them and to assess the physical meaningfulness of Summhammer's proposal.

3. Differential Information Gain

3.1. Finite Number of Tosses

For the prior distribution, we employ the symmetric beta distribution, which serves as the conjugate prior for the binomial distribution:

$$\Pr(p|I) = \frac{p^{\alpha}(1-p)^{\alpha}}{B(\alpha+1,\alpha+1)} \tag{10}$$

where $\alpha > -1$, and $B(\cdot,\cdot)$ is the beta function.

In general, the beta distribution is characterized by two parameters. However, as the prior over p is invariably taken to be symmetric about $p = 1/2$ (which follows from the desideratum that the prior be invariant under outcome relabelling), we use a symmetric, single-parameter beta distribution. This distribution encompasses a wide spectrum of priors, including the uniform distribution (when $\alpha = 0$) and the Jeffreys binomial prior (when $\alpha = -0.5$).

The differential information gain of the $(N+1)$th toss is (see Appendix A)

$$\begin{aligned} I_{\text{diff}} = {} & \psi(h_N+\alpha+2) - \psi(N+2\alpha+3) \\ & + \frac{h_N}{h_N+\alpha+1} - \frac{N}{N+2\alpha+2} + \ln\frac{N+2\alpha+2}{h_N+\alpha+1} \end{aligned} \tag{11}$$

where ψ is the digamma function (the digamma function can be defined in terms of the gamma function: $\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$), and h_N is the number of heads in the first N tosses.

In this context, we assume that t_{N+1} = 'Head'. There is also a corresponding $I_{\text{diff}}(t_{N+1} = \text{'Tail'})$, but there is no loss of generality since we consider all possible values of T_N and since the expressions for both cases (Head and Tail) are symmetric.

I_{diff} is a function of h_N and α, and h_N ranges from 0 to N. In the following, we select a specific value for α and calculate all the $N+1$ values of I_{diff} for each N (see Figure 3).

3.1.1. Positivity of I_{diff}

Returning to our initial question—"Will more data lead to more knowledge?"—if we use the term "knowledge" to represent the differential information gain and use I_{diff} to quantify the information gained in each measurement, the question becomes rather straightforward: "Is I_{diff} always positive?"

In Figure 3, we present the results of numerical calculations for various values of N. Upon close examination of the graph, it becomes evident that I_{diff} is not always positive, except under specific conditions. In the following sections, we will investigate the conditions that lead to exceptions.

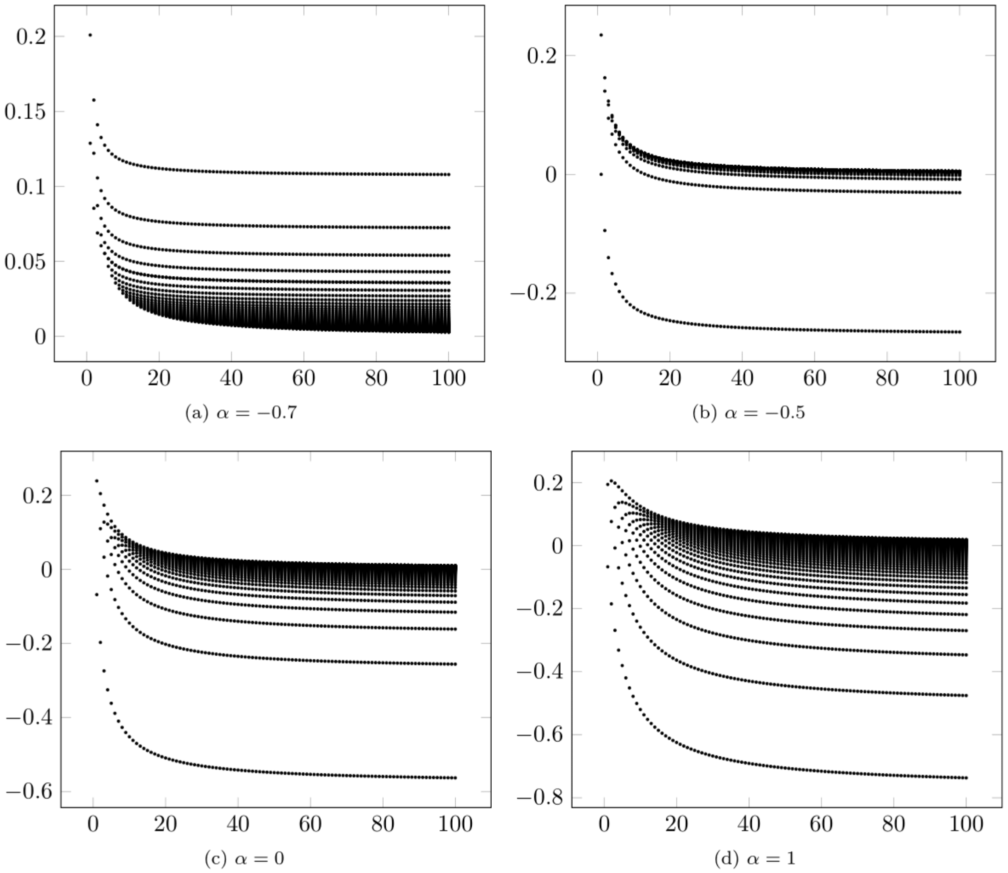


Figure 3. *Differential information gain (I_{diff}) vs. N for different priors.* Here, the y-axis represents the value of I_{diff}, and the x-axis corresponds to the value of N. In each graph, we fix the value of α to allow for a comparison of the behaviour of I_{diff} under different priors. Given N, there are $N+1$ points in the vertical direction as h_N ranges from 0 to N. Notably, for $\alpha = -0.7$, all points lie above the x-axis, while for other priors, negative points are present, and the fraction of negative points becomes constant as N increases. The asymptotic behaviour of this fraction is shown in Figure 4. Moreover, it appears that the graph is most concentrated when $\alpha = -0.5$, whereas for $\alpha < -0.5$ and $\alpha > -0.5$, the graph becomes more dispersed.

For certain priors, the differential information gain is consistently positive (Figure 3a), while for other priors, both positive and negative regions exist (Figure 3b–d). We note that for priors leading to negative regions, the lowest line exhibits greater dispersion compared to the other data lines. This lower line represents the scenario where the first N tosses all result in tails, but the $(N+1)$th toss yields a head. This situation is akin to a black swan event, and negative information gain in this extreme case holds significant meaning—if we have tossed a coin N times and obtaining all tails, we anticipate another tail in the next toss; hence, receipt of heads on the next toss raises the degree of uncertainty about the outcome of the next toss, leading to a reduction in information about the coin's bias.

3.1.2. Fraction of Negatives

In order to illustrate the variations in the positivity of information gain under different priors, we introduce a new quantity that we refer as to as the *Fraction of Negatives* (FoN), which represents the ratio of the number of h_N values that lead to negative I_{diff} and $N+1$. For instance, if, for a given α, $N = 10$ and $I_{\text{diff}} < 0$ when $h_N = 0, 1, 2, 3$, the FoN under this α and N is $\frac{4}{11}$.

From Figure 4, we identify a critical point, denoted as α_p, which is approximately -0.7. For any $\alpha \leq \alpha_p$, I_{diff} is guaranteed to be positive for all N and h_N values.

