# Toward a Theory of Information Processing

Sinan Sinanović and Don H. Johnson*
Computer and Information Technology Institute
Department of Electrical and Computer Engineering
Rice University
Houston, Texas 77251–1892
713.348.4956 713.348.5686 (FAX)
sinan@rice.edu, dhj@rice.edu

February 28, 2004

**Abstract**

Information processing theory endeavors to quantify how well signals encode information and how well systems, by acting on signals, process information. We use information-theoretic distance measures, the Kullback-Leibler distance in particular, to quantify how well signals represent information. The ratio of distances between a system's output and input quantifies the system's information processing properties. Using this approach, we derive the fundamental processing capabilities of simple system architectures without detailing the component systems and the signals they process.

# 1 Introduction

In trying to understand how neurons encode information,[1] neuroscientists must cope with many uncertainties. How the component systems are interconnected is in most cases only broadly known. Consequently, the importance or role recorded signal(s) play within the neural information processing structures are unknown. Given only a broad context, neuroscientists want to determine whether recorded signals represent information and if so, what information is represented and what is the quality of the representation (to what fidelity can the information be extracted). Classic information theory [26] has few answers to these representation and fidelity questions. For example, entropy is commonly used to assess a signal's "uncertainty," but signals have an entropy regardless of whether they are bear information or not and whether the information is relevant or not. Assuming an answer to the information-representation question can be found, rate distortion theory [5, 20, 26] provides an approach to answering the fidelity question. Unfortunately, using this approach requires knowledge of the joint probability distribution of the encoded information and its decoded version. Because of the elusive nature of the question "What is information?" and because the signal being studied may or may not be the ultimate output, determining the required joint probability function is quite difficult. Furthermore, rate distortion theory rests on specifying a distortion measure and on computing the mutual information between information and signals. One of the reasons neuroscientists are making recordings is to infer the inherent distortion measure. More importantly, computing mutual information requires a probabilistic model be assigned to information. We question the ultimate validity of creating a probabilistic model for information. Is the information contained in Shannon's paper random? What we mean by information here is not the sequence of words and phrases, but rather the *meaning* of the prose. If the information —the meaning— is random, what kind of random quantity is it and from what probability distribution was it drawn? For all of these problems and other reasons, we had to rethink how to approach the analysis of information processing systems. We need an "information probe" that can examine any kind of signal found in a complex system and determine how well it conveys relevant information.

In this paper, we frame a theory that recasts how one approaches understanding the information representation of signals and the information processing capabilities of systems. Our theory weds results from information theory and statistical signal processing. This preliminary theory complements classic information theory, which answers questions different than the ones posed here. Warren Weaver described in his introduction to Shannon's classic paper [27] that while Shannon's

[1] In neuroscience, the terms "encode" and "decode" are commonly used. In communication theory and signal processing, these terms are most akin to "modulation" and "demodulation."

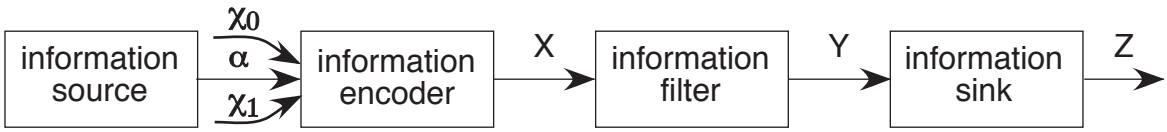


**Figure 1**: An information source produces information considered relevant, an abstract quantity represented by the symbol $\alpha$. This information, as well as extraneous information represented by $\chi_0$ and $\chi_1$ is encoded (modulated) onto the signal $X$, which we assume to be stochastic. A system, having its input-output relationship defined by the conditional probability function $p_{Y|X}(Y|X)$, serves as an information filter, changing the fidelity with which the information is represented by its output, possibly accentuating some aspects of the information while suppressing others. The information sink responds to the information encoded in $Y$ by exhibiting an action $Z$.

work was *the* mathematical theory of communication, it did not touch the entire realm of information processing notions that require analysis. He stratified communication problems on three levels: *technical*, *semantic*, and *influential*. He cast Shannon's work as the engineering or technical side of the problem because it did not deal with the extraction of meaning and with taking actions in response to the transmitted information. Our theory takes the approach that the meaning of signals can only be determined from actions taken as a result of interpreting the information. Thus, we merge semantic interpretation into quantifying the effectiveness of action (i.e., influence). In our framework, information processing ultimately results in actions, and these actions result in signals of some sort.

Tishby [28] has developed the information bottleneck approach, wherein rate-distortion theory, the Data Processing Theorem, and compression play major roles. Here, mutual information is the primary quantity, which we show in the Summary to be an inconvenient choice from both theoretical and empirical viewpoints. We use information-theoretic generalized distance measures between probability distributions that describe stochastic signals. We require viable distance measures to be analytically tractable, empirically measurable from *any* kind of signal, and related to the optimal performance of the information processing scenarios of classification and estimation. Using one particular measure, the Kullback-Leibler distance, we have created a system theory of information processing structures which quantifies the inherent processing capabilities of several system architectures for information processing.

## 2 Assumptions

Our theory rests on several fundamental assumptions. We use a standard communication-like model as a backdrop (figure 1).

- *Information can have any form and need not be stochastic*.

In general, we make no assumption about what form information takes. We represent it by the symbol $\alpha$ that we take to be a member of a set (either countable or uncountable). This symbol represents the *meaning* of the information that signals encode. Information uninteresting to the information sink may also be present; the symbols $\chi_0$, $\chi_1$ in figure 1 represent extraneous information.

- *Signals encode information.*
  Information does not exist in a tangible form; rather information is always *encoded* into a signal. Signals representing the same information can take on very different forms. For example, the text you are reading now could also have been read aloud to you. Thus, the text's alphabetic/punctuation symbol sequence, the image of a page and the air pressure at your ears each represent the same information. In the first case, the signal is a sequence drawn from a discrete, finite alphabet; in the latter two, the signal is analog. Any viable information processing theory must place a variety of signals on the same footing and allow comparison of the various ways information can be encoded.

- *What may or may not be information is determined by the information sink.*
  The recipient of the information —the information sink —determines what about a signal it receives is informative. Additional, often extraneous, information $\chi_0$, $\chi_1$ is represented by signals. Returning to the example describing the various ways in which the information in this paper could be represented, information extraneous to the paper's central theme is also contained in the signal. The text comes from a Roman alphabet expressing English prose. The page image reveals page numbers, fonts and details of the mathematical notation. The spoken text encodes the gender of the reader and dialect, which provides clues as to his/her region of origin. Such information may or may not be considered extraneous by some succeeding processing. For example, speaker identification systems are presumably designed to ignore what is being said. It is for the recipient to determine what information is. An immediate consequence of this assumption is that *no single objective measure can quantify the information contained in a signal*. "Objective" here means analysis of a signal out of context and without regard to the recipient.

- *When systems act on their input signal(s) and produce output signal(s), they indirectly perform information processing.*
  Systems map signals from their input space onto their output space. Because information does not exist as an entity, systems affect the information encoded in their inputs by their

input-output map. We can assess the effect a system has on information-bearing signals by comparing the fidelity of the relevant information represented by the input and output signals.

- *The result of processing information is an action.*
  An action is a quantifiable behavior taken by the information sink when it interprets the signal's information content. We assume that every action is a signal. For example, the influence of a voice command can be measured by observing the recipient's consequent behavior.

# 3 Quantifying Information Processing

In this preliminary information processing theory, we consider information to be a parameter $\alpha$ or a parameter vector $\boldsymbol{\alpha}$ that controls signal features. In this paper, we take the parameters to be real-valued, but in general they could be complex or even symbolic. While parameterizing information reduces the generality of the approach, it helps us develop results that can be applied to many information-processing systems.

## 3.1 Quantifying signal encoding

As shown in figure 1, let $X$ denote a signal that *encodes* information represented by the parameter $\alpha$ as well as $\chi_0$ and $\chi_1$. We assume that signals in our theory are stochastic, with their statistical structures completely specified by the probability functions (either a density or a mass function) $p_X(x;\alpha)$ that depends on the information parameter $\alpha$. Implicitly, this probability function also depends on the extraneous information. This notation for a signal's probability function could be misinterpreted as suggesting that signals are simple random variables. Rather, the notation is meant to succinctly express that signals could be of any kind. We assume that signals are observed over some interval in the signal's domain and that its probability function expresses the joint probability distribution of these observations. The dimensionality of the signal domain is suppressed notationally because our results do not depend on it. Consequently, speech, image and video signals are all represented by the same generic symbol. For discrete domains, we use the joint distribution of the signal's values for the required probability function. We include continuous-domain processes having a Karhunen-Loève expansion [3, §1.4] because they are specified by the joint distribution of the expansion coefficients (see Appendix Appendix B). Those that don't have a Karhunen-Loève expansion are also included because of a result due to Kolmogorov [3, §1.14]. Later, we will need to explicitly specify when the signal has multiple components (i.e., multi-channel signals); in this case the signal will be written in boldface.

Because we have argued that analyzing signals for their relevant information content statically is a hopeless task, we conceptually (or in reality for empirical work) consider controlled changes of the relevant and/or irrelevant information and determine how well the signals encode this information change [14]. By considering induced information changes, we essentially specify what constitutes information. We quantify how well the information is represented by calculating how different are the signals corresponding to the two information states. In this way, we have what amounts to an information "probe" that can be used anywhere in an information system architecture to quantify the processing. Because the signals can have an arbitrary form, usual choices for assessing signal difference like mean-squared error make little sense. Instead, we rely on distance measures that quantify how different the signals' probabilistic descriptions are. Basseville reviews of distance measures for signal processing applications [4].

