Information Theory and Coding

Computer Science Tripos Part II, Lent Term 1997
12 Lectures by J G Daugman

1. Overview: What is Information Theory? Historical Origins
2. Mathematical Foundations; Probability Rules; Bayes' Theorem
3. Entropies Defined, and Why They Are Measures of Information
4. Source Coding Theorem; Prefix, Variable-, & Fixed-Length Codes
5. Channel Types, Properties, Noise, and Channel Capacity
6. Continuous Information; Density; Noisy Channel Coding Theorem
7. Fourier Series, Convergence, Orthogonal Representation
8. Useful Fourier Theorems; Transform Pairs; Sampling; Aliasing
9. Discrete Fourier Transform. Fast Fourier Transform Algorithms
10. The Quantized Degrees-of-Freedom in a Continuous Signal
11. Gabor-Heisenberg-Weyl Uncertainty Relation. Optimal "Logons"
12. Kolmogorov Complexity. Minimal Description. Time Series.

Summary of Course Content

Foundations and uncertainty. How the movements and transformations of information, just like those of a fluid, are law-governed. How signals, channels, encoders and decoders are constrained in how much information they can provide. How the formal concepts of information are grounded in the principles and rules of probability; and how randomness, redundancy, compressibility, noise, bandwidth, and uncertainty are intricately connected to information.

Probability and information measures. Ensembles, random variables, marginal and conditional probabilities. Bayes' rule. Marginal entropy, joint entropy, conditional entropy, and the Chain Rule for entropy. Mutual information between ensembles of random variables. Why entropy is the fundamental measure of information content.

Coding theorems, channels, and communication capacity. Shannon's Source Coding Theorem. Prefix, variable-, and fixed-length symbol codes. Capacity of a noiseless discrete channel. Error correcting codes. Perfect communication through a noisy channel. Shannon's Noisy Channel Coding Theorem. Continuous information, signal-to-noise ratio, and power spectral density. Gaussian channels. SNR versus frequency. The Shannon Limit.

Fourier analysis and information. Representation of continuous or discrete data by complex exponentials. Fourier series and Fourier transforms in multiple dimensions. Transform pairs and useful theorems. FFT algorithms. Nyquist's sampling theorem; aliasing. The relation between spectral properties and statistical ones. Filters, correlation, modulation, and coherence. Wiener analysis and the information in a time-series.

Quantized degrees-of-freedom in signals; complexity. Why a continuous signal of finite bandwidth and duration has a fixed number of degrees of freedom. The Gabor-Heisenberg-Weyl uncertainty relation. Gabor's optimal expansion basis. Logons. Multi-resolution wavelet codes, and their extensions to images. Minimal description length; Kolmogorov complexity.

Reference book:

Cover, T.M. & Thomas, J.A. (1991). *Elements of Information Theory.*
New York: Wiley.

Information Theory and Coding

Computer Science Tripos Part II, Lent Term 1997
12 lectures by J G Daugman

---

1. Overview: What is Information Theory?

Key idea: The movements and transformations of information, just like those of a fluid, are constrained by mathematical and physical laws. These laws have deep connections with:

- probability theory, statistics, and combinatorics
- thermodynamics (statistical physics)
- spectral analysis, Fourier (and other) transforms
- sampling theory, prediction, estimation theory
- electrical engineering (bandwidth; signal-to-noise ratio)
- complexity theory (minimal description length)
- signal processing, representation, compressibility

As such, information theory addresses and answers the two fundamental questions of communication theory:

1. What is the ultimate data compression?
   (answer: the entropy of the data, H, is its compression limit.)
2. What is the ultimate transmission rate of communication?
   (answer: the channel capacity, C, is its rate limit.)

All communication schemes lie in between these two limits on the compressibility of data and the capacity of a channel. Information theory can suggest means to achieve these theoretical limits. But the subject also extends far beyond communication theory.

Important questions... to which Information Theory offers answers:

- How should information be measured?
- How much additional information is gained by some reduction in uncertainty?
- How do the *a priori* probabilities of possible messages determine the informativeness of receiving them?
- What is the information content of a random variable?
- How does the noise level in a communication channel limit its capacity to transmit information?
- How does the bandwidth (in cycles/second) of a communication channel limit its capacity to transmit information?
- By what formalism should prior knowledge be combined with incoming data to draw formally justifiable inferences from both?
- How much information in contained in a strand of DNA?
- How much information is there in the firing pattern of a neurone?

Historical origins and important contributions:

- Ludwig BOLTZMANN (1844-1906), physicist, showed in 1877 that thermodynamic entropy (defined as the energy of a statistical ensemble [such as a gas] divided by its temperature: ergs/degree) is related to the statistical distribution of molecular configurations, with increasing entropy corresponding to increasing randomness. He made this relationship precise with his famous formula $S = k \log W$ where S defines entropy, W is the total number of possible molecular configurations, and k is the constant which bears Boltzmann's name: k =1.38 x 10^{-16} ergs per degree centigrade. (The above formula appears as an epitaph on Boltzmann's tombstone.) This is equivalent to the definition of the information ("negentropy") in an ensemble, all of whose possible states are equiprobable, but with a minus sign in front (and when the logarithm is base 2, k=1.) The deep connections between Information Theory and that branch of physics concerned with thermodynamics and statistical mechanics, hinge upon Boltzmann's work.

- Leo SZILARD (1898-1964) in 1929 identified entropy with information. He formulated key information-theoretic concepts to solve the thermodynamic paradox known as "Maxwell's demon" (a thought-experiment about gas molecules in a partitioned box) by showing that the amount of information required by the demon about the positions and velocities of the molecules was equal (negatively) to the demon's entropy increment.

- James Clerk MAXWELL (1831-1879) originated the paradox called "Maxwell's Demon" which greatly influenced Boltzmann and which led to the watershed insight for information theory contributed by Szilard. At Cambridge, Maxwell founded the Cavendish Laboratory in which we are now assembled.

- R V HARTLEY in 1928 founded communication theory with his paper *Transmission of Information.* He proposed that a signal (or a communication channel) having bandwidth Ω over a duration T has a limited number of degrees-of-freedom, namely $2\Omega T$, and therefore it can communicate at most this quantity of information. He also defined the information content of an equiprobable ensemble of N possible states as equal to $\log_2 N$.

- Norbert WIENER (1894-1964) unified information theory and Fourier analysis by deriving a series of relationships between the two. He invented "white noise analysis" of non-linear systems, and made the definitive contribution to modeling and describing the information content of stochastic processes known as *Time Series.*

- Dennis GABOR (1900-1979) crystallized Hartley's insight by formulating a general *Uncertainty Principle* for information, expressing the trade-off for resolution between bandwidth and time. (Signals that are well specified in frequency content must be poorly localized in time, and those that are well localized in time must be poorly specified in frequency content.) He formalized the "Information Diagram" to describe this fundamental trade-off, and derived the continuous family of functions which optimize (minimize) the conjoint uncertainty relation. In 1974 Gabor won the Nobel Prize in Physics for his work in Fourier optics, including the invention of holography.

- Claude SHANNON (together with Warren WEAVER) in 1949 wrote the definitive, classic, work in information theory: *Mathematical Theory of Communication.* Divided into separate treatments for continuous-time and discrete-time signals, systems, and channels, this book laid out all of the key concepts and relationships that define the field today. In particular, he proved the famous Source Coding Theorem and the Noisy Channel Coding Theorem, plus many other related results about channel capacity.

- S KUHLBACK and R A LIEBLER (1951) defined *relative entropy* (also called *cross entropy*, or *information for discrimination.*)

- E T JAYNES (since 1957) developed *maximum entropy* methods for inference, hypothesis-testing, and decision-making, based on the physics of statistical mechanics. Others have inquired whether these principles impose fundamental physical limits to computation itself.

- A N KOLMOGOROV in 1965 proposed that the *complexity* of a string of data can be defined by the length of the shortest binary program for computing the string. Thus the complexity of data is its *minimal description length,* and this specifies the ultimate compressibility of data. The "Kolmogorov complexity" K of a string is approximately equal to its Shannon entropy H, thereby unifying the theory of descriptive complexity and information theory.

2. Mathematical Foundations; Probability Rules; Bayes' Theorem

What are random variables? What is probability?

Random variables are variables that take on values determined by probability distributions. They may be discrete or continuous, in either their domain or their range. For example, a stream of ASCII encoded text characters in a transmitted message is a discrete random variable, with a known probability distribution for any given natural language. An analog speech signal represented by a voltage or sound pressure waveform as a function of time (perhaps with added noise), is a continuous random variable having a continuous probability density function.

Most of Information Theory involves probability distributions of random variables, and conjoint or conditional probabilities defined over ensembles of random variables. Indeed, the information content of a symbol or event is defined by its (im)probability. Classically, there are two different points of view about what probability actually means:

- *relative frequency:* sample the random variable a great many times and tally up the fraction of times that each of its different possible values occurs, to arrive at the probability of each.
- *degree-of-belief:* probability is the plausibility of a proposition or the likelihood that a particular state (or value of a random variable) might occur, even if its outcome can only be decided once (e.g. the outcome of a particular horse-race).

The first view, the "frequentist" or operationalist view, is the one that predominates in statistics and in information theory. However, by no means does it capture the full meaning of probability. For example, the proposition that `"The moon is made of green cheese"` is one which surely has a probability that we should be able to attach to it. We could assess its probability by degree-of-belief calculations which combine our prior knowledge about physics, geology, and dairy products. Yet the "frequentist" definition of probability could only assign a probability to this proposition by performing (say) a large number of repeated trips to the moon, and tallying up the fraction of trips on which the moon turned out to be a dairy product....

In either case, it seems sensible that the less probable an event is, the more information is gained by observing its occurence. (Surely discovering that the moon IS made of green cheese would be more "informative" than merely learning that it is made only of earth-like rocks.)

Probability Rules

Most of probability theory was laid down by theologians: Blaise PASCAL (1623-1662) who gave it the axiomatization that we accept today; and Thomas BAYES (1702-1761) who expressed one of its most important and widely-applied propositions relating conditional probabilities.

Probability Theory rests upon two rules:

Product Rule:

$$p(A,B) = \text{“joint probability of } \textit{both } A \textit{ and } B\text{”}$$
$$= p(A|B)p(B)$$

or equivalently,

$$= p(B|A)p(A)$$

Clearly, in case A and B are *independent* events, they are not conditionalized on each other and so

$$P(A|B) = P(A)$$
$$\text{and } P(B|A) = P(B),$$

in which case their joint probability is simply $P(A,B) = P(A)P(B)$.

$\boxed{\text{Sum Rule:}}$
If event A is conditionalized on a number of other events B, then the total probability of A is the sum of its joint probabilities with all B:

$$p(A) = \sum_B p(A, B) = \sum_B p(A|B)p(B)$$

From the Product Rule and the symmetry that $p(A, B) = p(B, A)$, it is clear that $p(A|B)p(B) = p(B|A)p(A)$. Bayes' Theorem then follows:

<u>Bayes' Rule:</u>

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}$$

The importance of Bayes Rule is that it allows us to reverse the conditionalizing of events, and to compute $p(B|A)$ from knowledge of $p(A|B)$, $p(A)$, and $p(B)$. Often these are expressed as *prior* and *posterior* probabilities, or as the conditionalizing of hypotheses upon data.

<u>Worked Example:</u>
Suppose that a dread disease affects 1/1000th of all people. If you actually have the disease, a test for it is positive 95% of the time, and negative 5% of the time. If you don't have the disease, the test is positive 5% of the time. We wish to know how to interpret test results.

Suppose you test positive for the disease. What is the likelihood that you actually have it?

We use the above rules, with the following substitutions of "data" D and "hypothesis" H instead of A and B:

D = data: the test is positive
H = hypothesis: you have the disease
$\bar{H}$ = the other hypothesis: you do not have the disease

Before acquiring the data, we know only that the *a priori* probability of having the disease is .001, which sets $p(H)$. This is called a *prior*. We also need to know $p(D)$.

From the Sum Rule, we can calculate that the *a priori* probability $p(D)$ of testing positive, whatever the truth may actually be, is:

$$p(D) = p(D|H)p(H) + p(D|\bar{H})p(\bar{H}) = (.95)(.001) + (.05)(.999) = .051$$

and from Bayes' Rule, we can conclude that the probability that you actually have the disease given that you tested positive for it, is much smaller than you may have thought:

$$p(H|D) = \frac{p(D|H)p(H)}{p(D)} = \frac{(.95)(.001)}{(.051)} = \boxed{0.019} \qquad \text{(less than 2\%)}.$$

This quantity is called the *posterior probability* because it is computed after the observation of data; it tells us how likely the hypothesis is, given what we have observed. (Note: it is an extremely common human fallacy to confound $p(H|D)$ with $p(D|H)$. In the example given, most people would react to the positive test result by concluding that the likelihood that they have the disease is .95, since that is the "hit rate" of the test. They confound $p(D|H) = .95$ with $p(H|D) = .019$, which is what actually matters.)

A nice feature of Bayes Theorem is that it provides a simple mechanism for repeatedly updating our assessment of the hypothesis as more data continues to arrive. We can apply the rule recursively, using the latest *posterior* as the new *prior* for interpreting the next set of data. In Artificial Intelligence, this feature is important because it allows the systematic and real-time construction of interpretations that can be updated continuously as more data arrive in a time series, such as a flow of images or spoken sounds that we wish to understand.

3. Entropies Defined, and Why They are Measures of Information

The information content I of a single event or message is defined as the base-2 logarithm of its probability p:

$$I = \log_2 p \tag{1}$$

and its *entropy* H is the negative of this. Entropy can be regarded intuitively as "uncertainty," or "disorder." To gain information is to lose uncertainty by the same amount, so I and H differ only in sign: $H = -I$. Entropy and information have units of *bits.*

Note that I as defined in Eqn (1) is never positive: it ranges between 0 and $-\infty$ as p varies from 1 to 0. It is the antithesis of uncertainty, which is always positive or 0.

No information is gained (no uncertainty is lost) by the appearance of an event or the receipt of a message that was completely certain anyway ($p = 1$, so $I = 0$). Intuitively, the more improbable an event is, the more informative it is; and so the monotonic behaviour of Eqn (1) seems appropriate. <u>But why the logarithm?</u>

The logarithmic measure is justified by the desire for information to be <u>additive</u>. We want the algebra of our measures to reflect the Rules of Probability. When independent packets of information arrive, we would like to say that the total information received is the sum of the individual pieces. But the probabilities of independent events <u>multiply</u> to give their combined probabilities, and so we must take logarithms in order for the joint probability of independent events or messages to contribute additively to the information gained.

This principle can also be understood in terms of the combinatorics of state spaces. Suppose we have two independent problems, one with n

possible solutions (or states) each having probability p_n, and the other with m possible solutions (or states) each having probability p_m. Then the number of combined states is mn, and each of these has probability $p_m p_n$. We would like to say that the information gained by specifying the solution to *both* problems is the *sum* of that gained from each one. This desired property is achieved:

$$I_{mn} = \log_2(p_m p_n) = \log_2 p_m + \log_2 p_n = I_m + I_n \tag{2}$$

A Note on Logarithms:

In information theory we often wish to compute the base-2 logarithms of quantities, but most calculators (and tools like xcalc) only offer Napierian (base 2.718...) and decimal (base 10) logarithms. So the following conversions are useful:

$$\log_2 X = 1.443 \log_e X = 3.322 \log_{10} X$$

Henceforward we will omit the subscript; base-2 is always presumed.

Intuitive Example of the Information Measure (Eqn 1):

Suppose I choose at random one of the 26 letters of the alphabet, and we play the game of "25 questions" in which you must determine which letter I have chosen. I will only answer 'yes' or 'no.' What is the minimum number of such questions that you must ask in order to guarantee finding the answer? (What form should such questions take? e.g., "Is it A?" "Is it B?" ...or is there some more intelligent way to solve this problem?)

The answer to a Yes/No question having equal probabilities conveys one bit worth of information. In the above example with equiprobable states, you never need to ask more than 5 (well-phrased!) questions to discover the answer, even though there are 26 possibilities. Appropriately, Eqn (1) tells us that the uncertainty removed as a result of the solving this problem is about -4.7 bits.

Entropy of Ensembles

We now move from considering the information content of a single event or message, to that of an *ensemble*. An ensemble is the set of outcomes of one or more random variables. The outcomes have probabilities attached to them. In general these probabilities are non-uniform, with event i having probability p_i, but they must sum to 1 because all possible outcomes are included; hence they form a probability distribution:

$$\sum_i p_i = 1 \tag{3}$$

The *entropy of an ensemble* is simply the average entropy of all the elements in it. We can compute their average entropy by weighting each of the $\log p_i$ contributions by its probability p_i:

$$H = -I = -\sum_i p_i \log p_i \tag{4}$$

Eqn (4) allows us to speak of the information content or the entropy of a random variable, from knowledge of the probability distribution that it obeys. *(Entropy does not depend upon the actual values taken by the random variable! – Only upon their relative probabilities.)*

Let us consider a random variable that takes on only two values, one with probability p and the other with probability $(1-p)$. Entropy is a concave function of this distribution, and equals 0 if $p = 0$ or $p = 1$:

Example of entropy as average uncertainty:

The various letters of the written English language have the following relative frequencies (probabilities), in descending order:

E	T	O	A	N	I	R	S	H	D	L	C ...
.105	.072	.066	.063	.059	.055	.054	.052	.047	.035	.029	.023 ...

If they had been equiprobable, the entropy of the ensemble would have been $\log_2(\frac{1}{26}) = 4.7$ bits. But their non-uniform probabilities imply that, for example, an `E` is nearly five times more likely than a `C`; surely this prior knowledge is a reduction in the uncertainty of this random variable. In fact, the distribution of English letters has an entropy of only 4.0 bits. This means that as few as only four 'Yes/No' questions are needed, in principle, to identify one of the 26 letters of the alphabet; not five.

How can this be true?

That is the subject matter of Shannon's SOURCE CODING THEOREM (so named because it uses the "statistics of the source," the *a priori* probabilities of the message generator, to construct an optimal code.)

Note the important assumption: that the "source statistics" are known!

Several further measures of entropy need to be defined, involving the marginal, joint, and conditional probabilities of random variables. Some key relationships will then emerge, that we can apply to the analysis of communication channels.

Notation: We use capital letters X and Y to name random variables, and lower case letters x and y to refer to their respective outcomes. These are drawn from particular sets A and B: $x \in \{a_1, a_2, ...a_J\}$, and $y \in \{b_1, b_2, ...b_J\}$. The probability of a particular outcome $p(x = a_i)$ is denoted p_i, with $0 \leq p_i \leq 1$ and $\Sigma_i p_i = 1$.

An *ensemble* is just a random variable X, whose entropy was defined in Eqn (4). A *joint ensemble* 'XY' is an ensemble whose outcomes are ordered pairs x, y with $x \in \{a_1, a_2, ...a_I\}$ and $y \in \{b_1, b_2, ...b_J\}$. The joint ensemble XY defines a probability distribution $p(x, y)$ over all possible joint outcomes x, y.

Marginal probability: From the Sum Rule, we can see that the probability of X taking on a particular value $x = a_i$ is the sum of the joint probabilities of this outcome for X and all possible outcomes for Y:

$$p(x = a_i) = \sum_y p(x = a_i, y)$$

We can simplify this notation to: $p(x) = \sum_y p(x, y)$

and similarly: $p(y) = \sum_x p(x, y)$

Conditional probability: From the Product Rule, we can easily see that the conditional probability that $x = a_i$, given that $y = b_j$, is:

$$p(x = a_i | y = b_j) = \frac{p(x = a_i, y = b_j)}{p(y = b_j)}$$

We can simplify this notation to: $p(x|y) = \dfrac{p(x, y)}{p(y)}$

and similarly: $p(y|x) = \dfrac{p(x, y)}{p(x)}$

It is now possible to define various entropy measures for joint ensembles:

$\boxed{\text{Joint entropy of } XY}$

$$H(X, Y) = \sum_{x,y} p(x, y) \log \frac{1}{p(x, y)} \tag{5}$$

(Note that in comparison with Eqn (4), we have replaced the '–' sign in front by taking the reciprocal of p inside the logarithm).

From this definition, it follows that joint entropy is <u>additive</u> if X and Y are independent random variables:

$$H(X,Y) = H(X) + H(Y) \quad \text{iff} \;\; p(x,y) = p(x)p(y)$$

Prove this.

