

A FRAMEWORK OF A THEORY OF INFORMATION PROCESSING

Don H. Johnson and Sinan Sinanović

Computer and Information Technology Institute
 Department of Electrical and Computer Engineering
 Rice University
 Houston, Texas 77251–1892
 dhj@rice.edu, sinan@rice.edu

ABSTRACT

Information processing is performed when a system preserves aspects of the input related to the information the input represents while it removes other aspects. To describe a system’s information processing capability, input and output need to be compared in a way invariant to the what the signals are and the way they represent information. This comparison should be general and include scalar and vector signals, numeric or symbolic signals, and mixtures of these. We describe an approach to quantifying information processing based on applying controlled changes to the information, assessing how much the information-bearing signal changed, and if the signal serves as the input to a system, using the same assessment on the output. We calculate information-theoretic distances on information-bearing signals between these two conditions. We select the kind of distance according to its ability to characterize how well optimal signal processing systems can extract information from the signal. By computing the ratio of the output and input distances, we evaluate the system’s information processing capabilities. Properties of this ratio are used to derive fundamental information processing properties of systems and interconnected systems.

1. INTRODUCTION

In 1949, Warren Weaver’s introduction to Shannon’s information theory [10], stated that while Shannon’s 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 casted Shannon’s work as the engineering or technical side of the problem because it did not deal with the extraction of meaning. Less well understood fifty years ago is the concentration of

Shannon’s theory on digital channels: His work applied to channels used to send information that arose from digital sources. Many a researcher dealing with existing communications systems that are not digital (sensorineural systems, for example) has calculated a channel capacity when the notion is at best remote and more than likely useless in quantifying how the information is affected. Consequently, some issues remain on the technical level. ‘Semantic’ problems concern being able to determine what was the information transmitted by the communication system. Here, it is not enough to worry about source compression or error correcting codes; the primary semantic concern is determining the extent to which the receiver understands the communication. At the ‘influential’ level, we want to determine how nearly the best actions we taken based on analyzing the received information. In this paper, we frame here a theory of information processing that spans all three levels and complements classic information theory.

Signals represent information. When systems act on their input signal(s) and produce an output, they perform information processing, enhancing certain aspects of their input(s) and suppressing others. Systems that re-represent the input signal without loss, such as an ideal amplifier or the Fourier transform, preserve information completely and thus do not perform any information processing. Many systems, like non-ideal filters that attempt remove out-of-band noise, affect the information bearing component of their input(s) while attempting to produce a “cleaner” output. True information processing systems extract selected aspects of their input(s) and re-represent components that remain in output signal(s). To develop a measure that would characterize a system’s information processing capability, we need to compare input(s) and output(s) somehow. In linear systems, one uses the transfer function or cross-correlation. However, in quantifying the processing or more complex systems, non-linearity and

non-Gaussian effects cause classical methods to fail at capturing all a system does.

In earlier work [4, 11], we first described our approach. We conceptually (or in reality) induce controlled changes in the information represented by a system's input and calculate distances between inputs and outputs individually. Controlled change is required because information content cannot be judged solely by a signal's appearance or properties. Take the case of a multiuser channel: one user's signal is another's interference. By considering information-related changes in the signal, we are essentially specifying what signal components or aspects convey information. By measuring how different the two signals are, we can quantify how well the information is represented. By comparing input and output differences, we can quantify how well a system processes relevant information.

2. QUANTIFYING INFORMATION PROCESSING

We represent information by the quantity θ . Information in simple cases could be a collection of parameters, and θ would be a parameter vector. More generally, we do not need to restrict θ to be a collection; it could just symbolically represent information. Let \mathbf{X} represent a system's input and \mathbf{Y} its output. The independent variable is suppressed and these signals could be vectors of real or symbolic values. According to the data processing theorem [2], if $\theta \rightarrow \mathbf{X} \rightarrow \mathbf{Y}$ (if θ , \mathbf{X} , and \mathbf{Y} form Markov chain), then $I(\theta; \mathbf{X}) \geq I(\theta; \mathbf{Y})$, where $I(\cdot; \cdot)$ represents mutual information. The Markov chain assumption means that θ encapsulates *all* the information represented by the signal vector \mathbf{X} and that this vector represents *all* information-bearing inputs to some system that has \mathbf{Y} as its output signal vector. To recast the data processing theorem into a more relevant form, let $\mathbf{X}(\theta_0)$, $\mathbf{X}(\theta_1)$ represent input signals having different information content with $\mathbf{Y}(\theta_0)$, $\mathbf{Y}(\theta_1)$ representing the corresponding outputs. Many distance measures $d(\cdot, \cdot)$, which satisfy a data processing theorem in the following sense, are said to be *information-theoretic*.

