Learning Guide and Examples: Information Theory and Coding

12 lectures by J Daugman, Michaelmas Term

Prerequisite courses: Continuous Mathematics, Probability, Discrete Mathematics

Overview and Historical Origins: Foundations and Uncertainty. Why the movements and transformations of information, just like those of a fluid, are law-governed. How concepts of randomness, redundancy, compressibility, noise, bandwidth, and uncertainty are intricately connected to information. Origins of these ideas and the various forms that they take.

Mathematical Foundations; Probability Rules; Bayes' Theorem. The meanings of probability. Ensembles, random variables, marginal and conditional probabilities. How the formal concepts of information are grounded in the principles and rules of probability.

Entropies Defined, and Why They Are Measures of Information. Marginal entropy, joint entropy, conditional entropy, and the Chain Rule for entropy. Mutual information between ensembles of random variables. Why entropy is a fundamental measure of information content.

Source Coding Theorem; Prefix, Variable-, & Fixed-Length Codes. Symbol codes. Binary symmetric channel. Capacity of a noiseless discrete channel. Error correcting codes.

Channel Types, Properties, Noise, and Channel Capacity. Perfect communication through a noisy channel. Capacity of a discrete channel as the maximum of its mutual information over all possible input distributions.

Continuous Information; Density; Noisy Channel Coding Theorem. Extensions of the discrete entropies and measures to the continuous case. Signal-to-noise ratio; power spectral density. Gaussian channels. Relative significance of bandwidth and noise limitations. The Shannon rate limit and efficiency for noisy continuous channels.

Fourier Series, Convergence, Orthogonal Representation. Generalized signal expansions in vector spaces. Independence. Representation of continuous or discrete data by complex exponentials. The Fourier basis. Fourier series for periodic functions. Examples.

Useful Fourier Theorems; Transform Pairs. Sampling; Aliasing. The Fourier transform for non-periodic functions. Properties of the transform, and examples. Nyquist's Sampling Theorem derived, and the cause (and removal) of aliasing.

Discrete Fourier Transform. Fast Fourier Transform Algorithms. Efficient algorithms for computing Fourier transforms of discrete data. Computational complexity. Filters, correlation, modulation, demodulation, coherence.

The Quantized Degrees-of-Freedom in a Continuous Signal. Why a continuous signal of finite bandwidth and duration has a fixed number of degrees-of-freedom. Diverse illustrations of the principle that information, even in such a signal, comes in quantized, countable, packets.

Gabor-Heisenberg-Weyl Uncertainty Relation. Optimal "Logons." Unification of the time-domain and the frequency-domain as endpoints of a continuous deformation. The Uncertainty Principle and its optimal solution by Gabor's expansion basis of "logons." Multi-resolution wavelet codes. Extension to images, for analysis and compression.

Kolmogorov Complexity and Minimal Description Length. Definition of the algorithmic complexity of a data sequence, and its relation to the entropy of the distribution from which the data was drawn. Shortest possible description length, and fractals.

Recommended book:

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

Worked Example Problems

Information Theory and Coding: Example Problem Set 1

Let X and Y represent random variables with associated probability distributions $p(x)$ and $p(y)$, respectively. They are not independent. Their conditional probability distributions are $p(x|y)$ and $p(y|x)$, and their joint probability distribution is $p(x, y)$.

1. What is the marginal entropy $H(X)$ of variable X, and what is the mutual information of X with itself?

2. In terms of the probability distributions, what are the conditional entropies $H(X|Y)$ and $H(Y|X)$?

3. What is the joint entropy $H(X, Y)$, and what would it be if the random variables X and Y were independent?

4. Give an alternative expression for $H(Y) - H(Y|X)$ in terms of the joint entropy and both marginal entropies.

5. What is the mutual information $I(X; Y)$?

Model Answer – Example Problem Set 1

1. $H(X) = -\sum_x p(x) \log_2 p(x)$ is both the marginal entropy of X, and its mutual information with itself.

2. $H(X|Y) = -\sum_y p(y) \sum_x p(x|y) \log_2 p(x|y) = -\sum_x \sum_y p(x,y) \log_2 p(x|y)$
   $H(Y|X) = -\sum_x p(x) \sum_y p(y|x) \log_2 p(y|x) = -\sum_x \sum_y p(x,y) \log_2 p(y|x)$

3. $H(X,Y) = -\sum_x \sum_y p(x,y) \log_2 p(x,y)$.
   If X and Y were independent random variables, then $H(X,Y) = H(X) + H(Y)$.