$\boxed{\text{Conditional entropy of an ensemble } X\text{, given that } y = b_j}$

measures the uncertainty remaining about random variable X after specifying that random variable Y has taken on a particular value $y = b_j$. It is defined naturally as the entropy of the probability distribution $p(x|y = b_j)$:

$$H(X|y = b_j) = \sum_x p(x|y = b_j) \log \frac{1}{p(x|y = b_j)} \tag{6}$$

If we now consider the above quantity *averaged* over all possible outcomes that Y might have, each weighted by its probability $p(y)$, then we arrive at the...

$\boxed{\text{Conditional entropy of an ensemble } X\text{, given an ensemble } Y\text{:}}$

$$H(X|Y) = \sum_y p(y) \left[\sum_x p(x|y) \log \frac{1}{p(x|y)} \right] \tag{7}$$

and we know from the Sum Rule that if we move the $p(y)$ term from the outer summation over y, to inside the inner summation over x, the two probability terms combine and become just $p(x,y)$ summed over all x, y. Hence a simpler expression for this conditional entropy is:

$$H(X|Y) = \sum_{x,y} p(x,y) \log \frac{1}{p(x|y)} \tag{8}$$

This measures the average uncertainty that remains about X, when Y is known.

Chain Rule for Entropy

The joint entropy, conditional entropy, and marginal entropy for two ensembles X and Y are related by:

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \tag{9}$$

It should seem natural and intuitive that the joint entropy of a pair of random variables is the entropy of one plus the conditional entropy of the other (the uncertainty that it adds once its dependence on the first one has been discounted by conditionalizing on it). You can derive the Chain Rule from the earlier definitions of these three entropies.

Corollary to the Chain Rule:

If we have three random variables X, Y, Z, the conditionalizing of the joint distribution of any two of them, upon the third, is also expressed by a Chain Rule:

$$H(X,Y|Z) = H(X|Z) + H(Y|X,Z) \tag{10}$$

"Independence Bound on Entropy"

A consequence of the Chain Rule for Entropy is that if we have many different random variables $X_1, X_2, ..., X_n$, then the sum of all their individual entropies is an upper bound on their joint entropy:

$$H(X_1, X_2, ..., X_n) \leq \sum_{i=1}^{n} H(X_i) \tag{11}$$

Their joint entropy only reaches this upper bound if all of the random variables are independent.

Mutual Information between X and Y

The *mutual information* between two random variables measures the amount of information that one conveys about the other. Equivalently, it measures the average reduction in uncertainty about X that results from learning about Y. It is defined:

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{12}$$

Clearly X says as much about Y as Y says about X. Note that in case X and Y are independent random variables, then the numerator inside the logarithm equals the denominator. Then the log term vanishes, and the mutual information equals zero, as one should expect.

Non-negativity: mutual information is always ≥ 0. In the event that the two random variables are perfectly correlated, then their mutual information is the entropy of either one alone. (Another way to say this is: $I(X;X) = H(X)$: the mutual information of a random variable with itself is just its entropy. For this reason, the entropy $H(X)$ of a random variable X is sometimes referred to as its *self-information.*)

These properties are reflected in three equivalent definitions for the mutual information between X and Y:

$$I(X;Y) = H(X) - H(X|Y) \tag{13}$$

$$I(X;Y) = H(Y) - H(Y|X) = I(Y;X) \tag{14}$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{15}$$

In a sense the mutual information $I(X;Y)$ is the <u>intersection</u> between $H(X)$ and $H(Y)$, since it represents their statistical dependence. In the Venn diagram given at the top of page 19, the portion of $H(X)$ that does not lie within $I(X;Y)$ is just $H(X|Y)$. The portion of $H(Y)$ that does not lie within $I(X;Y)$ is just $H(Y|X)$.

$\boxed{\text{Distance } D(X,Y) \text{ between } X \text{ and } Y}$

The amount by which the joint entropy of two random variables exceeds their mutual information is a measure of the "*distance*" between them:

$$D(X,Y) = H(X,Y) - I(X;Y) \tag{16}$$

Note that this quantity satisfies the standard axioms for a distance: $D(X,Y) \geq 0$, $D(X,X) = 0$, $D(X,Y) = D(Y,X)$, and $D(X,Z) \leq D(X,Y) + D(Y,Z)$.

$\boxed{\text{Relative entropy, or Kullback-Leibler distance}}$

Another important measure of the "distance" between two random variables, although it does not satisfy the above axioms for a distance metric, is the *relative entropy* or *Kullback-Leibler distance.* It is also called the *information for discrimination.* If $p(x)$ and $q(x)$ are two probability distributions defined over the same set of outcomes x, then their relative entropy is:

$$D_{KL}(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \tag{17}$$

Note that $D_{KL}(p\|q) \geq 0$, and in case $p(x) = q(x)$ then their distance $D_{KL}(p\|q) = 0$, as one might hope. However, this metric is not strictly a "distance," since in general it lacks symmetry: $D_{KL}(p\|q) \neq D_{KL}(q\|p)$.

The relative entropy $D_{KL}(p\|q)$ is a measure of the "inefficiency" of assuming that a distribution is $q(x)$ when in fact it is $p(x)$. If we have an optimal code for the distribution $p(x)$ (meaning that we use on average $H(p(x))$ bits, its entropy, to describe it), then the number of <u>additional</u> bits that we would need to use if we instead described $p(x)$ using an optimal code for $q(x)$, would be their relative entropy $D_{KL}(p\|q)$.

Venn Diagram: Relationship between entropy and mutual information.

Fano's Inequality

We know that conditioning reduces entropy: $H(X|Y) \leq H(X)$. It is clear that if X and Y are perfectly correlated, then their conditional entropy is 0. It should also be clear that if X is any deterministic function of Y, then again, there remains no uncertainty about X once Y is known and so their conditional entropy $H(X|Y) = 0$.

Fano's Inequality relates the probability of error P_e in guessing X from knowledge of Y to their conditional entropy $H(X|Y)$, when the number of possible outcomes is $|\mathcal{A}|$ (e.g. the length of a symbol alphabet):

$$P_e \geq \frac{H(X|Y) - 1}{\log |\mathcal{A}|} \tag{18}$$

The lower bound on P_e is a linearly increasing function of $H(X|Y)$.

The "Data Processing Inequality"

If random variables X, Y, and Z form a Markov chain (i.e. the conditional distribution of Z depends only on Y and is independent of X), which is normally denoted as $X \to Y \to Z$, then the mutual information must be monotonically decreasing over steps along the chain:

$$I(X;Y) \geq I(X;Z) \tag{19}$$

We turn now to applying these measures and relationships to the study of communications channels. (The following material is from McAuley.)

4 Information and Entropy

Consider observing the discrete outputs of some source which is emitting symbols at periodic time intervals[2]. This can be modelled as a discrete random variable S, which takes on *symbols* from an finite *alphabet*:

$$\mathcal{S} = \{s_1, s_2, \ldots, s_n\} \tag{37}$$

with associated probabilities:

$$P(S = s_i) = p_i \ , \ i = 1, 2, \ldots n \tag{38}$$

with the necessary condition:

$$\sum_{i=1}^{n} p_i = 1 \tag{39}$$

If the symbols emitted are statistically independent, we describe the source as a *discrete memoryless source*. We wish to consider how much information is conveyed by the emission of a symbol. If for some k, $p_k = 1$ (i.e $p_i = 0$ for $i \neq k$), then no information is conveyed – we knew what was going to happen. Information is only conveyed when there is some uncertainty about what could be emitted.

What properties would we desire of a measure of information. We can consider both the information of a symbol $I(s_i)$

[2]The following can be extended to symbols of arbitrary duration with only slight complication to the treatment.

1. $I(s_i) = 0$ if $p_i = 1$,
   certainty gives no information,

2. $I(s_i) \geq 0$ for $0 \leq p_i \leq 1$,
   an event cannot cause us to lose information,

3. $I(s_i) > I(s_j)$ if $p_i < p_j$,
   the less probable an event the more we learn from it,

4. $I(s_i s_j) = I(s_i) + I(s_j)$ if the events are statistically independent
   the information from two random events is the sum of their individual information

and the average information per symbol $H(s_1, s_2, \ldots, s_n)$ (sometimes written $H(\mathcal{S})$), or $H(p_1, p_2, \ldots, p_n)$:

5. H should be continuous and symmetric in the p_i,

6. if all p_i are equal (i.e. $p_i = 1/n$), then H should be monotonic increasing function of n.

7. decomposition:

$$\begin{aligned} & H(p_1, \ldots p_k, p_{k+1}, \ldots s_n) = \\ & \qquad H(P_k, Q_k) + P_k H(p_1/P_k, \ldots p_k/P_k) + Q_k H(p_{k+1}/Q_k, \ldots p_n/Q_k) \end{aligned} \tag{40}$$

$$\begin{aligned} \text{with: } P_k &= \sum_{i=1}^{k} p_i \\ \text{and: } Q_k &= \sum_{i=k+1}^{n} p_i \end{aligned}$$

This last property allows us to decompose events into successive choices. For example consider language {A, B, C, D} with associated probabilities $\{1/2, 1/4, 1/8, 1/8\}$:

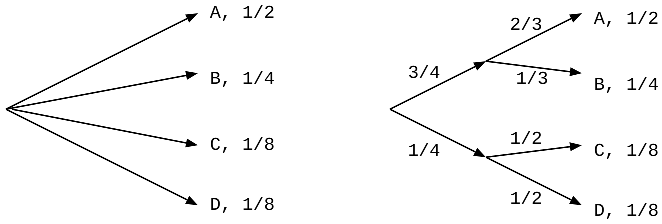


Figure 14: Decomposition of four choices into two successive choices of two.

The average information content per symbol then is simply the value of the expectation of the information per symbol or *entropy*:

$$H(\mathcal{S}) = \sum_{i=1}^{n} p_i I(s_i) \tag{41}$$

The properties of I and H can be shown to require:

$$\begin{aligned}
I(s_i) &= K\log(\frac{1}{p_i}) \\
H(\mathcal{S}) &= K\sum_{i=1}^{n} p_i \log\left(\frac{1}{p_i}\right) \\
&= -K\sum_{i=1}^{n} p_i \log(p_i)
\end{aligned}$$

(See Shannon and Weaver, Appendix 2 for derivation.)

The base of the logarithm is arbitrary, as is the factor K. However consider a source of two symbols $\{s_1, s_2\}$ with probabilities $\{p, 1-p\}$ (see fig 15). It would seem natural to choose base 2 and $K = 1$ and view the amount of information obtained when one of two equiprobable events occurs to be 1. This unit is called the *bit* – it is important to understand that in this case the *bit* is a measure of *information* not a contraction for *b*inary dig*it*.

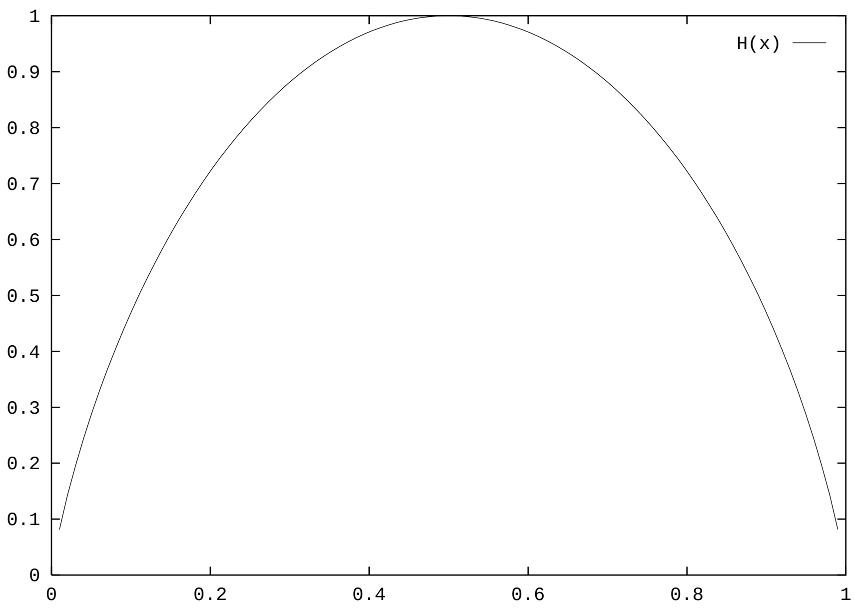


Figure 15: Entropy in the case of two symbols with probabilities p and $1-p$.

The entropy of the decompositions example in figure 14 can then be calculated. First without decomposition:

$$\begin{aligned}
H &= 2 \times \frac{1}{8}\log 8 + 1 \times \frac{1}{4}\log 4 + 1 \times \frac{1}{2}\log 2 \\
&= \frac{7}{4}
\end{aligned}$$

$$(42)$$

then by decomposition:

$$\begin{aligned} H &= \frac{3}{4}\log\frac{4}{3} + \frac{1}{4}\log 4 + \frac{3}{4}\left\{\frac{2}{3}\log\frac{3}{2} + \frac{1}{3}\log 3\right\} + \frac{1}{4}\left\{\frac{1}{2}\log 2 + \frac{1}{1}\log 2\right\} \\ &= \frac{7}{4} \end{aligned}$$

Observe that the obvious representation of the four events uses two binary digits, but they only convey 7/4 bits of information.

4.1 Properties

We note some properties of the entropy $H(\mathcal{S})$ of a source.

1. For an alphabet $\mathcal{S}$ of N symbols,

$$0 \leq H(\mathcal{S}) \leq log_2 N$$

   Equality with the lower limit is achieved when one of the associated probabilities $p_k = 1$ for some k

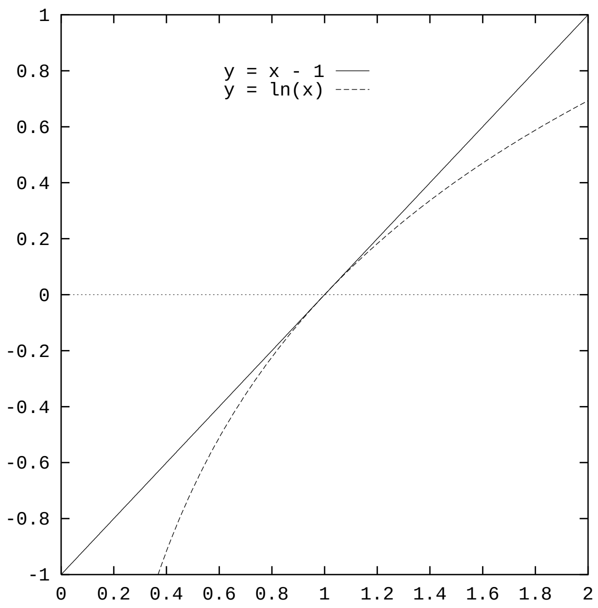


Figure 16: Functions $x - 1$ and $\ln(x)$

   For the upper limit we make use of the inequality:

$$\ln(x) < x - 1 \tag{43}$$

and consider two possible probability distributions p_i, q_i on an alphabet $\mathcal{S}$:

$$\begin{aligned}
\sum_{i=1}^{N} p_i \log_2 \left(\frac{q_i}{p_i}\right) &= \frac{1}{\log_2 e} \sum_{i=1}^{N} p_i \ln \left(\frac{q_i}{p_i}\right) \\
&\leq \frac{1}{\log_2 e} \sum_{i=1}^{N} p_i \left(\frac{q_i}{p_i} - 1\right) \\
&\leq \frac{1}{\log_2 e} \sum_{i=1}^{N} (q_i - p_i) \\
&\leq \frac{1}{\log_2 e} \left(\sum_{i=1}^{N} q_i - \sum_{i=1}^{N} p_i\right) \\
\sum_{i=1}^{N} p_i \log_2 \left(\frac{q_i}{p_i}\right) &\leq 0 \qquad (44)
\end{aligned}$$

This provides equality only if $p_i = q_i$. Suppose $\forall i$, $q_i = 1/N$, the entropy is then:

$$\begin{aligned}
\sum_{i=1}^{N} q_i \log_2 \left(\frac{1}{q_i}\right) &= \log_2 N \\
\sum_{i=1}^{N} p_i \log_2 \left(\frac{1}{p_i}\right) &\leq \log_2 N \\
H(\mathcal{S}) &\leq \log_2 N
\end{aligned}$$

2. Often we wish to consider blocks of symbols from the alphabet $\mathcal{S}$; for blocks of length n, we derive a new alphabet of symbol blocks $\mathcal{S}^n$. If the occurrence of symbols are independent then we obtain:

$$H(\mathcal{S}^n) = nH(\mathcal{S})$$

For our example in figure 14, we obtain the new alphabet $\{\sigma_0, \sigma_1, \ldots, \sigma_f\}$ of pairs $\{s_i s_j\}$ with associated probabilities:

	s_0	s_1	s_2	s_3
s_0	$\sigma_0 1/4$	$\sigma_1 1/8$	$\sigma_2 1/16$	$\sigma_3 1/16$
s_1	$\sigma_4 1/8$	$\sigma_5 1/16$	$\sigma_0 1/32$	$\sigma_7 1/32$
s_2	$\sigma_8 1/16$	$\sigma_9 1/32$	$\sigma_a 1/64$	$\sigma_b 1/64$
s_3	$\sigma_c 1/16$	$\sigma_d 1/32$	$\sigma_e 1/64$	$\sigma_f 1/64$

Calculation for this extended language will show an entropy of 7/2 bits.

3. If symbols are generated at frequency f_s, we can define the information rate, or entropy per second (in units of bits per second), as Hf_s.

4.2 Information sources with memory

We also wish to consider sources with memory, so we also consider Markov processes. Our four event process (a symbol is generated on each edge) is shown graphically together with a two state Markov process for the alphabet {A, B, C, D, E} in figure 17. We can then solve for the state occupancy using flow equations (this example is trivial).

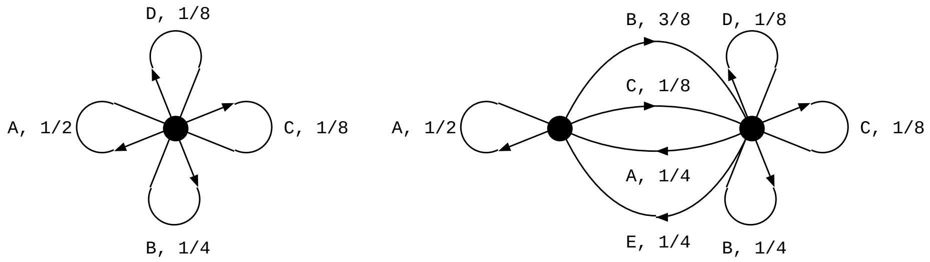


Figure 17: Graphs representing memoryless source and two state Markov process

In general then we can define for a finite state process with states $\{S_1, S_2, \ldots S_n\}$, with transition probabilities $p_i(j)$ being the probability of moving from state S_i to state S_j (with the emission of some symbol). First we can define the entropy of each state in the normal manner:

$$H_i = -\sum_j p_i(j) \log_2 p_i(j)$$

and then the entropy of the system to be the sum of these individual state entropy values weighted with the state occupancy (calculated from the flow equations):

$$\begin{aligned} H &= \sum_i P_i H_i \\ &= -\sum_i \sum_j P_i p_i(j) \log p_i(j) \end{aligned} \tag{45}$$

Clearly for a single state, we have the entropy of the memoryless source.

4.3 The Source Coding theorem

Often we wish to efficiently represent the symbols generated by some source. We shall consider encoding the symbols as binary digits.

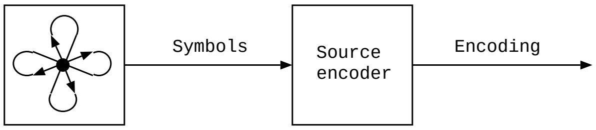


Figure 18: Discrete memoryless source and encoder

4.3.1 Fixed length codes

Consider encoding the N symbols $\{s_i\}$, entropy H, as a fixed length (R) block binary digits. To ensure we can decode these symbols we need:

$$R = \begin{cases} log_2(N) & \text{N a power of 2} \\ \lfloor log_2(N) \rfloor + 1 & \text{otherwise} \end{cases}$$

where $\lfloor X \rfloor$ is the largest integer less than X. The *code rate* is then R bits per symbol, and as we know $H \leq log_2(N)$ then $H \leq R$. The efficiency of the coding η is given by:

$$\eta = \frac{H}{R}$$

When the N symbols are equiprobable, and N is a power of two, $\eta = 1$ and $H = R$. Note that if we try to achieve $R < H$ in this case we must allocate at least one encoding to more than one symbol – this means that we are incapable of uniquely decoding.

