Basics of Information Processing

Don H. Johnson
Computer & Information Technology Institute
Department of Electrical & Computer Engineering
Rice University, MS366
Houston, TX 77251

Abstract

I describe the basics of probability theory, statistical signal processing and information theory, and how these disciplines inter-relate to form the foundations of a theory of information processing. Examples are drawn from point-process applications.

1 Probability theory

Random variables can be either discrete or continuous valued. Discrete random variables can assume one of a countable set of values, and these values could be numeric (the integers, for example) or symbolic (color values). In either case, the probability of a discrete random variable X taking on the value x $\Pr[X = x]$ is denoted by $P_X(x)$. This *probability function* has the properties

- $P_X(x) \geq 0$
- $\sum_x P_X(x) = 1$

An example of a discrete random variable is the *Poisson*, which is characterized by the probability function

$$P_X(x) = \frac{(\lambda T)^x e^{-\lambda T}}{x!}$$

We use the shorthand notation $X \sim \mathcal{P}(\lambda T)$ to say that X is a Poisson random variable having parameter λT.

Continuous-valued random variables take on a continuum of values over some range. The probability that a continuous random variable X has a value somewhere in the interval $[a, b]$ is given by

$$\Pr[a \leq X \leq b] = \int_a^b p_X(x)\, dx \ ,$$

with $p_X(x)$ known as a *probability density function*. It has the properties

- $p_X(x) \geq 0$
- $\int_{-\infty}^{\infty} p_X(x)\, dx = 1$

An important aspect of the probability density function is that it is *not* dimensionless. By definition, "probability" is dimensionless. Consequently, probability functions for discrete random variables are dimensionless, but density functions have dimensions of the reciprocal of the random variable's dimensions. The most common example of a continuous-valued random variable is the *Gaussian*.

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-m)^2/2\sigma^2}$$

The shorthand notation for a Gaussian random variable is $X \sim \mathcal{N}(m, \sigma^2)$, where m and σ^2 are the distribution's parameters.

We can study the properties of several random variables at once by collecting them into a *random vector* $\mathbf{X} = \mathrm{col}[X_1, X_2, \ldots, X_N]$. When the components are continuous-valued random variables, their probabilistic structure is determined by the so-called *joint* probability density function $p_{\mathbf{X}}(\mathbf{x})$.

$$\Pr[a_1 \leq X_1 \leq b_1 \text{ and } a_2 \leq X_2 \leq b_2 \cdots a_N \leq X_N \leq b_N] = \int_{a_1}^{b_1}\int_{a_2}^{b_2}\cdots\int_{a_N}^{b_N} p_{\mathbf{X}}(\mathbf{x})\, dx_1 dx_2 \cdots dx_N$$

The densities of individual random variables and the joint densities of a subset of the original random variable set can be obtained by integration. In the context of jointly defined random variables, the individual densities are known as *marginals*. For example,

$$\int_{-\infty}^{\infty} p_{X_1,X_2}(x_1, x_2)\, dx_1 = p_{X_2}(x_2) \ .$$

Jointly defined random variables are said to be statistically independent if the joint density equals the product of its marginals: $p_{\mathbf{X}}(\mathbf{x}) = \prod_n p_{X_n}(x_n)$. The most prevalent example of a random vector is the Gaussian, which has the joint density

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{2\pi \det(\mathbf{K})}} e^{-(\mathbf{x}-\mathbf{m})'\mathbf{K}^{-1}(\mathbf{x}-\mathbf{m})/2}$$

Here, $()'$ denotes matrix transpose. The notation indicating a Gaussian random vector is $\mathbf{X} \sim \mathcal{N}(\mathbf{m}, \mathbf{K})$.

