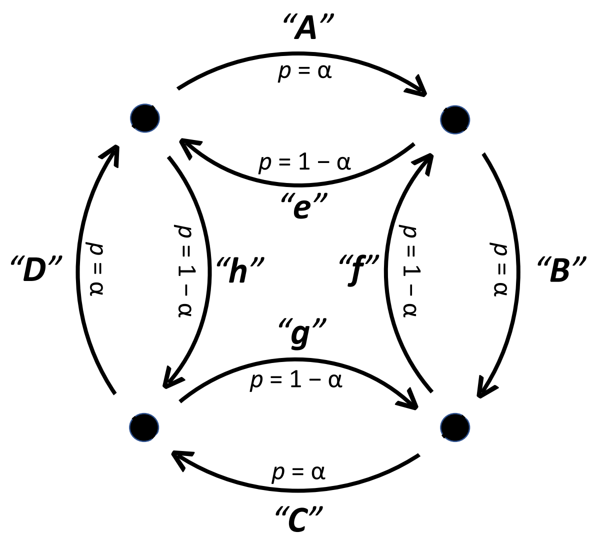
Solutions to Information Theory Exercise Problems 5–9

Exercise 5

Consider the four-state Markov process graphed below. It emits the eight letters $\{A, B, C, D, e, f, g, h\}$ with probabilities and changes of state as shown, but note the sequence constraints. (For example, an A can only be followed by a B or an e.) Letter emissions with clockwise state transitions occur with probability $p = \alpha$, and the others with probability $p = 1 - \alpha$, where $0 < \alpha < 1$.



(a) First imagine a one-state Markov process that emits any of eight letters with equal probabilities. What is its entropy?

(b) For the four-state Markov process shown, having parameter α, what is the long-term probability distribution across the eight letters?

(c) In terms of parameter α, what is the overall entropy $H(\alpha)$ of this four-state Markov process?

(d) Sketch a plot of $H(\alpha)$ as a function of the probability parameter α. Compare its maximum value with your earlier answer in (a) for the one-state Markov process that emits eight letters with equal probabilities, and explain the differences even in the case $\alpha = 0.5$ for this four-state Markov process.

Solution:

(a) The entropy of a one-state Markov process that emits eight letters with equal probabilities is:

$$H = -\sum_i p_i \log_2(p_i) = -\sum_1^8 \frac{1}{8} \log_2\left(\frac{1}{8}\right) = 3 \text{ bits.}$$

(b) The four-state Markov process emits a letter from the set $\{A, B, C, D\}$ with probability α, and it emits a letter from the set $\{e, f, g, h\}$ with probability $1-\alpha$, so the long-term probability distribution across these eight letters is:

letter	A	B	C	D	e	f	g	h
probability	$\frac{\alpha}{4}$	$\frac{\alpha}{4}$	$\frac{\alpha}{4}$	$\frac{\alpha}{4}$	$\frac{1-\alpha}{4}$	$\frac{1-\alpha}{4}$	$\frac{1-\alpha}{4}$	$\frac{1-\alpha}{4}$

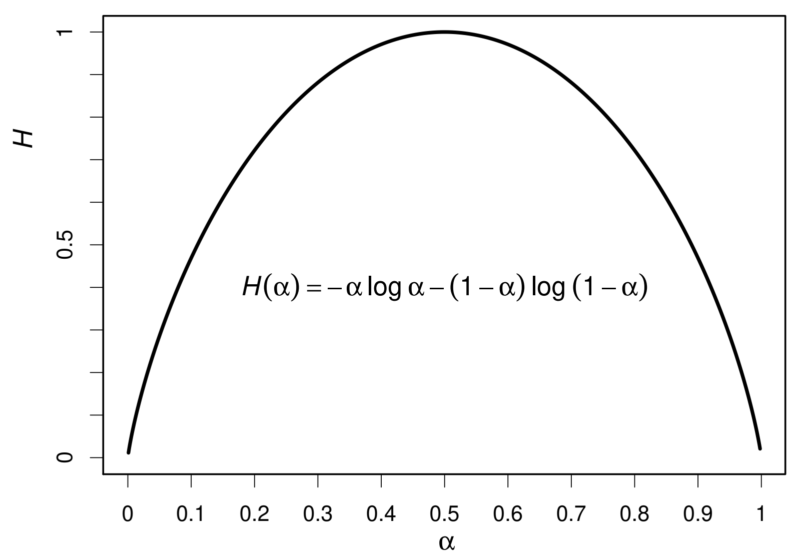
(c) The overall entropy of a multi-state Markov process is the average of the entropies of all states, weighted by their occupancy probabilities (which are all the same in this case). The entropy of each state is:

$$H = -\sum_i p_i \log_2(p_i) = -\alpha \log_2(\alpha) - (1-\alpha)\log_2(1-\alpha).$$

