

Article

Generalizing Information to the Evolution of Rational Belief

Jed A. Duersch * and Thomas A. Catanach * 

Sandia National Laboratories, Livermore, CA 94550, USA

* Correspondence: jaduers@sandia.gov (J.A.D.); tacatan@sandia.gov (T.A.C.)

Received: 19 November 2019; Accepted: 11 January 2020; Published: 16 January 2020



Abstract: Information theory provides a mathematical foundation to measure uncertainty in belief. Belief is represented by a probability distribution that captures our understanding of an outcome's plausibility. Information measures based on Shannon's concept of entropy include realization information, Kullback–Leibler divergence, Lindley's information in experiment, cross entropy, and mutual information. We derive a general theory of information from first principles that accounts for evolving belief and recovers all of these measures. Rather than simply gauging uncertainty, information is understood in this theory to measure change in belief. We may then regard entropy as the information we expect to gain upon realization of a discrete latent random variable. This theory of information is compatible with the Bayesian paradigm in which rational belief is updated as evidence becomes available. Furthermore, this theory admits novel measures of information with well-defined properties, which we explored in both analysis and experiment. This view of information illuminates the study of machine learning by allowing us to quantify information captured by a predictive model and distinguish it from residual information contained in training data. We gain related insights regarding feature selection, anomaly detection, and novel Bayesian approaches.

Keywords: information; Bayesian inference; entropy; self information; mutual information; Kullback–Leibler divergence; Lindley information; maximal uncertainty; proper utility

1. Introduction

This work integrates essential properties of information embedded within Shannon's derivation of entropy [1] and the Bayesian perspective [2–4], which identifies probability with plausibility. We pursued this investigation in order to understand how to rigorously apply information-theoretic concepts to the theory of inference and machine learning. Specifically, we wanted to understand how to quantify the evolution of predictions given by machine learning models. Our findings are general, however, and bear implications for any situation in which states of belief are updated. We begin in Section 1.1 with an experiment that illustrates shortcomings with the way standard information measures would partition prediction information and residual information during machine learning training.

1.1. Shortcomings with Standard Approaches

Let us examine a typical MNIST [5] classifier. This dataset comprises a set of images of handwritten digits paired with labels. Let both x and y denote random variables corresponding, respectively, to an image and a label in a pair. In this situation, the training dataset contains independent realizations of such pairs from an unknown joint probability distribution. We would like to obtain a measurement of prediction information that quantifies a shift in belief from an uninformed initial state $\mathbf{q}_0(y)$ to model predictions $\mathbf{q}_1(y|x)$. The symmetric uninformed choice for $\mathbf{q}_0(y)$ is uniform probability over all outcomes. Note that both $\mathbf{q}_0(y)$ and $\mathbf{q}_1(y|x)$ are simply hypothetical states of belief. Some architectures

may approximate Bayesian inference, but we cannot always interpret these as the Bayesian prior and posterior.

Two measurements that are closely related to Shannon’s entropy are the Kullback–Leibler (KL) divergence [6,7] and Lindley’s information in experiment [8], which are computed respectively as

$$D_{KL}[\mathbf{q}_1(y|x) \parallel \mathbf{q}_0(y)] = \int dy \mathbf{q}_1(y|x) \log\left(\frac{\mathbf{q}_1(y|x)}{\mathbf{q}_0(y)}\right) \quad \text{and}$$

$$D_L[\mathbf{q}_1(y|x) \parallel \mathbf{q}_0(y)] = \int dy \mathbf{q}_1(y|x) \log(\mathbf{q}_1(y|x)) - \int dy \mathbf{q}_0(y) \log(\mathbf{q}_0(y)).$$

Whatever we choose, we would like to use a consistent construction to understand how much information remains unpredicted. After viewing a label outcome \check{y} , we let $\mathbf{r}(y|\check{y})$ represent our new understanding of the actual state of affairs, which is a realization assigning full probability to the specified outcome. This distribution captures our most updated knowledge about y , and therefore constitutes rational belief. A consistent information measurement should then quantify residual information as the shift in belief from $\mathbf{q}_1(y|x)$ to $\mathbf{r}(y|\check{y})$. For example, the KL version would be $D_{KL}[\mathbf{r}(y|\check{y}) \parallel \mathbf{q}_1(y|x)]$.

In order to demonstrate shortcomings with each approach, some cases are deliberately mislabeled during model testing. We first compute information measurements while assuming the incorrect labels hold. Mislabeled cases are then corrected to \hat{y} with corrected belief given by $\mathbf{r}(y|\hat{y})$. This allows us to compare our first information measurements with corrected versions. An example of each belief state is shown in Figure 1 where the incorrect label 3 is changed to 0.

Digit	0	1	2	3	4	5	6	7	8	9
$\mathbf{q}_0(y)$	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
$\mathbf{q}_1(y x)$	0.952	0.0	0.045	0.001	0.0	0.001	0.0	0.0	0.0	0.002
$\mathbf{r}(y \check{y})$	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
$\mathbf{r}(y \hat{y})$	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

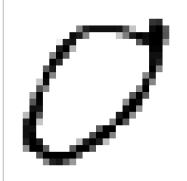


Figure 1. Example of the evolution of plausible labels for an image. Without evidence, the probability distribution $\mathbf{q}_0(y)$ assigns equal plausibility to all outcomes. A machine learning model processes the image and produces predictions $\mathbf{q}_1(y|x)$. The incorrect label 3 is represented by $\mathbf{r}(y|\check{y})$. After observing the image, shown on the right, the label is corrected to 0 in $\mathbf{r}(y|\hat{y})$.

Ideally, the sum of prediction information and residual information would be a conserved quantity, which would allow us to understand training as simply shifting information from a residual partition to the predicted partition. More importantly, however, prediction information should clearly capture prediction quality. Figure 2 shows information measurements corresponding to the example given.

Information Type	Original Label			Corrected Label		
	Prediction	Residual	Sum	Prediction	Residual	Sum
Kullback–Leibler	3.02	10.44	13.46	3.02	0.07	3.09
Lindley	3.02	0.30	3.32	3.02	0.30	3.32

Figure 2. Information measurements before and after label correction. Neither construction of prediction information allows the computation to account for claimed labels. Residual information, however, decreases in the KL construction when the label is corrected. The Lindley forms are totally unaffected by relabeling.

Mislabeled and relabeled shows us that neither formulation of prediction information captures prediction quality. This is because these constructions simply have no affordance to account for our understanding of what is actually correct. Large KL residual information offers some indication of mislabeling and decreases when we correct the label, but total information is not conserved. As such,

there is no intuitive notion for what appropriate prediction and residual information should be for a given problem. The Lindley formulation is substantially less satisfying. Although total information is conserved, neither metric changes and we have no indication of mislabeling.

The problem with these constructions is they do not recognize the gravity of the role of expectation. That is, reasonable expectation must be consistent with rational belief. We hold that our most justified understanding of what may be true provides a sound basis to measure changes in belief. Figure 3 gives a preview of information measurements in the framework of this theory. Total information, $\log_2(10)$ bits in this case, is conserved, and both prediction information and residual information react intuitively to mislabeling. Determining whether the predictive information is positive or negative provides a clear indication of whether the prediction was informative.

Information Type	Original Label			Corrected Label		
	Prediction	Residual	Sum	Prediction	Residual	Sum
Proposed	-7.11	10.44	3.32	3.25	0.07	3.32

Figure 3. Information measurements using our proposed framework. Total information is a conserved quantity, and when our belief changes, so do the information measurements. Negative prediction information forewarns either potential mislabeling or a poor prediction.

1.2. Our Contributions

In the course of pursuing a consistent framework in which information measurements may be understood, we have derived a theory of information from first principles that places all entropic information measures in a unified, interpretable context. By axiomatizing the properties of information we desire, we show that a unique formulation follows that subsumes critical properties of Shannon's construction of entropy.

This theory fundamentally understands entropic information as a form of reasonable expectation that measures the change between hypothetical belief states. Expectation is not necessarily taken with respect to the distributions that represent the shift in belief, but rather with respect to a third distribution representing our understanding of what may actually be true. We found compelling foundations for this perspective within the Bayesian philosophy of probability as an extended logic for expressing and updating uncertainty [4,9]. Our understanding of what may be true, and therefore the basis for measuring information, should be rational belief. Rational belief [10–14] begins with probabilistically coherent prior knowledge, and is subsequently updated to account for observations using Bayes' theorem. As a consequence, information associated with a change in belief is not a fixed quantity. Just as rational belief must evolve as new evidence becomes available, so also does the information we would reasonably assign to previous shifts in belief. By emphasizing the role of rational belief, this theory recognizes that the degree of validity we assign to past states of belief is both dynamic and potentially subjective as our state of knowledge matures.

As a consequence of enforcing consistency with rational belief, a second additivity property emerges; just as entropy can be summed over independent distributions, information gained over a sequence of observations can be summed over intermediate belief updates. Total information over such a sequence is independent of how results are grouped or ordered. This provides a compelling solution to the thought experiment above. Label information in training data is a conserved quantity and we motivate a formulation of prediction information that is directly tied to prediction quality.

Soofi, Ebrahimi, and others [15–18] identified key contributions to information theory in the decade following Shannon's paper that are intrinsically tied to entropy. These are the Kullback–Leibler divergence, Lindley's information in experiment, and Jaynes' construction of entropy-maximizing distributions that are consistent with specified expectations. We show how this theory recovers these measures of information and admits new forms that may not have been previously associated with entropic information, such as the log pointwise posterior predictive measure of model accuracy [19]. We also show how this theory admits novel information-optimal probability distributions analogous to that of Jaynes' maximum uncertainty. Having a consistent interpretation of information illuminates

how it may be applied and what properties will hold in a given context. Moreover, this theoretical framework enables us to solve multiple challenges in Bayesian learning. For example, one such challenge is understanding how efficiently a given model incorporates new data. This theory provides bounds on the information gained by a model resulting from inference and allows us to characterize the information provided by individual observations.

The rest of this paper is organized as follows. Section 2 discusses notation and background regarding entropic information, Bayesian inference, and reasonable expectation. Section 3 contains postulates that express properties of information we desire, the formulation of information that follows, and other related measures of information. Section 4 analyzes general consequences and properties of this formulation. Section 5 discusses further implications with respect to Bayesian inference and machine learning. Section 6 explores negative information with computational experiments that illustrate when it occurs, how it may be understood, and why it is useful. Section 7 summarizes these results and offers a brief discussion of future work. Appendix A proves our principal result. Appendix B contains all corollary proofs. Appendix C provides key computations used in experiments.

2. Background and Notation

Shannon's construction of entropy [1] shares a fundamental connection with thermodynamics. The motivation is to facilitate analysis of complex systems which can be decomposed into independent subsystems. The essential idea is simple—when probabilities multiply, entropy adds. This abstraction allows us to compose uncertainties across independent sources by simply adding results. Shannon applied this perspective to streams of symbols called channels. The number of possible outcomes grows exponentially with the length of a symbol sequence, whereas entropy grows linearly. This facilitates a rigorous formulation of the rate of information conveyed by a channel and analysis of what is possible in the presence of noise.

The property of independent additivity is used in standard training practices for machine learning. Just as thermodynamic systems and streams of symbols break apart, so does an ensemble of predictions over independent observations. This allows us to partition training sets into batches and compute cross-entropy [20] averages. MacKay [21] gives a comprehensive discussion of information in the context of learning algorithms. Tishby [22] examines information trends during neural network training.

A second critical property of entropy, which is implied by Shannon and further articulated by both Barnard [23] and Rényi [24], is that entropy is an expectation. Given a latent random variable z , we denote the probability distribution over outcomes as $\mathbf{p}(z)$. Stated as an expectation, entropy is defined as

$$S[\mathbf{p}(z)] = \int dz \mathbf{p}(z) \log\left(\frac{1}{\mathbf{p}(z)}\right) = \mathbb{E}_{\mathbf{p}(z)} \log\left(\frac{1}{\mathbf{p}(z)}\right).$$

Following Shannon, investigators developed a progression of divergence measures between general probability distributions, $\mathbf{q}_0(z)$ and $\mathbf{q}_1(z)$. Notable cases include the Kullback–Leibler divergence, Rényi's information of order- α [24], and Csiszár's f -divergence [25]. Ebrahimi, Soofi, and Soyer [18] offered an examination of these axiomatic foundations and generalizations with a primary focus on entropy and the KL divergence. Recent work on axiomatic foundations for generalized entropies [26] includes constructions that are suitable for strongly-interacting systems [27] and axiomatic derivations of other forms of entropy, including Sharma–Mittal and Frank–Daffertshofer entropies [28]. Further work uses group theory to relate properties of systems to corresponding notions of entropy and correlation laws [29].

2.1. Bayesian Reasoning

The Bayesian view of probability, going back to Laplace [2] and championed by Jeffreys [3] and Jaynes [4], focuses on capturing our beliefs. This perspective considers a probability distribution

as an abstraction that attempts to model these beliefs. This view subsumes all potential sources of uncertainty and provides a comprehensive scope that facilitates analysis in diverse contexts.

In the Bayesian framework, the prior distribution $\mathbf{p}(z)$ expresses initial beliefs about some latent variable z . Statisticians, scientists, and engineers often have well-founded views about real-world systems that form the basis for priors. Examples include physically realistic ranges of model parameters or plausible responses of a dynamical system. In the case of total ignorance, one applies the principle of insufficient reason [30]—we should not break symmetries of belief without justification. Jaynes' construction of maximally uncertainty distributions [31] generalizes this principle, which we discuss further in Section 4.6.

As observations x become available, we update belief from the prior distribution to obtain the posterior distribution $\mathbf{p}(z|x)$, which incorporates this new knowledge. This update is achieved by applying Bayes' theorem

$$\mathbf{p}(z|x) = \frac{\mathbf{p}(x|z)\mathbf{p}(z)}{\mathbf{p}(x)} \quad \text{where} \quad \mathbf{p}(x) = \int dz \mathbf{p}(x|z)\mathbf{p}(z).$$

The likelihood distribution $\mathbf{p}(x|z)$ expresses the probability of observations given any specified value of z . The normalization constant $\mathbf{p}(x)$ is also the probability of x given the prior belief that has been specified. Within Bayesian inference, this is also called model evidence and it is used to evaluate a model structure's plausibility for generating the observations.

Shore and Johnson [32,33] provide an axiomatic foundation for updating belief that recovers the principles of maximum entropy and minimum cross-entropy when prior evidence consists of known expectations. For reference, we summarize these axioms as

1. *Uniqueness.* When belief is updated with new observations, the result should be unique.
2. *Coordinate invariance.* Belief updates should be invariant to arbitrary choices of coordinates.
3. *System independence.* The theory should yield consistent results when independent random variables are treated either separately or jointly.
4. *Subset independence.* When we partition potential outcomes into disjoint subsets, the belief update corresponding to conditioning on subset membership first should yield the same result as updating first and conditioning on the subset second.

Jizba and Korbel [34] investigated generalizations of entropy for which the maximum entropy principle satisfies these axioms.

Integrating the maturing notion of belief found within the Bayesian framework with information theory recognizes that our perception of how informative observations are depends on how our beliefs develop, which is dynamic as our state of knowledge grows.

2.2. Probability Notation

Random variables are denoted in boldface; for example, x . Typically, x and y will imply observable measurements and z will indicate either a latent explanatory variable or unknown observable. Each random variable is implicitly associated with a corresponding probability space, including the set of all possible outcomes Ω_z , a σ -algebra \mathcal{F}_z of measurable subsets, and a probability measure \mathcal{P}_z which maps subsets of events to probabilities. We then express the probability measure as a distribution function $\mathbf{p}(z)$.

A realization, or specific outcome, will be denoted with either a check \check{z} , or, for discrete distributions only, a subscript z_i , where $i \in [n]$ and $[n] = \{1, 2, \dots, n\}$. If it is necessary to emphasize the value of a distribution at a specific point or realization, we will use the notation $\mathbf{p}(z = \check{z})$. Conditional dependence is denoted in the usual fashion as $\mathbf{p}(z|x)$. The joint distribution is then $\mathbf{p}(x, z) = \mathbf{p}(z|x)\mathbf{p}(x)$, and marginalization is obtained by $\mathbf{p}(x) = \int dz \mathbf{p}(x, z)$. When two distributions are equivalent over all subsets of nonzero measure, we use notation $\mathbf{q}_0(z) \equiv \mathbf{p}(z)$ or $\mathbf{q}_1(z) \equiv \mathbf{p}(z|x)$.

