Exercise Problems: Information Theory and Coding

Prerequisite courses: Mathematical Methods for CS; Probability

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 = $\underline{\text{3/8 bits}}$. - 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 $\underline{\text{3/8 bits}}$.

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 $\underline{H(Y) = 1 \text{ bit}}$, 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 $H(X,Y) = -\sum_{x,y} p(x,y)\log_2 p(x,y)$

$= -(1-\epsilon)\log(0.5(1-\epsilon)) - \epsilon\log(0.5\epsilon) = (1-\epsilon) - (1-\epsilon)\log(1-\epsilon) + \epsilon - \epsilon\log(\epsilon)$

$= \underline{1 - \epsilon\log(\epsilon) - (1-\epsilon)\log(1-\epsilon)}$

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

5. In the two cases of $\underline{\epsilon = 0}$ and $\underline{\epsilon = 1}$ (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 $\underline{\epsilon = 0.5}$, the channel capacity is minimal and equal to $\underline{0}$.

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?

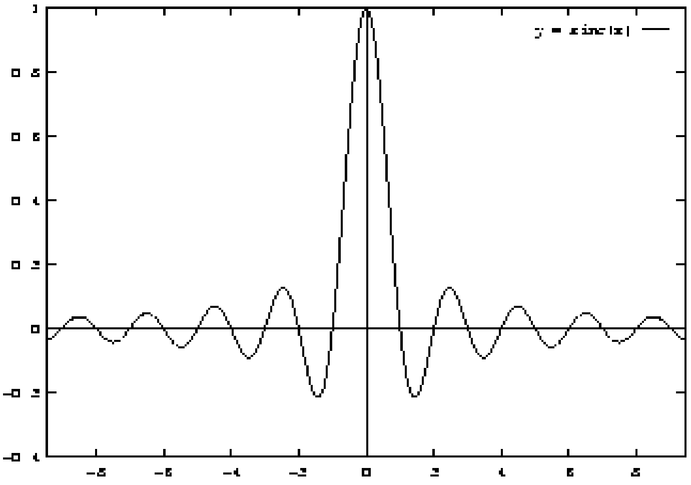
2. Name, define algebraically, and sketch a plot of the function you would need to use in order to recover completely the continuous signal transmitted, using just such a finite list of discrete periodic samples of it.

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. $2\omega\mathrm{T} = \underline{20{,}000}$ discrete samples are required.

2. The sinc function is required to recover the signal from its discrete samples, defined as: $\mathrm{sinc}(x) = \dfrac{\sin(\pi x)}{\pi x}$. Each sample point is replaced by scaled copies of this function.



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 $\dfrac{P}{N_0 W}$ 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

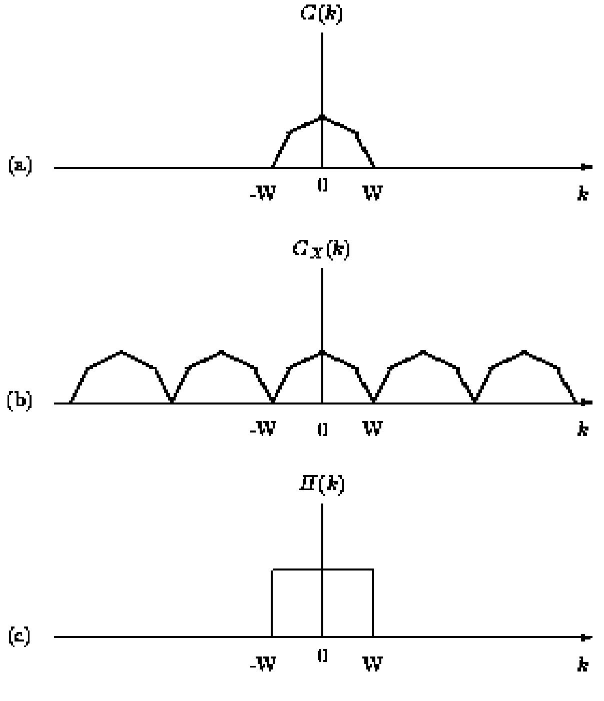
$$C = W \log_2 \left(1 + \frac{P}{N_0 W}\right)$$

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.



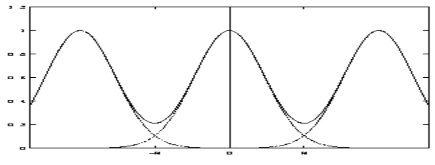


Figure 5: Band-limited signals: (a) Spectrum of $g(x)$. (b) Spectrum of $g_X(x)$. (c) Ideal filter response for reconstruction

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 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 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:"

$$f(x) = e^{-ik_0x}e^{-(x-x_0)^2/a^2} \tag{11}$$

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:

$$C = W \log_2 \left(1 + \frac{P}{N_0W}\right) \tag{12}$$

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: $R = (2 \text{ bits})(1/4)+(3 \text{ bits})(1/8)+(1 \text{ bit})(1/2)+(3 \text{ bits})(1/8) = \underline{1.75 \text{ bits}}$.

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 and a high signal-to-noise ratio, 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?

$$\left\{ \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}, \frac{1}{32} \right\}.$$

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:
`0110101000001001111001100110011111110011 10...`
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 coherent 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.

Information Theory and Coding: Example Problem Set 12

A.

State and explain (without proving) two different theorems about signal encoding that both illustrate the following principle: strict bandlimiting (either lowpass or bandpass) of a continuous signal reduces the information that it contains from potentially infinite to a finite discrete set of data, and allows exact reconstruction of the signal from just a sparse set of sample values. For both of your examples, explain what the sample data are, and why bandlimiting a signal has such a dramatic effect on the amount of information required to represent it completely.

