ASYMPTOTIC RATES OF THE INFORMATION TRANSFER RATIO

Sinan Sinanović and Don H. Johnson

Computer and Information Technology Institute
Department of Electrical and Computer Engineering
Rice University
Houston, Texas 77005–1892
sinan@rice.edu, dhj@rice.edu

ABSTRACT

Information processing is performed when a system preserves aspects of the input related to what 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 way signals represent information. Kullback-Leibler distance, information-theoretic measure which reflects the data processing theorem, is calculated on the input and output separately and compared to obtain information transfer ratio. We consider the special case where input serves several parallel systems and show that this configuration has the capability to represent the input information without loss. We also derive bounds for asymptotic rates at which the loss decreases as more parallel systems are added and show that the rate depends on the *input* distribution.

1. INTRODUCTION

Signals represent information. By operating on its input signal(s), systems perform information processing. Most systems have an information loss and act as "information filters." In quantifying the processing of arbitrary systems, non-linearities and mixed signal varieties means that classical methods, including using mutual information, fail to capture all a system does.

In our earlier work [6, 10], we first described our approach. We conceptually (or in reality, for empirical work) induce controlled changes of the information represented by a system's input and probe how well the system preserves these changes in its output. By measuring how different the two inputs and the corresponding outputs are, we calculate the information transfer ratio: the ratio of the distances between the outputs and the inputs. Because of the Data Processing Theorem (DPT), this ratio must be between zero and one, with the maximum value meaning the input change is entirely preserved in the output (no information loss).

This paper concerns the special case wherein the input signal serves as the input to several parallel systems (see Figure 1) each of which processes the signal separately from the other. We assume that the systems are stochastically identical: given the input, each output has the same probability distribution. The output signals do differ; they are members of the same ensemble. This generic model describes MIMO communication systems and simple neural populations. This paper determines how well the input information is represented by the collective output. We show that under very general conditions, this simple distributed, non-cooperative (the systems do not interact with each other) processing system will asymptotically preserve the input's information in the collective output. We explicitly determine bounds on the rate at which the information transfer ratio approaches one, and show that the bounds depend on the probabilistic structure of the *input*, not on that of the system's output. Our approach is to consider an optimal processing system that collects the outputs to yield an estimate of the input (see Figure 1). We then calculate the asymptotic distribution of the estimate, derive the distance between estimates that result from the two inputs, and find the information transfer ratio between the input and the

This work was supported by the National Science Foundation under Grant CCR-0105558.

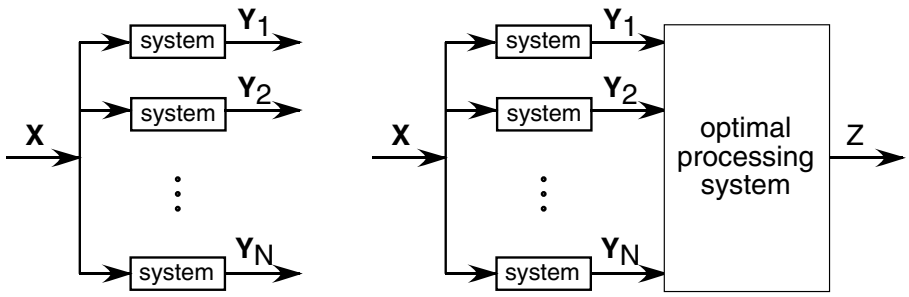

Figure 1: The left figure shows the set-up of our problem: systems transform input $\mathbf{X}$ in an arbitrary way to produce outputs $\mathbf{Y}$, which are conditionally iid. The figure to the right shows the set-up we use to find the asymptotic rate of the information transfer ratio. In the discrete case, optimal processing is the likelihood ratio detector. In the continuous case, optimal processing is the maximum likelihood estimator of $\mathbf{X}$.

estimate. Because of the DPT, this ratio forms a lower bound on the information transfer ratio between the input and the parallel system's collective output.