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

5. $I(X;Y) = \sum_x \sum_y p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$
   or: $\sum_x \sum_y p(x,y) \log_2 \frac{p(x|y)}{p(x)}$
   or: $I(X;Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X,Y)$

Information Theory and Coding: Example Problem Set 2

1. This is an exercise in manipulating conditional probabilities. Calculate the probability that if somebody is "tall" (meaning taller than 6 ft or whatever), that person must be male. Assume that the probability of being male is $p(M) = 0.5$ and so likewise for being female $p(F) = 0.5$. Suppose that 20% of males are T (i.e. tall): $p(T|M) = 0.2$; and that 6% of females are tall: $p(T|F) = 0.06$. So this exercise asks you to calculate $p(M|T)$.

If you know that somebody is male, how much information do you gain (in bits) by learning that he is also tall? How much do you gain by learning that a female is tall? Finally, how much information do you gain from learning that a tall person is female?

2. The input source to a noisy communication channel is a random variable X over the four symbols a, b, c, d. The output from this channel is a random variable Y over these same four symbols. The joint distribution of these two random variables is as follows:

	$x = a$	$x = b$	$x = c$	$x = d$
$y = a$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{4}$
$y = b$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$	0
$y = c$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{16}$	0
$y = d$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{16}$	0

(a) Write down the marginal distribution for X and compute the marginal entropy $H(X)$ in bits.

(b) Write down the marginal distribution for Y and compute the marginal entropy $H(Y)$ in bits.

(c) What is the joint entropy $H(X, Y)$ of the two random variables in bits?

(d) What is the conditional entropy $H(Y|X)$ in bits?

(e) What is the mutual information $I(X; Y)$ between the two random variables in bits?

(f) Provide a lower bound estimate of the channel capacity C for this channel in bits.

Model Answer – Example Problem Set 2

1. Bayes' Rule, combined with the Product Rule and the Sum Rule for manipulating conditional probabilities (see pages 7 - 9 of the Notes), enables us to solve this problem. First we must calculate the marginal probability of someone being tall:

$$p(T) = p(T|M)p(M) + p(T|F)p(F) = (0.2)(0.5) + (0.06)(0.5) = 0.13$$

Now with Bayes' Rule we can arrive at the answer that:

$$p(M|T) = \frac{p(T|M)p(M)}{p(T)} = \frac{(0.2)(0.5)}{(0.13)} = \underline{0.77}$$

The information gained from an event is -$\log_2$ of its probability.

Thus the information gained from learning that a male is tall, since $p(T|M) = 0.2$, is 2.32 bits.

The information gained from learning that a female is tall, since $p(T|F) = 0.06$, is 4.06 bits.

Finally, the information gained from learning that a tall person is female, which requires us to calculate the fact (again using Bayes' Rule) that $p(F|T) = 0.231$, is 2.116 bits.

2. (a) Marginal distribution for X is $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$.

Marginal entropy of X is 1/2 + 1/2 + 1/2 + 1/2 = 2 bits.

(b) Marginal distribution for Y is $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$.

Marginal entropy of Y is 1/2 + 1/2 + 3/8 + 3/8 = 7/4 bits.

(c) Joint Entropy: sum of $-p \log p$ over all 16 probabilities in the joint distribution (of which only 4 different non-zero values appear, with the following frequencies): (1)(2/4) + (2)(3/8) + (6)(4/16) + (4)(5/32) = 1/2 + 3/4 + 3/2 + 5/8 = 27/8 bits.

(d) Conditional entropy $H(Y|X)$: (1/4)H(1/2, 1/4, 1/8, 1/8) + (1/4)H(1/4, 1/2, 1/8, 1/8) + (1/4)H(1/4, 1/4, 1/4, 1/4) + (1/4)H(1, 0, 0, 0) = (1/4)(1/2 + 2/4 + 3/8 + 3/8) + (1/4)(2/4 + 1/2 + 3/8 + 3/8) + (1/4)(2/4 + 2/4 + 2/4 + 2/4) + (1/4)(0) = (1/4)(7/4) + (1/4)(7/4) + 1/2 + 0 = (7/8) + (1/2) = 11/8 bits.

(e) There are three alternative ways to obtain the answer:
$I(X;Y) = H(Y) - H(Y|X) = 7/4$ - $11/8 = \underline{3/8 \text{ bits}}$. - Or,

$I(X;Y) = H(X) - H(X|Y) = 2$ - $13/8 = 3/8$ bits. - Or,

$I(X;Y) = H(X) + H(Y) - H(X,Y) = 2 + 7/4$ - $27/8 = (16+14-27)/8 = 3/8$ bits.