B.

A variable length, uniquely decodable code which has the prefix property, and whose N binary code word lengths are

$$n_1 \leq n_2 \leq n_3 \leq \cdots \leq n_N$$

must satisfy what condition on these code word lengths?

(State both the condition on the code word lengths, and the name for this condition, but do not attempt to prove it.)

C.

For a discrete data sequence consisting of the N uniformly-spaced samples

$$\{g_n\} = \{g_0,\ g_1,\ ...,\ g_{N-1}\}$$

define both the Discrete Fourier Transform $\{G_k\}$ of this sequence, and its Inverse Transform, which recovers $\{g_n\}$ from $\{G_k\}$.

Model Answer – Example Problem Set 12

(Subject areas: Signal encoding; variable-length prefix codes; discrete FT.)

A.

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

$$F(k) = 0 \text{ for } |k| > W$$

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,

$$\text{sinc}(x) = \frac{\sin(\pi x)}{\pi x}$$

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:

$$f(x) = \sum_n f_n\left(\frac{n\pi}{W}\right)\frac{\sin(Wx - n\pi)}{(Wx - n\pi)}$$

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

$$k_{max} \leq 2k_{min}$$

i.e. $F(k)$ the Fourier Transform of $f(x)$ obeys

$$F(|k| > k_{max} = 2k_{min}) = 0$$

and

$$F(|k| < k_{min}) = 0$$

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.

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

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).

(continued...)

B.

The N binary code word lengths $n_1 \leq n_2 \leq n_3 \leq \cdots \leq n_N$ must satisfy the *Kraft-McMillan Inequality* if they are to constitute a uniquely decodable prefix code:

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

C.

The Discrete Fourier Transform $\{G_k\}$ of the regular sequence $\{g_n\} = \{g_0,\ g_1,\ ...,\ g_{N-1}\}$ is:

$$\{G_k\} = \sum_{n=0}^{N-1} g_n \exp\left(-\frac{2\pi i}{N} kn\right), \quad (k = 0, 1, ..., N-1)$$

The Inverse Transform (or synthesis equation) which recovers $\{g_n\}$ from $\{G_k\}$ is:

$$\{g_n\} = \frac{1}{N} \sum_{k=0}^{N-1} G_k \exp\left(\frac{2\pi i}{N} kn\right), \quad (n = 0, 1, ..., N-1)$$

Information Theory and Coding: Example Problem Set 13

A.
A Hamming Code allows reliable transmission of data over a noisy channel with guaranteed error correction as long as no more than one bit in any block of 7 is corrupted. What is the maximum possible rate of information transmission, in units of (data bits reliably received) per (number of bits transmitted), when using such an error correcting code?

In such a code, what type of Boolean operator on the data bits is used to build the syndromes? Is this operator applied before transmission, or upon reception?

B.
For each of the four classes of signals in the following table,

Class	*Signal Type*
1.	continuous, aperiodic
2.	continuous, periodic
3.	discrete, aperiodic
4.	discrete, periodic

identify its characteristic spectrum from the following table:

Class	*Spectral Characteristic*
A.	continuous, aperiodic
B.	continuous, periodic
C.	discrete, aperiodic
D.	discrete, periodic

("Continuous" here means supported on the reals, i.e. at least piecewise continuous but not necessarily everywhere differentiable. "Periodic" means that under multiples of some finite shift the function remains unchanged.) Give your answer just in the form 1-A, 2-B, etc. Note that you have 24 different possibilities.

For each case, name one example of such a function and its Fourier transform.

C.
Give two reasons why Logan's Theorem about the richness of zero-crossings for encoding and recovering all the information in a one-octave signal may not be applicable to images as it is for one-dimensional signals.

Model Answer – Example Problem Set 13

(Subject areas: Error correcting codes. Signals and spectra. Zero-crossings.)

A.

A Hamming Code transmits 7 bits in order to encode reliably 4 data bits; the 3 non-data bits are added to guarantee detection and correction of 1 erroneous bit in any such block of 7 bits transmitted. Thus the maximum rate of information transmission is 4/7ths of a bit per bit transmitted.

Syndromes are constructed by taking the Exclusive-OR of three different subsets of 4 bits from the 7 bits in a block. This Boolean operation is performed upon reception. (Before transmission, the XOR operator is also used to build the three extra error-correcting bits from the four actual data bits: each error-correcting bit is the XOR of a different triple of bits among the four data bits.) Upon reception, if the three syndrome bits computed (by XORing different subsets of 4 of the 7 bits received) are all 0, then there was no error; otherwise they identify which bit was corrupted, so that it can be inverted.

B.

1-A. Example: a Gaussian function, whose Fourier transform is also Gaussian.
2-C. Example: a sinusoid, whose Fourier transform is two discrete delta functions.
3-B. Example: a delta function, whose Fourier transform is a complex exponential.
4-D. Example: a comb sampling function, whose Fourier Transform is also a comb function.

C.

1. The zero-crossings in a two- (or higher-) dimensional signal, such as an image, are not denumerable. 2. The extension of the one-octave bandlimiting constraint to the Fourier plane does not seem to be possible in an isotropic manner. If applied isotropically (i.e. a one-octave annulus centred on the origin of the Fourier plane), then in fact both the vertical and horizontal frequencies are each low-pass, not bandpass. But if applied in a bandpass manner to each of the four quadrants, thereby selecting four disjoint square regions in the Fourier plane, then the different orientations in the image are treated differently (anisotropically).