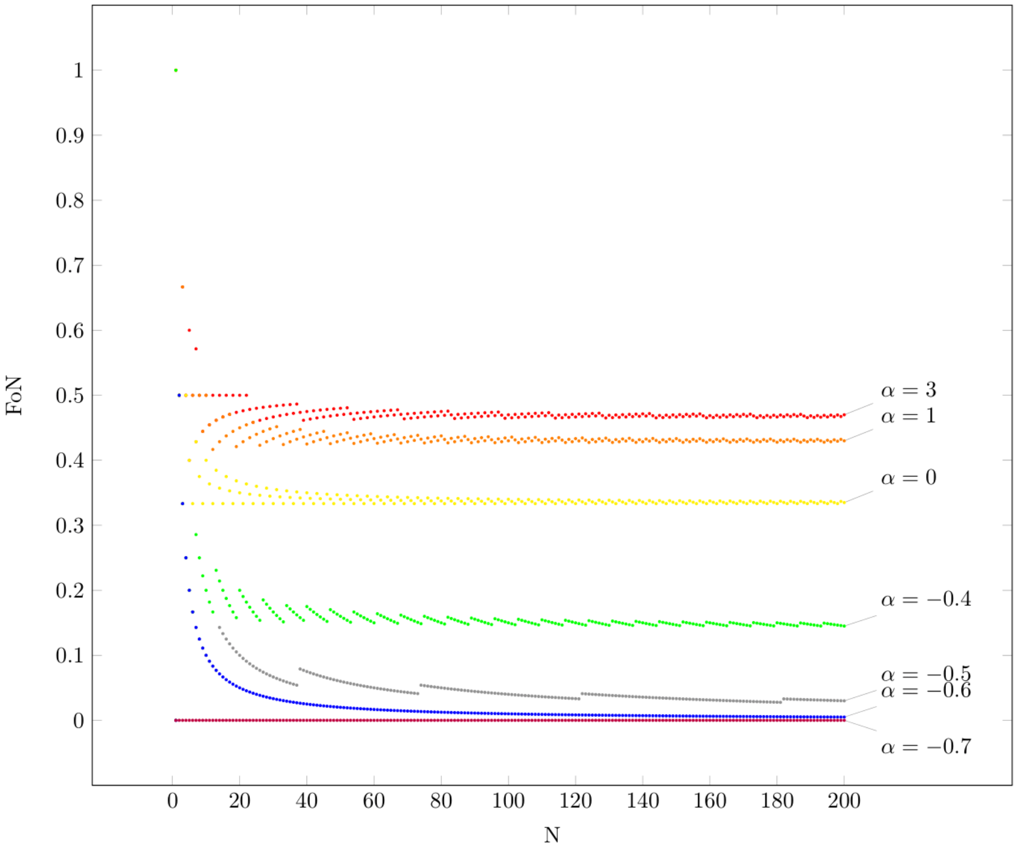


Figure 4. *Fraction of Negatives (FoN) vs. N under different values of α.* In Figure 3, we can observe that larger α values lead to more dispersed lines and an increased number of negative values for each N. We use FoN to quantify this fraction of negative points. It appears that for $\alpha \leq -0.7$, FoN is consistently zero, indicating that I_{diff} is always positive. For $\alpha \leq -0.5$ FoN decreases and tends to be zero as N becomes large, while for $\alpha > -0.5$, FoN tends to a constant as N increases, and this constant grows with increasing values of α.

If $\alpha > \alpha_p$, negative terms exist for some h_N; however, the patterns of these negative terms differ across various α values.

Additionally, we notice the presence of a turning point, $\alpha_0 = -0.5$. For $\alpha \leq \alpha_0$, FoN tends to zero as N increases, whereas for $\alpha > \alpha_0$, FoN approaches a constant as N grows.

A clearer representation of the critical point α_p and the turning point α_0 can be found in Figure 5, where the critical point α_p is approximately -0.68.

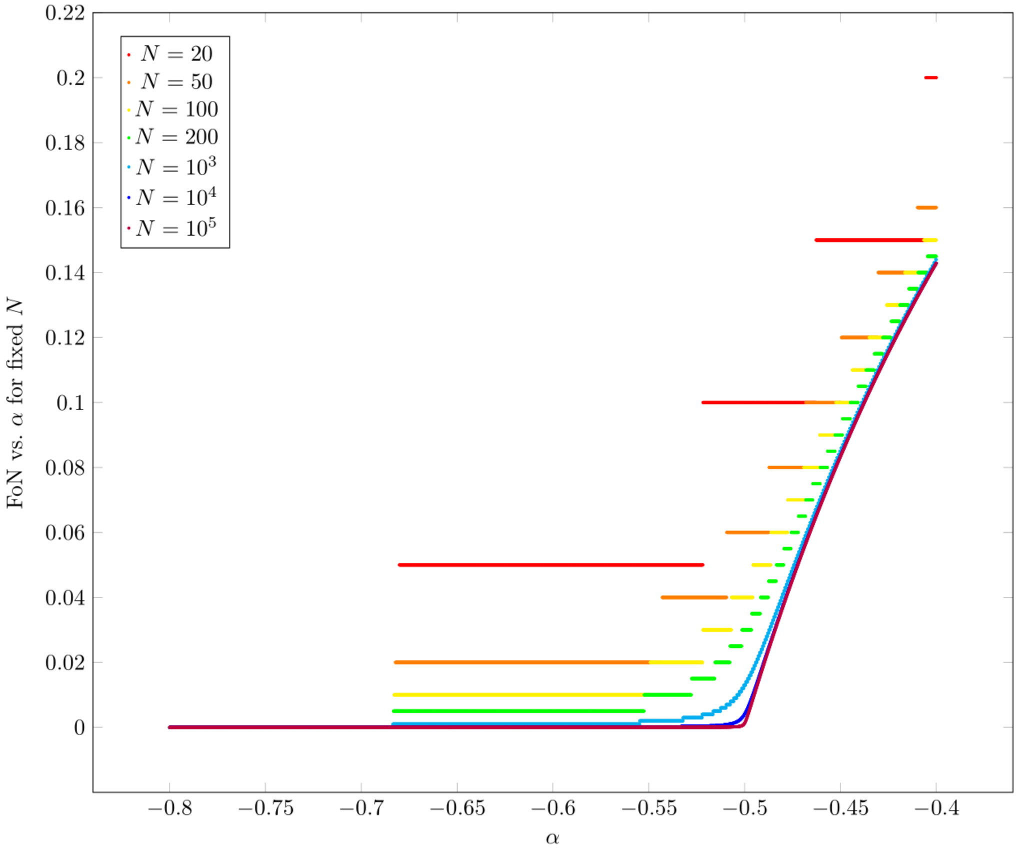


Figure 5. *Fraction of Negatives (FoN) vs. α for different values of N.* We identify a critical point, denoted as α_p, where the FoN equals zero when $\alpha \leq \alpha_p$. The critical point exhibits a gradual variation with respect to N following these patterns: (i) for small N, α_p is in close proximity to -0.68; (ii) for large N, α_p tends to -0.5.

3.1.3. Robustness of I_{diff}

In Figure 3, different priors not only exhibit varying degrees of positivity but also display varying degrees of variation in I_{diff} for different values of h_N; we refer to this as *divergence*. The divergence depends upon the choice of prior. To better understand this dependence, we quantify the dependence of I_{diff} on h_N by the standard deviation of I_{diff} across different values of h_N. Figure 6 illustrates how the standard deviation changes with respect to α while keeping N constant.

It is evident that when α is close to -0.5, the standard deviation is at its minimum. Reduced dependence of I_{diff} on h_N enhances its robustness against the effects of nature, as we attribute h_N to natural factors, while N is determined by human measurement choices. As N increases, the minimum point approaches -0.5. In the limit of large N, this minimum point will eventually converge to $\alpha = -\frac{1}{2}$, which means that under this specific choice of prior, I_{diff} depends minimally on h_N and primarily on N.

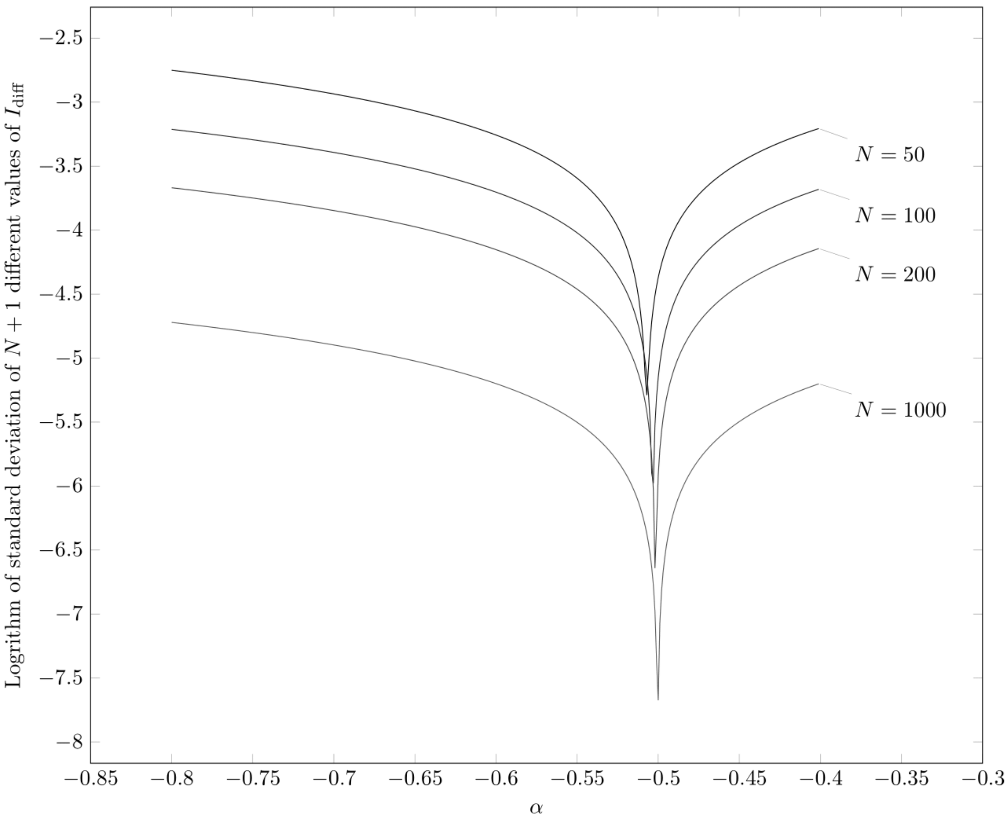


Figure 6. *Robustness of differential information gain (I_{diff}).* The y-axis represents the logarithm of the standard deviation of I_{diff} over all possible h_N values, while the x-axis depicts various selections of α. A smaller standard deviation indicates that different h_N values lead to the same result, implying greater independence of I_{diff} from h_N. This independence signifies the robustness of I_{diff} with respect to the natural variability in h_N, as we consider h_N to be solely determined by nature. The standard deviation, given a fixed N, is notably influenced by α, and there exists an α value at which the dependence on h_N is minimized. This particular α value approaches -0.5 as N increases.

