

Information theory and signal processing in neuroscience

Alexander Dimitrov
Center for Computational Biology
Montana State University
January, 2003

This is a rather brief outline of the topic of information theory and its applications to neuroscience. If you find this interesting, there are two main books to delve deeper:

- Elements of Information Theory. Cover and Thomas '91
A rather complete description of basic information theory.
- Entropy and Information Theory. Robert Gray '90
Encompasses less, but more technical and precise.

There is also a slew of papers on applications, which will be listed separately.

Signal processing and information theory in neuroscience

There are two principal uses of information theory in neuroscience:

- As an explanatory theory for neural function.

In this setting one usually advances the hypothesis that neural systems are optimal under certain information-theoretic measures. This hypothesis is either tested, or its validity is assumed and neural structures are predicted based on that.

- As a method for analysis of physiological data.

Here a neural system is analyzed as a communication channel. One usually estimates certain information-theoretic quantities in different operating modes of the system, and then draws conclusions about the system. A particular problem there is the question of neural coding: how is the activity of a set of neurons representing their inputs?

Example 1 (Neural Function). *Barlow (1967) proposed, and Atick (1992) studied the assumption that early sensory processing is concerned with redundancy reduction. The system under consideration is $X \rightarrow Y$, where X is a sensory stimulus, and Y – neural response. Atick defined the redundancy of the system as*

$$\mathcal{R} = 1 - \frac{I(X; Y)}{C(X)}.$$

and optimized the conditional probability $p(y|x) \equiv L \in \mathcal{L}(X, Y)$ for a fixed stimulus ensemble $p(x)$. The so obtained optimal L^ was compared to observed neural structure, and system performance.*

Example 2 (Neural Coding). Consider again the problem $X \rightarrow Y$, where X is the input (not necessarily sensory), and Y the output of a neural system. The field of neural coding attempts to answer the question

How does the input X produce the observed system response Y ?

or, equivalently

What is the observed system response Y telling us about the input X ?

Formally, we are looking for $q(y|x)$ (coding) or $q(x|y)$ (decoding) under some assumptions about $(X, p(x))$ or $(Y, p(y))$.

A popular assumption is that Y is a set of integers (number of spikes per unit time), and $p(y)$ is Poisson.

Probability theory

Information theory has probability theory at its heart. Random variables (sources) used in information theory are maps between probability measure spaces. A *probability measure space* is the triplet (Ω, \mathcal{B}, P) consisting of a sample space Ω , a σ -field \mathcal{B} of subsets of Ω , and a (probability) measure P , with the usual properties:

$$\begin{array}{ll} P(F) \geq 0, \forall F \in \mathcal{B}; & \text{non-negative} \\ P(\Omega) = 1; & \text{bounded} \\ P\left(\bigcup_{i=1}^{\infty} F_i\right) = \sum_{i=1}^{\infty} P(F_i), & \text{countably additive} \\ \forall F_i \in \mathcal{B} \text{ disjoint.} & \end{array}$$

The basic object in information theory is an *information source* or a random variable (measurable function). Let $(\Omega, \mathcal{B}), (\mathcal{X}, \mathcal{O})$ be two measurable spaces. A random variable is the mapping

$$X : (\Omega, \mathcal{B}) \rightarrow (\mathcal{X}, \mathcal{O}),$$

such that,

$$\text{if } F \in \mathcal{O}, \text{ then } X^{-1}(F) \in \mathcal{B}$$

If \mathcal{X} is countable, X is a discrete random variable (simple function). If $\mathcal{X} \subset \mathbb{R}^n$, X is a continuous random variable.

An example of a discrete random variable is the Poisson, which is characterized by the probability measure $P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$ on the set of positive integers. Our favorite continuous random variable is the Gaussian on the reals, characterized by $p(x) = \frac{1}{\sqrt{(2\pi\sigma^2)}} \exp(-(x - m)^2/2\sigma^2)$.

Expectations

Let f be a measurable function on (Ω, \mathcal{B}, p) . The *expectation* $E_p f$ of f is the integral of f under this measure (if it exists).

$$E_p f \equiv \int f dp$$

If f is a simple function (discrete random variable), the expectation is naturally represented as a sum

$$E_p f = \sum_i f_i p(F_i)$$

where $F_i \in \mathcal{B}$ is the set on which f takes a value f_i .

Frequently used expectations are the mean $\bar{x} = E_p x$, second moment $\bar{x^2} = E_p x^2$, variance $\sigma = E_p (x - \bar{x})^2$.

Information Theory

The basic object in information theory is an *information source* or a random variable (measurable function)

$$X : (\Omega, \mathcal{B}) \rightarrow (\mathcal{X}, \mathcal{O}),$$

where \mathcal{O} is the probability space of symbols produced by X , a representation of the elements of the probability space Ω . A source X is a mathematical model for a physical system that produces a succession of symbols $\{X_1, X_2, \dots, X_n\}$ in a manner which is unknown to us and is treated as random.

$\{X_1, X_2, \dots, X_n\}$ is said to be *i.i.d* or *identically and independently distributed* if X_i are mutually independent and if the probability density of X_i , is the same for every i , $p(X_i) = p(X)$.

$\{X_i\}$ is *stationary* if for each n and k , (X_0, \dots, X_n) and (X_k, \dots, X_{k+n}) have the same probability density. In other words, $\{X_i\}$ are stationary if no matter when one starts observing the sequence of random variables, the resulting observation has the same probabilistic structure.