Information Theory and Coding: Example Problem Set 14

A.

Consider an alphabet of 8 symbols whose probabilities are as follows:

A	B	C	D	E	F	G	H
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$	$\frac{1}{128}$

1. If someone has selected one of these symbols and you need to discover which symbol it is by asking 'yes/no' questions that will be truthfully answered, what would be the most efficient sequence of such questions that you could ask in order to discover the selected symbol?

2. By what principle can you claim that each of your proposed questions is maximally informative?

3. On average, how many such questions will need to be asked before the selected symbol is discovered?

4. What is the entropy of the above symbol set?

5. Construct a uniquely decodable prefix code for the symbol set, and explain why it is uniquely decodable and why it has the prefix property.

6. Relate the bits in your prefix code to the 'yes/no' questions that you proposed in (1).

B.

Explain the meaning of "self-Fourier," and cite at least two examples of mathematical objects having this property.

C.

Explain briefly:

1. Sensation limit

2. Critical band

3. Bark scale

4. Which different aspects of perception do Weber's law and Steven's law model?

Model Answer – Example Problem Set 14

A.

A	B	C	D	E	F	G	H
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$	$\frac{1}{128}$

1. For this symbol distribution, the most efficient sequence of questions to ask (until a 'yes' is obtained) would be just: (1) Is it A? (2) Is it B? (3) Is it C? (Etc.)

2. Each such 1-bit question is maximally informative because the remaining uncertainty is reduced by half (1 bit).

3. The probability of terminating successfully after exactly N questions is 2^{-N}. At most 7 questions might need to be asked. The weighted average of the interrogation durations is:
$$\frac{1}{2} + (2)(\frac{1}{4}) + (3)(\frac{1}{8}) + (4)(\frac{1}{16}) + (5)(\frac{1}{32}) + (6)(\frac{1}{64}) + (7)(\frac{2}{128}) = 1\frac{126}{128}$$
In other words, on average just slightly less than <u>two</u> questions need to be asked in order to learn which of the 8 symbols it is.

4. The entropy of the above symbol set is calculated by the same formula, but over all 8 states (whereas at most 7 questions needed to be asked):
$$H = -\sum_{i=1}^{8} p_i \log_2 p_i = 1\frac{126}{128}$$

5. A natural code book to use would be the following:

A	B	C	D	E	F	G	H
1	01	001	0001	00001	000001	0000001	0000000

6. It is uniquely decodable because each code corresponds to a unique letter rather than any possible combination of letters; and it has the prefix property because the code for no letter could be confused as the prefix for another letter.

7. The bit strings in the above prefix code for each letter can be interpreted as the history of answers to the 'yes/no' questions.

B.

Functions which have exactly the same form as their Fourier transforms are called "self-Fourier." Examples of such pairs include: the Gaussian; the Gabor wavelet; the sampling Comb function; and the hyperbolic secant.

C.

1. The sensation limit of a sense is the lowest amplitude of a stimulus that can be perceived.

2. If two audio tones fall within the same critical band, the ear is unable to recognize two separate tones and perceives a single tone with the average of their frequency instead. (The human ear has approximately 24 non-overlapping critical bands.)

3. The Bark scale is a non-linear transform of an audible frequency into the number range 0 to 24, such that if two frequencies are less than 1 apart on this scale, they are within the same critical band.

4. Weber's law is concerned with how the difference limit, the smallest amplitude change of a stimulus that can be distinguished, depends on the amplitude of the stimulus. (It states that the two are proportional, except for a small correction near the sensation limit.) Steven's law on the other hand is concerned with how the amplitude of a stimulus is perceived in relation to other amplitudes, for example how much must the amplitude raise such that the stimulus is perceived as being twice as strong. (It states a power-law relationship between amplitude and perceived stimulus strength.)

Information Theory and Coding: Example Problem Set 15

A.

A variable length, uniquely decodable code which has the prefix property, and whose N binary code word lengths are $n_1 \leq n_2 \leq n_3 \leq \cdots \leq n_N$ must satisfy what condition on code word lengths? (State the condition, and name it.)

B.

You are asked to compress a collection of files, each of which contains several thousand photographic images. All images in a single file show the same scene. Everything in this scene is static (no motion, same camera position, etc.) except for the intensity of the five light sources that illuminate everything. The intensity of each of the five light sources changes in completely unpredictable and uncorrelated ways from image to image. The intensity of each pixel across all photos in a file can be described as a linear combination of the intensity of these five light sources.

1. Which one of the five techniques *discrete cosine transform*, *μ-law coding*, *2-D Gabor transform*, *Karhunen-Loève transform* and *Golomb coding* would be best suited to remove redundancy from these files, assuming your computer is powerful enough for each?
2. Explain briefly this transform and why it is of use here.

Model Answer – Example Problem Set 15

A.

The N binary code word lengths $n_1 \leq n_2 \leq n_3 \leq \cdots \leq n_N$ must satisfy the Kraft-McMillan Inequality in order to form a uniquely decodable prefix code:

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

B.

