Information as a Measure of Variation

William A. Dembski

Conceptual Foundations of Science
Baylor University, Box 7130
Waco, Texas 76798

William_Dembski@baylor.edu

Version 3.11, 11 December 2004

Abstract

In many applications of information theory, information measures the reduction of uncertainty that results from the knowledge that an event has occurred. Even so, an item of information learned need not be the occurrence of an event but, rather, the change in probability distribution associated with an ensemble of events. This paper examines the basic account of information, which focuses on events, and reviews how it may be naturally generalized to probability distributions/measures. The resulting information measure is special case of the Rényi information divergence (also known as the Rényi entropy). This information measure, herein dubbed the *variational information*, meaningfully assigns a numerical bit-value to arbitrary state transitions of physical systems. The information topology of these state transitions is characterized canonically by a right and left continuity spectrum defined in terms of the Kantorovich-Wasserstein metric. These continuity spectra provide a theoretical framework for characterizing the informational continuity of evolving systems and for rigorously assessing the degree to which such systems exhibit, or fail to exhibit, continuous change.

1 The Fundamental Intuition

Ordinarily, information refers to the meaning or semantic content of a message. Getting a handle on the meaning of a message, however, has proven difficult mathematically. Thus, when mathematicians speak of information, they are concerned not so much with the meaning of a message as with the vehicle by which the message is transmitted from sender to receiver.

The most common vehicle for transmitting messages is the character string. The mathematical theory of information is largely about quantifying the complexity of such strings, characterizing their statistical properties when they are sent across a noisy communication channel (noise being represented as a stochastic process that disrupts the strings in statistically well-defined ways), preserving the strings despite the presence of noise (i.e., the theory of error-correcting codes), compressing the strings to improve efficiency, and transforming the strings into other strings to maintain their security (i.e., cryptography).

The underlying framework here can be generalized. A character string based on a given alphabet sits in a space of such character strings. This space constitutes a reference class of possibilities. In general, then, to communicate a message from a reference class of possibilities means selecting a subset from it, thereby identifying certain possibilities and ruling out the rest.

The fundamental intuition behind the mathematical theory of information is now readily stated. Robert Stalnaker (1984: 85) puts it this way: "To learn something, to acquire information, is to rule out possibilities. To understand the information conveyed in a communication is to know what possibilities would be excluded by its truth." To be told that it is either going to rain or not rain tomorrow is therefore to acquire no information. Rain-or-not-rain exhausts all possibilities, so learning that it is either going to rain or not rain is uninformative. Consequently, the only way to convey information is by restricting that range of possibilities. For instance, to be told that it will rain tomorrow does indeed communicate information because it excludes the possibility of not-rain.

Information always presupposes a range of possibilities, and conveying information means ruling out some of those possibilities. It follows that information can be quantified. Indeed, the more possibilities that get ruled out, the more information gets conveyed. Fred Dretske (1981: 4) elaborates: "Information theory identifies the amount of information associated with, or generated by, the occurrence of an event (or the realization of a state of affairs) with the reduction in uncertainty, the elimination of possibilities, represented by that event or state of affairs." Even so, to measure information it is not enough simply to count the number of possibilities that were eliminated and present that number as the relevant measure of information. The problem is that a simple enumeration of eliminated possibilities tells us nothing about how those possibilities were individuated.

Consider, for instance, the following individuation of poker hands: RF (a royal flush) and $\neg RF$ (all other poker hands). To learn that something other than a royal flush was dealt (i.e., possibility $\neg RF$) is clearly to acquire less information than to learn that a royal flush was dealt (i.e., possibility RF). A royal flush is highly specific. We have acquired a lot of information when we learn that a royal flush was dealt. On the other hand, we have acquired hardly any information when we learn that something other than a royal flush was dealt. Most poker hands are not royal flushes, and we expect to be dealt them only rarely. Nevertheless, if our measure of information is simply an enumeration of eliminated possibilities, the same numerical value must be assigned in both instances since, in each instance, a single possibility is eliminated.

It follows that how we measure information needs to be independent of whatever procedure we use to individuate the possibilities under consideration. The way to do this is not simply to count possibilities but to assign probabilities to those possibilities. For a thoroughly shuffled deck of cards, the probability of being dealt a royal flush (i.e., possibility RF) is approximately .000002 whereas the probability of being dealt anything other than a royal flush (i.e., possibility $\neg RF$) is approximately .999998.

Probabilities by themselves, however, are not information measures. Although probabilities distinguish possibilities by the amount of information they contain, probabilities are inconvenient for measuring information. There are two reasons for this. First, the scaling and directionality of the numbers assigned by probabilities needs to be recalibrated. We are clearly acquiring more information when we learn someone was dealt a royal flush than when we learn someone was not dealt a royal flush. And yet the probability of being dealt a royal flush (i.e., .000002) is minuscule compared to the probability of being dealt something other than a royal flush (i.e., .999998). Smaller probabilities signify more information, not less.

The second reason probabilities are inconvenient for measuring information is that they are multiplicative rather than additive. If we learn that Alice was dealt a royal flush playing poker at one Las Vegas casino and that Bob was dealt a royal flush playing poker at a different Las Vegas casino, the probability that both Alice and Bob were dealt royal flushes is the product of the individual probabilities. On the other hand, it is convenient for information to be measured additively so that the measure of information assigned to Alice and Bob jointly being dealt royal flushes equals the measure of information assigned to Alice being dealt a royal flush *plus* the measure of information assigned to Bob being dealt a royal flush. Now, there is a straightforward mathematical way to transform probabilities that circumvents both these difficulties, and that is to apply a negative logarithm to the probabilities. Applying a negative logarithm assigns more information to less probability and, because the logarithm of a product is the sum of the logarithms, transforms multiplicative probability measures into additive information measures.

Moreover, in deference to communication theorists, it is customary to use the logarithm to the base 2. The negative logarithm to the base 2 of a probability corresponds to the average number of binary digits, or bits, needed to identify an event of that probability. Shannon showed that the binary code provides the simplest and most cost-efficient way of handling information (in particular, it uses the least memory and bandwidth).[1] Hence, the most convenient way for communication theorists to measure information is in bits. Consequently, the logarithm to the base 2 has become the canonical logarithm for communication theorists. Given an event A of probability p, the information associated with A is therefore defined as

$$I(A) =_{def} -\log_2 p.$$

[1] For a nice discussion of the privileged place of the binary code, see von Baeyer (2004: 30–31).