When we change the information from a value of $\alpha_0$ to $\alpha_1$, the change in $X$ is quantified by the distance $d_X(\alpha_0, \alpha_1)$, which depends not on the signal itself, but on the probability functions $p_X(x;\alpha_0)$ and $p_X(x;\alpha_1)$. We require potential distance measures to satisfy $d(\alpha_0, \alpha_1) \geq 0$ with $d(\alpha, \alpha) = 0$. Conceptually, we want distance to be related to the ability to discern the information change from the signal. Small distances mean the changed aspect of the information would be difficult to ascertain from the signal; large distances would mean the information change would be easy to determine. We do not require viable distance measures to be monotonically related to the information change because doing so would imply that information has an ordering; since we include symbolic information, imposing an ordering would be overly restrictive. We envision calculating or estimating the distance for several values of $\alpha_1$ while maintaining a reference point $\alpha_0$. In this way, we can explore how information changes with respect to a reference correspond to distance changes. We have found that a systematic empirical study of a signal's information representation requires that several reference points be used. This requirement would seem excessive, but the geometries of both estimation and classification [2] clearly state that information extraction performance depends on the reference. Consequently, a thorough characterization of information encoding demands that the information space be explored in this fashion.

A widely use class of distances measures for probability functions is known as the Ali-Silvey class [1]. Distances in this class have the form $d_X(\alpha_0, \alpha_1) = f(\mathcal{E}_0[c(\Lambda(X))])$, where $\Lambda(\cdot)$ represents the likelihood ratio $p_X(\cdot;\alpha_1)/p_X(\cdot;\alpha_0)$, $c(\cdot)$ is convex, $\mathcal{E}_0[\cdot]$ denotes expected value with respect to the distribution given by the parameter $\alpha_0$ and $f(\cdot)$ is a non-decreasing function. Thus, each distance measure in this class are defined by the choices for $f(\cdot)$ and $c(\cdot)$. Because we require $d_X(\boldsymbol{\alpha}, \boldsymbol{\alpha}) = 0$, we restrict attention here on those distances that satisfy $f\big(c(1)\big) = 0$. Among the

many distance measures in this class are the Kullback-Leibler distance and the Chernoff distance. The Kullback-Leibler (KL) distance is defined to be

$$d_X(\alpha_0, \alpha_1) = \mathcal{D}_X(\alpha_1 \| \alpha_0) \equiv \int p_X(x;\alpha_1) \log \frac{p_X(x;\alpha_1)}{p_X(x;\alpha_0)} \, dx \ . \tag{1}$$

While the choice of the logarithm's base is arbitrary, $\log(\cdot)$ denotes the natural logarithm unless otherwise stated. The KL distance is not necessarily symmetric in its arguments, which means that it cannot serve as a metric. To create the Kullback-Leibler distance within this framework, $c(x) = x \log x$ and $f(x) = x$. In passing, Appendix Appendix A demonstrates that the Kullback-Leibler distance is information-theoretic. Another distance measure of importance here is the Chernoff distance [8].

$$\mathcal{C}_X(\alpha_0, \alpha_1) = \max_{0 \le t \le 1} -\log \mu(t), \ \mu(t) = \int [p_X(x;\alpha_0)]^{1-t} [p_X(x;\alpha_1)]^t \, dx \tag{2}$$

It too is in the Ali-Silvey class for each value of $t$, with $c(x) = -x^t$ and $f(x) = -\log(-x)$. Because of the optimization in its definition, the Chernoff distance itself is not in the Ali-Silvey class. A special case of the Chernoff distance is the Bhattacharyya distance [6, 19] $\mathcal{B}_X(\alpha_0, \alpha_1) = -\log \mu(\frac{1}{2})$, which is in the Ali-Silvey class.

As the Kullback-Leibler distance demonstrates, the term "distance" should not be taken rigorously; all of the distances defined here do not obey some of the fundamental axioms true distances must satisfy. The Kullback-Leibler distance is not symmetric, and the Chernoff and Bhattacharyya distances do not satisfy the triangle inequality [19]. In fact, $\mathcal{D}_X(\alpha_1 \| \alpha_0)$ is taken to mean the distance from $p_0$ to $p_1$; because of the asymmetry, the distance from $p_1$ to $p_0$, $\mathcal{D}_X(\alpha_0 \| \alpha_1)$, is usually different. Despite these difficulties, recent work has shown that the Kullback-Leibler distance is geometrically important [2, 12, 31]. If a manifold of probability distributions were created so that distribution pairs having equivalent optimal classifier defined manifold invariance, no distance metric can exist for the manifold because distance must be an asymmetric quantity. The Kullback-Leibler distance takes on that role for this manifold. Delving further into this geometry yields a relationship between the Kullback-Leibler and Chernoff distance measures [10]. On the manifold, the geodesic curve $p_t$ linking two given probability distributions $p_X(x;\alpha_0)$ and $p_X(x;\alpha_1)$ is given by

$$p_t(x) = \frac{[p_X(x;\alpha_0)]^{1-t}[p_X(x;\alpha_1)]^t}{\mu(t)} \ , 0 \le t \le 1,$$

where $\mu(t)$ is defined in equation (2).[2] Consider the halfway point on the manifold defined according to the Kullback-Leibler distance as the distribution equidistant from the endpoints:

[2]This geometric theory, though written in terms of marginal distributions, applies to joint probability distributions as well.

$\mathcal{D}(p_{t^*}\|p_0) = \mathcal{D}(p_{t^*}\|p_1)$. Equating these distances yields $t^*$ as the parameter value that maximizes $-\log \mu(t)$ with the halfway-distance being the Chernoff distance: $\mathcal{C}(p_0, p_1) = \mathcal{D}(p_{t^*}\|p_0)$. The Bhattacharyya distance essentially chooses "halfway" to mean $t = \frac{1}{2}$.

These distance measures satisfy the following important properties.

1. These three distances have the *additive* property: The distance between two joint distributions of statistically independent, identically distributed random variables equals the sum of the marginal distances. Note that because of the optimization step, the Chernoff distance is *not* additive when the random variables are not identically distributed; the Kullback-Leibler and Bhattacharyya distances are.

2. Through Stein's Lemma [9], the Kullback-Leibler and Chernoff distances are the exponential rates of optimal classifier performance probabilities. If $\mathbf{X}$ is a random vector having $N$ statistically independent and identically distributed components under both of the distributions $p_0$, $p_1$, the optimal (likelihood ratio) classifier results in error probabilities that obey the asymptotics

$$\begin{aligned}\lim_{N\to\infty} \frac{\log P_F}{N} &= -\mathcal{D}_X(\alpha_1\|\alpha_0) \text{ , fixed } P_M \\ \lim_{N\to\infty} \frac{\log P_e}{N} &= -\mathcal{C}_X(\alpha_0, \alpha_1) \\ \lim_{N\to\infty} \frac{\log P_e}{N} &\leq -\mathcal{B}_X(\alpha_0, \alpha_1)\end{aligned}$$

   Here, $P_F$, $P_M$, and $P_e$ are the false-alarm, miss, and average-error probabilities, respectively. Loosely speaking, Stein's Lemma suggests that these error probabilities decay exponentially in the amount of data available to the likelihood-ratio classifier: for example, $P_F \sim \exp\{-N\mathcal{D}_X(\alpha_1\|\alpha_0)\}$ for a Neyman-Pearson classifier [18]. The relevant distance determines the rate of decay. Whether all Ali-Silvey distances satisfy a variant of Stein's Lemma is not known. The *only* distances that can be asymptotically related directly to the error probabilities of optimal classifiers are the Kullback-Leibler and Chernoff distances. Because of the form of Stein's Lemma, these distances are known as the *exponential rates* of their respective error probabilities.

3. All Ali-Silvey distances have the property that the first partial of $d_X(\boldsymbol{\alpha}, \boldsymbol{\alpha}_0)$ with respect to each component of $\boldsymbol{\alpha}$ is zero and that their Hessians are proportional to the Fisher information

matrix when they are evaluated at $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$.

$$\left.\frac{\partial d_X(\boldsymbol{\alpha}, \boldsymbol{\alpha}_0)}{\partial \alpha_i}\right|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_0} = 0$$
$$\left.\frac{\partial^2 d_X(\boldsymbol{\alpha}, \boldsymbol{\alpha}_0)}{\partial \alpha_i \partial \alpha_j}\right|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_0} = f'\left(c(1)\right) c''(1) \left[\mathbf{F}_X(\boldsymbol{\alpha}_0)\right]_{i,j}$$

Consequently, for perturbational changes in the parameter vector, $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_0 + \boldsymbol{\delta\alpha}$, the distance between perturbed stochastic models is proportional to a quadratic form consisting of the perturbation and the Fisher information matrix evaluated at $\boldsymbol{\alpha}_0$.

$$d_X(\boldsymbol{\alpha}_0 + \boldsymbol{\delta\alpha}, \boldsymbol{\alpha}_0) \propto \boldsymbol{\delta\alpha}' \mathbf{F}(\boldsymbol{\alpha}_0) \boldsymbol{\delta\alpha} \tag{3}$$

The constant of proportionality equals $1/2$ for the Kullback-Leibler and Bhattacharyya distances, and equals $\left(t^* - (t^*)^2\right) /$ for the Chernoff distance. This property is known as the locally Gaussian property of distance measures. In the special case wherein $X$ is Gaussian with only the mean $m$ depending on the information parameter, all of the three distance measures of interest here (and many others) have the form

$$d_X(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_0) \propto \left(m(\boldsymbol{\alpha}_1) - m(\boldsymbol{\alpha}_0)\right)' \mathbf{F} \left(m(\boldsymbol{\alpha}_1) - m(\boldsymbol{\alpha}_0)\right)$$

for all choices of the information parameter (whether perturbationally different or not). Consequently, the distance between perturbationally different probability functions resembles the distance between Gaussian distributions differing in the mean. The sole difference is the Fisher information, which does depend on the probability function and how it depends on $\boldsymbol{\alpha}$. The reason this property is important because of the Cramér-Rao bound, a fundamental bound on the mean-squared estimation error. The bound states that for scalar parameter changes, the mean-squared error in estimating the parameter when it equals $\alpha_0$ can be no smaller than the reciprocal Fisher information. Thus, the larger the distance for a perturbational change in the parameter, the greater the Fisher information, and the smaller the estimation error can be.

The last two properties directly relate our distances to the performances of optimal classifiers and optimal parameter estimators. By analyzing distances, we at once assess how well two informationally different situations can be distinguished and how well the information parameter can be estimated through observation of the signal $X$.

We focus here on the KL distance and measures related to it. We use the KL distance because its computational and analytic properties are superior to those of the Chernoff distance. For example, estimating the Chernoff distance from data would require solving an optimization problem and the

Chernoff distance does not have this additivity property. We have a complete empirical theory [17] that frames how to estimate KL distances from data, examples of which can be found in [23, 25].