1. The Karhunen-Loève transform.
2. The Karhunen-Loève transform decorrelates random vectors. Let the values of the random vector $\mathbf{v}$ represent the individual images in one file. All vector elements being linear combinations of five values means that for each file there exists an orthonormal matrix M such that each image vector $\mathbf{v}$ can be represented as $\mathbf{v} = M\mathbf{t}$, where $\mathbf{t}$ is a new random vector whose covariance matrix is diagonal and in which all but the first five elements are zero. The Karhunen-Loève transform provides this matrix M by calculating the spectral decomposition of the covariance matrix of $\mathbf{v}$. The significant part of the transform result $M^\top \mathbf{v} = \mathbf{t}$ are only five numbers, which can be stored compactly for each image, together with the five relevant rows of M per file.

Information Theory and Coding: Example Problem Set 16

(*a*) For a binary symmetric communication channel whose input source is the alphabet $X = \{0, 1\}$ with probabilities $\{0.5, 0.5\}$ and whose output alphabet is $Y = \{0, 1\}$, having the following channel matrix where ϵ is the probability of transmission error:

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

(*i*) How much uncertainty is there about the input symbol once an output symbol has been received?

(*ii*) What is the mutual information $I(X;Y)$ of this channel?

(*iii*) What value of ϵ maximises the uncertainty $H(X|Y)$ about the input symbol given an output symbol?

(*b*) For a continuous (i.e. non-discrete) function $g(x)$, define:

(*i*) its continuous Fourier transform $G(k)$

(*ii*) the inverse Fourier transform that recovers $g(x)$ from $G(k)$

(*c*) What simplifications occur in the Fourier representation of a function if:

(*i*) the function is real-valued rather than complex-valued?

(*ii*) the function has even symmetry?

(*iii*) the function has odd symmetry?

(*d*) Give a bit-string representation of the number 13 in

(*i*) unary code for non-negative integers;
(*ii*) Golomb code for non-negative integers with parameter $b = 3$;
(*iii*) Elias gamma code for positive integers.

Model Answer – Example Problem Set 16

(*a*)

(*i*) The uncertainty about the input X given the observed output Y from the channel is the conditional entropy $H(X|Y)$, which is defined as:

$$H(X|Y) = -\sum_{x,y} p(x,y) \log p(x|y)$$

So, we need to calculate both the joint probability distribution $p(X,Y)$ and the conditional probability distribution $p(X|Y)$, and then combine their terms according to the above summation.

The 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 conditional probability distribution $p(X|Y)$ is

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

Combining these matrix elements accordingly gives us the conditional entropy:

$$H(X|Y) = -\left[0.5(1-\epsilon)\log(1-\epsilon) + 0.5\epsilon\log(\epsilon) + 0.5\epsilon\log(\epsilon) + 0.5(1-\epsilon)\log(1-\epsilon)\right]$$

$$= \underline{-(1-\epsilon)\log(1-\epsilon) - \epsilon\log(\epsilon)}$$

(*ii*) One definition of mutual information is $I(X;Y) = H(X) - H(X|Y)$. Since the two input symbols are equi-probable, clearly $H(X) = 1$ bit. We know from (i) above that $H(X|Y) = -(1-\epsilon)\log(1-\epsilon) - \epsilon\log(\epsilon)$, and so therefore, the mutual information of this channel is:

$$I(X;Y) = 1 + (1-\epsilon)\log(1-\epsilon) + \epsilon\log(\epsilon)$$

(*iii*) The uncertainty $H(X|Y)$ about the input, given the output, is maximised when $\underline{\epsilon = 0.5}$, in which case it is 1 bit.

(*b*) The analysis and synthesis (or forward and inverse) continuous Fourier transforms are, respectively:

(*i*) $G(k) = \int_{-\infty}^{+\infty} g(x)e^{-ikx}dx$

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

(*c*) The Fourier representation becomes simplified as follows:

(i) If the function is real-valued rather than complex-valued, then its Fourier transform has Hermitian symmetry: the real-part of the Fourier transform has even symmetry, and the imaginary part has odd-symmetry.

(ii) If the function has even symmetry, then its Fourier transform is purely real-valued.

(iii) If the function has odd symmetry, then its Fourier transform is purely imaginary-valued.

(d)

(i) 11111111111110 = $1^{13}0$
The unary code word for 13 is simply 13 ones, followed by a final zero.

(ii) 1111010 = $1^4 0$ 10
We first divide $n = 13$ by $b = 3$ and obtain the representation $n = q \times b + r = 4 \times 3 + 1$ with remainder $r = 1$. We then encode $q = 4$ as the unary code word "11110". To this we need to attach an encoding of $r = 1$. Since r could have a value in the range $\{0, \ldots, b-1\} = \{0, 1, 2\}$, we first use all $\lfloor \log_2 b \rfloor = 1$-bit words that have a leading zero (here only "0" for $r = 0$), before encoding the remaining possible values of r using $\lceil \log_2 b \rceil = 2$-bit values that have a leading one (here "10" for $r = 1$ and "11" for $r = 2$).

(iii) 1110101 = $1^3 0$ 101
We first determine the length indicator $m = \lfloor \log_2 13 \rfloor = 3$ (because $2^3 \leq 13 < 2^4$) and encode it using the unary code word "1110", followed by the binary representation of 13 (1101_2) with the leading one removed: "101".

Information Theory and Coding: Example Problem Set 17

(*a*) For continuous random variables X and Y, taking on continuous values x and y respectively with probability densities $p(x)$ and $p(y)$ and with joint probability distribution $p(x, y)$ and conditional probability distribution $p(x|y)$, define:

(*i*) the *differential entropy* $h(X)$ of random variable X;

(*ii*) the *joint entropy* $h(X, Y)$ of the random variables X and Y;

(*iii*) the *conditional entropy* $h(X|Y)$ of X, given Y;

(*iv*) the *mutual information* $i(X; Y)$ between continuous random variables X and Y;

(*v*) how the *channel capacity* of a continuous channel which takes X as its input and emits Y as its output would be determined.