2 Entropy

Information theorists sometimes refer to the definition of information just given as the *surprisal* associated with a particular event (the smaller the event's probability, the bigger the "surprise" associated with its occurrence—see Dretske 1981: 10). Yet regardless of the designation, it is striking how little this notion comes up directly in the mathematical theory of information. If the information associated with a particular event A is to signify anything mathematically, then $I(A) = -\log_2 p$ is it. Nonetheless, this notion is almost entirely passed over in favor of a different notion, called *entropy*. Entropy, rather than being associated with a particular event, is associated with a partition of events for a given reference class of possibilities Ω. Given events $A_1, A_2, ..., A_m$ that are mutually exclusive and exhaust Ω, and given that the probability of A_i is p_i ($1 \leq i \leq m$, $p_1 + p_2 + \cdots + p_m = 1$, no $p_i = 0$), the entropy associated with this collection of events is

$$H =_{def} -\sum_{i=1}^{m} p_i \log_2 p_i.$$

To be sure, I is tacitly embedded in this definition since this equation can be rewritten as

$$H = \sum_{i=1}^{m} p_i I(A_i).$$

But this reformulation of entropy adds no new insight, and the terms $I(A_i)$ have no independent significance within the mathematical theory of information. Communication engineers interpret each of the A_is as a possible transmission from an information source and thus interpret the entropy H as the average information outputted by that source.

Why is H rather than I the preferred measure for information among communication theorists? Fred Dretske (1981: 11) explains,

> Communication engineers have no use for the surprisal value of a particular state of affairs; they use the formula for calculating the surprisal value only as a "stepping stone" in the calculation of the average information generated by a source. This preoccupation with averages is perfectly understandable. What the engineer wants is a concept that characterizes the whole statistical nature of the information source. He is not concerned with individual messages. A communication system must face the problem of handling any message that the source can produce.

To this Warren Weaver (1949: 14) adds,

> If it is not possible or practicable to design a system that can handle everything perfectly, then the system should be designed to handle well the jobs it is most likely to be asked to do.... This sort of consideration leads at once to the necessity of characterizing the statistical nature of the whole ensemble of messages which a given kind of source can and will produce.

The communication engineer's preoccupation with ensembles and averages, however, obscures the fundamental intuition behind information. Dretske (1981: 50) therefore remarks,

> Although the *surprisal* of a given event . . . and the amount of information carried by a particular signal . . . are not significant quantities in engineering applications of information theory (except, perhaps, as mathematical intermediaries for the calculation of entropy), these *are* the important quantities for the study of information, as commonly understood, and hence for the kind of cognitive studies that depend on a semantically related concept of information.

The bottom line is that communication engineers downplay the surprisal and focus principally on entropy, or average information.[2]

[2] For the purposes of this article it is enough to consider entropy as developed within information theory. The notion, however, has deep connections to other areas in mathematics and physics. With a change in logarithmic base, it is equivalent to S, the Maxwell-Botzmann-Gibbs entropy: let W be the number of ways of arranging N atoms in cells numbered 1 through m with N_i atoms in cell i. Then $S = \log_e W$, where $W = \left[N!/\prod_{1\leq i\leq m} N_i!\right]^{1/N}$. If we now let $p_i = N_i/N$, then by Stirling's formula S comes out to approximately $-\sum_{1\leq i\leq m} p_i \log_e p_i$. See Yockey (1992: 66–67).

Entropy also plays an important role in ergodic theory, where it provides a necessary and sufficient condition for two Bernoulli shifts to be metrically isomorphic (these are shift automorphisms on infinite product spaces where the factors of a given product space are all identical and each constitutes a fixed finite set). In 1970 Donald Ornstein showed that entropy completely classifies Bernoulli shifts up to isomorphism, thereby creating a revolution in ergodic theory. In particular, he showed that if the entropies for the two spaces are the same, then there is a measure-preserving map that also preserves the Bernoulli shifts. For a textbook proof, see Cornfeld et al. (1982: 258–280). For the original article by Ornstein, see Ornstein (1970).

Entropy also comes up in statistical decision theory. Consider two hypotheses, H_1 and H_2, that induce probability measures μ_1 and μ_2 on Ω and are absolutely continuous with respect to some privileged measure λ (not necessarily a probability) so that by the Radon-Nikodym theorem, $\mu_1 = f_1 d\lambda$ and $\mu_2 = f_2 d\lambda$. Statisticians then define the information of μ_1 with respect to μ_2 as follows: $I(\mu_1, \mu_2) =_{def} \int_\Omega f_1(x) \log \dfrac{f_1(x)}{f_2(x)} d\lambda(x)$. This integral is interpreted as the mean information per observation under μ_1 that discriminates in favor of H_1 against H_2. If Ω is finite and λ is the counting measure, then $f_i(x)$ is the probability of x under μ_i. Letting $f_1(x) = p_x$ and $f_2(x) = q_x$, this integral becomes $\sum p_x \log p_x - \sum p_x \log q_x$. Clearly, entropy assumes pride of place in this expression. See Kullback (1997: 5).

The most common way to represent entropy is in terms of random variables (for simplicity we'll just consider finite probability spaces). Consider random variables $X : \Omega \longrightarrow \mathfrak{X}$ and $Y : \Omega \longrightarrow \mathfrak{Y}$ with probability $\mathbf{P}$ on Ω. Let $p(x) = \mathbf{P}(X = x)$, $p(y) = \mathbf{P}(Y = y)$, $p(x, y) = \mathbf{P}(X = x, Y = y)$, and $p(x|y) = \mathbf{P}(X = x|Y = y)$ for $x \in \mathfrak{X}$ and $y \in \mathfrak{Y}$. Then it is customary to define the following types of entropy (individual, joint, and conditional):

$$H(X) =_{def} -\sum_{x\in\mathfrak{X}} p(x)\log_2 p(x),$$

$$H(X,Y) =_{def} -\sum_{x\in\mathfrak{X}}\sum_{y\in\mathfrak{Y}} p(x,y)\log_2 p(x,y),$$

$$H(X|Y) =_{def} -\sum_{x\in\mathfrak{X}}\sum_{y\in\mathfrak{Y}} p(x,y)\log_2 p(x|y).$$

With these definitions in hand, it is also customary to define the *mutual information* of X with respect to Y as follows:

$$I(X:Y) =_{def} -\sum_{x\in\mathfrak{X}}\sum_{y\in\mathfrak{Y}} p(x,y)\log_2 \frac{p(x,y)}{p(x)p(y)}.$$

The mutual information $I(X : Y)$ is typically interpreted as measuring how much the uncertainty in X is reduced by knowing Y (see Cover and Thomas 1991: 20). It can also be rewritten as $H(X) + H(Y) - H(X, Y) = H(X) - H(X|Y)$, which, by symmetry, also equals $H(Y) - H(Y|X) = I(Y : X)$. Accordingly, the reduction in uncertainty in X given Y is the same as the reduction in uncertainty in Y given X. Or, as Cover and Thomas (1991: 20) put it, "X says as much about Y as Y says about X."[3] Note that $I(X : X) = H(X) - H(X|X) = H(X)$ (because $H(X|X) = 0$), indicating that the amount of uncertainty reduced in X is maximal when X itself is known (this is as it should be). Mutual information plays a key role in determining the capacity of communication channels (see Cover and Thomas 1991: ch. 8).