A measurable transformation $\varphi : \Omega \rightarrow \Omega$ is *measure preserving* if $p(\varphi^{-1}A) = p(A)$ for all $A \in \mathcal{O}$. A set $A \in \mathcal{O}$ is *invariant* if $\varphi^{-1}A = A$. Let $\mathcal{I} = \{A | A \text{ is invariant}\}$. The measurable transformation φ is *ergodic* if for every $A \in \mathcal{I}$, $p(A) \in \{0, 1\}$. The source $X_i = X \circ \varphi^i$ is said to be ergodic if φ is ergodic.

The basic concepts of information theory are *entropy* and *mutual information*. In information theory, entropy is described as a measure of the uncertainty, or of the self information, of a source, and is defined as

$$H(X) = -E_p \log p(x).$$

The *conditional* and *joint* entropy respectively given an information channel (X, Y) are defined respectively as

$$\begin{aligned} H(Y|X) &= -E_p \log p(y|x) \\ H(X, Y) &= -E_p \log p(x, y). \end{aligned}$$

It is easy to show that

$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y).$$

An *information channel* is a pair of information sources (X, Y) .

The notion of *mutual information* $I(X; Y)$ is introduced as a measure of the degree of dependence between a pair of sources in an information channel (X, Y) :

$$\begin{aligned} I(X; Y) &= H(X) + H(Y) - H(X, Y) \\ &= E_{x,y} \log \frac{p(x, y)}{p(x)p(y)} \end{aligned}$$

Both entropy and mutual information are special cases of a more general quantity – the *Kullback-Leibler directed divergence* or *relative entropy* between two probability measures, p and q , on the same event space:

$$D_{KL}(p||q) = E_p \log \left(\frac{p(x)}{q(x)} \right).$$

D_{KL} is always nonnegative and it is zero iff $p(x) = q(x)$ a.e. However, it is not symmetric and so it is not a proper distance on a set of probability measures. In spite of this it provides a sense of how different two probability measures are.

The second distance measure of interest here is the Chernoff distance.

$$D_C(p, q) = -\log \max_{0 \leq t \leq 1} E_p \left[\left(\frac{q}{p} \right)^t \right]$$

The Chernoff distance is always symmetric.

The information quantities H, I, D_{KL} and D_C depend only on the underlying probability distributions and not on the structure of X and Y . This allows us to evaluate them in cases where more traditional statistical measures (e.g. variance, correlation, etc.) do not exist.

Note that entropy is well defined for discrete random variables only, and does not make much sense for continuous. The more appropriate quantity in the continuous case is $H_q(p) \equiv D_{KL}(q||p)$: the relative entropy of p with respect to another measure. For a fixed $q \ll p$ (p is absolutely continuous with respect to q), this quantity retains most of the properties of the discrete entropy.

The data processing inequality

One of the most interesting results frames how systems affect the statistical dependence between random variables. Suppose we have a cascade of two systems with the random variables representing the input and output signals dependent on each other as $X \rightarrow Y \rightarrow Z$. In technical terms, these random variables form a Markov chain ($p(x, y, z) = p(z|y)p(y|x)p(x)$).

Theorem 1 (Data Processing Theorem). *If $X \rightarrow Y \rightarrow Z$ form a Markov chain, $I(X; Z) \leq I(X; Y)$.*

Larger mutual information means greater statistical dependence between a system's input and output. The introduction of a second stage of processing can never increase this dependence; in fact, it could lessen it. More processing is not necessarily beneficial!

There are similar results for H , D_{KL} and D_C as well.

Applicability of information-theoretic quantities

Let $\{Y_1, Y_2, \dots, Y_n\}$ be i.i.d. observations from an information source Y . Then the Strong Law of Large Numbers provides theoretical justification for making inference about population parameters (e.g. response parameters) from data collected experimentally. In particular, the Shannon Entropy Theorem in this case assures that the entropy (and hence the mutual information) calculated from data taken experimentally converges to the true population entropy as the amount of data available increases.

Theorem 2 (Shannon Entropy Theorem). (1948) If $\{Y_i\}$ are i.i.d. then

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log p(Y_1, Y_2, \dots, Y_n) = H(Y) \text{ a.s.}$$

Proof. The random variables $\{\log p(Y_i)\}_{i=1}^n$ are i.i.d. and so by the Strong Law of Large Numbers

$$\begin{aligned} E(\log(p(Y))) &= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \log p(Y_i) \\ &= \lim_{n \rightarrow \infty} \frac{1}{n} \log \prod_{i=1}^n p(Y_i) \\ &= \lim_{n \rightarrow \infty} \frac{1}{n} \log p(Y_1, Y_2, \dots, Y_n) \end{aligned}$$

almost surely. □

In many instances, as in the case of physiological recordings from a biological sensory system, the data $\{Y_1, Y_2, \dots, Y_n\}$ are not i.i.d.. For example, in the data presented in this thesis, a single, “long” recording of a neural response is partitioned into observations of length, say, 10 ms. Inference made about population parameters from data collected this way is justified if we can assume that Y is stationary ergodic. Now we may appeal to the Ergodic Theorem and the Shannon-McMillan-Breiman Theorem to justify the use of information theoretic quantities.

Theorem 3 (Ergodic Theorem). (Birkhoff, 1931) *If φ is a measure preserving transformation on (Ω, \mathcal{O}) and Y is a r.v. with $E(Y) < \infty$. Then*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=0}^{n-1} Y(\varphi^i \omega) = E(Y|\mathcal{I}) \text{ a.s.}$$

Remark 1. *If φ is ergodic, then $E(Y|\mathcal{I}) = E(Y)$. The Ergodic Theorem in this instance can be interpreted as a Strong Law of Large Numbers for ergodic processes.*

Theorem 4 (Shannon-McMillan-Breiman Theorem). (1948, 1953, 1957) *If Y_n for an integer n is an ergodic stationary sequence taking values in a finite set \mathcal{Y} , then*

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log p(Y_0, Y_1, \dots, Y_{n-1}) = H$$

where $H \equiv \lim_{n \rightarrow \infty} E(-\log p(Y_n | Y_{n-1}, \dots, Y_0))$ is the entropy rate of $\{Y_i\}$.