$$d(\mathbf{X}(\theta_0), \mathbf{X}(\theta_1)) \geq d(\mathbf{Y}(\theta_0), \mathbf{Y}(\theta_1))$$

All distances in the Ali-Silvey class [1] have this property by construction. Distances in this class have the form $d(\mathbf{X}(\theta_0), \mathbf{X}(\theta_1)) = f(\mathcal{E}_0[c(\Lambda(\mathbf{X}))])$ where $\Lambda(\mathbf{X})$ represents the likelihood ratio of the probability distributions that statistically characterize the inputs, $c(\cdot)$ is convex, \mathcal{E}_0 is expected value with respect to the distribution describing $\mathbf{X}(\theta_0)$ and $f(\cdot)$ is a non-decreasing

function. Some distance measures not in the Ali-Silvey class satisfy the data processing theorem as well.

In choosing a distance measure, we do not want to restrict the kind of signal that represents information. We seek distance measures that quantify the difference between information-induced changes in a signal's structure. A measure like mean-squared error will not suffice because it depends on the signal being real-valued, and many signals are not (point processes and symbolic signals [8], for example). Instead of selecting a measure that is a direct function of the signal, we use measures that are functions of the signal's joint probability law. Examples of such measures are those in the Ali-Silvey class. In this way, we can assess how different two signals are regardless of their structure so long as they are stochastic.

Among possible distance measures that depend on the probability law, viable ones must satisfy the data processing theorem. In addition, we need the distance measure to express the performance capabilities of information processing systems, which fall into two broad categories: classification and estimation systems. In estimation, one fundamental bound on the mean-squared estimation error is the Cramér-Rao bound, which depends on the inverse of the Fisher information matrix $\mathbf{F}(\theta)$. Many distance measures have the property that when the information parameter vector is real-valued,

$$d(\mathbf{X}(\theta_0), \mathbf{X}(\theta_0 + \delta\theta)) \approx K \cdot \delta\theta' \mathbf{F}_{\mathbf{X}}(\theta_0) \delta\theta$$

for sufficiently small perturbations $\delta\theta$. Here K is a constant that depends only the distance measure. This property is known as the *locally Gaussian property*: When θ is the mean of a Gaussian distribution, this result applies regardless of the perturbation size. For a given small perturbation, large distance means a large Fisher information, which in turn means a small estimation error is possible. Many distance measures have this property. Fewer distance measures relate to the performance of optimal classification systems. One result known as Stein's Lemma states that a Neyman-Pearson detector's performance (probability of a type I or type II error) decays exponentially in the distance between the two signals [7].

$$\Pr[\text{error}] \sim 2^{-d(\mathbf{X}(\theta_0), \mathbf{X}(\theta_1))}$$

Thus, a large distance means a more rapid exponential decrease in the error probability. The perturbational result implies that all binary detection problems wherein two hypotheses differ by a small amount have a Gaussian-like performance character. Note that the optimal detector must be used and that it may not be locally Gaussian.

One distance measure known to satisfy Stein’s Lemma and have the locally Gaussian property is a particular Ali-Silvey distance, the Kullback-Leibler (KL) distance [9].

$$d_{\text{KL}}(\mathbf{X}(\theta_0), \mathbf{X}(\theta_1)) = \int p_{\mathbf{X}(\theta_0)}(\mathbf{x}) \log \frac{p_{\mathbf{X}(\theta_0)}(\mathbf{x})}{p_{\mathbf{X}(\theta_1)}(\mathbf{x})} d\mathbf{x}$$

By computing the KL distance, we can infer how easily two signals differing in information content can be discriminated and how well information parameters can be estimated. This quantity is not necessarily symmetric in its arguments, which means that it cannot serve as a distance in the strict sense. Be that as it may, constructs that create a geometric framework for classification problems indicate that *no* distance measure can exist on the manifold of probability measures [3]. Consequently, using a distance measure that cannot be formally used as a metric (in a Hilbert space setting, for example) makes sense, but does not restrict the mathematical structure that can be used in our theory of information processing.