(*b*) For a time-varying continuous signal $g(t)$ which has Fourier transform $G(k)$, state the *modulation theorem* and explain its role in AM radio broadcasting. How does modulation enable many independent signals to be encoded into a common medium for transmission, and then separated out again via tuners upon reception?

(*c*) Briefly define

(*i*) The *Differentiation Theorem* of Fourier analysis: if a function $g(x)$ has Fourier transform $G(k)$, then what is the Fourier transform of the n^{th} derivative of $g(x)$, denoted $g^{(n)}(x)$?

(*ii*) If discrete symbols from an alphabet $\mathcal{S}$ having entropy $H(\mathcal{S})$ are encoded into blocks of length n, we derive a new alphabet of symbol blocks $\mathcal{S}^n$. If the occurrence of symbols is independent, then what is the entropy $H(\mathcal{S}^n)$ of the new alphabet of symbol blocks?

(*iii*) If symbols from an alphabet of entropy H are encoded with a *code rate* of R bits per symbol, what is the *efficiency* η of this coding?

(*d*) Briefly explain
(*i*) how a signal amplitude of 10 V is expressed in dBμV;
(*ii*) the YCrCb coordinate system.

Model Answer – Example Problem Set 17

(*a*)

(*i*) The *differential entropy* $h(X)$ is defined as:

$$h(X) = \int_{-\infty}^{+\infty} p(x) \log \left(\frac{1}{p(x)} \right) dx$$

(*ii*) The *joint entropy* $h(X, Y)$ of random variables X and Y is:

$$h(X, Y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(x, y) \log \left(\frac{1}{p(x, y)} \right) dxdy$$

(*iii*) The *conditional entropy* $h(X|Y)$ of X, given Y, is:

$$h(X|Y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(x, y) \log \left(\frac{p(y)}{p(x, y)} \right) dxdy$$

$$= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(x, y) \log \left(\frac{1}{p(x|y)} \right) dxdy$$

(*iv*) The *mutual information* $i(X; Y)$ between continuous random variables X and Y is:

$$i(X; Y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(x, y) \log \left(\frac{p(x|y)}{p(x)} \right) dxdy$$

$$= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right) dxdy$$

(*v*) The *capacity* of a continuous communication channel is computed by finding the maximum of the above expression for mutual information $i(X; Y)$ over all possible input distributions for X.

(*b*) The continuous signal $g(t)$ is *modulated* into a selected part of the frequency spectrum, defined by a transmitter carrier frequency. The signal is just multiplied by that carrier frequency (in complex form, i.e. as a complex exponential of frequency ω). The modulation theorem asserts that then the Fourier transform of the original signal is merely <u>shifted</u> by an amount equal to that carrier frequency ω:

$$g(t)e^{i\omega t} \rightleftharpoons G(k - \omega)$$

Many different signals can each be thus modulated into their own frequency bands and transmitted together over the electromagnetic spectrum using a common antenna. Upon reception, the reverse operation is performed by a tuner, i.e. multiplication of the received signal by the complex conjugate complex exponential $e^{-i\omega t}$ [and filtering away any other transmitted frequencies], thus restoring the original signal $g(t)$.

(*c*)

(*i*) The Fourier transform of the n^{th} derivative of $g(x)$ is: $(ik)^n G(k)$

(*ii*) The entropy of the new alphabet of symbol blocks is simply n times the entropy of the original alphabet:

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

(*iii*) The *efficiency* of the coding is defined as

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

(*d*)

(*i*) 10 V = 10^7 μV = (20 × 7) dBμV = 140 dBμV

(*ii*) Human colour vision splits the red/green/blue input signal into separate luminosity and colour channels. Compression algorithms can achieve a simple approximation of this by taking a linear combination of about 30% red, 60% green, and 10% blue as the luminance signal $Y = 0.3R + 0.6G + 0.1B$ (the exact coefficients differ between standards and do not matter here). The remaining colour information can be preserved, without adding redundancy, in the form of the difference signals $R - Y$ and $B - Y$. These are usually encoded scaled as $Cb = (B - Y)/2 + 0.5$ and $Cr = (R - Y)/1.6 + 0.5$, such that the colour cube remains, after this "rotation", entirely within the encoded unit cube, assuming that the original RGB values were all in the interval $[0, 1]$.

Information Theory and Coding: Example Problem Set 18

(a)
Suppose we know the conditional entropy $H(X|Y)$ for two slightly correlated discrete random variables X and Y. We wish to guess the value of X, from knowledge of Y. There are $\mathcal{N}$ possible values of X. Give a lower bound estimate for the probability of error, when guessing X from knowledge of Y. What is the name of this relationship?

(b)
In an error-correcting (7/4) Hamming code, under what circumstance is there still a residual error rate? (In other words, what event causes this error-correction scheme to fail?)

(c)
Broadband noise whose power spectrum is flat is "white noise." If the average power level of a white noise source is σ^2 and its excursions are zero-centred so its mean value is $\mu = 0$, give an expression describing the probability density function $p(x)$ for excursions x of this noise around its mean, in terms of σ. What is the special relationship between the entropy of a white noise source, and its power level σ^2?

(d)
Explain the phenomenon of aliasing when a continuous signal whose total bandwidth extends to $\pm W$ is sampled at a rate of $f_s < 2W$. If it is not possible to increase the sampling rate f_s, what can be done to the signal before sampling it that would prevent aliasing?