We can define averages of all sorts through *expected values*. If $g(\cdot)$ maps from the range-space of a random variable (or random vector) to the reals, its expected value is

$$\mathcal{E}\left[g(X)\right] = \begin{cases} \displaystyle\sum_x g(x) P_X(x) & \text{discrete-valued} \\ \displaystyle\int_{-\infty}^{\infty} g(x) p_X(x)\, dx & \text{continuous-valued} \end{cases} \qquad \mathcal{E}\left[g(\mathbf{X})\right] = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g(\mathbf{x}) p_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} \ .$$

When the random variable is numeric, we can define its expected value as

$$\mathcal{E}\left[X\right] = \sum_x x P_X(x) \quad \text{or} \quad \mathcal{E}\left[X\right] = \int_{-\infty}^{\infty} x p_X(x)\, dx \ .$$

The *variance* of a numeric-valued random variable is defined to be $\mathcal{V}[X] = \mathcal{E}\left[(X - \mathcal{E}\left[X\right])^2\right]$, which can be easily shown to be equal to $\mathcal{E}\left[X^2\right] - (\mathcal{E}\left[X\right])^2$. For the Poisson case, $\mathcal{E}\left[X\right] = \lambda T$ and $\mathcal{V}[X] = \lambda T$, and for the Gaussian case, $\mathcal{E}\left[X\right] = m$ and $\mathcal{V}[X] = \sigma^2$. For random vectors, expected values can also be defined. In particular, the expected value of a random vector is taken to be the vector of component expected values. The *covariance* of a random vector is defined to be $\mathcal{E}\left[(X - \mathcal{E}\left[X\right])(X - \mathcal{E}\left[X\right])'\right]$. In the Gaussian random vector case, the expected value is $\mathbf{m}$ and the covariance is the matrix $\mathbf{K}$.

2 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 function depends on the parameter (or parameters). This simplification reveals much; recasting the results to the case of signals rather than random variables is amazingly straightforward.

2.1 Classification

Assume that the probability function or proabability density depends on the parameter α as $P_X(x,\alpha)$ or $p_X(x,\alpha)$. 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 occuring. Note that $\Pr[\alpha_1] = 1 - \Pr[\alpha_0]$.

A *classifier* is a system having X as its input and its classification $\widehat{\alpha}$ as its output. In this binary case, $\widehat{\alpha}$ equals either α_0 or α_1. In most cases, the classifier is deterministic: Each value of the random variable corresponds without fail 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

$$P_e = \Pr[\alpha = \alpha_0] \underbrace{\Pr[\widehat{\alpha} = \alpha_1 \mid \alpha = \alpha_0]}_{P_F} + \Pr[\alpha = \alpha_1] \underbrace{\Pr[\widehat{\alpha} = \alpha_0 \mid \alpha = \alpha_1]}_{P_M} ,$$

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 classify that produces the best possible classification performance. Across a broad variety of criteria—minimizing P_e and minimizing

the false-alarm probability without completely sacrificing the miss probability among many—the optimal classification rule amounts to the *likelihood ratio test*.

$$\text{If } \frac{p_X(x,\alpha_1)}{p_X(x,\alpha_0)} > \gamma,\ \widehat{\alpha} = \alpha_1$$
$$\text{If } \frac{p_X(x,\alpha_1)}{p_X(x,\alpha_0)} < \gamma,\ \widehat{\alpha} = \alpha_0$$

The ratio of the two probability densities having different parameter values is known as the likelihood ratio. Each of these, generically denoted by $p_X(x,\alpha)$, 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 $\mathcal{E}[X] = m_0$ and α_1 to m_1. Let the variance be the same in each case.

$$\frac{p_X(x,\alpha_1)}{p_X(x,\alpha_0)} = \frac{e^{-(x-m_1)^2/2\sigma^2}}{e^{-(x-m_0)^2/2\sigma^2}}$$

Because we only need be concerned with the likelihood ratio's value relative to the threshold, we can simplify our work by taking the logarithm and comparing it to $\ln\gamma$. This manipulation will not affect the classifier's performance because the logarithm is a monotonic function.

$$\begin{aligned}\ln\frac{p_X(x,\alpha_1)}{p_X(x,\alpha_0)} &= -\frac{(x-m_1)^2}{2\sigma^2} + \frac{(x-m_0)^2}{2\sigma^2} \\ &= \frac{2x(m_1-m_0) - (m_1^2 - m_0^2)}{2\sigma^2}\end{aligned}$$

Because the parameters m_0, m_1, and σ^2 are assumed known, the classifier's decision rule is neatly summarized as comparing the observed value of the Gaussian random variable to a threshold.