Despite the Kullback-Leibler distance's computational and theoretical advantages, what becomes a nuisance in applications is its lack of symmetry. We have found a simple geometric relationship among the distances measures described here. This relationship also leads to a symmetric distance measure related to the Kullback-Leibler distance. Although Jeffreys [13] did not develop it to symmetrize the Kullback-Leibler distance, the so-called $J$-divergence equals the average of the two possible Kullback-Leibler distances between two probability distributions. [3] Assuming the component Kullback-Leibler distances exist,

$$\mathcal{J}_X(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1) = \frac{\mathcal{D}_X(\boldsymbol{\alpha}_0 \| \boldsymbol{\alpha}_1) + \mathcal{D}_X(\boldsymbol{\alpha}_1 \| \boldsymbol{\alpha}_0)}{2} .$$

We now have a symmetric quantity that is easily calculated and estimated and is in the Ali-Silvey class ($c(x) = \frac{x-1}{2} \log x$). However, its relation to classifier performance is more tenuous than the other distances [4, 19].

$$\lim_{N \to \infty} \frac{\log P_e}{N} \geq -\mathcal{J}_{p_0}(p_1,$$

) We have found this bound to be loose, with it not indicating well the exponential rate of the average-error probability $P_e$ (which is equal to the Chernoff distance). Beyond simple averaging are the geometric and harmonic means. The geometric mean $\mathcal{G}_X(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1) = \sqrt{\mathcal{D}_X(\boldsymbol{\alpha}_0 \| \boldsymbol{\alpha}_1)\, \mathcal{D}_X(\boldsymbol{\alpha}_1 \| \boldsymbol{\alpha}_0)}$ does not seem have as interesting properties as the harmonic mean. We define a new symmetric distance, what we call the *resistor-average* distance, via the harmonic mean.

$$\frac{1}{\mathcal{R}_X(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1)} \equiv \frac{1}{\mathcal{D}_X(\boldsymbol{\alpha}_0 \| \boldsymbol{\alpha}_1)} + \frac{1}{\mathcal{D}_X(\boldsymbol{\alpha}_0 \| \boldsymbol{\alpha}_1)} \tag{4}$$

This quantity gets its name from the formula for the equivalent resistance of a set of parallel resistors: $1/R_{\text{equiv}} = \sum_n 1/R_n$. It equals the harmonic sum (half the harmonic mean) of the component Kullback-Leibler distances. The resistor-average is not an Ali-Silvey distance, but because of its direct relationship to the Kullback-Leibler distance, it is locally Gaussian (as is the $J$-divergence and the geometric mean). The resistor-average distance is not additive in either the Markov or the statistically independent cases; because it is directly computed from quantities that are (Kullback-Leibler distances), it shares the computational and interpretative attributes that additivity offers.

The relation between the various distance measures can be visualized graphically (Figure 2). Because $-\log \mu(t)$ is concave, the resistor-average distance upper bounds the Chernoff distance:

[3] Many authors, including Jeffreys, define the $J$-divergence as the sum rather than the average. Using the average fits more neatly into the graphical relations developed subsequently.

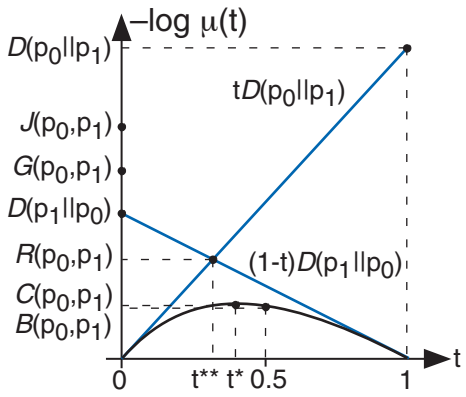


**Figure 2**: This figure portrays relations among many of the most frequently used information-theoretic distances. The focus is the function $-\log \mu(t)$ used to define the Chernoff distance (2). This curve's derivatives at $t = 0$ and $t = 1$ are $\mathcal{D}_X(\boldsymbol{\alpha}_0 \| \boldsymbol{\alpha}_1)$ and $-\mathcal{D}_X(\boldsymbol{\alpha}_1 \| \boldsymbol{\alpha}_0)$, respectively. The values of these tangent lines at their extremes thus correspond to the Kullback-Leibler distances. The tangent curves intersect at $t = t^{**}$, the value of which corresponds to the resistor-average distance defined in (4). The Bhattacharyya distance $\mathcal{B}_{p_0}(p_1,)$ equals $-\log \mu\left(\frac{1}{2}\right)$. The $J$-divergence $\mathcal{J}_{p_0}(p_1,)$ equals the average of the two Kullback-Leibler distances, with the geometric mean $\mathcal{G}_{p_0}(p_1,)$ lying somewhere between the $J$-divergence and the smaller of the Kullback-Leibler distances.

$\mathcal{R}_X(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1) \geq \mathcal{C}_X(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1)$. Consequently, $\lim_{N\to\infty} \frac{\log P_e}{N} \geq -\mathcal{R}_X(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1)$. The various distance measures described here have a specific ordering that applies not matter what the component probability distributions may be. $\max\{\mathcal{D}_X(\boldsymbol{\alpha}_0\|\boldsymbol{\alpha}_1), \mathcal{D}_X(\boldsymbol{\alpha}_1\|\boldsymbol{\alpha}_0)\} \geq \mathcal{J}_X(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1) \geq \mathcal{G}_X(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1) \geq \min\{\mathcal{D}_X(\boldsymbol{\alpha}_0\|\boldsymbol{\alpha}_1), \mathcal{D}_X(\boldsymbol{\alpha}_1\|\boldsymbol{\alpha}_0)\} \geq \mathcal{R}_X(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1) \geq \mathcal{C}_X(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1) \geq \mathcal{B}_X(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1)$. In many realistic examples the Chernoff distance roughly equals half the resistor-average distance.[4] Consequently, we have an easily computed quantity that can approximate the more difficult to compute Chernoff distance. The $J$-divergence can differ much more from the Chernoff distance while the more difficult-to-compute Bhattacharyya distance can be quite close. This graphical depiction of these distance measures suggests that as the two Kullback-Leibler distances differ more and more, the greater the discrepancy between the Chernoff distance from the $J$-divergence and the Bhattacharyya distance.]

### 3.2 Quantifying system performance

To analyze how well systems process information, we let $Y$ denote the output of a system that has $X$ as its input as show in figure 1. Because the input is stochastic and encodes information, $Y$ is also a stochastic process that encodes information. Mathematically, the input-output relationship for a system is defined by the conditional probability function $p_{Y|X}(y|x)$, which means that $(X, Y)$

[4]Computer experiments show that this relationship, found in analytic examples, does not apply in general. The curve $-\log \mu(t)$ can lie close to the $t$-axis, leaving the Chernoff and resistor-average distances far apart, and can hug its tangent lines so that the Chernoff and resistor-average distances are numerically close.

form a Markov chain $X \rightarrow Y$: the input signal is all that is needed to calculate the output. Note that this formulation means that the system cannot access the information parameter $\alpha$ other than through the encoded input $X$.

We require the distance measure to be what we call *information theoretic*: it must obey the so-called Data Processing Theorem [7, 10], which loosely states that a system operating on a signal cannot produce an output containing more information than that encoded in its input. The Data Processing Theorem is usually stated in terms of mutual information. Here, we require any viable distance measure to have the property that for any information change, when $X \rightarrow Y$, $d_X(\alpha_0, \alpha_1) \geq d_Y(\alpha_0, \alpha_1)$: output distance cannot exceed input distance. All distances in the Ali-Silvey class [1], as well as many others, have this property by construction. Importantly, the Chernoff and the resistor-average distances have this property.

To evaluate how well systems process information, we explore the quantity $\gamma$, the *information transfer ratio*, defined as the ratio of the distance between the two output distributions and the distance between the corresponding input distributions.

$$\gamma_{X,Y}(\alpha_0, \alpha_1) \equiv \frac{d_Y(\alpha_0, \alpha_1)}{d_X(\alpha_0, \alpha_1)}$$

This ratio is always less than or equal to one for information-theoretic distances. A value of one means that the information expressed by the input is perfectly reproduced in the output; a value of zero means none of the information is represented by the output. Note that achieving a value of one does not require the output signal be of the same kind as the input. Kullback [21] showed that if the information transfer ratio equals one *for all* $\alpha_0$, $\alpha_1$, $Y$ is sufficient statistic for $X$. More generally, the information transfer ratio can equal one only for some choices of $\alpha_1$ for each reference value $\alpha_0$ of the information parameter.[5] In such cases, the information transfer ratio quantifies those aspects of the information expressed perfectly in the system's output and those that aren't. A systematic study of a system's information processing demands that $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_1$ vary systematically, with $\boldsymbol{\alpha}_0$ usually taken as a reference point.

For the special case wherein the information parameter is perturbed ($\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_0 + \boldsymbol{\delta\alpha}$) and the distance measure satisfies the locally Gaussian property (3), we can explicitly write the information transfer ratio.

$$\gamma_{X,Y}(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_0 + \boldsymbol{\delta\alpha}) = \frac{\boldsymbol{\delta\alpha}' \mathbf{F}_Y(\boldsymbol{\alpha}_0) \boldsymbol{\delta\alpha}}{\boldsymbol{\delta\alpha}' \mathbf{F}_X(\boldsymbol{\alpha}_0) \boldsymbol{\delta\alpha}}$$

[5] Consider a statistically independent sequence of Gaussian random variables wherein the mean and variance constitute the components of the parameter vector. Let a system compute the sample average. Changes in the mean will yield an information transfer ratio of one while changes in the variance will not.

Note that the information transfer ratio for perturbational changes does not depend on the choice of distance measure. A subtlety emerges here in the multiparameter case: the value of $\gamma$ depends, in general, on the direction of the perturbation $\boldsymbol{\delta\alpha}$, which means that $\lim_{\boldsymbol{\delta\alpha}\to 0} \gamma_{X,Y}(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_0 + \boldsymbol{\delta\alpha})$ does not exist. This is not surprising since it tells us that the information transfer ratio differs in general for each component parameter $\alpha_i$. For example, there is no reason to believe that a system would preserve the information about amplitude and phase to the same fidelity. Examples show that when $\alpha_1$ and $\alpha_0$ differ substantially, the information transfer ratio's value *does* depend on distance measure choice. For example, symmetric and non-symmetric measures must yield different ratios. Because of its general utility, we use the KL distance here.