Remark 2. *Theorem 2 is a special case of Theorem 4 when $\{Y_i\}$ are i.i.d..*

Why are entropy and mutual information valid measures to use when analyzing an information channel between X and Y ?

Entropy has meaning through Shannon's Source Coding Theorem. This result prescribes the how random variables can be represented digitally as a sequence of bits.

Theorem 5 (Source Coding Theorem). *(Shannon, 1948) If a discrete random variable X is represented by a bit sequence, wherein each value of the random variable is represented by a sequence of bits, there exists a uniquely decodable bit sequence having an average length \bar{N} that satisfies*

$$H(X) \leq \bar{N} < H(X) + 1$$

What is the significance of this result?

This is a statement about compression: Any discrete (digital) source, irrespective of how complicated its probability distribution is, may be transmitted with a code, the average length of which is determined solely by the entropy of the source.

Interpreting this result, we see that entropy defines how complex a probability law is and how much "work" it takes to represent it. That work equals the entropy. Because entropy is maximized with a uniform probability law, this is the most complex from a communication viewpoint.

How can we understand this result better?

Instead of considering the full space \mathcal{Y} of all of the symbols elicited by Y , let us consider only a subset of \mathcal{Y} which one "typically observes." This set is defined rigorously in the following way. Each element of the output space \mathcal{Y} can be modeled as a sequence of symbols of a random variable

$$Z : (\Omega_Z, \mathcal{O}_Z) \rightarrow (\mathcal{Z}, \mathcal{B}_Z)$$

Let $Y = Z^k$, the k -th extension of Z , be the set of all sequences of length k of symbols from $Z \in \mathcal{Z}$.

There is a limited number of distinct messages which can be transmitted with sequences of length k from the source Z . These are the typical sequences of Z .

Definition 1. *The typical set A_ϵ^k with respect to probability density $p(Z)$ on Z is the set of sequences $(z_1, z_2, \dots, z_k) \in Y$ for which*

$$2^{-k(H(Z)+\epsilon)} \leq p(z_1, z_2, \dots, z_k) \leq 2^{-k(H(Z)-\epsilon)}.$$

$(z_1, z_2, \dots, z_n) \in A_\epsilon^k$ is called a typical sequence.

The typical set has the following properties:

Theorem 6 (Asymptotic Equipartition Property). *If Z is stationary ergodic, then*

1. *If $(z_1, z_2, \dots, z_k) \in A_\epsilon^k$ then*

$$H(Z) - \epsilon \leq -\frac{1}{k} \log p(z_1, z_2, \dots, z_k) \leq H(Z) + \epsilon$$

2. *$Pr\{A_\epsilon^k\} > 1 - \epsilon$ for k sufficiently large*

3. *$(1 - \epsilon)2^{k(H(Z) - \epsilon)} \leq |A_\epsilon^k| \leq 2^{k(H(Z) + \epsilon)}$ for k sufficiently large.
Here $|A|$ is the number of elements in set A .*

Thus the typical set has probability nearly 1, typical sequences are nearly equiprobable (with probability nearly $2^{-kH(Z)}$), and the number of typical sequences is nearly $2^{kH(Z)}$

Mutual information essentially measures how dependent two random variables are. If they are independent, the mutual information is zero; increasing dependence increases the mutual information, with a maximum achieved when the random variables equal each other.

Let X represent a transmitter's digital output to a communications channel and Y the corresponding channel output. To characterize the channel, Shannon defined the *channel capacity* C to be the maximum value of mutual information with respect to the probability distribution of X .

$$C = \max_{p(x)} I(X; Y).$$

Capacity depends only on the conditional probability function $p(y|x)$, which defines the channel's characteristics. Thus, capacity more correctly measures the encoding capabilities of a given system. Note that capacity can also be defined for continuous-valued random variables.

Perhaps Shannon's crowning achievement is the Noisy Channel Coding Theorem and its converse. It is this result that pointedly solves the technical problem of communication: the accuracy to which information can be transmitted.

Theorem 7 (Noisy Channel Coding Theorem). *(Shannon, 1948). There exists an error-correcting code for any rate less than capacity ($R < C$) so that as the code's blocklength (the number of data bits encoded together) approaches infinity, the probability of not being able to correct any errors that occur goes to zero. Furthermore, if $R > C$, errors will occur with probability one.*

Thus, Shannon's Noisy Channel Coding Theorem defines what is meant by reliable communication. It says that despite the fact that a digital channel introduces errors, if sufficient and able error correction is provided for a given channel, all information can be transmitted with no errors. This is an astounding result: Digital communication systems offer the possibility of error-free transmission over error-prone channels. However, Shannon's proof was not constructive: It provides no hint as to what error correcting codes might enable reliable communication.

Newer techniques allow us to have a better understanding of a communication channel. Rewrite X as a sequence of k symbols of a random variable

$$W : (\Omega_W, \mathcal{O}_W) \rightarrow (W, \mathcal{B}_W),$$

so that $X = W^k$. The next theorem considers the behavior of the pair (W, Z) .