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

Example: Now consider a Possion case with two parameter possibilities λ_0, λ_1. Again calculating the log likelihood ratio,

$$\ln\frac{p_X(x,\alpha_1)}{p_X(x,\alpha_0)} = x\ln\frac{\lambda_1}{\lambda_0} - (\lambda_1 - \lambda_0)T$$

Simplifying yields

$$x \underset{<}{\overset{>}{}} \frac{\ln\gamma + (\lambda_1 - \lambda_0)T}{\ln\frac{\lambda_1}{\lambda_0}}\ , \quad \lambda_1 > \lambda_0$$

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; for the Poisson, no closed-form expression can be found. An asymptotic expression does exist, but we need some information theoretic results first.

2.2 Estimation

In estimation, the observed random variable's (or vector's) probability law 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 ($\epsilon = \widehat{\alpha} - \alpha$). A little jargon: If the expected value of the estimation error is zero ($\mathcal{E}[\epsilon] = 0$), 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 $\boldsymbol{\alpha}$ and speak of the error vector $\boldsymbol{\epsilon}$.

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 does not depend on the criterion) as well as the probabilistic model. The most frequently used criterion is the mean-squared error: Find $\widehat{\alpha}$ that minimizes $\mathcal{E}\left[\epsilon^2\right]$ in the scalar case and find $\widehat{\boldsymbol{\alpha}}$ that minimizes $\mathcal{E}[\boldsymbol{\epsilon}'\boldsymbol{\epsilon}]$, the sum of the component mean-squared errors. 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 (an infrequent occurrence), the optimal mean-squared error estimator is the conditional expectation: $\widehat{\alpha}_{\mathrm{MS}} = \mathcal{E}[\alpha \mid x]$. 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 as we shall see 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.

$$\widehat{\boldsymbol{\alpha}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\alpha}} \ln p_X(x, \boldsymbol{\alpha}) \ .$$

A lower bound on the mean-squared estimation error $\mathcal{E}\left[\epsilon^2\right]$ can be found regardless of the

estimator used. Known as the *Cramér-Rao bound*, it states that for all unbaised estimators

$$\text{Scalar case:} \quad \mathcal{E}\left[\epsilon^2\right] \geq \frac{1}{F_X(\alpha)}$$

$$\text{Vector case:} \quad \mathcal{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\right] \geq \left[\mathbf{F}_{\mathbf{X}}(\boldsymbol{\alpha})\right]^{-1}$$

where $\mathbf{F}_{\mathbf{X}}(\boldsymbol{\alpha})$ is known as the *Fisher information* matrix. The quantity $\mathcal{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\right]$ is the error covariance matrix $\mathbf{K}_{\boldsymbol{\epsilon}}$ in the unbiased case.

$$\text{Scalar case:} \quad F_X(\alpha) = \mathcal{E}\left[\left(\frac{d}{d\alpha}\ln p_X(x,\alpha)\right)^2\right]$$

$$\text{Vector case:} \quad \mathbf{F}_{\mathbf{X}}(\boldsymbol{\alpha}) = \mathcal{E}\left[\left(\nabla_{\boldsymbol{\alpha}}\ln p_{\mathbf{X}}(\mathbf{x},\boldsymbol{\alpha})\right)\left(\nabla_{\boldsymbol{\alpha}}\ln p_{\mathbf{X}}(\mathbf{x},\boldsymbol{\alpha})\right)'\right]$$

One matrix being greater than another means that the difference between them is a positive-definite matrix. In order for this difference to be positive-definite, the diagonal values must be positive. This matrix property means that each parameter's mean-squared error, an entry on the diagonal of the error covariance matrix, must be greater than the corresponding entry in the inverse of the Fisher information matrix.

$$\mathcal{E}\left[\epsilon_i^2\right] = \left[\mathbf{K}_{\boldsymbol{\epsilon}}\right]_{ii} \geq \left[\mathbf{F}^{-1}(\boldsymbol{\alpha})\right]_{ii}$$

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.