The standard interpretation of mutual information raises two interesting points about the interpretation of information generally. For one, it suggests that information is properly defined as a relation between two items, one providing the backdrop against which the other provides novel input. The second point is whether information should be interpreted as a reduction or an addition. The mutual information $I(X : Y)$ identifies the amount by which the uncertainty in X is reduced by knowing Y, with the maximal reduction coming

[3] Christof Adami (2004: 6) elaborates on this symmetry: "The colon between X and Y in the notation for the [mutual] information is standard; it is supposed to remind the reader that information is a symmetric quantity: what X knows about Y, Y also knows about X."

when Y actually equals X. But it makes sense also to interpret information not reductively but additively, so that what is measured is the amount by which X extends, or adds to, our knowledge of Y. In that case, information is minimal when X merely repeats Y but grows as X diverges from Y. $H(X|Y)$ fits that bill, and could be interpreted as such a measure of information (though it is usually just called conditional entropy).

I want next to characterize an information measure that is both relational and additive in the sense just described, but that calculates the amount of information associated with specific possibilities as opposed to the amount associated with averages of possibilities, the latter being what information measures defined in terms of entropy always do. This measure seems much closer to our fundamental intuitions about information—indeed, it provides a very natural generalization for the information measure described in section 1.[4] Moreover, by being defined within a Hilbert space formalism, this information measure is mathematically tractable. The burden of this paper is to indicate how it might be readily and widely applied.

3 Information as a Modified Variance

In generalizing the information measure described in section 1, let's start by reexamining that measure and then reformulating it in a way that makes clear how the generalization should proceed. Given a probability space Ω and an event A in Ω, we defined the information in A as the negative logarithm to the base 2 of the probability of A. Let's write this as

$$I(A|\Omega) =_{def} -\log_2 \mathbf{P}(A),$$

where $\mathbf{P}$ is the relevant probability on Ω. Given an additional event B in Ω, this definition generalizes readily to the conditional information of A given B:

$$I(A|B) =_{def} -\log_2 \mathbf{P}(A|B) = -\log_2 \frac{\mathbf{P}(AB)}{\mathbf{P}(B)}.$$

Next, consider two real-valued functions defined on Ω: $f_1 =_{def} \frac{1}{\mathbf{P}(AB)} 1_{AB}$ and $f_2 =_{def} \frac{1}{\mathbf{P}(B)} 1_B$ (1_{AB} and 1_B are indicator functions, equal to 1 on the set in question, 0 outside). In addition, consider the two probability measures induced by these functions: $\mu_1 = f_1 d\mathbf{P}$ and $\mu_2 = f_2 d\mathbf{P}$. Then, $\mu_1 \ll \mu_2$ (i.e., μ_1 is absolutely continuous with respect to μ_2), and thus, by the Radon-Nikodym theorem, there is a function $\frac{d\mu_1}{d\mu_2}$ (the Radon-Nikodym derivative of μ_1 with respect to μ_2) such that $\frac{d\mu_1}{d\mu_2} d\mu_2 = d\mu_1$, or equivalently, $\frac{d\mu_1}{d\mu_2} f_2 d\mathbf{P} = f_1 d\mathbf{P}$. Substituting into this last equation for f_1 and f_2, we then get $\frac{d\mu_1}{d\mu_2} \frac{1}{\mathbf{P}(B)} 1_B d\mathbf{P} = \frac{1}{\mathbf{P}(AB)} 1_{AB} d\mathbf{P}$, which implies that $\frac{d\mu_1}{d\mu_2} = \frac{\mathbf{P}(B)}{\mathbf{P}(AB)} 1_{AB}$.

[4] See footnote 2.

Accordingly, we can rewrite $I(A|B)$ as follows:

$$\begin{aligned} I(A|B) &= -\log_2 \frac{\mathbf{P}(AB)}{\mathbf{P}(B)} \\ &= \log_2 \frac{\mathbf{P}(B)}{\mathbf{P}(AB)} \\ &= \log_2 \int_\Omega \frac{\mathbf{P}(B)}{\mathbf{P}(AB)^2} 1_{AB} d\mathbf{P} \\ &= \log_2 \int_\Omega \left(\frac{\mathbf{P}(B)}{\mathbf{P}(AB)} 1_{AB} \right)^2 \frac{1}{\mathbf{P}(B)} 1_B d\mathbf{P} \\ &= \log_2 \int_\Omega \left(\frac{d\mu_1}{d\mu_2} \right)^2 d\mu_2 \end{aligned}$$

This suggests that for measures μ_1 and μ_2 with $\mu_1 \ll \mu_2$, we can define the information of μ_1 with respect to μ_2 as follows (assuming the Radon-Nikodym derivative is square integrable):

$$I(\mu_1|\mu_2) =_{def} \log_2 \int_\Omega \left(\tfrac{d\mu_1}{d\mu_2} \right)^2 d\mu_2.$$

For reasons that will become clear momentarily, let us refer to this information measure as the *variational information* of μ_1 given μ_2.

How should we interpret this measure of information? Let $\mathbf{E}_\mu$ denote the expectation operator for integrable functions on Ω with respect to μ, i.e., $\mathbf{E}_\mu(f) = \int_\Omega f d\mu$. Additionally, let $\mathbf{V}_\mu$ denote the variance operator for square-integrable functions on Ω with respect to μ, i.e., $\mathbf{V}_\mu(f) = \mathbf{E}_\mu([f - \mathbf{E}_\mu(f)]^2)$. Then, because μ_1 and μ_2 are probability measures and $\mathbf{E}_{\mu_2}(\frac{d\mu_1}{d\mu_2}) = \int_\Omega \frac{d\mu_1}{d\mu_2} d\mu_2 = \int_\Omega d\mu_1 = 1$, it follows that

$$\int_\Omega \left(\tfrac{d\mu_1}{d\mu_2} \right)^2 d\mu_2 = \int_\Omega \left(\tfrac{d\mu_1}{d\mu_2} - 1 \right)^2 d\mu_2 + 1 = \mathbf{V}_{\mu_2}(\tfrac{d\mu_1}{d\mu_2}) + 1.^5$$

In other words,

$$I(\mu_1|\mu_2) = \log_2[\mathbf{V}_{\mu_2}(\tfrac{d\mu_1}{d\mu_2}) + 1],$$

making $I(\mu_1|\mu_2)$, in essence, a disguised form of variance, measuring how much μ_1 varies or diverges from μ_2.[6]

[5] Note that $\int_\Omega \left(\frac{d\mu_1}{d\mu_2} \right)^2 d\mu_2$ can also be rewritten as $\int_\Omega \frac{d\mu_1}{d\mu_2} d\mu_1$, which can be useful for certain purposes. Further, suppose that μ_1 and μ_2 are absolutely continuous with respect to a measure λ. Then this last integral can be rewritten as $\int_\Omega \left[\frac{d\mu_1}{d\lambda} / \frac{d\mu_2}{d\lambda} \right] \frac{d\mu_1}{d\lambda} d\lambda$.

