INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

**Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600**

UMI[®]

RICE UNIVERSITY

Toward a Theory of Information Processing

by

Sinan Sinanović

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Master of Science

APPROVED, THESIS COMMITTEE:

Dr. Don H. Johnson, Chairman
Professor
Electrical and Computer Engineering

Dr. Richard Baraniuk
Associate Professor
Electrical and Computer Engineering

Dr. Robert Nowak
Assistant Professor
Electrical and Computer Engineering

Houston, Texas

April, 2000

UMI Number: 1399307

UMI[®]

UMI Microform 1399307

Copyright 2000 by Bell & Howell Information and Learning Company.

All rights reserved. This microform edition is protected against
unauthorized copying under Title 17, United States Code.

Bell & Howell Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

Toward a Theory of Information Processing

Sinan Sinanović

Abstract

Information processing is performed when a system preserves some aspects of the information encoded by its input while it suppresses others. To describe a system's information processing capability, input and output need to be compared in a way invariant to the signal's form and how it represents information. We describe an approach to quantify information processing based on applying controlled changes to the input and observing the corresponding output. Information-theoretic distance measures — those that reflect the data processing theorem — are calculated on the input and output separately and compared. The best candidate for the distance measure to describe signal processing is the information theoretic distance measure called the Kullback-Leibler distance (also known as the relative entropy). Properties of the resulting information transfer ratio are used to derive fundamental information processing properties of systems and interconnected systems.

Acknowledgments

I would like to thank my advisor, Dr. Don Johnson, for numerous ideas, constant support, and dialogue which often lead to better understanding of the problems (and more dialogue:). I thank my committee members, Dr. Richard Baraniuk and Dr. Robert Nowak, for suggestions, comments and questions — including those during my talks at pizza and group meetings. I would like to thank the graduate students in the ECE Dept. — especially those from the Systems group — for discussing with me about my research.

I also thank my family for their support, patience and understanding.

Contents

Abstract	ii
Acknowledgments	iii
List of Illustrations	vi
1 Introduction	1
2 Background	3
2.1 Controlled Change in the Input of Nerves	3
2.2 Understanding Animal Communication	3
2.3 The Meaning of the Honey-bees' Waggle Dance	4
2.4 A Unified View of Information Processing	5
3 Theory	7
3.1 Quantifying Information Processing	7
3.2 Invariance of γ with Respect to Distance Measure	9
3.3 A System Theory for Information Processing	10
3.3.1 Cascaded Systems	11
3.3.2 Multiple Inputs	11
3.3.3 Multiple Outputs	13
3.4 Systems that re-represent information	14
3.4.1 Continuous Input, Discrete Output	14
3.4.2 Non-symmetric Binary Channel	16

3.4.3	Information Transfer Ratio of a Low Pass Filter	18
4	Conclusions	20
A	Significance of the Kullback-Leibler Distance	23
A.1	Some Properties of the Kullback-Leibler Distance	23
A.2	Kullback-Leibler distance between two Gaussian random processes . .	25
B	Proofs for Section 3.3	27
B.1	Cascaded Systems	27
B.2	Multiple Inputs	27
B.3	Multiple Outputs	28
B.4	Information Transfer Ratio in the Case of N Outputs	28
C	Proofs for Section 3.4	30
C.1	Continuous Input, Discrete Output	30
D	Proof of Local Invariance of γ with Respect to Distance Measure	32
	Bibliography	34

Illustrations

3.1	Two systems in cascade	11
3.2	Multiple inputs, one output	12
3.3	One input, two outputs	13
3.4	Information transfer ratio: continous input, discrete output	15
3.5	Gaussian input, Poisson conditional output	16
3.6	Non-symmetric binary channel	17
3.7	Plot of information transfer ratio vs. p in binary channel case	18
3.8	Information transfer ratio for low-pass filter	19

Chapter 1

Introduction

In general, information processing is performed when a system enhances certain aspects of its input as it suppresses 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. Most systems do have a loss and act as “information filters.” 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, we use the transfer function or cross-correlation. However, in quantifying the processing of more complex systems, non-linearities and non-Gaussian effects cause classical methods to fail to capture all a system does. We considered using mutual information as a measure of information processing, but it turns out to be inadequate.

In our theory, we consider information to be a parameter vector θ that could represent the parameters (such as source amplitude or direction) that control signal features. Here, we take the parameters to be real numbers, but in general they could be complex or even symbolic (like the components of DNA). Unlike classical Shannon’s work [10], we do not assume that information is necessarily stochastic, because this is sometimes an unrealistic assumption. For example, if a person cannot hear very well, one may have to repeat the same sentence for that person to understand. Here, the information one is trying to convey did not change, but the signal representing it did change. While parameterizing information reduces the full generality of the problem,

it helps us develop results which can still be applied in many information processing systems.

We conceptually (or in reality, for empirical work) induce controlled changes of the information represented by a system's input and compare "distances" between inputs and outputs, respectively [6]. Controlled change is required because information content cannot be judged solely by a signal's appearance or properties. For example, the same information can be encoded in two different bit-streams or in neurons or as frequency-modulated signal. This principle of information change to probe a system's processing is rather fundamental and an important concept in science in general: it is the idea behind the method of experiment. It is also important to realize that the content of information is relative to the observer; in other words, information is in the eye of the beholder. Take the case of a multiuser channel; one user's signal is another's interference. Therefore, by considering information changes that we induce, we essentially specify what conveys information. Finally, by measuring the difference between two inputs before and after the change and comparing this difference to the corresponding output difference, we quantify how a system processes relevant information.

Chapter 2

Background

We have already mentioned that the principle of information change to probe system's processing is rather fundamental and important concept in science in general. Let us look at some of the more interesting applications of this principle.

2.1 Controlled Change in the Input of Nerves

It is well known that different kinds of injuries to the brain have helped scientists learn more about the responsibilities assigned to different parts of the brain. For example, injury to a particular part of the cortex may result in blindness, although there is nothing wrong with the eyes or the pathway. We could conclude that the injured cortical center is responsible for the processing of the visual information. Also, scientists can now noninvasively observe cortical activity using fMRI (functional magnetic resonance imaging). It is clear that the controlled "input" in this case has helped in revealing information about the operation of the neural system and brain.

2.2 Understanding Animal Communication

In understanding animal communication, change in the input is definitely necessary to determine the way animals "speak." For example, if a monkey notices a snake, it will signal the danger to other monkeys by using specific type of screaming. On the other hand, if the monkey notices a lion, it will scream in a different way to warn its group. These differences are easily noticed in the frequency domain representation of

monkey's screaming. This fact confirms the hypothesis that monkey screams are actually more than just a scream: they are type of relatively developed communication. Controlling the environment in this experiment is obviously crucial step in gaining insight into animal communication.