Definition 2. *The set A_ϵ^k of jointly typical sequences $\{(w^k, z^k)\}$ with respect to the joint distribution $p(w, z)$ on $W \times Z$ is the set*

$$A_\epsilon^k = \left\{ (w^k, z^k) \in W^k \times Z^k : \begin{aligned} &2^{-k(H(W)+\epsilon)} \leq p(w^k) \leq 2^{-k(H(W)-\epsilon)}, \\ &2^{-k(H(Z)+\epsilon)} \leq p(z^k) \leq 2^{-k(H(Z)-\epsilon)}, \\ &2^{-k(H(W,Z)+\epsilon)} \leq p(w^k, z^k) \leq 2^{-k(H(W,Z)-\epsilon)} \end{aligned} \right\},$$

Theorem 8 (Asymptotic Equipartition Property for jointly typical sequences). *Let (W^k, Z^k) be a pair of i.i.d. sources. Then*

1. $Pr(A_\epsilon^k) > 1 - \epsilon.$

2. $(1 - \epsilon)2^{k(H(W,Z)-\epsilon)} \leq |A_\epsilon^k| \leq 2^{k(H(W,Z)+\epsilon)}$ for n sufficiently large.

3. *If $(\tilde{W}^k, \tilde{Z}^k)$ are a pair of random variables with joint probability $p(w^k, z^k) = p(w^k)p(z^k)$ (i.e. \tilde{W}^k and \tilde{Z}^k are independent with the same marginal distributions as W^k and Z^k), then for sufficiently large k ,*

$$(1 - \epsilon)2^{-k(I(W;Z)+3\epsilon)} \leq Pr\left((\tilde{W}^k, \tilde{Z}^k) \in A_\epsilon^k\right) \leq 2^{-k(I(W;Z)-3\epsilon)}.$$

Thus, the jointly typical set has probability close to 1. The number of jointly typical sequences is nearly $2^{kH(W,Z)}$ and they are each nearly equiprobable (with probability close to $2^{-kH(W;Z)}$). Cover and Thomas give the following argument to ascertain the number of distinguishable signals W^k given a signal Z^k :

For each (typical) n -sequence from X , there are approximately $2^{nH(Y|X)}$ possible Y sequences, all of them (approximately) equally likely. We wish to ensure that no two X sequences produce the same Y output sequence. Otherwise we will not be able to decide which sequence from X was sent.

The total number of possible (typical) Y sequences is $\approx 2^{nH(Y)}$. This set has to be divided into sets of size $2^{nH(Y|X)}$ corresponding to the different X sequences. Hence, the total number of disjoint sets is equal to $2^{n(H(Y)-H(Y|X))} = 2^{nI(X;Y)}$. Hence we can send at most $\approx 2^{nI(X;Y)}$ distinguishable sequences of length n without an error.

Mutual information has been used to characterize signal encoding. Here, X represents some input and Y a (noisy) representation of the input. For example, X might represent a set of stimulus conditions and Y – the neural response to them. Experimentally, the conditional probability function $p(y|x_0)$ can be measured by presenting the stimulus represented by x_0 repeatedly and accumulating an estimate of the resulting response's probability distribution. The mutual information can be computed according to the above formula, but what to use for $p(x)$? The numerical value of mutual information clearly depends on these stimulus probabilities. Because of this dependence, mutual information does not measure the properties of the encoder; it is a joint measure of input probabilities and the conditional output probabilities.

Also note that both entropy and mutual information are concerned with *exact* reproduction of the input sources (through lossless compression, or communication). This is also hardly the case in a nervous system, where there is much evidence of irreversible loss of signal on the way.

So, should neuroscientists forget about Information Theory?

Quantization

A source Y can be related to another random variable Y_N through the process of *quantization* (lossy compression). Y_N is referred to as the *reproduction* of Y . The process is defined by a map

$$q(Y_N|Y) : \mathcal{Y} \rightarrow \mathcal{Y}_N,$$

called a *quantizer*. In general, quantizers can be stochastic: q assigns to $y \in \mathcal{Y}$ the probability that the response y belongs to an abstract class $y_N \in Y_N$. A *deterministic quantizer* (simple function) is a special case in which q takes the values of 0 or 1 only. It can be shown that the mutual information $I(X; Y)$ is the least upper bound of $I(X_M; Y_N)$ over all possible reproductions (X_M, Y_N) of (X, Y) . Hence, the original mutual information can be approximated with arbitrary precision using carefully chosen reproduction spaces.

Rate Distortion Theory

Rate distortion theory is concerned with reduced representations of random variables (lossy compression). The quality of reproduction (fidelity) is assessed through a *distortion function*.

Consider the quantization $X \rightarrow X_N$.

Definition 3. A (pointwise) *distortion function*, or *distortion measure* is a mapping

$$d : X \times X_N \rightarrow R^+$$

from the set of source/reproduction pairs into the set of non-negative reals. The distortion is a measure of the “error” made by representing the symbol $x \in X$ with $x_n \in X_N$.

Example 3 (Squared error distortion). $d(x, x_n) = (x - x_n)^2$.

Definition 4 (Expected (mean) distortion function).

$$D(X; X_N) = E_{p(x, x_n)} d(x, x_n).$$

Definition 5 (Rate distortion problem). *The information rate distortion function $R(D)$ for a source X with a distortion measure $d(x, x_n)$ is defined as*

$$R(D) = \min_{q(x_n|x): D(X; X_N) \leq D} I(X; X_N)$$

where the minimization is over all conditional probabilities $q(x_n|x)$ for which the joint distribution $p(x, x_n) = q(x_n|x)p(x)$ satisfies the expected distortion constraint. Equivalently, one may consider the distortion rate problem

$$D(R) = \min_{q(x_n|x): I(X; X_N) \leq R} D(X; X_N).$$

With this in mind, here is a possible model of the functionality of a sensory (sub)-system:

$$X \xrightarrow{d} X_N \xrightarrow{R} Y$$

that is, the system quantizes the stimulus X to a reproduction X_N according to its own distortion function $d(x, x_n)$, and then sends the reproduction over a communication channel with rate R , to produce the response Y . Interesting questions in this formalism are

- What is the system's distortion function?
- How is the reproduction encoded for transmission to other systems? To be noise-resistive? To help further computations?