(f) Channel capacity is the maximum, over all possible input distributions, of the mutual information that the channel establishes between the input and the output. So one lower bound estimate is simply any particular measurement of the mutual information for this channel, such as the above measurement which was $\underline{3/8 \text{ bits}}$.

Information Theory and Coding: Example Problem Set 3

A. Consider a binary symmetric communication channel, whose input source is the alphabet $X = \{0, 1\}$ with probabilities $\{0.5, 0.5\}$; whose output alphabet is $Y = \{0, 1\}$; and whose channel matrix is

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

where ϵ is the probability of transmission error.

1. What is the entropy of the source, $H(X)$?

2. What is the probability distribution of the outputs, $p(Y)$, and the entropy of this output distribution, $H(Y)$?

3. What is the joint probability distribution for the source and the output, $p(X, Y)$, and what is the joint entropy, $H(X, Y)$?

4. What is the mutual information of this channel, $I(X; Y)$?

5. How many values are there for ϵ for which the mutual information of this channel is maximal? What are those values, and what then is the capacity of such a channel in bits?

6. For what value of ϵ is the capacity of this channel minimal? What is the channel capacity in that case?

B. The Fourier transform (whether continuous or discrete) is defined in the general case for complex-valued data, which gets mapped into a set of complex-valued Fourier coefficients. But often we are concerned with purely real-valued data, such as sound waves or images, whose Fourier transforms we would like to compute. What simplification occurs in the Fourier domain as a consequence of having real-valued, rather than complex-valued, data?

Model Answer – Example Problem Set 3

A.
1. Entropy of the source, $H(X)$, is 1 bit.

2. Output probabilities are $p(y = 0) = (0.5)(1 - \epsilon) + (0.5)\epsilon = 0.5$ and $p(y = 1) = (0.5)(1 - \epsilon) + (0.5)\epsilon = 0.5$. Entropy of this distribution is $H(Y) = 1$ bit, just as for the entropy $H(X)$ of the input distribution.

3. Joint probability distribution $p(X, Y)$ is

$$\begin{pmatrix} 0.5(1-\epsilon) & 0.5\epsilon \\ 0.5\epsilon & 0.5(1-\epsilon) \end{pmatrix}$$

and the entropy of this joint distribution is $H(X,Y) = -\sum_{x,y} p(x,y) \log_2 p(x,y)$
$= -(1-\epsilon)\log(0.5(1-\epsilon)) - \epsilon\log(0.5\epsilon) = (1-\epsilon) - (1-\epsilon)\log(1-\epsilon) + \epsilon - \epsilon\log(\epsilon)$

$= 1 - \epsilon\log(\epsilon) - (1-\epsilon)\log(1-\epsilon)$

4. The mutual information is $I(X;Y) = H(X) + H(Y) - H(X,Y)$, which we can evaluate from the quantities above as: $1 + \epsilon\log(\epsilon) + (1-\epsilon)\log(1-\epsilon)$.

5. In the two cases of $\epsilon = 0$ and $\epsilon = 1$ (perfect transmission, and perfectly erroneous transmission), the mutual information reaches its maximum of 1 bit and this is also then the channel capacity.

6. If $\epsilon = 0.5$, the channel capacity is minimal and equal to 0.

B. Real-valued data produces a Fourier transform having Hermitian symmetry: the real-part of the Fourier transform has even-symmetry, and the imaginary part has odd-symmetry. Therefore we need only compute the coefficients associated with (say) the positive frequencies, because then we automatically know the coefficients for the negative frequencies as well. Hence the two-fold "reduction" in the input data by being real- rather than complex-valued, is reflected by a corresponding two-fold "reduction" in the amount of data required in its Fourier representation.

Information Theory and Coding: Example Problem Set 4

1. Consider a noiseless analog communication channel whose bandwidth is 10,000 Hertz. A signal of duration 1 second is received over such a channel. We wish to represent this continuous signal exactly, at all points in its one-second duration, using just a finite list of real numbers obtained by sampling the values of the signal at discrete, periodic points in time. What is the length of the shortest list of such discrete samples required in order to guarantee that we capture all of the information in the signal and can recover it exactly from this list of samples?

2. Name, define algebraically, and sketch a plot of the function you would need to use in order to recover completely the continuous signal transmitted, using just such a finite list of discrete periodic samples of it.