2.3 The Meaning of the Honey-bees' Waggle Dance

The famous honey-bees' waggle-dance is also understood by observing the subtle changes in the way bees "dance." It has been observed that some bees, upon returning to the hive, perform an unusual ritual: they walk in a certain direction on the hive wall and, at the same time, produce a buzzing noise (of varying length at different times). Upon careful analysis, biologists realized that the direction in which a bee walks actually reveals the angle from the sun at which the observer bees should fly to reach nectar-rich flowers. Furthermore, the length of the buzz is proportional to the distance at which the flowers are located: the longer the buzz, the farther the flowers. Mathematically speaking, the bees use a polar coordinate system to define the position of the target (i.e. nectar). As if this is not enough, the "dancer" bee's body retains pollen from the flower it just visited, which enables other bees to sense the scent and identify which kind of nectar and flowers they need to find. Thus, if it was not for the careful observation of the changes in the bee behavior, we would not be able to understand how bees communicate the information about the location of the nectar-rich flowers. Also, the unusual waggle-dance would appear meaningless to the unlearned observer; this fact confirms the claim that information is in the eye of the beholder.

2.4 A Unified View of Information Processing

In signal processing, one witnesses a great variety of signals: analog (video and audio), discrete (sampled and digital), point-processes (neural recordings) and multi-channel. In every area, different signal processing algorithms have been developed, but there is no single, unified view for evaluating the performances of these various methods. One needs to go up one level in abstraction: from signal processing to information processing. We would like to concentrate more on quantifying the information processing where the signal is less important, it being only the carrier of information. Even though the signal which carries the information is processed by the system, we are ultimately interested in the information and not its carrier. We propose to use the information-theoretic distance measures between probability distributions in order to quantify information transfer. The key result from information theory which provides this notion of an information “filter” is the *Data Processing Theorem*.

Each area of signal processing has developed its own measures for quantifying the distinguishability between probability distributions. For the purpose of quantifying information processing, the choice of distance measure is limited to those measures that satisfy the data processing theorem [4]. Therefore, the most logical choice is the Ali-Silvey family of distance measures which has the desired property [1]. Practically all useful distance measures, including Kullback-Leibler (KL), J-divergence, Chernoff, and Kolmogorov distances, belong to Ali-Silvey family. In a good comprehensive summary of the distance measures, Basseville warns that it is difficult to give a strong recommendation for the choice of a distance measure for some application [2]. However, J-divergence (which is, roughly speaking, a symmetrized version of the KL distance) is recommended in general since it is used both in theory and in practice. We maintain that the Kullback-Leibler distance should be used because of its importance

and convenience. For example, the Kullback-Leibler distance corresponds to the exponential rate of the false-alarm probability of the optimal Neyman-Pearson detector. On the other hand, for small parametric changes in the input, the Kullback-Leibler distance equals a quadratic form consisting of the incremental parametric change serving as the vector and the Fisher information matrix serving as the kernel. Chernoff distance is also related to the optimal detector performance via Stein's lemma (see Appendix A.1) and to the Fisher information matrix; however, Chernoff distance is not as practical as the Kullback-Leibler distance since its calculation involves optimization and it does not have additive property (see Appendix A.1) for the random variables which are *not* identically distributed. Therefore, the fact that the Kullback-Leibler distance is related to both detection and estimation theory and the fact that it is relatively easily calculated in practice (by using its additive property) provide strong reason for its use in quantifying information processing.

Chapter 3

Theory

In this chapter we develop a new framework for quantifying the information processing. We start with quite general assumptions on types of systems which will be explored. First, the data processing theorem — a fundamental result from information theory on which our measure for information transfer is based — is introduced. We then demonstrate a few basic results which illustrate how collections of systems process information. Cases of multiple inputs and multiple outputs are also discussed.

3.1 Quantifying Information Processing

We represent information by a collection of parameters coalesced into the vector $\boldsymbol{\theta}$. Let $\mathbf{X}$ represent a system's input signal and $\mathbf{Y}$ its output. We suppress the independent variable (e.g., time); also, these signals could be vectors of real or symbolic values. According to the Data Processing Theorem [4], if $\boldsymbol{\theta} \rightarrow \mathbf{X} \rightarrow \mathbf{Y}$ (i.e., if $\boldsymbol{\theta}$, $\mathbf{X}$, and $\mathbf{Y}$ form a Markov chain), then $I(\boldsymbol{\theta}; \mathbf{X}) \geq I(\boldsymbol{\theta}; \mathbf{Y})$ where $I(\boldsymbol{\theta}; \mathbf{X})$ is the mutual information: $I(\boldsymbol{\theta}; \mathbf{X}) = \mathcal{E}_{p(\boldsymbol{\theta}, \mathbf{X})}[\log_2 \frac{p(\boldsymbol{\theta}, \mathbf{X})}{p(\boldsymbol{\theta})p(\mathbf{X})}]$. Equality holds if and only if $\boldsymbol{\theta} \rightarrow \mathbf{Y} \rightarrow \mathbf{X}$ form a Markov chain (notice that $\mathbf{X}$ and $\mathbf{Y}$ changed place). Let $\mathbf{X}(\boldsymbol{\theta}_0)$, $\mathbf{X}(\boldsymbol{\theta}_1)$ represent input signals having different information content, with $\mathbf{Y}(\boldsymbol{\theta}_0)$, $\mathbf{Y}(\boldsymbol{\theta}_1)$ representing the corresponding outputs. Many distance measures, which we generically write as $d(\cdot, \cdot)$, also satisfy the data processing theorem in the sense that

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

Thus, the data processing theorem clearly states that neither deterministic nor statistical processing of $\mathbf{X}$ can increase the information that $\mathbf{X}$ contains about $\boldsymbol{\theta}$. Ali-Silvey distances [1], which have the form $d(\mathbf{X}(\boldsymbol{\theta}_0), \mathbf{X}(\boldsymbol{\theta}_1)) = f(\mathcal{E}_0[c(\Lambda(\mathbf{X}))])$, where $\Lambda(\cdot) = p_1(\cdot)/p_0(\cdot)$ represents the likelihood ratio of the probability distributions that statistically characterize the inputs, $c(\cdot)$ is convex, $\mathcal{E}_0[\cdot]$ is expected value with respect to the $p_0(\cdot)$ (distribution describing $\mathbf{X}(\boldsymbol{\theta}_0)$) and for $f(\cdot)$ is a non-decreasing function, satisfy the data processing theorem by construction.