Still with equiprobable symbols, but when N is not a power of two, this coding is inefficient; to try to overcome this we consider sequences of symbols of length J and consider encoding each possible sequence of length J as a block of binary digits, then we obtain:

$$R = \frac{\lfloor J \log_2 N \rfloor + 1}{J}$$

where R is now the average number of bits per symbol. Note that as J gets large, $\eta \to 1$.

4.3.2 Variable length codes

In general we do not have equiprobable symbols, and we would hope to achieve some more compressed form of encoding by use of variable length codes – an example of such an encoding is Morse code dating from the days of telegraphs. We consider again our simple four symbol alphabet and some possible variable length codes:

X	P(X)	Code 1	Code 2	Code 3
A	1/2	1	0	0
B	1/4	00	10	01
C	1/8	01	110	011
D	1/8	10	111	111

We consider each code in turn:

1. Using this encoding, we find that presented with a sequence like 1001, we do not know whether to decode as ABA or DC. This makes such a code unsatisfactory. Further, in general, even a code where such an ambiguity could be resolved uniquely by looking at bits further ahead in the stream (and backtracking) is unsatisfactory.

   Observe that for this code, the coding rate, or average number of bits per symbol, is given by:

   $$\begin{aligned}\textstyle\sum_i s_i b_i &= \tfrac{1}{2} \times 1 + \tfrac{1}{4} \times 2 + 2 \times \tfrac{1}{8} \times 2 \\ &= \tfrac{3}{2}\end{aligned}$$

   which is less than the entropy.

2. This code is uniquely de-codable; further this code is interesting in that we can decode *instantaneously* – that is no backtracking is required; once we have the bits of the encoded symbol we can decode without waiting for more. Further this also satisfies the *prefix* condition, that is there is no code word which is prefix (i.e. same bit pattern) of a longer code word. In this case the coding rate is equal to the entropy.

3. While this is de-codable (and coding rate is equal to the entropy again), observe that it does not have the prefix property and is not an instantaneous code.

Shannon's first theorem is the *source-coding theorem* which is:

> For a discrete memoryless source with finite entropy H; for any (positive) ϵ it is possible to encode the symbols at an average rate R, such that:
>
> $$R = H + \epsilon$$

(For proof see Shannon & Weaver.) This is also sometimes called the noiseless coding theorem as it deals with coding without consideration of noise processes (i.e. bit corruption etc).

The entropy function then represents a fundamental limit on the number of bits on average required to represent the symbols of the source.

4.3.3 Prefix codes

We have already mentioned the prefix property; we find that for a prefix code to exist, it must satisfy the *Kraft-McMillan* inequality. That is, necessary and sufficient condition for a code having binary code words with lengths $n_1 \leq n_2 \leq \cdots \leq n_N$ to satisfy the prefix condition is:

$$\sum_{i=1}^{N} \frac{1}{2^{n_i}} \leq 1$$

We consider a binary tree of depth n_N. We choose a node at depth n_1 as a code word; to ensure this is not then the prefix to anything we prune the tree at this point, removing $2^{n_N - n_1}$ nodes in the process. We repeat this for each n_i. Then as the number of nodes we have thrown away is less than the total number:

$$\begin{aligned} \sum_{i=1}^{N} 2^{n_N - n_i} &\leq 2^{n_N} \\ \sum_{i=1}^{N} \frac{1}{2^{n_i}} \leq 1 \end{aligned}$$

From this we can deduce a corollary of the source coding theorem:

> For a discrete memoryless source with finite entropy H; we can construct a prefix code that has a rate R such that:
>
> $$H \leq R < H + 1$$

We can establish these bounds from the Kraft-McMillan inequality. Consider the lower bound:

$$\begin{aligned} H - R &= \sum_{i=1}^{N} p_i \log \frac{1}{p_i} - \sum_{i=1}^{N} p_i n_i \\ &= \sum_{i=1}^{N} p_i \log \frac{2^{-n_i}}{p_i} \\ 2^{n_N - n_i} &\leq 2^{n_N} \end{aligned}$$

Using $\ln x \leq x - 1$ again:

$$\begin{aligned} H - R &\leq \log_2(e) \sum_{i=1}^{N} p_i \left\{ \frac{2^{-n_i}}{p_i} - 1 \right\} \\ &\leq log_2(e) \sum_{i=1}^{N} \{2^{-n_i}\} - 1 \\ &\leq 0 \end{aligned}$$

The upper bound is derived from choosing n_i such that $2^{-n_i} \leq p_i < 2^{-n_i + 1}$. Summing $2^{-n_i} \leq p_i$ with respect to i we again obtain the Kraft-McMillan inequality showing that such a prefix code exists; then rewriting $p_i < 2^{-n_i+1}$ as:

$$n_i = 1 - \log p_i$$

we multiply by p_i and sum with respect to i we obtain the upper bound.

Together with multi-symbol coding using prefix codes, we can then generate a prefix code with a rate arbitrarily close to the entropy.

4.4 Discrete Memoryless Channel

We have considered the discrete source, now we consider a channel through which we wish to pass symbols generated by such a source by some appropriate encoding mechanism; we also introduce the idea of noise into the system – that is we consider the channel to modify the input coding and possibly generate some modified version.

We should distinguish between systematic modification of the encoded symbols, i.e. distortion, and noise. Distortion is when an input code always results in the the same output code; this process can clearly be reversed. Noise on the other hand introduces the element of randomness into the resulting output code.

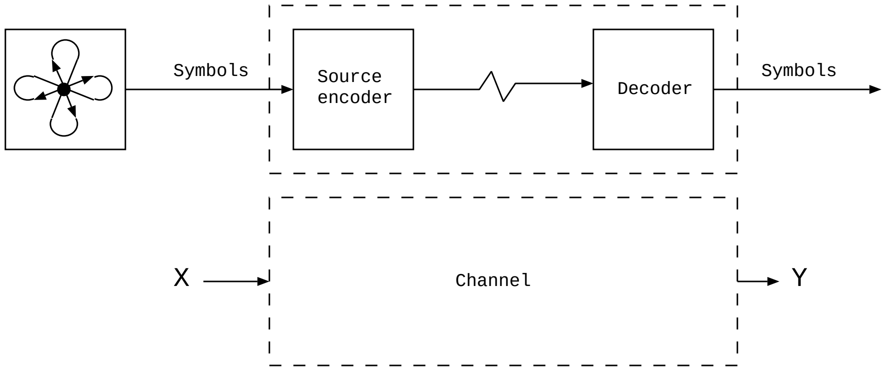


Figure 19: Coding and decoding of symbols for transfer over a channel.

We consider an input alphabet $\mathcal{X} = \{x_1, \ldots, x_J\}$ and output alphabet $\mathcal{Y} = \{y_1, \ldots, y_K\}$ and random variables X and Y which range over these alphabets. Note that J and K need not be the same – for example we may have the binary input alphabet $\{0, 1\}$ and the output alphabet $\{0, 1, \perp\}$, where $\perp$ represents the decoder identifying some error. The discrete memoryless channel can then be represented as a set of transition probabilities:

$$p(y_k|x_j) = P(Y = y_k|X = x_j)$$

That is the probability that if x_j is injected into the channel, y_k is emitted; $p(y_k|x_j)$ is the conditional probability. This can be written as the *channel matrix*:

$$[P] = \begin{pmatrix} p(y_1|x_1) & p(y_2|x_1) & \ldots & p(y_K|x_1) \\ p(y_1|x_2) & p(y_2|x_2) & \ldots & p(y_K|x_2) \\ \vdots & \vdots & \ddots & \vdots \\ p(y_1|x_J) & p(y_2|x_J) & \ldots & p(y_K|x_J) \end{pmatrix}$$

Note that we have the property that for every input symbol, we will get something out:

$$\sum_{k=1}^{K} p(y_k|x_j) = 1$$

Next we take the output of a discrete memoryless source as the input to a channel. So we have associated with the input alphabet of the channel the probability distribution of output from a memoryless source $\{p(x_j), j = 1, 2, \ldots J\}$. We then obtain the joint probability distribution of the random variables X and Y:

$$
\begin{aligned}
p(x_j, y_k) &= P(X = x_k, Y = y_k) \\
&= p(y_k|x_j)p(x_j)
\end{aligned}
$$

We can then find the marginal probability distribution of Y, that is the probability of output symbol y_k appearing:

$$
\begin{aligned}
p(y_k) &= P(Y = y_k) \\
&= \sum_{j=1}^{J} p(y_k|x_j)p(x_j)
\end{aligned}
$$

If we have $J = K$, we often identify each output symbol as being the desired result of some input symbol. Or we may select some subset of output symbols, for example in the input $\{0, 1\}$ and output $\{0, 1, \perp\}$.We then define the average probability of symbol error as:

$$
\begin{aligned}
P_e &= \sum_{k=1,k\neq j}^{K} P(Y = y_k|X \neq x_k) \\
&= \sum_{k=1}^{K} \sum_{j=1,j\neq k}^{J} p(y_k|x_j)p(x_j) \qquad (46)
\end{aligned}
$$

and correspondingly, the average probability of correct reception as $1 - P_e$.

4.4.1 Binary symmetric channel

The binary symmetric channel has two input and output symbols (usually written $\{0, 1\}$) and a common probability, p, of "incorrect" decoding of an input at the output; this could be a simplistic model of a communications link, figure 20a.

However, to understand the averaging property of the error rate P_e described above, consider the figure 20b, where we have 10^6 symbols, of which the first has a probability of being received in error (of 0.1), and the remainder are always received perfectly. Then observing that most of the terms in the sum on the right of equation 46 are zero:

$$
\begin{aligned}
P_e &= p(y_1|x_0)p(x_0) \\
&= 0.1 \times 10^{-6} \\
&= 10^{-7} \qquad (47)
\end{aligned}
$$

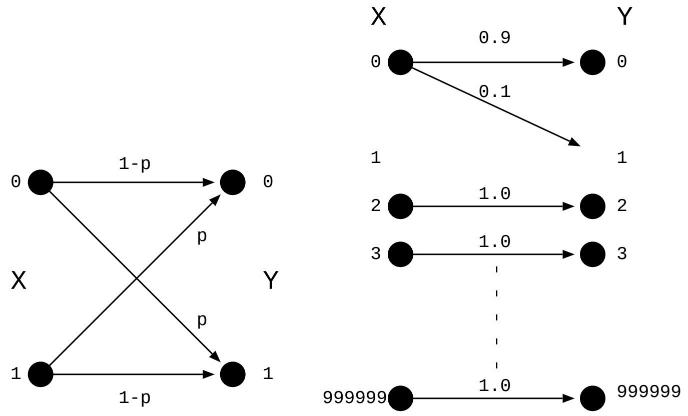


Figure 20: a) Binary symmetric channel, b) Tale of the unlucky symbol

4.5 Mutual information and entropy

Extending the ideas of information and entropy for the discrete source, we can now consider the information about X obtained by observing the value of the Y. We define the entropy after observing $Y = y_k$:

$$H(\mathcal{X}|Y = y_k) = \sum_{j=1}^{J} p(x_j|y_k) \log\left(\frac{1}{p(x_j|y_k)}\right)$$

this is a random variable, so we can take the average again:

$$\begin{aligned} H(\mathcal{X}|\mathcal{Y}) &= \sum_{k=1}^{K} H(\mathcal{X}|Y = y_k)p(y_k) \\ &= \sum_{k=1}^{K}\sum_{j=1}^{J} p(x_j|y_k) \log\left(\frac{1}{p(x_j|y_k)}\right) p(y_k) \\ &= \sum_{k=1}^{K}\sum_{j=1}^{J} p(x_j, y_k) \log\left(\frac{1}{p(x_j|y_k)}\right) \end{aligned} \quad (48)$$

$H(\mathcal{X}|\mathcal{Y})$ is the *conditional entropy*. We then write the *mutual information* of the channel:

$$I(\mathcal{X};\mathcal{Y}) = H(\mathcal{X}) - H(\mathcal{X}|\mathcal{Y})$$

This provides us with a measure of the amount of information or uncertainty removed about X after observing the output Y of a channel fed by X. The mutual information tells us something about the channel.

An alternative viewpoint is to think about the channel together with a correction device fed with information from observations of both the input and output of the channel, e.g. figure 21. One might ask how much information must be

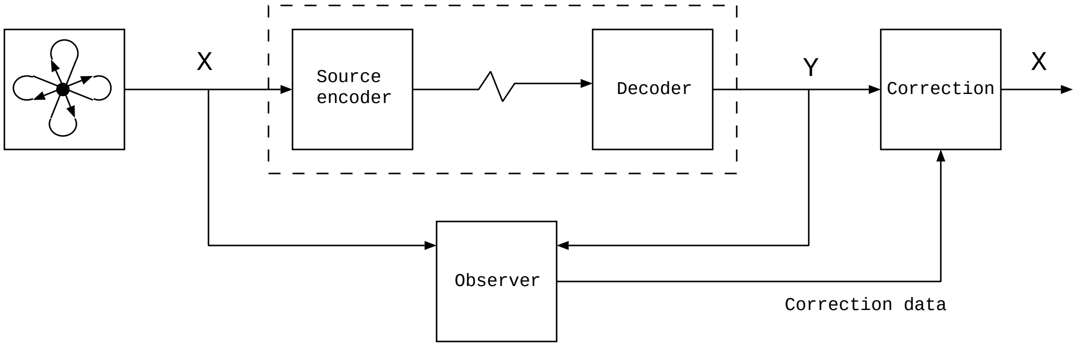


Figure 21: Correction system

passed along the correction channel to reconstruct the input; this turns out to be $H(X|Y)$.

We can further rewrite the entropy function $H(\mathcal{X})$ as a sum over the joint probability distribution:

$$\begin{aligned}
H(\mathcal{X}) &= -\sum_{j=1}^{J} p(x_j)\log(p(x_j)) \times 1 \\
&= -\sum_{j=1}^{J} p(x_j)\log(p(x_j)) \times \sum_{k=1}^{K} p(y_k|x_j) \\
&= -\sum_{j=1}^{J}\sum_{k=1}^{K} p(y_k|x_j)p(x_j)\log(p(x_j)) \\
&= -\sum_{j=1}^{J}\sum_{k=1}^{K} p(x_j, y_k)\log(p(x_j))
\end{aligned}$$

Hence we obtain an expression for the mutual information:

$$I(\mathcal{X};\mathcal{Y}) = \sum_{j=1}^{J}\sum_{k=1}^{K} p(x_j, y_k)\log\left(\frac{p(x_j|y_k)}{p(x_j)}\right)$$

We can deduce various properties of the mutual information:

1. $I(\mathcal{X};\mathcal{Y}) \geq 0$.
   To show this, we note that $p(x_j|y_k)p(y_k) = p(x_j, y_k)$ and substitute this in equation 4.5.

$$\begin{aligned}
I(\mathcal{X};\mathcal{Y}) &= \sum_{k=1}^{K}\sum_{j=1}^{J} p(x_j, y_k)\log\left(\frac{p(x_j, y_k)}{p(x_j)p(y_k)}\right) \\
&\geq 0 \qquad (49)
\end{aligned}$$

   by use of the inequality we established previously in equation 44.

2. $I(\mathcal{X};\mathcal{Y}) = 0$ if X and Y are statistically independent.
   If X and Y are independent, $p(x_j, y_k) = p(x_j)p(y_k)$, hence the log term becomes zero.

3. I is symmetric, $I(\mathcal{X};\mathcal{Y}) = I(\mathcal{Y};\mathcal{X})$.
   Using $p(x_j|y_k)p(y_k) = p(y_k|x_j)p(x_j)$, we obtain:

$$\begin{aligned} I(\mathcal{X};\mathcal{Y}) &= \sum_{j=1}^{J}\sum_{k=1}^{K} p(x_j, y_k) \log\left(\frac{p(x_j|y_k)}{p(x_j)}\right) \\ &= \sum_{j=1}^{J}\sum_{k=1}^{K} p(x_j, y_k) \log\left(\frac{p(y_k|x_j)}{p(y_k)}\right) \\ &= I(\mathcal{Y};\mathcal{X}) \end{aligned} \tag{50}$$

4. The preceding leads to the obvious symmetric definition for the mutual information in terms of the entropy of the output:

$$I(\mathcal{X};\mathcal{Y}) = H(\mathcal{Y}) - H(\mathcal{Y}|\mathcal{X})$$

5. We define the joint entropy of two random variables in the obvious manner, based on the joint probability distribution:

$$H(\mathcal{X},\mathcal{Y}) = -\sum_{j=1}^{J}\sum_{k=1}^{K} p(x_j, y_k) \log(p(x_j, y_k))$$

   The mutual information is the more naturally written:

$$I(\mathcal{X};\mathcal{Y}) = H(\mathcal{X}) + H(\mathcal{Y}) - H(\mathcal{X},\mathcal{Y})$$

4.5.1 Binary channel

Consider the conditional entropy and mutual information for the binary symmetric channel. The input source has alphabet $\mathcal{X} = \{0, 1\}$ and associated probabilities $\{1/2, 1/2\}$ and the channel matrix is:

$$\begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}$$

Then the entropy, conditional entropy and mutual information are given by:

$$\begin{aligned} H(\mathcal{X}) &= 1 \\ H(\mathcal{X}|\mathcal{Y}) &= -p\log(p) - (1-p)\log(1-p) \\ I(\mathcal{X};\mathcal{Y}) &= 1 + \text{p}\log(p) + (1-p)\log(1-p) \end{aligned}$$

Figure 22a, shows the capacity of the channel against transition probability. Note that the capacity of the channel drops to zero when the transition probability is 1/2, and is maximized when the transition probability is 0; or 1 – if we reliably transpose the symbols on the wire we also get the maximum amount of information through! Figure 22b shows the effect on mutual information for asymmetric input alphabets.

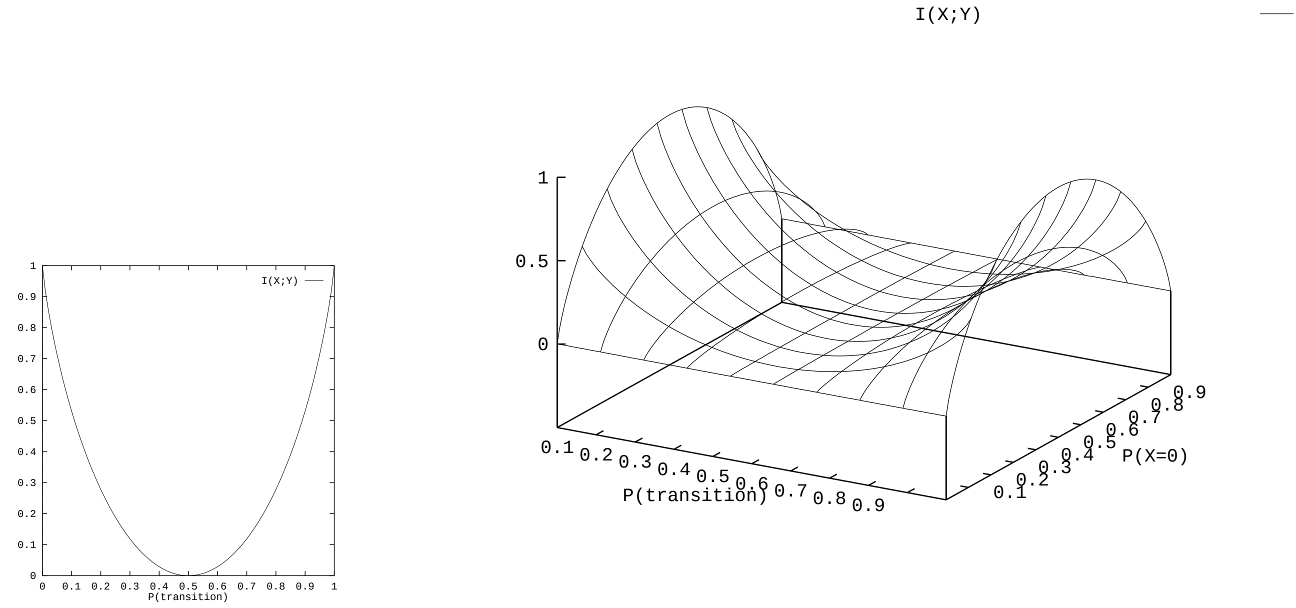


Figure 22: Capacity of binary channel a) symmetric, b) asymmetric

4.6 Channel capacity

We wish to define the capacity of a channel, using the model of a free input alphabet and dependent output alphabet (given by the channel matrix). We note that for a given channel, if we wish to maximize the mutual information, we must perform a maximization over all probability distributions for the input alphabet X. We define the *channel capacity*, demoted by C, as:

$$C = \max_{\{p(x_j)\}} I(\mathcal{X};\mathcal{Y}) \tag{51}$$

When we achieve this rate we describe the source as being matched to the channel.