The probability measure allows us to compute expectations over functions $f(z)$ which are denoted

$$\mathbb{E}_{\mathbf{p}(z)}f(z) = \int dz \mathbf{p}(z)f(z).$$

The support of integration or summation is implied to be the same as the support of $\mathbf{p}(z)$; that is, the set of outcomes for which $\mathbf{p}(z) > 0$. For example, in both the discrete case above and continuous cases, such as a distribution on the unit interval $z \in \mathbb{R}_{[0,1]}$, the integral notation should be interpreted respectively as

$$\int dz \mathbf{p}(z)f(z) = \sum_{i=1}^n \mathbf{p}(z = z_i)f(z_i) \quad \text{and} \quad \int dz \mathbf{p}(z)f(z) = \int_0^1 d\check{z} \mathbf{p}(z = \check{z})f(\check{z}).$$

2.3. Reasonable Expectation and Rational Belief

The postulates and theory in this work concern the measurement of a shift in belief from an initial state $\mathbf{q}_0(z)$ to an updated state $\mathbf{q}_1(z)$. In principle, these are any hypothetical states of belief. For example, they could be predictions given by the computational model in Section 1.1, previous beliefs held before observing additional data, or convenient approximations of a more informed state of belief. A third state $\mathbf{r}(z)$, *rational belief*, serves a distinct role as the distribution over which expectation is taken. When we wish to emphasize this role, we also refer to $\mathbf{r}(z)$ as the *view of expectation*.

To understand the significance of rational belief, we briefly review work by Cox [9] regarding reasonable expectation from two perspectives on the meaning of probability. The first perspective understands probability as a description of relative frequencies in an ensemble. If we prepare a large ensemble of independent random variables, $Z = \{z_i | i \in [n]\}$, and each is realized from a proper (normalized) probability distribution $\mathbf{p}(z)$, then the relative frequency of outcomes within each subset $\omega \in \Omega_z$ will approach the probability measure $\mathcal{P}_z(\omega)$ for large n . It follows that the ensemble mean of any transformation $f(z)$ will approach the expectation

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n f(z_i) = \mathbb{E}_{\mathbf{p}(z)}f(z).$$

The difficulty arises when we distinguish *what is true* from *what may be known*, given limited evidence. This falls within the purview of the second perspective, the Bayesian view, regarding probability as an extended logic. To illustrate, suppose z is the value of an unknown real mathematical constant. The true probability distribution would be a Dirac delta $\mathbf{p}(z) \equiv \delta(z - \check{z})$ assigning unit probability to the unknown value \check{z} . Accordingly, each element in the ensemble above would take the same unknown value. If we have incomplete knowledge $\mathbf{r}(z)$ regarding the distribution of plausible values, then we can still compute an expectation $\mathbb{E}_{\mathbf{r}(z)}f(z)$, but we must bear in mind that the result only approximates the unknown true expectation. Since the expectation is limited by the credibility of $\mathbf{r}(z)$, we seek to drive belief towards the truth as efficiently as possible from available evidence to fulfill this role.

Within Bayesian Epistemology, rational belief is defined as a belief that is unsusceptible to a Dutch Book. When an agent's beliefs correspond to their willingness to place bets, a Dutch Book [11–14,35] means that it is possible for a bookie to construct a table of bets that the agent finds acceptable but also guarantees that the agent will lose money. Therefore the existence of such a table corresponds to the agent holding an irrational state of belief. When multiple bets are allowed to be conditioned on a sequence of outcomes, it has been shown that the agent must use Bayes' Theorem to account for previous outcomes in the sequence to update beliefs regarding subsequent outcomes to avoid irrationality [14].

For our purposes, it is sufficient to say that if we have a coherent prior belief in a latent variable and a likelihood function that implies beliefs about observations, Bayes' theorem incorporates observational

evidence to from the posterior distribution representing rational belief. For example, we could measure inference information from prior belief $\mathbf{q}_0(z) \equiv \mathbf{p}(z)$ to the first posterior $\mathbf{q}_1(z) \equiv \mathbf{p}(z|x)$ conditioned on an observation x . When we have additional evidence y that complements x , then rational belief must correspond to a second inference $\mathbf{r}(z) \equiv \mathbf{p}(z|x, y)$, because retaining the belief $\mathbf{q}_1(z)$ would not account for y . Likewise, if z is an observable realization, then rational belief must assign full probability to the observed outcome \check{z} . This case is specifically denoted as $\mathbf{r}(z|\check{z})$, and in continuous settings it is equivalent to the Dirac delta function $\mathbf{r}(z|\check{z}) \equiv \delta(z - \check{z})$.

2.4. Remarks on Bayesian Objectivism and Subjectivism

Within the Bayesian philosophy, we may disagree about whether or not rational belief is unique. This disagreement corresponds to objectivist versus subjectivist views of Bayesian epistemology; see [4,36] for a discussion. Note that this is not the same as the more general view of objective versus subjective probabilities.

In the objectivist's view, one's beliefs must be consistent with the entirety of evidence and prior knowledge must be justified by sound principles of reason. Therefore, anyone with the same body of evidence must hold the same rational belief. In contrast, the subjectivist holds that one's prior beliefs do not need justification. Provided evidence is taken into account using Bayes' theorem, the resulting posterior is rational for any prior as long as the prior is coherent. Note that the subjectivist view does not imply that all beliefs are equally valid. It simply allows validity in the construction of prior belief to be derived from other notions of utility, such as computational feasibility.

While the following postulates in Section 3 and derivation of Theorem 1 do not require adoption of either perspective, these philosophies influence how we understand reasonable expectation. The objective philosophy implies that an information measurement is justified to the same degree as the view of expectation that defines it, whereas the subjective philosophy entertains information analysis with any view of expectation.

3. Information and Evolution of Belief

In order to provide context for comparison, we begin by presenting the properties of entropic information originally put forward by Shannon using our notation.

3.1. Shannon's Properties of Entropy

1. Given a discrete probability distribution $\mathbf{p}(z)$ for which $z \in \{z_i \mid i \in [n]\}$, the entropy $S[\mathbf{p}(z)]$ is continuous in the probability of each outcome $\mathbf{p}(z = z_i)$.
2. If all outcomes are equally probable, namely, $\mathbf{p}(z = z_i) = 1/n$, then $S[\mathbf{p}(z)]$ is monotonically increasing in n .
3. The entropy of a joint random variable $S[\mathbf{p}(z, w)]$ can be decomposed using a chain rule expressing conditional dependence

$$S[\mathbf{p}(z, w)] = S[\mathbf{p}(z)] + \mathbb{E}_{\mathbf{p}(z)} S[\mathbf{p}(w|z)].$$

The first point is aimed at extending Shannon's derivation, which employs rational probabilities, to real-valued probabilities. The second point drives at understanding entropy as a measure of uncertainty; as the number of possible outcomes increases, each realization becomes less predictable. This results in entropy taking positive values. The third point is critical, as—not only does it encode independent additivity, it implies that entropic information is computed as an expectation.

We note that Fadeeva [37] gives a simplified set of postulates. R enyi [24] generalizes information by replacing the last point with a weaker version, which simply requires independent additivity, but not conditional expectation. This results in α -divergences. Csisz ar [25] generalizes this further using convex functions f to obtain f -divergences.

3.2. Postulates

Rather than repeating direct analogs of Shannon's properties in the context of evolving belief, it is both simpler and more illuminating to be immediately forthcoming regarding the key requirement of information in the perspective of this theory.

Postulate 1. Entropic information associated with the change in belief from $\mathbf{q}_0(z)$ to $\mathbf{q}_1(z)$ is quantified as an expectation over belief $\mathbf{r}(z)$, which we call the view of expectation. As an expectation, it must have the functional form

$$\mathbb{I}_{\mathbf{r}(z)}[\mathbf{q}_1(z) \parallel \mathbf{q}_0(z)] = \int dz \mathbf{r}(z) f(\mathbf{r}(z), \mathbf{q}_1(z), \mathbf{q}_0(z)).$$

Postulate 2. Entropic information is additive over independent belief processes. Taking joint distributions associated with two independent random variables z and w to be $\mathbf{q}_0(z, w) = \mathbf{q}_0(z)\mathbf{q}_0(w)$, $\mathbf{q}_1(z, w) = \mathbf{q}_1(z)\mathbf{q}_1(w)$, and $\mathbf{r}(z, w) = \mathbf{r}(z)\mathbf{r}(w)$ gives

$$\mathbb{I}_{\mathbf{r}(z)\mathbf{r}(w)}[\mathbf{q}_1(z)\mathbf{q}_1(w) \parallel \mathbf{q}_0(z)\mathbf{q}_0(w)] = \mathbb{I}_{\mathbf{r}(z)}[\mathbf{q}_1(z) \parallel \mathbf{q}_0(z)] + \mathbb{I}_{\mathbf{r}(w)}[\mathbf{q}_1(w) \parallel \mathbf{q}_0(w)].$$

Postulate 3. If belief does not change then no information is gained, regardless of the view of expectation,

$$\mathbb{I}_{\mathbf{r}(z)}[\mathbf{q}_0(z) \parallel \mathbf{q}_0(z)] = 0.$$

Postulate 4. The information gained from any normalized prior state of belief $\mathbf{q}_0(z)$ to an updated state of belief $\mathbf{r}(z)$ in the view of $\mathbf{r}(z)$ must be nonnegative

$$\mathbb{I}_{\mathbf{r}(z)}[\mathbf{r}(z) \parallel \mathbf{q}_0(z)] \geq 0.$$

The first postulate requires information to be reassessed as belief changes. The most justified state of belief, based on the entirety of observations, will correspond to the most justified view of information. The second postulate is the additive form of Shore and Johnson's Axiom 3, system independence. That is, we need some law of composition, addition in this case, that allows independent random variables to be treated separately and arrive at the same result as treating them jointly.

By combining the first two postulates, it is possible to show that $f(r, q, p) = \log(r^\gamma q^\alpha p^\beta)$ for constants α, β, γ . See Appendix A for details. The third postulate constrains these exponential constants and the fourth simply sets the sign of information.

3.3. Principal Result

Theorem 1. Information as a Measure of Change in Belief. Information measurements that satisfy these postulates must take the form

$$\mathbb{I}_{\mathbf{r}(z)}[\mathbf{q}_1(z) \parallel \mathbf{q}_0(z)] = \alpha \int dz \mathbf{r}(z) \log \left(\frac{\mathbf{q}_1(z)}{\mathbf{q}_0(z)} \right) \quad \text{for some } \alpha > 0.$$

Proof is given in Appendix A. As Shannon notes regarding entropy, α corresponds to a choice of units. Typical choices are natural units $\alpha = 1$ and bits $\alpha = \log(2)^{-1}$. We employ natural units in analysis and bits in experiments.

Although it would be possible to combine Postulate 1 and Postulate 2 into an analog of Shannon's chain rule as a single postulate, doing so would obscure the reasoning behind the construction. We leave the analogous chain rule as a consequence in Corollary 1. Regarding Shannon's proof that entropy is the only construction that satisfies properties he provides, we observe that he has restricted attention to functionals acting upon a single distribution. The interpretation of entropy is discussed in Section 4.1.

Normalization of $\mathbf{r}(z)$ is a key property of rational belief and reasonable expectation. As for $\mathbf{q}_0(z)$ and $\mathbf{q}_1(z)$, however, nothing postulated prevents analysis respecting improper or non-normalizable probability distributions. In the Bayesian context, such distributions merely represent relative plausibility among subsets of outcomes. We caution that such analysis is a further abstraction, which requires additional care for consistent interpretation.

We remark that although R enyi and Csisz ar were able to generalize divergence measures by weakening Shannon’s chain rule to independent additivity, inclusion of the first postulate prevents such generalizations. We suspect, however, that if we replace Postulate 1 with an alternative functional that incorporates rational belief into information measurements, or we replace Postulate 2 with an alternative formulation of system independence, then other compelling information theories would follow.

3.4. Regarding the Support of Expectation

The proof given assumes $\mathbf{q}_0(z)$ and $\mathbf{q}_1(z)$ take positive values over the support of the integral, which is also the support of $\mathbf{r}(z)$. In the Bayesian context, we also have

$$\mathbf{q}_1(z) = \frac{\mathbf{p}(x|z)\mathbf{q}_0(z)}{\mathbf{p}(x)} \quad \text{and} \quad \mathbf{r}(z) = \frac{\mathbf{p}(y|z,x)\mathbf{q}_1(z)}{\mathbf{p}(y|x)}.$$

Accordingly, if for some \check{z} we have $\mathbf{q}_1(\check{z}) = 0$, it follows that $\mathbf{r}(\check{z}) = 0$. Likewise, $\mathbf{q}_0(\check{z}) = 0$ would imply both $\mathbf{q}_1(\check{z}) = 0$ and $\mathbf{r}(\check{z}) = 0$. This forbids information contributions that fall beyond the scope of the proof. Even so, the resulting form is analytic and admits analytic continuation.

Since both $\lim_{\varepsilon \rightarrow 0} [\varepsilon \log \varepsilon] = 0$ and $\lim_{\varepsilon \rightarrow 0} [\varepsilon \log \varepsilon^{-1}] = 0$, limits of information of the form

$$\lim_{\varepsilon \rightarrow 0} \varepsilon \log \left(\frac{q_1}{q_0} \right), \quad \lim_{\varepsilon \rightarrow 0} \varepsilon \log \left(\frac{q_1}{\varepsilon} \right), \quad \lim_{\varepsilon \rightarrow 0} \varepsilon \log \left(\frac{\varepsilon}{q_0} \right), \quad \text{and} \quad \lim_{\varepsilon \rightarrow 0} \varepsilon \log \left(\frac{\varepsilon}{\varepsilon} \right)$$

are consistent with restricting the domain of integration (or summation) to the support of $\mathbf{r}(z)$. We gain further insight by considering limits of the form

$$\lim_{\varepsilon \rightarrow 0} r \log \left(\frac{\varepsilon}{q} \right) \quad \text{and} \quad \lim_{\varepsilon \rightarrow 0} r \log \left(\frac{q}{\varepsilon} \right).$$

Information diverges to $-\infty$ in the first case and $+\infty$ in the second. This is consistent with the fact that no finite amount of data will recover belief over a subset that has been strictly forbidden from consideration, which bears ramifications for how we understand rational belief.

If belief is not subject to influence from evidence, it is difficult to credibly construe an inferred outcome as having rationally accounted for that evidence. Lindley calls this Cromwell’s rule [38]; we should not eliminate a potential outcome from consideration unless it is logically false. The principle of insufficient reason goes further by avoiding unjustified creation of information that is not influenced by evidence.

3.5. Information Density

The Radon–Nikodym theorem [39] formalizes the notion of density that relates two measures. If we assign both probability and a second measure to any subset within a probability space, then there exists a density function, unique up to subsets of measure zero, such that the second measure is equivalent to the integral of said density over any subset.

Definition 1. Information Density. We take the Radon–Nikodym derivative to obtain information density of the change in belief from $\mathbf{q}_0(z)$ to $\mathbf{q}_1(z)$.

$$\mathbb{D}[\mathbf{q}_1(z) \parallel \mathbf{q}_0(z)] = \frac{d\mathbb{I}_{\mathbf{r}(z)}[\mathbf{q}_1(z) \parallel \mathbf{q}_0(z)]}{d\mathbf{r}(z)} = \log\left(\frac{\mathbf{q}_1(z)}{\mathbf{q}_0(z)}\right).$$

The key property we find in this construction is independence from the view of expectation. As such, information density encodes all potential information outcomes one could obtain from this theory. Furthermore, this formulation is amenable to analysis of improper distributions. For example, it proves useful to consider information density corresponding to constant unit probability density $\mathbf{q}_1(z) \equiv 1$, which is discussed further in Section 4.1.