3. Consider a noisy analog communication channel of bandwidth Ω, which is perturbed by additive white Gaussian noise whose power spectral density is N_0. Continuous signals are transmitted across such a channel, with average transmitted power P (defined by their expected variance). What is the channel capacity, in bits per second, of such a channel?

Model Answer – Example Problem Set 4

1. 20,000 discrete samples are required.
2. The sinc function is required to recover the signal from its discrete samples, defined as: $\text{sinc}(x) = \dfrac{\sin(\pi x)}{\pi x}$
3. The channel capacity is $\Omega \log_2 \left(1 + \dfrac{P}{N_0 \Omega}\right)$ bits per second.

Information Theory and Coding: Example Problem Set 5

A. Consider Shannon's third theorem, the *Channel Capacity Theorem,* for a continuous communication channel having bandwidth W Hertz, perturbed by additive white Gaussian noise of power spectral density N_0, and average transmitted power P.

1. Is there any limit to the capacity of such a channel if you increase its signal-to-noise ratio $\frac{P}{N_0 W}$ without limit? If so, what is that limit?

2. Is there any limit to the capacity of such a channel if you can increase its bandwidth W in Hertz without limit, but while not changing N_0 or P? If so, what is that limit?

B. Explain why smoothing a signal, by low-pass filtering it *before* sampling it, can prevent aliasing. Explain aliasing by a picture in the Fourier domain, and also show in the picture how smoothing solves the problem. What would be the most effective low-pass filter to use for this purpose? Draw its spectral sensitivity.

C. Suppose that women who live beyond the age of 70 outnumber men in the same age bracket by three to one. How much information, in bits, is gained by learning that a certain person who lives beyond 70 happens to be male?

Model Answer – Example Problem Set 5

A.

1. The capacity of such a channel, in bits per second, is

$$C = W \log_2 \left(1 + \frac{P}{N_0 W}\right)$$

Increasing the quantity $\frac{P}{N_0 W}$ inside the logarithm without bounds causes the capacity to increase monotonically and without bounds.

2. Increasing the bandwidth W alone causes a monotonic increase in capacity, but only up to an asymptotic limit. That limit can be evaluated by observing that in the limit of small α, the quantity $\ln(1+\alpha)$ approaches α. In this case, setting $\alpha = \frac{P}{N_0 W}$ and allowing W to become arbitrarily large, C approaches the limit $\frac{P}{N_0} \log_2(e)$. Thus there are vanishing returns from endless increase in bandwidth, unlike the unlimited returns enjoyed from improvement in signal-to-noise ratio.

B.

The Nyquist Sampling Theorem tells us that aliasing results when the signal contains Fourier components higher than one-half the sampling frequency. Thus aliasing can be avoided by removing such frequency components from the signal, by low-pass filtering it, before sampling the signal. The ideal low-pass filter for this task would have a strict cut-off at frequencies starting at (and higher than) one-half the planned sampling rate.

C.

Since p(female|old)=3*p(male|old), and since p(female|old)+p(male|old)=1, it follows that p(male|old) = 0.25. The information gained from an observation is $-\log_2$ of its probability. Thus the information gained by such an observation is 2 bits.

Information Theory and Coding: Example Problem Set 6

The information in continuous but bandlimited signals is *quantized*, in that such continuous signals can be completely represented by a finite set of discrete numbers. Explain this principle in each of the following four important contexts or theorems. Be as quantitative as possible:

1. The Nyquist Sampling Theorem.

2. Logan's Theorem.

3. Gabor Wavelet Logons and the Information Diagram.

4. The Noisy Channel Coding Theorem
   (relation between channel bandwidth W, noise power spectral density N_0, signal power P or signal-to-noise ratio P/N_0W, and channel capacity C in bits/second).

Model Answer – Example Problem Set 6

1. 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.

2. 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, 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.

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

Obviously there is only a finite and countable number of zero-crossings in any given length of the bandlimited signal, and yet these "quanta" suffice to recover the original continuous signal completely (up to a scale constant).

3. Gabor Wavelet Logons and 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:

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

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

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

$$(\Delta x)(\Delta k) \geq \frac{1}{4\pi} \tag{10}$$

This is the Gabor-Heisenberg-Weyl Uncertainty Principle. It is another 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. Therefore any continuous signal can contain only a fixed number of information "quanta" in the Information Diagram. 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. Dennis Gabor named such minimal areas "logons." 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:"

$$f(x) = e^{-ik_0 x} e^{-(x-x_0)^2/a^2} \tag{11}$$

localized at epoch x_0, modulated by frequency k_0, and with size constant a.