We use one particular Ali-Silvey distance, the Kullback-Leibler (KL) distance, extensively because of its convenience and importance. (See Appendix A for important properties of the Kullback-Leibler distance.) The Kullback-Leibler distance belongs to Ali-Silvey class for $c(\Lambda) = \Lambda \log(\Lambda)$. The Kullback-Leibler distance, also known as relative entropy, between two probability mass functions p_1 and p_0 is defined as:

$$\mathcal{D}(p_1||p_0) = \sum_x p_1(x) \log \frac{p_1(x)}{p_0(x)}.$$

For continuous probability density functions, the definition is:

$$\mathcal{D}(p_1||p_0) = \int p_1(x) \log \frac{p_1(x)}{p_0(x)} dx.$$

Based on continuity arguments, the following convention is used: $0 \cdot \log(0/q) = 0$ and $p \cdot \log(p/0) = \infty$. The Kullback-Leibler distance can also be expressed as the expected value with respect to $p_1(\cdot)$ of $\log_2 \frac{p_1(x)}{p_0(x)}$

$$\mathcal{D}(p_1||p_0) = \mathcal{E}_{p_1} \left[\log_2 \frac{p_1(X)}{p_0(X)} \right].$$

The fundamental quantity we use to quantify information processing is γ , the *information transfer ratio*, defined as the ratio of KL distance between the two output distributions and the KL distance of the corresponding two input distributions

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

This ratio is bounded between zero and one: zero means none of the information change $\theta_0 \rightarrow \theta_1$ is represented by the output and one means perfect reproduction of the input information change whether or not the input and the output signals encode the information differently.

The special case when $\gamma_{\mathbf{X},\mathbf{Y}}(\theta_0, \theta_1) = 1$ is related to the concept of *sufficient statistic*. Assume we have a sample $\mathbf{X}$ from a distribution which belongs to a family $\{p_{\theta}(x)\}$ indexed by θ . Now, let $T(\mathbf{X})$ be any statistic (i.e., a function of the sample). Then $\theta \rightarrow \mathbf{X} \rightarrow T(\mathbf{X})$, and from the data processing theorem, it follows that

$$d(\mathbf{X}(\theta_0), \mathbf{X}(\theta_1)) \geq d(T(\mathbf{X}(\theta_0)), T(\mathbf{X}(\theta_1))) \quad (3.1)$$

We already know that if equality holds then no information is lost. Roughly speaking, a statistic $T(\mathbf{X})$ is called sufficient for θ if it contains all the information about θ in $\mathbf{X}$. Formally, a statistic $T(\mathbf{X})$ is called sufficient statistic relative to the family $\{p_{\theta}(x)\}$ if $\mathbf{X}$ is independent of θ given $T(\mathbf{X})$ (i.e. $\theta \rightarrow T(\mathbf{X}) \rightarrow \mathbf{X}$ forms a Markov chain). In [1], it is shown that the equality holds in (3.1) (i.e., the information transfer ratio $\gamma_{\mathbf{X},\mathbf{Y}}(\theta_0, \theta_1) = 1$) if and only if $T(\mathbf{X})$ is sufficient statistic for θ .

3.2 Invariance of γ with Respect to Distance Measure

The special case wherein the information parameter is slightly 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 as follows (See Appendix D for details of the proof.)

$$\gamma_{\mathbf{X},\mathbf{Y}}(\theta, \theta + \delta\theta) = \frac{f(c(1)) + \frac{1}{2}f'(c(1))c''(1)\delta\theta'\mathbf{F}_{\mathbf{Y}}(\theta)\delta\theta}{f(c(1)) + \frac{1}{2}f'(c(1))c''(1)\delta\theta'\mathbf{F}_{\mathbf{X}}(\theta)\delta\theta}. \quad (3.2)$$

Here, $\mathbf{F}(\cdot)$ is the Fisher information matrix with index denoting the output or the input side:

$$\mathbf{F}_{\mathbf{X}}(\theta) = \mathcal{E}[(\nabla_{\theta} \log p(X|\theta))(\nabla_{\theta} \log p(X|\theta))'].$$

(In the last expression, $\nabla_{\boldsymbol{\theta}}$ denotes the gradient, $(\cdot)'$ denotes the transpose and $\mathcal{E}[\cdot]$ is the expected value with respect to $p(x|\boldsymbol{\theta})$.) With the reasonable assumption that $d(\mathbf{X}(\boldsymbol{\theta}_0), \mathbf{X}(\boldsymbol{\theta}_0)) = 0$ (which is true for Kullback-Leibler distance, J -divergence, Chernoff distance, Hellinger distance and corresponds to the intuition that the distance between same distributions is zero), it follows that $f(c(1)) = 0$ and expression (3.2) simplifies to

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

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. A subtlety emerges here in the multiparameter case: the value of γ depends, in general, on the direction of the perturbation $\delta\boldsymbol{\theta}$. This is not surprising since it tells us that the information transfer ratio of each parameter θ_i in the parameter vector $\boldsymbol{\theta}$ is different in general. For example, there is no reason to believe that a system would (in general) preserve the information about the amplitude and the phase with the same fidelity. In the special case where we perturb only component θ_i and keep others fixed, we have:

$$\gamma_{\mathbf{X}, \mathbf{Y}}(\boldsymbol{\theta}, \boldsymbol{\theta} + [0, 0, \dots, \delta\theta_i, \dots, 0]') = \frac{[\mathbf{F}_{\mathbf{Y}}(\boldsymbol{\theta})]_{ii}}{[\mathbf{F}_{\mathbf{X}}(\boldsymbol{\theta})]_{ii}}$$

Index ii denotes the i th element on the diagonal of the Fisher information matrix. Finally, examples show that when $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_0$ differ substantially, the ratio's value *does* depend on distance measure choice. For example, symmetric and non-symmetric measures must yield different ratios.

3.3 A System Theory for Information Processing

We have several results that use the information transfer ratio to quantify how collections of systems process information. (See Appendix B for the proofs.)

Figure 3.1: Two systems in cascade: the overall information transfer ratio is the product of the component ratios

3.3.1 Cascaded Systems

If two systems are in cascade, as in Figure 3.1, 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, then

$$\gamma_{\mathbf{X},\mathbf{Z}}(\theta_0, \theta_1) = \gamma_{\mathbf{X},\mathbf{Y}}(\theta_0, \theta_1) \cdot \gamma_{\mathbf{Y},\mathbf{Z}}(\theta_0, \theta_1).$$

Notice that this result holds for general distance measures (not necessarily the KL distance). This result trivially generalizes for any number of the systems in cascade. For example, an analog signal, which is first digitized, then compressed and finally transmitted over the wireless channel, will lose certain amount of its information content at each stage. Since the component ratios are numbers between zero and one, the overall ratio quickly approaches zero when the number of cascaded lossy systems is large.

3.3.2 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 (see Figure 3.2)

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

Figure 3.2: System with multiple inputs and one output

This result does not hold for general distance measures since it is derived by using the additivity of the Kullback-Leibler distance (see Appendix B). Note that each $\gamma_{X_i, Y}$ can exceed one because each (X_i, Y) does not necessarily form a Markov chain. For example, for N independent and identically distributed Gaussian variables at the input of the system whose output is the sum of the inputs, every $\gamma_{X_i, Y} = N$, while $\gamma_{\mathbf{X}, Y} = 1$. (The fact that overall information transfer ratio is equal to one is not surprising if one remembers that the sum of the observations is the sufficient statistic for the family of Gaussian distributions parameterized by the mean.) Also notice that

$$\gamma_{\mathbf{X}, Y}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) \leq \min_i [\gamma_{X_i, Y}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)].$$

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 every input. Also notice that a change in only one input results, in general, in a change at the output which in turn results in changing *all* of the individual information transfer ratios. This fact is not strange if one remembers that the individual inputs do not form a Markov chain with the output.