To quantify how systems process information, we explore the quantity γ , the *information transfer ratio*, defined as the ratio of the distance between the two output distributions and the distance between the corresponding input distributions.

$$\gamma_{\mathbf{X}, \mathbf{Y}}(\theta_0, \theta_1) = \frac{d(\mathbf{Y}(\theta_0), \mathbf{Y}(\theta_1))}{d(\mathbf{X}(\theta_0), \mathbf{X}(\theta_1))}$$

This ratio is always less than or equal to one, with one meaning perfect reproduction of the information expressed by the input in the output and with zero meaning none of the information is represented in the output. The special case wherein the information parameter is perturbed ($\theta_1 = \theta_0 + \delta\theta$) yields interesting results. When the distance measure is in the Ali-Silvey class, we can explicitly write the information transfer ratio.

$$\gamma_{\mathbf{X}, \mathbf{Y}} = \frac{f(c(1)) + \frac{1}{2}f'(c(1))c''(1)\delta\theta' \mathbf{F}_{\mathbf{Y}}(\theta)\delta\theta}{f(c(1)) + \frac{1}{2}f'(c(1))c''(1)\delta\theta' \mathbf{F}_{\mathbf{X}}(\theta)\delta\theta}$$

With the reasonable assumption that $d(\mathbf{X}(\theta_0), \mathbf{X}(\theta_0)) = 0$ (which corresponds to the intuition that the distance between same distributions is zero, and holds for the Kullback-Leibler distance and many other distance measures), it follows that $f(c(1)) = 0$ and this expression simplifies to

$$\gamma_{\mathbf{X}, \mathbf{Y}} = \frac{\delta\theta' \mathbf{F}_{\mathbf{Y}}(\theta)\delta\theta}{\delta\theta' \mathbf{F}_{\mathbf{X}}(\theta)\delta\theta}.$$

We refer to this result as the *local invariance property*: The information transfer ratio for perturbational changes is invariant to the choice of distance measure.

Using the information transfer ratio, we can quantify how various system organizations can affect 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.

Systems in cascade

If two systems are in cascade, with the first system’s output serving as the second system’s input, the overall information transfer ratio is the product of the component ratios. Specifically, if $\theta \rightarrow \mathbf{X} \rightarrow \mathbf{Y} \rightarrow \mathbf{Z}$ form a Markov chain, $\gamma_{\mathbf{X}, \mathbf{Z}} = \gamma_{\mathbf{X}, \mathbf{Y}} \cdot \gamma_{\mathbf{Y}, \mathbf{Z}}$. Because information transfer ratios cannot exceed one, this result means that once a system reduces γ for some information change, that loss of information representation capability cannot be recovered.

Multi-input systems

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.

$$\frac{1}{\gamma_{\mathbf{X}, \mathbf{Y}}(\theta_0, \theta_1)} = \sum_i \frac{1}{\gamma_{X_i, \mathbf{Y}}(\theta_0, \theta_1)}.$$

One consequence of this result is that the overall information transfer ratio is bounded by the smallest component ratios: $\gamma_{\mathbf{X}, \mathbf{Y}}(\theta_0, \theta_1) \leq \min_i \gamma_{X_i, \mathbf{Y}}(\theta_0, \theta_1)$. Thus, inefficient use of one input by a multi-input system dominates the overall system’s information processing capability. Note, however, that each of these component γ s can exceed one because the Markov chain assumption of the data processing theorem does not hold for them. The Markov chain assumption is presumed to hold for the overall information transfer ratio (all inputs are contained in the vector \mathbf{X}).

Multi-output systems

Let a system have one input and two outputs: $\mathbf{Y} = (Y_1, Y_2)$. The overall information transfer ratio is related to the component ratios as

$$\gamma_{\mathbf{X}, \{Y_1, Y_2\}}(\theta_0, \theta_1) = \gamma_{\mathbf{X}, Y_1}(\theta_0, \theta_1) + \gamma_{\mathbf{X}, Y_2|Y_1}(\theta_0, \theta_1),$$

where the second term uses the distance between the output’s second component conditioned on first. Thus, a distributed representation of information can help (increase the information transfer ratio) provided that the additional outputs convey additional information.

Parallel systems

Generalizing the previous example, consider N outputs that are *conditionally* independent given the input. This assumption amounts to assuming that \mathbf{X} serves as the input to N parallel systems. We want to explore how the information transfer ratio changes as the number of systems increases. We calculated the information transfer ratio for two special cases. In the first, the input is exponentially distributed and each output has conditional mass function given as $p_{Y_i|X}(n) = x^n e^{-x} / n!$.

$$\gamma_{\mathbf{X}, \mathbf{Y}}(\theta_0, \theta_1) = \frac{\ln \frac{\theta_1}{\theta_0} + \frac{\theta_1 + N}{\theta_1} \ln \frac{\theta_0 + N}{\theta_1 + N}}{\ln \frac{\theta_1}{\theta_0} + \frac{\theta_0 - \theta_1}{\theta_1}}$$

It is not hard to show that when $N \rightarrow \infty$, $\gamma_{\mathbf{X}, \mathbf{Y}}(\theta_0, \theta_1) \rightarrow 1$ and that the differential increase in γ , defined to be $\gamma(N+1) - \gamma(N)$, is proportional to $1/N^2$. A similar result holds when we consider a system with a Gaussian input ($\mathbf{X} \sim \mathcal{N}(m, \sigma^2)$), and the output repeats the input with first-order Gauss-Markov noise added ($\mathbf{Y} = \mathbf{X} \cdot \mathbf{1} + \mathbf{Z}$ where $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{K}_N)$). Again, asymptotically the differential gain is $1/N^2$.