The information transfer ratio quantifies the information-processing performance of systems and allows us to think of them as *information filters*. Let $\boldsymbol{\alpha}_0$ define a multiparameter operating point and allow $\boldsymbol{\alpha}_1$ to vary. For some systems, changes might yield a dramatic reduction in $\gamma$ while other changes leave $\gamma$ near its maximal value of one. In this sense, the system, by acting on its input signals, reduces the ability to discern certain information changes relative to other ones. A plot of $\gamma_{X,Y}(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1)$ as a function of $\boldsymbol{\alpha}_1$ reveals the system's information transfer ability. The "passband" occurs when the information transfer ratio is relatively large and the "stopband" when it is much smaller. Note that in general, this ability varies with operating point defined by $\boldsymbol{\alpha}_0$. The maximum value of the information transfer ratio defines the system's information gain and quantifies how well the system's output can be used to extract information relative to the input.

To illustrate the information filter concept, we consider an array-processing example. Figure 3 shows a five-sensor linear array that attempts to determine the waveform of a propagating sinusoidal signal by using classic delay-and-sum beamforming [16]. With the information parameters being the amplitude and propagation angle, the distance between inputs corresponding to the reference and to other amplitudes and angles is given by (see Appendix Appendix B)

$$\mathcal{D}_{\mathbf{X}}(\boldsymbol{\alpha}_1 \| \boldsymbol{\alpha}_0) = \frac{T}{2N_0}\left[M\left(A_0^2 + A_1^2\right) - 2A_0A_1\frac{\sin M\pi f\left(\tau(\theta_1) - \tau(\theta_0)\right)}{\sin \pi f\left(\tau(\theta_1) - \tau(\theta_0)\right)}\right],$$

where $T$ is the observation interval, $N_0/2$ is the spectral height of the white noise, and $A, f, \theta$ are the propagating sinusoid's amplitude, frequency and angle, respectively. $\tau(\theta) = \frac{d}{c}\sin\theta$ is the propagation delay between sensors spaced $d$ apart with $c$ being the propagation speed. $M$ is the number of sensors in the linear array. Distance variations with angle reflect the signal's sinusoidal nature. The analog beamformer, having processing delay $\tau^* = \tau(\theta^*)$, yields an output distance $\mathcal{D}_Y(\boldsymbol{\alpha}_1 \| \boldsymbol{\alpha}_0)$ of

$$\frac{T - (M-1)\tau^*}{2MN_0}\left[A_1\frac{\sin M\pi f\left(\tau(\theta_1) - \tau^*\right)}{\sin \pi f\left(\tau(\theta_1) - \tau^*\right)} - A_0\frac{\sin M\pi f\left(\tau(\theta_0) - \tau^*\right)}{\sin \pi f\left(\tau(\theta_0) - \tau^*\right)}\right]^2$$

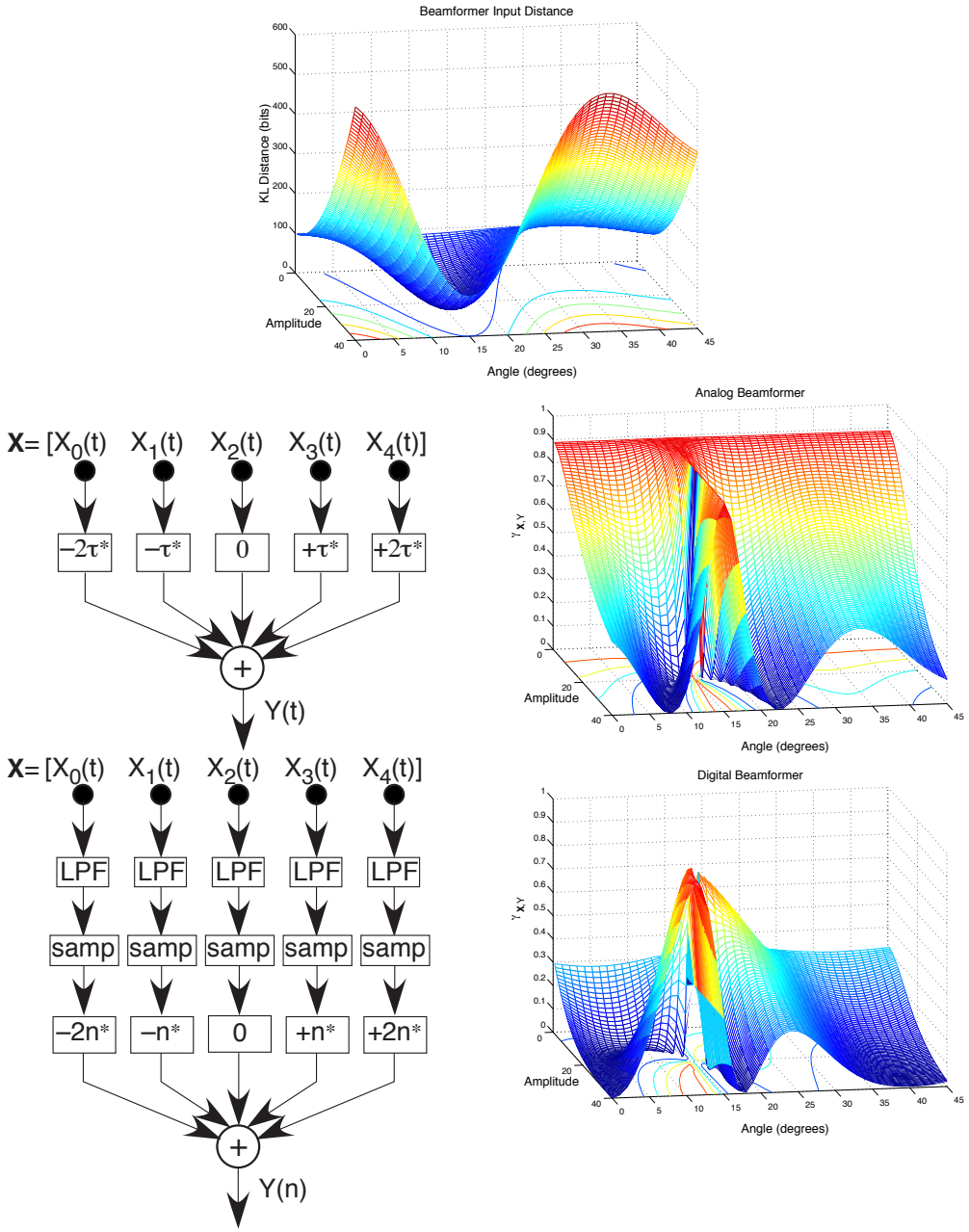


**Figure 3**: A sinusoidal signal propagates toward a five-sensor array. Each sensor measures the propagating field at a particular spatial location in the presence of noise, which we assume to be Gaussian and white both temporally and spatially. The left column shows analog and digital beamformers, each of which delays each sensor's output by an amount designed to cancel the propagation delay corresponding to a particular angle, here equal to $\theta_0$. The digital system's delays are quantized versions of these delays. The delayed signals are then averaged (the summers in the block diagrams represent averagers). For analyzing the information processing capabilities of these beamformers, the input is the vector of sensor outputs over a time interval lasting $T = 100$ s and the output is defined over a shorter time interval of $T - 4\tau^*$. The propagating signal is $A\sin(2\pi ft)$, with $f = 0.1$, and the noise at each sensor had spectral height 500. In the digital case, each output is lowpass-filtered then sampled (bandwidth equal to $0.33$ Hz and sampling interval $\Delta = 1.5$ s). The information vector contains the propagating signal's amplitude and propagation angle: $\boldsymbol{\alpha} = [A, \theta]$. The top plot shows the KL distance between inputs measured in bits (choosing base 2 logarithms in (1)), with the reference being $\boldsymbol{\alpha}_0 = [20, 15°]$. The right column shows the information transfer ratios for the two beamformers.

Figure 3 shows the resulting information transfer ratio. The design behind a beamformer is to produce an estimate of the waveform propagating from a specific direction $\theta^*$. Signals propagating from other directions should be suppressed to some degree. Because the beamformer is linear, it does not affect signal amplitude. The information transfer ratio should reflect these signal processing capabilities, but in addition indicate how the processing affects the ability of optimal processors to determine information represented by the propagating signal. Because we selected amplitude and propagation angle as the information parameters, we can examine how the beamformer reduces the ability of optimal systems to extract these parameters. First of all, the analog beamformer's information transfer ratio maximal value does not achieve the largest possible value of one; its maximal value is $1 - (M-1)\tau^*/T$. This fundamental reduction occurs because each sensor produces a signal over the same time interval that must be delayed relative to the others before averaging. Since averaging only makes sense over a time interval when all delayed signals overlap, the signal-to-noise ratio at the output is smaller than that of the vector of sensor outputs. The maximal information transfer ratio is maintained for two conditions. The ridge extending along the angle $\theta^*$ for amplitudes larger than the reference indicates the intended spatial filtering properties of the array. Signals propagating from other directions are suppressed. For smaller amplitudes than the reference, the information transfer ratio is essentially constant for all propagation angles. This property, which is certainly not part of the beamformer's design, means that the beamformer affects little the ability to detect smaller amplitude signals.

The KL distance at the output of the digital beamformer expresses the loss due to sampling.

$$\frac{\Delta\left(\left\lfloor\frac{T}{\Delta}\right\rfloor-(M-1)n^*\right)}{2MN_0}\left[A_1\frac{\sin M\pi f\left(\tau(\theta_1)-n^*\Delta\right)}{\sin \pi f\left(\tau(\theta_1)-n^*\Delta\right)}-A_0\frac{\sin M\pi f\left(\tau(\theta_0)-n^*\Delta\right)}{\sin \pi f\left(\tau(\theta_0)-n^*\Delta\right)}\right]^2$$

The sampling interval is $\Delta$. The digital beamformer's information transfer ratio is smaller because the sampling restricts the angles to those wherein the sampled delay $n^* = \left\lfloor\frac{\tau^*}{\Delta}\right\rfloor$ is an integer. The sampling interval and reference propagation direction were selected in this example so that the required delay would not be an integer number of samples. The beamformer no longer provides maximal information at an angle corresponding to that of the reference signal, and the information transfer ratio shown in figure 3 reflects this loss.