[6] A variance form of information is not without precedent. Indeed, it predates Shannon's formulation of information by twenty years. For a probability density $f(x;\theta)$ indexed by a

It is significant that the variance term in $I(\mu_1|\mu_2)$ takes the form $\int_\Omega \left(\frac{d\mu_1}{d\mu_2} - 1\right)^2 d\mu_2$. Because the Radon-Nikodym derivative of a measure with respect to itself is always identically 1, this variance term can be rewritten as $\int_\Omega \left(\frac{d\mu_1}{d\mu_2} - \frac{d\mu_2}{d\mu_2}\right)^2 d\mu_2$. In other words, $I(\mu_1|\mu_2)$ measures the mean square variation of $\frac{d\mu_1}{d\mu_2}$ from $\frac{d\mu_2}{d\mu_2}$ with respect to μ_2. But because $\frac{d\mu_2}{d\mu_2} \equiv 1$, $\frac{d\mu_2}{d\mu_2}$ is the uniform probability density with respect to μ_2. We may therefore think of $I(\mu_1|\mu_2)$ as the mean square variation of the probability density of μ_1 from the uniform density with respect to μ_2.

The variational information has the requisite properties that we have come to expect from an information measure. Because the variance is always non-negative, $\mathbf{V}_\mu(\frac{d\mu_1}{d\mu_2}) + 1 = \int_\Omega \left(\frac{d\mu_1}{d\mu_2}\right)^2 d\mu_2$ is always greater than or equal to 1. Moreover, by Jensen's inequality, this quantity is strictly greater than 1 when $\frac{d\mu_1}{d\mu_2}$ differs from 1 on a set whose μ_2-measure is greater than 0. It follows that $I(\mu_1|\mu_2)$ is always greater than or equal to 0 and strictly greater than zero so long as $\mu_1 \ll \mu_2$ and these two measures are distinct.

Additivity of the variational information also follows. Suppose $\mathcal{A}_1, \mathcal{A}_2, ..., \mathcal{A}_n$ are independent σ-algebras on Ω with respect to a probability measure μ (i.e., for $A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2, ..., A_n \in \mathcal{A}_n, \mu(A_1A_2 \cdots A_n) = \mu(A_1)\mu(A_2)\cdots\mu(A_n)$). Suppose that these σ-algebras together generate the σ-algebra $\mathcal{A}$, and suppose that μ and the additional probability measure ν are both defined on $\mathcal{A}$. For $1 \leq i \leq n$, let μ_i be the restriction of μ to $\mathcal{A}_i$ and suppose that each $\mu_i \ll \nu$. Then, μ equals the product measure $\mu_1 \otimes \mu_2 \cdots \otimes \mu_n$, and therefore

$$\begin{aligned}
I(\mu|\nu) &= I(\mu_1 \otimes \mu_2 \cdots \otimes \mu_n|\nu) \\
&= \log_2 \int_\Omega \left(\frac{d\mu_1}{d\nu}\frac{d\mu_2}{d\nu}\cdots\frac{d\mu_n}{d\nu}\right)^2 d\nu \\
&= \log_2 \int_\Omega \left(\frac{d\mu_1}{d\nu}\right)^2 \left(\frac{d\mu_2}{d\nu}\right)^2 \cdots \left(\frac{d\mu_n}{d\nu}\right)^2 d\nu \\
&= \log_2 \left[\int_\Omega \left(\frac{d\mu_1}{d\nu}\right)^2 d\nu\right]\left[\int_\Omega \left(\frac{d\mu_2}{d\nu}\right)^2 d\nu\right]\cdots\left[\int_\Omega \left(\frac{d\mu_n}{d\nu}\right)^2 d\nu\right] \qquad [*] \\
&= \log_2 \int_\Omega \left(\frac{d\mu_1}{d\nu}\right)^2 d\nu + \log_2 \int_\Omega \left(\frac{d\mu_2}{d\nu}\right)^2 d\nu + \cdots + \log_2 \int_\Omega \left(\frac{d\mu_n}{d\nu}\right)^2 d\nu \\
&= I(\mu_1|\nu) + I(\mu_2|\nu) + \cdots I(\mu_n|\nu)
\end{aligned}$$

parameter θ, Ronald Fisher defined what is now known as the Fisher information: $J(\theta) = \int \left[\frac{\partial}{\partial\theta} \ln f(x;\theta)\right]^2 f(x;\theta)dx$. By the Cramér-Rao inequality, the mean squared error for any unbiased esitmator T of the parameter θ is bounded below by the reciprocal of the Fisher information, i.e., $\mathbf{V}_\theta(T) \geq \dfrac{1}{J(\theta)}$. Indeed, if $f(x;\theta)$ is normally distributed with mean θ and variance σ^2, then $J(\theta) = \dfrac{1}{\sigma^2}$. Note that in this case the information increases as the variance decreases. The opposite is the case with variational information. Variational information and Fisher information are therefore not equivalent. For more on Fisher information, see Thomas and Cover (1991: 326–331). For the article in which Fisher originally formulated Fisher information, see Fisher (1925).

Note that [*] follows from the previous line by independence and Fubini's theorem.

Although the variational information was, strictly speaking, defined only for pairs of probability measures μ_1 and μ_2 such that $\mu_1 \ll \mu_2$ and such that the Radon-Nikodym derivative $\frac{d\mu_1}{d\mu_2}$ was square integrable with respect to μ_2, in fact the variational information can be defined for all probability measures μ_1 and μ_2 on Ω: in case $\mu_1 \ll \mu_2$ but $\frac{d\mu_1}{d\mu_2}$ is not square integrable with respect to μ_2, define $I(\mu_1|\mu_2) = \infty$; in case μ_1 is not absolutely continuous with respect to μ_2, also define $I(\mu_1|\mu_2) = \infty$ (assigning infinity in the latter case makes good intuitive sense because for μ_1 to assign nonzero probability to an event of probability zero with respect to μ_2 is to render probable under μ_1 what is impossible under μ_2—such an eventuality suggests an infinite infusion of information).

The variational information is a special case of the *Rényi information divergence* (also known as the *Rényi entropy*). For a random variable X defined on Ω and density f induced by X on $\mathbb{R}$, Alfred Rényi defined the quantity

$$h_r(X) =_{def} \tfrac{1}{1-r} \log_2 \textstyle\int_\Omega [f(x)]^r dx$$

for $0 < r < \infty$ and $r \neq 1$. For measures μ_1 and μ_2 such that $\mu_1 \ll \mu_2$, this readily generalizes to

$$h_r(\mu_1|\mu_2) =_{def} \tfrac{1}{1-r} \log_2 \textstyle\int_\Omega \left(\frac{d\mu_1}{d\mu_2}\right)^r d\mu_2.$$

The Rényi information divergence has, as r varies, a wealth of informational measures embedded in it. Most of these, however, are not physically significant. There are two notable exceptions: in the limit as r goes to 1, this quantity is just the standard Shannon entropy of section 2 (as formulated for densities, however, rather than for partitions of Ω). The other exception is the Rényi divergence for $r = 2$, which, obviously, is equivalent to the variational information.[7]