Example: Suppose we want to estimate the parameter λ of a Poisson random variable. Usually, we don't have a probability distribution for this parameter, so we use maximum likelihood.

$$\widehat{\alpha} = \arg\max_{\lambda}\left[x\ln\lambda T - \lambda T - \ln x!\right]$$

Calculating the derivative and setting it equal to zero yields

$$\frac{d}{d\lambda}\left[x\ln\lambda T - \lambda T - \ln x!\right] = \frac{x}{\lambda} - T = 0 \implies \widehat{\lambda}_{\text{ML}} = \frac{x}{T}$$

Because $\mathcal{E}\left[\widehat{\lambda}_{\mathrm{ML}}\right] = \lambda$, the estimate is unbiased, which means the Cramér-Rao bound applies. The Fisher information is found to be

$$F_X(\lambda) = \mathcal{E}\left[\left(\frac{x}{\lambda} - T\right)^2\right] = \frac{T}{\lambda}\,.$$

Calculating the mean-squared error yields the same result, which means that no other estimation procedure can have a smaller mean-squared error than the maximum likelihood estimator in this case.

3 Information (Shannon) theory

"Classic" information theory originated with the 1948 publication of Claude Shannon's paper "A Mathematical Theory of Communication." As pointed out in the introduction of the subsequently published book *The Mathematical Theory of Communication*, which contains a corrected version of that paper (with a subtle title change), Shannon's theory is just a beginning. It concerns the efficient encoding and transmission of digital signals. As Warren Weaver, who wrote the book's introduction as a paper in *Scientific American* (1949), stated

> In communication there seems to be problems at three levels: 1) technical, 2) semantic, and 3) influential.
>
> The technical problems are concerned with the accuracy of transference of information from sender to receiver. . . . The semantic problems are concerned with the interpretation of meaning by the receiver, as compared with the intended meaning of the sender. . . . The problems of influence or effectiveness are concerned with the success with which the meaning conveyed to the receiver leads to the desired conduct on his part.

In this light, what we now call information theory concerns the "technical" problem, and as the publication titles suggest, Shannon's work concerns communication, not information, which carries connotations of meaning. Later on, Weaver says

> The concept of information developed in this theory at first seems disappointing and bizarre—disappointing because it has nothing to do with meaning, and bizarre because it deals not with a single message but rather with the statistical character of a whole ensemble of messages, bizarre also because in these statistical terms the words information and uncertainty find themselves partners.

> But we have seen upon further examination of the theory that this analysis has so penetratingly cleared the air that one is now perhaps for the first time ready for a real theory of meaning.

Despite Weaver's optimism, no theory of information (meaning) has been developed, although the name "information theory" has stuck. To develop a true theory of information, we need to understand Shannon theory, the modern term for results surrounding entropy, mutual information, and capacity.

3.1 Entropy

Shannon defined the entropy of the random variable X, denoted as $\mathcal{H}(X)$, to be

$$\mathcal{H}(X) = -\sum_{x} P_X(x) \log P_X(x) = \mathcal{E}\left[-\log P_X(X)\right]$$

The base of the logarithm determines the units for entropy. In this paper, $\log(\cdot)$ denotes base-2 logarithm and, somewhat unconventionally, $\exp(\cdot)$ denotes base-2 exponential. Entropy has the following properties.

- $\mathcal{H}(X) \geq 0$.
- $\mathcal{H}(X) = 0$ when the probability function concentrates all its probability at one value: $P_X(x_0) = 1$ for some x_0, which makes the other probability values zero.
- When the number of possible values of the random variable is finite, $\mathcal{H}(X)$ is maximized for a uniform probability law.
- When the number of possible values is infinite, $\mathcal{H}(X)$ is maximized by the geometric distribution ($P_X(x) = (1-a)a^x$, $x = 0, 1, \ldots, 0 \leq a < 1$).

Example: Let $X \sim \mathcal{P}(\lambda T)$. This random variable's entropy cannot be calculated in closed form.

$$\begin{aligned}\mathcal{H}(X) &= \mathcal{E}\left[-x \ln \lambda T + \lambda T + \ln x!\right] / \ln 2 \\ &= \left(\lambda T(1 - \ln \lambda T) + \mathcal{E}\left[\ln x!\right]\right) / \ln 2\end{aligned}$$

The expected value $\mathcal{E}\left[\ln x!\right]$ has no closed form expression.