3.6. Information Pseudometrics

The following pseudometrics admit interpretations as notions of distance between belief states that remain compatible with Postulate 1. This is achieved by simply taking the view of expectation $\mathbf{r}(z)$ to be the weight function in weighted- L^p norms of information density. These constructions then satisfy useful properties of pseudometrics:

1. *Positivity*, $\mathbb{I}_{\mathbf{r}(z)}^p[\mathbf{q}_1(z) \parallel \mathbf{q}_0(z)] \geq 0$;
2. *Symmetry*, $\mathbb{I}_{\mathbf{r}(z)}^p[\mathbf{q}_1(z) \parallel \mathbf{q}_0(z)] = \mathbb{I}_{\mathbf{r}(z)}^p[\mathbf{q}_0(z) \parallel \mathbf{q}_1(z)]$;
3. *Triangle inequality*,

$$\mathbb{I}_{\mathbf{r}(z)}^p[\mathbf{q}_2(z) \parallel \mathbf{q}_0(z)] \leq \mathbb{I}_{\mathbf{r}(z)}^p[\mathbf{q}_2(z) \parallel \mathbf{q}_1(z)] + \mathbb{I}_{\mathbf{r}(z)}^p[\mathbf{q}_1(z) \parallel \mathbf{q}_0(z)].$$

Definition 2. L^p Information Pseudometrics. We may construct pseudometrics that measure distance between states of belief $\mathbf{q}_0(z)$ and $\mathbf{q}_1(z)$ with the view of expectation $\mathbf{r}(z)$, by taking weighted- L^p norms of information density where the view of expectation serves as the weight function

$$\mathbb{I}_{\mathbf{r}(z)}^p[\mathbf{q}_1(z) \parallel \mathbf{q}_0(z)] = \left(\int dz \mathbf{r}(z) \left| \log\left(\frac{\mathbf{q}_1(z)}{\mathbf{q}_0(z)}\right) \right|^p \right)^{1/p} \quad \text{for some } p \geq 1.$$

Note that taking $p = 1$ results in a pseudometric that is also a pure expectation. The *homogeneity* property of seminorms, $\|\alpha x\| = |\alpha| \|x\|$ for $\alpha \in \mathbb{R}$, implies that these constructions retain the units of measure of information density; if information density is measured in bits, these distances have units of bits as well. Symmetry is obvious from inspection, and the other properties follow by construction as seminorms. Specifically, positivity follows from the fact that $|\cdot|^p$ is a convex function for $p \geq 1$. The lower bound immediately follows from Jensen's inequality:

$$\mathbb{I}_{\mathbf{r}(z)}^p[\mathbf{q}_1(z) \parallel \mathbf{q}_0(z)] \geq \left| \mathbb{I}_{\mathbf{r}(z)}[\mathbf{q}_1(z) \parallel \mathbf{q}_0(z)] \right| \geq 0.$$

A short proof of the triangle inequality is given in Appendix B.

We observe that if $\mathbf{q}_0(z)$ and $\mathbf{q}_1(z)$ are measurably distinct over the support of $\mathbf{r}(z)$, then the measured distance must be greater than zero. We may regard states of belief $\mathbf{q}_0(z)$ and $\mathbf{q}_1(z)$ as weakly equivalent in the view of $\mathbf{r}(z)$ if their difference is immeasurable over the support of $\mathbf{r}(z)$. That is, if $\mathbf{q}_0(z)$ and $\mathbf{q}_1(z)$ only differ over subsets of outcomes that are deemed by $\mathbf{r}(z)$ to be beyond plausible consideration, then in the view of $\mathbf{r}(z)$, they are equivalent. As such, these pseudometrics could be regarded as subjective metrics in the view of $\mathbf{r}(z)$. The natural definition of information variance also satisfies the properties of a pseudometric and is easily interpreted as a standard statistical construct.

Definition 3. Information Variance. Information variance between belief states $\mathbf{q}_0(z)$ and $\mathbf{q}_1(z)$ in the view of expectation $\mathbf{r}(z)$ is simply the variance of information density

$$\text{Var}_{\mathbf{r}(z)}[\mathbf{q}_1(z) \parallel \mathbf{q}_0(z)] = \int dz \mathbf{r}(z) \left(\log \left(\frac{\mathbf{q}_1(z)}{\mathbf{q}_0(z)} \right) - \varphi \right)^2,$$

where $\varphi = \mathbb{I}_{\mathbf{r}(z)}[\mathbf{q}_1(z) \parallel \mathbf{q}_0(z)]$.

4. Corollaries and Interpretations

The following corollaries examine primary consequences of Theorem 1. Note that multiple random variables may be expressed as a single joint variable such as $z = (z_1, z_2, \dots, z_n)$. The following corollaries explore one or two components at a time, such as variables z_1 and z_2 or observations x and y . Extensions to multiple random variables easily follow.

Note that the standard formulation of conditional dependence holds for all probability distributions in Corollary 1. That is, given an arbitrary joint distribution $\mathbf{q}(z_1, z_2)$, we can compute the marginalization as $\mathbf{q}(z_1) \equiv \int dz_2 \mathbf{q}(z_1, z_2)$ and conditional dependence follows by the Radon–Nikodym derivative to obtain $\mathbf{q}(z_2|z_1) \equiv \frac{\mathbf{q}(z_1, z_2)}{\mathbf{q}(z_1)}$. All proofs are contained in Appendix B.

Corollary 1. Chain Rule of Conditional Dependence. Information associated with joint variables decomposes as

$$\begin{aligned} \mathbb{I}_{\mathbf{r}(z_1, z_2)}[\mathbf{q}_1(z_1, z_2) \parallel \mathbf{q}_0(z_1, z_2)] &= \mathbb{I}_{\mathbf{r}(z_1)}[\mathbf{q}_1(z_1) \parallel \mathbf{q}_0(z_1)] \\ &+ \mathbb{E}_{\mathbf{r}(z_1)} \mathbb{I}_{\mathbf{r}(z_2|z_1)}[\mathbf{q}_1(z_2|z_1) \parallel \mathbf{q}_0(z_2|z_1)]. \end{aligned}$$

Corollary 2. Additivity Over Belief Sequences. Information gained over a sequence of belief updates is additive within the same view. Given initial belief $\mathbf{q}_0(z)$, intermediate states $\mathbf{q}_1(z)$ and $\mathbf{q}_2(z)$, and the view $\mathbf{r}(z)$, we have

$$\mathbb{I}_{\mathbf{r}(z)}[\mathbf{q}_2(z) \parallel \mathbf{q}_0(z)] = \mathbb{I}_{\mathbf{r}(z)}[\mathbf{q}_2(z) \parallel \mathbf{q}_1(z)] + \mathbb{I}_{\mathbf{r}(z)}[\mathbf{q}_1(z) \parallel \mathbf{q}_0(z)].$$

Corollary 3. Antisymmetry. Information from $\mathbf{q}_1(z)$ to $\mathbf{q}_0(z)$ is the negative of information from $\mathbf{q}_0(z)$ to $\mathbf{q}_1(z)$

$$\mathbb{I}_{\mathbf{r}(z)}[\mathbf{q}_0(z) \parallel \mathbf{q}_1(z)] = -\mathbb{I}_{\mathbf{r}(z)}[\mathbf{q}_1(z) \parallel \mathbf{q}_0(z)].$$

4.1. Entropy

Shannon’s formalization of entropy as uncertainty may be consistently understood as the expectation of information gained by realization. We first reconstruct information contained in realization. We then define the general form of entropy in the discrete case, which is cross entropy, and finally, the standard form of entropy follows.

Corollary 4. Realization Information (Discrete). Let z be a discrete random variable $z \in \{z_i \mid i \in [n]\}$. Information gained by realization \check{z} from $\mathbf{q}(z)$ in the view of realization $\mathbf{r}(z|\check{z})$ is

$$\mathbb{I}_{\mathbf{r}(z|\check{z})}[\mathbf{r}(z|\check{z}) \parallel \mathbf{q}(z)] = \mathbb{D}[1 \parallel \mathbf{q}(z = \check{z})].$$

Corollary 5. Cross Entropy (Discrete). Let z be a discrete random variable $z \in \{z_i \mid i \in [n]\}$ and \check{z} be a hypothetical realization. Expectation over the view $\mathbf{r}(\check{z})$ of information gained by realization from belief $\mathbf{q}(z)$ recovers cross entropy

$$\mathbb{E}_{\mathbf{r}(\check{z})} \mathbb{I}_{\mathbf{r}(z|\check{z})}[\mathbf{r}(z|\check{z}) \parallel \mathbf{q}(z)] = \mathbb{I}_{\mathbf{r}(z)}[1 \parallel \mathbf{q}(z)] = S_{\mathbf{r}(z)}[\mathbf{q}(z)].$$

Corollary 6. Entropy (Discrete). Let z be a discrete random variable $z \in \{z_i \mid i \in [n]\}$ and \check{z} be a hypothetical realization. Expectation over plausible realizations $\mathbf{q}(\check{z})$ of information gained by realization from belief $\mathbf{q}(z)$ recovers entropy

$$\mathbb{E}_{\mathbf{q}(\check{z})} \mathbb{I}_{\mathbf{r}(z|\check{z})} [1 \parallel \mathbf{q}(z)] = \mathbb{I}_{\mathbf{q}(z)} [1 \parallel \mathbf{q}(z)] = S[\mathbf{q}(z)].$$

Shannon proved that this is the only construction as a functional acting on a single distribution $\mathbf{q}(z)$ that satisfies his properties. As mentioned earlier, the information notation $\mathbb{I}_{\mathbf{q}(z)} [1 \parallel \mathbf{q}(z)]$ requires some subtlety of interpretation. Probability density 1 over all discrete outcomes $z \in \Omega_z$ is not generally normalized. Although these formulas are convenient abstractions that result from formal derivations as expectations in the discrete case, nothing prevents us from applying them in continuous settings, which recovers the typical definitions in such cases.

In the continuous setting, we must emphasize that this definition of entropy is not consistent with taking the limit of a sequence of discrete distributions that converges in probability density to a continuous limiting distribution. The entropy of such a sequence diverges to infinity, which matches our intuition; the number of bits required to specify a continuous (real) random variable also diverges.

4.2. Information in an Observation

As discussed in Section 2.3, we may regard $\mathbf{q}_0(z) \equiv \mathbf{p}(z)$ as prior belief and $\mathbf{q}_1(z) \equiv \mathbf{p}(z|x)$ as the posterior conditioned on the observation of x . Without any additional evidence, we must hold $\mathbf{r}(z) \equiv \mathbf{p}(z|x)$ to be rational belief and we recover the Kullback–Liebler divergence as the rational measure of information gained by the observation of x , but with a caveat: once we obtain additional evidence y , then information in the observation of x must be recomputed as $\mathbb{I}_{\mathbf{p}(z|x,y)} [\mathbf{p}(z|x) \parallel \mathbf{p}(z)]$. In contrast, this theory holds that Lindley’s corresponding measure,

$$D_L[\mathbf{p}(z|x) \parallel \mathbf{p}(z)] = S[\mathbf{p}(z)] - S[\mathbf{p}(z|x)],$$

is not the information gained by the observation of x ; it is simply the difference in uncertainty before and after the observation.

4.3. Potential Information

We now consider expectations over hypothetical future observations w that would influence belief in z as a latent variable. Given belief $\mathbf{p}(z)$, the probability of an observation w is $\mathbf{p}(w) = \int dz \mathbf{p}(w|z)\mathbf{p}(z)$ as usual.

Corollary 7. Consistent Future Expectation. Let the view $\mathbf{p}(z)$ express present belief in the latent variable z and w represent a future observation. The expectation over plausible w of information in the belief-shift from $\mathbf{q}_0(z)$ to $\mathbf{q}_1(z)$ in the view of rational future belief $\mathbf{p}(z|w)$ is equal to information in the present view

$$\mathbb{E}_{\mathbf{p}(w)} \mathbb{I}_{\mathbf{p}(z|w)} [\mathbf{q}_1(z) \parallel \mathbf{q}_0(z)] = \mathbb{I}_{\mathbf{p}(z)} [\mathbf{q}_1(z) \parallel \mathbf{q}_0(z)].$$

Corollary 8. Mutual Information. Let the view $\mathbf{p}(z)$ express present belief in the latent variable z and w represent a future observation. Expectation of information gained by a future observation w is mutual information

$$\mathbb{E}_{\mathbf{p}(w)} \mathbb{I}_{\mathbf{p}(z|w)} [\mathbf{p}(z|w) \parallel \mathbf{p}(z)] = \mathbb{I}_{\mathbf{p}(z,w)} [\mathbf{p}(z,w) \parallel \mathbf{p}(z)\mathbf{p}(w)].$$

Corollary 9. Realization Limit. Let z be a latent variable and \check{z} be the limit of increasing observations to obtain arbitrary precision over plausible values of z . Information gained from $\mathbf{q}_0(z)$ to $\mathbf{q}_1(z)$ in the realization limit $\mathbf{r}(z|\check{z})$ is pointwise information density

$$\mathbb{I}_{\mathbf{r}(z|\check{z})} [\mathbf{q}_1(z) \parallel \mathbf{q}_0(z)] = \mathbb{D}[\mathbf{q}_1(z = \check{z}) \parallel \mathbf{q}_0(z = \check{z})].$$

4.4. Consistent Optimization Analysis

Bernardo [40] shows that integrating entropy-like information measures with Bayesian inference provides a logical foundation for rational experimental design. He considers potential utility functions, or objectives for optimization, which are formulated as kernels of expectation over posterior belief updated by the outcome of an experiment. Bernardo then distinguishes the belief a scientist reports from belief that is justified by inference.

For a utility function to be *proper*, the Bayesian posterior must be the unique optimizer of expected utility over all potentially reported beliefs. In other words, a proper utility function must not provide an incentive to lie. His analysis shows that Lindley information is a proper utility function. Corollary 10 holds that information in this theory also provides proper utility. Thus, information measures are not simply ad hoc objectives; they facilitate consistent optimization-based analysis that recovers rational belief.

Corollary 10. Information Is a Proper Utility Function. *Taking the rational view $\mathbf{p}(z|x)$ over the latent variable z conditioned upon an experimental outcome x , the information $\mathbb{I}_{\mathbf{p}(z|x)}[\mathbf{q}(z) \parallel \mathbf{p}(z)]$ from prior belief $\mathbf{p}(z)$ to reported belief $\mathbf{q}(z)$ is a proper utility function. That is, the unique optimizer recovers rational belief*

$$\mathbf{q}^*(z) \equiv \underset{\mathbf{q}(z)}{\operatorname{argmax}} \mathbb{I}_{\mathbf{p}(z|x)}[\mathbf{q}(z) \parallel \mathbf{p}(z)] \equiv \mathbf{p}(z|x).$$

We would like to go a step further and show that when information from $\mathbf{q}_0(z)$ to $\mathbf{q}_1(z)$ is positive in the view of $\mathbf{r}(z)$, we may claim that $\mathbf{q}_1(z)$ is closer to $\mathbf{r}(z)$ than $\mathbf{q}_0(z)$. For this claim to be consistent, we must show that any perturbation that unambiguously drives belief $\mathbf{q}_1(z)$ toward the view $\mathbf{r}(z)$ must also increase information. The complementary perturbation response with respect to $\mathbf{q}_0(z)$ immediately follows by Corollary 3.

Corollary 11. Proper Perturbation Response. *Let $\mathbf{q}_1(z)$ be measurably distinct from the view $\mathbf{r}(z)$ and $\mathbb{I}_{\mathbf{r}(z)}[\mathbf{q}_1(z) \parallel \mathbf{q}_0(z)]$ be finite. Let the perturbation $\eta(z)$ preserve normalization and drive belief toward $\mathbf{r}(z)$ on all measurable subsets. It follows that*

$$\lim_{\varepsilon \rightarrow 0} \frac{\partial}{\partial \varepsilon} \mathbb{I}_{\mathbf{r}(z)}[\mathbf{q}_1(z) + \varepsilon \eta(z) \parallel \mathbf{q}_0(z)] > 0.$$

It bears repeating, by Corollary 8, that mutual information captures expected proper utility, which provides a basis for rational experimental design and feature selection.

4.5. Discrepancy Functions

Ebrahimi, Soofi, and Soyer [18] discuss information discrepancy functions, which have two key properties. First, a discrepancy function is nonnegative $\mathcal{D}[\mathbf{q}_1(z) \parallel \mathbf{q}_0(z)] \geq 0$ with equality if and only if $\mathbf{q}_1(z) \equiv \mathbf{q}_0(z)$. Second, if we hold $\mathbf{q}_0(z)$ fixed, then $\mathcal{D}[\mathbf{q}_1(z) \parallel \mathbf{q}_0(z)]$ is convex in $\mathbf{q}_1(z)$. One of the reasons information discrepancy functions are useful is that they serve to identify independence. Random variables x and z are independent if and only if $\mathbf{p}(x) \equiv \mathbf{p}(x|z)$. Therefore, we have $\mathcal{D}[\mathbf{p}(x,z) \parallel \mathbf{p}(x)\mathbf{p}(z)] \geq 0$ with equality if and only if x and z are independent, noting that $\mathbf{p}(x,z) \equiv \mathbf{p}(x|z)\mathbf{p}(z)$. This has implications regarding sensible generalizations of mutual information.