4 Continuity Spectra

The temporal dynamics of many physical systems can be represented as trajectories of measures μ_t for t in some real interval $[a, b]$. In such cases, it makes sense to consider the variational information along these trajectories: $I(\mu_t|\mu_s)$ for $a \leq s < t \leq b$. For instance, within classical mechanics the trajectory of a particle can be represented as a continuous path $x(t)$ in a manifold Ω. As a consequence, it can also be equivalently represented as a continuous path of probability measures qua point masses $\delta_{x(t)}$ (continuity here being according to

[7] For a brief overview of the Rényi entropy/information divergence, see Cover and Thomas (1991: 499–501). For the original formulation by Rényi, see Rényi (1961).

the weak topology—measures converge in the weak topology iff their integrals converge for all bounded continuous real-valued functions). Now, assuming that $x(t)$ doesn't halt or double back on itself (i.e., it is one-to-one), it follows that for all $s \neq t$ in the interval $[a, b]$, $\delta_{x(t)}$ is not absolutely continuous with respect to $\delta_{x(s)}$ and hence $I(\delta_{x(t)}|\delta_{x(s)}) = \infty$. This, however, seems countertintuitive since $\lim_{s\uparrow t} \delta_{x(s)} = \delta_{x(t)}$ in the weak topology. The problem, then, is that the variational information is not keeping track of any topological structure associated with the underlying probability space Ω. Our next task, therefore, is to coordinate the variational information with the topological structure of Ω.

Most of the interesting mathematical work on probability theory focuses on separable metric spaces (specifically, on separable topological spaces that can be metrized with a complete metric—these are known as *Polish* spaces).[8] In the sequel, we therefore focus on the separable metric space Ω with metric D whose open sets induce the Borel σ-algebra $\mathcal{B}$. If we now define $\mathbf{M}(\Omega)$ as the set of all probability measures on $(\Omega, \mathcal{B})$, then $\mathbf{M}(\Omega)$ is itself a separable metric space in the Kantorovich-Wasserstein metric $\overline{D}$ (which induces the weak topology on $\mathbf{M}(\Omega)$). For Borel probability measures μ and v on Ω,

$$\begin{aligned}\overline{D}(\mu, \nu) &= \inf\left\{\textstyle\int D(x, y)\zeta(dx, dy) : \zeta \in \mathbf{P}_2(\mu, \nu)\right\} \\ &= \sup\left\{\left|\textstyle\int f(x)\mu(dx) - \int f(x)\nu(dx)\right| : \|f\|_L \leq 1\right\}\end{aligned}$$

where, in the first equation, $\mathbf{P}_2(\mu, \nu)$ is the collection of all Borel probability measures on $\Omega \times \Omega$ with marginal distributions μ on the first factor and ν on the second, and where, in the second equation, f ranges over all continuous real-valued functions on Ω for which the Lipschitz seminorm is ≤ 1 ($\|f\|_L = \sup\{|f(x) - f(y)|/D(x, y) : x, y \in \Omega,\ x \neq y\}$). Both the infimum and the supremum on the right of these two equations define metrics. The first is called the Wasserstein metric, the second the Kantorovich metric. Though the two expressions appear quite different, they are known to be equal (see Dudley 1976).

The Kantorovich-Wasserstein metric $\overline{D}$ is the canonical extension to $\mathbf{M}(\Omega)$ of the metric D on Ω. It is fair to say that it extends the metric structure of Ω as fully as possible to $\mathbf{M}(\Omega)$. For instance, if δ_x and δ_y are point masses in $\mathbf{M}(\Omega)$, then $\overline{D}(\delta_x, \delta_y) = D(x, y)$. It follows that the canonical embedding of Ω into $\mathbf{M}(\Omega)$, i.e., $x \mapsto \delta_x$, is in fact an isometry. But perhaps the best way to see that $\overline{D}$ scrupulously extends the metric structure of Ω to $\mathbf{M}(\Omega)$ is to consider the following reformulation of this metric.

Let $\mathbf{M}_{av}(\Omega) = \{\frac{1}{n}\sum_{1\leq i\leq n} \delta_{x_i} : x_i \in \Omega,\ n \text{ a positive integer}\}$. It is readily seen that $\mathbf{M}_{av}(\Omega)$ is dense in $\mathbf{M}(\Omega)$ in the weak topology. Note that the x_is are not required to be distinct, implying that $\mathbf{M}_{av}(\Omega)$ consists of all convex

[8] See chapter 8 of Cohn (1996), which is devoted to Polish spaces and analytic sets. See also Billingsley (1999)—the interesting theorems here on the convergence of probability measures are proved for separable metric spaces.

linear combinations of point masses with rational weights; note also that such combinations, when restricted to a countable dense subset of Ω, form a countable dense subset of $\mathbf{M}(\Omega)$ in the weak topology, showing that $\mathbf{M}(\Omega)$ is itself separable in the weak topology.

Now, for any measures μ and v in $\mathbf{M}_{av}(\Omega)$, it is possible to find a positive integer n such that $\mu = \frac{1}{n}\sum_{1\leq i\leq n}\delta_{x_i}$ and $\nu = \frac{1}{n}\sum_{1\leq i\leq n}\delta_{y_i}$. Next, define

$$\overline{D}_{perm}(\tfrac{1}{n}\textstyle\sum_{1\leq i\leq n}\delta_{x_i}, \tfrac{1}{n}\sum_{1\leq i\leq n}\delta_{y_i}) =_{def} \min\{\tfrac{1}{n}\sum_{1\leq i\leq n} D(x_i, y_{\sigma i}) : \sigma \in \mathbf{S}_n\}$$

where $\mathbf{S}_n$ is the symmetric group on the numbers 1 to n. $\overline{D}_{perm}$ looks for the best way to match up point masses for any pair of measures in $\mathbf{M}_{av}(\Omega)$ vis-a-vis the metric D. It is straightforward to show that $\overline{D}_{perm}$ is well-defined and constitutes a metric on $\mathbf{M}_{av}(\Omega)$. The only point in need of proof here is whether for arbitrary measures $\frac{1}{n}\sum_{1\leq i\leq n}\delta_{x_i}$ and $\frac{1}{n}\sum_{1\leq i\leq n}\delta_{y_i}$ in $\mathbf{M}_{av}(\Omega)$, and for any measures $\frac{1}{mn}\sum_{1\leq i\leq mn}\delta_{z_i} = \frac{1}{n}\sum_{1\leq i\leq n}\delta_{x_i}$and $\frac{1}{mn}\sum_{1\leq i\leq mn}\delta_{w_i} = \frac{1}{n}\sum_{1\leq i\leq n}\delta_{y_i}$,

$$\begin{aligned}&\min\{\tfrac{1}{n}\textstyle\sum_{1\leq i\leq n} D(x_i, y_{\sigma i}) : \sigma \in \mathbf{S}_n\} = \\ &\qquad\qquad \min\{\tfrac{1}{mn}\textstyle\sum_{1\leq i\leq mn} D(z_i, w_{\rho i}) : \rho \in \mathbf{S}_{mn}\}.\end{aligned}$$

This equality does in fact hold. Crucial in its proof is Philip Hall's well-known "marriage lemma" from combinatorial theory.