2. QUANTIFYING INFORMATION PROCESSING

We symbolically represent information by parameter $\boldsymbol{\theta}$. Let $\mathbf{X}$ represent a system's input signal and $\mathbf{Y}$ its output. The form of these signals is arbitrary but they must have a probabilistic description. All Ali-Silvey distances [1] satisfy the Data Processing Theorem by construction. Expressed in terms of distances, this theorem [3] states that if $\boldsymbol{\theta}$, $\mathbf{X}$, and $\mathbf{Y}$ form a Markov chain, then

$$d\big(\mathbf{X}(\boldsymbol{\theta}_0), \mathbf{X}(\boldsymbol{\theta}_1)\big) \geq d\big(\mathbf{Y}(\boldsymbol{\theta}_0), \mathbf{Y}(\boldsymbol{\theta}_1)\big) \qquad (1)$$

We use one particular Ali-Silvey distance, the Kullback-Leibler (KL) distance, extensively because of its convenience and importance.

$$d\big(\mathbf{X}(\boldsymbol{\theta}_0), \mathbf{X}(\boldsymbol{\theta}_1)\big) = \mathcal{E}_0\left[\log p(\mathbf{X}(\boldsymbol{\theta}_0))/p(\mathbf{X}(\boldsymbol{\theta}_1))\right]$$

We define the quantity γ, the *information transfer ratio*, as the ratio of KL distances between the two output distributions and the corresponding input distributions.

$$\gamma_{\mathbf{X},\mathbf{Y}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_0) = \frac{d\big(\mathbf{Y}(\boldsymbol{\theta}_1), \mathbf{Y}(\boldsymbol{\theta}_0)\big)}{d\big(\mathbf{X}(\boldsymbol{\theta}_1), \mathbf{X}(\boldsymbol{\theta}_0)\big)}$$

The larger γ is, the greater the fidelity with which the output represents the change in the input. Note that this quantity can be defined regardless of the nature of the signals $\mathbf{X}$ and $\mathbf{Y}$, and regardless of how $\boldsymbol{\theta}$ is represented by $\mathbf{X}$ and $\mathbf{Y}$.

3. ASYMPTOTIC RATES OF THE INFORMATION TRANSFER RATIO

3.1. Discrete input distribution case

Let $\mathbf{X}$ be drawn from a set and have discrete probability distribution. We are interested in the asymptotic (in N, the number of parallel systems) behavior of the information transfer ratio:

$$\gamma_{\mathbf{X},\mathbf{Y}^{(N)}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_0) = \frac{d\big(\mathbf{Y}^{(N)}(\boldsymbol{\theta}_1), \mathbf{Y}^{(N)}(\boldsymbol{\theta}_0)\big)}{d\big(\mathbf{X}(\boldsymbol{\theta}_1), \mathbf{X}(\boldsymbol{\theta}_0)\big)}$$

Consider a categorization problem where the output $\mathbf{Y}^{(N)} = \{\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_N\}$ is observed to determine which letter of the input alphabet occured. We use an optimal classifier for this purpose. Let $M = |\mathbf{X}|$ and let $\mathbf{Z}$ be the output decision (see Figure 1). 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, we know that the error probabilities in this crossover diagram do *not* depend on the *a priori* symbol probabilities so long as they are non-zero. Let π_m^i denote the *a priori* probability of $\mathbf{X}_m$ under $\boldsymbol{\theta}_i$ and $\epsilon_m^j = \Pr[\mathbf{Z}_m|\mathbf{X}_j]$ the crossover probability. Then, the output symbol probabilities are

$$\Pr[\mathbf{Z}_m|\boldsymbol{\theta}_i] = \pi_m^i(1 - \sum_{j \neq m} \epsilon_j^m) + \sum_{k \neq m} \pi_k^i \epsilon_m^k$$

Note that $\epsilon_m^m \to 1$ as $N \to \infty$. This expression for $\Pr[\mathbf{Z}_m|\boldsymbol{\theta}_i]$ is written in terms of the crossover probabilities ϵ_i^j, $i \neq j$ that tend to 0 with increasing N. Now, we compute the output Kullback-Leibler distance for

$\mathbf{Z}$ and approximate it for small crossover probabilities.

$$d\big(\mathbf{Z}(\boldsymbol{\theta}_1),\mathbf{Z}(\boldsymbol{\theta}_0)\big) = d\big(\mathbf{X}(\boldsymbol{\theta}_1),\mathbf{X}(\boldsymbol{\theta}_0)\big) + \sum_{j,m} \pi_j^1 \left(1 - a_m/a_j + \log(a_m/a_j)\right) \epsilon_m^j + o(\epsilon_{max}) \quad (2)$$

where $a_j = \pi_j^1/\pi_j^0$ and

$$\epsilon_{max} = f(N) \exp\left\{-N \min_{i\neq j} C(p(\mathbf{Y}|\mathbf{X}_i), p(\mathbf{Y}|\mathbf{X}_j))\right\}$$

with $\mathbf{Y}$ representing one system's output, $f(\cdot)$ a slowly varying function in the sense that