Note that entropy cannot be defined for continuous-valued random variables, and that none of the properties just listed apply if one forced the definition. If a definition were extant, it

would be $-\int p_X(x) \log p_X(x)\, dx$. Because probability densities have units equal to the reciprocal units of the random variable, the logarithm of a density makes little sense. For example, when $X \sim \mathcal{N}(m, \sigma^2)$, its entropy is $\frac{1}{2}\ln(2\pi\sigma^2 e)/\ln 2$. Note that the logarithm of the variance will depend on the units of the random variable: If you make a voltage measurement in volts, scaling it to be in millivolts results in a different, possibly negative, even zero, entropy.

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.

Source Coding Theorem. If a discrete random variable 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 (you can determine the value of the random variable from the bit sequence) having an average length $\bar{N}$ that satisfies

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

Furthermore, if $\bar{N} < \mathcal{H}(X)$, no uniquely decodable sequence exists.

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.

3.2 Mutual Information and Capacity

Before defining the information theoretic quantities mutual information and capacity, the concept of error correction needs to be understood. Communication engineers long recognized that in the case of digital transmission, errors that occur could be corrected by the receiver. For example, suppose each bit is sent three times. With this *repetition code*, the sequence `011` is transmitted as `000111111`. We define the *rate* R of a code to be the reciprocal of the number of bits sent for each data bit: our repetition code example has a rate of $1/3$. If the original sequence were transmitted and an error occurred (so that `111` were received, for example), the receiver would have no method of determining if an error occurred much less correcting it. If transmitted with the repetition code and `100111110` were received, the receiver can not only detect errors but also correct them by a simple majority vote: Whichever bit occurs most frequently in a block of three bits, take that as being what the transmitter sent. One can easily show that if the probability of a bit being received in error is less than $\frac{1}{2}$, the repetition code reduces the probability of an error. However, this so-called *channel coding* scheme does not make the probability of error zero; it merely reduces it. Engineers

long wondered what the limits of error correction were and what the best error correcting codes were. Shannon entered the stage and gave a profound answer.

Shannon defined the *mutual information* between two random variables X, Y to be

$$\mathcal{I}(X;Y) = \sum_{x,y} P_{X,Y}(x,y) \log \frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)} = \mathcal{E}\left[\log \frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)}\right] \tag{1}$$

Mutual information can be defined for continuous-valued random variables in a similar way. Mutual information has the properties

- $\mathcal{I}(X;Y) \geq 0$, equaling zero only when X and Y are statistically independent random variables.
- For *discrete* random variables, $\mathcal{I}(X;Y) = \mathcal{H}(X) - \mathcal{H}(X \mid Y)$, where the latter quantity equals

  $$\mathcal{H}(X \mid Y) = -\sum_{x,y} P_{X,Y}(x,y) \log P_{X|Y}(x \mid y) \ .$$

  Also $\mathcal{I}(X;Y) = \mathcal{H}(Y) - \mathcal{H}(Y \mid X)$. This formula is not meaningful in the case of continuous-valued random variables because of the problems associated with entropy.
- Because of this property, in the discrete-random variable case, $\mathcal{I}(X;Y) \leq \mathcal{H}(X)$ because conditional entropy is non-negative. Mutual information achieves this upper bound when the two random variables are equal. In the case of *continuous* random variables, mutual information has no upper bound ($\mathcal{I}(X;X) = \infty$).

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.

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 $P_{Y|X}(y \mid x_0)$ can be measured by presenting the stimulus represented by x_0 repeatedly and acumulating an estimate of the resulting response's probability distribution. The mutual information can be computed according to the above formula (1), but what to use for $P_X(x)$? 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.

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

$$\begin{aligned} C &\equiv \max_{P_X(\cdot)} \mathcal{I}(X;Y) \\ &= \max_{P_X(\cdot)} \sum_{x,y} P_{Y|X}(y \mid x) P_X(x) \log \frac{P_{Y|X}(y \mid x)}{\sum_\alpha P_{Y|X}(y \mid \alpha) P_X(\alpha)} \end{aligned}$$

This latter expression for the mutual information shows that capacity depends only on the conditional probability function $P_{Y|X}(y \mid x)$, 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.

Noisy Channel Coding Theorem. 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 with 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. It is known that our simple repetition code won't work; Shannon's result says some other rate $1/3$ code will, so long as the capacity is greater.

As it stands, capacity has units of bits/transmission. If we divide by the time T taken for a transmission, we express capacity in bits/second: $C' = C/T$. Thus expressed, the capacity determines the maximum datarate that can be used for a given channel if we hope to have reliable communication.