Proposition. *$\overline{D}_{perm} = \overline{D}$ on $M_{av}(\Omega)$.*

Remark. *Because $M_{av}(\Omega)$ is dense in $M(\Omega)$, it follows that $\overline{D}_{perm}$ extends uniquely to $\overline{D}$ on all of $M(\Omega)$.*

Proof. *For $\overline{D}$, let us use $\inf\left\{\int D(x,y)\zeta(dx,dy) : \zeta \in \mathbf{P}_2(\mu,\nu)\right\}$ (i.e., the Wasserstein as opposed to Kantorovich version of the metric). Let $\mu = \frac{1}{n}\sum_{1\leq i\leq n}\delta_{x_i}$ and $\nu = \frac{1}{n}\sum_{1\leq i\leq n}\delta_{y_i}$ be arbitrary measures in $M_{av}(\Omega)$ represented with a common n. Consider $\zeta = \frac{1}{n}\sum_{1\leq i\leq n}\delta_{(x_i,y_i)}$. Then $\zeta \in \mathbf{P}_2(\mu,\nu)$. This is true for any ordering of indices. We therefore assume that the indices are so chosen that $\overline{D}_{perm}(\mu,\nu) = \frac{1}{n}\sum_{1\leq i\leq n} D(x_i,y_i)$. But this is precisely $\int D(x,y)\zeta(dx,dy)$. It follows that $\overline{D}_{perm} \geq \overline{D}$.*

To prove the reverse inequality, consider an arbitrary $\zeta \in \mathbf{P}_2(\mu,\nu)$. ζ is then of the form $\frac{1}{n}\sum_{1\leq i,j\leq n} a_{ij}\delta_{(x_i,y_j)}$ where the $n\times n$ matrix $[a_{ij}]$ is doubly stochastic (this is clear when the x_is and y_is are distinct among themselves; for repetitions we may, by suitably averaging, choose corresponding rows and columns identical, thus yielding a doubly stochastic matrix in the general case). By the Birkhoff theorem, $[a_{ij}]$ can be written as a convex combination of permutation matrices. Thus, for some $t_1,\ldots,t_m > 0$ such that $t_1+\cdots+t_m = 1$ and $n\times n$ permutation matrices $\Pi_1,\ldots,\Pi_m$,

$$[a_{ij}] = t_1\Pi_1 + \cdots + t_m\Pi_m.$$

Corresponding to the permutation matrices $\Pi_1, \ldots, \Pi_m$ *are permutations* $\sigma^1, \ldots,$ σ^m *respectively for which*

$$\begin{aligned}\int D(x,y)\zeta(dx,dy) &= \frac{1}{n}\sum_{1\le i,j\le n} a_{ij}D(x_i,y_j)\\ &= \frac{1}{n}\sum_{1\le r\le m} t_r \sum_{1\le i\le n} D(x_i, y_{\sigma^r i}).\end{aligned}$$

But since $\overline{D}_{perm}(\mu,\nu) = \min\{\frac{1}{n}\sum_{1\le i\le n} D(x_i,y_{\sigma i}) : \sigma \in S_n\}$*, it follows that*

$$\begin{aligned}\int D(x,y)\zeta(dx,dy) &= \frac{1}{n}\sum_{1\le r\le m} t_r \sum_{1\le i\le n} D(x_i, y_{\sigma^r i})\\ &= \sum_{1\le r\le m} t_r \left[\frac{1}{n}\sum_{1\le i\le n} D(x_i, y_{\sigma^r i})\right]\\ &\ge \sum_{1\le r\le m} t_r \overline{D}_{perm}(\mu,\nu)\\ &= \overline{D}_{perm}(\mu,\nu).\end{aligned}$$

Thus $\overline{D}_{perm}(\mu,\nu) \le \overline{D}(\mu,\nu)$*. This proves the result.* ■

With the Kantorovich-Wasserstein metric in hand, our task now is to coordinate it with the variational information. Thus, for $\varepsilon > 0$ and probability measures μ and ν in $\mathbf{M}(\Omega)$, let us define

$$I^{\varepsilon}(\mu|\nu) =_{def} \inf\big\{I(\zeta|\xi) : \zeta,\xi \in \mathbf{M}(\Omega),\ \overline{D}(\zeta,\mu)\le\varepsilon,\ \overline{D}(\xi,\nu)\le\varepsilon\big\}.$$

We call I^{ε} the *variational information modulo* ε. From this definition, it follows that for all positive ε, $I^{\varepsilon}(\mu|\nu) \le I(\mu|\nu)$ and $I^{\varepsilon}(\mu|\nu)$ increases as $\varepsilon \downarrow 0$.

Conjecture. $\lim_{\varepsilon\downarrow 0} I^{\varepsilon}(\mu|\nu) = I(\mu|\nu)$ *for all* μ *and* ν *in* $M(\Omega)$.

This conjecture seems likely to be true, but has no simple proof. The problem is that it is not enough to find measures ζ_n and ξ_n that converge to μ and ν in the weak topology (and thus with respect to the Kantorovich-Wasserstein metric $\overline{D}$). Rather, any such ζ_n and ξ_n must also be suitably chosen so that the former is, in each case, absolutely continuous with respect to the latter, and in a way that keeps $I(\zeta_n|\xi_n)$ as small as possible.

To see what's at stake, consider two point masses on the real line: δ_0 and δ_η for η positive and very close to zero. Then δ_0 and δ_η, though neither is absolutely continuous with respect to the other, are very close in the $\overline{D}$ metric: $\overline{D}(\delta_0, \delta_\eta) = |0 - \eta| = \eta$. Now consider two normal distributions on the real line: $N(0, \varepsilon)$ and $N(\eta, \varepsilon)$—normal distributions with means 0 and η respectively and variance ε (ε much bigger than η). These measures are mutually absolutely continuous. Moreover, they concentrate most of their mass at 0 and η respectively. Because the probability densities of these distributions with respect to Lebesgue measure are $\frac{1}{\sqrt{2\pi\varepsilon}}e^{-\frac{x^2}{\varepsilon}}$ and $\frac{1}{\sqrt{2\pi\varepsilon}}e^{-\frac{(x-\eta)^2}{\varepsilon}}$, it follows that

$$\begin{aligned} I(N(0,\varepsilon)|N(\eta,\varepsilon)) &= \log_2 \int_{-\infty}^{\infty} \left[\left(\frac{1}{\sqrt{2\pi\varepsilon}} e^{-\frac{x^2}{\varepsilon}} \right)^2 \Big/ \left(\frac{1}{\sqrt{2\pi\varepsilon}} e^{-\frac{(x-\eta)^2}{\varepsilon}} \right) \right] dx \\ &= \log_2 \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\varepsilon}} e^{-\frac{(x^2+2x\eta-\eta^2)}{\varepsilon}} dx \end{aligned}$$

This last integral converges to 1 as $\eta \downarrow 0$, and so the variational information of $N(0, \varepsilon)$ given $N(\eta, \varepsilon)$ goes to 0 as $\eta \downarrow 0$. Note that to obtain this result, we needed to convert point masses, which were not mutually absolutely continuous, to suitable probability measures concentrated around those points that were mutually absolutely continuous. To prove the full conjecture requires generalizing this approach to arbitrary measures on arbitrary separable metric spaces.