3.2. Large N Approximation

Utilizing a recurrence relation and a large x approximation, the digamma function can be approximated as:

$$\psi(x) = \frac{1}{x-1} + \psi(x-1) \approx \frac{1}{x-1} + \ln(x-1) - \frac{1}{2(x-1)} = \frac{1}{2(x-1)} + \ln(x-1) \quad (12)$$

As a result, the large N approximation for the differential information gain in Equation (11) becomes:

$$I_{\text{diff}} = \frac{2h_N + 1}{2(h_N + \alpha + 1)} - \frac{2N + 1}{2(N + 2\alpha + 2)} \quad (13)$$

Using this approximation, when $\alpha = -\frac{1}{2}$, $I_{\text{diff}} = \frac{1}{2(N+1)}$, which shows that I_{diff} solely depends on N. This finding aligns with Figure 3, which demonstrates that I_{diff} is most concentrated when $\alpha = -0.5$ and is also consistent with the results of [23].

In Figure 4, we observe that the FoN tends to become constant for very large values of N. These constants can be estimated using the large N approximation of I_{diff} in Equation (13) (see Table 1). If $I_{\text{diff}} \leq 0$, then

$$h_N \leq \frac{2N\alpha + N + \alpha + 1}{4\alpha + 3}, \quad (14)$$

and we obtain:

$$\text{FoN} = \frac{1}{N+1}\frac{2N\alpha + N + \alpha + 1}{4\alpha + 3} \approx \frac{2\alpha + 1}{4\alpha + 3} \tag{15}$$

This equation aligns with the asymptotic lines in Figure 4, providing support for the observation mentioned in Figure 3: namely, that for $\alpha = -0.7$, all points lie above the x-axis, while for other priors, negative points are present, and the fraction of negative points becomes constant.

Table 1. Fraction of Negatives (FoN) under selected priors. A comparison between numerical results and asymptotic results demonstrates their agreement.

α	FoN (Numerical Result, $N = 1000$)	FoN (Asymptotic Result)	Discrepancy between the Two Results
−0.7	0	0	0
−0.6	0.001	0	0.1%
−0.5	0.013	0	1.3%
−0.4	0.144	0.143	0.1%
0	0.334	0.333	0.1%
1	0.429	0.429	0
3	0.467	0.467	0

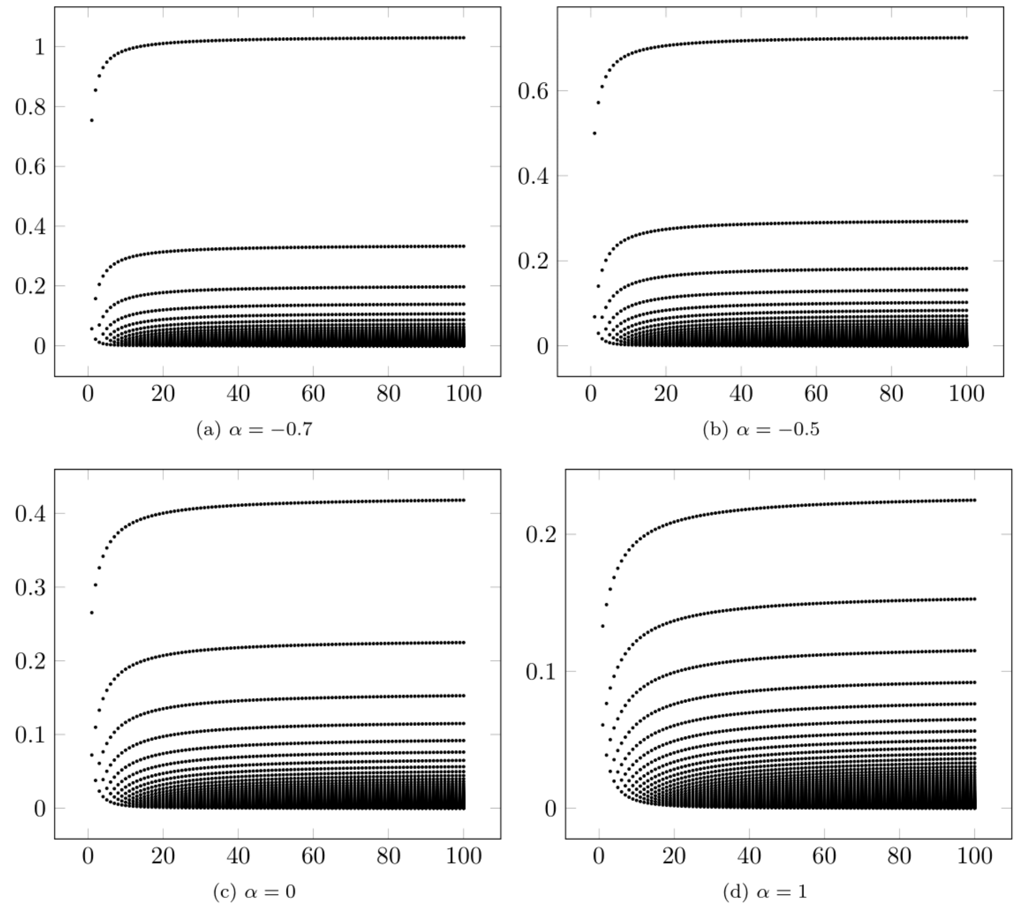
4. Relative Information Gain

The second form of information gain in a single toss is relative information gain, which represents the KL divergence from the posterior after N tosses to the posterior after $N+1$ tosses. We continue to use the one-parameter beta distribution prior in the form of Equation (10). The relative information gain is (see Appendix B):

$$I_{\text{rel}}(t_{N+1} = \text{‘Head’}) = \psi(h_N + \alpha + 2) - \psi(N + 2\alpha + 3) + \ln\frac{N + 2\alpha + 2}{h_N + \alpha + 1} \tag{16}$$

Relative information gain exhibits entirely different behaviour compared to differential information gain. Due to the properties of KL divergence, relative information gain is always non-negative, eliminating the need to consider negative values. We explore the dependence of relative information gain on priors and the interpretation of information gain in extreme cases.

In Figure 7, it becomes evident that, under different priors, the data lines exhibit similar shapes. This suggests that relative information gain is relatively insensitive to the choice of priors. On each graph, the top line represents the extreme case where the first N tosses result in tails and the $(N+1)$th toss results in a head. This line is notably separated from the other data lines, indicating that relative information gain behaves more like a measure of the degree of surprise associated with this additional data. In this black swan event, the posterior after $N+1$ tosses differs significantly from the posterior after N tosses.



(a) $\alpha = -0.7$

(b) $\alpha = -0.5$

(c) $\alpha = 0$

(d) $\alpha = 1$

Figure 7. *Relative information gain (I_{rel}) over different priors.* The y-axis represents the value of I_{rel}, while the x-axis represents N. For each N, there are $N+1$ different values of I_{rel}. It is important to note that I_{rel} is consistently positive across these selected priors. Similar to the differential information gain, each graph displays numerous divergent lines. However, the shape of these divergent lines remains remarkably consistent across varying values of α. The majority of these lines fall within the range of I_{rel} between 0 and 0.2.

For small values of N, both the average value and the standard deviation of I_{rel} exhibit a clear monotonic relationship with α, meaning that larger values of α result in smaller average values and standard deviations. However, as N becomes large, all priors converge and become indistinguishable. Nonetheless, it is important to note that relative information gain remains heavily independent on the specific data sequences (h_N). Figure 8 illustrates how the standard deviation of I_{rel} under different priors converges to the same value as N increases.