(e) Prove that the sinc function,

$$\text{sinc}(x) = \frac{\sin(\pi x)}{\pi x}$$

is invariant under convolution with itself: in other words that the convolution of a sinc function with itself is just another sinc function. You might find it useful to recall that the Fourier transform of a sinc function is the rectangular pulse function:

$$\Pi(k) = \begin{cases} \frac{1}{2\pi} & |k| \leq \pi \\ 0 & |k| > \pi \end{cases}$$

Model Answer – Example Problem Set 18

(a) The error probability has lower bound:

$$P_e \geq \frac{H(X|Y) - 1}{\log_2 \mathcal{N}}$$

This relationship is Fano's Inequality.

(b) In an error-correcting (7/4) Hamming code, errors will fail to be corrected if more than 1 bit in a block of 7 bits was corrupted.

(c) The probability density function for excursions x of the white noise source around its mean value of 0, with average power level (or variance) σ^2, is:

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

The special relationship is that for all possible noise power distributions having average power σ^2, the white noise source is the one with the greatest entropy.

(d) Sampling a signal effectively multiplies it with a "comb" function. This causes its Fourier spectrum to be reproduced completely at each "tyne" of another comb function in the frequency domain, where the tynes are separated from each other by the sampling frequency f_s. Provided that $f_s \geq 2W$ then all of these reproduced copies of the signal's spectrum can still be perfectly separated from each other. We can recover the original signal's spectrum just by ideal low-pass filtering, to discard everything outside of $\pm W$. But this is no longer possible if $f_s < 2W$, since in that case the reproduced copies of the original spectrum overlap and become partly superimposed, and thus they can no longer be separated from each other by low-pass filtering. To prevent aliasing when it not possible to increase the sampling rate f_s, the signal should first be low-pass filtered before it is sampled, reducing its frequency composition to be within $\pm W_0$ such that the condition $f_s \geq 2W_0$ is then satisfied.

(e) When two functions are convolved together, their Fourier transforms are just multiplied together to give the Fourier transform of the result of the convolution. In this case, convolving the sinc function with itself means that the Fourier transform of the result would be the product of the rectangular pulse function with itself; which is, of course, just another rectangular pulse function. Hence the result of the convolution is just another sinc function.

As a slightly modified version of this question: What happens when two different sinc functions (differing in their frequency parameter) are convolved together?

Answer: By the same reasoning as above, the result is always just whichever sinc function had the lower frequency! Hence, somewhat bizarrely, convolution implements the "select the lower frequency" operation on sinc functions...

Information Theory and Coding: Example Problem Set 19

(*a*) Suppose that the following sequence of Yes/No questions was an optimal strategy for playing the "Game of 7 questions" to learn which of the letters $\{A, B, C, D, E, F, G\}$ someone had chosen, given that their *a priori* probabilities were known:

"Is it A?"	"No."
"Is it a member of the set $\{B, C\}$?"	"No."
"Is it a member of the set $\{D, E\}$?"	"No."
"Is it F?"	"No."

(*i*) Write down a probability distribution for the 7 letters, $p(A), ..., p(G)$, for which this sequence of questions was an optimal strategy.

(*ii*) What was the uncertainty, in bits, associated with each question?

(*iii*) What is the entropy of this alphabet?

(*iv*) Now specify a variable length, uniquely decodable, prefix code for this alphabet that would minimise the average code word length.

(*v*) What is your average coding rate R for letters of this alphabet?

(*vi*) How do you know that a more efficient code could not be developed?

(*b*) An invertible transform generates projection coefficients by integrating the product of a signal onto each of a family of functions. In a reverse process, expansion coefficients can be used on those same functions to reproduce the signal. If the functions in question happen to form an orthonormal set, what is the consequence for the projection coefficients and the expansion coefficients?

(*c*) In the Information Diagram (a plane whose axes are time and frequency), why does the Gabor-Heisenberg-Weyl *Uncertainty Principle* imply that information is *quantised* – *i.e.* that it exists in only a limited number of independent quanta?

Model Answer – Example Problem Set 19

(*a*)

(*i*) Under the Asymptotic Equipartition Theorem, the following *a priori* probability distribution would make the given questioning strategy an optimal one:

$p(A)$	$p(B)$	$p(C)$	$p(D)$	$p(E)$	$p(F)$	$p(G)$
1/2	1/8	1/8	1/16	1/16	1/16	1/16

(*ii*) Each Yes/No question had 1 bit entropy (uncertainty), because both possible answers were equiprobable in each case.

(*iii*) Since entropy $= -\sum_i p_i \log_2 p_i$, the entropy of this alphabet is 2.25 bits.

(*iv*) One possible variable length, uniquely decodable, prefix code is:

A	B	C	D	E	F	G
0	110	111	1000	1001	1010	1011

(*v*) Summing over all the letters, the probability of each letter times its code word length in bits, gives us $R = (1/2)(1) + (2/8)(3) + (4/16)(4) = 2.25$ bits per letter on average.

(*vi*) Because the coding rate equals the entropy of the source alphabet, and Shannon's Source Coding Theorem tells us that this is the lower bound for the coding rate, we know that no more efficient code could be developed.

(*b*) In the case that the functions used for projection and expansion are an orthonormal set, then the projection coefficients and the expansion coefficients will be the same.