$$\lim_{N\to\infty} [\ln f(N)]/N = 0$$

and $C(\cdot,\cdot)$ denoting Chernoff infomation [8]. Since $1 - x + \log x \leq 0 \ \forall x > 0$, the term inside the parentheses (2) is non-positive. Therefore, we have that

$$d\big(\mathbf{Z}(\boldsymbol{\theta}_1),\mathbf{Z}(\boldsymbol{\theta}_0)\big) \geq d\big(\mathbf{X}(\boldsymbol{\theta}_1),\mathbf{X}(\boldsymbol{\theta}_0)\big) - K\epsilon_{max} + o(\epsilon_{max})$$

where

$$K = -\sum_{j,m} \pi_j^1 (1 - a_m/a_j + \log(a_m/a_j)) \geq 0$$

Since according to the DPT (see (1)), $d\big(\mathbf{Y}(\boldsymbol{\theta}_1),\mathbf{Y}(\boldsymbol{\theta}_0)\big) \geq d\big(\mathbf{Z}(\boldsymbol{\theta}_1),\mathbf{Z}(\boldsymbol{\theta}_0)\big)$, we have:

$$\gamma_{\mathbf{X},\mathbf{Y}}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_0) \geq 1 - \frac{K}{d\big(\mathbf{X}(\boldsymbol{\theta}_1),\mathbf{X}(\boldsymbol{\theta}_0)\big)} f(N) \times \exp\left\{-N \min_{i\neq j} C(p(\mathbf{Y}|\mathbf{X}_i), p(\mathbf{Y}|\mathbf{X}_j))\right\}.$$

We conclude that for the case of discrete input distribution with the 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 $N \to \infty$.

3.2. Continuous input distribution case

Let the probability distribution of the input, $\mathbf{X}$, be continuous. To determine the rate of increase of the information transfer ratio, we use the same approach but with $\mathbf{Z}$ being the maximum likelihood estimator (MLE) of $\mathbf{X}$. Under certain regularity conditions [4] and because the $\mathbf{Y}_i$'s are conditionally independent and identically distributed, we know that the MLE is asymptotically Gaussian. We can now obtain probability density of $\mathbf{Z}$:

$$p_{\mathbf{Z}}(\mathbf{z}) = \int p_{\mathbf{X}}(\mathbf{x}) \left(\det\left\{\frac{N\mathbf{F}_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{\theta})}{2\pi}\right\}\right)^{1/2} \times \exp\left(-\frac{N(\mathbf{z}-\mathbf{x})'\mathbf{F}_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{\theta})(\mathbf{z}-\mathbf{x})}{2}\right) d\mathbf{x}$$

where $\mathbf{F}_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{\theta})$ is the conditional Fisher information. If the third derivative of the input probability density function, $p_{\mathbf{X}}(\cdot)$, is bounded, we can expand $p_{\mathbf{X}}(\cdot)$ in a Taylor series around $\mathbf{z}$, up to the third-order term and then perform term-by-term integration. This amounts to the *Laplace approximation* for an integral. The probability density of $\mathbf{Z}$ can be then expressed as

$$p_{\mathbf{Z}}(\mathbf{z}) = p_{\mathbf{X}}(\mathbf{z}) + \frac{1}{2}\text{tr}\{\mathbf{H}(\mathbf{z})\mathbf{F}^{-1}_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{\theta})\}\frac{1}{N} + O\left(\frac{1}{N^{3/2}}\right)$$

where $\mathbf{H}(\mathbf{z})$ is the Hessian of $p_{\mathbf{X}}(\cdot)$ evaluated at $\mathbf{z}$. For two input densities, governed by $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$, the two corresponding output densities, $p_{\mathbf{Z}_0}(\mathbf{z})$ and $p_{\mathbf{Z}_1}(\mathbf{z})$, are obtained. Letting the coefficients of $1/N$ be $r_i(\mathbf{z}) = \frac{1}{2}\text{tr}\{\mathbf{H}(\mathbf{z})\mathbf{F}^{-1}_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{\theta})\}$ for $i = 0, 1$, the Kullback-Leibler distance between those two output distributions can be calculated as:

$$d\big(\mathbf{Z}(\boldsymbol{\theta}_1),\mathbf{Z}(\boldsymbol{\theta}_0)\big) = d\big(\mathbf{X}(\boldsymbol{\theta}_1),\mathbf{X}(\boldsymbol{\theta}_0)\big) - \frac{K}{N} + O\left(\frac{1}{N^{3/2}}\right)$$

where

$$K = -\int r_1(\mathbf{z}) + r_1(\mathbf{z}) \log \frac{p_{\mathbf{X}_1}(\mathbf{z})}{p_{\mathbf{X}_0}(\mathbf{z})} - \frac{p_{\mathbf{X}_1}(\mathbf{z}) r_0(\mathbf{z})}{p_{\mathbf{X}_0}(\mathbf{z})} d\mathbf{z}.$$