# 4 A System Theory for Information Processing

Using the information transfer ratio based on the KL distance, we can quantify how various system organizations affect the ability to perform information processing. We emphasize that these results

make few assumptions about the systems —they can be linear or nonlinear —and about the signals they process and produce.

**Cascaded Systems** If two systems are in cascade (figure 1), with $X \rightarrow Y \rightarrow Z$, the overall information transfer ratio is the product of the component ratios.

$$\gamma_{X,Z}(\alpha_0, \alpha_1) = \gamma_{X,Y}(\alpha_0, \alpha_1) \cdot \gamma_{Y,Z}(\alpha_0, \alpha_1) \tag{5}$$

Notice that this result holds for general distance measures (not just the KL distance). Because information transfer ratios cannot exceed one, this result means that once a system reduces $\gamma$ for some information change, *that loss of information representation capability cannot be recovered by subsequent processing*. This result easily generalizes for any number of the systems in cascade. However, insertion of a pre-processing system can increase the information transfer ratio. For example, consider a memoryless system that center-clips: if $|X| < X_0$, $Y = 0$; if $|X| \geq X_0$, $Y = X$. Let the parameter $\alpha$ correspond to the input signal's amplitude. If $\alpha_0 < X_0$, the output is zero and, for some range of amplitude changes, the output remains zero. In this case, $\gamma_{X,Y}(\alpha_0, \alpha_1) = 0$, and no amount of post-processing will change this. On the other hand, if we insert an amplifier of sufficient gain before the center-clipper, $\gamma$ will be non-zero, making the overall information transfer ratio larger than its original value, but smaller than that of the amplifier. Though (5) may bear resemblance to transfer function formulas in linear system theory wherein systems don't "load" each other and system order doesn't matter, these properties don't apply to information processing systems in general. In particular, our theory suggests that pre-processing can improve information processing; post-processing cannot.

**Multiple Inputs** When the input consists of several statistically independent components, the overall information transfer ratio is related to individual transfer ratios by an expression identical to the parallel resistor formula.

$$\frac{1}{\gamma_{\mathbf{X},Y}(\alpha_0, \alpha_1)} = \sum_n \frac{1}{\gamma_{X_n,Y}(\alpha_0, \alpha_1)}$$

This result applies only to those distance measures having the additivity property. Note that each $\gamma_{X_n,Y}$ can exceed one because each $(X_n, Y)$ does not necessarily form a Markov chain. For example, if the input consists of $N$ independent and identically distributed Gaussian variables that encode information in the mean and the system's output equals the sum of the inputs, every $\gamma_{X_n,Y} = N$, while $\gamma_{\mathbf{X},Y} = 1$.[6] The parallel resistor formula implies $\gamma_{\mathbf{X},Y}(\alpha_0, \alpha_1) \leq \min_n \gamma_{X_n,Y}(\alpha_0, \alpha_1)$.

[6]The information transfer ratio equaling one is not surprising because the sum of the observations is the sufficient statistic for Gaussian distributions parameterized by the mean.

While this bound may not be tight (as we have seen, the right side can actually be greater than one), it tells us that the system's overall information transfer ratio is *smaller* than the individual ratios for each input.

**Multiple Outputs** Let a system have one input and $N$ outputs $\mathbf{Y} = [Y_1, \ldots, Y_N]$. When we use the KL distance, the overall information transfer ratio is related to the component ratios as

$$\gamma_{X,[Y_1,\ldots,Y_N]}(\alpha_0, \alpha_1) = \gamma_{X,Y_1}(\alpha_0, \alpha_1) + \sum_{n=2}^{N} \gamma_{X,[Y_n|Y_1,\ldots,Y_{n-1}]}(\alpha_0, \alpha_1) \ . \tag{6}$$

The conditional information transfer ratio in the summation derives from the KL distance's property[7] $\mathcal{D}_{[Y_1,Y_2]}(\alpha_1 \| \alpha_0) = \mathcal{D}_{Y_1}(\alpha_1 \| \alpha_0) + \mathcal{D}_{Y_2|Y_1}(\alpha_1 \| \alpha_0)$, where

$$\mathcal{D}_{Y_2|Y_1}(\alpha_1 \| \alpha_0) = \int p_{Y_1,Y_2}(y_1, y_2; \alpha_1) \log \frac{p_{Y_2|Y_1}(y_2|y_1; \alpha_1)}{p_{Y_2|Y_1}(y_2|y_1; \alpha_0)} \, dy_1 dy_2 \ .$$

Since the indexing of the outputs is arbitrary, result (6) applies to all permutations of the outputs. Also note that this result applies regardless of the systems, the nature of the signals $Y_n$, and how each signal represents the information. Result (6) says that the total information transfer ratio equals the sum of one individual transfer ratio plus the sum of the incremental ratios that quantify how each additional output increases the ability to discriminate the two information states. Since $\gamma \leq 1$, the sum of incremental information transfer ratios must attain some asymptotic value, which means that beyond some point, additional outputs may not significantly increase the information expressed by the aggregate output.

One important special case has several systems in parallel, each of which is operating on the same input. Mathematically, the $N$ outputs are *conditionally* independent of each other given the input: $p_{\mathbf{Y}|X}(\mathbf{y}|x) = \prod_n p_{Y_n|X}(y_n|x)$. We consider two cases. The simplest has each system with its own input that is statistically independent of the other system's input, as in the common model for distributed detection [30]. Because of the additivity property, the input and output KL distances equal the sum of the individual distances. Simple bounding arguments [15] show that the overall information transfer ratio $\gamma(N)$ is bounded below by the smallest information transfer ratio expressed by the individual systems and above by the largest. Consequently, this structure shows no information gain as the number of systems grow (i.e., little variation with $N$). A more interesting case occurs when the systems have a common input, wherein one signal forms the input to all $N$ systems (figure 4). When each system's individual information transfer ratio is greater than zero, we show in Appendix Appendix A that as the number of systems increases, $\gamma(N)$ will *always* increase

[7] This result is a generalization of the additivity property.

strictly monotonically and will have a limiting value of one as long as the systems being added don't have information transfer ratios that decrease too rapidly. This result applies widely: we require no assumption about the nature of the signals involved or about the systems (other than $\gamma > 0$ for each system). *Any* information processing system having the structure shown in figure 4 regardless of what the systems do, what signal serves as the common input, or what kinds of output signals are produced, this result holds. For example, a sufficiently large population of neurons operating independently on their inputs will convey every information-bearing aspect of their inputs regardless of the neural code employed [15] and that space-time coding systems, wherein the coding and the channels are independent of each other, can transmit information asymptotically error-free.

Somewhat surprisingly, the asymptotic rates at which the information transfer ratio increases depend *only* on the nature of the signal $X$ and not on the nature of $Y$ (see Appendix Appendix A).

$$\gamma(N) \overset{N \to \infty}{=} \begin{cases} 1 - k_1 \exp\{-k_2 N\} & X \text{ is discrete-valued} \\ 1 - \dfrac{k}{N} & X \text{ is continuous-valued} \end{cases} \tag{7}$$

As a simple illustrative case, consider parallel systems that share a Gaussian input ($X \sim \mathcal{N}(m, \sigma_X^2)$) with each adding a Gaussian random variable (mean zero and variance $\sigma_n^2$) statistically independent of the input and that added by the other systems. Let the mean $m$ encode the information. Each output $Y_n$ has a Gaussian distribution and individual information transfer ratio calculated with the KL distance equals $\sigma_X^2/(\sigma_X^2 + \sigma_n^2)$. The collective output $\mathbf{Y}$ is a Gaussian random vector having mean $m\mathbf{1}$ and covariance $\text{diag}[\sigma_1^2, \ldots, \sigma_N^2] + \sigma_X^2 \mathbf{1}\mathbf{1}'$, where $\mathbf{1} = \text{col}[1, \ldots, 1]$ and $'$ denotes matrix transpose. The overall information transfer ratio calculated with the KL distance equals

$$\gamma_{X,\mathbf{Y}} = \frac{\sigma_X^2 \sum_n \frac{1}{\sigma_n^2}}{1 + \sigma_X^2 \sum_n \frac{1}{\sigma_n^2}} \ .$$

If the sum $\sum_n \frac{1}{\sigma_n^2}$ converges, the information transfer ratio does not achieve its maximal value of one. Convergence only occurs when $\lim_{n \to \infty} \sigma_n^2 = \infty$; in other words, when the systems become increasingly noisy at a sufficiently rapid rate. When the sum of reciprocal variances diverges (when, for example, $\sigma_n^2$ is constant), the collective output can be used to extract the mean with the same fidelity as the mean can be estimated from the input. When the noise variances are bounded, we have

$$\frac{1}{\frac{\max_n \sigma_n^2}{N\sigma_X^2} + 1} \leq \gamma_{X,\mathbf{Y}} \leq \frac{1}{\frac{\min_n \sigma_n^2}{N\sigma_X^2} + 1}$$

which means that predicted hyperbolic asymptotic behavior indeed occurs.