Thus the overall entropy of this four-state Markov process is also

$$H = -\alpha \log_2(\alpha) - (1-\alpha)\log_2(1-\alpha)$$

(d) Plot:



This entropy peaks at 1 bit (when $\alpha = 0.5$), whereas the one-state Markov process giving this same long-term distribution of eight equiprobable letters had entropy 3 bits. The *sequence constraints* have reduced the entropy of this four-state Markov process compared to the one-state process, whose output sequence was unconstrained. The entropy of the four-state Markov process will be further reduced, below 1 bit, if α moves away from 0.5 in either direction.

Exercise 6

(a) An error-correcting (7/4) Hamming code combines four data bits b_3, b_5, b_6, b_7 with three error-correcting bits: $b_1 = b_3 \oplus b_5 \oplus b_7$, $b_2 = b_3 \oplus b_6 \oplus b_7$, and $b_4 = b_5 \oplus b_6 \oplus b_7$. The 7-bit block is then sent through a noisy channel, which corrupts one of the seven bits. The following incorrect bit pattern is received:

b_1	b_2	b_3	b_4	b_5	b_6	b_7
1	1	0	1	0	0	0

Evaluate three syndromes that can be derived upon reception of this corrupted 7-bit block: $s_1 = b_1 \oplus b_3 \oplus b_5 \oplus b_7$, $s_2 = b_2 \oplus b_3 \oplus b_6 \oplus b_7$, $s_4 = b_4 \oplus b_5 \oplus b_6 \oplus b_7$, and provide the corrected 7-bit block that was the original input to this noisy channel.

Solution:

(a) The three syndromes that are derived upon reception of the 7-bit block evaluate to:

$$\begin{aligned} s_1 &= b_1 \oplus b_3 \oplus b_5 \oplus b_7 = 1 \\ s_2 &= b_2 \oplus b_3 \oplus b_6 \oplus b_7 = 1 \\ s_4 &= b_4 \oplus b_5 \oplus b_6 \oplus b_7 = 1 \end{aligned}$$

Because they are not all 0 they show that a 1-bit error did occur, namely at position $b_{s_4 s_2 s_1} = b_7$.

So, the error-corrected 7-bit block, inferred to be the original bit pattern that was the input to this noisy channel, is:

b_1	b_2	b_3	b_4	b_5	b_6	b_7
1	1	0	1	0	0	1

(b) Consider a binary symmetric channel with error probability ϵ that any bit may be flipped. Two possible error-correcting coding schemes are available: Hamming, or simple repetition.

(i) Without any error-correcting coding scheme in place, state all the conditions that would maximise the channel capacity. Include conditions on the error probability ϵ and also on the probability distribution of the binary source input symbols.

Solution:

(i) With no error-correcting coding scheme in place, the capacity of this channel would be maximised if: (1) the binary source had probabilities $(0.5, 0.5)$ for the two input symbols; and (2) the bit error probability was either $\epsilon = 0$, or $\epsilon = 1$.

(ii) If a (7/4) Hamming code is used to deliver error correction for up to one flipped bit in any block of seven bits, provide an expression for the residual error probability P_e that such a scheme would fail.

Solution:

(ii) A (7/4) Hamming code can always correct errors provided that any block of 7 bits contains no more than 1 flipped bit. It will fail if 2 or more bits have flipped. Thus, its residual error rate P_e is simply the probability that *two or more* bits in 7 have flipped (and the rest have not), summed over all possible "ways" of choosing 2 or more from 7. Hence we have a combinatorial term times a probability term, in a binomial series:

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

(iii) If repetition were used to try to achieve error correction by repeating every message an odd number of times $N = 2m+1$, for some integer m followed by majority voting, provide an expression for the residual error probability P_e that the repetition scheme would fail.

Solution:

(iii) If a repetition strategy is used instead, the majority voting scheme will fail if more than half of the number $N = 2m+1$ of repeated transmissions (namely $m+1$ of them) had errors. Thus we again have a binomial remainder series for the residual error rate:

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

A crucial point is that N-fold repetition dilutes the *code rate* by a factor of N.

Exercise 7

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

Solution:

(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) What does the Fourier power spectrum of this class of signals look like?