Figure 3.3: System with one input and two outputs

3.3.3 Multiple Outputs

Let a system have one input and two outputs as in Figure 3.3: $\mathbf{Y} = [Y_1, Y_2]$. 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, in the second term, the distance between the output's second component conditioned on the first is used in the numerator. Thus, the conditional distribution change is what increases information processing capability.

Generalizing the previous example, consider N outputs that are *conditionally* independent given the input (N parallel systems is one example). We want to explore how the information transfer ratio changes as more outputs are added. We can calculate the information transfer ratio for two special cases (see Appendix B.4 for calculations). In the first, the input is exponentially distributed and each output has conditional mass function given as $p_{Y|X}(n) = x^n e^{-x} / n!$ (Poisson distribution is often used to model a phenomenon in which we are waiting for an occurrence such as a spike in a neuron.) Then,

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

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 asymptotic differential increase in γ is proportional to $1/N^2$. When we consider a second system

with a Gaussian input ($X \sim N(m, \sigma_x^2)$), and the output repeats the input with first-order Gauss-Markov noise added ($\mathbf{Y} = X \cdot \mathbf{1} + \mathbf{Z}$ where $\mathbf{1}$ is a vector with N elements which are all equal to 1, $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{K}_N)$ and $\mathbf{K}_N = 1/(1 - a^2)[K_{i,j}]$, $K_{i,j} = a^{|i-j|}$, $|a| < 1$). Information transfer ratio for two inputs which differ in mean, after some linear-algebraic manipulations, is shown to be

$$\gamma_{X,\mathbf{Y}}(m_0, m_1) = \frac{[(N-2)(a-1)^2 - 2(a-1)]\sigma_x^2}{1 + [(N-2)(a-1)^2 - 2(a-1)]\sigma_x^2},$$

and simplifies considerably in the case of white Gaussian noise ($a \rightarrow 0$)

$$\gamma_{X,\mathbf{Y}}(m_0, m_1) = \frac{N\sigma_x^2}{1 + N\sigma_x^2}.$$

Again, the differential gain asymptotically is $1/N^2$ and when $N \rightarrow \infty$, then it follows that $\gamma_{X,\mathbf{Y}}(m_0, m_1) \rightarrow 1$. These results indicate that the addition of more statistically identical outputs quickly reaches a point of diminishing returns.

3.4 Systems that re-represent information

Interesting results are obtained when we consider simple examples in which the system transforms information from one representation to another (from an analog waveform to a point process, for example). (See Appendix C for the calculations.)

3.4.1 Continuous Input, Discrete Output

In the case when the input distribution is exponential $p_X(x) = \theta e^{-\theta x}$ and the system's conditional output is Poisson $p_{Y|X}(n|x) = x^n e^{-x}/n!$, the information transfer ratio calculated using Kullback-Leibler distances is

$$\gamma_{\mathbf{X},\mathbf{Y}}(\theta_0, \theta_1) = \frac{\ln \frac{\theta_1}{\theta_0} + \frac{\theta_1 + 1}{\theta_1} \ln \frac{1 + \theta_0}{1 + \theta_1}}{\ln \frac{\theta_1}{\theta_0} + \frac{\theta_0 - \theta_1}{\theta_1}}.$$

Figure 3.4: Information transfer ratio for a system that converts a real-valued exponential random variable into a discrete-valued one — a conditionally Poisson output given the input. The white area along the equality line corresponds to the case where the ratio is indeterminant and equals $0/0$. The white line (where $\theta_1 \rightarrow \theta_0$) decays as $1/(1 + \theta_0)$.

Notice that it is not possible to express this KL ratio as a function of θ_1/θ_0 . Figure 3.4 shows a plot of $\gamma_{\mathbf{X},\mathbf{Y}}(\theta_0, \theta_1)$. Notice the non-symmetric shape of the information transfer ratio. Also notice that the ratio is the largest when both θ_0 and θ_1 are close to zero. This example demonstrates that only when the input parameter is very small is efficient information transfer obtained. Note how the ratio easily deals with the case where input distribution is continuous while the output one is discrete.

In the next example, input distribution is Gaussian — $X \sim N(m, \sigma^2)$ — and the system's conditional distribution is Poisson: $p_{Y|X}(n|x) = x^n e^{-x}/n!$.¹ A closed-form expression for the unconditional output distribution cannot be obtained. However, a series expansion can be found, and we calculated the Kullback-Leibler distance

¹Only the cases with positive m and with high SNR will be considered, because negative rate for the conditionally Poisson output does not make sense.

Figure 3.5: The left panel shows the output distribution for the Gaussian-input, Poisson-output example. The right panel shows how the information transfer ratio varies with mean (considering it as θ_1 varying relative to θ_0 , a mean value of 10.)

between outputs using symbolic manipulation software (Mathcad). In Figure 3.5, output probability mass function is shown, with input mean $m = 15$ and variance $\sigma^2 = 1$. In figure 3.5, the information transfer ratio is shown where we varied the mean of the input signal. The fixed distribution at the input has the mean 10 which is the reason for the dip artifact at $m = 10$, and corresponding output distribution is used as the fixed distribution in KL distance calculation at the output. Notice that *increase* in the input mean, which corresponds to an *increased* Kullback-Leibler distance for the input, corresponds to the *decrease* in the information transfer ratio.

3.4.2 Non-symmetric Binary Channel

We now explore dependence of information transfer ratio on the input a priori probabilities. In particular, we searched for the input parameter p (see Figure 3.6) whose differential change will yield largest information transfer ratio. The capacity-achieving probability is $p = \frac{2^{H(\epsilon)}}{2^{H(\epsilon)} + 2^{H(\delta)}}$, where $H(\cdot)$ denotes entropy of a binary-valued random variable (in bits):

$$H(\epsilon) = -\epsilon \log_2(\epsilon) - (1 - \epsilon) \log_2(1 - \epsilon).$$

Figure 3.6: Non-symmetric binary channel

The resulting capacity is $C = \log_2(2^{-H(\epsilon)} + 2^{-H(\delta)})$. To use our approach, let θ_0 correspond to p and $\theta_1 = p + \Delta p$ (where Δp is very small). Information transfer ratio in this case can easily be calculated from its definition by using Taylor series expansion for log function or by applying the method of section 3.2. Then,

$$\gamma_{\mathbf{X},\mathbf{Y}}(p, p + \Delta p) = (1 - \epsilon - \delta)^2 \frac{p(1-p)}{[p(1-\epsilon-\delta) + \epsilon][1-p(1-\epsilon-\delta) - \epsilon]}.$$

Notice that γ is independent of Δp . The maximum of $\gamma_{\mathbf{X},\mathbf{Y}}(p + \Delta p, p)$ with respect to p occurs, in general, at value of input probability which is *different* from the value of p which determines capacity. For $\epsilon \neq \delta$, we have

$$p = \frac{1}{\epsilon - \epsilon^2 - \delta + \delta^2} \left(\epsilon - \epsilon^2 - \sqrt{(\epsilon - \epsilon^2)(\delta - \delta^2)} \right)$$

and the maximum information transfer equals

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

where $V = \sqrt{\epsilon(\epsilon - 1)\delta(\delta - 1)}$. In the case of binary symmetric channel (BSC) ($\epsilon = \delta$), $p = 1/2$ for both capacity and maximum γ . Here, the maximum information transfer is

$$\gamma_{\mathbf{X},\mathbf{Y}} = (1 - 2\epsilon)^2.$$

Figure 3.7 shows examples of how $\gamma_{\mathbf{X},\mathbf{Y}}$ varies with p .