Theorem 1 does not satisfy information discrepancy properties unless the view of expectation is taken to be $\mathbf{r}(z) \equiv \mathbf{q}_1(z)$, which is the KL divergence. We note, however, that information pseudometrics and information variance given in Section 3.6 satisfy a weakened formulation. Specifically, $\mathbb{I}_{\mathbf{r}(z)}^p[\mathbf{q}_1(z) \parallel \mathbf{q}_0(z)] \geq 0$ with equality if and only if $\mathbf{q}_0(z)$ and $\mathbf{q}_1(z)$ are weakly equivalent in the view of $\mathbf{r}(z)$. Likewise, these formulations are convex in information density $\mathbb{D}[\mathbf{q}_1(z) \parallel \mathbf{q}_0(z)]$.

4.6. Jaynes Maximal Uncertainty

Jaynes uses entropy to analytically construct a unique probability distribution for which uncertainty is maximal while maintaining consistency with a specified set of expectations. This construction avoids unjustified creation of information and places the principle of insufficient reason into an analytic framework within which the notion of symmetry generalizes to informational symmetries conditioned upon observed expectations.

We review how Jaynes constructs the resulting distribution $\mathbf{r}^*(z)$. Let such kernels of expectation be denoted $f_i(z)$ for $i \in [n]$ and the observed expectations be $\mathbb{E}_{\mathbf{r}(z)}[f_i(z)] = \varphi_i$. The objective of optimization is

$$\mathbf{r}^*(z) = \underset{\mathbf{r}(z)}{\operatorname{argmax}} S[\mathbf{r}(z)] \quad \text{subject to} \quad \mathbb{E}_{\mathbf{r}(z)} f_i(z) = \varphi_i \quad \forall i \in [n].$$

The Lagrangian, which captures both the uncertainty objective and expectation constraints, is

$$\mathcal{L}[\mathbf{r}(z), \lambda] = \int dz \mathbf{r}(z) \left(\log \left(\frac{1}{\mathbf{r}(z)} \right) - \sum_{i=1}^n \lambda_i (f_i(z) - \varphi_i) \right),$$

where $\lambda \in \mathbb{R}^n$ is the vector of Lagrange multipliers. This Lagrangian formulation satisfies the variational principle in both $\mathbf{r}(z)$ and λ . Variational analysis yields the optimizer

$$\mathbf{r}^*(z) \propto \exp \left(\sum_{i=1}^n \lambda_i f_i(z) \right).$$

Information-Critical Distributions

Rather than maximizing entropy, we may minimize $\mathbb{I}_{\mathbf{r}(z)}[\mathbf{r}(z) \parallel \mathbf{q}_0(z)]$ while maintaining consistency with specified expectations. Since the following corollary holds for general distributions $\mathbf{q}_0(z)$, including the improper case $\mathbf{q}_0(z) \equiv 1$, this includes Jaynes' maximal uncertainty as a minimization of negative entropy.

Corollary 12. Minimal Information. *Given kernels of expectation $f_i(z)$ and specified expectations $\mathbb{E}_{\mathbf{r}(z)}[f_i(z)] = \varphi_i$ for $i \in [n]$, the distribution $\mathbf{r}^*(z)$ that satisfies these constraints while minimizing information $\mathbb{I}_{\mathbf{r}(z)}[\mathbf{r}(z) \parallel \mathbf{q}_0(z)]$ is given by*

$$\mathbf{r}^*(z) \propto \mathbf{q}_0(z) \exp \left(\sum_{i=1}^n \lambda_i f_i(z) \right) \quad \text{for some } \lambda \in \mathbb{R}^n.$$

4.7. Remarks on Fisher Information

Fisher provides an analytic framework to assess the suitability of a pointwise latent description of a probability distribution [41]. As Kullback and Leibler note, the functional properties of information in Fisher's construction are quite different from Shannon's, and thus, we do not regard Fisher information as a form of entropic information. Fisher's construction, however, can be rederived and understood within this theory. He begins with the assumption that there is some latent realization \check{z} for which $\mathbf{p}(x|\check{z})$ is an exact description of the true distribution of x . We can then define the *Fisher score* as the gradient of information from any independent prior belief $\mathbf{q}_0(x)$ to a pointwise latent description $\mathbf{p}(x|z)$, in the view $\mathbf{p}(x|\check{z})$

$$f = \nabla_z \mathbb{I}_{\mathbf{p}(x|\check{z})}[\mathbf{p}(x|z) \parallel \mathbf{q}_0(x)].$$

Note that \check{z} is fixed by assumption, despite remaining unknown. By the variational principle, the score must vanish at the optimizer z^* . By Corollary 10, the optimizer must be $z^* = \check{z}$. We can then assess

the sensitivity of information to the parameter z at the optimizer z^* by computing the Hessian. This recovers an equivalent construction of the Fisher matrix within this theory:

$$F_{ij} = \frac{\partial^2}{\partial z_i \partial z_j} \mathbb{I}_{\mathbf{p}(x|z)}[\mathbf{p}(x|z) \parallel \mathbf{q}_0(x)].$$

The primary idea behind this construction is that high-curvature in z implies that a pointwise description is both suitable and a well-conditioned optimization problem.

Generalized Fisher Matrix

We may eliminate the assumption of an exact pointwise description and generalize analogous formulations to arbitrary views of expectation.

Definition 4. Generalized Fisher Score. Let $\mathbf{r}(x)$ be the view of expectation regarding an observable x . The gradient with respect to z of information from independent prior belief $\mathbf{q}_0(x)$ to a pointwise description $\mathbf{p}(x|z)$ gives the score

$$f = \nabla_z \mathbb{I}_{\mathbf{r}(x)}[\mathbf{p}(x|z) \parallel \mathbf{q}_0(x)].$$

Definition 5. Generalized Fisher Matrix. Let $\mathbf{r}(x)$ be the view of expectation regarding an observable x . The Hessian matrix with respect to components of z of information from independent prior belief $\mathbf{q}_0(x)$ to the pointwise description $\mathbf{p}(x|z)$ gives the generalized Fisher matrix

$$F_{ij} = \frac{\partial^2}{\partial z_i \partial z_j} \mathbb{I}_{\mathbf{r}(x)}[\mathbf{p}(x|z) \parallel \mathbf{q}_0(x)].$$

Again, a local optimizer z^* must satisfy the variational principle and yield a score of zero. The generalized Fisher matrix would typically be evaluated at such an optimizer z^* .

5. Information in Inference and Machine Learning

We now examine model information and predictive information provided by inference. Once we have defined these information measurements, we derive upper and lower bounds between them that we anticipate being useful for future work. Finally, we show how inference information may be constrained, which addresses some challenges in Bayesian inference.

5.1. Machine Learning Information

Akaike [42] first introduced information-based complexity criteria as a strategy for model selection. These ideas were further developed by Schwarz, Burnham, and Gelman [19,43,44]. We anticipate these notions will prove useful in future work to both understand and control the problem of memorization in machine learning training. Accordingly, we discuss how this theory views model complexity and distinguishes formulations of predictive information and residual information.

In machine learning, observations correspond to matched pairs of inputs and labels $Y = \{(x^{(j)}, y^{(j)}) \mid j \in [T]\}$. For each sample j of T training examples, we would like to map the input $x^{(j)}$ to an output label $y^{(j)}$. Latent variables θ are unknown model parameters from a specified model family or computational structure. A *model* refers to a specific parameter state and the predictions that the model computes are $\mathbf{p}(y^{(j)}|x^{(j)}, \theta)$. Since the definitions and derivations that follow hold with respect to either single cases or the entire training ensemble, we will use shorthand notation $\mathbf{p}(y|\theta)$ to refer to both scenarios.

We denote the initial state of belief in model parameters as $\mathbf{q}_0(\theta)$ and update belief during training as $\mathbf{q}_i(\theta)$ for $i \in [n]$. We can then compute predictions from any state of model belief by marginalization $\mathbf{q}_i(y) \equiv \int d\theta \mathbf{p}(y|\theta) \mathbf{q}_i(\theta)$.

Definition 6. Model Information. Model information from initial belief $\mathbf{q}_0(\theta)$ to updated belief $\mathbf{q}_i(\theta)$ in the view of $\mathbf{r}(\theta)$ is given by

$$\mathbb{I}_{\mathbf{r}(\theta)}[\mathbf{q}_i(\theta) \parallel \mathbf{q}_0(\theta)] \quad \text{for } i \in [n].$$

When we compute information contained in training labels, the label data obviously provide the rational view. This is represented succinctly by $\mathbf{r}(y|\tilde{y})$, which assigns full probability to specified outcomes. Again, if we need to be explicit, then this could be written as $\mathbf{r}(y|x^{(i)}, y^{(j)})$ for each case in the training set.

Definition 7. Predictive Label Information. The realization of training labels is the rational view $\mathbf{r}(y|\tilde{y})$ of label plausibility. We compute information from prior predictive belief $\mathbf{q}_0(y)$ to predictive belief $\mathbf{q}_i(y)$ in this view as

$$\mathbb{I}_{\mathbf{r}(y|\tilde{y})}[\mathbf{q}_i(y) \parallel \mathbf{q}_0(y)].$$

In the continuous setting, this formulation is closely related to log pointwise predictive density [19]. We can also define complementary label information that is not contained in the predictive model.

Definition 8. Residual Label Information (Discrete). Residual information in the label realization view $\mathbf{r}(y|\tilde{y})$ is computed as

$$\mathbb{I}_{\mathbf{r}(y|\tilde{y})}[\mathbf{r}(y|\tilde{y}) \parallel \mathbf{q}_i(y)].$$

Residual information is equivalent to cross-entropy if the labels are full realizations. We note, however, that if training labels are probabilistic and leave some uncertainty, then replacing both occurrences of $\mathbf{r}(y|\tilde{y})$ above with a general distribution $\mathbf{r}(y)$ would correctly calibrate residual information, so that if predictions were to match label distributions, then residual information would be zero.

As a consequence of Corollary 2, the sum of predictive label information and residual label information is always constant. This allows us to rigorously frame predictive label information as a fraction of the total information contained in training labels. Moreover, Corollary 11 assures us that model perturbations that drive predictive belief toward the label view must increase predictive information. In the continuous setting, just as the limiting form of entropy discussed in Section 4.1 diverges, so too does residual information diverge. Predictive label information, however, remains a finite alternative. This satisfies our initial incentive for this investigation.

There is a second type of predictive information we may rationally construct, however. Rather than considering predictive information with respect to specified label outcomes, we might be interested in the information we expect to obtain about new samples from the generative process. If we regard marginalized predictions $\mathbf{q}_i(y)$ as our best approximation of this process, then we would simply measure change in predictive belief in this view.

Definition 9. Predictive Generative Approximation. We may approximate the distribution of new outcomes from model belief $\mathbf{q}_i(\theta)$ using the predictive marginalization $\mathbf{q}_i(y) \equiv \int d\theta \mathbf{p}(y|\theta)\mathbf{q}_i(\theta)$. If we hold this to be the rational view of new outcomes from the generative process, predictive information is

$$\mathbb{I}_{\mathbf{q}_i(y)}[\mathbf{q}_i(y) \parallel \mathbf{q}_0(y)].$$

5.2. Inference Information Bounds

In Bayesian inference, we have prior belief in model parameters $\mathbf{q}_0(\theta) \equiv \mathbf{p}(\theta)$ and the posterior inferred from training data $\mathbf{q}_1(\theta) \equiv \mathbf{p}(\theta|\tilde{y})$. The predictive marginalizations are called the *prior predictive* and *posterior predictive* distributions respectively:

$$\mathbf{p}(y) \equiv \int d\theta \mathbf{p}(y|\theta)\mathbf{p}(\theta) \quad \text{and} \quad \mathbf{p}(y|\tilde{y}) \equiv \int d\theta \mathbf{p}(y|\theta)\mathbf{p}(\theta|\tilde{y}).$$

We can derive inference information bounds for Bayesian networks [45]. Let y , θ_1 , and θ_2 represent a directed graph of latent variables. In general, the joint distribution can always be written as $\mathbf{p}(y, \theta_1, \theta_2) = \mathbf{p}(\theta_2|\theta_1, y)\mathbf{p}(\theta_1|y)\mathbf{p}(y)$. The property of *local conditionality* [46] means $\mathbf{p}(\theta_2|\theta_1, \check{y}) \equiv \mathbf{p}(\theta_2|\theta_1)$. That is, belief dependence in θ_2 is totally determined by that of θ_1 just as belief in θ_1 is computed from \check{y} .

Corollary 13. Joint Local Inference Information. *Inference information in θ_1 gained by having observed \check{y} is equivalent to the inference information in both θ_1 and θ_2 .*

$$\mathbb{I}_{\mathbf{p}(\theta_1, \theta_2|\check{y})}[\mathbf{p}(\theta_1, \theta_2|\check{y}) \parallel \mathbf{p}(\theta_1, \theta_2)] = \mathbb{I}_{\mathbf{p}(\theta_1|\check{y})}[\mathbf{p}(\theta_1|\check{y}) \parallel \mathbf{p}(\theta_1)].$$

Corollary 14. Monotonically Decreasing Local Inference Information. *Inference information in θ_2 gained by having observed \check{y} is bound above by inference information in θ_1 .*

$$\mathbb{I}_{\mathbf{p}(\theta_2|\check{y})}[\mathbf{p}(\theta_2|\check{y}) \parallel \mathbf{p}(\theta_2)] \leq \mathbb{I}_{\mathbf{p}(\theta_1|\check{y})}[\mathbf{p}(\theta_1|\check{y}) \parallel \mathbf{p}(\theta_1)].$$

This shows that an inference yields nonincreasing information as we compound the inference with locally conditioned latent variables, which is relevant for sequential predictive computational models, such as neural networks. We observe that the inference sequence from training data \check{y} to model parameters θ , to new predictions y , is also a locally conditioned sequence. If belief in a given latent variable is represented as a probability distribution, this places bounds on what transformations are compatible with the progression of information. For example, accuracy measures which snap the maximum probability outcome of a neural network to unit probability impose an unjustified creation of information.

Corollary 15. Inferred Information Upper Bound. *Model information in the posterior view is less than or equal to predictive label information resulting from inference*

$$\mathbb{I}_{\mathbf{p}(\theta|\check{y})}[\mathbf{p}(\theta|\check{y}) \parallel \mathbf{p}(\theta)] \leq \mathbb{I}_{\mathbf{r}(y|\check{y})}[\mathbf{p}(y|\check{y}) \parallel \mathbf{p}(y)].$$

This is noteworthy because it tells us that inference always yields a favorable tradeoff between increased model complexity and predictive information. Combining Corollary 14 and Corollary 15, we have upper and lower bounds on model information due to inference

$$\mathbb{I}_{\mathbf{p}(y|\check{y})}[\mathbf{p}(y|\check{y}) \parallel \mathbf{p}(y)] \leq \mathbb{I}_{\mathbf{p}(\theta|\check{y})}[\mathbf{p}(\theta|\check{y}) \parallel \mathbf{p}(\theta)] \leq \mathbb{I}_{\mathbf{r}(y|\check{y})}[\mathbf{p}(y|\check{y}) \parallel \mathbf{p}(y)].$$

5.3. Inference Information Constraints

Practitioners of Bayesian inference often struggle when faced with inference problems for models structures that are not well suited to the data. An under-expressive model family is not capable of representing the process being modeled. As a consequence, the posterior collapses to a small set of outcomes that are least inconsistent with the evidence. In contrast, an over-expressive model admits multiple sufficient explanations of the process.

Both model and predictive information measures offer means to understand and address these challenges. By constraining the information gained by inference, we may solve problems associated with model complexity. In this section, we discuss explicit and implicit approaches to enforcing such constraints.

5.3.1. Explicit Information Constraints

Our first approach to encode information constraints is to explicitly solve a distribution that satisfies expected information gained from the prior to the posterior. We examine how information-critical distributions can be constructed from arbitrary states of belief $\mathbf{q}_i(\theta)$ for $i \in [n]$.

Again, we may obtain critical distributions with respect to uncertainty by simply setting $\mathbf{q}_0(\theta) \equiv 1$. By applying this to an inference, so that $n = 1$ and $\mathbf{q}_1(\theta) \equiv \mathbf{p}(\theta|y)$, we recover likelihood annealing as a means to control model information.