4. The Noisy Channel Coding Theorem asserts that for a channel with bandwidth W, and a continuous input signal of average power P, added channel noise of power spectral density N_0, or a signal-to-noise ratio P/N_0W, the capacity of the channel to communicate information reliably is limited to a discrete number of "quanta" per second. Specifically, its capacity C in bits/second is:

$$C = W \log_2 \left(1 + \frac{P}{N_0 W}\right) \tag{12}$$

This capacity is clearly "quantized" into a finite number of bits per second, even though the input signal is continuous.

Information Theory and Coding: Example Problem Set 7

(a) What is the entropy H, in bits, of the following source alphabet whose letters have the probabilities shown?

A	B	C	D
1/4	1/8	1/2	1/8

(b) Why are fixed length codes inefficient for alphabets whose letters are not equiprobable? Discuss this in relation to Morse Code.

(c) Offer an example of a uniquely decodable prefix code for the above alphabet which is optimally efficient. What features make it a uniquely decodable prefix code?

(d) What is the coding rate R of your code? How do you know whether it is optimally efficient?

(e) What is the maximum possible entropy H of an alphabet consisting of N different letters? In such a maximum entropy alphabet, what is the probability of its most likely letter? What is the probability of its least likely letter?

Model Answer – Example Problem Set 7

(a) The entropy of the source alphabet is

$$H = -\sum_{i=1}^{4} p_i \log_2 p_i = (1/4)(2) + (1/8)(3) + (1/2)(1) + (1/8)(3)$$

$= \underline{1.75 \text{ bits}}$.

(b) Fixed length codes are inefficient for alphabets whose letters are not equiprobable because the cost of coding improbable letters is the same as that of coding more probable ones. It is more efficient to allocate fewer bits to coding the more probable letters, and to make up for the fact that this would cover only a few letters, by making longer codes for the less probable letters. This is exploited in Morse Code, in which (for example) the most probable English letter, e, is coded by a single dot.

(c) A uniquely decodable prefix code for the letters of this alphabet:
Code for A: 10
Code for B: 110
Code for C: 0
Code for D: 111 (the codes for B and D could also be interchanged)

This is a uniquely decodable prefix code because even though it has variable length, each code corresponds to a unique letter rather than any possible combination of letters; and the code for no letter could be confused as the prefix for another letter.

(d) Multiplying the bit length of the code for each letter times the probability of occurence of that letter, and summing this over all letters, gives us a coding rate of:
R = (2 bits)(1/4)+(3 bits)(1/8)+(1 bit)(1/2)+(3 bits)(1/8) = $\underline{1.75 \text{ bits}}$.

This code is optimally efficient because $R = H$: its coding rate equals the entropy of the source alphabet. Shannon's Source Coding Theorem tells us that this is the lower bound for the coding rate of all possible codes for this alphabet.

(e) The maximum possible entropy of an alphabet consisting of N different letters is $H = \log_2 N$. This is only achieved if the probability of every letter is $1/N$. Thus $1/N$ is the probability of both the "most likely" and the "least likely" letter.

Information Theory and Coding: Example Problem Set 8

(a) What class of continuous signals has the greatest possible entropy for a given variance (or power level)? What probability density function describes the excursions taken by such signals from their mean value?

(b) What does the Fourier power spectrum of this class of signals look like? How would you describe the entropy of this distribution of spectral energy?

(c) An error-correcting Hamming code uses a 7 bit block size in order to guarantee the detection, and hence the correction, of any single bit error in a 7 bit block. How many bits are used for error correction, and how many bits for useful data? If the probability of a single bit error within a block of 7 bits is $p = 0.001$, what is the probability of an error correction failure, and what event would cause this?

(d) Suppose that a continuous communication channel of bandwidth W Hertz, which is perturbed by additive white Gaussian noise of constant power spectral density, has a channel capacity of C bits per second. Approximately how much would C be degraded if suddenly the added noise power became 8 times greater?

(e) You are comparing different image compression schemes for images of natural scenes. Such images have strong statistical correlations among neighbouring pixels because of the properties of natural objects. In an efficient compression scheme, would you expect to find strong correlations in the compressed image code? What statistical measure of the code for a compressed image determines the amount of compression it achieves, and in what way is this statistic related to the compression factor?