Figure 3.7: Binary symmetric channel case ($\epsilon = \delta = 0.1$) and two non-symmetric cases: asym1 has $\epsilon = 0.15$ and $\delta = 0.05$, while asym2 has $\epsilon = 0.02$ and $\delta = 0.2$. Figure 3.7 shows all three γ vs. p plots.

3.4.3 Information Transfer Ratio of a Low Pass Filter

Consider the following system: sinusoid mixed with white Gaussian noise (WGN) passes through the low pass filter and white Gaussian noise is added after the filter. After passing through the low pass filter (LPF) — with transfer function $H(z) = \frac{1-a}{1-az^{-1}}$ — the noise becomes colored with correlation $R_{Y'}(n) = (1-a)/(1+a)\sigma^2 a^{|n|}$. The correlation matrix of the input noise *after* it passed through the filter is given by $\mathbf{K}' = [R_{Y'}(i-j)]_{i,j}$, $i, j = 0, 1, \dots, N-1$. On the other hand, N samples of the sinusoid $A \sin(\omega n)$, $n = 0, 1, \dots, N-1$, $\omega = 2\pi/N$, after passing through the LPF, become:

$\mathbf{V} = [V_n]_{N \times 1}$, where

$$V_n = A(1-a)/(1-2a \cos(\omega) + a^2)(a^{n+1} \sin(\omega) - a \sin(n+1)\omega + \sin(n\omega)).$$

Thus, the output of the LPF is distributed normally as $N(\mathbf{V}, \mathbf{K}')$. After adding white Gaussian noise with variance b , we obtain $\mathbf{K}$ ($\mathbf{K}' + b\mathbf{I} = \mathbf{K}$). Finally, information

Figure 3.8: Both plots are obtained with the same set of parameters but the figure on the right-hand side is plotted on log-log scale. Notice the sharp drop in $\gamma_{X,Y}$ as b increases initially; the drop is due to the fact that adding the noise to the filtered input signal greatly reduces the ability to extract the information from the output.

transfer ratio can be computed:

$$\gamma_{X,Y}(A_0, A_1) = \frac{\|\mathbf{V}^{(1)} - \mathbf{V}^{(0)}\|_{\mathbf{K}^{-1}}^2}{\|\mathbf{U}^{(1)} - \mathbf{U}^{(0)}\|_{1/\sigma^2 \mathbf{I}}^2},$$

where $\mathbf{U}^{(i)}$'s are mean vectors of the input distribution ($i = 0, 1$) which correspond to two different amplitudes of the input sine and $\|\cdot\|_{\mathbf{M}}^2$ denotes the quadratic form with matrix $\mathbf{M}$ as the kernel. For the case where $N = 5$, $a = 0.5$, the plot for $\gamma_{X,Y}(s)$ vs. b_s (s indexes values of b), the variance of the WGN added after the LPF, is in figure 3.8. In the case that no noise is added ($b = 0$), γ would be one which means that no information filtering is performed. However, as the variance of the noise added after the filtering increases, the information transfer ratio decays rapidly which implies that the ability to extract the amplitude information also decreases rapidly.

Chapter 4

Conclusions

The approach of using information change to probe a system's processing represents an important concept. Without change, more static measures such as mutual information could be used. However, such a measure considers input and output as indivisible quantities, and does not directly measure what is information-bearing and what is not. For us, the information source must reveal itself by changing information content. We then employ distance measures that reflect the data processing theorem and apply them separately to the input and output. In this way, assessing systems that process and/or represent information can be analyzed in a coherent framework. We have represented information here as a parameter vector, which may be numeric- or symbolic-valued. This assumption is not general; information can have many more forms. We justify it here as an initial assumption to initiate a theory of information processing.

The non-symmetric binary channel example indicates that the information 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 non-symmetric binary channel example shows that maximizing information transfer does not occur at the same input distribution that yields capacity. The probability distribution that maximizes

information transfer ratio describes the best conveyor of information. 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.

One can envisage an information processing system as an *information filter*, where certain parameter changes about some operating point have larger gains (information transfer ratios) than others. 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 examples suggest that the systems that re-represent information, by extracting information content from their inputs and re-encoding it in their outputs, may not achieve the data processing theorem's upper bound. In one study of neural systems [7], for example, very small transfer ratios were measured ($\approx 10^{-3}$), but analysis showed that ratios could not be much larger for an ideal system performing a similar information transformation. In another recent study, it has been found that information coding in neural systems is dynamic (the representation of information changes with time but not the fidelity) and that neurons can multiplex information with differing degrees of fidelity [5].

We have seen in numerous examples that the information transfer ratio behaves well and that it gives meaningful answers. Many results do not strongly depend on which distance measure is used to compute the ratio. Especially critical here is that the distance measure be computed from probability distributions. For example, all Ali-Silvey distances have this property. By relying on probability distribution, 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. Moreover, we have shown (see Appendix A.2) that when Gaussian processes are sampled, the information transfer ratio — based on Kullback-Leibler distance before and after sampling — is asymptotically equal to one (as the sampling rate becomes large). No matter which distance measure is used, our system theory shows that information can be easily degraded as a succession of systems process it and the representation changes from one form to another.

Appendix A

Significance of the Kullback-Leibler Distance

A.1 Some Properties of the Kullback-Leibler Distance

While the Kullback-Leibler (KL) distance is not a true distance, since it is non-symmetric and does not satisfy triangle inequality, it is useful to think of it as “distance” between distributions. KL distance is primarily important because of its relations to estimation theory (via Fisher information matrix) and detection theory (through Stein’s lemma). It also possesses additive property which is desirable from both theoretical and practical points of view.

1. Kullback-Leibler distance is always non-negative:

$$\mathcal{D}(p_1||p_0) \geq 0.$$

It is equal to zero if and only if $p_1 = p_0$ in discrete case (or $p_1 = p_0$ almost everywhere in continuous case).