(*c*) The Gabor-Heisenberg-Weyl *Uncertainty Principle* asserts that in the Information Diagram, there is a lower bound on the size of the smallest area that can be occupied by any signal or filter. In other words, resolution of information along both axes at once, is fundamentally limited. The fact that there is a smallest possible occupied area, or quantum, means that information is quantised; there exists only a limited number of independent quanta of data in any given piece of this plane.

Information Theory and Coding: Example Problem Set 20

(*a*) Suppose that X is a random variable whose entropy $H(X)$ is 8 bits. Suppose that $Y(X)$ is a deterministic function that takes on a different value for each value of X.

(*i*) What then is $H(Y)$, the entropy of Y?

(*ii*) What is $H(Y|X)$, the conditional entropy of Y given X?

(*iii*) What is $H(X|Y)$, the conditional entropy of X given Y?

(*iv*) What is $H(X, Y)$, the joint entropy of X and Y?

(*v*) Suppose now that the deterministic function $Y(X)$ is not invertible; in other words, different values of X may correspond to the same value of $Y(X)$. In that case, what could you say about $H(Y)$?

(*vi*) In that case, what could you say about $H(X|Y)$?

(*b*) Write down the general functional form for a 1-D Gabor wavelet, and explain how particular choices for the values of its parameters would turn it into either the Fourier basis or the delta function sampling basis, as two special cases.

(*c*) Show that the set of all Gabor wavelets is closed under convolution. *I.e.* show that the convolution of any two Gabor wavelets is also a Gabor wavelet. Comment on how this property relates to the fact that these wavelets are also closed under multiplication, and that they are also self-Fourier.

(*d*) We wish to compute the Fourier Transform of a data sequence of 1,024 samples:

(*i*) Approximately how many multiplications would be needed if the Fourier integral expressions were to be computed literally (as written mathematically) and without a clever algorithm?

(*ii*) Approximately how many multiplications would be needed if an FFT algorithm were used?

Model Answer – Example Problem Set 20

(*a*)

(*i*) The entropy of Y: $H(Y) = 8$ bits also.

(*ii*) The conditional entropy of Y given X: $H(Y|X) = 0$

(*iii*) The conditional entropy of X given Y: $H(X|Y) = 0$ also.

(*iv*) The joint entropy $H(X, Y) = H(X) + H(Y|X) = 8$ bits

(*v*) Since now different values of X may correspond to the same value of $Y(X)$, the distribution of Y has lost entropy and so $H(Y) < 8$ bits.

(*vi*) Now knowledge of Y no longer determines X, and so the conditional entropy $H(X|Y)$ is no longer zero: $H(X|Y) > 0$

(*b*) The general functional form for a 1-D Gabor wavelet is:

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

In the case that we set the parameter a very large ($a \to \infty$), then this becomes the classical Fourier basis (the complex exponentials). In the case that we set a small ($a \to 0$) and $k_0 = 0$, then this becomes the classical Dirac delta function sampling basis.

(*c*) The Fourier Transform of a 1-D Gabor wavelet has exactly the same functional form, but with the parameters simply interchanged or inverted:

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

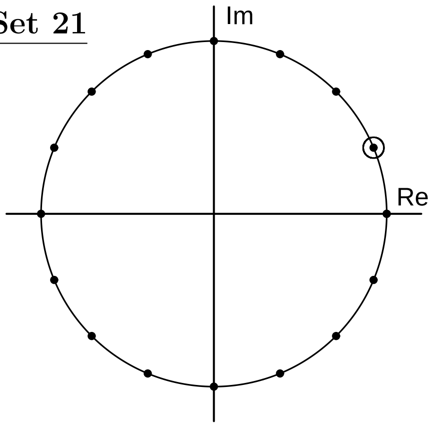
(In other words, Gabor wavelets are self-Fourier.) It is obvious that the product of any two Gabor wavelets $f(x)$ will still have the functional form of a Gabor wavelet. Therefore the product's Fourier transform will also preserve this general form. Hence (using the convolution theorem of Fourier analysis), it follows that the family of Gabor wavelets are also closed under convolution.

(*d*)

(*i*) A numerically literal computation of the Fourier transform of a sequence of 1,024 data samples would require on the order of 1 million multiplications.

(*ii*) If instead we used a Fast Fourier Transform algorithm, $\mathcal{O}(N \log N)$ or only about 5,000 to 10,000 multiplications would be required, about 1% as many.

Information Theory and Coding: Example Problem Set 21



Fast Fourier Transform algorithms use factorisation of discrete complex exponentials to avoid repeated multiplications by common factors. The diagram on the right shows a unit circle in the complex plane. The unit circle represents a continuous complex exponential (one orbit around it spans one cycle), and the 16 dots represent discrete samples of this Fourier component which need to be multiplied by 16 data points and summed to compute one discrete Fourier coefficient.

(i) The circled dot $\boxed{e^{2\pi i/n}}$ is a *primitive* n^{th}-root of unity, where for this diagram $n = 16$. Write down a similar expression for the full set of all the n^{th}-roots of unity, indexed by k, where $1 \leq k \leq n$.

(ii) The 16 frequency components needed to compute the discrete Fourier transform of 16 data points are obtained by undersampling the dots; *e.g.* the 2^{nd} frequency uses every 2^{nd} dot and orbits twice. Explain the redundancy that occurs when multiplying these discrete complex exponentials by the data points.

(iii) For n data points, roughly how many multiplications are needed in a Fast Fourier Transform algorithm that avoids these redundancies?