Shannon calculated the capacity of a specific channel: one that adds white Gaussian noise to the transmissions.

$$C' = W \log\left(1 + \frac{\mathrm{S}}{\mathrm{N}}\right) \text{ bits/s}$$

Here, W denotes the available bandwidth and $\frac{\mathrm{S}}{\mathrm{N}}$ denotes the signal-to-noise ratio in that band. Capacity is sometimes known as the "digital bandwidth."

Kabanov derived the capacity of the point process channel in 1975. He showed that of all point processes, the Poisson process achieved capacity; if a Poisson process is used for digital communication in such a way that the event rate $\lambda(t)$ has average $\bar{\lambda}$ and ranges between $\lambda_{\min}$ and $\lambda_{\max}$,

$$C' = \begin{cases} \dfrac{\lambda_{\min}}{\ln 2}\left[\frac{1}{e}\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{\lambda_{\max}/(\lambda_{\max}-\lambda_{\min})} - \ln\left\{\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{\lambda_{\max}/(\lambda_{\max}-\lambda_{\min})}\right\}\right] & \bar{\lambda} - \lambda_{\min} \geq \lambda^\circ \\ \dfrac{\bar{\lambda} - \lambda_{\min}}{\ln 2} \ln \dfrac{\lambda_{\max} - \lambda_{\min}}{\bar{\lambda} - \lambda_{\min}} & \bar{\lambda} - \lambda_{\min} < \lambda^\circ \end{cases}$$

where

$$\lambda^\circ = \frac{\lambda_{\min}}{e}\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{\lambda_{\max}/(\lambda_{\max}-\lambda_{\min})}$$

In the interesting special case when $\lambda_{\min} = 0$, the capacity expression simplifies greatly.

$$C' = \begin{cases} \frac{\lambda_{\max}}{e \ln 2} = 0.53\lambda_{\max} & \bar{\lambda} > \lambda^\circ = \lambda_{\max}/e \\ \dfrac{\bar{\lambda}}{\ln 2} \ln \dfrac{\lambda_{\max}}{\bar{\lambda}} & \bar{\lambda} < \lambda^\circ \end{cases}$$

It is popular to quote the capacity of a point process channel in bits/event by dividing this result by the average event rate $\bar{\lambda}$.

$$\frac{C'}{\bar{\lambda}} = \begin{cases} 0.53\dfrac{\lambda_{\max}}{\bar{\lambda}} & \bar{\lambda} > \lambda_{\max}/e \\ \dfrac{1}{\ln 2} \ln \dfrac{\lambda_{\max}}{\bar{\lambda}} & \bar{\lambda} < \lambda_{\max}/e \end{cases}$$

Capacity expressed this way depends only on the ratio of the maximum and average event rates (when $\lambda_{\min} = 0$).

The Noisy Channel Coding Theorem also applies to analog (continuous-valued) communication, but in a much more complicated way. Let the input be represented by the random variable X. It is encoded in Z, and this quantity is transmitted, received as Z', and the result decoded as X'.

Thus, we have the chain $X \rightarrow Z \rightarrow Z' \rightarrow X'$. Shannon defines what is now known as a *distortion measure* ν according to

$$\nu = \int\int \rho(x, x') p_{X,X'}(x, x')\, dx\, dx' ,$$

where $\rho(x, x')$ is the *rate distortion function* that measures how similar the original and decoded signals are. For example, $\rho(x, x')$ could be $(x - x')^2$, which gives the mean-squared distortion. He then defines the rate R at which information can be reproduced to a given distortion ν_0 as

$$R = \min_{p_{X,X'}(x,x')} \mathcal{I}(X; X') \text{ with } \nu = \nu_0 .$$

This constrained minimization with respect to the joint probability function $p_{X,X'}(x, x')$ can be quite difficult to compute. The Noisy Channel Coding Theorem now becomes that so long as $R < C$, where capacity is computed with respect to the pair (Z, Z'), the input can be encoded in such a way that the required distortion criterion is met. The more stringent the criterion (the smaller ν_0), the larger R becomes until one cannot send information through the channel and meet the distortion criterion. Said another way, finding the capacity of a continuous-valued channel, like the Poisson, yields a limit on how effectively a source can be encoded according any distortion criterion. However, translating a capacity into a value for a distortion measure can be very difficult.