Solution:

(b) The Fourier power spectrum of this class of signals is flat (uniform over all frequencies), so it is called "white" in analogy with light. Hence the term "white noise".

(c) Consider a noisy continuous communication channel of bandwidth $W = 1$ MHz, which is perturbed by additive white Gaussian noise whose total spectral power is $N_0 W = 1$. Continuous signals are transmitted across such a channel, with average transmitted power $P = 1{,}000$. Give a numerical estimate for the *channel capacity*, in bits per second, of this noisy channel. Then, for a channel having the same bandwidth W but whose signal-to-noise ratio $\frac{P}{N_0 W}$ is four times better, repeat your numerical estimate of capacity in bits per second.

Solution:

(c) The channel capacity of a noisy continuous communication channel having bandwidth W in Hertz, perturbed by additive white Gaussian noise whose whose power spectral density is N_0, when transmitting a signal with average power P (defined by its expected variance), is:

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

Therefore, for the parameter values given, the channel capacity is about 10^7 bits per second.

If the signal-to-noise ratio $\frac{P}{N_0 W}$ of this channel were improved by a factor of four, then its channel capacity would increase to about 12 million bits per second. (Note that $4,000 \approx 2^{12}$).

(d) Suppose that for such a continuous channel with added white Gaussian noise, the ratio of signal power to noise power is given as **30 decibels**, and the frequency bandwidth W of this channel is 10 MHz. Roughly what is the information capacity C of this channel, in bits/second?

Solution:

(d) 30 decibels means that the ratio of signal power to noise power (SNR) is 1,000:1, whose base-2 logarithm is about 10. Thus with the channel's frequency bandwidth now $W = 10$ MHz, we conclude that this channel's information capacity is now $C = 100$ million bits/second.

(e) With no constraints on the parameters of such a channel, is there any limit to its capacity if you increase its signal-to-noise ratio $\dfrac{P}{N_0 W}$ without limit? If so, what is that limit?

Solution:

(e) 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 signal-to-noise ratio, the quantity $\dfrac{P}{N_0 W}$ inside the logarithm, without bounds increases the channel's capacity monotonically without limit (but at a decelerating rate).

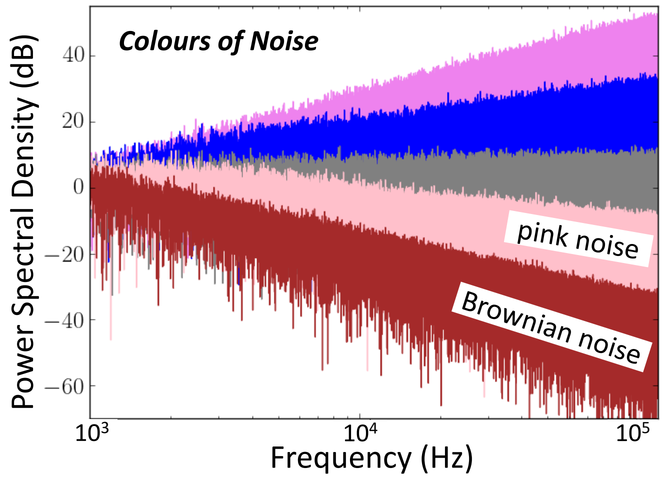
(f) Is there any limit to the capacity of such a channel if you can increase its spectral bandwidth W (in Hertz) without limit, while not changing N_0 or P? If so, what is that limit?

Solution:

(f) 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 x, the quantity $\log_e(1 + x)$ is approximately x. In this case, setting $x = \frac{P}{N_0 W}$ 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.

Exercise 8

Shannon's *Noisy Channel Coding Theorem* showed how the capacity C of a continuous communication channel is limited by added white Gaussian noise; but other colours of noise are available. Among the "power-law" noise profiles shown in the figure as a function of frequency ω, Brownian noise has power that attenuates as $(\frac{\omega}{\omega_0})^{-2}$, and pink noise as $(\frac{\omega}{\omega_0})^{-1}$, above some minimum ω_0.



Consider three channels suffering from either white, pink, or Brownian noise. At frequency $\omega = \omega_0$ all three channels have the same signal-to-noise ratio $\text{SNR}(\omega_0)$ and it remains at this level for the white channel, but at higher frequencies ω it improves as $(\frac{\omega}{\omega_0})$ for the pink channel and as $(\frac{\omega}{\omega_0})^2$ for the Brownian channel. Show that across any frequency band $[\omega_1, \omega_2]$ $(\omega_0 < \omega_1 < \omega_2)$ the Brownian and the pink noise channels have higher capacity than the white noise channel, and show that as frequency grows large the Brownian channel capacity approaches **twice** that of the pink channel.