Model Answer – Example Problem Set 8

(a) The family of continuous signals having maximum entropy per variance (or power level) are Gaussian signals. Their probability density function for excursions x around a mean value μ, when the power level (or variance) is σ^2, is:

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

(b) The Fourier power spectrum of this class of signals is flat, or white. Hence these signals correspond to "white noise." The distribution of spectral energy has uniform probability over all possible frequencies, and therefore this continuous distribution has maximum entropy.

(c) An error-correcting Hamming code with a 7 bit block size uses 3 bits for error correction and 4 bits for data transmission. It would fail to correct errors that affected more than one bit in a block of 7; but in the example given, with $p = 0.001$ for a single bit error in a block of 7, the probability of two bits being corrupted in a block would be about 1 in a million.

(d) The channel capacity C in bits per second would be reduced by about $3W$, where W is the channel's bandwidth in Hertz, if the noise power level increased eight-fold. This is because the channel capacity, in bits per second, is

$$C = W \log_2 \left(1 + \frac{P}{N_0 W}\right)$$

If the signal-to-noise ratio (the term inside the logarithm) were degraded by a factor of 8, then its logarithm is reduced by -3, and so the overall capacity C is reduced by $3W$. The new channel capacity C' could be expressed either as:

$$C' = C - 3W$$

or as a ratio that compares it with the original undegraded capacity C:

$$\frac{C'}{C} = 1 - \frac{3W}{C}$$

(e) In an efficient compression scheme, there would be few correlations in the compressed representations of the images. Compression depends upon decorrelation. An efficient scheme would have low entropy; Shannon's Source Coding Theorem tells us a coding rate R as measured in bits per pixel can be found that is nearly as small as the entropy of the image representation. The compression factor can be estimated as the ratio of this entropy to the entropy of the uncompressed image (i.e. the entropy of its pixel histogram).

Information Theory and Coding: Example Problem Set 9

A. Prove that the information measure is additive: that the information gained from observing the combination of N independent events, whose probabilities are p_i for $i = 1....N$, is the *sum* of the information gained from observing each one of these events separately and in any order.

B. What is the shortest possible code length, in bits per average symbol, that could be achieved for a six-letter alphabet whose symbols have the following probability distribution?

$$\{ \tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{8}, \tfrac{1}{16}, \tfrac{1}{32}, \tfrac{1}{32} \}.$$

C. Suppose that ravens are black with probability 0.6, that they are male with probability 0.5 and female with probability 0.5, but that male ravens are 3 times more likely to be black than are female ravens.

If you see a non-black raven, what is the probability that it is male?

How many bits worth of information are contained in a report that a non-black raven is male?

Rank-order for this problem, from greatest to least, the following uncertainties:
(i) uncertainty about colour;
(ii) uncertainty about gender;
(iii) uncertainty about colour, given only that a raven is male;
(iv) uncertainty about gender, given only that a raven is non-black.

D. If a continuous signal $f(t)$ is *modulated* by multiplying it with a complex exponential wave $\exp(i\omega t)$ whose frequency is ω, what happens to the Fourier spectrum of the signal?

Name a very important practical application of this principle, and explain why modulation is a useful operation.

How can the original Fourier spectrum later be recovered?

E. Which part of the 2D Fourier Transform of an image, the amplitude spectrum or the phase spectrum, is indispensable in order for the image to be intelligible?

Describe a demonstration that proves this.

Model Answer – Example Problem Set 9

A. The information measure assigns $\log_2(p)$ bits to the observation of an event whose probability is p. The probability of the combination of N independent events whose probabilities are $p_1....p_N$ is $\prod_{i=1}^{N} p_i$
Thus the information content of such a combination is:

$$\log_2(\prod_{i=1}^{N} p_i) = \log_2(p_1) + \log_2(p_2) + \cdots + \log_2(p_N)$$

which is the sum of the information content of all of the separate events.

B.
Shannon's *Source Coding Theorem* tells us that the entropy of the distribution is the lower bound on average code length, in bits per symbol. This alphabet has entropy

$$H = -\sum_{i=1}^{6} p_i \log_2 p_i = (1/2)(1) + (1/4)(2) + (1/8)(3) + (1/16)(4) + (1/32)(5) + (1/32)(5) =$$

$1\frac{15}{16}$ or $\frac{31}{16}$ bits per average symbol (less than 2 bits to code 6 symbols!)