4 Beyond Classic Information Theory

The field now known as information theory is quite broad. The subfield known as Shannon theory has elaborated Shannon's work in many ways. Some pertinent results for information processing theory, an attempt to bring semantics and influential considerations to the stage, follow. What is meant by "pertinent" is relating information theoretic quantities to those occuring in statistical signal processing. In statistical signal processing, we assume that signals, abstracted here to be random variables, represent a parameter α. Representing a random variable with a bit stream or reliably communicating it through a noisy channel makes no reference to the underlying parameter, which we take here to ultimately be important. General relations between mutual information/entropy and optimal classifier/estimator performances are not known. We need quantities that are related. Furthermore, a viable theory must deal with continuous-valued signals on the same footing as discrete-valued ones. The problems of meaningfully defining entropy and the complexity of the Noisy Channel Coding Theorem in the continuous case mean that a different approach is required.

4.1 Data Processing Theorem

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 $X \rightarrow Y \rightarrow Z$. Thus, X is the first system's input, Y is the first system's output and second system's input, and Z is the second system's output. In technical terms, these random variables form a Markov chain.

Data Processing Theorem. If $X \rightarrow Y \rightarrow Z$ form a Markov chain, $\mathcal{I}(X;Y) \geq \mathcal{I}(X;Z)$.

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. From the Data Processing Theorem, we see that more processing is not necessarily beneficial.

4.2 Information-theoretic distances

Work has shown that linkages between information theory and signal processing exist through quantifying how "different" two probability functions (or densities) might be. For technical reasons, no distance measure having the true properties of a distance, in particular symmetry and the triangle inequality, make this linkage well. We should not term the following quantities distances, but we shall to avoid jumping through semantic hoops.

Two distances are most important for us. Let P_X and Q_X denote two probability functions defined for the same random variable. The *Kullback-Leibler* distance from Q to P is defined to be

$$\mathcal{D}\left(P\|Q\right) = \sum_x P_X(x) \log \frac{P_X(x)}{Q_X(x)} = \mathcal{E}_P\left[\log \frac{P}{Q}\right]$$

The Kullback-Leibler distance between densities can also be defined. The notation $\mathcal{E}_P[\cdot]$ means the expected value of the indicated quantity with respect to the probability function P. The Kullback-Leibler distance has the properties

- $\mathcal{D}(P\|Q) \geq 0$, equaling zero only when $P = Q$.
- If $\mathbf{X}$ has statistically independent components under both probability laws, $\mathcal{D}(P_{\mathbf{X}}\|Q_{\mathbf{X}}) = \sum_i \mathcal{D}(P_{X_i}\|Q_{X_i})$.
- The Kullback-Leibler distance acts like a squared distance.

- The Kullback-Leibler distance is usually not symmetric in the two probability functions: $\mathcal{D}(P\|Q) \neq \mathcal{D}(Q\|P)$.

The Kullback-Leibler distance quantifies how two probability distributions defined for the same random variable differ. Zero distance means no difference, and the distance can be infinite.

Example: Let X be Gaussian, with $p = \mathcal{N}(m_1, \sigma^2)$ and $q = \mathcal{N}(m_0, \sigma^2)$. $\mathcal{D}(p\|q) = \frac{(m_1 - m_0)^2}{2\sigma^2}$. The Gaussian is one of the few yielding a symmetric distance.

Example: If $P = \mathcal{P}(\lambda_1 T)$, $Q = \mathcal{P}(\lambda_0 T)$ be two Poisson distributions. $\mathcal{D}(P\|Q) = \frac{T}{\ln 2}\left[\lambda_1 \ln \frac{\lambda_1}{\lambda_0} + (\lambda_0 - \lambda_1)\right]$. In this case, the distance is asymmetric.

Using the Kullback-Leibler distance, we can rephrase the Data Processing Theorem in more pointed terms. If $X \rightarrow Y$, with P_X resulting in P_Y and Q_X yielding Q_Y, then $\mathcal{D}(P_X\|Q_X) \geq \mathcal{D}(P_Y\|Q_Y)$. Thus, systems can only bring two different input distributions closer together.