4.7 Channel coding

To overcome the problem of noise in the system, we might consider adding redundancy during the encoding process to overcome possible errors. The examples that are used here are restricted to sources which would naturally be encoded in a noiseless environment as fixed size block codes – i.e. a source alphabet $\mathcal{X}$, which has 2^n equiprobable symbols; however, the discussion applies to more general sources and variable length coding schemes.

One particular aspect to be considered in real uses of channel coding is that many sources which we are interested in encoding for transmissions have a significant amount of redundancy already. Consider sending a piece of syntactically correct and semantically meaningful English or computer program text through a channel which randomly corrupted on average 1 in 10 characters (such as might be introduced by transmission across a rather sickly Telex system). e.g.:

1. Bring reinforcements, we're going to advance
2. It's easy to recognise speech

Reconstruction from the following due to corruption of 1 in 10 characters would be comparatively straight forward:

1. Brizg reinforce ents, we're going to advance
2. It's easy mo recognise speech

However, while the redundancy of this source protects against such random character error, consider the error due to a human miss-hearing:

1. Bring three and fourpence, we're going to a dance.
2. It's easy to wreck a nice peach.

The coding needs to consider the error characteristics of the channel and decoder, and try to achieve a significant "distance" between plausible encoded messages.

4.7.1 Repetition Codes

On of the simplest codes is a *repetition code*. We take a binary symmetric channel with a transition probability p; this gives a channel capacity $C = 1 + p\log(p) + (1-p)\log(1-p)$. The natural binary encoding for this is then $\{0, 1\}$ – the repetition code will repeat these digits an odd number of times and perform majority voting.

Hence we transmit $n = 2m+1$ bits per symbol, and will obtain an error if $m+1$ or more bits are received in error, that is:

$$P_e = \sum_{i=m+1}^{2m+1} \binom{2m+1}{i} p^i (1-p)^{2m+1-i}$$

Considering a transition probability of 0.01. The channel capacity as given by equation 51 is $C = 0.9192$ (figure 23a). The code rate of the repetition technique against the residual probability of error is demonstrated in figure 23b.

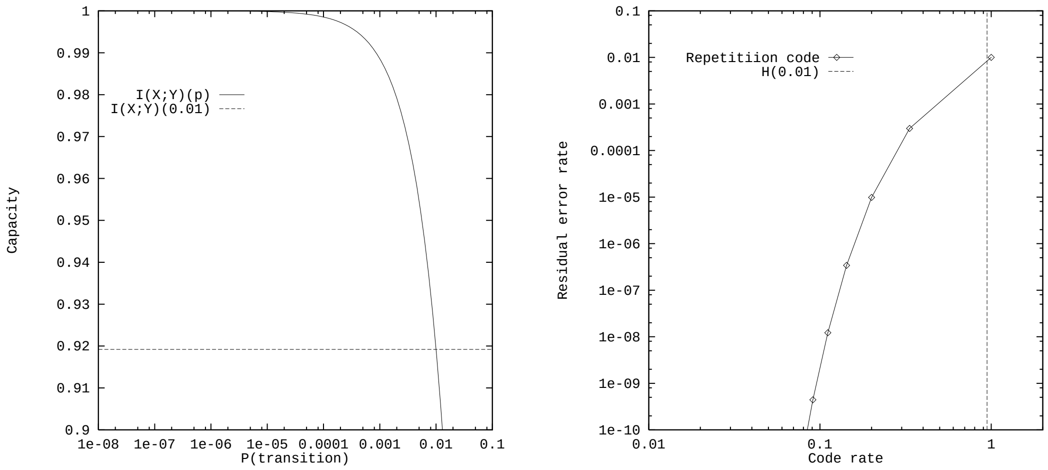


Figure 23: a) Capacity of binary symmetric channel at low loss rates, b) Efficiency of repetition code for a transition probability of 0.01

4.8 Channel Coding Theorem

We arrive at Shannon's second theorem, the *channel coding theorem*:

> For a channel of capacity C and a source of entropy H; if $H \leq C$, then for arbitrarily small ϵ, there exists a coding scheme such that the source is reproduced with a residual error rate less than ϵ.

Shannon's proof of this theorem is an existence proof rather than a means to construct such codes in the general case. In particular the choice of a good code is dictated by the characteristics of the channel noise. In parallel with the noiseless case, better codes are often achieved by coding multiple input symbols.

4.8.1 An efficient coding

Consider a rather artificial channel which may randomly corrupt one bit in each block of seven used to encode symbols in the channel – we take the probability of a bit corruption event is the same as correct reception. We inject N equiprobable input symbols (clearly $N \leq 2^7$ for unique decoding). What is the capacity of this channel?

We have 2^7 input and output patterns; for a given input x_j with binary digit representation $b_1b_2b_3b_4b_5b_6b_7$, we have eight equiprobable (i.e. with 1/8 probability) output symbols (and no others):

$$
\begin{array}{c}
b_1b_2b_3b_4b_5b_6b_7 \\
\bar{b}_1b_2b_3b_4b_5b_6b_7 \\
b_1\bar{b}_2b_3b_4b_5b_6b_7 \\
b_1b_2\bar{b}_3b_4b_5b_6b_7 \\
b_1b_2b_3\bar{b}_4b_5b_6b_7 \\
b_1b_2b_3b_4\bar{b}_5b_6b_7 \\
b_1b_2b_3b_4b_5\bar{b}_6b_7 \\
b_1b_2b_3b_4b_5b_6\bar{b}_7
\end{array}
$$

Then considering the information capacity per symbol:

$$
\begin{aligned}
C &= max(H(\mathcal{Y}) - H(\mathcal{Y}|\mathcal{X})) \\
&= \frac{1}{7}\left(7 - \sum_j \sum_k p(y_k|x_j)\log\left(\frac{1}{p(y_k|x_j)}\right)p(x_j)\right) \\
&= \frac{1}{7}\left(7 + \sum_j 8(\frac{1}{8}\log\frac{1}{8})\frac{1}{N}\right) \\
&= \frac{1}{7}\left(7 - N(\frac{8}{8}\log\frac{1}{8})\frac{1}{N}\right) \\
&= \frac{4}{7}
\end{aligned}
$$

The capacity of the channel is 4/7 information bits per binary digit of the channel coding. Can we find a mechanism to encode 4 information bits in 7 channel bits subject to the error property described above?

The (7/4) Hamming Code provides a *systematic* code to perform this – a systematic code is one in which the obvious binary encoding of the source symbols is present in the channel encoded form. For our source which emits at each time interval 1 of 16 symbols, we take the binary representation of this and copy it to bits b_3, b_5, b_6 and b_7 of the encoded block; the remaining bits are given by b_4, b_2, b_1, and *syndromes* by s_4, s_2, s_1:

$$
\begin{aligned}
b_4 &= b_5 \oplus b_6 \oplus b_7 \text{ and,} \\
s_4 &= b_4 \oplus b_5 \oplus b_6 \oplus b_7 \\
b_2 &= b_3 \oplus b_6 \oplus b_7 \text{ and,} \\
s_2 &= b_2 \oplus b_3 \oplus b_6 \oplus b_7 \\
b_1 &= b_3 \oplus b_5 \oplus b_7 \text{ and,} \\
s_1 &= b_1 \oplus b_3 \oplus b_5 \oplus b_7
\end{aligned}
$$

On reception if the binary number $s_4s_2s_1 = 0$ then there is no error, else $b_{s_4s_2s_1}$ is the bit in error.

This Hamming code uses 3 bits to correct 7 ($= 2^3 - 1$) error patterns and transfer 4 useful bits. In general a Hamming code uses m bits to correct $2^m - 1$ error patterns and transfer $2^m - 1 - m$ useful bits. The Hamming codes are called *perfect* as they use m bits to correct $2^m - 1$ errors.

The Hamming codes exist for all pairs $(2^n - 1, 2^{n-1})$ and detect one bit errors. Also the Golay code is a $(23, 12)$ block code which corrects up to three bit errors, an unnamed code exists at $(90, 78)$ which corrects up to two bit errors.

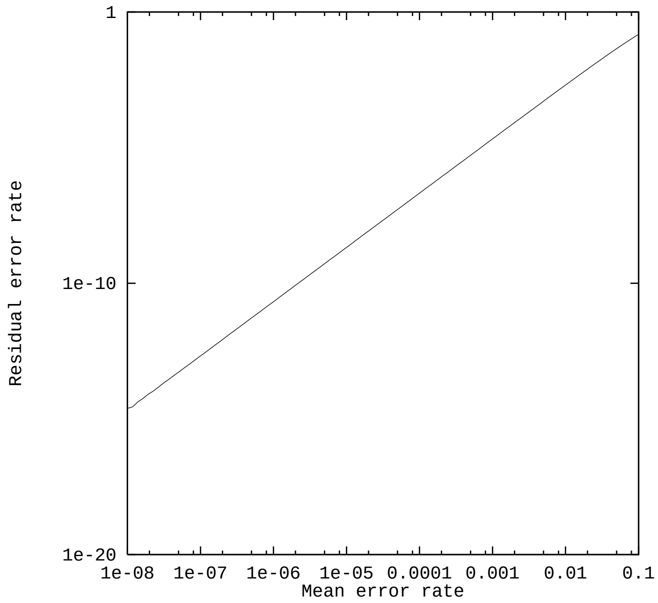


Figure 24: (7/4) Hamming code residual error rate

We can then consider the more general case, where we have random bit errors uniformly distributed (i.e. we allow the possibility of two or more bit errors per 7 bit block). Again we obtain the probability of residual error as the remainder of the binomial series:

$$P_e = \sum_{i=2}^{7} \binom{7}{i} p^i (1-p)^{7-i}$$

5 Continuous information

We now consider the case in which the signals or messages we wish to transfer are continuously variable; that is both the symbols we wish to transmit are continuous functions, and the associated probabilities of these functions are given by a continuous probability distribution.

We could try and obtain the results as a limiting process from the discrete case. For example, consider a random variable X which takes on values $x_k = k\delta x$, $k = 0, \pm 1, \pm 2, \ldots$, with probability $p(x_k)\delta x$, i.e. probability density function $p(x_k)$. We have the associated probability normalization:

$$\sum_k p(x_k)\delta x = 1$$

Using our formula for discrete entropy and taking the limit:

$$
\begin{aligned}
H(\mathcal{X}) &= \lim_{\delta x \to 0} \sum_k p(x_k)\delta x \log_2\left(\frac{1}{p(x_k)\delta x}\right) \\
&= \lim_{\delta x \to 0}\left[\sum_k p(x_k)\log_2\left(\frac{1}{p(x_k)}\right)\delta x - \log_2(\delta x)\sum_k p(x)\delta x\right] \\
&= \int_{-\infty}^{\infty} p(x)\log_2\left(\frac{1}{p(x)}\right)dx - \left(\lim_{\delta x \to 0}\log_2(\delta x)\right) \times \int_{-\infty}^{\infty} p(x)dx \\
&= h(\mathcal{X}) - \lim_{\delta x \to 0}\log_2(\delta x) \qquad (52)
\end{aligned}
$$

This is rather worrying as the latter limit does not exist. However, as we are often interested in the differences between entropies (i.e. in the consideration of mutual entropy or capacity), we define the problem away by using the first term only as a measure of *differential entropy*:

$$
h(\mathcal{X}) = \int_{-\infty}^{\infty} p(x)\log_2\left(\frac{1}{p(x)}\right)dx \qquad (53)
$$

We can extend this to a continuous random vector of dimension n concealing the n-fold integral behind vector notation and bold type:

$$
h(\mathbf{X}) = \int_{-\infty}^{\infty} p(\mathbf{x})\log_2\left(\frac{1}{p(\mathbf{x})}\right)d\mathbf{x} \qquad (54)
$$

We can then also define the joint and conditional entropies for continuous distributions:

$$
\begin{aligned}
h(\mathbf{X},\mathbf{Y}) &= \int\int p(\mathbf{x},\mathbf{y})\log_2\left(\frac{1}{p(\mathbf{x},\mathbf{y})}\right)d\mathbf{x}d\mathbf{y} \\
h(\mathbf{X}|\mathbf{Y}) &= \int\int p(\mathbf{x},\mathbf{y})\log_2\left(\frac{p(\mathbf{y})}{p(\mathbf{x},\mathbf{y})}\right)d\mathbf{x}d\mathbf{y} \\
h(\mathbf{Y}|\mathbf{X}) &= \int\int p(\mathbf{x},\mathbf{y})\log_2\left(\frac{p(\mathbf{x})}{p(\mathbf{x},\mathbf{y})}\right)d\mathbf{x}d\mathbf{y}
\end{aligned}
$$

with:

$$
\begin{aligned}
p(x) &= \int p(\mathbf{x},\mathbf{y})d\mathbf{y} \\
p(y) &= \int p(\mathbf{x},\mathbf{y})d\mathbf{x}
\end{aligned}
$$

Finally, mutual information of continuous random variables X and Y is defined using the double integral:

$$
i(X;Y) = \int\int p(x,y)\log\left(\frac{p(x|y)}{p(x)}\right)dxdy
$$

with the capacity of the channel given by maximizing this mutual information over all possible input distributions for X.

5.1 Properties

We obtain various properties analagous to the discrete case. In the following we drop the bold type, but each distribution, variable etc should be taken to be of n dimensions.

1. Entropy is maximized with equiprobable "symbols". If x is limited to some volume v (i.e. is only non-zero within the volume) then $h(x)$ is maximized when $p(x) = 1/v$.

2. $h(x, y) \leq h(x) + h(y)$

3. What is the maximum differential entropy for specified variance – we choose this as the variance is a measure of average power. Hence a restatement of the problem is to find the maximum differential entropy for a specified mean power.

   Consider the 1-dimensional random variable X, with the constraints:

$$\begin{aligned} \int p(x) dx &= 1 \\ \int (x-\mu)^2 p(x) dx &= \sigma^2 \\ \text{where: } \mu &= \int x p(x) dx \end{aligned} \tag{55}$$

   This optimization problem is solved using Lagrange multipliers and maximizing:

$$\int \left(-p(x) \log p(x) + \lambda_1 p(x)(x-\mu)^2 + \lambda_2 p(x) \right) dx$$

   which is obtained by solving:

$$-1 - \log p(x) + \lambda_1 (x-\mu)^2 + \lambda_2 = 0$$

   so that with due regard for the constraints on the system:

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

   hence:

$$h(X) = \log(\sqrt{(2\pi e \sigma}) = \frac{1}{2} \log(2\pi e \sigma^2) \tag{56}$$

   We observe that: i) for any random variable Y with variance σ, $h(Y) \leq h(X)$, ii) the differential entropy is dependent only on the variance and is independent of the mean, hence iii) the greater the power, the greater the differential entropy.

   This extends to multiple dimensions in the obvious manner.

5.2 Ensembles

In the case of continuous signals and channels we are interested in functions (of say time), chosen from a set of possible functions, as input to our channel, with some perturbed version of the functions being detected at the output. These functions then take the place of the input and output alphabets of the discrete case.

We must also then include the probability of a given function being injected into the channel which brings us to the idea of *ensembles* – this general area of work is known as *measure theory.*

For example, consider the ensembles:

1. $$f_\theta(t) = \sin(t+\theta)$$

   Each value of θ defines a different function, and together with a probability distribution, say $P(\theta)$, we have an ensemble. Note that θ here may be discrete or continuous – consider phase shift keying.

2. Consider a set of random variables $\{a_i; i = 0, \pm 1, \pm 2, \ldots\}$ where each a_i takes on a random value according to a Gaussian distribution with standard deviation $\sqrt{N}$; then:

   $$f(\{a_i\}, t) = \sum_i a_i \mathrm{sinc}(2Wt - i)$$

   is the "white noise" ensemble, band limited to W Hertz and with average power N^2.

3. More generally, we have for random variables $\{x_i\}$ the ensemble of band-limited functions:

   $$f(\{x_i\}, t) = \sum_i x_i \mathrm{sinc}(2Wt - i)$$

   where of course we remember from the sampling theorem that:

   $$x_i = f\left(\frac{i}{2W}\right)$$

   If we also consider functions limited to time interval T, then we obtain only $2TW$ non-zero coefficients and we can consider the ensemble to be represented by an n-dimensional $(n = 2TW)$probability distribution $p(x_1, x_2, \ldots, x_n)$.

4. More specifically, if we consider the ensemble of limited power (by P), band- limited (to $\pm W$) and time- limited signals (non-zero only in interval $(0, T)$), we find that the ensemble is represented by an n-dimensional probability distribution which is zero outside the n-sphere radius $r = \sqrt{2WP}$.

By considering the latter types of ensembles, we can fit them into the finite dimensional continuouus differential entropy definitions given in section 5.

5.3 Channel Capacity

We consider channels in which noise is injected independently of the signal; the particular noise source of interest is the so called *additive white Gaussian noise*. Further we restrict considerations to the final class of ensemble.

We have a signal with average power P, time limited to T and bandwidth limited to W.

We then consider the $n = 2WT$ samples (X_k and Y_k) that can be used to characterise both the input and output signals. Then the relationship between the input and output is given by:

$$Y_k = X_k + N_k \ , \ k = 1, 2, \ldots n$$

where N_k is from the band limited Gaussian distribution with zero mean and variance:

$$\sigma^2 = N_0 W$$

where N_0 is the power spectral density.

As N is independent of X we obtain the conditional entropy as being solely dependent on the noise source, and from equation 56 find its value:

$$h(Y|X) = h(N) = \frac{1}{2} \log 2\pi e N_0 W$$

Hence:

$$i(X;Y) = h(Y) - h(N)$$

The capacity of the channel is then obtained by maximizing this over all input distributions – this is achieved by maximizing with respect to the distribution of Y subject to the average power limitation:

$$E[X_k^2] = P$$

As we have seen this is achieved when we have a Gaussian distribution, hence we find both X and Y must have Gaussian distributions. In particular X has variance P and Y has variance $P + N_0 W$.

We can then evaluate the channel capacity, as:

$$\begin{aligned} h(Y) &= \frac{1}{2} \log 2\pi e (P + N_0 W) \\ h(N) &= \frac{1}{2} \log 2\pi e (N_0 W) \\ C &= \frac{1}{2} \log \left(1 + \frac{P}{N_0 W}\right) \end{aligned} \tag{57}$$

This capacity is in bits per channel symbol, so we can obtain a rate per second by multiplication by n/T, i.e. from $n = 2WT$, multiplication by $2W$:

$$C = 2W \log_2 \left(1 + \frac{P}{N_0 W}\right) \text{ bit/s}$$

So Shannon's third theorem, the *channel capacity theorem*:

> The capacity of a channel bandlimited to W Hertz, perturbed by additive white Gaussian noise of power spectral density N_0 and bandlimited to W is given by:
>
> $$C = 2W \log_2 \left(1 + \frac{P}{N_0 W}\right) \text{ bit/s} \tag{58}$$
>
> where P is the average transmitted power.

5.3.1 Notes

The second term within the log in equation 58 is the signal to noise ratio (SNR).

1. Observe that the capacity increases monotonically and without bound as the SNR increases.

2. Similarly the capacity increases monotonically as the bandwidth increases but to a limit. Using Taylor's expansion for ln:

$$\ln(1 + \alpha) = \alpha - \frac{\alpha^2}{2} + \frac{\alpha^3}{3} - \frac{\alpha^4}{4} + \cdots$$

   we obtain:

$$C \rightarrow \frac{P}{N_0} \log_2 e$$

3. This is often rewritten in terms of *energy per bit*, E_b, which is defined by $P = E_b C$. The limiting value is then:

$$\frac{E_b}{N_0} \rightarrow \log_e 2 = 0.693$$

   This is called the Shannon Limit.