To show that this asymptotic behavior does not depend on the nature of the output distribution, we considered the case where the input is an exponentially distributed random variable and each

system's output is a Poisson random variable.

$$p_X(x;a) = a e^{-a\,x},\ x \geq 0$$
$$p_{Y_n}(y_n|X) = \frac{(G_n X)^{y_n} e^{-G_n X}}{y_n!}$$

Thus, the input is continuous-valued; despite the fact the output is discrete-valued, our theory predicts that the asymptotics of the information transfer ratio will resemble that of the Gaussian example examined above. Here, $G_n$ is the gain between the input and the $n^{\text{th}}$ output. The unconditional output probability function for each system is geometric:

$$p_{Y_n}(y_n;a) = \frac{a}{a+G_n}\left(\frac{G_n}{a+G_n}\right)^{y_n},\ y_n = 0,1,2,\ldots$$

For the aggregate, the input and output KL distances are

$$\mathcal{D}_X\left(a_1\|a_0\right) = \log\frac{a_1}{a_0} + \frac{a_0-a_1}{a_1}$$
$$\mathcal{D}_{\mathbf{Y}^{(N)}}\left(a_1\|a_0\right) = \log\frac{a_1}{a_0} + \left(1+\frac{1}{a_1}\sum_n G_n\right)\log\left(\frac{a_0+\sum_n G_n}{a_1+\sum_n G_n}\right)$$

For the information transfer ratio to achieve a value of one, the sum of the gains must diverge. In this case, plotting the information transfer ratio when $G_n = G$ reveals a hyperbolic asymptotic behavior.[8]

Changing the input in this Poisson example to a Bernoulli random variable results in exponential asymptotics. Here, $p_X(x) = p$ when $x = 0$ and equals $1-p$ when $x = 1$. The input and output KL distances are

$$\mathcal{D}_X\left(p_1\|p_0\right) = p_1\log\frac{p_1}{p_0} + (1-p_1)\log\frac{1-p_1}{1-p_0}$$
$$\mathcal{D}_{\mathbf{Y}^{(N)}}\left(p_1\|p_0\right) = \left(p_1+(1-p_1)e^{-\sum_n G_n}\right)\log\frac{p_1+(1-p_1)e^{-\sum_n G_n}}{p_0+(1-p_0)e^{-\sum_n G_n}} + (1-p_1)(1-e^{-\sum_n G_n})\log\frac{1-p_1}{1-p_0}$$

Thus, if the sum of gains diverge, the information transfer ratio has an asymptotic value of one and approaches it exponentially. Consequently, a change in the nature of the *input* changes the asymptotics of the information transfer ratio.

[8] While the hyperbolic asymptotic formula is valid, we found that the formula $\gamma(N) = \left(1+\frac{k}{N}\right)^{-1}$ more accurately characterizes how $\gamma$ varies with $N$ in several examples. This formula is exact in the Gaussian example and closely approximates the information transfer ratio in the Poisson example.

From a general perspective, a parallel structure of information processing by noisy systems can overcome component system noisiness and achieve a nearly perfectly effective information representation (an information transfer ratio close to one) with relatively few systems. The information transfer ratio for each system determines how many systems are needed: the smaller $\gamma_{X,Y_n}$ is, proportionally more systems are needed to compensate.

## 5 Summary and Conclusions

Our theory of information processing rests on the approach of using information change to probe how well a signal —*any* signal —represents information and how a system "filters" the information encoded in its input. In the theory's current form, signals must be completely described by a probability function. Continuous-time and discrete-time signals are certainly in this class. Signals could be symbolic-valued, and multichannel signals having a mixed character are embraced by our theory. Thus, we can within one framework, describe a wide variety of stochastic signals. The main idea is to use information theoretic distances, especially the Kullback-Leibler distance, to quantify how the signal's probability function changes as a consequence of an informational change. The choice of the KL distance follows from the desire for the distance to express information processing performance. In the current theory, information processing means either classification (distinguishing between informational states) or estimation (estimating the information). If one informational state change yields a distance of 2 bits and another a distance of 3 bits, the detector for the second change has error probabilities roughly a factor of 2 smaller than the one for the first. For perturbational changes of real-valued information parameters, the KL distance is proportional to the Fisher information matrix, thus quantifying how well the information parameter can be estimated in the mean-square sense. We used the information transfer ratio to quantify how the information filtering properties of analog and discrete-time beamformers differ (figure 3) and how effective a neural signal generator is [25]. Thus, our approach can cope with a variety of signals and systems, and analyze them in unified way.

Conventional wisdom suggests that static measures such as mutual information and entropy could be used to assess a signal's information content. Unfortunately, entropy does *not* quantify the information contained in a signal for two reasons. First of all, it considers a signal as an indivisible quantity, not reflecting what is information bearing and what is not, and not reflecting what information expressed by the signal is relevant. Paralleling our approach, one could consider measuring entropy changes when information changes. However, entropy does not obey the Data Processing Theorem. For example, increasing (decreasing) the amplitude of a signal increases (decreases) its

entropy, which means the output entropy is greater (less) than the input entropy. Secondly, entropy cannot be defined for all signals. The entropy of an analog signal can either be considered to be infinite or quantified by differential entropy [5]. Differential entropy has the problem that because it is not scale-invariant, it depends on the signal's amplitude units in a nonsensical way. For example, the differential entropy of a Gaussian random variable is $\frac{1}{2}\log 2\pi e\sigma^2$. Depending on whether the standard deviation is expressed in volts or millivolts, the entropy changes! The KL distance remains constant under all transformations that result in sufficient statistics, scale being only one such transformation. Differential entropy can be related to Fisher information (de Bruijn's identity; see [10, Theorem 16.6.2]), but this relationship is not particularly useful.

An alternative choice for the quantity upon which to base a theory of information processing is mutual information. Mutual information is a statistical similarity measure, quantifying how closely the probability functions of a system's input and output agree. Mutual information plays a central role in rate-distortion theory, where it quantifies how well the output expresses the input. The information bottleneck method uses this property to assess how well signals encode information when faced with an ultimate signal compression [28]. This approach not only requires specifying what is relevant information, but also the information must have a stochastic model. While in many problems information could be considered parametric (this paper uses many such examples), requiring a stochastic model is overly restrictive. We attempted to develop a parallel theory that used mutual information using notions described by Popoli and Mendel [24]. Not only is the relationship between mutual information and classification/estimation tenuous and indirect, but no simple structural theory could be obtained. Furthermore, using mutual information would require *all* empirical studies to have access to each system's input and output because the joint probability function between input and output is required. From our experience in neuroscience, this requirement can rarely be met in practice. Usually both can't be measured simultaneously and what comprises input and output is only broadly known. Furthermore, the well-known curse of dimensionality becomes apparent. If $O(N)$ data values are required to estimate either of the marginal probability functions to some accuracy, estimating the joint distribution requires $O\left(N^2\right)$. Estimating the KL distance only requires estimates of a *single* signal's probability function under two conditions, which is $O(N)$.

Ultimately, we believe Weaver was right [27]: information is about meaning and how processing information leads to effective action. On more technical grounds, using rate-distortion theory or the information bottleneck requires specification of a distortion function that measures the error between the encoded and decoded information. In empirical work wherein you gather data to study a part of an information processing system, usually the distortion function is both unknown and

irrelevant. Unknown because intermediate signals are just that and only encode the information: when the decoded information is not evident, distortion can't be measured. Irrelevant because we don't know how to quantify the distortion between intended and derived meaning. Information for us ultimately concerns actions derived from signal interpretation. If action is the measurable output, it certainly lies in a different space than information or information-bearing signals. In this case, the distortion function is again unknown. Our use of the KL distance encompasses both classification-related and mean-squared error distortion measures. In cases where information is parametric, the mean-squared error viewpoint is arbitrary and restrictive. However, when the "true" distortion measure is not known, mean-squared error at least represents a well-understood choice.

Because we chose information-theoretic distances, we can quantify how systems affect information. In using the ratio of distances, we quantify how well optimal information extraction systems could perform when provided the output signal relative to the performance when provided the input signal. Note that the input and output signals need not be defined over the same space. The KL distance "probe" can assess optimal information processing performance regardless of the nature of the signals involved. We have found this generality to be quite useful. In one study of neural systems [25], the input is a noisy analog signal and the output is a neural spike train (modeled as a point process). In this case, the input and output were measured simultaneously and separated using wavelet denoising techniques. We separately measured the KL distances of these very different signals and computed the information transfer ratios. Interestingly, we measured very small information transfer ratios ($10^{-4}$–$10^{-3}$). Analysis showed that an ideal system performing a similar information transformation could not yield a larger information transfer ratio because of the conversion between signal types and certain physical constraints (signal-to-noise ratio at the input and maximal event rate at the output).

Our analysis of various information processing system architectures indicates that structure—how systems are interconnected—affects information processing capabilities *regardless of the kinds of signals* and, in many cases, regardless of the specific information-theoretic distance chosen. The simple cascade result shows that the ability to extract information can be easily degraded as a succession of systems process it and the representation changes from one form to another. Once a loss occurs, such as in our neural system, it can never be regained if only the output is used in subsequent processing. This conclusion follows from the Data Processing Theorem and consequently is not that surprising. However, the other structures we have been able to analyze do represent new insights. Of particular interest is the parallel system result: when a sufficiently large number of systems operate separately on a common input, their aggregated output can express

whatever information the input expresses with little loss. Inhomogeneity among the systems and their outputs can occur without affecting this result. If instead these same systems view statistically independent inputs that encode the same information, the information transfer ratio cannot exceed the largest information transfer ratio provided by the component systems [15]. Consequently, this structure is ineffective in than it provides no more information processing capability than its best component system.

From a broader perspective, our information processing theory goes beyond Shannon's classic results that apply to communicating information. The entropy limit on source coding defines how to efficiently represent discrete-symbol streams. Channel capacity defines how to reliably communicate signals over noisy channels, be they discrete or continuous-valued. Rate-distortion theory assesses how compression affects signal fidelity. As useful as these entities are, they do not help analyze existing systems to assess how well they work. More penetrating theories must be concerned with what signals represent, how well they do so, and how classes of systems more general than channels affect that representation. Our theory tries to address Weaver's vision [27] of a broader theory that concerns information content. By requiring the information to be changed, we effectively probe the signal's semantic content. By using information-theoretic measures that reflect optimal-processing performance, we can quantify the effectiveness of *any* signal's information-bearing capability.

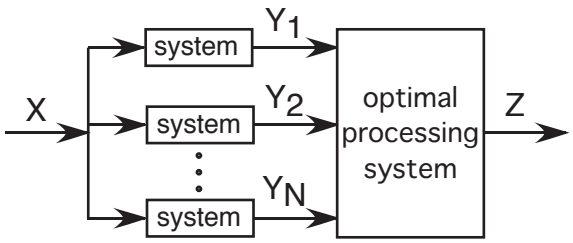


**Figure 4**: The initial portion leading to the output vector $\mathbf{Y}$ shows what we call a noncooperative structure: systems transform a common input $X$ to produce outputs $Y_n$ that are conditionally independent and identically distributed. To find the asymptotic behavior of the information transfer ratio, we append an optimal processing system. In the discrete case, the optimal processor is the likelihood ratio detector that indicates which value of $X$ occurred. In the continuous case, the optimal processor is the maximum likelihood estimator of $X$.