Corollary 16. Constrained Information. *Given states of belief $\mathbf{q}_i(\theta)$ and information constraints $\mathbb{I}_{\mathbf{r}(\theta)}[\mathbf{q}_i(\theta) \parallel \mathbf{q}_0(\theta)] = \varphi_i$ for $i \in [n]$, the distribution $\mathbf{r}^*(\theta)$ that satisfies these constraints while minimizing $\mathbb{I}_{\mathbf{r}(\theta)}[\mathbf{r}(\theta) \parallel \mathbf{q}_0(\theta)]$ has the form*

$$\mathbf{r}^*(\theta) \propto \mathbf{q}_0(\theta) \prod_{i=1}^n \left(\frac{\mathbf{q}_i(\theta)}{\mathbf{q}_0(\theta)} \right)^{\lambda_i} \quad \text{for some } \lambda \in \mathbb{R}^n.$$

Corollary 17. Information-Annealed Inference. *Annealed belief $\mathbf{r}(\theta)$ for which information gained from prior to posterior belief is fixed $\mathbb{I}_{\mathbf{r}(\theta)}[\mathbf{p}(\theta|\check{y}) \parallel \mathbf{p}(\theta)] = \varphi$ and information $\mathbb{I}_{\mathbf{r}(\theta)}[\mathbf{r}(\theta) \parallel \mathbf{p}(\theta)]$ is minimal must take the form*

$$\mathbf{r}(\theta) \propto \mathbf{p}(\check{y}|\theta)^\lambda \mathbf{p}(\theta) \quad \text{for some } \lambda \in \mathbb{R}.$$

Note that the bounds in Section 5.2 still apply if we simply include λ as a fixed model parameter in the definition of the likelihood function so that $\mathbf{p}(\check{y}|\theta) \mapsto \mathbf{p}(\check{y}|\theta, \lambda) \equiv \mathbf{p}(\check{y}|\theta)^\lambda$. This prevents the model from learning too much, which may be useful for under-expressive models or for smoothing out the posterior distribution to aid exploration during learning.

5.3.2. Implicit Information Constraints

Our second approach introduces hyper-parameters, λ and ψ , into the Bayesian inference problem, which allows us to define a prior on those hyper-parameters that implicitly encodes information constraints. This approach gives us a way to express how much we believe we can learn from the data and model that we have in hand. Doing so may prevent overconfidence when there are known modeling inadequacies or underconfidence from overly broad priors.

As above, λ parameters influence the likelihood and can be thought of as controlling annealing or an embedded stochastic error model. The ψ parameters control the prior on the model parameter θ . For example, these parameters could be the prior mean and co-variance if we assume a Gaussian prior distribution. Therefore, the inference problem takes the form

$$\mathbf{p}(\theta, \lambda, \psi|\check{y}) = \frac{\mathbf{p}(\check{y}|\theta, \lambda)\mathbf{p}(\theta|\psi)\mathbf{p}(\lambda, \psi)}{\mathbf{p}(\check{y})}.$$

In order to encode the information constraints, we must construct the hyper-prior distribution $\mathbf{p}(\lambda, \psi) = g(\varphi_\theta, \varphi_y)$ where

$$\begin{aligned} \varphi_\theta &= \mathbb{I}_{\mathbf{p}(\theta|\check{y}, \lambda, \psi)}[\mathbf{p}(\theta|\check{y}, \lambda, \psi) \parallel \mathbf{p}(\theta|\psi)] \quad \text{and} \\ \varphi_y &= \mathbb{I}_{\mathbf{r}(y|\check{y})}[\mathbf{p}(y|\check{y}, \lambda, \psi) \parallel \mathbf{p}(y|\lambda, \psi)] \end{aligned}$$

control model information and predictive information, respectively. The function $g(\varphi_\theta, \varphi_y)$ is the likelihood of λ and ψ given the specified model and prediction complexities. For example, this could be an indicator function as to whether the information gains are within some range. Note that we may also consider other forms of predictive information, such as the predictive generative approximation.

The posterior distribution on model parameters and posterior predictive distribution can be formed by marginalizing over hyper-parameters

$$\begin{aligned} \mathbf{p}(\theta|\check{y}) &= \int d\lambda d\psi \mathbf{p}(\theta, \lambda, \psi|\check{y}) \quad \text{and} \\ \mathbf{p}(y|\check{y}) &= \int d\lambda d\psi \mathbf{p}(y|\theta, \lambda)\mathbf{p}(\theta, \lambda, \psi|\check{y}). \end{aligned}$$

6. Negative Information

The possibility of negative information is a unique property of this theory in contrast to divergence measures such as Kullback–Leibler divergence, α -divergences, and f -divergences. It provides an easily interpreted notion of whether a belief update is consistent with our best understanding. Negative information can be consistently associated with misinformation in the view of rational belief. That is, if $\mathbb{I}_{\mathbf{r}(z)}[\mathbf{q}_1(z) \parallel \mathbf{q}_0(z)]$ is negative, then $\mathbf{q}_0(z)$ is a better approximation of rational belief than $\mathbf{q}_1(z)$. The consistency of this interpretation would be violated if we could construct $\mathbf{q}_1(z)$ that is unambiguously better than $\mathbf{q}_0(z)$ at approximating $\mathbf{r}(z)$. Corollary 11 shows, however, that this is not possible. If we construct $\mathbf{q}_1(z)$ by integrating perturbations from $\mathbf{q}_0(z)$ that drive belief towards $\mathbf{r}(z)$ on all measurable subsets, then information must be positive. The following experiments illustrate examples of negative information and motivate its utility.

6.1. Negative Information in Continuous Inference

In the following set of experiments we have a latent variable $\theta \in \mathbb{R}^2$, which is distributed as $\mathcal{N}(\theta|0, I)$. Each sample $y^{(j)} \in \mathbb{R}^2$ corresponds to realization of an independent latent variable $x^{(j)} \in \mathbb{R}^2$ so that $y^{(j)} = \theta + x^{(j)}$. Each $x^{(j)}$ is distributed as $\mathcal{N}(x^{(j)}|0, \sigma_1^2 I)$ where $\sigma_1 = 1/2$. Both prior belief in plausible values of θ and prior predictive belief in plausible values of y are visualized in Figure 4. Deciles separate annuli of probability 1/10. The model information we expect to gain by observing 10 samples of y , which is also mutual information from Corollary 8, is $\mathbb{I}_{\mathbf{p}(y, \theta)}[\mathbf{p}(y, \theta) \parallel \mathbf{p}(y)\mathbf{p}(\theta)] = 5.36$ bits. See Appendix C for details.

The first observation consists of 10 samples of y followed by inference of θ . Subsequent observations each add another 10, 20, and 40 samples respectively. A typical inference sequence is shown in Figure 5. Model information gained by inference from the first observation in the same view is 5.72 bits. As additional observations become available the model information provided by first inference is eventually refined to 5.10 bits. Typically, the region of plausible models θ resulting from each inference is consistent with what was previously considered plausible.

By running one million independent experiments, we constructed a histogram of the model information provided by first inference in subsequent views. That is shown in Figure 6. As a consequence of Postulate 4, the model information provided by first inference must always be positive before any additional observations are made. The change in model covariance in this experiment provides a stronger lower bound of 3.95 bits after first inference, which can be seen in the first view on the left. This bound is saturated in the limit when the inferred mean is unchanged. Additional observations may indicate that the first inference was less informative than initially believed. We may regard the rare cases showing negative information as being misinformed after first inference. The true value of the model θ may be known to arbitrary precision if we collect enough observations. Said value is the realization limit on the right. Under this experimental design, this limit converges to the Laplace distribution centered at mutual information $\mathcal{L}(\mu, (\log 2)^{-1})$ computed in the prior view.

From these million experiments, we can select the most unusual cases for which the information provided by first inference is later found to an extreme. Figure 7 visualizes the experiment for which the information provided by first inference is found to be the minimum after observing 160 total samples from the generative process. Although model information assessed following the first observation is a fairly typical value, additional samples quickly show that the first samples were unusual. This becomes highly apparent in the fourth view, which includes 80 samples in total.

Figure 8 visualizes the complementary case in which we select the experiment for which the information provided by first inference is later found to be the maximum. The explanatory characteristic of this experiment is the rare value that the true model has taken. High information in inference shows a high degree of unexpected content, given what the prior distribution deemed plausible. Each inference indicates a range of plausible values of θ that is quite distant from the plausible region indicated by prior belief. The change in belief due to first inference is confirmed by additional data in fourth inference.

Finally, we examine a scenario in which the first 10 samples are generated from a different process than subsequent samples. We proceed with inference as before and assume a single generative process, despite the fact that this assumption is actually false. Figure 9 shows the resulting inference sequence. After first inference, nothing appears unusual because there is no data that would contradict inferred belief. As soon as additional data become available, however, information in first inference becomes conspicuously negative. Note that the *one-in-a-million* genuine experiment exhibiting minimum information, Figure 7, gives -11.53 bits after 70 additional samples. In contrast, this experiment yields -47.91 bits after only 10 additional samples.

By comparing this result to the information distribution in the realization limit, we see that the probability of a genuine experiment exhibiting information this negative would be less than 2^{-155} . This shows how highly negative information may flag anomalous data. We explore this further in the next section.

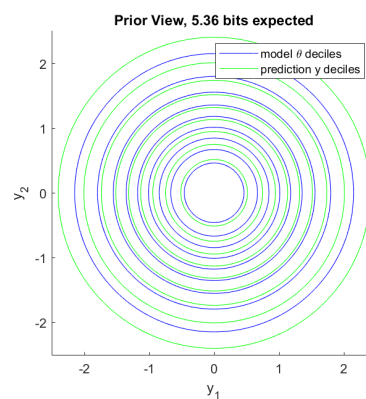


Figure 4. Prior distribution of θ and prior predictive distribution of individual y samples. The domain of plausible θ values is large before any observations are made.

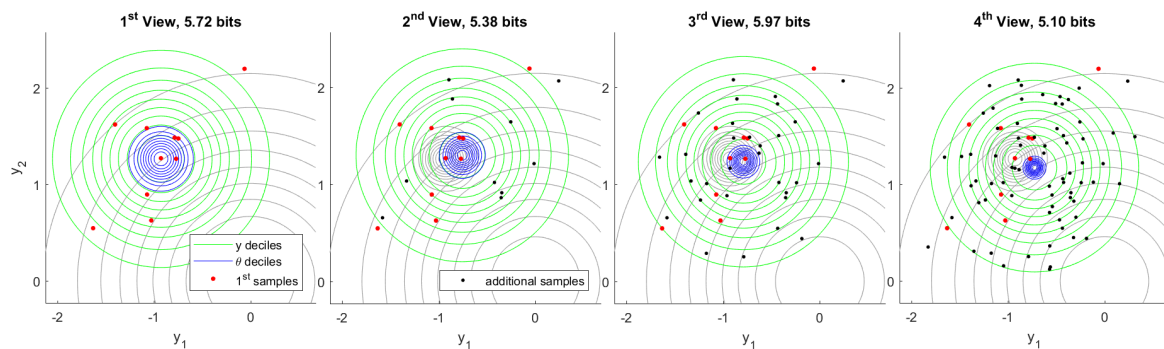


Figure 5. Typical inference of θ from observation of 10 samples of y (left) followed by 10, 20, and 40 additional samples, respectively. Both prior belief and first inference deciles of θ are shown in gray. As observations accumulate, the domain of plausible θ values tightens.

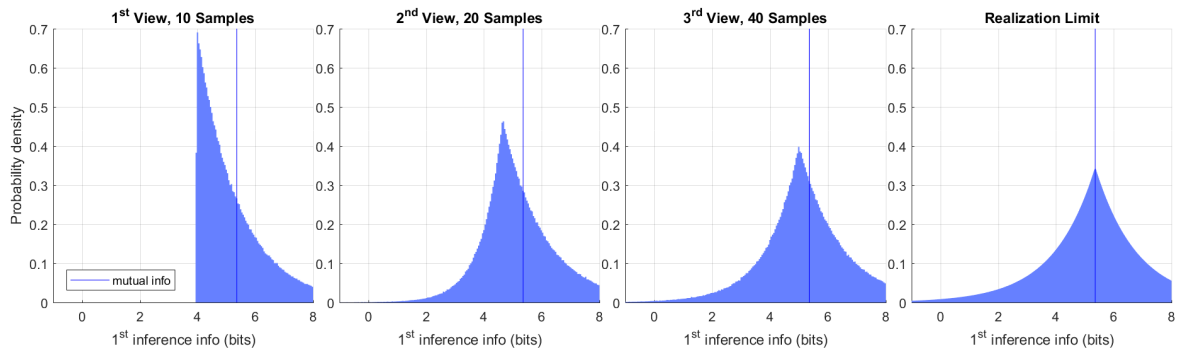


Figure 6. Histogram of the earliest inference information in the observation sequence. The vertical line at 5.56 bits is mutual information. Information is positive after first inference, but may drop with additional observations. The limiting view of infinite samples (realization) is shown on the right.

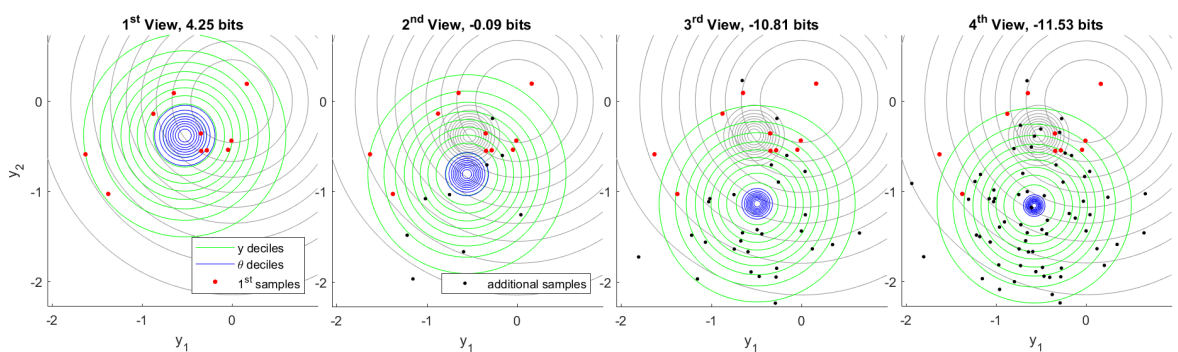


Figure 7. Minimum first inference information out of one million independent experiments. This particularly rare case shows how first samples can mislead inference, which is later corrected by additional observations. The fourth inference (right) bears remarkably little overlap with the first.

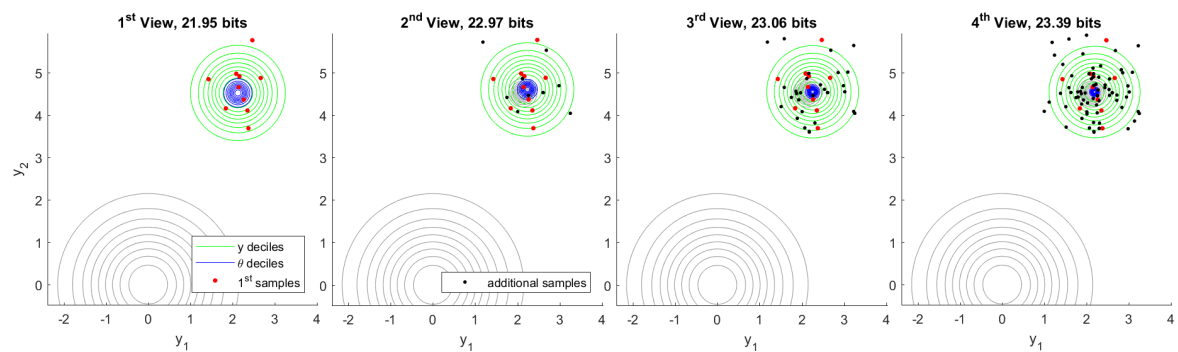


Figure 8. Maximum first inference information out of one million independent experiments. The true value of θ has taken an extremely rare value. As evidence accumulates, plausible ranges of θ confirm the first inference.

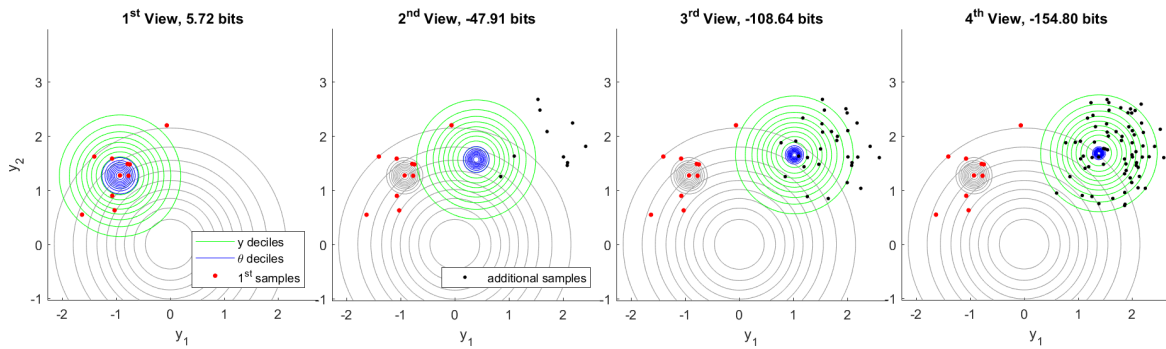


Figure 9. Inconsistent inference. The first 10 samples are drawn from a different ground truth than subsequent samples, but inference proceeds as usual. As additional data become available, first inference information becomes markedly negative.