Yet regardless of whether this conjecture is true, it does not affect how we assess the continuity of probability paths. We define the *continuity spectra* of a probability path μ_t for $t \in [a, b]$ as as two real-valued functions on this interval. The *left continuity spectrum* is the map that assigns to $t \in [a, b]$ the value

$$I^-(t) =_{def} \lim_{\varepsilon\downarrow 0} \lim_{s\uparrow t} I^\varepsilon(\mu_t|\mu_s).$$

Alternately, the *right continuity spectrum* assigns to $t \in [a, b]$ the value

$$I^+(t) =_{def} \lim_{\varepsilon\downarrow 0} \lim_{s\downarrow t} I^\varepsilon(\mu_s|\mu_t).$$

A probability path is by definition *informationally continuous* iff both the left and right continuity spectra are identically zero. It is *informationally left continuous* iff the left continuity spectrum is identically zero. It is *informationally right continuous* iff the right continuity spectrum is identically zero.

It is immediate that if $x(t)$ is a continuous path in a Riemannian manifold Ω with metric D induced by the Riemannian metric, then $\mu_t = \delta_{x(t)}$ is informationally continuous (in order to approximate the variational information modulo

ε, substitute $\frac{1}{\lambda(B_\varepsilon(x(t))}1_{B_\varepsilon(x(t))}d\lambda$ for $\delta_{x(t)}$; here λ is the measure derived from volume element induced by the Riemannian metric).

On the other hand, if μ_t describes the probability distribution associated with a state of a quantum system that is evolving according to the Schrödinger equation, the associated left continuity spectrum will be nonzero at precisely those places where the system assumes an eigenstate. For instance, consider a photon hitting a filter polarized at a 45 degree angle. Before hitting the filter, it had a fifty-fifty chance of passing through the filter. Suppose it passes through the filter. Then $I^-(t) = 1$ at t the instant that the photon hits the filter.

There is an irony here. In classical physics, the dynamics of a system is described in terms of discrete particles, characterized mathematically as points, that move continuously in a state space. Represented measure-theoretically, these points correspond to distinct point masses that are not absolutely continuous with respect to one another. And yet, the dynamics of these systems, when represented not just measure-theoretically but also in terms of the continuity spectra corresponding to the underlying topology, are informationally continuous. By contrast, in quantum physics, the dynamics of a system is described in terms of quantum states that induce probability densities that are everywhere nonzero because quantum processes cannot be meaningfully localized (see Gordon 2002). In consequence, the states induce probability measures that are always mutually absolutely continuous; and yet, when a measurement is taken, the new probability densities shift so dramatically that the dynamics becomes informationally discontinuous.

The intuition underlying the continuity spectra is straightforward: observers trying to assess information are always limited in their powers of observation by the degree to which they can discriminate between distinct states of affairs.[9] If two states of affairs, though distinct in actuality, nonetheless are indistinguishable because our powers of observation do not enable us to resolve the difference, then they effectively convey the same information, and, so, one state of affairs adds no new information to the other. Accordingly, our mathematics should be as parsimonious as possible in assessing information, always preferring among indistinguishable states of affairs those that assign less information. To do otherwise is to inflate our assessments of information because of idiosyncrasies in the way we mathematically represent states of affairs rather than because of any intrinsic difference that our powers of observation are able to ascertain and that our mathematics is able faithfully to capture. In practice, we characterize our powers of observation to resolve distinct states of affairs in terms of an ε-tolerance factor. Accordingly, measurements closer than some positive ε are effectively indistinguishable. This accounts for the definition of the continuity

[9] Consider the following statement in the experimental psychology literature: "It is a well-known fact that there are limits to the revolving power of the subject. Given a series of stimuli which differ with respect to some discriminable aspect—some psychological attribute—it is possible to select two stimuli which are so close together on the continuum that the subject cannot report with any confidence which is the greater." Quoted from Torgerson (1958: 132). This book, though dated, contains useful insights relevant to the continuity spectrum, especially in chapter 7, titled "The Differential-Sensitivity Methods."

spectra in terms of the variational information modulo ε.

To sum up, the variational information does not take into account the metric structure of the underlying probability space and therefore suggests far more discontinuity than is actually present in probability paths. It's therefore necessary to factor in the metric structure of the underlying probability space, and this is properly done by basing continuity spectra on the Kantorovich-Wasserstein metric. These continuity spectra, because they are based on the Kantorovich-Wasserstein metric and because this metric canonically extends the metric on the underlying probability space Ω, define the canonical information topology for probability paths. This information topology makes sense independently of any additional information geometry that may be defined on $\mathbf{M}(\Omega)$. It is an open question in what sense the information geometries defined to date are consistent with this information topology for probability paths (cf. Amari and Nagaoka 2000).

5 Practical and Scientific Significance

We began this study by examining how to measure the information of an event A with probability $\mathbf{P}(A)$ and concluded that the appropriate information measure in that case was $I(A) = -\log_2 \mathbf{P}(A)$. This definition extended naturally to pairs of events A and B, with $I(A|B) = -\log_2 \mathbf{P}(A|B)$. This measure, however, has only played an ancillary role in the mathematical theory of information as developed by communication engineers (notably the line of research initiated by Claude Shannon). Communication engineers need to assess the information contained not in individual events but rather in ensembles of events that partition a reference class of possibilities, where these possibilities are typically thought to correspond to possible messages. This accounts for the preeminent role of entropy in the mathematical theory of information, which provides an averaged measure of information. In place of events A and B, entropy looks to partitions of events $\mathfrak{A} = \{A_1, \ldots, A_m\}$ and $\mathfrak{B} = \{B_1, \ldots, B_m\}$ and assigns an averaged information measure $H(\mathfrak{A}) = -\sum_i \mathbf{P}(A_i) \log_2 \mathbf{P}(A_i) = \sum \mathbf{P}(A_i) I(A_i)$ and $H(\mathfrak{A}|B) = -\sum_{i,j} \mathbf{P}(A_i B_j) \log_2 \mathbf{P}(A_i|B_j) = \sum_{i,j} \mathbf{P}(A_i B_j) I(A_i|B_j)$.