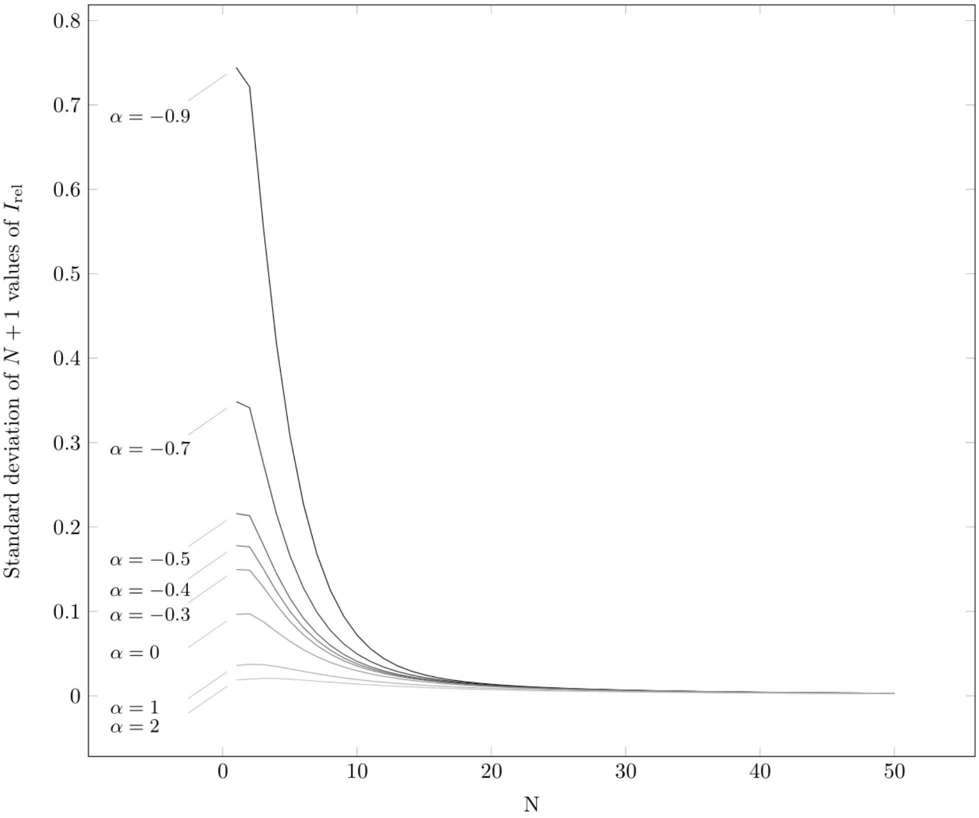


Figure 8. *Robustness of relative information gain (I_{rel}).* The y-axis represents the standard deviation of I_{rel} across all possible values of h_N. This demonstrates the substantial independence of I_{rel} from h_N. Additionally, as N increases, the standard deviations tend to approach zero for all priors.

By utilizing the aforementioned approximation of the digamma function, we obtain:

$$\begin{aligned} I_{\mathrm{rel}}(t_{N+1} = \text{`Head'}) &\approx \frac{1}{2(h_N + \alpha + 1)} - \frac{1}{2(N + 2\alpha + 2)} \\ &= \frac{N - h_N + \alpha + 1}{2(h_N + \alpha + 1)(N + 2\alpha + 2)} \end{aligned} \tag{17}$$

In the large N limit, I_{rel} becomes:

$$I_{\mathrm{rel}}(t_{N+1} = \text{`Head'}) \approx \frac{1}{2N}\left[\left(\frac{h_N}{N}\right)^{-1} - 1\right], \tag{18}$$

which is independent of α. Thus, it appears that the properties of relative information gain and differential information gain are complementary to each other. The differences between them are summarized in Table 2.

Table 2. Comparison of characteristics of two measures of information gain.

Information Gain Measure	**Asymptotic forms** **(t_{N+1} = 'Head')**	**Asymptotic Sensitivity to Prior**
Differential Information Gain	$I_{\text{diff}} \approx \frac{2h_N+1}{2(h_N+\alpha+1)} - \frac{2N+1}{2(N+2\alpha+2)}$	Heavily dependent upon prior. Independent of h_N for certain priors ($\alpha = -1/2$).
Relative Information Gain	$I_{\text{rel}} \approx \frac{1}{2(h_N+\alpha+1)} - \frac{1}{2(N+2\alpha+2)}$	Insensitive to prior. For large N, only affected by h_N.

5. Expected Information Gain

In this section, we discuss a new scenario: after N tosses but before the $(N+1)$th toss has been taken, can we predict how much information gain will occur in the next toss? The answer is affirmative, as discussed earlier.

After N tosses, we obtain a data sequence T_N with h_N heads. However, we can only estimate the probability p based on the posterior $\Pr(p|N, T_N, I)$. The expected value of p can be expressed as:

$$\langle p \rangle = \int_0^1 p \ \Pr(p|N, T_N, I) \, dp = \frac{h_N + \alpha + 1}{N + 2\alpha + 2} \tag{19}$$

Based on this expected value of p, we can calculate the average of the information gain in the $(N+1)$th toss. We define the expected differential information gain in the $(N+1)$th toss as:

$$\begin{aligned}
\overline{I_{\text{diff}}} &= \langle p \rangle \times I_{\text{diff}}(t_{N+1} = \text{'Head'}) + \langle 1-p \rangle \times I_{\text{diff}}(t_{N+1} = \text{'Tail'}) \\
&= \frac{h_N+\alpha+1}{N+2\alpha+2}\psi(h_N+\alpha+2) + \frac{N-h_N+\alpha+1}{N+2\alpha+2}\psi(N-h_N+\alpha+2) - \psi(N+2\alpha+3) \\
&\quad + \frac{h_N+\alpha+1}{N+2\alpha+2}\ln\frac{N+2\alpha+2}{h_N+\alpha+1} + \frac{N-h_N+\alpha+1}{N+2\alpha+2}\ln\frac{N+2\alpha+2}{N-h_N+\alpha+1}
\end{aligned} \tag{20}$$

$\overline{I_{\text{diff}}}$ represents the expected value of differential information gain in the $(N+1)$th toss. Similarly, we can define the expected relative information gain as:

$$\begin{aligned}
\overline{I_{\text{rel}}} &= \langle p \rangle \times I_{\text{rel}}(t_{N+1} = \text{'Head'}) + \langle 1-p \rangle \times I_{\text{rel}}(t_{N+1} = \text{'Tail'}) \\
&= \frac{h_N+\alpha+1}{N+2\alpha+2}\psi(h_N+\alpha+2) + \frac{N-h_N+\alpha+1}{N+2\alpha+2}\psi(N-h_N+\alpha+2) - \psi(N+2\alpha+3) \\
&\quad + \frac{h_N+\alpha+1}{N+2\alpha+2}\ln\frac{N+2\alpha+2}{h_N+\alpha+1} + \frac{N-h_N+\alpha+1}{N+2\alpha+2}\ln\frac{N+2\alpha+2}{N-h_N+\alpha+1}
\end{aligned} \tag{21}$$

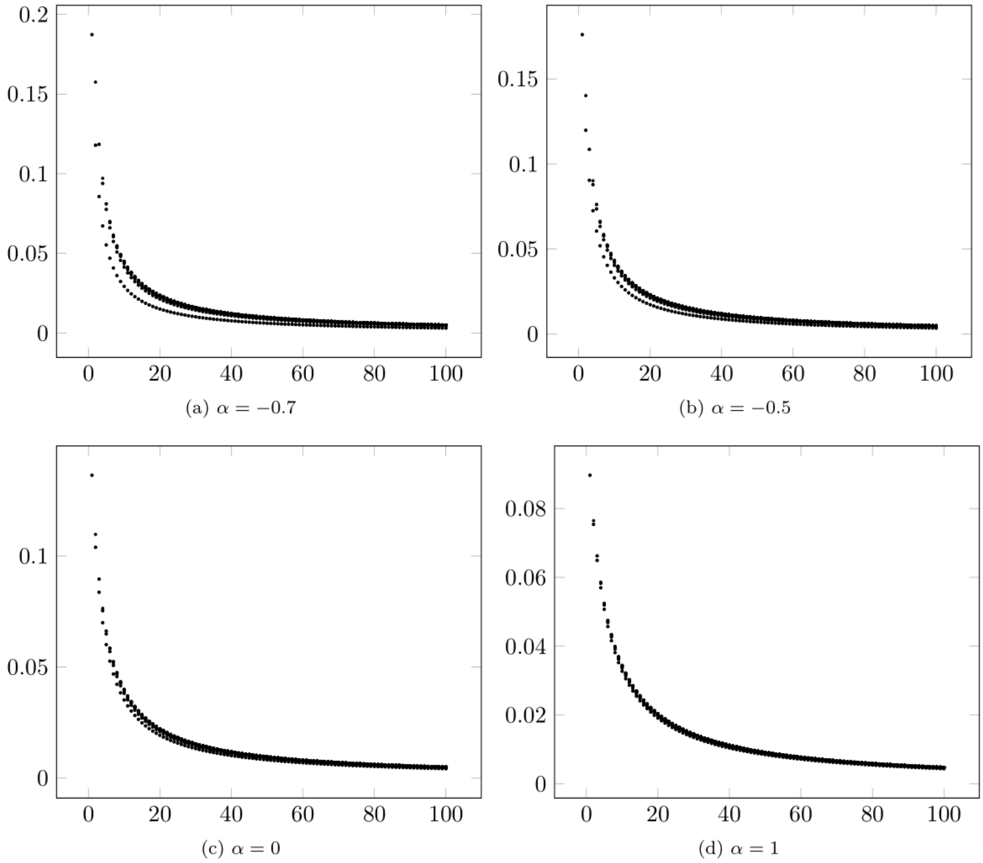
Surprisingly, $\overline{I_{\text{diff}}} = \overline{I_{\text{rel}}}$. This relationship holds true for any prior, not being limited to the beta distribution type prior, and furthermore holds for an arbitrary n-outcome probabilistic source. Please refer to Appendix C for a detailed proof. This suggests that there is only one choice for the expected information gain.