6.2. Negative Information in MNIST Model with Mislabeled Data

We also explored predictive label information in machine learning models by constructing a small neural network to predict MNIST digits [5]. This model was trained with 50,000 images with genuine labels. Training was halted using cross-validation from 10,000 images that also had genuine labels. To investigate how predictive label information serves as an indicator of prediction accuracy, we randomly mislabeled a fraction of unseen cases. Prediction information was observed on 10,000 images for which 50% had been randomly relabeled, which resulted in 5521 original labels and 4479 mismatched labels.

The resulting distribution of information outcomes is plotted in Figure 10, which shows a dramatic difference between genuine labels and mislabeled cases. In all cases, prediction information is quantified from the uninformed probabilities $q_0(y = y_i) = 1/10$ for all outcomes $i \in [10]$ to model predictions, which are conditioned on the image input $q_1(y|x)$, in the view of the label $r(y|y_i)$. Total label information—the sum of predictive label information and residual label information:

$$\mathbb{I}_{r(y|y_i)}[r(y|y_i) \parallel q_0(y)] = \mathbb{I}_{r(y|y_i)}[q_1(y|x) \parallel q_0(y)] + \mathbb{I}_{r(y|y_i)}[r(y|y_i) \parallel q_1(y|x)] = \log_2(10) \text{ bits,}$$

or roughly 3.32 bits for each case. Both Figures 11 and 12 show different forms of anomaly detection using negative information.

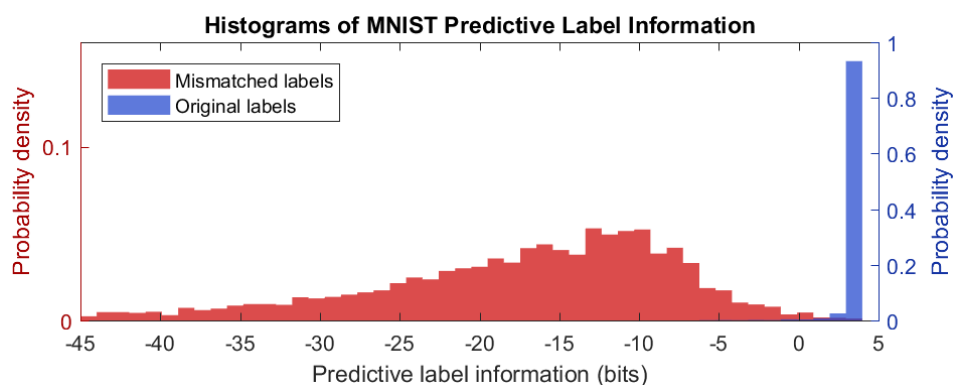


Figure 10. Histogram of information outcomes for mismatched and original labels. Correct label information is highly concentrated at 3.2 bits, which is 95.9% of the total information contained in labels. Mislabeled cases have mean information at -18 bits, and information is negative for 99.15% of mislabeled cases.

Figure 11 shows that genuine labels may exhibit negative information when predictions are poor. Only 1.1% of correct cases exhibit negative predictive label information. The distribution mean is 3.2

bits for this set. Notably, the first two images appear to be genuinely mislabeled in the original dataset, which underscores the ability of this technique to detect anomalies.

In contrast, over 99.1% of mislabeled cases exhibit negative information with the distribution mean at -18 bits. The top row of Figure 12 shows that information is most negative when the claimed label is not plausible and model predictions clearly match the image. Similarly to Section 6.1, strongly negative information indicates anomalous data. When incorrect predictions match incorrect labels, however, information can be positive, as shown in the bottom row. The cases appear to share identifiable features with the claim.

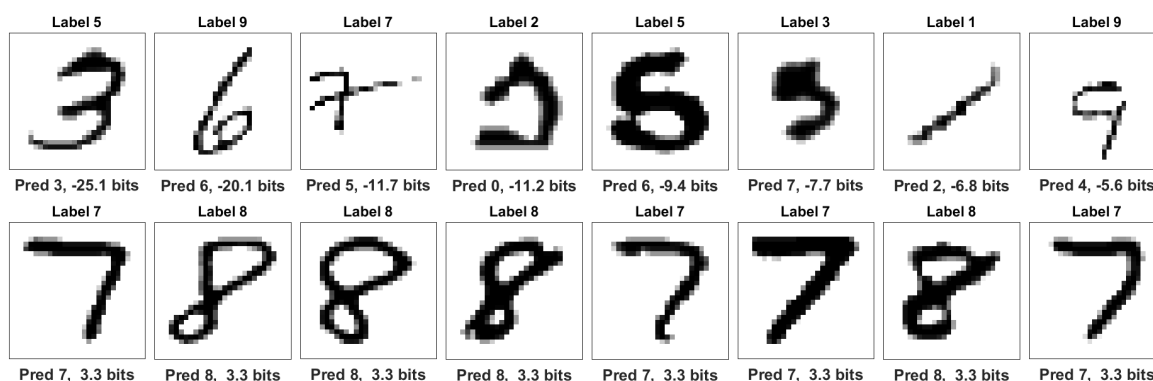


Figure 11. Original MNIST labels. The top row shows lowest predictive label information among original labels. Notably, the two leading images appear to be genuinely mislabeled in the original dataset. Subsequent predictions are poor. The bottom row shows the highest information among original labels. Labels and predictions are consistent in these cases.

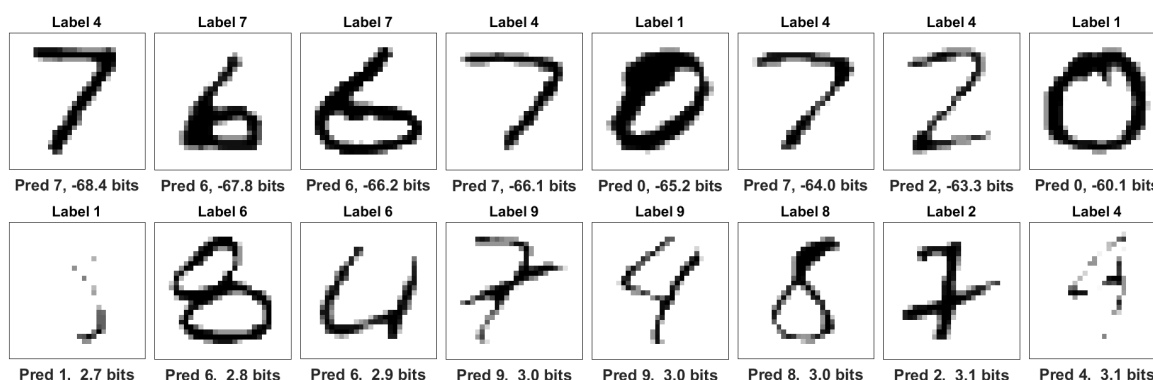


Figure 12. Mislabeled digits. The top row shows the lowest predictive label information among mislabeled cases. In each case, the claimed label is implausible and the prediction is correct. The bottom row shows the highest prediction information among mislabeled cases. Although claimed labels are incorrect, most images share identifiable features with the claim.

7. Conclusions

Just as belief matures with accumulation of evidence, we hold that the information associated with a shift in belief must also mature. By formulating principles that articulate how we may regard information as a reasonable expectation that measures change in belief, we derived a theory of information that places existing measures of entropic information in a coherent unified framework. These measures include Shannon’s original description of entropy, cross-entropy, realization information, Kullback–Leibler divergence, and Lindley information (uncertainty difference) due to an experiment.

Moreover, we found other explainable information measures that may be adapted to specific scenarios from first principles, including the log pointwise predictive measure of model accuracy.

We derived useful properties of information, including the chain rule of conditional dependence, additivity over belief updates, consistency with respect expected future observations, and expected information in future experiments as mutual information. We also showed how this theory generalizes information-critical probability distributions that are consistent with observed expectations analogous to those of Jaynes. In the context of Bayesian inference, we showed how information constraints recover and illuminate useful annealed inference practices.

We also examined the phenomenon of negative information, which occurs when a more justified point of view, based on a broader body of evidence, indicates that a previous change of belief was misleading. Experiments demonstrated that negative information reveals anomalous cases of inference or anomalous predictions in the context of machine learning.

The primary value of this theoretical framework is the consistent interpretation and corresponding properties of information that guide how it may be assessed in a given context. The property of additivity over belief updates within the present view allows us to partition information in a logically consistent manner. For machine learning algorithms, we see that total information from the uninformed state to a label-informed state is a constant that may be partitioned into the predicted component and the residual component. This insight suggests new approaches to model training, which will be the subject of continuing research.

Future Work

The challenges we seek to address with this theory relate to real-world applications of inference and machine learning. Although Bayesian inference provides a rigorous foundation for learning, poor choices of prior or likelihood can lead to results that elude or contradict human intuition when analyzed after the fact. This only becomes worse as the scale of learning problems increases, as in deep neural networks, where human intuition cannot catch inconsistencies. Information provides a metric to quantify how well a model is learning that may be useful when structuring learning problems. Some related challenges include:

1. Controlling model complexity in machine learning to avoid memorization;
2. Evaluating the influences of different experiments and data points to identify outliers or poorly supported inferences;
3. Understanding the impact of both the model-structure and fidelity of variational approximations on learnability.

Author Contributions: Conceptualization, J.A.D.; Formal Analysis, J.A.D.; Investigation, J.A.D. and T.A.C.; Software, J.A.D.; Validation, J.A.D. and T.A.C.; Visualization, J.A.D.; Writing–Original Draft, J.A.D.; Writing–Review & Editing, J.A.D. and T.A.C. All authors have read and agreed to the published version of the manuscript.

Funding: The Department of Homeland Security sponsored the production of this material under DOE Contract Number DE-NA0003525 for the management and operation of Sandia National Laboratories. This work was also funded in part by the Department of Energy Office of Advanced Scientific Computing Research.

Acknowledgments: We would like to extend our earnest appreciation to Michael Bierma, Christopher Harrison, Steven Holtzen, Philip Kegelmeyer, Jaideep Ray, and Jean-Paul Watson for helpful discussions on this topic. We also sincerely thank referees for providing expert feedback to improve this paper. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy’s National Nuclear Security Administration under contract DE-NA-0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

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

Appendix A. Proof of Principal Result

Lemma A1. Let $g(\cdot) : \mathbb{R}_+ \mapsto \mathbb{R}$ be a function such that $g(x_1 x_2) = g(x_1) + g(x_2)$ for all $x_1, x_2 > 0$. It follows that $g(x) = a \log(x)$ where a is a constant.

Proofs of Lemma A1 given by Erdős [47], Fadeev [37], Rényi [24].

Lemma A2. Let $f(\cdot, \cdot, \cdot) : \mathbb{R}_+^3 \mapsto \mathbb{R}$ be a function such that $f(r_1 r_2, q_1 q_2, p_1 p_2) = f(r_1, q_1, p_1) + f(r_2, q_2, p_2)$ for all $r_1, q_1, p_1, r_2, q_2, p_2 > 0$. It follows that $f(r, q, p) = \log(r^\gamma q^\alpha p^\beta)$.

Proof of Lemma A2. We begin by defining $g_1(x) \equiv f(x^{-1}, x, x)$, $g_2(y) \equiv f(y, y^{-1}, y)$, and $g_3(z) \equiv f(z, z, z^{-1})$. It follows that $g_1(x_1 x_2) = g_1(x_1) + g_1(x_2)$. From Lemma A1, we have $g_1(x) = a \log(x)$ for some constant a . Similarly, $g_2(y) = b \log(y)$ and $g_3(z) = c \log(z)$ for constants b and c . We may now construct positive quantities $x = \sqrt{qp}$, $y = \sqrt{pr}$, and $z = \sqrt{rq}$, and observe $f(r, q, p) = f(x^{-1}yz, xy^{-1}z, xyz^{-1}) = f(x^{-1}, x, x) + f(y, y^{-1}, y) + f(z, z, z^{-1}) = g_1(x) + g_2(y) + g_3(z)$. The desired result follows by identifying constants $\alpha = (c + a)/2$, $\beta = (a + b)/2$, and $\gamma = (b + c)/2$. □

Proof of Theorem 1. We proceed by combining Postulate 1 with Postulate 2, which gives

$$\begin{aligned} & \mathbb{I}_{\mathbf{r}(z)\mathbf{r}(w)}[\mathbf{q}_1(z)\mathbf{q}_1(w) \parallel \mathbf{q}_0(z)\mathbf{q}_0(w)] \\ &= \int dz dw \mathbf{r}(z)\mathbf{r}(w) f(\mathbf{r}(z)\mathbf{r}(w), \mathbf{q}_1(z)\mathbf{q}_1(w), \mathbf{q}_0(z)\mathbf{q}_0(w)) \\ &= \mathbb{I}_{\mathbf{r}(z)}[\mathbf{q}_1(z) \parallel \mathbf{q}_0(z)] + \mathbb{I}_{\mathbf{r}(w)}[\mathbf{q}_1(w) \parallel \mathbf{q}_0(w)] \\ &= \int dz \mathbf{r}(z) f(\mathbf{r}(z), \mathbf{q}_1(z), \mathbf{q}_0(z)) + \int dw \mathbf{r}(w) f(\mathbf{r}(w), \mathbf{q}_1(w), \mathbf{q}_0(w)) \\ &= \int dz dw \mathbf{r}(z)\mathbf{r}(w) [f(\mathbf{r}(z), \mathbf{q}_1(z), \mathbf{q}_0(z)) + f(\mathbf{r}(w), \mathbf{q}_1(w), \mathbf{q}_0(w))]. \end{aligned}$$

The last line follows by multiplying each term in the previous line by $\int dw \mathbf{r}(w) = 1$ and $\int dz \mathbf{r}(z) = 1$, respectively. Since this must hold for arbitrary $\mathbf{r}(z)\mathbf{r}(w)$, this implies

$$f(\mathbf{r}(z)\mathbf{r}(w), \mathbf{q}_1(z)\mathbf{q}_1(w), \mathbf{q}_0(z)\mathbf{q}_0(w)) = f(\mathbf{r}(z), \mathbf{q}_1(z), \mathbf{q}_0(z)) + f(\mathbf{r}(w), \mathbf{q}_1(w), \mathbf{q}_0(w)).$$

By Lemma A2, we have $f(\mathbf{r}(z), \mathbf{q}_1(z), \mathbf{q}_0(z)) = \log(\mathbf{r}(z)^\gamma \mathbf{q}_1(z)^\alpha \mathbf{q}_0(z)^\beta)$. From Postulate 3, we require

$$\int dz \mathbf{r}(z) \log(\mathbf{r}(z)^\gamma \mathbf{q}_0(z)^{\alpha+\beta}) = \gamma \int dz \mathbf{r}(z) \log \mathbf{r}(z) + (\alpha + \beta) \int dz \mathbf{r}(z) \log \mathbf{q}_0(z) = 0.$$

Since this must hold for arbitrary $\mathbf{r}(z)$ and $\mathbf{q}_0(z)$, this implies $\gamma = 0$ and $\beta = -\alpha$. Thus we see that $\mathbb{I}_{\mathbf{r}(z)}[\mathbf{r}(z) \parallel \mathbf{q}_0(z)] = \alpha D_{KL}[\mathbf{r}(z) \parallel \mathbf{q}_0(z)]$, which is a constant α times the Kullback–Leibler divergence [6]. Jensen’s inequality easily shows nonnegativity of the form

$$\begin{aligned} & \int dz \mathbf{r}(z) \log\left(\frac{\mathbf{r}(z)}{\mathbf{q}_0(z)}\right) = - \int dz \mathbf{r}(z) \log\left(\frac{\mathbf{q}_0(z)}{\mathbf{r}(z)}\right) \\ & \geq - \log\left(\int dz \mathbf{r}(z) \frac{\mathbf{q}_0(z)}{\mathbf{r}(z)}\right) = - \log(1) = 0. \end{aligned}$$

It follows from Postulate 4 that $\alpha > 0$. As Shannon notes, the scale is arbitrary and simply defines the unit of measure. □