In defining H in terms of I, communication engineers have not so much generalized the information measure I as merely applied it to multiple events. By contrast, extending I to the variational information significantly enriches I. It is truly a generalization because events A map canonically to probability measures $\frac{1}{\mathbf{P}(A)} 1_A d\mathbf{P}$, and the variational information applied to such probability measures equals the ordinary information measure applied to the corresponding events. It is, moreover, the canonical generalization in the sense that the variational information is the one instance of the Rényi information divergence that faithfully extends the ordinary information associated with individual events.

The variational information, though implicit as a special case in the Rényi information divergence ($r = 2$), has to date gone largely unnoticed and unappreciated. That's unfortunate because it comes up in perfectly ordinary contexts. Consider the prospect of rain or not rain. If our approach to information is limited to events, then given that we are in Seattle and given that the probability of rain is .99 and the probability of no rain is .01, the amount of information gained if we learn that it is sunny and not raining is $-\log_2 .01 = 6.6439$ bits. But now consider a different scenario. This time we learn not that it isn't raining, but rather that we were wrong in thinking that we were in Seattle in the first place, and that, instead, we are in the Sahara desert where the probability of rain is .01 and the probability of no rain is .99. In that case, a simple calculation shows that the variational information for the probability of rain or no rain in the Sahara desert given the earlier probability of rain or no rain in Seattle is $\log_2 98.0101 = 6.6149$ bits (the number 98.0101 being the integral of the square of the pertinent Radon-Nikodym derivative). In other words, we acquired over six and a half bits of information in learning that the pattern of rain is no longer Seattle's but rather the Sahara desert's. The reference to "bits" here is entirely appropriate: because the variational information is the canonical extension to probability measures of the ordinary information for events, it makes sense to think of the numerical values delivered by the variational information as bits.

The variational information not only gives a precise and privileged mathematical sense to the information we gain as our knowledge of probabilities changes but also makes good intuitive sense of the information associated with changing probabilities. Nevertheless, the real challenge facing the variational information is to turn it into a practical tool that's useful for science. What follows are four potentially fruitful areas of application:

Example 1 *Employing the variational information not just to track changes in probabilities but also to track changes in the structure, configuration, and dynamics of physical systems. These aspects of physical systems, though often capable of being modeled by means of probability measures, need not be interpreted probabilistically (in the sense of, for instance, sampling a random variable) but can instead be interpreted information-theoretically in terms of generalized bits. In the absence of a straightforward probabilistic interpretation, how helpful might this approach be for generating scientific insights?*

Example 2 *Deriving variational and least action solutions by minimizing (or maximizing) the variational information when it is applied to suitably indexed or parameterized probability measures. Roy Frieden (1998) has, mutatis mutandis, derived a good deal of physics from Fisher information, including many standard results from the calculus of variations. How much of physics can be derived from the variational information?*

Example 3 *Distinguishing scientific theories in terms of informational continuity and discontinuity. Classical physics consistently yields continuous information spectra. By contrast, quantum physics yields discontinuous information spectra. Likewise, classical evolutionary theories à la Darwin are gradualistic and suggest continuous information spectra whereas saltational approaches to evolution suggest discontinuous information spectra. To what extent can variational information make this distinction rigorous and provide genuine insights into the processes responsible for life's evolutionary history?*

Example 4 *Assessing the sensitivity to perturbation as well as the robustness of biophysical laws. Laws governing physics and biology seem fine-tuned to bring about interesting features that would be absent if the laws were slightly different. Alternatively, there are many features of the biophysical world that seem largely insensitive to the contingencies of natural history. For instance, paleontologist Simon Conway Morris (2003) finds that evolution reinvents the same organic structures over and over and concludes that evolution is robustly constrained to proceed along a limited number of fixed paths. Can the variational information be used to gain insight into the sensitivity to perturbation as well as the robustness of biophysical laws?*

Acknowledgment. I'm grateful to Cosma Shalizi for a helpful critique of an earlier draft of this paper and especially for pointing out the connection between the variational information and the Rényi information divergence. I also wish to thank the contributors to iscid.org, arn.org, and pandasthumb.org for their continued valuable feedback on my work.

References

[1] Adami, Christof, "Information Theory in Molecular Biology," *Physics of Life Reviews* 1 (2004): 3–22.

[2] Amari, Shun-ichi, and Hiroshi Nagaoka, *Methods of Information Geometry*, Translations of Mathematical Monographs, vol. 191 (Providence, R.I.: American Mathematical Society, 2000).

[3] Billingsley, Patrick, *Convergence of Probability Measures*, 2nd ed. (New York: Wiley, 1999).

[4] Cohn, Donald L., *Measure Theory* (Boston: Birkhäuser, 1996).

[5] Conway Morris, Simon, *Life's Solution: Inevitable Humans in a Lonely Universe* (Cambridge: Cambridge University Press, 2003).

[6] Cornfeld, I. P., S. V. Fomin, and Ya. G. Sinai, *Ergodic Theory* (New York: Springer, 1982).

[7] Cover, Thomas M., and Joy A. Thomas, *Elements of Information Theory* (New York: Wiley, 1991).

[8] Dretske, Fred, *Knowledge and the Flow of Information* (Cambridge, Mass.: MIT Press, 1981).

[9] Dudley, R. M., *Probability and Metrics* (Aarhus: Aarhus University Press, 1976).

[10] Fisher, Ronald A., "Theory of Statistical Estimation," *Proceedings of the Cambridge Philosophical Society* 22 (1925): 700–725.

[11] Frieden, B. Roy, *Physics from Fisher Information: A Unification* (Cambridge: Cambridge University Press, 1998).

[12] Gordon, Bruce, "Maxwell-Boltzmann Statistics and the Metaphysics of Modality," *Synthese* 133(3) (2002): 393–417.

[13] Kullback, Solomon, *Information Theory and Statistics* (New York: Dover, 1997).

[14] Ornstein, Donald S., "Bernoulli Shifts with the Same Entropy Are Isomorphic," *Advances in Mathematics* 4 (1970): 337–352.

[15] Rennolls, Keith, "Likelihood, Entropy, and Species Diversity; Some Comparisons in a Sumatran Forest," typescript, 2000, available online at http://cms1.gre.ac.uk/conferences/iufro/proceedings/Rennolls1Diversity.pdf (last accessed 11 August 2004).

[16] Rényi, Alfred, "On Measures of Information and Entropy," in J. Neyman, ed., *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1 (Berkeley, Calif.: University of California Press): 547–561.

[17] Stalnaker, Robert, *Inquiry* (Cambridge, Mass.: MIT Press, 1984).

[18] Torgerson, Warren S., *Theory and Methods of Scaling* (New York: Wiley, 1958).

[19] von Baeyer, Hans Christian, *Information: The New Language of Science* (Cambridge, Mass.: Harvard University Press, 2004).

[20] Weaver, Warren, "Recent Contributions to the Mathematical Theory of Communication," the introductory essay in Claude Shannon and Warren Weaver, *The Mathematical Theory of Communication* (Urbana, Ill.: University of Illinois Press, 1949).

[21] Yockey, Hubert, *Information Theory and Molecular Biology* (Cambridge: Cambridge University Press, 1992).