4. The capacity of the channel is achieved when the source "looks like noise". This is the basis for spread spectrum techniques of modulation and in particular Code Division Multiple Access (CDMA).

1 Fourier Series and Transforms

Consider real valued periodic functions $f(x)$, i.e. for some a and $\forall x$, $f(x+a) = f(x)$. Without loss of generality we take $a = 2\pi$.

We observe the orthogonality properties of $\sin(mx)$ and $\cos(nx)$ for integers m and n:

$$\begin{aligned}
\int_0^{2\pi} \cos(nx)\cos(mx)dx &= \begin{cases} 2\pi & \text{if } m = n = 0 \\ \pi\delta_{mn} & \text{otherwise} \end{cases} \\
\int_0^{2\pi} \sin(nx)\sin(mx)dx &= \begin{cases} 0 & \text{if } m = n = 0 \\ \pi\delta_{mn} & \text{otherwise} \end{cases} \\
\int_0^{2\pi} \sin(nx)\cos(mx)dx &= 0 \quad \forall m, n
\end{aligned} \tag{1}$$

We use the Kronecker δ function to mean:

$$\delta_{mn} = \begin{cases} 1 & m = n \\ 0 & \text{otherwise} \end{cases}$$

Then the Fourier Series for $f(x)$ is:

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos(nx) + b_n \sin(nx)) \tag{2}$$

where the Fourier Coefficients are:

$$a_n = \frac{1}{\pi}\int_0^{2\pi} f(x)\cos(nx)dx \quad n \geq 0 \tag{3}$$

$$b_n = \frac{1}{\pi}\int_0^{2\pi} f(x)\sin(nx)dx \quad n \geq 1 \tag{4}$$

We hope that the Fourier Series provides an expansion of the function $f(x)$ in terms of cosine and sine terms.

1.1 Approximation by least squares

Let $S'_N(x)$ be any sum of sine and cosine terms:

$$S'_N(x) = \frac{a'_0}{2} + \sum_{n=1}^{N-1} (a'_n \cos(nx) + b'_n \sin(nx)) \tag{5}$$

and $S_N(x)$, the truncated Fourier Series for a function $f(x)$:

$$S_N(x) = \frac{a_0}{2} + \sum_{n=1}^{N-1} (a_n \cos(nx) + b_n \sin(nx))$$

where a_n and b_n are the Fourier coefficients. Consider the integral giving the discrepancy between $f(x)$ and $S'_n(x)$ (assuming $f(x)$ is well behaved enough for the integral to exist):

$$\int_0^{2\pi} \{f(x) - S'_n(x)\}^2 \, dx$$

which simplifies to:

$$\begin{aligned} & \int_0^{2\pi} \{f(x)\}^2 \, dx \\ - \quad & 2\pi a_0^2 - \pi \sum_{n=1}^{N-1} (a_n^2 + b_n^2) \\ + \quad & 2\pi (a'_0 - a_0)^2 + \pi \sum_{n=1}^{N-1} \left\{ (a'_n - a_n)^2 + (b'_n - b_n)^2 \right\} \end{aligned} \tag{6}$$

Note that the terms involving a' and b' are all ≥ 0 and vanish when $a'_n = a_n$ and $b'_n = b_n$. *The Fourier Series is the best approximation, in terms of mean squared error, to f that can be achieved using these circular functions.*

1.2 Requirements on functions

The Fourier Series and Fourier coeffieients are defined as above. However we may encounter some problems:

1. integrals in equations 3, 4 fail to exist. e.g.:

   $$f(x) = \frac{1}{x}$$

   or

   $$f(x) = \begin{cases} 1 & x \text{ rational} \\ 0 & x \text{ irrational} \end{cases}$$

2. although a_n, b_n exist, the series does not converge,

3. even though the series converges, the result is not $f(x)$

   $$f(x) = \begin{cases} +1 & 0 \leq x < \pi \\ -1 & \pi \leq x < 2\pi \end{cases}$$

   then:

   $$a_n = 0, \ b_n = \frac{2}{\pi} \int_0^{\pi} sin(nx) dx = \begin{cases} \frac{4}{n\pi} & n \text{ odd} \\ 0 & n \text{ even} \end{cases}$$

   so:

   $$f(x) \stackrel{?}{=} \frac{4}{\pi} \left[\sin(x) + \frac{\sin(3x)}{3} + \frac{\sin(5x)}{5} + \cdots \right]$$

   but series gives $f(n\pi) \stackrel{?}{=} 0$.

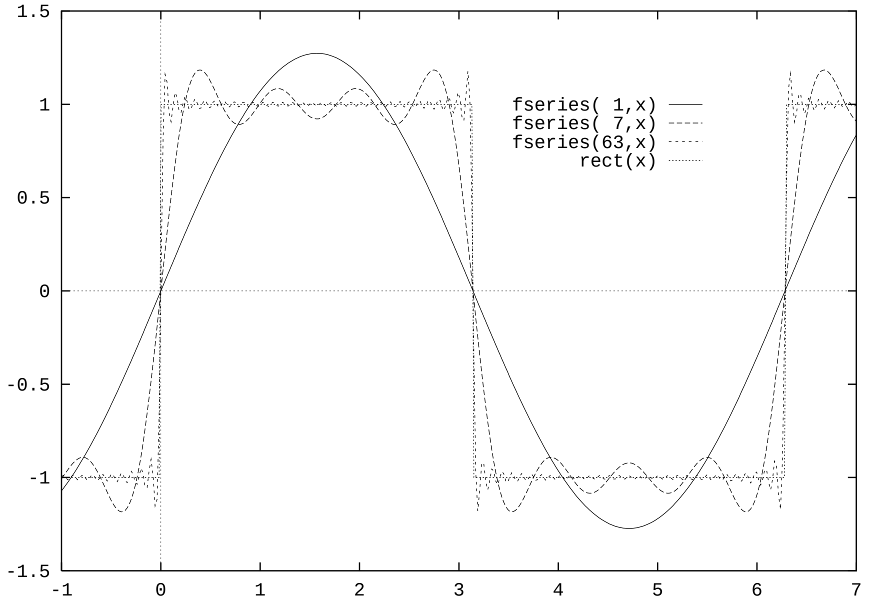


Figure 1: Approximations to rectangular pulse

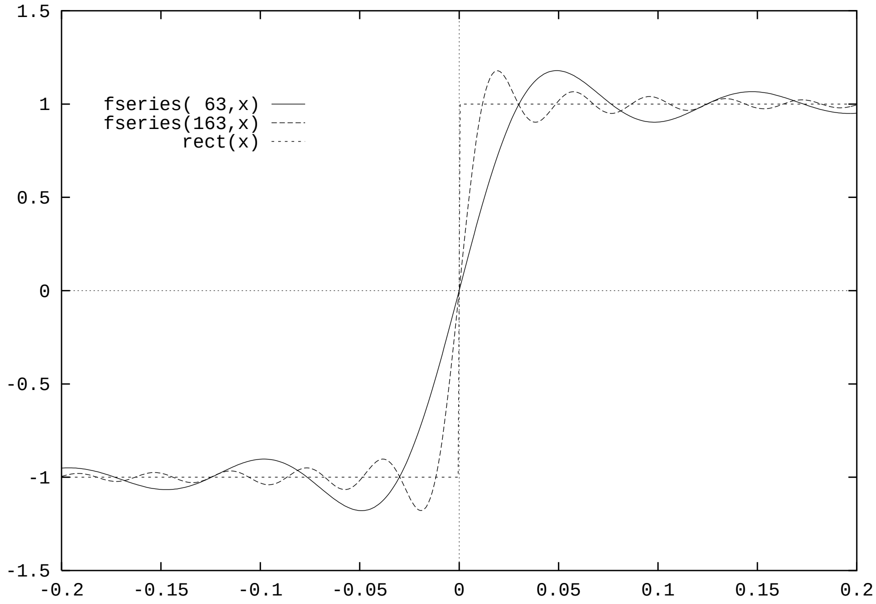


Figure 2: Gibbs phenomenon

However, in real life examples encountered in signal processing things are simpler as:

1. If $f(x)$ is bounded in $(0, 2\pi)$ and piecewise continuous, the Fourier series converges to $\{f(x_-) + f(x_+)\}/2$ at interior points and $\{f(0_+) + f(2\pi_-)\}/2$ at 0 and 2π. [1]
2. If $f(x)$ is also continuous on $(0, 2\pi)$, the sum of the Fourier Series is equal to $f(x)$ at all points.
3. a_n and b_n tend to zero a least as fast as $1/n$.

1.3 Complex form

Rewrite using e^{inx} in the obvious way from the formula for sin and cos:

$$\begin{aligned}
&\frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos(nx) + b_n \sin(nx)) \\
&= \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \frac{(e^{inx} + e^{-inx})}{2} + b_n \frac{(e^{inx} - e^{-inx})}{2} \right) \\
&= \sum_{-\infty}^{\infty} c_n e^{inx} \qquad (7)
\end{aligned}$$

with:

$$\begin{array}{rrcl}
 & c_0 & = & a_0/2 \\
n > 0 & c_n & = & (a_n - ib_n)/2 \\
 & c_{-n} & = & (a_n + ib_n)/2 \\
\text{observe:} & c_{-n} & = & c_n^*
\end{array} \qquad (8)$$

where c_* denotes the complex conjugate, and:

$$c_n = \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-inx} dx \qquad (9)$$

1.4 Approximation by interpolation

Consider the value of a periodic function $f(x)$ at N discrete equally spaced values of x:

$$x_r = r\phi \quad (\phi = \frac{2\pi}{N}\ ,\ r = 0, 1, \ldots, N-1)$$

try to find coefficients c_n such that:

$$f(r\phi) = \sum_{n=0}^{N-1} c_n e^{i\phi rn} \qquad (10)$$

[1] A notation for limits is introduced here $f(x_-) = \lim_{\epsilon \to 0} f(x - \epsilon)$ and $f(x_+) = \lim_{\epsilon \to 0} f(x + \epsilon)$.

Multiply by $e^{-i\phi rm}$ and sum with respect to r:

$$\sum_{r=0}^{N-1} f(r\phi)e^{-i\phi rm} = \sum_{r=0}^{N-1}\sum_{n=0}^{N-1} c_n e^{i\phi r(n-m)}$$

but by the sum of a geometric series and blatant assertion:

$$\sum_{r=0}^{N-1} e^{i\phi r(n-m)} = \begin{cases} \dfrac{1-e^{i\phi N(n-m)}}{1-e^{i\phi(n-m)}} = 0 & n \neq m \\ N & n = m \end{cases}$$

so:

$$c_m = \frac{1}{N}\sum_{r=0}^{N-1} f(r\phi)e^{-i\phi rm} \tag{11}$$

Exercise for the reader: show inserting equation 11 into equation 10 satisfies the original equations.

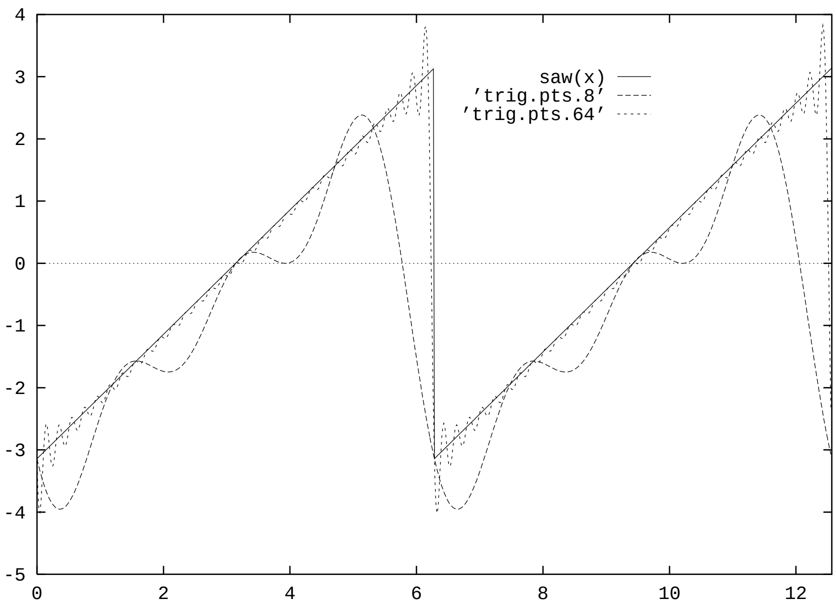


Figure 3: 8 and 64 point interpolations to sawtooth function

We can translate this back into cos and sin series:

$$f(x_r) = a_0 + \sum_{n=1}^{N/2-1} (a_n \cos(\frac{2\pi rn}{N}) + b_n \sin(\frac{2\pi rn}{N}))$$

where (for $0 < n < N/2$):

$$a_0 = \frac{1}{N}\sum_{r=0}^{N-1} f(\frac{2\pi r}{N})$$

$$
\begin{aligned}
a_n &= \frac{2}{N}\sum_{r=0}^{N-1} f(\frac{2\pi r}{N})\cos(\frac{2\pi rn}{N}) \\
b_n &= \frac{2}{N}\sum_{r=0}^{N-1} f(\frac{2\pi r}{N})\sin(\frac{2\pi rn}{N}) \qquad (12)
\end{aligned}
$$

As we increase N we can make the interpolation function agree with $f(x)$ at more and more points. Taking a_n as an example, as $n \to \infty$ (for well behaved $f(x)$):

$$
a_n = \frac{1}{\pi}\sum_{r=0}^{N-1} f(\frac{2\pi r}{N})\cos(\frac{2\pi rn}{N})\frac{2\pi}{N} \quad \to \quad \frac{1}{\pi}\int_0^{2\pi} f(x)\cos(nx)dx \qquad (13)
$$

1.5 Cosine and Sine Series

Observe that if $f(x)$ is symmetric, that is $f(-x) = f(x)$ then:

$$
\begin{aligned}
a_0 &= \frac{2}{\pi}\int_0^{\pi} f(x)dx \\
a_n &= \frac{2}{\pi}\int_0^{\pi} f(x)\cos(nx)dx \\
b_n &= 0
\end{aligned}
$$

hence the Fourier series is simply:

$$
\frac{a_0}{2} + \sum_{n>0} a_n \cos(nx)
$$

On the other hand if $f(x) = -f(-x)$: then we get the corresponding sine series:

$$
\sum_{n>0} b_n \sin(nx)
$$

with:

$$
b_n = \frac{2}{\pi}\int_0^{\pi} f(x)\sin(nx)dx
$$

Example, take $f(x) = 1$ for $0 < x < \pi$. We can extend this to be periodic with period 2π either symmetrically, when we obtain the cosine series (which is simply $a_0 = 1$ as expected), or with antisymmetry (square wave) when the sine series gives us:

$$
1 = \frac{4}{\pi}\left[\sin(x) + \frac{\sin(3x)}{3} + \frac{\sin(5x)}{5} + \cdots\right] \text{ (for } 0 < x < \pi)
$$

Another example, take $f(x) = x(\pi - x)$ for $0 < x < \pi$. The cosine series is:

$$
f(x) = \frac{\pi^2}{6} - \cos(2x) - \frac{\cos(4x)}{2^2} - \frac{\cos(6x)}{3^2} - \cdots
$$

hence as $f(x) \to 0$ as $x \to 0$:

$$\frac{\pi^2}{6} = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \cdots$$

the corresponding sine series is:

$$x(\pi - x) = \frac{8}{\pi}\left[\sin(x) + \frac{\sin(3x)}{3^3} + \frac{\sin(5x)}{5^3} + \cdots\right]$$

and at $x = \pi/2$:

$$\pi^3 = 32\left[1 - \frac{1}{3^3} + \frac{1}{5^3}\cdots\right]$$

Observe that one series converges faster than the other, when the values of the basis functions match the values of the function at the periodic boundaries.

1.6 Fourier transform

Starting from the complex series in equation 9, make a change of scale – consider a periodic function $g(x)$ with period $2X$; define $f(x) = g(xX/\pi)$, which has period 2π.

$$\begin{aligned} c_n &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-inx}dx \\ &= \frac{1}{2\pi}\frac{\pi}{X}\int_{-X}^{X} g(y)e^{-in\pi y/X}dy \\ c(k) &= \frac{1}{2\pi}\int_{-X}^{X} g(y)e^{-iky}dy \end{aligned} \tag{14}$$

where $k = n\pi/X$, and $c(k) = Xc_n/\pi$. Hence:

$$\begin{aligned} f(x) &= \sum_n c_n e^{inx} \\ g(y) &= \sum_k c(k)e^{iky}\frac{\pi}{X} \\ &= \sum_k c(k)e^{iky}\delta k \end{aligned} \tag{15}$$

writing δk for the step π/X. Allowing $X \to \infty$, then we hope:

$$\sum_k c(k)e^{iky}\delta k \to \int_{-\infty}^{\infty} G(k)e^{iky}dk$$

Hence we obtain the Fourier Transform pair:

$$g(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} G(k)e^{ikx}dk \tag{16}$$

$$G(k) = \int_{-\infty}^{\infty} g(x)e^{-ikx}dx \tag{17}$$

Equation 16 expresses a function as a spectrum of frequency components. Taken together (with due consideration for free variables) we obtain:

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(y)e^{ik(x-y)}dkdy \tag{18}$$

Fourier's Integral Theorem is a statement of when this formula holds; if $f(x)$ is of bounded variation, and $|f(x)|$ is integrable from $-\infty$ to ∞, then equation 18 holds (well more generally the double integral gives $(f(x_+)+f(x_-))/2$).