C.
Givens: $p(B|m) = 3p(B|f)$, $p(m) = p(f) = 0.5$, $p(B) = 0.6$ and so $p(NB) = 0.4$ where m means male, f means female, B means black and NB means non-black. From these givens plus the Sum Rule fact that $p(m)p(B|m) + p(f)p(B|f) = p(B) = 0.6$, it follows that $p(B|f) = 0.3$ and $p(B|m) = 0.9$, and hence that $p(NB|m) = 1 - 0.9 = 0.1$

Now we may apply Bayes Rule to calculate that

$$p(m|NB) = \frac{p(NB|m)p(m)}{p(NB)} = \frac{(0.1)(0.5)}{(0.4)} = 0.125 = 1/8$$

From the information measure $\log_2(p)$, there are 3 bits worth of information in discovering that a non-black raven is male.

(i) The colour distribution is { 0.6, 0.4 }
(ii) The gender distribution is { 0.5, 0.5 }
(iii) The (colour | male) distribution is { 0.9, 0.1 }
(iv) The (gender | non-black) distribution is { 0.125, 0.875 }

Uncertainty of a random variable is greater, the closer its distribution is to uniformity. Therefore the rank-order of uncertainty, from greatest to least, is: ii, i, iv, iii.

D. Modulation of the continuous signal by a complex exponential wave $\exp(i\omega t)$ will shift its entire frequency spectrum upwards by an amount ω.

All of AM broadcasting is based on this principle. It allows many different communications channels to be multi-plexed into a single medium, like the electromagnetic spectrum, by shifting different signals up into separate frequency bands.

The original Fourier spectrum of each of these signals can then be recovered by demodulating them down (this removes each AM carrier). This is equivalent to multiplying the transmitted signal by the conjugate complex exponential, $\exp(-i\omega t)$.

E. The phase spectrum is the indispensable part. This is demonstrated by crossing the amplitude spectrum of one image with the phase spectrum of another one, and *vice versa*. The new image that you see looks like the one whose phase spectrum you are using, and not at all like the one whose amplitude spectrum you've got.

Information Theory and Coding: Example Problem Set 10

1.
Consider n different discrete random variables, named $X_1, X_2, ..., X_n$, each of which has entropy $H(X_i)$.

Suppose that random variable X_j has the smallest entropy, and that random variable X_k has the largest entropy.

What is the upper bound on the joint entropy $H(X_1, X_2, ..., X_n)$ of all these random variables?

Under what condition will this upper bound be reached?

What is the lower bound on the joint entropy $H(X_1, X_2, ..., X_n)$ of all these random variables?

Under what condition will the lower bound be reached?

2.
Define the Kolmogorov algorithmic complexity K of a string of data.

What relationship is to be expected between the Kolmogorov complexity K and the Shannon entropy H for a given set of data?

Give a reasonable estimate of the Kolmogorov complexity K of a fractal, and explain why it is reasonable.

3.
The signal-to-noise ratio SNR of a continuous communication channel might be different in different parts of its frequency range. For example, the noise might be predominantly high frequency hiss, or low frequency rumble. Explain how the information capacity C of a noisy continuous communication channel, whose available bandwidth spans from frequency ω_1 to ω_2, may be defined in terms of its signal-to-noise ratio as a function of frequency, $SNR(\omega)$. Define the bit rate for such a channel's information capacity, C, in bits/second, in terms of the $SNR(\omega)$ function of frequency.

(Note: This question asks you to generalise beyond the material lectured.)

Model Answer – Example Problem Set 10

1.
The upper bound on the joint entropy $H(X_1, X_2, ..., X_n)$ of all the random variables is:

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

This upper bound is reached only in the case that all the random variables are independent.

The lower bound on the joint entropy $H(X_1, X_2, ..., X_n)$ is the largest of their individual entropies:

$$H(X_1, X_2, ..., X_n) \geq H(X_k)$$

(But note that if all the random variables are some deterministic function or mapping of each other, so that if any one of them is known there is no uncertainty about any of the other variables, then they all have the same entropy and so the lower bound is equal to $H(X_j)$ or $H(X_k)$.)

2.
The Kolmogorov algorithmic complexity K of a string of data is defined as the length of the shortest binary program that can generate the string. Thus the data's Kolmogorov complexity is its "Minimal Description Length."

The expected relationship between the Kolmogorov complexity K of a set of data, and its Shannon entropy H, is that approximately $K \approx H$.
Because fractals can be generated by extremely short programs, namely iterations of a mapping, such patterns have Kolmogorov complexity of nearly $K \approx 0$.