## Appendix A Multi-output and parallel systems

First of all, we prove that the information transfer ratio monotonically increases as the number of systems increases regardless of the systems and the input encoding for the separate system case shown in figure 4: $\gamma(N) \overset{N\to\infty}{\to} 1$. This result rests on the *log-sum inequality* [10]:

$$\int p_1(x) \log \frac{p_1(x)}{p_0(x)}\, dx \geq \int p_1(x)\, dx \log \frac{\int p_1(x)\, dx}{\int p_0(x)\, dx}$$

with $p_0(x)$, $p_1(x)$ being probability functions. The integral can be a sum in the case of discrete probability functions. Equality occurs only when $p_0(x) = p_1(x)$. To apply this result, note that the probability function for each system's output is given by

$$p_{Y_n}(y_n;\alpha) = \int p_{Y_n|X}(y_n|x) p_X(x;\alpha)\, dx\ .$$

Note that $p_{Y_n|X}(y_n|x)$, which defines each system's input-output relation, does *not* depend on $\alpha$. The KL distance $\mathcal{D}_{\mathbf{Y}^{(N)}}(\alpha_1\|\alpha_0)$ between the outputs responding to the two information conditions $\alpha_0$ and $\alpha_1$ equals

$$\int \left[ \int \prod_{n=1}^{N} p_{Y_n|X}(y_n|x) p_X(x;\alpha_1)\, dx \right] \times \log \frac{\int \prod_n p_{Y_n|X}(y_n|x) p_X(x;\alpha_1)\, dx}{\int \prod_n p_{Y_n|X}(y_n|x) p_X(x;\alpha_0)\, dx}\, d\mathbf{y}$$

By applying the log-sum inequality with respect to the input, we upper-bound this distance and demonstrate the Data Processing Theorem: $\mathcal{D}_{\mathbf{Y}^{(N)}}(\alpha_1\|\alpha_0) \leq \mathcal{D}_X(\alpha_1\|\alpha_0)$. Applying the log-sum inequality to the integral over $y_N$, we find that $\mathcal{D}_{\mathbf{Y}^{(N)}}(\alpha_1\|\alpha_0) > \mathcal{D}_{\mathbf{Y}^{(N-1)}}(\alpha_1\|\alpha_0)$, with strict inequality arising because the individual system KL distances are not zero. Thus, as the population size increases, the KL distance strictly increases. In what follows, we find a lower bound on the rates of increase that indicates that $\gamma(N)$ approaches one asymptotically under mild conditions.

To determine the rate at which the information transfer ratio approaches one, we consider an optimal processing system that collects the outputs $\mathbf{Y}$ to yield an estimate $Z$ of the input $X$ (figure 4). We then calculate the asymptotic distribution of the estimate, derive the distance between the estimates, and find the information transfer ratio between the input and the estimate. Because of the cascade property (5), this ratio forms a lower bound on the information transfer ratio between the input and the parallel system's collective output.

### Discrete-valued inputs

Let $X$ be drawn from a set and have a discrete probability distribution. We are interested in the asymptotic (in $N$, the number of parallel systems) behavior of the information transfer ratio $\gamma_{X,Z}(\alpha_0, \alpha_1) = d_Z(\alpha_0, \alpha_1)/d_X(\alpha_0, \alpha_1)$. Because the distance between the inputs does not vary with $N$, we need only consider the numerator distance. Because the input is discrete, the "optimal processing system" of figure 4 is a likelihood ratio classifier that uses $\mathbf{Y}^{(N)} = [Y_1, \ldots, Y_N]$ to determine which value of $X$ occurred. Let $M = |X|$ and $Z$ be the output decision. The probabilistic relation between the input set and the decision set can be expressed by an $M$-ary crossover diagram. Since we will consider asymptotics in $N$, here equal to the number of systems, we know that the error probabilities in this crossover diagram do *not* depend on the *a priori* symbol probabilities $p_X(x;\alpha)$ so long as the symbol probabilities are non-zero. Let $\epsilon_m^j = \Pr[Z = z_m | X = x_j]$ denote the crossover probabilities. Then, the output symbol probabilities are

$$p_Z(z_m;\alpha) = p_X(x_m;\alpha)\left(1 - \sum_{j\neq m} \epsilon_j^m\right) + \sum_{k\neq m} p_X(x_k;\alpha)\epsilon_m^k$$

Note that $\epsilon_m^m \to 1$ as $N \to \infty$. This expression for $p_Z(z_m;\alpha)$ is written in terms of the crossover probabilities $\epsilon_i^j$, $i \neq j$ that all tend to 0 with increasing $N$. The crossover probabilities do depend on the distribution of the system output but don't vary with the information parameter. We can collect these crossover probabilities into the set $\boldsymbol{\epsilon}$ and explicitly indicate that the distance between classifier outputs depends on $\boldsymbol{\epsilon}$ as $d_Z(\alpha_0, \alpha_1; \boldsymbol{\epsilon})$. As long as the distance is differentiable with respect to the probabilities $p_Z(z_m;\alpha)$, the distance can be represented by a Taylor series approximation about the origin.

$$d_Z(\alpha_0, \alpha_1; \boldsymbol{\epsilon}) = d_Z(\alpha_0, \alpha_1; \mathbf{0}) + \boldsymbol{\epsilon}' \nabla_{\boldsymbol{\epsilon}} d_Z(\alpha_0, \alpha_1; \boldsymbol{\epsilon})\big|_{\boldsymbol{\epsilon}=0} + O(\epsilon_{\max}^2)$$

The first term equals the distance $d_X(\alpha_0, \alpha_1)$ since having zero crossover probabilities corresponds to a classifier that makes no errors. When the distance is information theoretic, the remaining terms must total a negative quantity because of the Data Processing Theorem. The maximum of all the

crossover probabilities, $\epsilon_{\max}$, asymptotically equals [22]

$$\epsilon_{\max} = f(N) \exp\left\{ -\min_{i \neq j} \mathcal{C}\left( p(\mathbf{Y}^{(N)}|X_i), p(\mathbf{Y}^{(N)}|X_j) \right) \right\}$$

with $f(\cdot)$ a slowly varying function in the sense that $\lim_{N\to\infty} [\ln f(N)]/N = 0$, and $\mathcal{C}(\cdot,\cdot)$ the Chernoff distance. Because $\epsilon_{\max}$ dominates the performance of the optimal classifier as the number of systems increases, the distance $d_Z(\alpha_0, \alpha_1; \boldsymbol{\epsilon})$ decreases linearly as the crossover probabilities decrease, with the largest of these dominating the decrease. Consequently,

$$\gamma_{X,\mathbf{Y}^{(N)}}(\alpha_0, \alpha_1) \geq 1 - \frac{K f(N)}{d_X(\alpha_0, \alpha_1)} \exp\left\{ -\min_{i \neq j} \mathcal{C}\left( p(\mathbf{Y}^{(N)}|X_i), p(\mathbf{Y}^{(N)}|X_j) \right) \right\} . \quad (8)$$

Here, $K$ is the negative gradient of the distance with respect to $\epsilon_{\max}$. Whenever the Chernoff distance diverges as $N \to \infty$, the information transfer ratio approaches one asymptotically. Typically, the Chernoff distance for each system's output $Y_n$ is bounded from above and below by constants: $\mathcal{C}_{\max} \geq \mathcal{C}\left(p(Y_n|X_i), p(Y_n|X_j)\right) \geq \mathcal{C}_{\min} > 0$. This special case means that new systems added in parallel do not have a systematically larger or smaller discrimination abilities than the others. In this case, the Chernoff distance term in the exponent of (8) increases linearly in $N$. We thus conclude that for the case of discrete input distribution with finite support, the asymptotic increase in the information transfer ratio (as we increase number of parallel outputs) is exponential (or greater) and that the information transfer ratio reaches 1 as $N \to \infty$ so long as the systems being added don't have too rapid a systematic diminished ability to discriminate which input value occurred. If the regularity condition (differentiability with respect to the probabilities of signal values) is satisfied, our result applies to *any* distance measure used to define the information transfer ratio.

### Continuous-valued inputs

To determine the rate of increase of the information transfer ratio when $X$ has a probability function defined over an interval or over a Cartesian product of intervals, we use the approach shown in figure 4 with $Z$ being the maximum likelihood estimator (MLE) of $X$. Under certain regularity conditions [11] and because the $Y_n$'s are conditionally independent, the MLE is asymptotically (in the number of systems $N$) Gaussian with a mean equal to $X$ and variance equal to the reciprocal of the Fisher information $F_{\mathbf{Y}|X}$. This result applies when the systems aren't homogeneous so long as the third absolute moment of the score function of $Y_n$ is bounded and the Fisher information $F_{\mathbf{Y}|X}$ diverges as $N \to \infty$. Because the quantity $\mathbf{Y}$ has conditionally independent components, $F_{\mathbf{Y}|X} = \sum_{n=1}^{N} F_{Y_n|X}$. Thus, if this sum diverges, we can obtain the probability density of $Z$.

$$p_Z(z;\alpha) = \int p_X(x;\alpha) \left( \frac{\sum_n F_{Y_n|X}}{2\pi} \right)^{1/2} \times \exp\left( -\frac{(z-x)^2 \sum_n F_{Y_n|X}}{2} \right) dx$$

If the third derivative of the input probability density function, $p_X(\cdot;\alpha)$, is bounded, we can expand $p_X(\cdot;\alpha)$ in a Taylor series around $z$, up to the third-order term and then perform term-by-term integration. This procedure amounts to the *Laplace approximation* for an integral. The probability density of $Z$ can be then expressed as

$$p_Z(z;\alpha) = p_X(z;\alpha) + \frac{H(z;\alpha)}{2\sum_{n=1}^{N} F_{Y_n|X}} + O\left(\frac{1}{\left(\sum_{n=1}^{N} F_{Y_n|X}\right)^{3/2}}\right) , \tag{9}$$

where $H(z;\alpha)$ is the second derivative of $p_X(\cdot;\alpha)$ evaluated at $z$: $H(z;\alpha) = \left.\frac{\partial^2 p_X(x;\alpha)}{\partial x^2}\right|_{x=z}$. Result (9) says that the distance between MLEs corresponding to the two information states equals the distance between the input probability densities perturbed in a particular way. If $F_{\max} \geq F_{Y_n|X} \geq F_{\min}$, then the sum of Fisher information terms increases linearly in $N$. In this case, a Taylor-like series can be written for the distance between perturbed input probability functions using the Gateaux derivative, a generalized directional derivative.

$$d_Z(\alpha_0,\alpha_1) = d_X(\alpha_0,\alpha_1) - \frac{K}{N} + O\left(\frac{1}{N^{3/2}}\right)$$

Here $K$ summarizes the negative first Gateaux derivative of the distance in the perturbational direction. Consequently, for all information-theoretic distances for which the first-order Gateaux derivatives exist (all Ali-Silvey distances are in this class), the information transfer ratio between the input and the output $\mathbf{Y}$ is lower-bounded by the quantity above divided by $d_X(\alpha_0,\alpha_1)$.

$$\gamma_{X,\mathbf{Y}^{(N)}}(\alpha_0,\alpha_1) \geq 1 - \frac{K}{N d_X(\alpha_0,\alpha_1)} + O\left(\frac{1}{N^{3/2}}\right) .$$

This result states that the asymptotic information transfer ratio approaches one at least as fast as a hyperbola. The Gaussian example described previously had this asymptotic behavior, which means this hyperbolic behavior is tight.