1.7 FT Properties

Writing $g(x) \rightleftharpoons G(k)$ to signify the transform pair, we have the following properties:

1. Symmetry and reality
   - if $g(x)$ real then $G(-k) = G^*(k)$
   - if $g(x)$ real and even:
     $$G(k) = 2\int_0^{\infty} f(x)\cos(kx)dx$$
   - if $g(x)$ real and odd
     $$G(k) = -2i\int_0^{\infty} f(x)\sin(kx)dx$$

   The last two are analogues of the cosine and sine series – cosine and sine transforms.
2. Linearity; for constants a and b, if $g_1(x) \rightleftharpoons G_1(k)$ and $g_2(x) \rightleftharpoons G_2(k)$ then
   $$ag_1(x) + bg_2(x) \rightleftharpoons aG_1(k) + bG_2(k)$$
3. Space/time shifting $g(x-a) \rightleftharpoons e^{-ika}G(k)$
4. Frequency shifting $g(x)e^{i\lambda x} \rightleftharpoons G(k-\lambda)$
5. Differentiation once $g'(x) \rightleftharpoons ikG(k)$
6. ... and n times, $g^{(n)}(x) \rightleftharpoons (ik)^n G(k)$

1.8 Convolutions

Suppose $f(x) \rightleftharpoons F(k)$, $g(x) \rightleftharpoons G(k)$; what function $h(x)$ has a transform $H(k) = F(k)G(k)$? Inserting into the inverse transform:

$$\begin{aligned} h(x) &= \frac{1}{2\pi}\int F(k)G(k)e^{ikx}dk \\ &= \frac{1}{2\pi}\iiint f(y)g(z)e^{ik(x-y-z)}dkdydz \\ &= \frac{1}{2\pi}\iiint f(y)g(\xi-y)e^{ik(x-\xi)}dkdyd\xi \end{aligned}$$

From equation 18, we then obtain the *convolution* of f and g:

$$\begin{aligned} h(x) &= \int_{-\infty}^{\infty} f(y)g(x-y)dy && (19) \\ &= f(x) \star g(x) && (20) \end{aligned}$$

or:

$$\int_{-\infty}^{\infty} f(y)g(x-y)dy \rightleftharpoons F(k)G(k)$$

Similarly:

$$f(y)g(y) \rightleftharpoons \int_{-\infty}^{\infty} F(\lambda)G(k-\lambda)d\lambda$$

As a special case convolve the real functions $f(x)$ and $f(-x)$, then:

$$\begin{aligned} F(-k) &= F^*(k) \\ H(k) &= F(k)F^*(k) \\ &= |F(k)|^2 && (21) \\ \rho_f(x) &= \int_{-\infty}^{\infty} f(y)f(y-x)dy && (22) \end{aligned}$$

The function ρ_f in Equation 22 is the *autocorrelation* function (of f), while $|F(k)|^2$ in equation 21 is the *spectral density*. Observe Parseval's theorem:

$$\rho(0) = \int_{-\infty}^{\infty} [f(y)]^2 dy = \frac{1}{2\pi}\int_{-\infty}^{\infty} |F(k)|^2 dk$$

1.9 Some FT pairs

Some example Transform pairs:

1. Simple example:

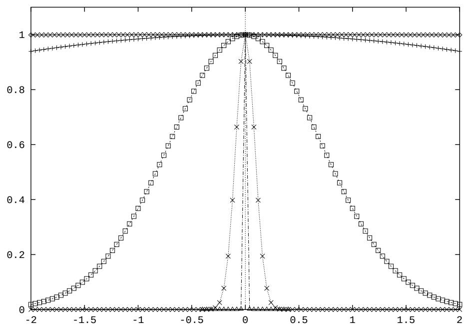
$$f(x) = \begin{cases} e^{-ax} & x > 0 \\ 0 & x < 0 \end{cases}, \; F(k) = \frac{1}{a+ik}$$

   If we make the function symmetric about 0:

$$f(x) = e^{-a|x|}\;,\; F(k) = \frac{2a}{a^2+k^2}$$

2. Gaussian example (see figure 4):

$$\begin{aligned} f(x) &= e^{-\lambda^2 x^2} \\ F(k) &= \int e^{-\lambda^2 x^2 - ikx} dx \\ &= e^{-\frac{k^2}{4\lambda^2}} \int e^{-\left(\lambda x + \frac{ik}{2\lambda}\right)^2} dx \\ &= \frac{e^{-\frac{k^2}{4\lambda^2}}}{\lambda} \int_{-\infty+\frac{ik}{2\lambda}}^{\infty+\frac{ik}{2\lambda}} e^{-u^2} du \\ &= \frac{\sqrt{\pi}}{\lambda} e^{-\frac{k^2}{4\lambda^2}} && (23) \end{aligned}$$



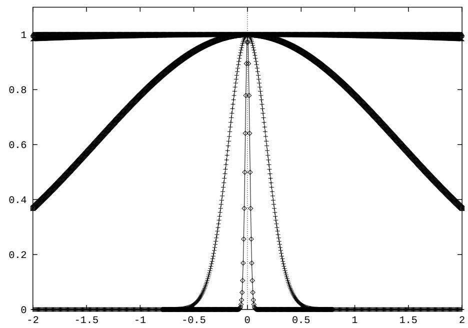


Figure 4: Gaussian distributions and their transforms

Gaussian function is self dual. Consider what happens as $\lambda \to \infty$.

3. Rectangular pulse function:

$$r(x) = \begin{cases} \frac{1}{2a} & |x| < a \\ 0 & |x| > a \end{cases} \quad , \; R(k) = \frac{\sin(ka)}{ka} \tag{24}$$

Notice that as $a \to 0$, and observe $R(k) \to 1$; what happens to $r(x)$?

1.10 The Dirac delta function

We shall find the Dirac δ-function to be of considerable use in the consideration of sampling. However, this "function" is not a function at all in the classical analysis sense, and we will play fast and loose with its properties as a full discussion is beyond the scope of this course. The treatment below is purely formal.

Considering the rectangular pulse example above in equation 24 we are interested in the properties of $r(x)$ (replacing a by ϵ) as $\epsilon \to 0$.

$$\int_{-\infty}^{y} r(x)dx = \begin{cases} 1 & y > \epsilon \\ 0 & y < -\epsilon \end{cases} \tag{25}$$

On the wild assumption that it exists:

$$\lim_{\epsilon \to 0} r(x) = \delta(x)$$

Some properties:

1. conceptually
$$\delta(x) = \begin{cases} `\infty' & x = 0 \\ 0 & x \neq 0 \end{cases}$$
2. taking the limit of equation 25:
$$\int_{-\infty}^{y} \delta(x)dx = \begin{cases} 1 & y > 0 \\ 0 & y < 0 \end{cases}$$
3. assuming $f(x)$ is continuous in the interval $(c-\epsilon, c+\epsilon)$ and using the displaced pulse function $r_\epsilon(x-c)$; by the intermediate value theorem for some ξ in the interval $(c-\epsilon, c+\epsilon)$:
$$\int_{-\infty}^{\infty} f(x)r(x-c)dx = f(\xi)$$
Hence with usual disregard for rigorousness, taking the limit:
$$\int_{-\infty}^{\infty} f(x)\delta(c-x)dx = f(c)$$
Observe this is a convolution ...
4. further note $g(x)\delta(x-c) = g(c)\delta(x-c)$.

In some ways the δ-function and its friends can be considered as analogous to extending the rationals to the reals; they both enable sequences to have limits.

We define the *ideal sampling function* of interval X as:
$$\delta_X(x) = \sum_n \delta(x - nX)$$
we can Fourier transform this (exercise for the reader) and obtain:
$$\Delta_X(k) = \frac{1}{X}\sum_m \delta(kX - 2\pi m)$$
with $\delta_X \rightleftharpoons \Delta_X$.

1.11 Sampling Theorem

Our complex and cosine/sine series gave us a discrete set of Fourier coefficients for a periodic function; or looking at it another way for a function which is non-zero in a finite range, we can define a periodic extension of that function. The symmetric nature of the Fourier transform would suggest that something interesting might happen if the Fourier transform function is also non-zero only in a finite range.

Consider sampling a signal $g(x)$ at intervals X to obtain the discrete time signal:

$$\begin{aligned} g_X(x) &= \sum_n g(nX)\delta(x-nX) \\ &= g(x)\delta_X(x) \end{aligned} \tag{26}$$

Remember the properties of convolutions, we know that a product in the x domain is a convolution in the transform domain. With $g(x) \rightleftharpoons G(k)$ and $g_X(x) \rightleftharpoons G_X(k)$:

$$\begin{aligned} G_X(k) &= G(k) \star \Delta_X(k) \\ &= \frac{1}{X}\sum_m G(k) \star \delta(kX - 2\pi m) \\ &= \frac{1}{X}\sum_m G(k - \frac{2\pi m}{X}) \end{aligned} \tag{27}$$

From this we note that for any $g(x) \rightleftharpoons G(k)$ the result of sampling at intervals X in the x domain results in a transform $G_X(k)$ which is the periodic extension of $G(k)$ with period $2\pi/X$.

Conversely if $g(x)$ is strictly band-limited by W (radians per x unit), that is $G(k)$ is zero outside the interval $(-W, W)$, then by sampling at intervals $2\pi/2W$ in the x domain:

$$G(k) = \frac{2\pi}{2W}G_X(k)\ ,\ (-W < k < W)$$

as shown in figure 5. Performing the Fourier transform on equation 26, we obtain:

$$G(k) = \frac{2\pi}{2W}\sum_n g(\frac{2\pi n}{2W})e^{-ikn2\pi/2W}\ ,\ (-W < k < W) \tag{28}$$

But we know $G(k) = 0$ for all $|k| > W$; therefore the sequence $\{g(n/2W)\}$ completely defines $G(k)$ and hence $g(x)$. Using equation 28 and performing the inverse transform:

$$\begin{aligned} g(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} G(k)e^{ikx}dk \\ &= \frac{1}{2\pi}\int_{-W}^{W}\frac{2\pi}{2W}\sum_n g(\frac{n\pi}{W})e^{-ikn\pi/W}e^{ikx}dk \\ &= \frac{1}{2W}\sum_n g(\frac{n\pi}{2W})\int_{-W}^{W} e^{ik(x-n\pi/W)}dk \\ &= \sum_n g(\frac{n\pi}{W})\frac{\sin(Wx - n\pi)}{Wx - n\pi} \end{aligned} \tag{29}$$

Taking the example of time for the X domain, we can rewrite this in terms of the sampling frequency f_s normally quoted in Hertz rather than radians per sec:

$$\begin{aligned} g(t) &= \sum_n g(\frac{n}{f_s})\frac{\sin(\pi(f_s t - n))}{\pi(f_s t - n)} \\ &= \sum_n g(\frac{n}{f_s})\mathrm{sinc}(f_s t - n) \end{aligned} \tag{30}$$

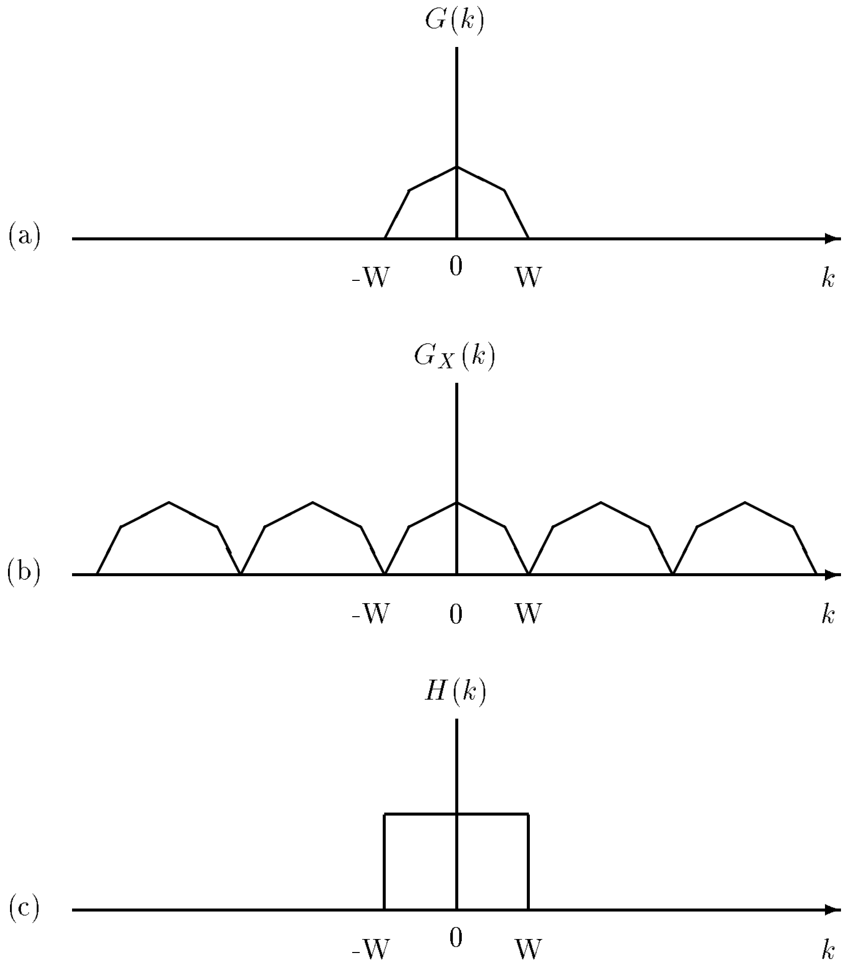


Figure 5: Band-limited signals: (a) Spectrum of $g(x)$. (b) Spectrum of $g_X(x)$. (c) Ideal filter response for reconstruction

The *Nyquist* rate: a signal band-limited by W (Hertz) can be uniquely determined by sampling at a rate of $f_s \geq 2W$. The minimum sampling rate $f_s = 2W$ is the *Nyquist* rate. The sampling theorem is sometimes called the *Nyquist* theorem.

1.12 Aliasing

In reality is is not possible to build the analogue filter which would have the perfect response required to achieve the Nyquist rate (response shown figure 5(c)).

Figure 7 demonstrates the problem if the sampling rate is too low for a given filter. We had assumed a signal band-limited to W and sampled at Nyquist rate $2W$, but the signal (or a badly filtered version of the signal) has non-zero frequency components at frequencies higher than W which the periodicity of the transform $G_X(k)$ causes to be added in. In looking only in the interval

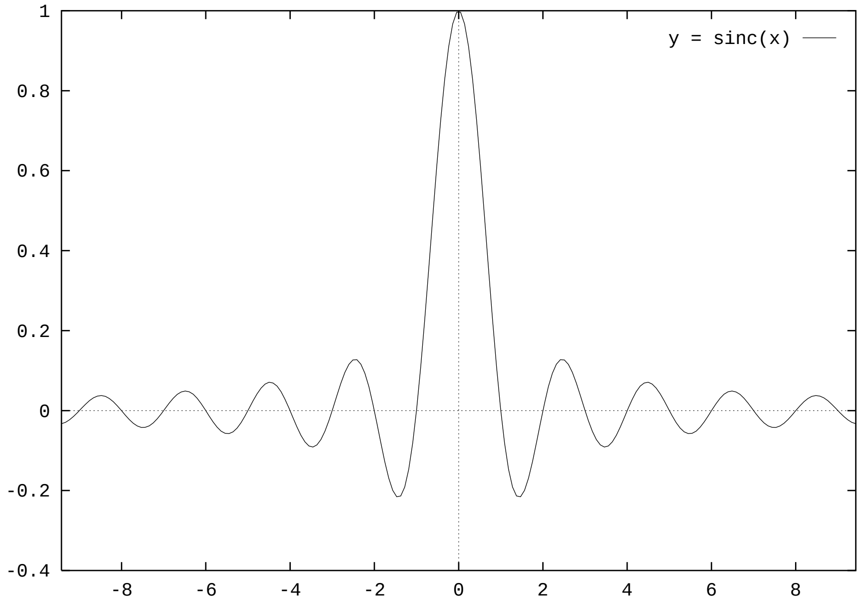


Figure 6: The sinc function, $\frac{sin(\pi x)}{\pi x}$

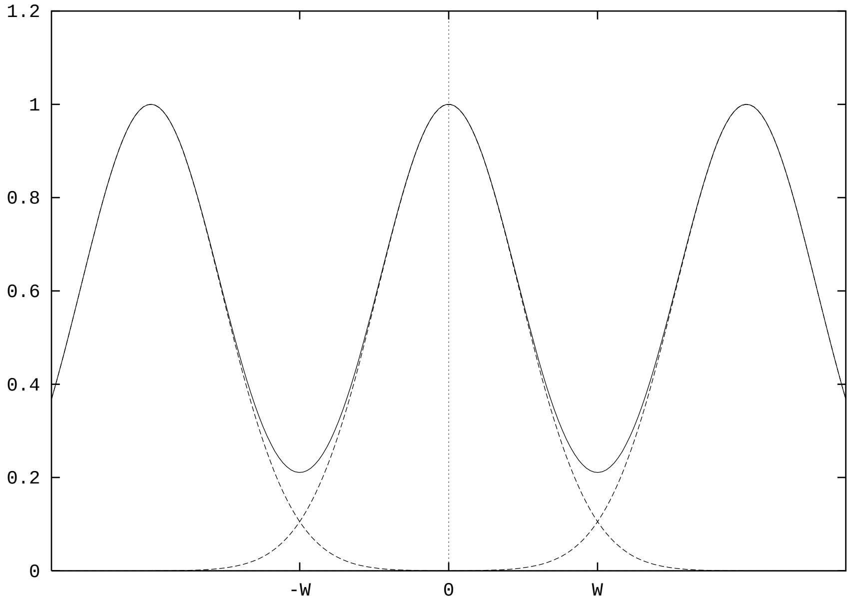


Figure 7: Aliasing effect example

$(-W, W)$ it appears that the tails have been folded over about $x = -W$ and $x = W$ with higher frequencies being reflected as lower frequencies.

To avoid this problem, it is normal to aim to sample well above the Nyquist rate; for example standard 64Kbps CODECs used in the phone system aim to band-limit the analogue input signal to about 3.4kHz and sample at 8kHz.

Examples also arise in image digitization where spatial frequencies higher than the resolution of the scanner or video camera cause aliasing effects.

2 The Discrete Fourier Transform

We have seen how we can describe a periodic signal by a discrete set of Fourier coefficients and conversely how a discrete set of points can be used to describe a band-limited signal.

Time to come to earth; we shall concern ourselves with band-limited periodic signals and consider the *Discrete Fourier Transform* as this lends itself to computation. Furthermore the DFT is also amenable to efficient implementation as the *Fast Fourier Transform*.

2.1 Definitions

Consider a data sequence $\{g_n\} = \{g_0, g_1, \dots g_{N-1}\}$. For example these could represent the values sampled from an analogue signal $s(t)$ with $g_n = s(nT_s)$.

The Discrete Fourier Transform is:

$$G_k = \sum_{n=0}^{N-1} g_n e^{-\frac{2\pi i}{N}kn} \;, \; (k = 0, 1, \dots, N-1) \tag{31}$$

and its inverse:

$$g_n = \frac{1}{N}\sum_{k=0}^{N-1} G_k e^{\frac{2\pi i}{N}kn} \;, \; (n = 0, 1, \dots, N-1) \tag{32}$$

One major pain of the continuous Fourier Transform and Series now disappears, there is no question of possible convergence problems with limits as these sums are all finite.

2.2 Properties

The properties of the DFT mimic the continuous version:

1. Symmetry and reality

- if g_n real then $G_{-k} = G_k^*$
- as g_n is periodic, $G_{(N/2)-k} = G^*_{(N/2)+k}$

2. Linearity; $ag_n + bh_n$ has DFT $aG_k + bH_k$ in the obvious manner.
3. Shifting; observe a shift in the g_n values is really a rotation. For the rotated sequence g_{n-n_0} the DFT is $G_k e^{-2\pi i k n_0/N}$.

There is also the parallel of convolution. The *circular convolution* of sequences g_n and h_n is defined by:

$$y_n = \sum_{r=0}^{N-1} g_r h_{n-r} \ , \ (n = 0, 1, \ldots N-1)$$

The DFT of y_n is then:

$$\begin{aligned}
Y_k &= \sum_{n=0}^{N-1} y_n e^{-\frac{2\pi i}{N}kn} \\
&= \sum_{n=0}^{N-1}\sum_{r=0}^{N-1} g_r h_{n-r} e^{-\frac{2\pi i}{N}kn} \\
&= \sum_{r=0}^{N-1} g_r e^{-\frac{2\pi i}{N}kr} \sum_{r=0}^{N-1} h_{n-r} e^{-\frac{2\pi i}{N}k(n-r)} \\
&= G_k H_k \qquad (33)
\end{aligned}$$

2.3 Fast Fourier Transform

A simplistic implementation of the DFT would require N^2 complex multiplications and $N(N-1)$ complex additions. However the symmetry of the DFT can be exploited and a divide and conquer strategy leads to the Fast Fourier Transform when N is a power of 2.