We conjecture that this result applies widely when the parallel systems have identical statistical input-output relations (in other words, they share the same conditional probability $p_{Y_i|X}$). We can show that as more systems are added, the information transfer ratio increases so long as the additional system is not functionally equivalent to the others. This condition generalizes the previous result describing the processing efficacy of adding a second output to create a multi-output system. If the differential increase in γ is indeed proportional to $1/N^2$, adding more systems quickly reaches a point of diminishing returns. From another 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.

3. RELATIONS TO CLASSIC INFORMATION THEORY

We have found some interesting differences between calculations for channel capacity and maximizing the information transfer ratio with respect to the input's probability distribution. While the capacity-achieving input probability does not have much meaning for communications systems, the one that maximizes γ does. This distribution maximizes the system's ability to represent information in its output.

Consider a binary channel that is not necessarily symmetric. We maximized the information transfer ratio with respect to the input *a priori* probabilities, searching in particular for the input parameter p whose differential change will yield largest information transfer ratio. Letting ϵ_1, ϵ_2 denote the cross-over probabilities, the capacity-achieving input probability is $p = \frac{2^{H(\epsilon_1)}}{2^{H(\epsilon_1)} + 2^{H(\epsilon_2)}}$, where $H(\cdot)$ denotes entropy of a binary-valued random variable. The resulting capacity is

$$C = \log_2(2^{-H(\epsilon_1)} + 2^{-H(\epsilon_2)}) .$$

Using our approach, let θ_0 correspond to p and $\theta_1 = p + \delta p$. Maximizing $\gamma_{\mathbf{X}, \mathbf{Y}}(p + \delta p, p)$ occurs at a different value of input probability:

$$p = \frac{1}{\epsilon_1 - \epsilon_1^2 - \epsilon_2 + \epsilon_2^2} (\epsilon_1 - \epsilon_1^2 - \sqrt{(\epsilon_1 - \epsilon_1^2)(\epsilon_2 - \epsilon_2^2)})$$

with the maximum information transfer equaling

$$\gamma_{\mathbf{X}, \mathbf{Y}} = \frac{(\epsilon_2^2 - \epsilon_2 + V)(\epsilon_1^2 - \epsilon_1 + V)}{(\epsilon_2 \epsilon_1 - \epsilon_2 + V)(\epsilon_2 \epsilon_1 - \epsilon_1 + V)} ,$$

where $V = \sqrt{\epsilon_1(\epsilon_1 - 1)\epsilon_2(\epsilon_2 - 1)}$. The two ways of determining the optimal p can yield quite different results. Because capacity and information transfer ratio are very different quantities (γ is a normalized quantity and capacity is not), comparing their values for a given system doesn't seem fruitful. We point out that capacity cannot, in general, be normalized (capacity has no theoretical upper bound for continuous-amplitude signals).

4. CONCLUSIONS

We have seen in numerous examples that the information transfer ratio behaves well and that it gives meaningful answers. Several results presented here do not strongly depend on which distance measure is used to compute the ratio. What is critical is that the distance measure be computed from probability distributions, as do Ali-Silvey distances. By using such distance measures, the distance calculation does not rely on any assumption about the kind of signal that represents the information. Analog signals, discrete-time signals, point processes and symbolic sequences can all be dealt with on the same footing.

The approach of using information change to probe a system's processing represents an important concept. Without change, static measures such as mutual information and entropy could be used. However, such measures consider a signal as an indivisible quantity, and do

not directly measure what is information-bearing and what is not. For us, the information source must reveal itself by changing information content. Because information can have a complex structure, different changes can reveal differing information representation efficacies and information processing capabilities. For example, in sensory systems like the visual system, one pervasive information processing strategy is extraction of features from the input signal. Some visual system components extract edge information while others extract color information. We employ distance measures that reflect the data processing theorem and apply them separately to the input and output. By studying how various information changes affect the information transfer ratio, we can quantify the processing capabilities of these subsystems.