Statistical signal processing

The fundamental assumption underlying information processing is that signals represent information. This representation may be direct: each information "value" corresponds to a unique signal, or much more complicated, as with speech conveying meaning or intent. Statistical approaches assume that the signals themselves are stochastic or that statistical interference confounds determining the information. We assume that information can be represented by a parameter or a collection of parameters (a parameter vector). The parameter value could be the information itself or it could indicate what the information is (serve as an index).

When the parameter value is one of several values of a finite set, the signal processing approach is classification: optimally classify the signal as being one of several. When the parameter is numeric and continuous-valued, the approach is estimation: estimate the parameter's value from the observed signal.

In what follows, we summarize the fundamentals of classification and estimation theory. To simplify the discussion, we assume that the signal is just a random variable whose probability measure depends on parameter (or parameters).

Classification

Assume that we have a family of probability measures depending on a parameter α as $p(x; \alpha)$. Assume that the parameter can take on one of two possible values, α_0 or α_1 , and that $Pr[\alpha_0]$, $Pr[\alpha_1]$ denote the probability of these values occurring.

A classifier is a system having X as its input and its classification $\hat{\alpha}$ as its output. In this binary case, $\hat{\alpha}$ equals either α_0 or α_1 . In most cases, the classifier is deterministic: Each value of the random variable corresponds to one of the output values. The average probability of error P_e , the probability that the classifier makes the wrong classification, is given by

$$P_e \equiv Pr[\alpha = \alpha_0] \underbrace{Pr[\hat{\alpha} = \alpha_1 | \alpha = \alpha_0]}_{P_E} + Pr[\alpha = \alpha_1] \underbrace{Pr[\hat{\alpha} = \alpha_0 | \alpha = \alpha_1]}_{P_M}$$

where P_F, P_M are known as the false-alarm and miss probabilities respectively. These probabilities detail all of the possible errors a classification system can make.

To construct optimal classifiers, we need to find the rule that produces the best possible classification performance. Across a broad variety of criteria: minimizing P_e and minimizing P_F with bound on P_M among many, the optimal classification rule amounts to the likelihood ratio test.

$$\begin{aligned} \text{If } \frac{p(x; \alpha_1)}{p(x; \alpha_0)} &> \gamma; \hat{\alpha} = \alpha_1 \\ \text{If } \frac{p(x; \alpha_1)}{p(x; \alpha_0)} &< \gamma; \hat{\alpha} = \alpha_0 \end{aligned}$$

The ratio of the two probability densities having different parameter values is known as the likelihood ratio. Each of these, generically denoted by $p(x; \alpha)$, is known as the likelihood function. The optimal classification system observes the value of the random variable, substitutes it into an expression for the likelihood ratio, compares that result to a threshold, and produces an output depending whether the likelihood ratio was greater or smaller than the threshold. The threshold γ depends on the optimization criterion.

Example: Consider the Gaussian case wherein the expected value is one of two possibilities: α_0 corresponds to $E_p[X] = m_0$ and α_1 to m_1 . Let the variance be the same in each case. Then

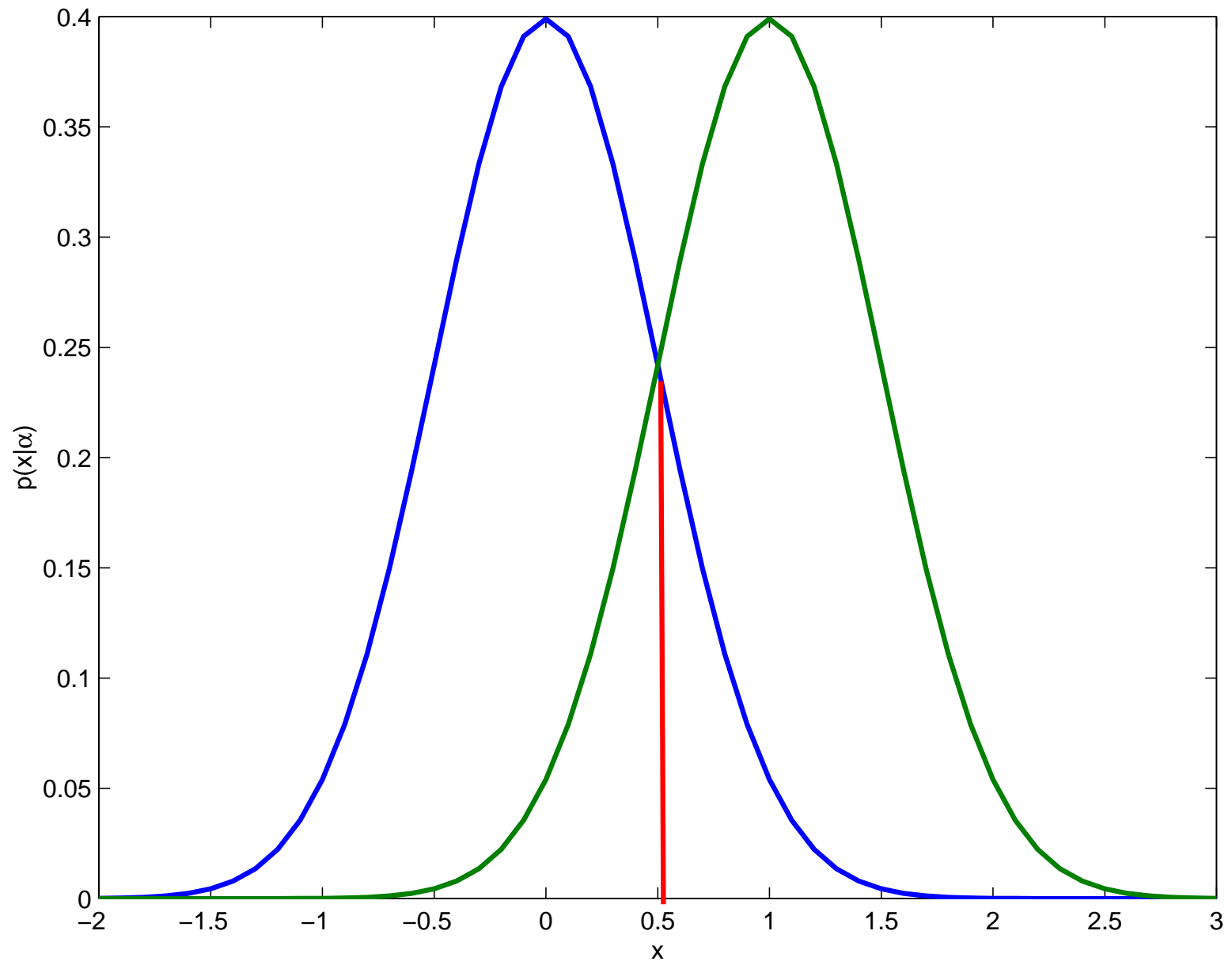
$$\frac{p(x; \alpha_1)}{p(x; \alpha_2)} = \frac{e^{-(x-m_1)^2/2\sigma^2}}{e^{-(x-m_0)^2/2\sigma^2}}$$