For simplicity we write $\omega = e^{-2\pi i/N}$, then the DFT for an even number $N = 2L$, becomes:

$$\begin{aligned}
G_k &= \sum_{n=0}^{2L-1} g_n \omega^{nk} \ , \ k = 0, 1, \ldots 2L-1 \qquad (34) \\
&= \sum_{n=0}^{L-1} g_n \omega^{nk} + \sum_{n=L}^{2L-1} g_n \omega^{nk} \\
&= \sum_{n=0}^{L-1} (g_n + g_{n+L}\omega^{kL})\omega^{kn} \\
&= \sum_{n=0}^{L-1} (g_n + g_{n+L}(-1)^k)\omega^{kn} \qquad (35)
\end{aligned}$$

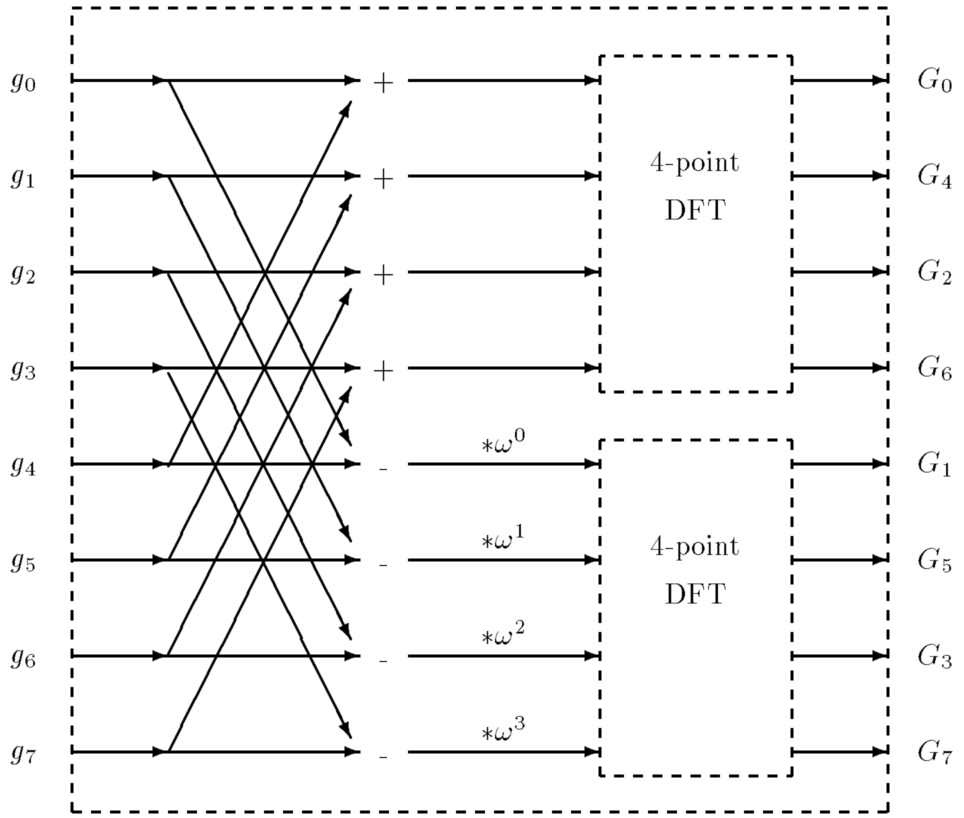


Figure 8: Division of 8-point into two 4-point DFTs

as, $\omega^L = -1$, hence $\omega^{kL} = (-1)^k$. We can then simply separate out even and odd terms of G_k and we obtain:

$$
\begin{aligned}
G_{2l} &= \sum_{n=0}^{L-1} (g_n + g_{n+L})(\omega^2)^{ln} \;,\; l = 0, 1, \ldots, L-1 \\
G_{2l+1} &= \sum_{n=0}^{L-1} ((g_n - g_{n+L})\omega^n)(\omega^2)^{ln} \;,\; l = 0, 1, \ldots, L-1
\end{aligned}
$$

Observe that these are two l-point DFTs of the sequences $\{g_n + g_{n+L}\}$ and $\{(g_n + g_{n+L})\omega^n\}$.

If N is a power of 2, then we can decompose in this manner $\log_2 N$ times until we obtain N single point transforms.

The diagram in figure 8 shows the division of an 8-point DFT into two 4-point DFTs. Recursive division in this manner results in the Fast Fourier Transform (FFT).

Figure 9 shows the a repetitive pattern formed between certain pairs of points – this *butterfly* pattern highlights some features of the FFT:

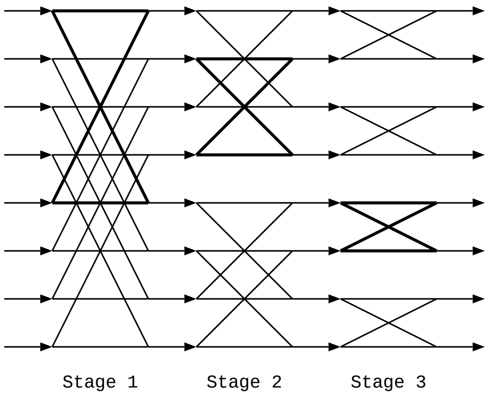


Figure 9: An example *butterfly* pattern at each of the three stages is shown in bold

1. Each butterfly requires one complex multiplication and two additions – hence FFT requires $(N/2)\log_2 N$ complex multiplications and $Nlog_2N$ complex additions – we have reduced the $O(N^2)$ DFT process to an $O(N\log_2 N)$ one.

2. At each iteration within each stage, we need consider only two input coefficients which generate two output coefficients – if we were to store the coefficients in an array, the two outputs can occupy the same locations as the two inputs (this of course destroys the input in the process). Even if we store the output in another array this algorithm is still only $O(N)$ in space terms.

3. To find the location of G_k in the FFT output array, take k as an binary number of $\log_2 N$ bits, reverse them and treat as index into array.

2.4 Inverse DFT by FFT

The Inverse DFT is given by:

$$g_n = \frac{1}{N}\sum_{k=0}^{N-1} G_k e^{\frac{2\pi i}{N}kn} \ , \ (n = 0, 1, \ldots, N-1)$$

This can be rewritten as:

$$Ng_n^* = \sum_{k=0}^{N-1} G_k^* \omega^{kn} \ , \ (n = 0, 1, \ldots, N-1)$$

This is seen to be a DFT of the complex conjugates of the Fourier coefficients; thus the FFT can be used as an inverse DFT after appropriate massaging of the coefficients on input and output.

2.5 More dimensions

When operating on images, we often wish to consider the two dimensional version of the Fourier Series / Transform / Discrete Transform. For the DFT that means describing a function of two variables $f(x, y)$ as components $e^{-2\pi i(nx/N+my/M)}$:

$$F_{k,l} = \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} f_{m,n} e^{-2\pi i(\frac{mk}{M}+\frac{nl}{N})}$$

Figure 10 shows some of the cosine basis functions. For a sequence of images, we might consider 3 dimensions (two space, one time).

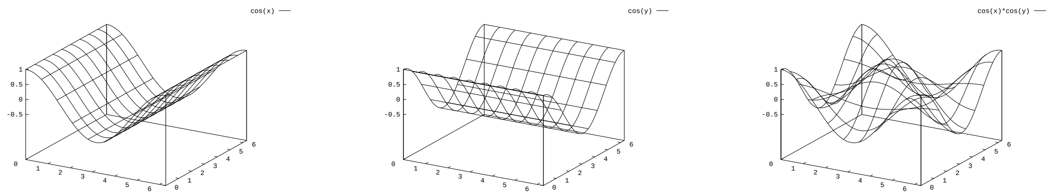


Figure 10: 2-D cosine basis functions

The FFT algorithm applies equally well in higher dimensions; for dimension d, we obtain an $O(N^d \log N)$ algorithm rather than the $O(N^{d+1})$.

3 Discrete Transforms and other animals

We have migrated from Fourier Series and Transforms of continuous space or time functions which express signals in frequency space to discrete transforms of discrete functions. However, we can recast the DFT in a more general framework. Working in two dimensions, we write $[f]$ to represent the matrix:

$$[f] = \begin{pmatrix} f_{0,0} & f_{0,1} & \cdots & f_{0,N-1} \\ f_{1,0} & f_{1,1} & \cdots & f_{1,N-1} \\ \vdots & & & \vdots \\ f_{M-1,0} & f_{M-1,1} & \cdots & f_{M-1,N-1} \end{pmatrix}$$

then writing $[]^{-1}$ for the inverse of the relevant matrix, the DFT pair is the obvious matrix products:

$$[F] = [\Phi_{M,M}][f][\Phi_{N,N}]$$

$$[f] = [\Phi_{M,M}]^{-1}[F][\Phi_{N,N}]^{-1}$$

where $[\Phi_{J,J}]$ is the $J \times J$ matrix with element (m, n) given by:

$$\frac{1}{J} e^{-\frac{2\pi i}{J} mn} \ , \ m, n = 0, 1, \ldots, J-1$$

In general then for non-singular matrices $[P]$ and $[Q]$, we can consider the generalised discrete transform pair given by:

$$[F] = [P][f][Q]$$

$$[f] = [P]^{-1}[F][Q]^{-1}$$

At first sight this style of presentation for a 2D function would appear to indicate that we must be dealing with an $O(N^3)$ process to evaluate the transform matrix, but in many cases of interest (as we have seen the FFT) use of symmetry and orthogonality relations provides a more efficient algorithm.

3.1 The Discrete Cosine Transform

The DCT is now widely used in image compression (JPEG and MPEG I&II); an image is carved up into small square tiles of $N \times N$ pixels and the DCT applied to each tile (after luminance/chrominance separation). By appropriate quantization and encoding of the coefficients a highly compressed form of the image can be generated (note that this is not usually a loss free process).

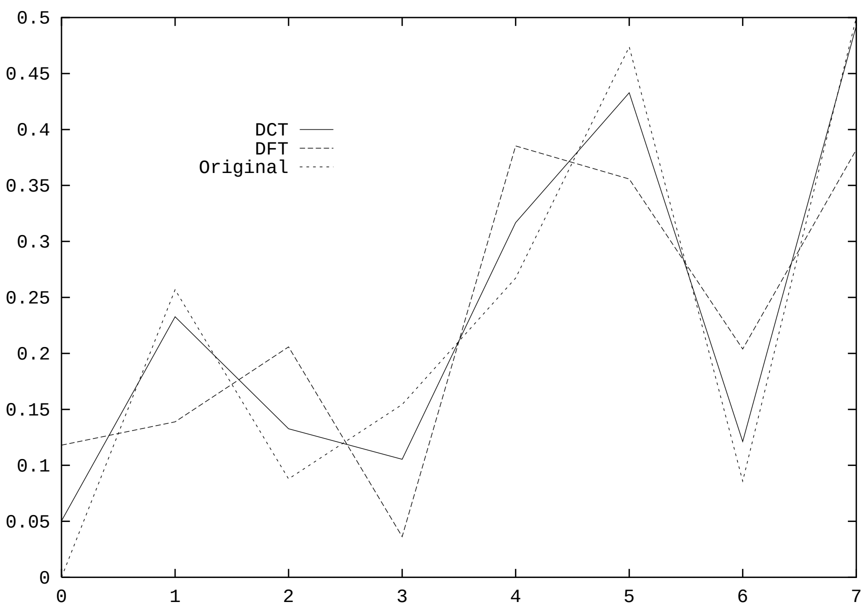


Figure 11: DCT v. DFT to avoid blocking artifacts - original signal is random

The tile is extended in a symmetric manner to provide a function which repeats in both x and y coordinates with period $2N$. This leads to only the cosine terms being generated by the DFT. The important property of this is that at the $N \times N$ pixel tile edges, the DCT ensures that $f(x_-) = f(x_+)$. Using the DFT on the original tile leads to *blocking artifacts*, where after inversion (if we have approximated the high frequency components), the pixel tile edges turn out to have the mean value of the pixels at the end of each row or column. Figure 11 shows the effect.

3.2 The Hadamard or Walsh Transform

A (2D) *Hadamard* matrix, $[H_{J,J}]$, is a symmetric $J \times J$ matrix whose elements are either ± 1 and where the rows (and hence necessarily columns) are mutually orthogonal. For example:

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{pmatrix}$$

As with the DFT, the Hadamard tranform has a *fast* version, as have Haar matrices (with coefficients $0, \pm 1$). Note that the transformations here are somewhat simpler to compute with than the complex terms of the DFT.

3.3 Orthonormal functions

In general we are looking for a set of orthornormal functions (we have used sine, cosine, complex expoentials, and finally Hadamard) with which too represent a function.

The examples given so far are all a prescribed set of functions independant of the function which we are trying to represent; however, it is also possible to select orthonormal basis functions which are optimised for a particular function, or more interestingly a family of functions.

The Karhunen-Loève theorem describes such a technique in which *eigenvectors (matrices)* of the autocorrelation function are the basis. It can be shown that this decomposition is the best achievable. However the computation of these eigenvectors is expensive, $O(N^4)$ for a 2D image, although faster versions do exists based on FFT – the derivation of which is not for the squeamish (see Rosenfeld and Kak, Digital Picture Processing).

3.4 Convolution

Many processes in signal recognition involve the use of convolution.

3.4.1 Gradient

Consider the problem of edge detection in a 2D image, where the edge is not necessarily vertical or horizontal. This involves a consideration of the gradient of the function $f(x,y)$ along paths on the surface; that is the vector of magnitude, $\sqrt{(\partial f/\partial x)^2 + (\partial f/\partial y)^2}$, and direction $\tan^{-1}((\partial f/\partial y)/(\partial f/\partial x))$.

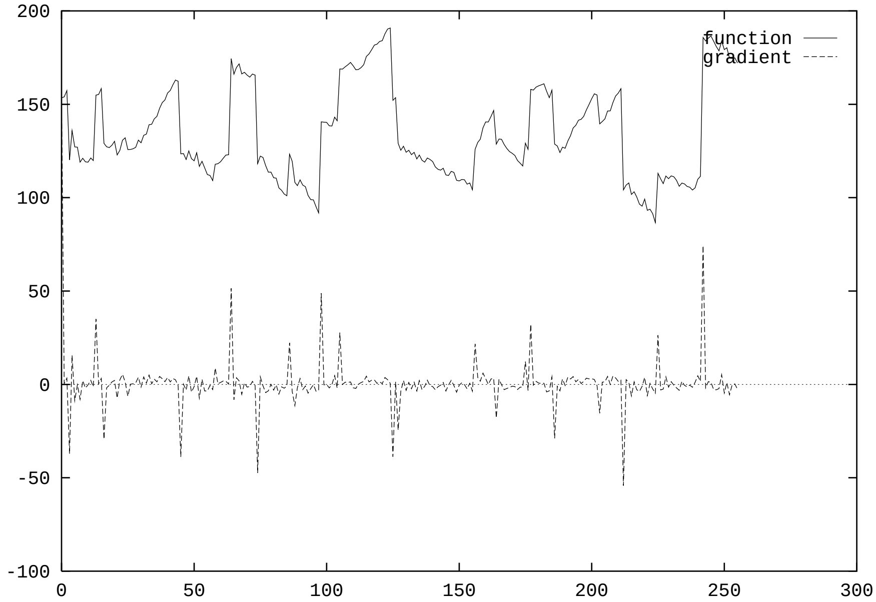


Figure 12: 1D edge detection by convolution

For a digital representation this indicates a difference equation:

$$\begin{aligned}(\Delta_x f)(x,y) &\equiv f(x,y) - f(x-1,y)\\(\Delta_y f)(x,y) &\equiv f(x,y) - f(x,y-1)\end{aligned}$$

Note that this is a convolution of f in the x and y directions with $\{g\}$ given by:

$$\{-1, 1, 0, 0, \ldots\}$$

3.4.2 Laplacian

We are interested in reconstructing a 2D image (which had initial pixel values $f(x,y)$), from blurred images observed some time later; the blurred image at time t is $g(x,y,t)$. We might model the bluring of an image as a diffusion process over time – if so we need to consider the solution to the well know partial differential equation involving the Laplacian:

$$\frac{\partial^2 g}{\partial x^2} + \frac{\partial^2 g}{\partial y^2} = \frac{1}{k}\frac{\partial g}{\partial t}$$

$$\nabla^2 g = \frac{1}{k}\frac{\partial g}{\partial t} \tag{36}$$

We had at time $t = 0$, $g(x, y, 0)$ the unblurred image of $f(x, y)$, but at some time $t = \tau$ we have observed the blurred picture $g(x, y, \tau)$. Using the Taylor expansion:

$$g(x, y, 0) = g(x, y, \tau) - \tau\frac{\partial g}{\partial t}(x, y, \tau) + \tau^2\frac{\partial^2 g}{\partial t^2}(x, y, \tau) - \cdots$$

To the first order we obtain:

$$f = g - k\tau\nabla^2 g$$

Again in difference equations:

$$\nabla^2 f(x, y) = f(x+1, y) + f(x-1, y) + f(x, y+1) + f(x, y-1) - 4f(x, y)$$

and the convolution kernel:

$$\begin{bmatrix} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{bmatrix}$$

3.4.3 Noise removal

If we know there is some periodic signal embedded in a very noisy signal, we can use the autocorrelation function to extract the signal. We take a single sine wave in uniform white noise with a SNR or 0.1 or -10dB:

The noise is of course uncorrelated at points other than $n = 0$. In the case of a set of superimposed periodic signals, we can use the autocorrelation function to remove the noise and then perform the DFT to extract the individual frequency components. Such techniques have applicability in passive sonar.

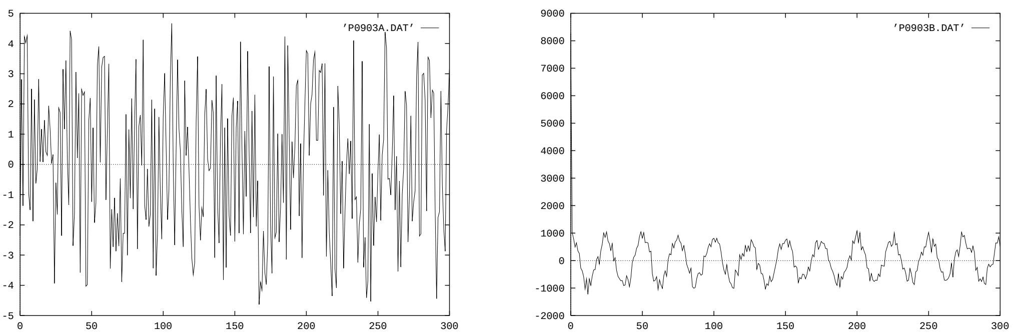


Figure 13: (a) sine wave in noise and (b) autocorrelation

10 Quantized Degrees-of-Freedom in a Continuous Signal

We have now encountered several theorems expressing the idea that even though a signal is continuous and dense in time (i.e. the value of the signal is defined at each real-valued moment in time), nevertheless a finite and countable set of discrete numbers suffices to describe it completely, and thus to reconstruct it, provided that its frequency bandwidth is limited.

Such theorems may seem counter-intuitive at first: How could a finite sequence of numbers, at discrete intervals, capture exhaustively the continuous and uncountable stream of numbers that represent all the values taken by a signal over some interval of time?

In general terms, the reason is that bandlimited continuous functions are *not as free* to vary as they might at first seem. Consequently, specifying their values at only certain points, suffices to *determine* their values at all other points.

Three examples that we have already seen are:

- **Nyquist's Sampling Theorem:** If a signal $f(x)$ is strictly bandlimited so that it contains no frequency components higher than W, i.e. its Fourier Transform $F(k)$ satisfies the condition
$$F(k) = 0 \text{ for } |k| > W \tag{1}$$
then $f(x)$ is completely determined just by sampling its values at a rate of at least $2W$. The signal $f(x)$ can be exactly recovered by using each sampled value to fix the amplitude of a $\text{sinc}(x)$ function,
$$\text{sinc}(x) = \frac{\sin(\pi x)}{\pi x} \tag{2}$$
whose width is scaled by the bandwidth parameter W and whose location corresponds to each of the sample points. The continuous signal $f(x)$ can be perfectly recovered from its discrete samples $f_n(\frac{n\pi}{W})$ just by adding all of those displaced $\text{sinc}(x)$ functions together, with their amplitudes equal to the samples taken:
$$f(x) = \sum_n f_n\left(\frac{n\pi}{W}\right)\frac{\sin(Wx - n\pi)}{(Wx - n\pi)} \tag{3}$$
Thus we see that any signal that is limited in its bandwidth to W, during some duration T has at most $2WT$ degrees-of-freedom. It can be completely specified by just $2WT$ real numbers (Nyquist, 1911; R V Hartley, 1928).

- **Logan's Theorem:** If a signal $f(x)$ is strictly bandlimited to one octave or less, so that the highest frequency component it contains is no greater than twice the lowest frequency component it contains
$$k_{max} \leq 2k_{min} \tag{4}$$
i.e. $F(k)$ the Fourier Transform of $f(x)$ obeys
$$F(|k| > k_{max} = 2k_{min}) = 0 \tag{5}$$
and
$$F(|k| < k_{min}) = 0 \tag{6}$$
and if it is also true that the signal $f(x)$ contains no complex zeroes in common with its Hilbert Transform (too complicated to explain here, but this constraint serves to

exclude families of signals which are merely amplitude-modulated versions of each other), then the original signal $f(x)$ can be perfectly recovered (up to an amplitude scale constant) merely from knowledge of the set $\{x_i\}$ of zero-crossings of $f(x)$ alone!

$$\{x_i\} \text{ such that } f(x_i) = 0 \tag{7}$$

Notes:
(1) This is a very complicated, and surprising, and rather recent result (W F Logan, 1977). (2) Only an existence theorem has been proven. There is so far no stable constructive algorithm for actually making this work – i.e. no known procedure that can actually recover $f(x)$ in all cases, within a scale factor, from the mere knowledge of its zero-crossings $f(x) = 0$; only the existence of such algorithms is proven. (3) The "Hilbert Transform" constraint (where the Hilbert Transform of a signal is obtained by convolving it with a hyperbola, $h(x) = 1/x$, or equivalently by shifting the phase of only the negative frequency components of the signal $f(x)$ by phase angle $-\pi$), serves to exclude ensembles of signals such as $a(x)\sin(\omega x)$ where $a(x)$ is a purely positive function $a(x) > 0$. Clearly $a(x)$ modulates the amplitudes of such signals, but it could not change any of their zero-crossings, which would always still occur at $x = 0, \frac{\pi}{\omega}, \frac{2\pi}{\omega}, \frac{3\pi}{\omega}, ...$, and so such signals could not be uniquely represented by their zero-crossings. (4) It is very difficult to see how to generalize Logan's Theorem to two-dimensional signals (such as images). In part this is because the zero-crossings of two-dimensional functions are non-denumerable (uncountable): they form continuous "snakes," rather than a discrete and countable set of points. Also, it is not clear whether the one-octave bandlimiting constraint should be isotropic (the same in all directions), in which case the projection of the signal's spectrum onto either frequency axis is really low-pass rather than bandpass; or anisotropic, in which case the projection onto both frequency axes may be strictly bandpass but the different directions are treated differently. (5) Logan's Theorem has been proposed as a significant part of a "brain theory" by David Marr and Tomaso Poggio, for how the brain's visual cortex processes and interprets retinal image information! The zero-crossings of bandpass-filtered retinal images constitute edge information within the image.