We first show the numerical results of expected information gain under different priors. It is evident that all data points are above the x-axis, indicating that the expected information gain is positive-definite, as anticipated. Since both I_{rel} and $\langle p \rangle$ are positive, it follows that $\overline{I_{\text{rel}}}$ must also be positive.

As with the discussions of differential information gain and relative information gain, we are also interested in examining the dependence of expected information gain on α or h_N. However, such dependence appears to be weak, as illustrated in Figures 9 and 10. Expected information gain demonstrates strong robustness concerning variations in α and h_N.

The asymptotic expression of expected information gain is

$$\overline{I_{\text{diff}}} = \overline{I_{\text{rel}}} = \frac{1}{2N} \tag{22}$$



(a) $\alpha = -0.7$

(b) $\alpha = -0.5$

(c) $\alpha = 0$

(d) $\alpha = 1$

Figure 9. *Expected information gain vs. N for fixed α.* The y-axis represents the value of expected information, while the x-axis represents the value of N. Notably, all expected information gain values are positive. The shapes of each graph exhibit remarkable similarity, with a limited number of divergent lines. As α increases, the number of divergent lines decreases.

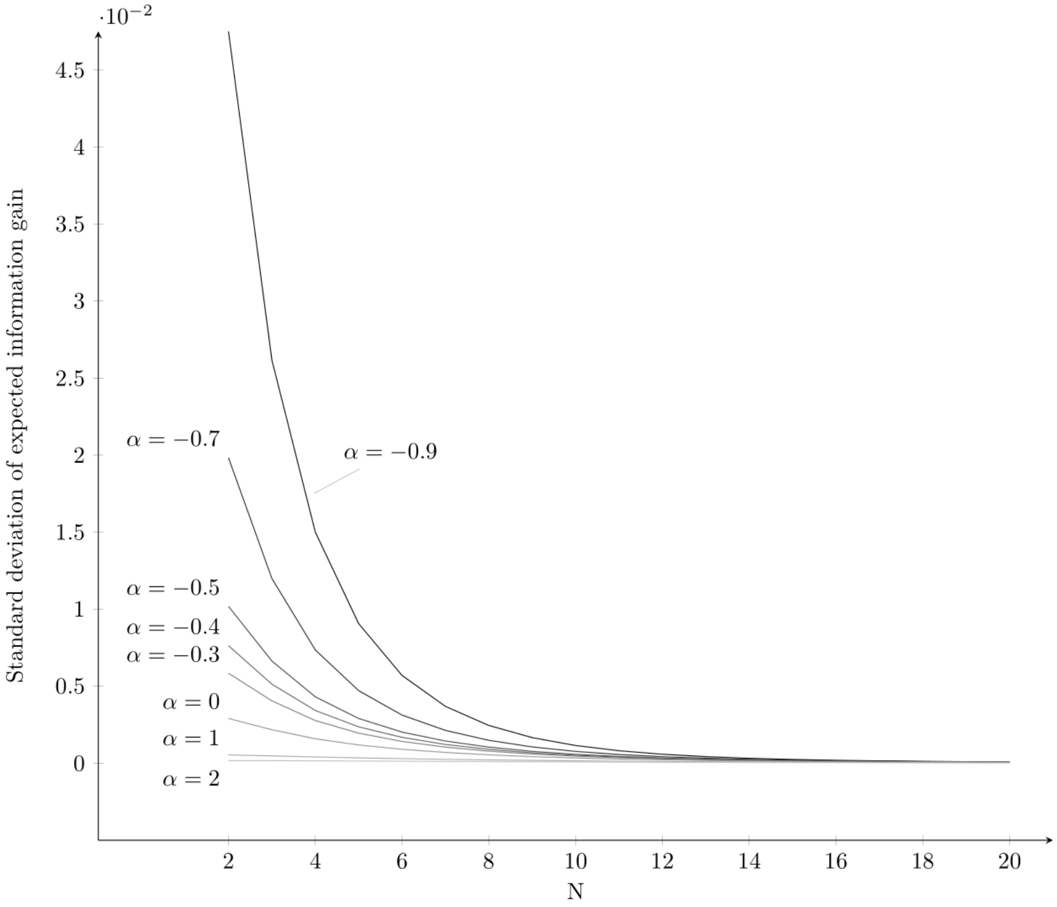


Figure 10. *Robustness of expected information gain.* The y-axis represents the standard deviation of the expected information gain over all possible values of h_N, while the x-axis represents the value of N. As N increases, and even for relatively small values of N, the standard deviation tends toward zero for all priors.

6. Comparison of Three Information Gain Measures, and the Information Increase Principle

From an operational perspective, the information measures we have considered can be categorized into two types: differential information gain and relative information gain pertain to a measurement that has *already been made*, while expected information gain pertains to a measurement that has *yet to be conducted.*

Regarding positivity, which is tied to the fundamental question of "Will acquiring more data from measurements lead to a deeper understanding of the system?": for relative information gain and expected information gain, the answer is affirmative, but differential information gain is positive only under certain specific prior conditions.

All three measures are functions of variables denoted as N, α, and h_N, which characterize the size of the data sequences, the prior information, and the existing data sequence, respectively. How sensitive are these measures to these parameters, particularly for large values of N? As we have shown, differential information gain is heavily influenced by all three parameters. It becomes nearly independent of h_N only when $\alpha = -0.5$. Relative information gain is not highly sensitive to the choice of priors. In the case of large values of N, relative information gain is affected by both h_N and N, whereas expected information gain depends solely on N. The comparison between them is summarized in Table 3.

Table 3. Comparison of three information gain measures.

Type of Information Gain	Positivity	Robustness about T_N
Differential	Strictly positive when $\alpha < \alpha_p$ where $\alpha_p \approx -0.68$. Asymptotically positive when $\alpha \leq -0.5$.	Robustness exists only when $\alpha = -0.5$ of beta distribution prior.
Relative	Strictly positive for all priors.	No significant differences of robustness among beta distribution priors.
Expected	Strictly positive for all priors.	No significant differences of robustness among beta distribution priors.

At first, one might have expected that the idea that *more data from measurements lead to more knowledge about the system* would hold strictly: namely, that the information gain from additional data would always be strictly positive. However, our perspective has been challenged by the observation of black swan events. In the extreme scenario where the first N tosses all result in tails and the $(N+1)$th toss yields a head, a negative information gain in this $(N+1)$th toss may be a more reasonable interpretation. To address this, we propose the

> **Principle of Information Increase:** *In a series of interrogations of an n-outcome probabilistic source, the information gain from additional data should tend towards positivity in the asymptotic limit. However, in the extreme case where the first N data points are identical and the data of the $(N+1)$th trial is contrary to the previous data, the information gain in this exceptional case should be negative.*

Applying this criterion, the choice of using the differential information gain becomes more appropriate for measuring the extent of knowledge contributed by additional data. For the beta distribution prior, it should be constrained within the range of approximately $-0.68 \lesssim \alpha \leq -0.5$. If we also consider the robustness of information gain under various given data scenarios, then the Jeffreys binomial prior ($\alpha = -0.5$) emerges as the most favourable choice.

7. Related Work

7.1. Information Increase Principle and the Jeffreys Binomial Prior

In [18,19], Summhammer introduces the idea that *more measurements lead to more knowledge about a physical quantity* and quantifies the level of knowledge regarding a quantity by assessing its uncertainty range after a series of repeated measurements. Quantified in this manner, the notion can be summarized as: "The uncertainty range of a physical quantity should decrease as the number of measurements increases." For a quantity θ, the uncertainty range $\Delta\theta$ is a function of the number of measurements:

$$\Delta\theta(N+1) < \Delta\theta(N) \tag{23}$$

If this quantity is determined by the probability of a two-outcome measurement, such as the probability of obtaining heads (p) in a coin toss, then there exists a relationship between the uncertainty range of θ and that of p,

$$\Delta\theta = \left|\frac{\partial\theta}{\partial p}\right|\Delta p \tag{24}$$

In large N approximation, $\Delta p = \sqrt{p(1-p)/N}$, so that

$$\Delta\theta = \left|\frac{\partial\theta}{\partial p}\right|\sqrt{p(1-p)/N}. \tag{25}$$

One intuitive way to ensure Equation (23) holds is by forcing $\Delta\theta$ to be purely a function of N. Observing the relationship between $\Delta\theta$ and Δp, the simplest solution would be to set $\Delta\theta = \frac{\text{const.}}{\sqrt{N}}$. Under this solution, the relationship between p and θ takes the following form:

$$\left|\frac{\partial\theta}{\partial p}\right|\sqrt{p(1-p)} = \text{const.}, \tag{26}$$

which yields Malus' law $p(\theta) = \cos^2(m(\theta-\theta_0)/2)$, with $m \in \mathbb{Z}$.