One can envisage an information processing system as an *information filter*, wherein certain information changes about some operating point have larger gains (information transfer ratios) than others. For example, changing the edge gradient may have little effect on a color-sensitive system’s output and yield a small γ ; color changes would yield a larger γ . The resulting surface defined over parameter perturbations can be likened to a transfer function, relating information fidelity rather than sinusoidal signal amplitude. Such transfer functions measure how a system processes information and allows comparison of information loss with theoretical upper limits (the data processing theorem).

Our analysis of various system architectures indicates that 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, it can never be regained unless earlier information-bearing signals can serve as inputs to later stages. Underlying our use of the KL distance is its relation to the performance of optimal processing systems. Obtaining a small information transfer ratio can mean ineffective processing. However, we have results that show that in some cases the maximal value of γ can be much less than one. In our study of neural systems [6], for example, we measured very small transfer ratios ($\approx 10^{-3}$), but analysis showed that ratios could not be much bigger for an ideal system performing a similar information transformation. Here, the input is an analog signal and the output is a neural spike train, which is accurately modeled as a point process. When we imposed output rate restrictions on our analytic model similar to those we measured, the small value for γ resulted. If this restriction is relaxed, larger values for the information transfer ratio can be obtained.

The digital channel example indicates that the in-

formation transfer ratio and channel capacity measure different aspects of an information system’s behavior. Capacity measures the maximum data rate that can sustain reliable communication through some channel. The information transfer ratio quantifies a system’s information processing capability, which measures how well it extracts some information-bearing aspects and suppresses others. The example shows that maximizing information transfer does not occur at the same input distribution that yields capacity. We are working to understand this dichotomy. Because of the local invariance property (the information transfer ratio does not vary with Ali-Silvey distance measure), we do know that the optimal transfer ratio and the optimizing input distribution do not depend on choice of distance measure.

The major drawback in our approach is that calculating KL distances analytically can be difficult, even impossible. However, we have a complete empirical theory [5] that frames how to estimate KL distances from data. Consequently, we can *measure* information transfer ratios for real systems and compare them with theoretical predictions made from analysis or simulation.

From a broader perspective, our information processing theory goes beyond Shannon’s classic results. The entropy limit on source coding and the notion of channel capacity are catholic: They formulate how to efficiently represent and reliably communicate discrete-symbol streams and place no restriction on whether the stream expresses information or not. More penetrating theories must be concerned with what digital and more general streams represent and how classes of systems more general than channels affect what they represent. Our theory tries to address Weaver’s vision 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 distance measures that reflect optimal-processing performance, we can quantify the effectiveness of *any* signal’s information-bearing capability. If information processing systems do result in behaviors, these behaviors result in signals of some sort. If so, systems that take in signals, analyze them, and produce behaviors are just like any other system, and we can use our framework to quantify how well the original information influences decisions and consequent actions. In this way, we address Weaver’s desire for “influential” analysis. Consequently, it appears that one formalism can address the last two components of Weaver’s communication system analysis program.

REFERENCES

- [1] S.M. Ali and D. Silvey. A general class of coefficients of divergence of one distribution from another, *J. Roy. Stat. Soc. B*, 28: 131–142, 1966.
- [2] T.M. Cover and J.A. Thomas. *Elements of Information Theory*, John Wiley and Sons, Inc., 1991.
- [3] A. Dabak and D.H. Johnson. A geometry for detection theory. *Conf. Info. Sciences and Systems*, Princeton, NJ, March 1992.
- [4] D.H. Johnson. Toward a theory of signal processing, IT Workshop on Detection, Estimation, Classification, and Imaging, Santa Fe, NM, USA, Feb. 24-26, 1999.
- [5] D.H. Johnson, C. Gruner, K. Baggerly, C. Seshagiri. Information-theoretic analysis of neural coding. To appear in *J. Comp. Neuroscience*, 2000.
- [6] D.H. Johnson, C.M. Gruner, R.M. Glantz. Quantifying information transfer in spike generation. *Computational Neuroscience '99*, Pittsburgh, PA, July 18–25, 1999.
- [7] D.H. Johnson and G.C. Orsak. Relation of signal set choice to the performance of optimal non-Gaussian detectors. *IEEE Trans. Communication*, 41: 1319–1328, 1993.
- [8] D.H. Johnson and W. Wang. Symbolic signal processing. *ICASSP '99*, Phoenix, Arizona, March, 1999.
- [9] S. Kullback. *Information Theory and Statistics*, Dover Publications, NY, 1967.
- [10] C.E. Shannon and W. Weaver. *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, Illinois, 1949.
- [11] S. Sinanović and D.H. Johnson. Toward a theory of information processing. *Intern. Sym. Info. Theory*, Sorrento, Italy, 2000.