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

$$\mathcal{C}(P,Q) = -\log \max_{0 \leq t \leq 1} \sum_x [P_X(x)]^{(1-t)} [Q_X(x)]^t = -\log \max_{0 \leq t \leq 1} \mathcal{E}_P\left[\left(\frac{Q}{P}\right)^t\right]$$

The Chernoff distance is also well-defined for densities. Note that the Chernoff distance is always symmetric. The Data Processing Theorem can be expressed in terms of it as well.

4.3 Relations to signal processing

For likelihood ratio classifiers that minimize the miss probability while holding the false-alarm probability constant, the miss probability has the asymptotic form

$$P_M \sim \exp\{-\mathcal{D}(P_{\mathbf{X}}(\mathbf{x}, \alpha_0)\|P_{\mathbf{X}}(\mathbf{x}, \alpha_1))\} \text{ for all } P_F \ .$$

Thus, as the dimension of the random vector increases, the miss probability always decreases exponentially with increasing Kullback-Leibler distance.

A similar result applies to the average error probability, except that the Chernoff distance is most important. For likelihood ratio classifiers, as the dimension of the random vector increases,

$$P_e \sim \exp\{-\mathcal{C}(P_{\mathbf{X}}(\mathbf{x}, \alpha_0), P_{\mathbf{X}}(\mathbf{x}, \alpha_1))\} \text{ [natural logarithm]}.$$

We can relate both distance measures to estimation performance. Let the random variable depend on a parameter vector $\boldsymbol{\alpha}$, and let two probability functions be defined when the parameter

value equals $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}_0$. The matrix of mixed second derivatives of each distance with respect to $\boldsymbol{\alpha}$ is proportional to the Fisher information matrix.

$$\left.\frac{\partial^2 \mathcal{D}(P_{\mathbf{X}}(\mathbf{x},\boldsymbol{\alpha})\|P_{\mathbf{X}}(\mathbf{x},\boldsymbol{\alpha}_0))}{\partial\alpha_i\partial\alpha_j}\right|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_0} = [\mathbf{F}_{\mathbf{X}}(\boldsymbol{\alpha}_0)]_{ij} \text{ [natural logarithm]}$$

$$\left.\frac{\partial^2 \mathcal{C}(P_{\mathbf{X}}(\mathbf{x},\boldsymbol{\alpha}),P_{\mathbf{X}}(\mathbf{x},\boldsymbol{\alpha}_0))}{\partial\alpha_i\partial\alpha_j}\right|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_0} = \frac{1}{2}[\mathbf{F}_{\mathbf{X}}(\boldsymbol{\alpha}_0)]_{ij} \text{ [natural logarithm]}$$

This means that for small changes in the parameter vector ($\boldsymbol{\alpha} = \boldsymbol{\alpha}_0 + \delta\boldsymbol{\alpha}$), the Kullback-Leibler and Chernoff distances are proportional to Fisher information.

$$\mathcal{D}(P_{\mathbf{X}}(\mathbf{x},\boldsymbol{\alpha}+\delta\boldsymbol{\alpha})\|P_{\mathbf{X}}(\mathbf{x},\boldsymbol{\alpha})) \approx \frac{(\delta\boldsymbol{\alpha})'\mathbf{F}_{\mathbf{X}}(\boldsymbol{\alpha})(\delta\boldsymbol{\alpha})}{2} \text{ [natural logarithm]}$$

$$\mathcal{C}(P_{\mathbf{X}}(\mathbf{x},\boldsymbol{\alpha}),P_{\mathbf{X}}(\mathbf{x},\boldsymbol{\alpha}+\delta\boldsymbol{\alpha})) \approx \frac{(\delta\boldsymbol{\alpha})'\mathbf{F}_{\mathbf{X}}(\boldsymbol{\alpha})(\delta\boldsymbol{\alpha})}{4} \text{ [natural logarithm]}$$

These distance measures form the linkage between the Data Processing Theorem and the two main problems in statistical signal processing. A large distance means that classification error probabilities decrease more rapidly and that estimation errors will be small (Fisher information is inversely related to mean-squared estimation error). We would like information processing systems to maintain as large a distance as possible to yield as small errors as possible; the limit on errors is determined by those of optimal processing algorithms operating on the system's input.