Summhammer does not employ information theory to quantify "knowledge about a physical quantity" but instead utilizes the statistical uncertainty associated with the quantity. However, viewed from the Bayesian perspective, if we assume that the prior distribution of the physical quantity, θ, is uniform, the difference between θ and p in Equation (25) implies that the prior distribution of the probability follows the Jeffreys binomial prior:

$$\Pr(p|I) = \left|\frac{\partial\theta}{\partial p}\right|\Pr(\theta|I) = \frac{1}{\pi}\frac{1}{\sqrt{p(1-p)}} \tag{27}$$

Thus, in the large N approximation, Summhammer's result can be interpreted to mean that the prior associated with the probability of a uniformly distributed physical quantity must adhere to the Jeffreys binomial prior.

Goyal [23] introduces an *asymptotic* Principle of Information Gain, which states that "In n interrogations of a N-outcome probabilistic source with an unknown probabilistic vector $\vec{P}$, the amount of Shannon–Jaynes information provided by the data about $\vec{P}$ remains independent of $\vec{P}$ for all $\vec{P}$ in the limit as $n \to \infty$." Goyal establishes the equivalence between this principle and the Jeffreys rule. Under his Principle of Information Gain, the Jeffreys multinomial prior is then derived. In the case of a two-outcome probabilistic model, the Jeffreys multinomial prior reduces to the Jeffreys binomial prior. Asymptotic analysis reveals that Shannon–Jaynes information is not only independent of the probability vector $\vec{P}$ but also monotonically increases with the number of interrogations. It is worth noting that Shannon–Jaynes information can be viewed as the accumulation of differential information gain. This asymptotic result aligns with our findings: under the Jeffreys binomial prior, the differential information gain is solely dependent on the number of measurements.

7.2. Other Information-Theoretical Motivations of the Jeffreys Binomial Prior

Wootters [20] introduces a novel perspective on the Jeffreys binomial prior, where quantum measurement is employed as a communication channel. In this framework, Alice aims to transmit a continuous variable, denoted as θ, to Bob. Instead of directly sending θ to Bob, Alice transmits a set of identical coins to Bob, where the probability of getting heads, $p(\theta)$, in each toss is a function of θ. Bob's objective is to maximize the information about θ that he can extract from a finite number of tosses. The measure of information used in this context is the mutual information between θ and the total number of heads, n, in N tosses.

$$I(n:\theta) = H(n) - H(n|\theta) = -\sum_{n=0}^{N} p(n)\ln P(n) - \left\langle -\sum_{n=0}^{N} p(n|p(\theta))\ln p(n|p(\theta))\right\rangle \tag{28}$$

However, the function $p(\theta)$ is unknown, and the optimization process begins with a set of discrete values, $p_1, p_2, \ldots, p_L$ rather than utilizing the continuous function $p(\theta)$. For each discrete value, p_k, there is an associated weight, w_k. The mutual information can be expressed as follows:

$$I(n:\theta) = -\sum_{n=0}^{N} p(n) \ln P(n) + \sum_{k=1}^{L} w_k \sum_{n=0}^{N} p(n|p_k) \ln p(n|p_k) \tag{29}$$

In the large N approximation, it is found that the weight w takes on a specific form:

$$w(p) = \frac{1}{\pi\sqrt{p(1-p)}} \tag{30}$$

which serves a role akin to the prior probability of p. Remarkably, this prior probability aligns with the Jeffreys binomial prior. A similar procedure can be extended to the Jeffreys multinomial prior distribution. Wootters' approach shares similarities with the concept of a reference prior, where the selected prior aims to maximize mutual information, which can be viewed as the expected information gain across all data. The outcome is consistent with the reference prior for multinomial data [24], thus revealing another informational interpretation of the Jeffreys prior.

8. Conclusions

In this paper, motivated by recent work in quantum reconstruction and quantum state tomography, we have investigated the concept of information gain for a two-outcome probabilistic source from an operational perspective. We have introduced an informational postulate, the Principle of Information Increase, which serves as a criterion for selecting the appropriate measure to quantify the extent of information gained from measurements and to guide the choice of prior. We have shown that differential information gain is the most physically meaningful measure when compared to the other contender: the relative information gain. We have also uncovered the unanticipated and rather remarkable result that the *expected* value of these two measures of information gain are *equal* for any prior and for any n-outcome probabilistic source.

Within the set of symmetric beta distributions, we have shown that the Jeffreys binomial prior exhibits notable characteristics. Both Summhammer's work and ours demonstrate that, under this prior, the intuitive notion that *more data from measurements leads to more knowledge about the system* holds true, as confirmed by two distinct methods of quantifying knowledge. Additionally, Wootters shows that this prior enables the communication of maximal information, further highlighting its significance. Here, we have formulated the novel notion of *robustness* and have shown that the Jeffreys binomial prior displays maximal robustness within the set of symmetric beta distributions. Our work raises the intriguing question of whether this feature could be extended to the multinomial Jeffreys prior and whether it would be possible to lift the initial restriction to the set of beta distributions. We also speculate that a deeper understanding of the robustness of the Jeffreys prior remains to be uncovered.

Author Contributions: Conceptualization, P.G.; methodology, P.G.; software, Y.Y.; formal analysis, P.G. and Y.Y..; investigation, Y.Y.; data curation, Y.Y.; writing—original draft preparation, Y.Y.; writing—review and editing, P.G.; visualization, P.G. and Y.Y.; supervision, P.G.; project administration, P.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting the findings of this study are derived from analytical expressions and numerical calculations detailed within the paper. All calculations were conducted under the specified parameters presented in each graph. The code used for these calculations is straightforward, reflecting direct implementation of the analytical expressions provided. While there are no external data sources, the code utilized for generating the data are available from the authors upon reasonable request.

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

Appendix A. Derivation of Differential Information Gain

The posterior is determined by T_N and a prior. For the sake of simplicity, we set that the prior belongs to the family of beta distributions:

$$\Pr(p|I) = \frac{p^{\alpha}(1-p)^{\alpha}}{B(\alpha+1,\alpha+1)} \tag{A1}$$

where $\alpha > -1, B(x,y)$ is the beta function.

Given N, there are 2^N different values of T_N. However, we may not need to calculate all 2^N sequences. Suppose every toss is independent—this happens in quantum mechanics—then this coin tossing model would become a binomial distribution. Let h_N be the number of heads inside T_N; the posterior $\Pr(p|N,T_N,I)$ is equivalent to $\Pr(p|N,h_N,I)$, and the likelihood will be

$$\Pr(h_N|N,p,I) = \binom{N}{h_N} p^{h_N}(1-p)^{N-h_N}. \tag{A2}$$

Hence, the posterior after N tosses is

$$\begin{aligned}\Pr(p|N,h_N,I) &= \frac{\Pr(h_N|N,p,I)\Pr(p|I)}{\int \Pr(h_N|N,p,I)\Pr(p|I)dp} \\ &= \frac{p^{h_N+\alpha}(1-p)^{N-h_N+\alpha}}{B(h_N+\alpha+1,N-h_N+\alpha+1)}\end{aligned} \tag{A3}$$

The information gain in the $(N+1)$th toss would be

$$I_{\text{diff}} = D_{\text{KL}}(\Pr(p|N+1,\{T_N,t_{N+1}\},I)||\Pr(p|I)) - D_{\text{KL}}(\Pr(p|N,h_N,I)||\Pr(p|I)) \tag{A4}$$

I_{diff} is determined by h_N and the prior, and the result of the $(N+1)$th toss t_{N+1}. t_{N+1} could be either "Head" or "Tail"; then, the posterior after $N+1$ tosses could be

$$\Pr(p|N+1,\{T_N,t_{N+1} = \text{‘Head''Head’}\},I) = \frac{p^{h_N+\alpha+1}(1-p)^{N-h_N+\alpha}}{B(h_N+\alpha+2,N-h_N+\alpha+1)} \tag{A5}$$

$$\Pr(p|N+1,\{T_N,t_{N+1} = \text{‘Tail’}\},I) = \frac{p^{h_N+\alpha}(1-p)^{N-h_N+\alpha+1}}{B(h_N+\alpha+1,N-h_N+\alpha+2)} \tag{A6}$$

Taking $t_{N+1} = \text{‘Head’}$, the first term in (A4) would become

$$\begin{aligned}&D_{\text{KL}}(\Pr(p|N+1,\{T_N,t_{N+1} = \text{‘Head’}\},I)||\Pr(p|I)) \\ &= \int_0^1 \Pr(p|N+1,h_N+1,I)\ln\frac{\Pr(p|N+1,h_N+1,I)}{\Pr(p|I)}dp \\ &= \int_0^1 \frac{p^{h_N+\alpha+1}(1-p)^{N-h_N+\alpha}}{B(h_N+\alpha+2,N-h_N+\alpha+1)}\ln\frac{p^{h_N+1}(1-p)^{N-h_N}B(\alpha+1,\alpha+1)}{B(h_N+\alpha+2,N-h_N+\alpha+1)}dp \\ &= \int_0^1 \frac{p^{h_N+\alpha+1}(1-p)^{N-h_N+\alpha}}{B(h_N+\alpha+2,N-h_N+\alpha+1)}\left\{\ln[p^{h_N+1}(1-p)^{N-h_N}] + \ln\frac{B(\alpha+1,\alpha+1)}{B(h_N+\alpha+2,N-h_N+\alpha+1)}\right\}dp \\ &= \int_0^1 \frac{p^{h_N+\alpha+1}(1-p)^{n-h_N+\alpha}}{B(h_N+\alpha+1,n-h_N+\alpha+1)}\ln[p^{h_N+1}(1-p)^{N-h_N}]dp + \ln\frac{B(\alpha+1,\alpha+1)}{B(h_N+\alpha+2,n-h_N+\alpha+1)}\end{aligned} \tag{A7}$$