Appendix B. Corollary Proofs

Proof of triangle inequality for Definition 2. To simplify notation, we use the following definitions

$$\begin{aligned} a(z) &= \log\left(\frac{\mathbf{q}_1(z)}{\mathbf{q}_0(z)}\right), & \alpha &= \mathbb{I}_{\mathbf{r}(z)}^p[\mathbf{q}_1(z) \parallel \mathbf{q}_0(z)] = \left(\int dz \mathbf{r}(z) |a(z)|^p\right)^{1/p}, \\ b(z) &= \log\left(\frac{\mathbf{q}_2(z)}{\mathbf{q}_1(z)}\right), & \beta &= \mathbb{I}_{\mathbf{r}(z)}^p[\mathbf{q}_2(z) \parallel \mathbf{q}_1(z)] = \left(\int dz \mathbf{r}(z) |b(z)|^p\right)^{1/p}. \end{aligned}$$

Applying homogeneity followed by Jensen’s inequality, we have

$$\begin{aligned} \mathbb{I}_{\mathbf{r}(z)}^p[\mathbf{q}_2(z) \parallel \mathbf{q}_0(z)] &= \left(\int dz \mathbf{r}(z) |a(z) + b(z)|^p \right)^{1/p} \\ &= (\alpha + \beta) \left(\int dz \mathbf{r}(z) \left| \left(\frac{\alpha}{\alpha + \beta} \right) \frac{a(z)}{\alpha} + \left(\frac{\beta}{\alpha + \beta} \right) \frac{b(z)}{\beta} \right|^p \right)^{1/p} \\ &\leq (\alpha + \beta) \left(\left(\frac{\alpha}{\alpha + \beta} \right) \int dz \mathbf{r}(z) \left| \frac{a(z)}{\alpha} \right|^p + \left(\frac{\beta}{\alpha + \beta} \right) \int dz \mathbf{r}(z) \left| \frac{b(z)}{\beta} \right|^p \right)^{1/p} \\ &= (\alpha + \beta) \left(\frac{\alpha}{\alpha + \beta} + \frac{\beta}{\alpha + \beta} \right)^{1/p} \\ &= \alpha + \beta. \end{aligned}$$

□

Proof of Corollary 1. We unpack Theorem 1 and write joint distributions as the marginalization times of the corresponding conditional distributions, such as $\mathbf{r}(z_1, z_2) \equiv \mathbf{r}(z_2|z_1)\mathbf{r}(z_1)$. This gives

$$\begin{aligned} \mathbb{I}_{\mathbf{r}(z_1, z_2)}[\mathbf{q}_1(z_1, z_2) \parallel \mathbf{q}_0(z_1, z_2)] &= \int dz_1 dz_2 \mathbf{r}(z_1, z_2) \log \left(\frac{\mathbf{q}_1(z_1, z_2)}{\mathbf{q}_0(z_1, z_2)} \right) \\ &= \int dz_1 \mathbf{r}(z_1) \int dz_2 \mathbf{r}(z_2|z_1) \log \left(\frac{\mathbf{q}_1(z_1) \mathbf{q}_1(z_2|z_1)}{\mathbf{q}_0(z_1) \mathbf{q}_0(z_2|z_1)} \right) \\ &= \int dz_1 \mathbf{r}(z_1) \log \left(\frac{\mathbf{q}_1(z_1)}{\mathbf{q}_0(z_1)} \right) \\ &\quad + \int dz_1 \mathbf{r}(z_1) \int dz_2 \mathbf{r}(z_2|z_1) \log \left(\frac{\mathbf{q}_1(z_2|z_1)}{\mathbf{q}_0(z_2|z_1)} \right) \\ &= \mathbb{I}_{\mathbf{r}(z_1)}[\mathbf{q}_1(z_1) \parallel \mathbf{q}_0(z_1)] + \mathbb{E}_{\mathbf{r}(z_1)} \mathbb{I}_{\mathbf{r}(z_2|z_1)}[\mathbf{q}_1(z_2|z_1) \parallel \mathbf{q}_0(z_2|z_1)]. \end{aligned}$$

□

Proof of Corollary 2. Again, we simply unpack Theorem 1 and apply the product property of the logarithm as

$$\begin{aligned} \mathbb{I}_{\mathbf{r}(z)}[\mathbf{q}_2(z) \parallel \mathbf{q}_0(z)] &= \int dz \mathbf{r}(z) \log \left(\frac{\mathbf{q}_2(z)}{\mathbf{q}_0(z)} \right) \\ &= \int dz \mathbf{r}(z) \log \left(\frac{\mathbf{q}_2(z) \mathbf{q}_1(z)}{\mathbf{q}_1(z) \mathbf{q}_0(z)} \right) \\ &= \int dz \mathbf{r}(z) \log \left(\frac{\mathbf{q}_2(z)}{\mathbf{q}_1(z)} \right) + \int dz \mathbf{r}(z) \log \left(\frac{\mathbf{q}_1(z)}{\mathbf{q}_0(z)} \right) \\ &= \mathbb{I}_{\mathbf{r}(z)}[\mathbf{q}_2(z) \parallel \mathbf{q}_1(z)] + \mathbb{I}_{\mathbf{r}(z)}[\mathbf{q}_1(z) \parallel \mathbf{q}_0(z)]. \end{aligned}$$

□

Proof of Corollary 3. Swapping $\mathbf{q}_0(z)$ and $\mathbf{q}_1(z)$ reciprocates the argument of the logarithm in Theorem 1, which gives the negative of the original ordering. □

Proof of Corollary 4. After realization $\check{z} = z_j$, probability is distributed as $\mathbf{r}(z = z_i|z_j) = \delta_{ij}$. Restricting support to $z = \check{z}$ gives

$$\mathbb{I}_{\mathbf{r}(z|\check{z})}[\mathbf{r}(z|\check{z}) \parallel \mathbf{q}(z)] = \log \left(\frac{1}{\mathbf{q}(z = \check{z})} \right) = \mathbb{D}[1 \parallel \mathbf{q}(z = \check{z})].$$

□

Proof of Corollary 5. Computing the expectation value as given, easily reconstructs the standard formulation of cross entropy

$$\begin{aligned} S_{\mathbf{r}(z)}[\mathbf{q}(z)] &= \mathbb{E}_{\mathbf{r}(z)} \mathbb{I}_{\mathbf{r}(z)}[\mathbf{r}(z) \parallel \mathbf{q}(z)] \\ &= \int dz \mathbf{r}(z) \log\left(\frac{1}{\mathbf{q}(z)}\right) \\ &= \mathbb{I}_{\mathbf{r}(z)}[1 \parallel \mathbf{q}(z)]. \end{aligned}$$

□

Proof of Corollary 6. This follows by simply replacing $\mathbf{r}(z)$ with $\mathbf{q}(z)$ in cross entropy. □

Proof of Corollary 7. Plausible joint values of z and w are $\mathbf{p}(z, w) = \mathbf{p}(z|w)\mathbf{p}(w)$. Marginalizing over w recovers present belief $\int dw \mathbf{p}(z|w)\mathbf{p}(w) = \mathbf{p}(z)$. It follows

$$\begin{aligned} \mathbb{E}_{\mathbf{p}(w)} \mathbb{I}_{\mathbf{p}(z|w)}[\mathbf{q}_1(z) \parallel \mathbf{q}_0(z)] &= \int dw \mathbf{p}(w) \int dz \mathbf{p}(z|w) \log\left(\frac{\mathbf{q}_1(z)}{\mathbf{q}_0(z)}\right) \\ &= \int dz \mathbf{p}(z) \log\left(\frac{\mathbf{q}_1(z)}{\mathbf{q}_0(z)}\right) \\ &= \mathbb{I}_{\mathbf{p}(z)}[\mathbf{q}_1(z) \parallel \mathbf{q}_0(z)]. \end{aligned}$$

□

Proof of Corollary 8. We compute the expectation value stated and simply rewrite the product of the marginalization and conditional distribution as the joint distribution $\mathbf{p}(z|w)\mathbf{p}(w) \equiv \mathbf{p}(z, w)$. This gives

$$\begin{aligned} \mathbb{E}_{\mathbf{p}(w)} \mathbb{I}_{\mathbf{p}(z|w)}[\mathbf{p}(z|w) \parallel \mathbf{p}(z)] &= \int dw \mathbf{p}(w) \int dz \mathbf{p}(z|w) \log\left(\frac{\mathbf{p}(z|w)\mathbf{p}(w)}{\mathbf{p}(z)\mathbf{p}(w)}\right) \\ &= \int dw dz \mathbf{p}(z, w) \log\left(\frac{\mathbf{p}(z, w)}{\mathbf{p}(z)\mathbf{p}(w)}\right) \\ &= \mathbb{I}_{\mathbf{p}(z, w)}[\mathbf{p}(z, w) \parallel \mathbf{p}(z)\mathbf{p}(w)]. \end{aligned}$$

□

Proof of Corollary 9. The limit of increasing precision yields the Dirac delta function $\mathbf{p}(z|\check{z}) \equiv \delta(z - \check{z})$. It follows

$$\mathbb{I}_{\mathbf{p}(z|\check{z})}[\mathbf{q}_1(z) \parallel \mathbf{q}_0(z)] = \int dz \delta(z - \check{z}) \log\left(\frac{\mathbf{q}_1(z)}{\mathbf{q}_0(z)}\right) = \log\left(\frac{\mathbf{q}_1(z = \check{z})}{\mathbf{q}_0(z = \check{z})}\right).$$

□

Proof of Corollary 10. We consider differential variations at the optimizer $\mathbf{q}_1(z) = \mathbf{q}^*(z) + \varepsilon\boldsymbol{\eta}(z)$ where variations $\boldsymbol{\eta}(z)$ must maintain normalization $\int dz \boldsymbol{\eta}(z) = 0$ and are otherwise arbitrary. Taking the Gâteaux derivative (with respect to the differential element ε) and applying the variational principle gives

$$0 = \int dz \boldsymbol{\eta}(z) \left[\frac{\mathbf{p}(z|x)}{\mathbf{q}^*(z)} \right].$$

To satisfy the normalization constraint for otherwise arbitrary $\boldsymbol{\eta}(z)$, the term in brackets must be constant. The stated result immediately follows. □

Proof of Corollary 11. Let measurable disjoint subsets of outcomes be $\Omega_{>} = \{z \mid \mathbf{r}(z) > \mathbf{q}_1(z) > 0\}$ and $\Omega_{<} = \{z \mid \mathbf{r}(z) < \mathbf{q}_1(z)\}$. If $\boldsymbol{\eta}(z)$ drives belief toward $\mathbf{r}(z)$ on all measurable subsets, then $\boldsymbol{\eta}(z) \geq 0$ for z almost everywhere in $\Omega_{>}$. Likewise, $\boldsymbol{\eta}(z) \leq 0$ almost everywhere in $\Omega_{<}$. Finally, $\boldsymbol{\eta}(z) = 0$ almost everywhere on the complement $\Omega_z \setminus (\Omega_{>} \cup \Omega_{<})$. In order to retain normalization, we note that $\int dz \boldsymbol{\eta}(z) = 0$. If information is finite, then we may express $\mathbf{r}(z) = \mathbf{q}_1(z)(1 + \delta(z))$ almost everywhere (except an immeasurable subset that is not contained in $\Omega_{>} \cup \Omega_{<}$ for which we could have $\mathbf{q}_1(z) = 0$ and $\mathbf{r}(z) > 0$) and observe that $\delta(z) > 0$ for $z \in \Omega_{>}$ just as $\delta(z) < 0$ for $z \in \Omega_{<}$. Since $\boldsymbol{\eta}(z)$ has the same sign as $\delta(z)$ almost everywhere, it follows

$$\begin{aligned} & \lim_{\varepsilon \rightarrow 0} \frac{\partial}{\partial \varepsilon} \mathbb{I}_{\mathbf{r}(z)}[\mathbf{q}_1(z) + \varepsilon \boldsymbol{\eta}(z) \parallel \mathbf{q}_0(z)] \\ &= \lim_{\varepsilon \rightarrow 0} \frac{\partial}{\partial \varepsilon} \int dz \mathbf{r}(z) \log\left(\frac{\mathbf{q}_1(z) + \varepsilon \boldsymbol{\eta}(z)}{\mathbf{q}_0(z)}\right) \\ &= \int dz \mathbf{r}(z) \frac{\boldsymbol{\eta}(z)}{\mathbf{q}_1(z)} \\ &= \int dz (1 + \delta(z)) \boldsymbol{\eta}(z) \\ &= \int dz \delta(z) \boldsymbol{\eta}(z) \\ &> 0. \end{aligned}$$

□

Proof of Corollary 12. We proceed by constructing the Lagrangian

$$\mathcal{L}[\mathbf{r}(z), \lambda] = \int dz \mathbf{r}(z) \left(\log\left(\frac{\mathbf{r}(z)}{\mathbf{q}_0(z)}\right) - \sum_{i=1}^n \lambda_i (f_i(z) - \varphi_i) \right).$$

We consider differential variations at the optimizer $\mathbf{r}(z) = \mathbf{r}^*(z) + \varepsilon \boldsymbol{\eta}(z)$ where variations $\boldsymbol{\eta}(z)$ must maintain normalization $\int dz \boldsymbol{\eta}(z) = 0$ and are otherwise arbitrary. Taking the Gâteaux derivative and applying the variational principle gives

$$0 = \int dz \boldsymbol{\eta}(z) \left[\log\left(\frac{\mathbf{r}^*(z)}{\mathbf{q}_0(z)}\right) + 1 - \sum_{i=1}^n \lambda_i (f_i(z) - \varphi_i) \right].$$

To satisfy the normalization constraint for otherwise arbitrary $\boldsymbol{\eta}(z)$, the variational principle requires the term in brackets to be constant. The stated result immediately follows. □

Proof of Corollary 13. We unpack Theorem 1 and apply the local conditionality property to write $\mathbf{p}(\theta_1, \theta_2 | \check{y}) \equiv \mathbf{p}(\theta_2 | \theta_1) \mathbf{p}(\theta_1 | \check{y})$. This gives

$$\begin{aligned} & \mathbb{I}_{\mathbf{p}(\theta_1, \theta_2 | \check{y})}[\mathbf{p}(\theta_1, \theta_2 | \check{y}) \parallel \mathbf{p}(\theta_1, \theta_2)] \\ &= \int d\theta_1 d\theta_2 \mathbf{p}(\theta_1, \theta_2 | \check{y}) \log\left(\frac{\mathbf{p}(\theta_1, \theta_2 | \check{y})}{\mathbf{p}(\theta_1, \theta_2)}\right) \\ &= \int d\theta_2 \mathbf{p}(\theta_2 | \theta_1) \int d\theta_1 \mathbf{p}(\theta_1 | \check{y}) \log\left(\frac{\mathbf{p}(\theta_2 | \theta_1) \mathbf{p}(\theta_1 | \check{y})}{\mathbf{p}(\theta_2 | \theta_1) \mathbf{p}(\theta_1)}\right) \\ &= \int d\theta_1 \mathbf{p}(\theta_1 | \check{y}) \log\left(\frac{\mathbf{p}(\theta_1 | \check{y})}{\mathbf{p}(\theta_1)}\right) \\ &= \mathbb{I}_{\mathbf{p}(\theta_1 | \check{y})}[\mathbf{p}(\theta_1 | \check{y}) \parallel \mathbf{p}(\theta_1)]. \end{aligned}$$

□

Proof of Corollary 14. As a consequence of local conditional dependence, we observe that $\mathbf{p}(\theta_2|\check{y}) = \int d\theta_1 \mathbf{p}(\theta_2|\theta_1)\mathbf{p}(\theta_1|\check{y})$. Then, we apply Jensen’s inequality followed by Bayes’ Theorem to obtain

$$\begin{aligned} & \mathbb{I}_{\mathbf{p}(\theta_2|\check{y})}[\mathbf{p}(\theta_2|\check{y}) \parallel \mathbf{p}(\theta_2)] \\ &= \int d\theta_2 \left[\int d\theta_1 \mathbf{p}(\theta_2|\theta_1)\mathbf{p}(\theta_1|\check{y}) \right] \log\left(\frac{\mathbf{p}(\theta_2|\check{y})}{\mathbf{p}(\theta_2)}\right) \\ &\leq \int d\theta_1 \mathbf{p}(\theta_1|\check{y}) \log\left(\int d\theta_2 \mathbf{p}(\theta_2|\theta_1)\frac{\mathbf{p}(\theta_2|\check{y})}{\mathbf{p}(\theta_2)}\right) \\ &= \int d\theta_1 \mathbf{p}(\theta_1|\check{y}) \log\left(\int d\theta_2 \frac{\mathbf{p}(\theta_1|\theta_2)\mathbf{p}(\theta_2|\check{y})}{\mathbf{p}(\theta_1)}\right) \\ &= \int d\theta_1 \mathbf{p}(\theta_1|\check{y}) \left[\log\left(\frac{\mathbf{p}(\theta_1|\check{y})}{\mathbf{p}(\theta_1)}\right) + \log\left(\int d\theta_2 \frac{\mathbf{p}(\theta_1|\theta_2)\mathbf{p}(\theta_2|\check{y})}{\mathbf{p}(\theta_1|\check{y})}\right) \right]. \end{aligned}$$