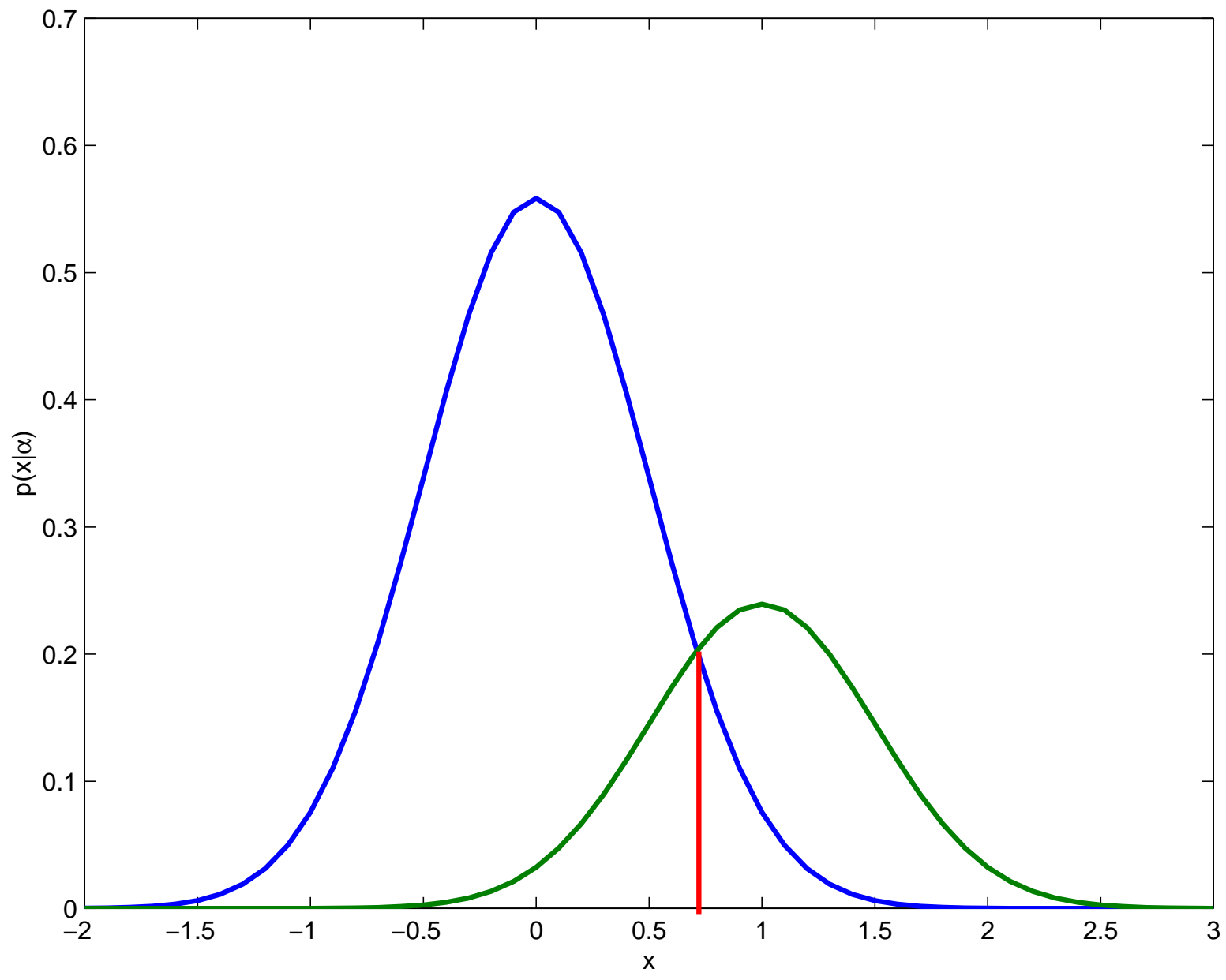
In this (and other) cases, taking the logarithm and comparing it to $\ln \gamma$ simplifies operations. This manipulation will not affect the classifier's performance because \ln is a monotonic function.