Solution:

Shannon's *Noisy Channel Coding Theorem* implies that the information capacity of a channel within a band of frequencies $[\omega_1, \omega_2]$ with added noise described by $\text{SNR}(\omega)$ is:

$$C = \int_{\omega 1}^{\omega 2} \log_2(1 + \text{SNR}(\omega))\ d\omega \ \ \text{bits/sec.}$$

Clearly for any band of frequencies above ω_0 the information capacity of the pink and the Brownian channels exceeds that of the white channel, because their $\text{SNR}(\omega)$ is larger. As frequency ω grows large, the "1+" term in the logarithm can be ignored and the capacity of the channel with added pink noise becomes

$$C = \int_{\omega 1}^{\omega 2} \log_2\left(\frac{\omega}{\omega_0}\right)\ d\omega \ \ \text{bits/sec}$$

and the capacity of the channel with added Brownian noise becomes

$$C = \int_{\omega 1}^{\omega 2} \log_2\left(\frac{\omega}{\omega_0}\right)^2\ d\omega \quad = \quad 2 \times \int_{\omega 1}^{\omega 2} \log_2\left(\frac{\omega}{\omega_0}\right)\ d\omega \ \ \text{bits/sec.}$$

We see that the capacity of the Brownian channel approaches twice that of the pink channel, and both are greater than that of the white noise channel having constant SNR.

Exercise 9

(a) An inner product space V is spanned by an orthonormal system of vectors $\{e_1, e_2, \ldots, e_n\}$ so that $\forall i \neq j$ the inner product $\langle e_i, e_j\rangle = 0$, but every e_i is a unit vector so that $\langle e_i, e_i\rangle = 1$. We wish to represent a data set consisting of vectors $u \in \text{span}\{e_1, e_2, \ldots, e_n\}$ in this space as a linear combination of the orthonormal vectors: $u = \sum_{i=1}^{n} a_i e_i$. Derive how the coefficients a_i can be determined for any vector u, and comment on the computational advantage of representing the data in an orthonormal system.

Solution:

(a) Given that $u = \sum_{i=1}^{n} a_i e_i$, when we project the vector u onto any vector e_i we find that all of the inner products generated by this sum equal 0 (due to orthogonality) except for one of them, which equals 1 (due to ortho*normality*):

$$\begin{aligned}\langle u, e_i\rangle &= \langle a_1 e_1 + a_2 e_2 + \cdots + a_n e_n, e_i\rangle \\ &= a_1\langle e_1, e_i\rangle + a_2\langle e_2, e_i\rangle + \cdots + a_n\langle e_n, e_i\rangle \\ &= a_i\,.\end{aligned}$$

Thus the required coefficients a_i which represent the data in this space can be obtained simply by taking the inner product of each data vector u with each of the orthonormal vectors: $a_i = \langle u, e_i\rangle$. A major computational advantage of representing data in an orthonormal system is that the expansion coefficients a_i are just these inner product projection coefficients.

(b) An inner product space containing complex functions $f(x)$ and $g(x)$ is spanned by a set of orthonormal basis functions $\{e_i\}$. Complex coefficients $\{\alpha_i\}$ and $\{\beta_i\}$ therefore exist such that $f(x) = \sum_i \alpha_i e_i(x)$ and $g(x) = \sum_i \beta_i e_i(x)$.

Show that the inner product $\langle f, g\rangle = \sum_i \alpha_i \overline{\beta_i}$.

Solution:

(b)

$$\begin{aligned}\langle f, g\rangle &= \left\langle \sum_i \alpha_i e_i, \sum_j \beta_j e_j \right\rangle \\ &= \sum_i \alpha_i \left\langle e_i, \sum_j \beta_j e_j \right\rangle \\ &= \sum_i \alpha_i \left(\sum_j \overline{\beta_j}\, \langle e_i, e_j\rangle \right) \\ &= \sum_i \alpha_i \overline{\beta_i}\end{aligned}$$

where the last step exploits the fact that $\langle e_i, e_j\rangle = 0$ for $i \neq j$ but $\langle e_i, e_j\rangle = 1$ if $i = j$, because $\{e_i\}$ is an orthonormal basis.

(c) 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?