By using the integral

$$\int_0^1 x^a(1-x)^b ln(x)dx = B(a+1,b+1)[\psi(a+1)-\psi(a+b+2)] \tag{A8}$$

where $\psi(x)$ is the digamma function, we can obtain the following result:

$$\begin{aligned} D_{\mathrm{KL}}(\Pr(p|N+1,\{T_N,t_{N+1}=\text{'Head'}\},I)||\Pr(p|I)) =& (h_N+1)\psi(h_N+\alpha+2)+(N-h_N)\psi(N-h_N+\alpha+1) \\ &-(N+1)\psi(N+2\alpha+3) \\ &+\ln\frac{B(\alpha+1,\alpha+1)}{B(h_N+\alpha+2,n-h_N+\alpha+1)} \end{aligned} \tag{A9}$$

The second term in (A4) would become

$$\begin{aligned} &D_{\mathrm{KL}}(\Pr(p|N,h_N,I)||\Pr(p|I)) \\ &=\int_0^1 \Pr(p|N,h_N,I)\ln\frac{\Pr(p|N,h_N,I)}{\Pr(p|I)}dp \\ &=\int_0^1 \frac{p^{h_N+\alpha}(1-p)^{N-h_N+\alpha}}{B(h_N+\alpha+1,N-h_N+\alpha+1)}\ln\frac{p^{h_N}(1-p)^{N-h_N}B(\alpha+1,\alpha+1)}{B(h_N+\alpha+1,N-h_N+\alpha+1)}dp \\ &=\int_0^1 \frac{p^{h_N+\alpha}(1-p)^{N-h_N+\alpha}}{B(h_N+\alpha+1,N-h_N+\alpha+1)}\left\{\ln[p^{h_N}(1-p)^{N-h_N}]+\ln\frac{B(\alpha+1,\alpha+1)}{B(h_N+\alpha+1,N-h_N+\alpha+1)}\right\}dp \\ &=\int_0^1 \frac{p^{h_N+\alpha}(1-p)^{n-h_N+\alpha}}{B(h_N+\alpha+1,n-h_N+\alpha+1)}\ln[p^{h_N}(1-p)^{N-h_N}]dp+\ln\frac{B(\alpha+1,\alpha+1)}{B(h_N+\alpha+1,n-h_N+\alpha+1)} \\ &=h_N\psi(h_N+\alpha+1)+(N-h_N)\psi(N-h_N+\alpha+1)-N\psi(N+2\alpha+2)+\ln\frac{B(\alpha+1,\alpha+1)}{B(h_N+\alpha+1,n-h_N+\alpha+1)} \end{aligned} \tag{A10}$$

Now, we obtain the final expression of (A4):

$$\begin{aligned} I_{\mathrm{diff}}(t_{N+1}=\text{'Head'}) &= D_{\mathrm{KL}}(\Pr(p|N+1,\{T_N,t_{N+1}=\text{'Head'}\},I)||\Pr(p|I))-D_{\mathrm{KL}}(\Pr(p|N,h_N,I)||\Pr(p|I)) \\ &=\psi(h_N+\alpha+2)-\psi(N+2\alpha+3) \\ &\quad+\frac{h_N}{h_N+\alpha+1}-\frac{N}{N+2\alpha+2}+\ln\frac{N+2\alpha+2}{h_N+\alpha+1} \end{aligned} \tag{A11}$$

Similarly, we can obtain I_{diff} when $t_{N+1}=$ 'Tail':

$$\begin{aligned} I_{\mathrm{diff}}(t_{N+1}=\text{'Tail'}) &= \psi(N-h_N+\alpha+2)-\psi(N-h_N+2\alpha+3) \\ &\quad+\frac{N-h_N}{N-h_N+\alpha+1}-\frac{N}{N+2\alpha+2}+\ln\frac{N+2\alpha+2}{N-h_N+\alpha+1} \end{aligned} \tag{A12}$$

This suggests that for fixed N and α, $I_{\mathrm{diff}}(t_{N+1}=$ 'Head') and $I_{\mathrm{diff}}(t_{N+1}=$ 'Tail') are symmetric since h_N ranges from 0 to N.

Appendix B. Derivation of Relative Information Gain

From Appendix A, we know that the posterior after N tosses is

$$\Pr(p|N,T_N,I)=\Pr(p|N,h_N,I)=\frac{p^{h_N+\alpha}(1-p)^{N-h_N+\alpha}}{B(h_N+\alpha+1,N-h_N+\alpha+1)} \tag{A13}$$

Therefore, the posterior after $N+1$ tosses would be

$$\Pr(p|N+1,T_{N+1},I)=\frac{\Pr(h_N,T_{N+1}|p,N+1,I)\Pr(p|I)}{\int_0^1 \Pr(h_N,T_{N+1}|p,N+1,I)\Pr(p|I)dp} \tag{A14}$$

Depending on different results for t_{N+1}, the posterior after $N+1$ tosses would be

$$\Pr(p|N+1,\{T_N,t_{N+1}=\text{'Head'}\},I)=\frac{p^{h_N+\alpha+1}(1-p)^{N-h_N+\alpha}}{B(h_N+\alpha+2,N-h_N+\alpha+1)} \tag{A15}$$

$$\Pr(p|N+1,\{T_N,t_{N+1}=\text{'Tail'}\},I)=\frac{p^{h_N+\alpha}(1-p)^{N-h_N+\alpha+1}}{B(h_N+\alpha+1,N-h_N+\alpha+2)} \tag{A16}$$

And the corresponding relative information gain would be

$$\begin{aligned}
&I_{\text{rel}}(t_{N+1} = \text{`Head'}) \\
&= D_{\text{KL}}(\Pr(p|N+1, \{T_N, t_{N+1} = \text{`Head'}\}, I)||\Pr(p|N, h_N, I)) \\
&= \int_0^1 \Pr(p|N+1, \{T_N, t_{N+1} = \text{`Head'}\}, I) \ln \frac{\Pr(p|N+1, \{T_N, t_{N+1} = \text{`Head'}\}, I)}{\Pr(p|N, h_N, I)} dp \\
&= \int_0^1 \frac{p^{h_N+\alpha+1}(1-p)^{N-h_N+\alpha}}{B(h_N+\alpha+2, N-h_N+\alpha+1)} \ln \frac{pB(h_N+\alpha+1, N-h_N+\alpha+1)}{B(h_N+\alpha+2, N-h_N+\alpha+1)} dp \\
&= \psi(h_N+\alpha+2) - \psi(N+2\alpha+3) + \ln \frac{N+2\alpha+2}{h_N+\alpha+1}
\end{aligned} \tag{A17}$$

$$I_{\text{rel}}(t_{N+1} = \text{`Tail'}) = \psi(N-h_N+\alpha+2) - \psi(N-h_N+2\alpha+3) + \ln \frac{N+2\alpha+2}{N-h_N+\alpha+1} \tag{A18}$$

Appendix C. Equivalence of Expected Differential Information Gain and Expected Relative Information Gain

In a n-outcome model, the probability of each outcome is p_i, and

$$p_1 + p_2 + \cdots + p_n = 1 \tag{A19}$$

After N "tosses", the data sequence has the form

$$D_N = (f_1, f_2, \cdots, f_n), \quad \sum_{i=1}^{n} f_i = N \tag{A20}$$

where f_i is the number of ith outcomes in these N tosses.

We may use a tuple $\vec{p} = (p_1, p_2, \cdots, p_n)$ to represent the probabilities of these outcomes. The prior is just $\Pr(\vec{p}|I)$, and the posterior based on the data D_N is $\Pr(\vec{p}|D_N, I)$.