The first term provides the upper bound we seek. It remains to show that the second term is bound from above by zero, which follows from a second application of Jensen’s inequality

$$\begin{aligned} & \mathbb{I}_{\mathbf{p}(\theta_2|\check{y})}[\mathbf{p}(\theta_2|\check{y}) \parallel \mathbf{p}(\theta_2)] \\ &\leq \mathbb{I}_{\mathbf{p}(\theta_1|\check{y})}[\mathbf{p}(\theta_1|\check{y}) \parallel \mathbf{p}(\theta_1)] + \log\left(\int d\theta_1 d\theta_2 \mathbf{p}(\theta_1|\theta_2)\mathbf{p}(\theta_2|\check{y})\right) \\ &= \mathbb{I}_{\mathbf{p}(\theta_1|\check{y})}[\mathbf{p}(\theta_1|\check{y}) \parallel \mathbf{p}(\theta_1)]. \end{aligned}$$

□

Proof of Corollary 15. The denominator of the first log argument is model evidence, and we apply Bayes’ Theorem to the second log argument. Denominators of log arguments cancel and Jensen’s inequality implies that the first term must be greater than the second.

$$\begin{aligned} & \mathbb{I}_{\mathbf{r}(y|\check{y})}[\mathbf{q}(y|\check{y}) \parallel \mathbf{p}(y)] - \mathbb{I}_{\mathbf{p}(\theta|\check{y})}[\mathbf{p}(\theta|\check{y}) \parallel \mathbf{p}(\theta)] \\ &= \int dy \delta(y - \check{y}) \log\left(\frac{\int d\theta \mathbf{p}(y|\theta)\mathbf{p}(\theta|\check{y})}{\int d\theta \mathbf{p}(y|\theta)\mathbf{p}(\theta)}\right) - \int d\theta \mathbf{p}(\theta|\check{y}) \log\left(\frac{\mathbf{p}(\theta|\check{y})}{\mathbf{p}(\theta)}\right) \\ &= \log\left(\frac{\int d\theta \mathbf{p}(\check{y}|\theta)\mathbf{p}(\theta|\check{y})}{\mathbf{p}(\check{y})}\right) - \int d\theta \mathbf{p}(\theta|\check{y}) \log\left(\frac{\mathbf{p}(\check{y}|\theta)}{\mathbf{p}(\check{y})}\right) \\ &= \log\left(\int d\theta \mathbf{p}(\check{y}|\theta)\mathbf{p}(\theta|\check{y})\right) - \int d\theta \mathbf{p}(\theta|\check{y}) \log(\mathbf{p}(\check{y}|\theta)) \geq 0. \end{aligned}$$

□

Proof of Corollary 16. The stated result immediately follows by taking $f_i(\theta) \equiv \log\left(\frac{\mathbf{q}_i(\theta)}{\mathbf{q}_0(\theta)}\right)$ and applying Corollary 12. □

Proof of Corollary 17. We take $n = 1$, $\mathbf{q}_0(\theta) \equiv \mathbf{p}(\theta)$, and $\mathbf{q}_1(\theta) \equiv \mathbf{p}(\theta|\check{y})$ and apply Corollary 16 with Bayes’ theorem to obtain

$$\mathbf{r}(\theta) \propto \mathbf{p}(\theta) \left(\frac{\mathbf{p}(\theta|\check{y})}{\mathbf{p}(\theta)}\right)^\lambda = \mathbf{p}(\theta) \left(\frac{\mathbf{p}(\check{y}|\theta)}{\mathbf{p}(\check{y})}\right)^\lambda.$$

□

Appendix C. Computation of Information in the Experiments

Appendix C.1. Inference

Prior belief is $\mathbf{p}(\theta) \equiv \mathcal{N}(\theta|0, A)$ and $\mathbf{p}(x^{(j)}) \equiv \mathcal{N}(x^{(j)}|0, \Sigma)$. Since $y^{(j)} = \theta + x^{(j)}$, we have $\mathbf{p}(y^{(j)}|\theta) \equiv \mathcal{N}(y^{(j)}|\theta, \Sigma)$. If we let n samples have an average \bar{y} , it easily follows that $\mathbf{p}(\bar{y}|\theta) \equiv \mathcal{N}(\bar{y}|\theta, \frac{1}{n}\Sigma)$. Bayes' rule gives $\mathbf{p}(\theta|\bar{y}) \propto \mathbf{p}(\bar{y}|\theta)\mathbf{p}(\theta)$ or

$$\begin{aligned}\mathbf{p}(\theta|\bar{y}) &\propto \exp\left[\frac{-1}{2}\theta^T A \theta - \frac{n}{2}(\theta - \bar{y})^T \Sigma (\theta - \bar{y})\right] \\ &\propto \exp\left[\frac{-1}{2}(\theta - \mu)^T B (\theta - \mu)\right],\end{aligned}$$

where $B = A + n\Sigma$ and $\mu = nB\Sigma \bar{y}$. Normalization yields $\mathbf{p}(\theta|\bar{y}) \equiv \mathcal{N}(\theta|\mu, B)$.

Appendix C.2. Mutual Information

We marginalize over plausible θ to obtain the corresponding probability of observing \bar{y} as $\mathbf{p}(\bar{y}) = \int d\theta \mathbf{p}(\theta)\mathbf{p}(\bar{y}|\theta)$. This gives $\mathbf{p}(\bar{y}) \equiv \mathcal{N}(\bar{y}|0, A + \frac{1}{n}\Sigma)$. Mutual information, which is the expected information gained by observing \bar{y} according to present belief, is computed this way:

$$\begin{aligned}&\int d\bar{y} d\theta \mathbf{p}(\bar{y}, \theta) \log\left(\frac{\mathbf{p}(\bar{y}, \theta)}{\mathbf{p}(\bar{y})\mathbf{p}(\theta)}\right) \\ &= \int d\theta \mathbf{p}(\theta) \int d\bar{y} \mathbf{p}(\bar{y}|\theta) \log\left(\frac{\mathbf{p}(\bar{y}|\theta)}{\mathbf{p}(\bar{y})}\right) \\ &= \int d\theta \mathbf{p}(\theta) \int d\bar{y} \mathbf{p}(\bar{y}|\theta) \log\left(\frac{|2\pi\frac{1}{n}\Sigma|^{-1/2} \exp(\frac{-n}{2}(\bar{y} - \theta)^T \Sigma (\bar{y} - \theta))}{|2\pi(A + \frac{1}{n}\Sigma)|^{-1/2} \exp(\frac{-1}{2}\bar{y}^T (A + \frac{1}{n}\Sigma)^{-1}\bar{y})}\right) \\ &= \frac{1}{2} \log \det(n\Sigma A + I).\end{aligned}$$

Appendix C.3. First Inference Information in a Subsequent View

Let the state of belief before an experiment be $\mathbf{p}(\theta) \equiv \mathcal{N}(\theta|0, A)$. After observing $\bar{y}^{(1)}$, inference gives $\mathbf{p}(\theta|\bar{y}^{(1)}) \equiv \mathcal{N}(\theta|\mu, B)$. Additional observations $\bar{y}^{(2)}$ yield $\mathbf{p}(\theta|\bar{y}^{(1)}, \bar{y}^{(2)}) \equiv \mathcal{N}(\theta|v, C)$. Information gained in the first observation in the view of inference following the second observation is computed as

$$\begin{aligned}&\mathbb{I}_{\mathbf{p}(\theta|\bar{y}^{(1)}, \bar{y}^{(2)})} \left[\mathbf{p}(\theta|\bar{y}^{(1)}) \parallel \mathbf{p}(\theta) \right] \\ &= \int d\theta \mathcal{N}(\theta|v, C) \log\left(\frac{|2\pi B|^{-1/2} \exp(\frac{-1}{2}(\theta - \mu)^T B (\theta - \mu))}{|2\pi A|^{-1/2} \exp(\frac{-1}{2}\theta^T A \theta)}\right) \\ &= \frac{1}{2} \left(\log \det(AB) + \text{tr}((A - B)C) + v^T A v - (v - \mu)^T B (v - \mu) \right).\end{aligned}$$

References

1. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [[CrossRef](#)]
2. LaPlace, P.S. Mémoire sur la probabilité des causes par les événements. *Mémoires de l'Academie Royale des Sciences Présentés par Divers Savan* **1774**, *6*, 621–656.
3. Jeffreys, H. *The Theory of Probability*; OUP Oxford: Oxford, UK, 1998.

4. Jaynes, E.T. *Probability Theory: The Logic of Science*; Cambridge University Press: Cambridge, UK, 2003.
5. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [[CrossRef](#)]
6. Kullback, S.; Leibler, R.A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [[CrossRef](#)]
7. Kullback, S. *Information Theory and Statistics*; Courier Corporation: North Chelmsford, MA, USA, 1997.
8. Lindley, D.V. On a measure of the information provided by an experiment. *Ann. Math. Stat.* **1956**, *27*, 986–1005. [[CrossRef](#)]
9. Cox, R.T. Probability, frequency and reasonable expectation. *Am. J. Phys.* **1946**, *14*, 1–13. [[CrossRef](#)]
10. Ramsey, F.P. Truth and probability. In *Readings in Formal Epistemology*; Springer: Berlin, Germany, 2016; pp. 21–45.
11. Lehman, R.S. On confirmation and rational betting. *J. Symbol. Logic* **1955**, *20*, 251–262. [[CrossRef](#)]
12. Adams, E.W. On rational betting systems. *Arch. Math. Logic* **1962**, *6*, 7–29. [[CrossRef](#)]
13. Freedman, D.A.; Purves, R.A. Bayes' method for bookies. *Ann. Math. Stat.* **1969**, *40*, 1177–1186. [[CrossRef](#)]
14. Skyrms, B. Dynamic Coherence and Probability Kinematics. *Philos. Sci.* **1987**, *54*, 1–20. [[CrossRef](#)]
15. Soofi, E.S. Capturing the intangible concept of information. *J. Am. Stat. Assoc.* **1994**, *89*, 1243–1254. [[CrossRef](#)]
16. Soofi, E.S. Principal information theoretic approaches. *J. Am. Stat. Assoc.* **2000**, *95*, 1349–1353. [[CrossRef](#)]
17. Ebrahimi, N.; Soofi, E.S.; Zahedi, H. Information properties of order statistics and spacings. *IEEE Trans. Inf. Theory* **2004**, *50*, 177–183. [[CrossRef](#)]
18. Ebrahimi, N.; Soofi, E.S.; Soyer, R. Information measures in perspective. *Int. Stat. Rev.* **2010**, *78*, 383–412. [[CrossRef](#)]
19. Gelman, A.; Hwang, J.; Vehtari, A. Understanding predictive information criteria for Bayesian models. *Stat. Comput.* **2013**, *24*, 997–1016. [[CrossRef](#)]
20. Good, I.J. Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *Ann. Math. Stat.* **1963**, *34*, 911–934. [[CrossRef](#)]
21. MacKay, D.J.C. *Information Theory, Inference and Learning Algorithms*; Cambridge University Press: Cambridge, UK, 2003.
22. Tishby, N.; Zaslavsky, N. Deep learning and the information bottleneck principle. In Proceedings of the 2015 IEEE Information Theory Workshop (ITW), Jerusalem, Israel, 26 April–1 May 2015.
23. Barnard, G. The theory of information. *J. R. Stat. Soc. Methodol.* **1951**, *13*, 46–64. [[CrossRef](#)]
24. Rényi, A. On Measures of Entropy and Information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*; University of California Press: Berkeley, CA, USA, 1961; pp. 547–561.
25. Csiszár, I. Information-type measures of difference of probability distributions and indirect observation. *Stud. Sci. Math. Hung.* **1967**, *2*, 229–318.
26. Jizba, P.; Arimitsu, T. Generalized statistics: Yet another generalization. *Physica A* **2004**, *340*, 110–116. [[CrossRef](#)]
27. Hanel, R.; Thurner, S. A comprehensive classification of complex statistical systems and an axiomatic derivation of their entropy and distribution functions. *EPL Europhysic. Lett.* **2011**, *93*, 20006. [[CrossRef](#)]
28. Ilić, V.M.; Stanković, M.S. Generalized Shannon–Khinchin axioms and uniqueness theorem for pseudo-additive entropies. *Physica A* **2014**, *411*, 138–145. [[CrossRef](#)]
29. Tempesta, P. Group entropies, correlation laws, and zeta functions. *Phys. Rev. E* **2011**, *84*, 021121. [[CrossRef](#)] [[PubMed](#)]
30. Bernoulli, J. *Ars Conjectandi*; Basileae Impensis Thurnisiorum Fratrum: Basel, Switzerland, 1713. Available online: https://books.google.com.hk/books?hl=en&lr=&id=XPOf7STJ3y4C&oi=fnd&pg=PA1&dq=Ars+conjectandi%3B+Impensis+Thurnisiorum,+fratrum,+1713.&ots=Lj-EfRRgbu&sig=KCYr2_EIoMa1ui-2fibrhQAV5aE&redir_esc=y&hl=zh-CN&sourceid=cndr#v=onepage&q=Ars%20conjectandi%3B%20Impensis%20Thurnisiorum%2C%20fratrum%2C%201713.&f=false (accessed on 15 January 2020).
31. Jaynes, E.T. Information theory and statistical mechanics. *Phys. Rev.* **1957**, *106*, 620. [[CrossRef](#)]
32. Shore, J.; Johnson, R. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans. Inf. Theory* **1980**, *26*, 26–37. [[CrossRef](#)]
33. Shore, J.; Johnson, R. Properties of cross-entropy minimization. *IEEE Trans. Inf. Theory* **1981**, *27*, 472–482. [[CrossRef](#)]

34. Jizba, P.; Korbel, J. Maximum Entropy Principle in statistical inference: Case for non-Shannonian entropies. *Phys. Rev. Lett.* **2019**, *122*, 120601. [[CrossRef](#)]
35. Hájek, A. Dutch book arguments. In *The Handbook of Rational and Social Choice*; Oxford University Press: Oxford, UK, 2008; pp. 173–196.
36. Weisberg, J. Varieties of Bayesianism. In *Inductive Logic*; Gabbay, D.M., Hartmann, S., Woods, J., Eds.; Elsevier: Amsterdam, The Netherlands, 2011; Volume 10, pp. 477–551.
37. Fadeev, D. Zum Begriff der Entropie einer endlichen Wahrscheinlichkeitsschemas. In *Arbeiten zur Informationstheorie I*; Deutscher Verlag der Wissenschaften: Berlin, Germany, 1957; pp. 85–90.
38. Lindley, D.V. *The Bayesian Approach to Statistics*; Technical Report; University of California, Berkeley, Operations Research Center: Berkeley, CA, USA, 1980.
39. Nikodym, O. Sur une généralisation des intégrales de MJ Radon. *Fundamenta Mathematicae* **1930**, *15*, 131–179. [[CrossRef](#)]
40. Bernardo, J.M. Expected information as expected utility. *Ann. Stat.* **1979**, *7*, 686–690. [[CrossRef](#)]
41. Fisher, R.A. Theory of statistical estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*; Cambridge University Press: Cambridge, UK, 1925; Volume 22, pp. 700–725.
42. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723. [[CrossRef](#)]
43. Schwarz, G. Estimating the Dimension of a Model. *Ann. Stat.* **1978**, *6*, 461–464. [[CrossRef](#)]
44. Burnham, K.P.; Anderson, D.R. Kullback-Leibler information as a basis for strong inference in ecological studies. *Wildl. Res.* **2001**, *28*, 111. [[CrossRef](#)]
45. Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*; Elsevier: Amsterdam, The Netherlands, 2014.
46. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
47. Erdős, P. On the Distribution Function of Additive Functions. *Ann. Math.* **1946**, *47*, 1. [[CrossRef](#)]



© 2020 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) license (<http://creativecommons.org/licenses/by/4.0/>).