Model Answer – Example Problem Set 21

(i) The set of all the n^{th}-roots of unity is described by

$$e^{(2\pi i/n)^k} \quad \text{or} \quad e^{2\pi ik/n}, \quad (1 \leq k \leq n)$$

These are the discrete samples of one cycle of a complex exponential, from which all higher frequencies (having integer multiples of this frequency) are obtained.

(ii) Successive discrete Fourier components are all constructed from the same set of the n^{th}-roots of unity as illustrated in the diagram, merely undersampled to construct higher frequencies. But the same complex numbers (dots in the diagram) are used again and again to multiply by the same data points within the inner product summations that compute each Fourier coefficient. In addition, for all frequencies higher than the first frequency, any given discrete sample (dot in the diagram) is used in successive cycles to multiply more than one data point. These repeated multiplications can be grouped together in successive factorisations using powers of the primitive root $\boxed{\omega = e^{2\pi i/n}}$ to implement the transform without redundant multiplications.

(iii) By eliminating the redundant multiplications through factorisation, a Fast Fourier Transform algorithm can compute the discrete Fourier transform of n data points with a number of multiplications that is on the order of $\mathcal{O}\ (n \log_2 n)$.

Information Theory and Coding: Example Problem Set 22

(*a*) Calculate the entropy in bits for each of the following random variables:

(*i*) Pixel values in an image whose possible grey values are all the integers from 0 to 255 with uniform probability.

(*ii*) Humans grouped by whether they are, or are not, mammals.

(*iii*) Gender in a tri-sexual insect population whose three genders occur with probabilities 1/4, 1/4, and 1/2.

(*iv*) A population of persons classified by whether they are older, or not older, than the population's median age.

(*b*) Let $p(x)$ and $q(x)$ be two probability distributions specified over integers x.

(*i*) What is the *Kullback-Leibler distance* (*KL*) between these distributions?

(*ii*) If we have devised an optimally compact code for the random variable described by $q(x)$, what does the *KL* tell us about the effectiveness of our code if the probability distribution is $p(x)$ instead of $q(x)$?

(*iii*) Which axiom of distance metrics is violated by this distance?

(*iv*) What happens to this metric if there are some forbidden values of x for which $p(x) = 0$, and other values of x for which $q(x) = 0$?

(*c*) Explain why the encoding of continuous signals into sequences of coefficients on Gabor wavelets encompasses, as special cases, both the delta function sampling basis and the Fourier Transform basis. Show how one particular parameter determines where a signal representation lies along this continuum that bridges from delta function sampling to the complex exponential.

(*d*) Explain why data can be compressed by encoding it into transforms (such as the DCT, Fourier or Gabor) which result in coefficients that have a more narrow, peaked, distribution than the original data. Without going into details about particular transforms, explain why the coefficients obtained have distributions with less entropy than the original signal or image, and why this enables compression.

Model Answer – Example Problem Set 22

(a) By definition, $H = -\sum_i p_i \log_2 p_i$ is the entropy in bits for a discrete random variable distributed over states whose probabilities are p_i. So:

(i) In this case each $p_i = 1/256$ and the ensemble entropy summation extends over 256 such equiprobable grey values, so $H = -(256)(1/256)(-8) = 8$ bits.

(ii) Since all belong to the single state (humans $\subset$ mammals), there is no uncertainty about this state and hence the entropy is 0 bits.

(iii) The entropy of this tri-state gender distribution is $-(1/4)(-2) - (1/4)(-2) - (1/2)(-1) = 1.5$ bits.

(iv) In this case both classes have probability 0.5, so the entropy is 1 bit.

(b) For $p(x)$ and $q(x)$ as probability distributions over the integers:
(i) The Kullback-Leibler distance between random variables is defined as

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

(ii) $D_{KL}(p\|q)$ reveals the inefficiency that arises from basing the code on the wrong distribution. It specifies the number of additional bits that would be needed per codeword, on average, if the actual distribution is $p(x)$ instead of $q(x)$.

(iii) The symmetry axiom for distance metrics is violated in this case.

(iv) The Kullback-Leibler distance becomes infinite, or undefined, when some values of x have zero probability for either $p(x)$, or $q(x)$, but not both.

(c) 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:

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

The parameter a (the space-constant in the Gaussian term) 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

$$\lim_{a\to\infty} f(x) = e^{-ik_0 x}$$

– the ordinary Fourier basis. If the frequency parameter k_0 and the size parameter a are 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:

$$\lim_{k_0, a\to 0} f(x) = \delta(x - x_0)$$

Hence the Gabor expansion basis "contains" both of the other two domains of signal representation simultaneously. It allows us to make a continuous deformation that selects a representation lying anywhere on a one-parameter continuum between the two domains that were hitherto distinct.

(d) Transforms such as the Discrete Cosine Transform, the Fourier, and Gabor Transform encode data into a sequence of coefficients on expansion functions. Because these expansion basis functions are decorrelating, the resulting coefficients have non-uniform (and usually quite peaked) distributions. For example, in many cases the peak of the distribution occurs at a coefficient value of 0, and only very few large coefficients (positive or negative) are encountered. Those few "do all the work" but because they are so sparsely distributed, the distribution typically has very low entropy. Shannon's Source Coding Theorem explains that codebooks can be constructed with an average code length per codeword that is no larger than the entropy of the distribution (which in this case is small). Thus even lossless compression (allowing perfect recovery or decompression by expansion) can yield large compression factors. When coarse quantisation of the computed coefficients is also allowed, as done in JPEG compression using the DCT basis or in JPEG-2000 compression using Daubechies wavelets, then even greater compression can be achieved without visible or significant error (e.g. typical image compression factors of 30:1 or more).