3.
The information capacity C of any tiny portion $\Delta\omega$ of this noisy channel's total frequency band, near frequency ω where the signal-to-noise ratio happens to be $SNR(\omega)$, is:

$$C = \Delta\omega \log_2(1 + SNR(\omega))$$

in bits/second. Integrating over all of these small $\Delta\omega$ bands in the available range from ω_1 to ω_2, the total capacity in bits/second of this variable-SNR channel is therefore:

$$C = \int_{\omega 1}^{\omega 2} \log_2(1 + SNR(\omega)) d\omega$$

Information Theory and Coding: Example Problem Set 11

1.
Construct an efficient, uniquely decodable binary code, having the prefix property and having the shortest possible average code length per symbol, for an alphabet whose five letters appear with these probabilities:

Letter	*Probability*
A	1/2
B	1/4
C	1/8
D	1/16
E	1/16

How do you know that your code has the shortest possible average code length per symbol?

2.
For a string of data of length N bits, what is the upper bound for its Minimal Description Length, and why?

Comment on how, or whether, you can know that you have truly determined the Minimal Description Length for a set of data.

3.
Suppose you have sampled a strictly bandlimited signal at regular intervals more frequent than the Nyquist rate; or suppose you have identified all of the zero-crossings of a bandpass signal whose total bandwidth is less than one octave. In either of these situations, provide some intuition for why you now also have knowledge about exactly what the signal must be doing at all points between these observed points.

4.
Explain how autocorrelation can remove noise from a signal that is buried in noise, producing a clean version of the signal. For what kinds of signals, and for what kinds of noise, will this work best, and why? What class of signals will be completely unaffected by this operation except that the added noise has been removed? Begin your answer by writing down the autocorrelation integral that defines the autocorrelation of a signal $f(x)$.

Some sources of noise are additive (the noise is just superimposed onto the signal), but other sources of noise are multiplicative in their effect on the signal. For which type would the autocorrelation clean-up strategy be more effective, and why?

Model Answer – Example Problem Set 11

1.
Example of one such code (there are others as well):

Letter	*Code*
A	1
B	01
C	001
D	0000
E	0001

This is a uniquely decodable code, and it also has the prefix property that no symbol's code is the beginning of a code for a different symbol.

The shortest possible average code length per symbol is equal to the entropy of the distribution of symbols, according to Shannon's Source Coding Theorem. The entropy of this symbol alphabet is:

$$H = -\sum_i p_i \log_2(p_i) = 1/2 + 2/4 + 3/8 + 4/16 + 4/16 = 1(7/8)$$

bits, and the average code length per symbol for the above prefix code is also (just weighing the length in bits of each of the above letter codes, by their associated probabilities of appearance): $1/2 + 2/4 + 3/8 + 4/16 + 4/16 = 1(7/8)$ bits. Thus no code can be more efficient than the above code.

2.
For a string of data of length N bits, the upper bound on its Minimal Description Length is N. The reason is that this would correspond to the worst case in which the shortest program that can generate the data is one that simply lists the string itself.

It is often impossible to know whether one has truly found the shortest possible description of a string of data. For example, the string:
`01101010000010011110011001100111111100111 0...`
passes most tests for randomness and reveals no simple rule which generates it, but it turns out to be simply the binary expansion for the irrational number $\sqrt{2} - 1$.

3.
The bandlimiting constraint (either just a highest frequency component in the case of Nyquist sampling, or the bandwidth limitation to one octave in the case of Logan's Theorem), is remarkably severe. It ensures that the signal cannot vary unsmoothly between the sample points (i.e. it must be everywhere a linear combination of shifted sinc functions in the Nyquist case), and it cannot remain away from zero for very long in Logan's case. Doing so would violate the stated frequency bandwidth constraint.

4.
The autocorrelation integral for a (real-valued) signal $f(x)$ is:

$$g(x) = \int f(y) f(x + y) dy$$

i.e. $f(x)$ is multiplied by a shifted copy of itself, and this product integrated, to generate a new signal as a function of the amount of the shift.

Signals differ from noise by tending to have some coherent, or oscillatory, component whose phase varies regularly; but noise tends to be incoherent, with randomly changing phase. The autocorrelation integral shifts the coherent component systematically from being in-phase with itself to being out-of-phase with itself. But this self-reinforcement does not happen for the noise, because of its randomly changing phase. Therefore the noise tends to cancel out, leaving the signal clean and reinforced. The process works best for purely coherently signals (sinusoids) buried in completely incoherent noise. Sinusoids would be perfectly extracted from the noise.

Autocorrelation as a noise removal strategy depends on the noise being just added to the signal. It would not work at all for multiplicative noise.