- **The Information Diagram:** The *Similarity Theorem* of Fourier Analysis asserts that if a function becomes narrower in one domain by a factor a, it necessarily becomes broader by the same factor a in the other domain:

$$f(x) \longrightarrow F(k) \tag{8}$$

$$f(ax) \longrightarrow |\frac{1}{a}|F(\frac{k}{a}| \tag{9}$$

The Hungarian Nobel-Laureate Dennis Gabor took this principle further with great insight and with implications that are still revolutionizing the field of signal processing (based upon wavelets), by noting that an *Information Diagram* representation of signals in a plane defined by the axes of time and frequency is fundamentally *quantized.* There is an irreducible, minimal, volume that any signal can possibly occupy in this plane. Its uncertainty (or spread) in frequency, times its uncertainty (or duration) in time, has an inescapable lower bound.

11 Gabor-Heisenberg-Weyl Uncertainty Relation. "Logons."

11.1 The Uncertainty Principle

If we define the "effective support" of a function $f(x)$ by its normalized variance, or the normalized second-moment

$$(\Delta x)^2 = \frac{\int_{-\infty}^{+\infty} f(x)f^*(x)(x-\mu)^2 dx}{\int_{-\infty}^{+\infty} f(x)f^*(x)dx} \tag{10}$$

where μ is the mean value, or first-moment, of the function

$$\mu = \int_{-\infty}^{+\infty} xf(x)f^*(x)dx \tag{11}$$

and if we similarly define the effective support of the Fourier Transform $F(k)$ of the function by its normalized variance in the Fourier domain

$$(\Delta k)^2 = \frac{\int_{-\infty}^{+\infty} F(k)F^*(k)(k-k_0)^2 dk}{\int_{-\infty}^{+\infty} F(k)F^*(k)dk} \tag{12}$$

where k_0 is the mean value, or first-moment, of the Fourier transform $F(k)$

$$k_0 = \int_{-\infty}^{+\infty} kF(k)F^*(k)dk \tag{13}$$

then it can be proven (by Schwartz Inequality arguments) that there exists a fundamental lower bound on the product of these two "spreads," *regardless* of the function $f(x)$!

$$\boxed{(\Delta x)(\Delta k) \geq \frac{1}{4\pi}} \tag{14}$$

This is the famous Gabor-Heisenberg-Weyl Uncertainty Principle. Mathematically it is exactly identical to the uncertainty relation in quantum physics, where (Δx) would be interpreted as the position of an electron or other particle, and (Δk) would be interpreted as its momentum or deBroglie wavelength. We see that this is not just a property of nature, but more abstractly a property of all functions and their Fourier Transforms. It is thus a still further, and more lofty, respect in which the information in continuous signals is quantized, since they must occupy an area in the Information Diagram (time - frequency axes) that is always greater than some irreducible lower bound.

11.2 Gabor "Logons"

Dennis Gabor named such minimal areas "logons" from the Greek word for information, or order: logōs. He thus established that the Information Diagram for any continuous signal can contain only a fixed number of information "quanta." Each such quantum constitutes an independent datum, and their total number within a region of the Information Diagram represents the number of independent degrees-of-freedom enjoyed by the signal.

The unique family of signals that actually achieve the lower bound in the Gabor-Heisenberg-Weyl Uncertainty Relation are the complex exponentials multiplied by Gaussians. These are sometimes referred to as "Gabor wavelets:"

$$f(x) = e^{-ik_0 x} e^{-(x-x_0)^2/a^2} \tag{15}$$

localized at "epoch" x_0, modulated by frequency k_0, and with size or spread constant a. It is noteworthy that such wavelets have Fourier Transforms $F(k)$ with exactly the same functional form, but with their parameters merely interchanged or inverted:

$$F(k) = e^{-ix_0 k} e^{-(k-k_0)^2 a^2} \tag{16}$$

Note that in the case of a wavelet (or wave-packet) centered on $x_0 = 0$, its Fourier Transform is simply a Gaussian centered at the modulation frequency k_0, and whose size is $1/a$, the reciprocal of the wavelet's space constant.

Because of the optimality of such wavelets under the Uncertainty Principle, Gabor (1946) proposed using them as an expansion basis to represent signals. In particular, he wanted them to be used in broadcast telecommunications for encoding continuous-time information. He called them the "elementary functions" for a signal. Unfortunately, because such functions are mutually non-orthogonal, it is very difficult to obtain the actual coefficients needed as weights on the elementary functions in order to expand a given signal in this basis! The first constructive method for finding such "Gabor coefficients" was developed in 1981 by the Dutch physicist Martin Bastiaans, using a dual basis and a complicated non-local infinite series.

The following diagrams show the behaviour of Gabor elementary functions both as complex wavelets, their separate real and imaginary parts, and their Fourier transforms. When a family of such functions are parameterized to be self-similar, i.e. they are dilates and translates of each other so that they all have a common template ("mother" and "daughter"), then they constitute a (non-orthogonal) *wavelet basis.* Today it is known that an infinite class of wavelets exist which can be used as the expansion basis for signals. Because of the self-similarity property, this amounts to representing or analyzing a signal at different scales. This general field of investigation is called *multi-resolution analysis.*

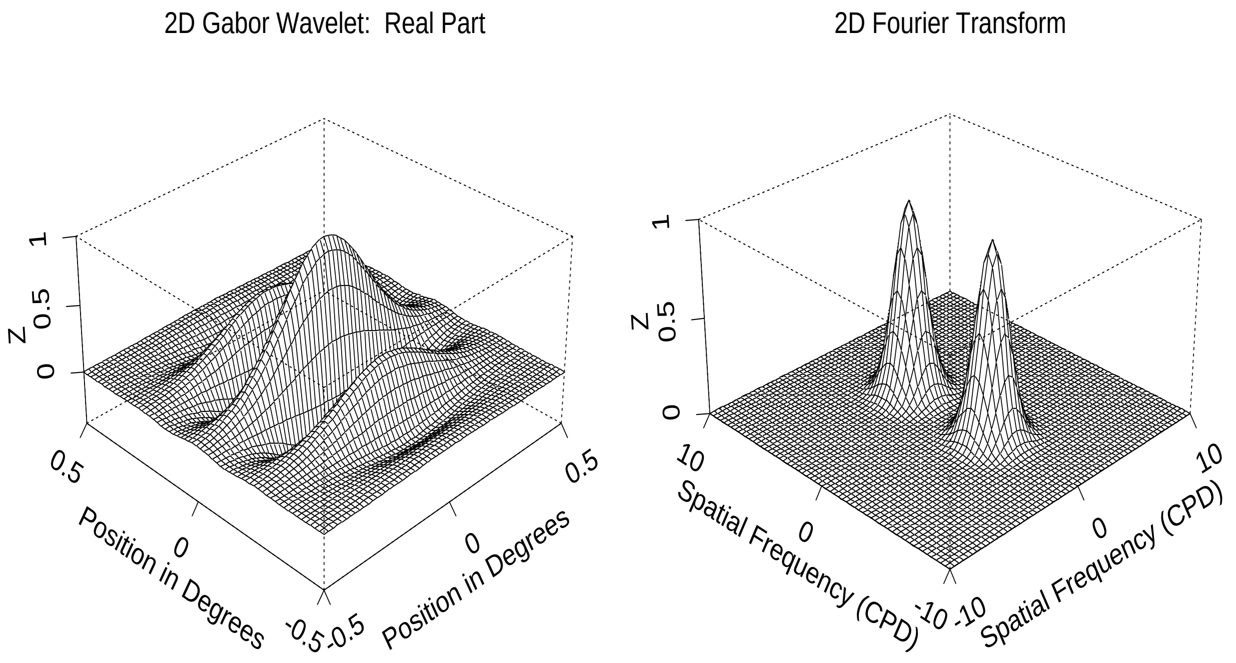


Figure 1: The real part of a 2-D Gabor wavelet, and its 2-D Fourier transform.

11.3 Generalization to Two Dimensional Signals

An effective strategy for extracting both coherent and incoherent image structure is the computation of two-dimensional Gabor coefficients for the image. This family of 2-D filters were originally proposed (Daugman, 1980) as a framework for understanding the orientation-selective and spatial-frequency-selective receptive field properties of neurons in the brain's visual cortex, and as useful operators for practical image analysis problems. These 2-D filters are conjointly optimal in providing the maximum possible resolution both for information about the spatial frequency and orientation of image structure (in a sense "what"), simultaneously with information about 2-D position ("where"). The 2-D Gabor filter family uniquely achieves the theoretical lower bound for joint uncertainty over these four variables, as dictated by the inescapable Uncertainty Principle when generalized to four-dimensional space.

These properties are particularly useful for texture analysis because of the 2-D spectral specificity of texture as well as its variation with 2-D spatial position. A rapid method for obtaining the required coefficients on these elementary expansion functions for the purpose of representing any image completely by its "2-D Gabor Transform," despite the non-orthogonality of the expansion basis, is possible through the use of a relaxation neural network. A large and growing literature now exists on the efficient use of this non-orthogonal expansion basis and its applications.

Two-dimensional Gabor filters over the image domain (x, y) have the functional form

$$f(x,y) = e^{-\pi\left[(x-x_0)^2/\alpha^2+(y-y_0)^2/\beta^2\right]}e^{-2\pi i[u_0(x-x_0)+v_0(y-y_0)]} \tag{17}$$

where (x_0, y_0) specify position in the image, (α, β) specify effective width and length, and (u_0, v_0) specify modulation, which has spatial frequency $\omega_0 = \sqrt{u_0^2 + v_0^2}$ and direction $\theta_0 = \arctan(v_0/u_0)$. (A further degree-of-freedom not included above is the relative orientation of the elliptic Gaussian envelope, which creates cross-terms in xy.) The 2-D Fourier transform $F(u, v)$ of a 2-D Gabor filter has exactly the same functional form, with parameters just interchanged or inverted:

$$F(u,v) = e^{-\pi\left[(u-u_0)^2\alpha^2+(v-v_0)^2\beta^2\right]}e^{-2\pi i[x_0(u-u_0)+y_0(v-v_0)]} \tag{18}$$

The real part of one member of the 2-D Gabor filter family, centered at the origin $(x_0, y_0) = (0,0)$ and with unity aspect ratio $\beta/\alpha = 1$ is shown in the figure, together with its 2-D Fourier transform $F(u,v)$.

2-D Gabor functions can form a complete self-similar 2-D wavelet expansion basis, with the requirements of orthogonality and strictly compact support relaxed, by appropriate parameterization for dilation, rotation, and translation. If we take $\Psi(x,y)$ to be a chosen generic 2-D Gabor wavelet, then we can generate from this one member a complete self-similar family of 2-D wavelets through the generating function:

$$\Psi_{mpq\theta}(x,y) = 2^{-2m}\Psi(x',y') \tag{19}$$

where the substituted variables (x', y') incorporate dilations in size by 2^{-m}, translations in position (p, q), and rotations through orientation θ:

$$x' = 2^{-m}[x\cos(\theta) + y\sin(\theta)] - p \tag{20}$$

$$y' = 2^{-m}[-x\sin(\theta) + y\cos(\theta)] - q \tag{21}$$

It is noteworthy that as consequences of the similarity theorem, shift theorem, and modulation theorem of 2-D Fourier analysis, together with the rotation isomorphism of the 2-D Fourier transform, all of these effects of the generating function applied to a 2-D Gabor mother wavelet $\Psi(x,y) = f(x,y)$ have corresponding identical or reciprocal effects on its 2-D Fourier transform $F(u,v)$. These properties of self-similarity can be exploited when constructing efficient, compact, multi-scale codes for image structure.

11.4 Grand Unification of Domains: an *Entente Cordiale*

Until now we have viewed "the space domain" and "the Fourier domain" as somehow opposite, and incompatible, domains of representation. (Their variables are reciprocals; and the Uncertainty Principle declares that improving the resolution in either domain must reduce it in the other.) But we now can see that the "Gabor domain" of representation actually embraces and unifies both of these other two domains! To compute the representation of a signal or of data in the Gabor domain, we find its expansion in terms of elementary functions having the form

$$f(x) = e^{-ik_0x}e^{-(x-x_0)^2/a^2} \tag{22}$$

The single parameter a (the space-constant in the Gaussian term) actually builds a continuous bridge between the two domains: if the parameter a is made very large, then the second exponential above approaches 1.0, and so in the limit our expansion basis becomes

$$\lim_{a\to\infty} f(x) = e^{-ik_0x} \tag{23}$$

the ordinary Fourier basis! If the parameter a is instead made very small, the Gaussian term becomes the approximation to a delta function at location x_o, and so our expansion basis implements pure space-domain sampling:

$$\lim_{a\to 0} f(x) = \delta(x - x_0) \tag{24}$$

Hence the Gabor expansion basis "contains" both domains at once. It allows us to make a continuous deformation that selects a representation lying anywhere on a one-parameter continuum between two domains that were hitherto distinct and mutually unapproachable. A new *Entente Cordiale*, indeed.

Reconstruction of Lena: 25, 100, 500, and 10,000 Two-Dimensional Gabor Wavelets

Figure 2: Ilustration of the completeness of 2-D Gabor wavelets as basis functions.

12 Kolmogorov Complexity and Minimal Description Length

An idea of fundamental importance is the measure known as Kolmogorov complexity: the complexity of a string of data is defined as the length of the shortest binary program for computing the string. Thus the complexity is the data's "minimal description length."

It is an amazing fact that the Kolmogorov complexity K of a string is approximately equal to the entropy H of the distribution from which the string is a randomly drawn sequence. Thus Kolmogorov descriptive complexity is intimately connected with information theory, and indeed K defines the ultimate data compression. Reducing the data to a program that generates it exactly is obviously a way of compressing it; and running that program is a way of decompressing it. Any set of data can be generated by a computer program, even if (in the worst case) that program simply consists of data statements. The length of such a program defines its algorithmic complexity.

It is important to draw a clear distinction between the notions of *computational complexity* (measured by program execution time), and *algorithmic complexity* (measured by program length). Kolmogorov complexity is concerned with finding descriptions which minimize the latter. Little is known about how (in analogy with the optimal properties of Gabor's elementary logons in the 2D Information Plane) one might try to minimize simultaneously along *both* of these orthogonal axes that form a "Complexity Plane."

Most sequences of length n (where "most" considers all possible permutations of n bits) have Kolmogorov complexity K close to n. The complexity of a truly random binary sequence is as long as the sequence itself. However, it is not clear how to be certain of discovering that a given string has a much lower complexity than its length. It might be clear that the string

```
0101010101010101010101010101010101010101010101010101010101010101
```

has a complexity much less than 32 bits; indeed, its complexity is the length of the program: `Print 32 "01"s`. But consider the string

```
0110101000001001111001100110011111110011101111001100100100001000
```

which looks random and passes most tests for randomness. How could you discover that this sequence is in fact just the binary expansion for the irrational number $\sqrt{2}-1$, and that therefore it can be specified extremely concisely?

Fractals are examples of entities that look very complex but in fact are generated by very simple programs (i.e. iterations of a mapping). Therefore, the Kolmogorov complexity of fractals is nearly zero.

A sequence $x_1, x_2, x_3, ..., x_n$ of length n is said to be *algorithmically random* if its Kolmogorov complexity is at least n (i.e. the shortest possible program that can generate the sequence is a listing of the sequence itself):

$$K(x_1x_2x_3...x_n|n) \geq n \tag{25}$$

An infinite string is defined to be *incompressible* if its Kolmogorov complexity, in the limit as the string gets arbitrarily long, approaches the length n of the string itself:

$$\lim_{n\to\infty} \frac{K(x_1x_2x_3...x_n|n)}{n} = 1 \tag{26}$$

An interesting theorem, called the *Strong Law of Large Numbers for Incompressible Sequences,* asserts that the proportions of 0's and 1's in any incompressible string must be nearly equal! Moreover, any incompressible sequence must satisfy all computable statistical tests for randomness. (Otherwise, identifying the statistical test for randomness that the string failed would reduce the descriptive complexity of the string, which contradicts its incompressibility.) Therefore the algorithmic test for randomness is the ultimate test, since it includes within it all other computable tests for randomness.

13 Information in Time Series

Data unfolding in time as a random variable, or a series of random variables, are called *time series.* This topic represents an important unity between information theory and statistics, and the definitive treatment of this field was developed by Norber Wiener (1949). We will only introduce the notion of a Poisson Point Process in order to ask how different aspects of such a time series (pulse timing distributions; pulse count distributions) might be used as ways of encoding and transmitting information. This subject is immensely important today in *computational neuroscience*, since nerves are only able to compute and to communicate with each other by means of nerve impulses which constitute time series ("point processes").

14 A Short Bibliography

Cover, T M and Thomas, J A (1991) *Elements of Information Theory.* New York: Wiley.

McEliece, R J (1977) *The Theory of Information and Coding: A Mathematical Framework for Communication.* Addison-Wesley. Reprinted (1984) by Cambridge University Press in *Encyclopedia of Mathematics.*

Blahut, R E (1987) *Principles and Practice of Information Theory.* New York: Addison-Wesley.

Shannon, C and Weaver, W (1949) *Mathematical Theory of Communication.* Urbana: University of Illinois Press.

Wiener, N (1948) *Time Series.* Cambridge: MIT Press.

15 Some Final Thoughts About Redundancy

"Having chosen English as the preferred language in the EEC, the European Parliment has commissioned a feasibility study in ways of improving efficiency in communications between Government departments.

European officials have often pointed out that English spelling is unnecessarily difficult; for example cough, plough, rough, through and thorough. What is clearly needed is a phased programme of changes to iron out these anomalies. The programme would, of course, be administered by a committee staff at top level by participating nations.

In the first year, for example, the committee would suggest using 's' instead of the soft 'c'. Sertainly, sivil servants in all sites would reseive this news with joy. Then the hard 'c' could be replaced by 'k' sinse both letters are pronounsed alike. Not only would this klear up konfusion in the minds of klerikal workers, but typewriters kould be made with one less letter.

There would be growing enthusiasm when in the sekond year, it is anounsed that the troublesome 'ph' would henseforth be written 'f'. This would make words like 'fotograf' twenty persent shorter in print.

In the third year, publik akseptanse of the new spelling kan be expekted to reach the stage where more komplikated shanges are possible. Governments would enkourage the removal of double leters which have always ben a deterent to akurate speling.

We would al agre that the horible mes of silent 'e's in the language is disgrasful. Therefor we kould drop thes and kontinu to read and writ as though nothing had hapend. By this tim it would be four years sins the skem began and peopl would be reseptiv to steps sutsh as re-plasing 'th' by 'z'. Perhaps zen ze funktion of 'w' kould be taken on by 'v', vitsh is, after al, half a 'w'. Shortly after zis, ze unesary 'o' kuld b dropd from words kontaining 'ou'. Similar arguments vud of kors be aplid to ozer kombinations of leters.

Kontinuing zis proces yer after yer, ve vud eventuli hav a reli sensibl riten styl. After tventi yers zer vud b no more trubls, difikultis and evrivun vud fin it ezi tu understand ech ozer. Ze drems of ze Guvermnt vud finali hav kum tru."