The average value of the ith outcome probability is

$$\langle p_i \rangle = \int p_i \Pr(\vec{p}|D_N, I) dp_1 dp_2 \cdots dp_n \tag{A21}$$

Assume the $(N+1)$th toss is the ith outcome, and the posterior of these after this additional toss is

$$\begin{aligned}
\Pr(\vec{p}|D_N, d_{N+1} = \text{`}i\text{'}, I) &= \frac{p_i \Pr(\vec{p}|D_N, I)}{\int p_i \Pr(\vec{p}|D_N, I) dp_1 dp_2 \cdots dp_n} \\
&= \frac{p_i}{\langle p_i \rangle} \Pr(\vec{p}|D_N, I)
\end{aligned} \tag{A22}$$

Then we can write I_{diff} as

$$\begin{aligned}
I_{\text{diff}}(d_{N+1} = \text{`}i\text{'}) &= D_{\text{KL}}(\Pr(\vec{p}|D_N, d_{N+1} = \text{`}i\text{'}, I)|\Pr(\vec{p}|I)) - D_{\text{KL}}(\Pr(\vec{p}|D_N, I)|\Pr(\vec{p}|I)) \\
&= \int \frac{p_i}{\langle p_i \rangle} \Pr(\vec{p}|D_N, I) \ln \frac{p_i \Pr(\vec{p}|D_N, I)}{\langle p_i \rangle \Pr(\vec{p}|I)} dp_1 dp_2 \cdots dp_n - \int \Pr(\vec{p}|D_N, I) \ln \frac{\Pr(\vec{p}|D_N, I)}{\Pr(\vec{p}|I)} dp_1 dp_2 \cdots dp_n
\end{aligned} \tag{A23}$$

Then the expected differential information gain is given by

$$\begin{aligned}
\overline{I_{\text{diff}}} &= \sum_{i=1}^{n} \langle p_i \rangle \, I_{\text{diff}}(d_{N+1} = `i') \\
&= \sum_{i=1}^{n} \int p_i \Pr(\vec{p}|D_N, I) \ln \frac{p_i \Pr(\vec{p}|D_N, I)}{\langle p_i \rangle \Pr(\vec{p}|I)} dp_1 dp_2 \cdots dp_n - \sum_{i=1}^{n} \langle p_i \rangle \int \Pr(\vec{p}|D_N, I) \ln \frac{\Pr(\vec{p}|D_N, I)}{\Pr(\vec{p}|I)} dp_1 dp_2 \cdots dp_n \\
&= \sum_{i=1}^{n} \left[\int p_i \Pr(\vec{p}|D_N, I) \ln \frac{p_i}{\langle p_i \rangle} dp_1 dp_2 \cdots dp_n + \int p_i \Pr(\vec{p}|D_N, I) \ln \frac{\Pr(\vec{p}|D_N, I)}{\Pr(\vec{p}|I)} dp_1 dp_2 \cdots dp_n \right] \\
&\quad - \int \Pr(\vec{p}|D_N, I) \ln \frac{\Pr(\vec{p}|D_N, I)}{\Pr(\vec{p}|I)} dp_1 dp_2 \cdots dp_n \\
&= \sum_{i=1}^{n} \int p_i \Pr(\vec{p}|D_N, I) \ln \frac{p_i}{\langle p_i \rangle} dp_1 dp_2 \cdots dp_n + \int \sum_{i=1}^{n} p_i \Pr(\vec{p}|D_N, I) \ln \frac{\Pr(\vec{p}|D_N, I)}{\Pr(\vec{p}|I)} dp_1 dp_2 \cdots dp_n \\
&\quad - \int \Pr(\vec{p}|D_N, I) \ln \frac{\Pr(\vec{p}|D_N, I)}{\Pr(\vec{p}|I)} dp_1 dp_2 \cdots dp_n \\
&= \sum_{i=1}^{n} \int p_i \Pr(\vec{p}|D_N, I) \ln \frac{p_i}{\langle p_i \rangle} dp_1 dp_2 \cdots dp_n
\end{aligned} \tag{A24}$$

Similarly, I_{rel} can be written as

$$\begin{aligned}
I_{\text{rel}}(d_{N+1} = `i') =& D_{\text{KL}}(\Pr(\vec{p}|D_N, d_{N+1} = `i', I) | \Pr(\vec{p}|D_N, I)) \\
&= \int \frac{p_i}{\langle p_i \rangle} \Pr(\vec{p}|D_N, I) \ln \frac{p_i \Pr(\vec{p}|D_N, I)}{\langle p_i \rangle \Pr(\vec{p}|D_N, I)} dp_1 dp_2 \cdots dp_n \\
&= \int \frac{p_i}{\langle p_i \rangle} \Pr(\vec{p}|D_N, I) \ln \frac{p_i}{\langle p_i \rangle} dp_1 dp_2 \cdots dp_n
\end{aligned} \tag{A25}$$

Then the expected relative information gain is, accordingly,

$$\overline{I_{\text{rel}}} = \sum_{i=1}^{n} \langle p_i \rangle \, I_{\text{rel}}(d_{N+1} = `i') = \sum_{i=1}^{n} \int p_i \Pr(\vec{p}|D_N, I) \ln \frac{p_i}{\langle p_i \rangle} dp_1 dp_2 \cdots dp_n \tag{A26}$$

From (A24) and (A26), we can see that in this n-outcome model, the expected differential information gain $\overline{I_{\text{diff}}}$ and expected relative information gain $\overline{I_{\text{rel}}}$ are equal, irrespective of the choice of prior.

References

1. Patra, M.K. Quantum state determination: estimates for information gain and some exact calculations. *J. Phys. A Math. Theor.* **2007**, *40*, 10887–10902. [CrossRef]
2. Madhok, V.; Riofrío, C.A.; Ghose, S.; Deutsch, I.H. Information Gain in Tomography–A Quantum Signature of Chaos. *Phys. Rev. Lett.* **2014**, *112*, 014102. [CrossRef] [PubMed]
3. Quek, Y.; Fort, S.; Ng, H.K. Adaptive quantum state tomography with neural networks. *npj Quantum Inf.* **2021**, *7*, 105. [CrossRef]
4. Gupta, R.; Xia, R.; Levine, R.D.; Kais, S. Maximal Entropy Approach for Quantum State Tomography. *PRX Quantum* **2021**, *2*, 010318. [CrossRef]
5. McMichael, R.D.; Dushenko, S.; Blakley, S.M. Sequential Bayesian experiment design for adaptive Ramsey sequence measurements. *J. Appl. Phys.* **2021**, *130*, 144401. [CrossRef]
6. Placek, B.; Angerhausen, D.; Knuth, K.H. Analyzing Exoplanet Phase Curve Information Content: Toward Optimized Observing Strategies. *Astron. J.* **2017**, *154*, 154. [CrossRef]
7. Ma, C.W.; Ma, Y.G. Shannon information entropy in heavy-ion collisions. *Prog. Part. Nuclear Phys.* **2018**, *99*, 120–158. [CrossRef]
8. Grinbaum, A. Elements of information-theoretic derivation of the formalism of quantum theory. *Int. J. Quantum Inf.* **2003**, *1*, 289–300. [CrossRef]
9. Brukner, V.; Zeilinger, A. Information Invariance and Quantum Probabilities. *Foundations Phys.* **2009**, *39*, 677–689. [CrossRef]
10. Goyal, P.; Knuth, K.H.; Skilling, J. Origin of Complex Quantum Amplitudes and Feynman's Rules. *Phys. Rev. A* **2010**, *81*, 022109. [CrossRef]
11. Caticha, A. Entropic dynamics, time and quantum theory. *J. Phys. A Math. Theor.* **2011**, *44*, 225303. [CrossRef]

12. Masanes, L.; Müller, M.P.; Augusiak, R.; Pérez-García, D. Existence of an information unit as a postulate of quantum theory. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 16373–16377. [CrossRef] [PubMed]
13. De Raedt, H.; Katsnelson, M.I.; Michielsen, K. Quantum theory as plausible reasoning applied to data obtained by robust experiments. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2016**, *374*, 20150233. [CrossRef]
14. Höhn, P.A. Quantum Theory from Rules on Information Acquisition. *Entropy* **2017**, *19*, 98. [CrossRef]
15. Aravinda, S.; Srikanth, R.; Pathak, A. On the origin of nonclassicality in single systems. *J. Phys. A Math. Theor.* **2017**, *50*, 465303. [CrossRef]
16. Czekaj, L.; Horodecki, M.; Horodecki, P.; Horodecki, R. Information content of systems as a physical principle. *Phys. Rev. A* **2017**, *95*, 022119. [CrossRef]
17. Chiribella, G. Agents, Subsystems, and the Conservation of Information. *Entropy* **2018**, *20*, 358. [CrossRef] [PubMed]
18. Summhammer, J. Maximum predictive power and the superposition principle. *Int. J. Theor. Phys.* **1994**, *33*, 171–178. [CrossRef]
19. Summhammer, J. Maximum predictive power and the superposition principle. *arXiv* **1999**, arXiv.quant-ph/9910039.
20. Wootters, W.K. Communicating through Probabilities: Does Quantum TheoryOptimize the Transfer of Information? *Entropy* **2013**, *15*, 3130–3147. [CrossRef]
21. Cover, T.M.; Thomas, J.A. Differential Entropy. In *Elements of Information Theory*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2005; chapter 8, pp. 243–259. [CrossRef]
22. Jaynes, E.T. Information Theory and Statistical Mechanics. In *Statistical Physics*; Ford, K.W., Ed.; W. A. Benjamin, Inc.: Tokyo, Japan, 1963; pp. 181–218.
23. Goyal, P. Prior Probabilities: An Information-Theoretic Approach. *AIP Conf. Proc.* **2005**, *803*, 366–373. [CrossRef]
24. Berger, J.O.; Bernardo, J.M. Ordered Group Reference Priors with Application to the Multinomial Problem. *Biometrika* **1992**, *79*, 25–37. [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.