2. If two joint densities $\mathbf{p}_0$ and $\mathbf{p}_1$ consist of L independent and identically distributed variables, then the Kullback-Leibler distance between these two distributions is simply:

$$\mathcal{D}(\mathbf{p}_1||\mathbf{p}_0) = L\mathcal{D}(p_1||p_0),$$

where p_1 and p_0 are the marginal densities. This *additive* property of the Kullback-Leibler distance holds also for statistically independent random variables which are not necessarily identically distributed: the distance between

joint distributions is equal to the sum of the distances between marginal distributions. Specifically, if $p_j(\mathbf{x}) = \prod_i p(x_i)$, $j = 0, 1$, then

$$\mathcal{D}(p_1(\mathbf{x})||p_0(\mathbf{x})) = \sum_i \mathcal{D}(p_1(x_i)||p_0(x_i))$$

3. Mutual information is just a special case of KL distance:

$$I(X; Y) = \mathcal{D}(p_{x,y}||p_x p_y)$$

where $p_{x,y}$ is the joint probability density function and p_x, p_y are the marginal densities.

4. Through Stein's Lemma [4, 11], the Kullback-Leibler distance represents the exponential rate of optimal classifier performance probability. If $\mathbf{X}$ is a random vector having N statistically independent and identically distributed components,

$$\lim_{N \rightarrow \infty} \frac{\log P_F}{N} = -\mathcal{D}(p_1||p_0) \quad \text{for fixed } P_M,$$

where P_F and P_M are false-alarm and miss probabilities of the optimal Neyman-Pearson classifier. Roughly speaking, false-alarm probability asymptotically decays exponentially in the amount of data available to the optimal classifier and the rate of decay is equal to the Kullback-Leibler distance. Therefore, the Kullback-Leibler distance is related to detection theory via Stein's lemma.

5. For a small change in the parameter $\boldsymbol{\theta}$ which governs the probability density $p_{\boldsymbol{\theta}}$ of $X(\boldsymbol{\theta})$, the Kullback-Leibler distance is equal to:

$$\mathcal{D}(p_{\boldsymbol{\theta}}||p_{\boldsymbol{\theta}+\delta\boldsymbol{\theta}}) = \mathcal{D}(p_{\boldsymbol{\theta}+\delta\boldsymbol{\theta}}||p_{\boldsymbol{\theta}}) \approx \frac{1}{2}\delta\boldsymbol{\theta}'\mathbf{F}(\boldsymbol{\theta})\delta\boldsymbol{\theta},$$

where $\mathbf{F}(\cdot)$ is Fisher information matrix:

$$\mathbf{F}(\boldsymbol{\theta}) = \mathcal{E}[(\nabla_{\boldsymbol{\theta}} \log p(X|\boldsymbol{\theta}))(\nabla_{\boldsymbol{\theta}} \log p(X|\boldsymbol{\theta}))'].$$

In the last expression, ∇_{θ} denotes the gradient, $(\cdot)'$ denotes the transpose and $\mathcal{E}[\cdot]$ is the expected value with respect to $p(x|\theta)$. On the other hand, according to the Cramér-Rao bound, the best unbiased estimator $\hat{\theta}$ of θ —from observations $\mathbf{x}$ that belong to $p_{\theta}(\mathbf{x})$ —is lower-bounded by

$$\mathcal{E}[(\theta_i - \hat{\theta}_i)^2] \geq [\mathbf{F}^{-1}]_{ii}.$$

Thus, the Kullback-Leibler distance is related to the estimation theory via the Fisher information matrix that appears in the Cramér-Rao bound.

6. Kullback-Leibler distance also appears as a penalty in the length of the code which was designed using some wrong distribution q instead of the true one, p . While the expected length of the Huffman code is lower-bounded by the entropy of the source for the correct source distribution, expected length of the Huffman code for the wrong distribution is lower-bounded by the entropy of the source *plus* the Kullback-Leibler distance between the true and wrong distributions, $\mathcal{D}(p||q)$.

A.2 Kullback-Leibler distance between two Gaussian random processes

Consider two Gaussian random processes, $X^{(0)}(t)$ and $X^{(1)}(t)$ with two different functions of mean, $m_0(t)$ and $m_1(t)$ and the same correlation structure given by $K(\tau) = \sigma^2 e^{-a|\tau|}$. To calculate the KL distance between $X^{(0)}(t)$ and $X^{(1)}(t)$, we used the Karhunen-Loève expansion to represent processes in terms of *uncorrelated* random variables, $X_i^{(j)}$, $j = 0, 1$. Since these variables are also 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 two corresponding uncorrelated

components, $X_i^{(1)}$ and $X_i^{(0)}$:

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

The right-hand side becomes

$$\frac{1}{2} \sum_i \frac{|m_i^{(1)} - m_i^{(0)}|^2}{\lambda_i} = \frac{1}{2} \|m^{(1)}(t) - m^{(0)}(t)\|_Q^2$$

by using the result for the KL distance between two Gaussian densities with different means and the same variance. Here, λ_i 's are eigenvalues of $K(\cdot)$ and $Q(t, v)$ satisfies

$$\int_0^T K(t, u) Q(t, v) dt = \delta(u - v)$$

where $K(u, v) = K(u - v) = K(\tau)$. For the special case $K(\tau) = \sigma^2 e^{-a|\tau|}$

$$\mathcal{D}(X^{(1)}(t) \| X^{(0)}(t)) = \frac{1}{4\sigma^2} [a \int_0^T r^2(u) du + \frac{1}{a} \int_0^T [r'(u)]^2 du + r^2(0) + r^2(T)] \quad (\text{A.1})$$

where $r(t) = m_1(t) - m_0(t)$, and $r'(t)$ is the derivative of $r(t)$ [12].

We will now consider the KL distance between sampled version of the two processes, vectors $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(0)}$. The correlation matrix $\mathbf{K}$ is given by $K_{i,j} = \sigma^2 e^{-a|(i-j)\Delta|}$, where Δ is the sampling interval. In this case, KL distance is given as

$$\mathcal{D}(\mathbf{X}^{(1)} \| \mathbf{X}^{(0)}) = \frac{\mathbf{R}' \mathbf{K}^{-1} \mathbf{R}}{2},$$

where $\mathbf{R}$ denotes the vector of samples from $r(t)$, $(\cdot)'$ denotes transpose and $\mathbf{K}^{-1}$ denotes the inverse of $\mathbf{K}$. Also, $n = T/\Delta$ and the matrix $\mathbf{K}$ is n by n . The last expression becomes

$$\frac{1}{2\sigma^2(1 - e^{-2a\Delta})} \left[\sum_{i=1}^{n-1} (r_i - e^{-a\Delta} r_{i+1})^2 - (e^{-2a\Delta} - 1) r_n^2 \right]$$

If we let $\Delta \rightarrow 0$, then the last expression becomes the same as (A.1). Thus, the KL distance between sampled versions of two Gaussian processes with different means and the same (Gauss-Markov) correlation structure approaches the KL distance between two corresponding Gaussian random processes as the sampling rate increases.