Because of the data processing theorem, we know that $d\big(\mathbf{Y}(\boldsymbol{\theta}_1),\mathbf{Y}(\boldsymbol{\theta}_0)\big) \geq d\big(\mathbf{Z}(\boldsymbol{\theta}_1),\mathbf{Z}(\boldsymbol{\theta}_0)\big)$. Finally, we conclude that the information transfer ratio asymptotically approaches 1 at (at least) the rate proportional to $1/N$ when $N \to \infty$:

$$\gamma_{\mathbf{X},\mathbf{Y}^{(N)}}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_0) \geq 1 - \frac{K}{N d\big(\mathbf{X}(\boldsymbol{\theta}_1),\mathbf{X}(\boldsymbol{\theta}_0)\big)} + O\left(\frac{1}{N^{3/2}}\right).$$

4. CONCLUSION

We investigated the behavior of the information transfer ratio for a particularly interesting distributed processing system. Here each system processes its input in stochastically identical ways and the systems do not interact with each other. Our results show that regardless of the information encoding strategy or the nature of the input and output signals, this processing structure asymptotically yields a perfect representation of the input's information. The only assumption made is that the input information change does elicit a change in each system's output. Therefore, parallel systems need not "cooperate" to achieve perfect reproduction of the input.

Interestingly, how the information transfer ratio increases depends on whether the input distribution is discrete or continuous. In the discrete case, the information transfer ratio increases exponentially or faster, and in the continuous case it increases as $1/N$. Examples confirm this behavior. For instance, if the input is a Gaussian random variable with θ affecting the mean and each system simply adds a statistically independent Gaussian random variable having variance σ^2, the information transfer ratio equals $(1 + \sigma^2/(\sigma_x^2 N))^{-1} \approx 1 - \sigma^2/(N\sigma_x^2)$. Our results also mean that regardless of the system that processes the information-bearing signal $\mathbf{X}(\boldsymbol{\theta})$, encoding the information in signals that have a discrete distribution requires fewer non-cooperative systems to achieve a given level of fidelity (setting γ equal a criterion value) than would having a continuous distribution. In figure 2, a factor of two fewer systems are needed in the discrete case to satisfy the preformance criterion.

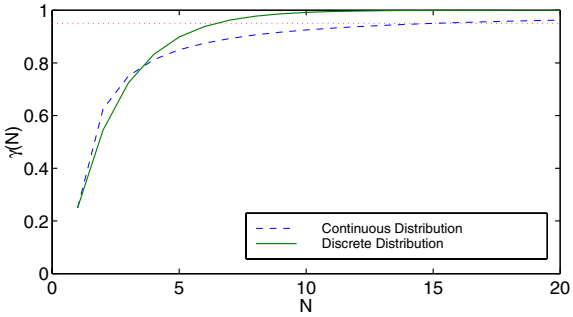

Figure 2: The two asymptotic formulas for the information transfer ratios are plotted as a function of the number of non-cooperative systems on the assumption the formulas apply for all N. Each formula had the same value for $\gamma(1)$. A criterion value of 0.95 is shown.

5. 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, Vol 28, No. 1, 1966, pp.131-142.

[2] M. Basseville, Distance Measures for Signal Processing and Pattern Recognition, *Signal Processing* 18, 1989, pp. 349-369.

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

[4] H. Cramér, *Mathematical Methods of Statistics*, Princeton University Press, Princeton, New Jersey, 1946.

[5] I. Cziszar and J. Korner, *Information Theory, Coding Theorems for Discrete Memoryless Systems*, Akademiai Kiado, Budapest, Hungary, 1986.

[6] 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.

[7] S. Kullback, *Information Theory and Statistics*, Dover Publications, NY, 1967.

[8] C. C. Leang and D. H. Johnson, On the Asymptotics of M-Hypothesis Bayesian Detection, *IEEE Trans. Inform. Theory*, vol. 43, No. 1, January 1997, pp.280-282.

[9] E.L. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd edition, Springer-Verlag, New York, 1998.

[10] S. Sinanović and D.H. Johnson, Toward a theory of information processing. International Symposium on Information Theory, Sorento, Italy, 2000.

[11] H.L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I*, John Wiley and Sons, New York, 1968.