$$\ln \frac{p(x; \alpha_1)}{p(x; \alpha_2)} = \frac{-(x - m_1)^2 + (x - m_0)^2}{2\sigma^2}$$

Because the parameters m_0, m_1, σ are assumed known, the classifier's decision rule is neatly summarized as comparing the observed value of the Gaussian random variable to a threshold.

$$x - \frac{m_1 + m_2}{2} > \frac{\sigma^2}{m_1 - m_0} \ln \gamma, \quad m_1 > m_2$$





Despite the fact that the likelihood ratio test is optimal, no general expression for its performance is known. For the Gaussian example, an error probability can be calculated. Not so for other cases. However, a general asymptotic expression does exist. For likelihood ratio classifiers that minimize the miss probability while holding the false-alarm probability constant, the miss probability has the asymptotic form

$$P_M \propto \exp\{-D_{KL}(P(x; \alpha_0) || p(x; \alpha_1))\} \text{ for fixed } P_F$$

Thus, the miss probability always decreases exponentially with increasing Kullback-Leibler distance.

A similar result applies to the average error probability, except that there the Chernoff distance is most important. For likelihood ratio classifiers,

$$P_e \propto \exp\{-D_C(P(x; \alpha_0) || P(x; \alpha_1))\}$$

Estimation

In estimation, the observed random variable's probability measure depends on a parameter that we would like to determine as accurately as possible. Denote by ϵ the estimation error: the difference between the estimate and the actual value ($\epsilon = \hat{\alpha} - \alpha$). If the expected value of the estimation error is zero ($E[\epsilon] = 0$), the estimate is said to be unbiased. In some cases, we want to estimate several parameters from a given set of data; for example, we need an estimate of the mean and variance. We form a parameter vector α and speak of the error vector ϵ .

In deriving optimal estimators, we establish a criterion (some function of the estimation error) and seek the estimate that minimizes it. The estimator depends heavily on the criterion (in contrast to classification where the optimal classifier in general does not depend on the criterion) as well as the probabilistic model (parametric family of measures). The most frequently used criterion is the mean-squared error: Find $\hat{\alpha}$ that minimizes $E[\epsilon^2]$. It is important to note that, in general, the mean-squared error depends on the parameter's actual value: It is usually not a constant.

If the parameter is itself a random variable, the optimal mean squared error estimator is the conditional expectation:

$$\hat{\alpha}_{MS} = E[\alpha|x].$$

If the parameter value is simply unknown (we don't have enough information to assign a probability distribution to it), minimizing the mean-squared error does not yield a meaningful answer. Rather, the ad hoc, but very powerful, maximum likelihood estimation procedure produces very accurate results across a broad range of problems. Here, we seek the function of the random variable that maximizes the log likelihood function.

$$\hat{\alpha}_{MS} = \arg \max_{\alpha} \ln p(x; \alpha).$$

A lower bound on the mean-squared estimation error $E[\epsilon^2]$ can be found regardless of the estimator used! Known as the Cramér-Rao bound, it states that for all unbiased estimators

$$E[\epsilon^2] \geq F_p(\alpha)^{-1}$$

where $F_p(\alpha)$ is known as the Fisher information matrix.

$$F_p(\alpha) \equiv E[\nabla_{\alpha} \ln p(x; \alpha) \nabla_{\alpha} \ln p(x; \alpha)']$$

(one matrix being greater than another means that the difference between them is a positive-definite matrix.)

Despite the ad hoc nature of the maximum likelihood estimate, it has the following properties. If the Cramér-Rao bound can be attained, the maximum likelihood estimate's mean-squared error will equal the lower bound. In such cases, the maximum likelihood estimator is optimal when the criterion is minimum mean-squared error.

As the amount of data grows (now a random vector is observed and its dimension increases), the maximum likelihood estimate will be unbiased and have a mean-squared error equal to the Cramér-Rao bound.

We can relate information-theoretic distance measures to estimation performance as well. Let the random variable depend on a parameter vector α , and let two probability functions be defined when the parameter value equals α and α_0 . The matrix of mixed second derivatives of each distance with respect to α is proportional to the Fisher information matrix:

$$\begin{aligned}\nabla_{\alpha}\nabla_{\alpha}D_{KL}(p(x;\alpha)||p(x,\alpha_0))(\alpha_0) &= F_p(\alpha_0) \\ \nabla_{\alpha}\nabla_{\alpha}D_C(p(x;\alpha)||p(x,\alpha_0))(\alpha_0) &= F_p(\alpha_0)/2\end{aligned}$$

This means that for small changes in the parameter vector ($\alpha = \alpha_0 + \delta\alpha$), the Kullback-Leibler and Chernoff distances are proportional to the Fisher information.

$$\begin{aligned}D_{KL}(p(x;\alpha)||p(x,\alpha_0)) &\approx \frac{\delta\alpha'F_p(\alpha_0)\delta\alpha}{2} \\ D_C(p(x;\alpha)||p(x,\alpha_0)) &\approx \frac{\delta\alpha'F_p(\alpha_0)\delta\alpha}{4}\end{aligned}$$

What about deterministic (dynamical) systems?

Definition 6. *An abstract dynamical system consists of a probability measure space (Ω, \mathcal{B}, P) together with a measurable transformation $\varphi : \Omega \rightarrow \Omega$. The quadruple $(\Omega, \mathcal{B}, P, \varphi)$ is called dynamical system in ergodic theory.*

This is still a far cry from ordinary differential equations, but uses many of the techniques of modern dynamical systems theory (maps, action on measurable sets, invariant measures, measure preserving transformations, etc.)