Appendix B

Proofs for Section 3.3

B.1 Cascaded Systems

If $\boldsymbol{\theta} \rightarrow \mathbf{X} \rightarrow \mathbf{Y} \rightarrow \mathbf{Z}$ form a Markov chain, then

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

and

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

Then,

$$\begin{aligned} \gamma_{\mathbf{X},\mathbf{Y}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) \cdot \gamma_{\mathbf{Y},\mathbf{Z}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) &= \frac{d(\mathbf{Z}(\boldsymbol{\theta}_0), \mathbf{Z}(\boldsymbol{\theta}_1))}{d(\mathbf{X}(\boldsymbol{\theta}_0), \mathbf{X}(\boldsymbol{\theta}_1))} \\ &= \gamma_{\mathbf{X},\mathbf{Z}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1). \end{aligned}$$

Thus, $\gamma_{\mathbf{X},\mathbf{Z}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) = \gamma_{\mathbf{X},\mathbf{Y}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) \cdot \gamma_{\mathbf{Y},\mathbf{Z}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$.

B.2 Multiple Inputs

When the input vector $\mathbf{X}$ consists of several statistically independent components X_i , the following set of equalities hold:

$$\begin{aligned} \frac{1}{\gamma_{\mathbf{X},\mathbf{Y}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)} &= \frac{d(\mathbf{X}(\boldsymbol{\theta}_0), \mathbf{X}(\boldsymbol{\theta}_1))}{d(\mathbf{Y}(\boldsymbol{\theta}_0), \mathbf{Y}(\boldsymbol{\theta}_1))} \\ &= \frac{\sum_i d(X_i(\boldsymbol{\theta}_0), X_i(\boldsymbol{\theta}_1))}{d(\mathbf{Y}(\boldsymbol{\theta}_0), \mathbf{Y}(\boldsymbol{\theta}_1))} \\ &= \sum_i \frac{1}{\gamma_{X_i,\mathbf{Y}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)}. \end{aligned} \tag{B.1}$$

The transformation in (B.1) is possible because the Kullback-Leibler distance between the joint distributions of statistically independent random variables equals the sum of the marginal distances (see Appendix A.1).

B.3 Multiple Outputs

The chain rule for Kullback-Leibler distance states that

$$d(\mathbf{Y}(\boldsymbol{\theta}_0), \mathbf{Y}(\boldsymbol{\theta}_1)) = d(Y_1(\boldsymbol{\theta}_0), Y_1(\boldsymbol{\theta}_1)) + d(Y_2|Y_1(\boldsymbol{\theta}_0), Y_2|Y_1(\boldsymbol{\theta}_1)).$$

By dividing both sides of the expression above by $d(\mathbf{X}(\boldsymbol{\theta}_0), \mathbf{X}(\boldsymbol{\theta}_1))$ we obtain

$$\gamma_{\mathbf{X},[Y_1,Y_2]}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) = \gamma_{\mathbf{X},Y_1}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) + \gamma_{\mathbf{X},Y_2|Y_1}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1).$$

Obviously, one can consider Y_1 conditioned on Y_2 on the right-hand side and equality would still hold (i.e. one can change the ordering of component outputs in the vector output). Furthermore, by considering Y_1 as a vector, the formula for N outputs can be easily derived:

$$\gamma_{\mathbf{X},[Y_1,\dots,Y_N]}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) = \gamma_{\mathbf{X},Y_1}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) + \sum_{i=2}^N \gamma_{\mathbf{X},[Y_i|Y_1,\dots,Y_{i-1}]}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1).$$

B.4 Information Transfer Ratio in the Case of N Outputs

For the exponentially distributed input and N outputs each of which is Poisson distributed (subsection 3.3.3), the proof is very similar to the one in Appendix C where the case of $N = 1$ is considered. In the second multiple output example, we have a system with a Gaussian input ($X \sim N(m, \sigma_x^2)$), and the output repeats the input with first-order Gauss-Markov noise added ($\mathbf{Y} = X \cdot \mathbf{1} + \mathbf{Z}$ where $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{K}_N)$ and $\mathbf{K}_N = 1/(1 - a^2)[K_{i,j}]$; $i, j = 1, \dots, N$; $K_{i,j} = a^{|i-j|}$; $|a| < 1$). Then, the output correlation matrix is

$$\mathbf{K}_Y = \mathbf{1}\sigma_x^2\mathbf{1}' + \mathbf{K}_N.$$

We will observe information transfer ratio for two inputs which differ in mean. Since both the input and the output are governed by Gaussian distributions, the Kullback-Leibler distances are

$$\text{at the input: } \mathcal{D}(X_{\delta m+m}||X_m) = \frac{(\delta m)^2}{2\sigma_x^2}$$

$$\text{at the output: } \mathcal{D}(Y_{\delta m+m}||Y_m) = \frac{(\delta m)^2}{2} \mathbf{1}' \mathbf{K}_Y^{-1} \mathbf{1}.$$

Hence, the information transfer ratio is

$$\gamma_{X,Y}(m, \delta m + m) = \sigma_x^2 \mathbf{1}' \mathbf{K}_Y^{-1} \mathbf{1}.$$

We now have to invert $\mathbf{K}_Y$:

$$\mathbf{K}_Y^{-1} = \mathbf{K}_N^{-1} \left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{1/\sigma_x^2 + \mathbf{1}' \mathbf{K}_N^{-1} \mathbf{1}} \mathbf{K}_N^{-1} \right).$$

If we let $\alpha = \mathbf{1}' \mathbf{K}_N^{-1} \mathbf{1}$ then

$$\begin{aligned} \gamma_{X,Y}(m, \delta m + m) &= \alpha \sigma_x^2 - \frac{\alpha^2 \sigma_x^4}{1 + \alpha \sigma_x^2} \\ &= \frac{\alpha \sigma_x^2}{1 + \alpha \sigma_x^2} \end{aligned} \tag{B.2}$$

Since we have

$$\mathbf{K}_N^{-1} = \begin{bmatrix} 1 & -a & 0 & 0 & 0 & \dots & 0 \\ -a & a^2 + 1 & -a & 0 & 0 & \dots & 0 \\ 0 & -a & a^2 + 1 & a & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \dots & 0 & -a & 1 \end{bmatrix}$$

it follows that α is equal to the sum of the elements of the matrix $\mathbf{K}_N^{-1}$:

$$\alpha = (N - 2)(a - 1)^2 - 2(a - 1)$$

By replacing the last expression in (B.2), we obtain the expression for information transfer ratio in this case.