# Appendix B Kullback-Leibler distance between two Gaussian random processes

Consider two Gaussian random processes, $X^{(0)}(t)$ and $X^{(1)}(t)$ with two different mean functions, $m_0(t)$ and $m_1(t)$ and the same covariance structure $K(t,u)$. To calculate the KL distance between $X^{(0)}(t)$ and $X^{(1)}(t)$ over some time interval, we used the Karhunen-Loève expansion [3, §1.4] to represent processes in terms of *uncorrelated* random variables, $X_i^{(j)}$, $j = 0, 1$. Since these variables are Gaussian, they are also statistically independent, which means that the KL distance between two

processes can be written as a sum of KL distances between the their representations $\{X_i^{(1)}\}$ and $\{X_i^{(0)}\}$:

$$\mathcal{D}\left(X^{(1)}(t)\|X^{(0)}(t)\right) = \sum_i \mathcal{D}\left(X_i^{(1)}\|X_i^{(0)}\right)$$

By using the result for the KL distance between two Gaussian densities with different means and the same variance [18], the right-hand side equals

$$\frac{1}{2}\sum_i \frac{1}{\lambda_i}\left(m_i^{(1)} - m_i^{(0)}\right)^2$$

Here, $\{\lambda_i\}$ are eigenvalues of $K(\cdot,\cdot)$ and $\{m_i^{(0)}\}, \{m_i^{(1)}\}$ are the coefficients of the series expansion of the means using the Karhunen-Loève basis. Defining $Q(t,u)$ to be the so-called inverse kernel of the covariance function [29] that satisfies $\int_0^T K(t,u)Q(t,v)\,dt = \delta(u-v)$, it has the same eigenfunctions as $K(\cdot,\cdot)$ with eigenvalues equal to $1/\lambda_i$. The KL distance between the processes has the following expression.

$$\begin{aligned}\mathcal{D}\left(X^{(1)}(t)\|X^{(0)}(t)\right) &= \frac{1}{2}\|m^{(1)}(t) - m^{(0)}(t)\|_Q^2 \\ &= \int_0^T\int_0^T \left(m^{(1)}(t) - m^{(0)}(t)\right)Q(t,u)\left(m^{(1)}(u) - m^{(0)}(u)\right)dt\,du\end{aligned}$$

When the process is white with spectral height $\frac{N_0}{2}$, $K(t,u) = \frac{N_0}{2}\delta(t-u)$ and $Q(t,u) = \frac{2}{N_0}\delta(t-u)$, in which case, the Kullback-Leibler distance between two white-noise processes differing in their means is $\|m^{(1)}(t) - m^{(0)}(t)\|^2/N_0$.

## References

[1] S.M. Ali and S.D. Silvey. A general class of coefficients of divergence of one distribution from another. *J. Roy. Stat. Soc.*, 28:131–142, 1966.

[2] S. Amari and H. Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, Rhode Island, 2000.

[3] R.B. Ash and M.F. Gardner. *Topics in Stochastic Processes*. Academic Press, New York, NY, 1975.

[4] M. Basseville. Distance measures for signal processing and pattern recognition. *Signal Processing*, 18:349–369, 1989.

[5] T. Berger. *Rate Distortion Theory*. Prentice-Hall, Englewood Cliffs, NJ, 1971.

[6] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35:99–109, 1943.

[7] R.E. Blahut. *Principles and Practice of Information Theory*. Addison-Wesley, Reading, MA, 1987.

[8] H. Chernoff. Measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Stat.*, 23:493–507, 1952.

[9] H. Chernoff. Large-sample theory: Parametric case. *Ann. Math. Stat.*, 27:1–22, 1956.

[10] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., 1991.

[11] H. Cramer. *Mathematical Methods of Statistics*. Princeton University Press, Princeton, NJ, 1946.

[12] A.G. Dabak. *A Geometry for Detection Theory*. PhD thesis, Dept. Electrical & Computer Engineering, Rice University, Houston, TX, 1992.

[13] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. A*, 186:453–461, 1946.

[14] D.H. Johnson. Toward a theory of signal processing. In *Info. Th. Worskhop on Detection, Estimation, Classification and Imaging*, Santa Fe, NM, 1999.

[15] D.H. Johnson. Neural population structures and consequences for neural coding. *J. Computational Neuroscience*, 2002. Submitted.

[16] D.H. Johnson and D.E. Dudgeon. *Array Signal Processing: Concepts and Techniques*. Prentice-Hall, 1993.

[17] D.H. Johnson, C.M. Gruner, K. Baggerly, and C. Seshagiri. Information-theoretic analysis of neural coding. *J. Computational Neuroscience*, 10:47–69, 2001.

[18] D.H. Johnson and G.C. Orsak. Relation of signal set choice to the performance of optimal non-Gaussian detectors. *IEEE Trans. Comm.*, 41:1319–1328, 1993.

[19] T. Kailath. The divergence and Bhattacharyya distance measures in signal selection. *IEEE Trans. on Comm. Tech.*, COM-15(1):52–60, 1967.

[20] A.N. Kolmogorov. On the Shannon theory of information transmission in the case of continuous signals. *IRE Trans. Info. Th.*, 3:102–108, 1956.

[21] S. Kullback. *Information Theory and Statistics*. Wiley, New York, 1959.

[22] C.C. Leang and D.H. Johnson. On the asymptotics of $M$-hypothesis Bayesian detection. *IEEE Trans. Info. Th.*, 43:280–282, 1997.

[23] C.S. Miller, D.H. Johnson, J.P. Schroeter, L. Myint, and R.M. Glantz. Visual responses of crayfish ocular motoneurons: An information theoretical analysis. *J. Computational Neuroscience*, 15:247–269, 2003.

[24] R.F. Popoli and J.M. Mendel. Relative sufficiency. *IEEE Trans. Automatic Control*, 38:826–828, 1993.

[25] C.J. Rozell, D.H. Johnson, and R.M. Glantz. Information processing during transient responses in the crayfish visual system. *Neurocomputing*, 52-54:53–58, 2003.

[26] C.E. Shannon. A mathematical theory of communication. *Bell System Tech. J.*, pages 379–423, 623–656, 1948.

[27] C.E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. U. Illinois Press, Urbana, IL, 1949.

[28] N. Tishby, F.C. Pereira, and W. Bialek. The information bottleneck method. In *Proc. $37^{th}$ Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.

[29] H.L. van Trees. *Detection, Estimation, and Modulation Theory, Part I*. Wiley, New York, 1968.

[30] P.K. Varshney. *Distributed Detection and Data Fusion*. Springer-Verlag, New York, NY, 1997.

[31] N.N. Čencov. *Statistical Decision Rules and Optimal Inference*, volume 14 of *Translations in Mathematics*. American Mathematical Society, Providence, RI, 1982.

## Figure Captions

**Figure 1** An information source produces information considered relevant, an abstract quantity represented by the symbol $\alpha$. This information, as well as extraneous information represented by $\chi_0$ and $\chi_1$ is encoded (modulated) onto the signal $X$, which we assume to be stochastic. A system, having its input-output relationship defined by the conditional probability function $p_{Y|X}(Y|X)$, serves as an information filter, changing the fidelity with which the information is represented by its output, possibly accentuating some aspects of the information while suppressing others. The information sink responds to the information encoded in $Y$ by exhibiting an action $Z$.

**Figure 3** A sinusoidal signal propagates toward a five-sensor array. Each sensor measures the propagating field at a particular spatial location in the presence of noise, which we assume to be Gaussian and white both temporally and spatially. The left column shows analog and digital beamformers, each of which delays each sensor's output by an amount designed to cancel the propagation delay corresponding to a particular angle, here equal to $\theta_0$. The digital system's delays are quantized versions of these delays. The delayed signals are then averaged (the summers in the block diagrams represent averagers). For analyzing the information processing capabilities of these beamformers, the input is the vector of sensor outputs over a time interval lasting $T = 100$ s and the output is defined over a shorter time interval of $T - 4\tau^*$. The propagating signal is $A\sin(2\pi ft)$, with $f = 0.1$, and the noise at each sensor had spectral height 500. In the digital case, each output is lowpass-filtered then sampled (bandwidth equal to $0.33$ Hz and sampling interval $\Delta = 1.5$ s). The information vector contains the propagating signal's amplitude and propagation angle: $\boldsymbol{\alpha} = [A, \theta]$. The top plot shows the KL distance between inputs measured in bits (choosing base 2 logarithms in (1)), with the reference being $\boldsymbol{\alpha}_0 = [20, 15°]$. The right column shows the information transfer ratios for the two beamformers.

**Figure 4** The initial portion leading to the output vector $\mathbf{Y}$ shows what we call a noncooperative structure: systems transform a common input $X$ to produce outputs $Y_n$ that are conditionally independent and identically distributed. To find the asymptotic behavior of the information transfer ratio, we append an optimal processing system. In the discrete case, the optimal processor is the likelihood ratio detector that indicates which value of $X$ occurred. In the continuous case, the optimal processor is the maximum likelihood estimator of $X$.