However, it is still an open question of how we map the functional description that we may uncover with probabilistic “black box” tools to actual neural structure!

Applications to neural systems analysis

Let us consider Atick's example again: Atick (1992) studied the assumption that early sensory processing is concerned with *redundancy reduction*. The system under consideration is

$$\left. \begin{array}{l} \text{Light} \\ \text{Noise} \end{array} \right\} \rightarrow X \rightarrow Y,$$

where X is the contrast of light at the retina, and Y – the neural response (e.g., spike rate). Atick defined the redundancy of the system as

$$\mathcal{R} = 1 - \frac{I(\text{Light}; Y)}{C(Y)}.$$

and considered the relation $y = L x$, where $L : X \rightarrow Y$ is a linear operator. He optimized L so that the second moment of Y , $E[Y Y']$ was proportional to the identity.

Issues:

- In Atick's paper, L was a bijection. According to a corollary of the data processing, a bijection leads to equality: $I(X; Z) = I(X; Y)$. So how is this helping reduce the redundancy $\mathcal{R} = 1 - I(\text{Light}, Y)/C(Y)$?
(*moral - careful with continuous random variables*)
- The input space was *assumed* to be the contrast of light at retinal position, the output space – the spike rate at this retinal position. What if these assumptions are incorrect?
- Is vision really concerned with light? What is the “correct” stimulus space for this sensory modality?
- And more broadly (nothing to do with Atick now), is sensation really concerned with processing whatever the sensors registered?

The capacity C can be defined under various constraints:

$$C = \max_{p(x) \in \mathcal{P}} I(X; Y),$$

where \mathcal{P} is a subset of all possible probabilities. A typical example is capacity with average power constraint:

$$C = \max_{p(x) | E_p(x^2) \leq P_{ave}} I(X; Y).$$

So instead of using mutual information, or unconstrained capacity, one may pose similar optimality problems with the *constrained* capacity, assuming certain constraint types (energy/power being an obvious one). Some very nice results in this area by Vijay Balasubramanian. (energy constraint considerations in coding pioneered by Simon Laughlin).

A big **big big** barrier for using information-theoretic quantities is lack of good estimators for them. They are more general, however they do depend on the underlying probability measure in a non-trivial way. Statistical estimators $E_p f \rightarrow 1/N \sum_i^N f(i)$ do not work here, since $f(i)$ still depends on the measure! Currently, all estimates are done through some estimate of the joint probability.

- Direct estimates (through a histogram): require **enormous** amounts of data. Strongly biased for small data sets (Trevis and Panzeri, '96).
- Estimates through models of the joint probability: everything depends on the model...
- Estimates through series expansion of $I(X;Y)$ (Panzeri and Schultz). Break down rather quickly.
- Estimates through quantization: $I(X_N; Y_M) \leq I(X; Y)$, etc. Less data than direct, but obtain only a lower bound of the quantity. We need to make inferences with bounds! So what if yours is bigger than mine? Nice results by Paninski on precision and uncertainty.

The direct method. (Strong et. al., '98)

The system is $X \rightarrow Y$, where X is a temporal sequence/function of light intensities, Y is a temporal sequence of spikes in the fly's brain. Task: to estimate $I(X; Y)$.

Algorithm:

- Generate a random light sequence.
- Repeatedly show it to the fly (until you are blue in the face, or the fly dies).
- Take subsets of this repeated sequence and obtain a histogram estimate of $p(\text{resp}|\text{stim})$. Use this to estimate $I(X; Y)$.

Issues

- Almost never have enough data, even if Rob de Ruyter records for a day.
- Even if there is enough data for the estimate of $I(X;Y)$ to be unbiased by the size of the dataset, because we have to repeat, we can't sample the stimulus space too well. This may bias the estimate.
- Even if we manage to obtain $I(X;Y)$, what do we learn from that? It depends on the stimulus as well!
- Aside: Say we get the unconstrained capacity somehow? Is the "optimal" stimulus set also "natural".

Do not despair! There are some successful formulations...

An interesting use of $I(X;Y)$ as a tool is in questions of independence and interactions: Is a system of neurons doing more than its individual components through interaction. For 2 cells, is

$$I(X; \{Y_1, Y_2\}) > I(X; Y_1) + I(X; Y_2)?$$

This uses properties of $I(X;Y)$ for independent distributions. Of course it relies on being able to estimate $I(X;Y)$ well ...

Alternate characteristics of system performance: Johnson's information processing functions.

Better estimators: Johnson, Victor, Paninski (and maybe us).

Conclusions

- Information theory is a powerful tool, with broad range of applicability.
- It seems well suited for neural systems analysis. After all, we assume neurons compute, and communicate.
- However, information theory has to be used *carefully*. It's not a panacea!
- Most estimates of information-theoretic quantities in the literature are unreliable.
- Nevertheless, progress seems closer than “just beyond the horizon”. More work is needed.

Acknowledgments

The following standard source were used to compile this tutorial:

- “Elements of Information Theory”, Cover and Thomas (1991), Wiley
- “Entropy and Information Theory”, RM Gray (1990), Springer-Verlag
- “Array Signal Processing”, Johnson and Dudgeon (1993), Prentice Hall

In addition, the author used material from the following sources

- “Solving the rate distortion problem”, Albert Parker (2003), doctoral dissertation, Montana State University
- Don Johnson’s marvelous tutorial, “Basics of Information Processing”, <http://www.ece.rice.edu/~dhj/cv.html#publications>