Appendix C

Proofs for Section 3.4

C.1 Continuous Input, Discrete Output

First, we consider the case where the input distribution is exponential $p_X(x) = \theta e^{-\theta x}$ and the system's conditional p.m.f. is Poisson $p_{Y|X}(n|x) = x^n e^{-x} / n!$. We need to find output p.m.f. It is much easier to calculate it in the frequency domain by applying the z -transform to the following identity:

$$p_Y(n) = \int p_{Y|X}(n|x) p_X(x) dx.$$

The z -transform of the system's conditional p.m.f. is easily found to be $e^{x(z-1)}$ and it follows that the z -transform of the output p.m.f. is $\theta/(\theta + 1 - z)$. Thus,

$$p_Y(n) = \frac{\theta}{1 + \theta} \left(\frac{1}{1 + \theta} \right)^n.$$

It is not hard to show that

$$d(X(\theta_0), X(\theta_1)) = \ln(\theta_1/\theta_0) + (\theta_0 - \theta_1)/\theta_1,$$

and

$$d(Y(\theta_0), Y(\theta_1)) = \ln \frac{\theta_1}{\theta_0} + \frac{\theta_1 + 1}{\theta_1} \ln \frac{1 + \theta_0}{1 + \theta_1}.$$

Finally, information transfer ratio is:

$$\gamma_{X,Y}(\theta_0, \theta_1) = \frac{\ln \frac{\theta_1}{\theta_0} + \frac{\theta_1 + 1}{\theta_1} \ln \frac{1 + \theta_0}{1 + \theta_1}}{\ln \frac{\theta_1}{\theta_0} + \frac{\theta_0 - \theta_1}{\theta_1}}.$$

Another example with continuous input and discrete output has Gaussian distribution at the input and system's conditional output is Poisson-distributed. Again, the trick of applying the z -transform to the identity

$$p_Y(n) = \int p_{Y|X}(n|x)p_X(x) dx$$

is used to simplify calculations. Output distribution is obtained by observing the coefficients of z^n (a_n 's) which correspond to $p_Y(n)$. Finally,

$$p_Y(n) = \exp\left(\frac{\sigma^2}{2} - m\right) \frac{1}{n!} \int_0^\infty x^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - (m - \sigma^2))^2}{2\sigma^2}\right) dx.$$

Even though a closed-form expression for the unconditional output distribution cannot be obtained, symbolic manipulation software is used to calculate Kullback-Leibler distance between outputs.

Appendix D

Proof of Local Invariance of γ with Respect to Distance Measure

We consider information transfer ratio for Ali-Silvey class of information theoretic distance measures. Here is the definition of Ali-Silvey distance family: $d(p_0, p_1) = f(\mathcal{E}_0[c(\Lambda(R))])$, where f is monotonically non-decreasing, $\mathcal{E}_0(\cdot)$ is expected value with respect to p_0 , $c(\cdot)$ is convex function and $\Lambda(\cdot)$ is the likelihood ratio: $p_1(\cdot)/p_0(\cdot)$.

We will consider ratio $\Lambda = p(x|\boldsymbol{\theta} + \delta\boldsymbol{\theta})/p(x|\boldsymbol{\theta})$, where $\delta\boldsymbol{\theta}$ represents differential change in input vector parameter $\boldsymbol{\theta}$. Then,

$$\Lambda = [p(x|\boldsymbol{\theta}) + \delta\boldsymbol{\theta}'\nabla_{\boldsymbol{\theta}}p(x|\boldsymbol{\theta})]/p(x|\boldsymbol{\theta}) + \text{higher order terms.}$$

Thus, $c(\Lambda)$ becomes

$$c(1 + \frac{\delta\boldsymbol{\theta}'\nabla_{\boldsymbol{\theta}}p(x|\boldsymbol{\theta})}{p(x|\boldsymbol{\theta})}) \approx c(1) + \frac{\delta\boldsymbol{\theta}'\nabla_{\boldsymbol{\theta}}p(x|\boldsymbol{\theta})}{p(x|\boldsymbol{\theta})}c'(1) + \frac{\delta\boldsymbol{\theta}'\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}p(x|\boldsymbol{\theta})\delta\boldsymbol{\theta}}{p(x|\boldsymbol{\theta})^2}\frac{c''(1)}{2}$$

After taking expected value with respect to $p(x|\boldsymbol{\theta})$,

$$\begin{aligned} d(p_0, p_1) &\approx f\left(c(1) + \frac{c''(1)}{2} \int \frac{\delta\boldsymbol{\theta}'\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}p(x|\boldsymbol{\theta})\delta\boldsymbol{\theta}}{p(x|\boldsymbol{\theta})} dx\right) \\ &= f\left(c(1) + \frac{c''(1)}{2} \delta\boldsymbol{\theta}'\mathbf{F}(\boldsymbol{\theta})\delta\boldsymbol{\theta}\right), \end{aligned}$$

where $\mathbf{F}(\cdot)$ is Fisher information matrix:

$$\mathbf{F}(\boldsymbol{\theta}) = \mathcal{E}[(\nabla_{\boldsymbol{\theta}} \log p(X|\boldsymbol{\theta}))(\nabla_{\boldsymbol{\theta}} \log p(X|\boldsymbol{\theta}))'].$$

Finally,

$$d(p_0, p_1) \approx f(c(1)) + \frac{1}{2}f'(c(1))c''(1)\delta\boldsymbol{\theta}'\mathbf{F}(\boldsymbol{\theta})\delta\boldsymbol{\theta}.$$

For the Kullback-Leibler distance $\mathcal{D}(p_{\boldsymbol{\theta}_1} \| p_{\boldsymbol{\theta}_0})$, $f(x) = x$ and $c(x) = x \log(x)$. (For the Kullback-Leibler distance $\mathcal{D}(p_{\boldsymbol{\theta}_0} \| p_{\boldsymbol{\theta}_1})$, $f(x) = x$ and $c(x) = -\log(x)$.) Even though Kullback-Leibler distance is asymmetric in general, for small changes in the input parameter, both distances take the same value. Thus, $d(p_0, p_1) \approx 1/2 \delta \boldsymbol{\theta}' \mathbf{F}(\boldsymbol{\theta}) \delta \boldsymbol{\theta}$ in Kullback-Leibler case. For the Ali-Silvey class of distances, information transfer ratio, for perturbational (small) difference between two input distributions, is:

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

This is the result we use in section 3.2, equation 3.2.

Bibliography

- [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] I. Csiszar and J. Korner, *Information Theory, Coding Theorems for Discrete Memoryless Systems*, Akademiai Kiado, Budapest, Hungary, 1986.
- [4] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, John Wiley and Sons, Inc., 1991.
- [5] C.M. Gruner, *Quantifying information coding limits in sensory systems*, Ph.D. thesis, Rice University, Houston, TX, July 1998.
- [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] D.H. Johnson, C.M. Gruner, R.M. Glantz, "Quantifying information transfer in spike generation," *Computational Neuroscience '99*, Pittsburgh, PA, July 18-25, 1999.
- [8] D.H. Johnson and G.C. Orsak, "Relation of signal set choice to the performance

of optimal non-Gaussian detectors,” *IEEE Trans. Commun.*, vol. 41, pp. 1319-1328, 1993.

- [9] S. Kullback, *Information Theory and Statistics*, Dover Publications, NY, 1967.
- [10] C. E. Shannon, “A mathematical theory of communication,” *Bell Sys. Tech. Journal*, 27: 379-423, 623-656, 1948.
- [11] C. Stein, “Information and comparison of experiments,” unpublished.
- [12] H. L. Van Trees, *Detection, Estimation, and Modulation Theory*, John Wiley and Sons, Inc., NY, 1968.