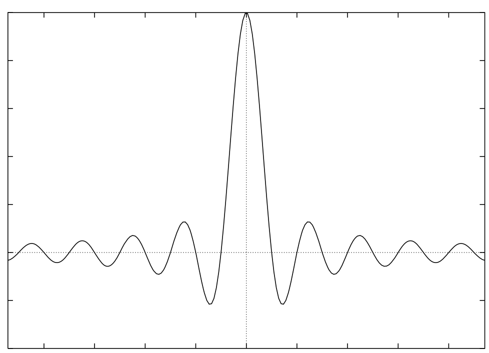
Solution:

(c) As the signal duration is $T = 1$ second and the channel bandwidth is $W = 10,000$ Hertz, the Nyquist Sampling Theorem tells us that $2WT = 20,000$ discrete (regularly spaced) samples are required for exact recovery of the signal, even at points in between the samples.

(d) 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.

Solution:

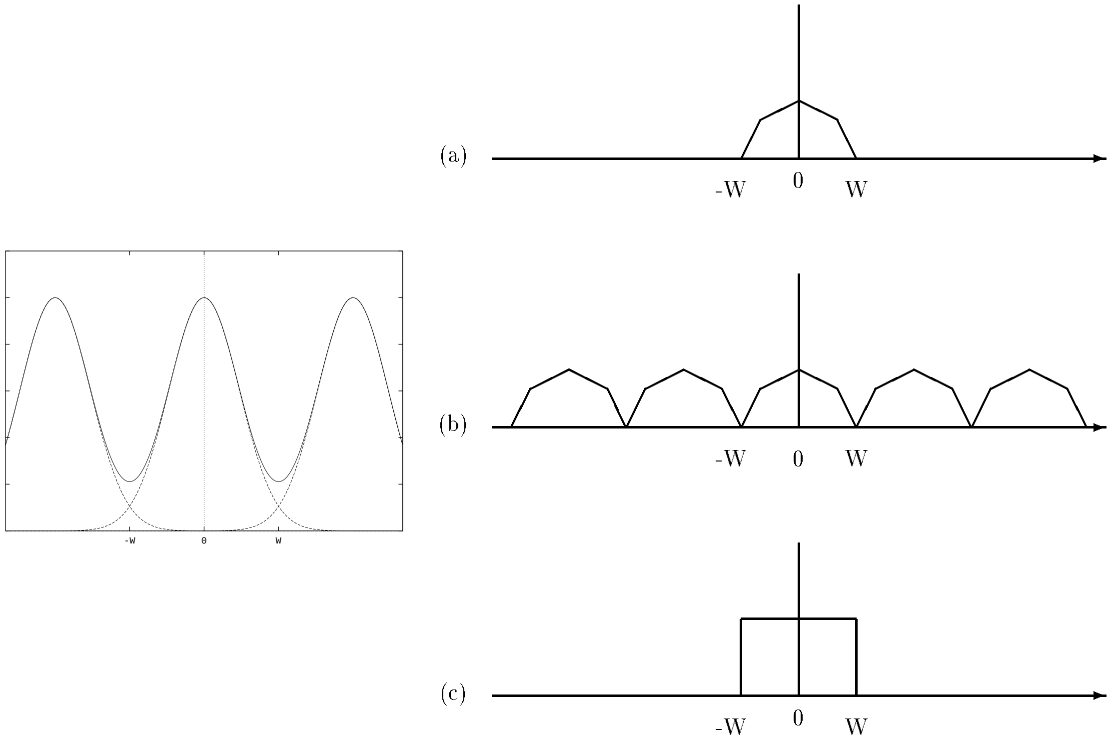
(d) The sinc function is required to recover the signal from its discrete samples, defined as: $\text{sinc}(t) = \dfrac{\sin(2\pi Wt)}{2\pi Wt}$. Each sample point is replaced by scaled copies of this function, scaled by the amplitude of the sample taken, and with its sign. These all superimpose to reproduce the signal exactly, even between the points where it was sampled. (!)



(e) 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.

Solution:

(e) The Nyquist Sampling Theorem tells us that aliasing results when the signal contains Fourier components higher than one-half the sampling frequency. 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 strictly reject all frequencies starting at one-half the planned sampling rate, as indicated below by the $\pm W$ in trace (c).



(*f*) 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?

<u>*Solution:*</u>

(*f*) 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.

Amplitude Modulation communication is based on this principle. It allows many different communications channels to be multiplexed into a single medium, namely the electromagnetic spectrum, by shifting different signals up into separate frequency bands.

Each original signal with its original Fourier spectrum can be recovered by demodulating it back down (this removes each AM carrier). This is equivalent to multiplying the transmitted signal by the conjugate complex exponential, $\exp(-i\omega t)$, in the band-selecting tuner.