

# Information Theory and Noise – PX214

## Weeks 16-19, 2001. 6 CATs

Lecturer: David Cobden

Lecture times:

Wed 14 Feb 12.00 L4	Thu 15 Feb 14.00 H051	Fri 16 Feb 10.00 and 15.00 L5
Wed 21 Feb 12.00 H051	Thu 22 Feb 14.00 H051	Fri 23 Feb 10.00 and 15.00 L5
Wed 28 Feb 12.00 H051	Thu 01 Mar 14.00 H051	Fri 02 Mar 10.00 and 15.00 L5
	Thu 08 Mar 14.00 H051	Fri 09 Mar 10.00 and 15.00 L5

This handout is unique to the first lecture. I don't plan any more handouts, except some question sheets.

### Objectives of the course

To understand principles of signals and transmitting information

To see how information can be quantified

To appreciate information theory in physics

To revise and apply Fourier theory and probability theory

To know the main sources of noise and their properties

### Reading material

Books are a tricky problem for this course. The best textbooks on information theory are out of print! There are lots of more recent texts spawned by information technology, but most contain only small sections that are relevant to us. Therefore, this course will come with a set of notes covering the syllabus that will be available on the web. Below is a selection of books and articles which are either useful, entertaining or easy to obtain; it is hard to be all these at the same time.

#### Introductory level

**Article on [Information Theory](#) on Britannica.com** *R. Gallagher*

Freely available, nice introduction to the field with very few equations.

**The Bit and the Pendulum: From Quantum Computing to M Theory-The New Physics of Information** *Tom Siegfried (Wiley 2000, ISBN: 0471399744, £9.20, not yet in library)*

An entertaining popular-science bedtime read; shows how wide-ranging information theory is nowadays.

**An introduction to Information Theory – Symbols, Signals and Noise**

*John R. Pierce (Dover 1981, ISBN: 0486240614, £9.90, Q360.P4)*

A nice wordy introduction; helpful, but not quite up to the maths level of this course.

**Introduction to quantum computing**

*Oxford Quantum Computing group web page, [http://www.qubit.org/Intros\\_Tuts.html](http://www.qubit.org/Intros_Tuts.html)*

An easy read, for your entertainment.

**Introduction to quantum teleportation**

*IBM research web page, <http://www.research.ibm.com/quantuminfo/teleportation/>*

Likewise.

#### Medium level (closest to this course)

**Information and Communication Theory**

*A. M. Rosie, (van Nostrand 1973, unfortunately out of print, TK5102.R6)*

Contains most of the material for this course, and at the right level, but it's hard to get hold of and a bit out of date.

### **Feynman Lectures on Computation**

*Robin W. Allen (Editor) (Penguin 1999; ISBN: 0140284516, £15.20, not yet in library)*

An inspiring book, as always with Feynman, who takes a unique but amazingly insightful approach to everything. Not very mathematical, but mind-stretching and quite relevant. Relies on some physics which you may not yet have met.

### **An Introduction to Information Theory**

*Fazlollah M. Reza (Dover 1994, ISBN: 0486682102, £11.16, TK5102.R3)*

Mathematically inclined and dry but not too difficult. Doesn't cover the signal analysis part of the course. Contains lots of examples.

### **The Mathematical Theory of Communication**

*Claude E. Shannon and W. Warren (University of Illinois 1949, out of print. Q 360.S4)*

The original and lucid authoritative work, virtually defining modern information theory.

### **Resource Letter ITP-1: Information Theory in Physics**

*W. T. Grandy, Jr., American Journal of Physics, vol 65, p. 466 (1997).*

This review paper is an excellent way both to get a summary of the history and look into the state of the art.

## **Advanced level**

### **Information Theory, Pattern Recognition and Neural Networks**

*Course notes by David MacKay at Cambridge (2001), <http://wol.ra.phy.cam.ac.uk/pub/mackay/itprnn/>*

Goes way beyond our course, but beautifully clear. This is where to learn information theory properly.

### **Information Theory and Statistics**

*Solomon Kullback (Dover 1997, ISBN: 0486696847, £9.55, QA276,K8)*

Strictly for mathematicians.

### **Information and Coding Theory**

*Gareth A. Jones and J. Mary Jones (Springer 2000; ISBN: 1852336226, £18.95, not yet in library)*

Even more so.

## **1. Introductory notes**

In the lecture (also on the web) we will draw a chart indicating the range and context of information theory. Information theory comes into physics at all levels and in many ways. It is a young science, having appeared only around the mid 20<sup>th</sup> century, where it was developed in response to the rapid growth of telecommunications.

Here's a quote from Wheeler (inventor of black holes) given in the intro of *The Bit and the Pendulum*:  
I think of my lifetime in physics as divided into three periods. In the first period, ... I was in the grip of the idea that Everything is Particles ... I call my second period Everything is Fields ... Now I am in the grip of a new vision, that Everything is Information.

Wheeler is a Messiah of information theory. He talks of the new "information paradigm", meaning the way it's now trendy to look at everything in the world in terms of the bits of information which represent it. This kind of resembles the way people once looked at everything in terms of elements (the Greeks), then energy and particles (after Newton), then waves and fields (after Maxwell, Bohr, Einstein *et al*). Is he really onto something? You can decide.

Quote from Simon Benjamin (Science Magazine 2000):

No information without representation! This is the fundamental principle behind quantum information, a new, rapidly evolving field of physics. Information cannot exist without a physical system to represent it, be it chalk marks on a stone tablet or aligned spins in atomic nuclei. And because the laws of physics govern any such system, physics ultimately determines both the nature of information and how it can be manipulated. Quantum physics enables fundamentally new ways of information processing, such as procedures for "teleporting" states between remote locations, highly efficient algorithms for seeking solutions to equations and for factorization, and protocols for perfectly secure data transmission.

# Syllabus

## Fourier analysis

Real and complex Fourier series.  
Negative frequency.  
Parseval's theorem.  
Fourier transforms, and related theorems.  
The frequency-shift and time-shift theorems.  
Convolution and the convolution theorem  
The impulse response and the transfer function of a filter.  
The Dirac delta-function and its Fourier properties.  
The sampling theorem, the cardinal series, and limitations imposed by causality.

## Elementary probability theory

The random variable, expectation for discrete and continuous random variables.  
Mean, mean-square, and variance.  
Common distributions  
Joint and conditional probabilities.  
Adding random variables; characteristic functions.  
The central limit theorem.

## Noise

Autocorrelation functions and the Weiner-Khintchine Theorem.  
Spectral properties of white and coloured noise.  
Shot noise - the autocorrelation function, power spectrum, and statistics.  
Thermal noise (Nyquist argument).  
1/f-noise in large and small structures.  
Generality of 1/f phenomena in nature.

## Information theory

Definition of information: sources, channels.  
Application to language.  
Memoryless (Markovian) processes.  
Classification entropy and Shannon's first theorem.  
Twenty questions.  
Redundancy.  
Huffman coding; lossless and lossy compression  
Error correction and message space.  
Conditional and mutual information.  
The channel matrix.  
The channel capacity.  
Shannon's noisy coding theorem.  
The Shannon-Hartley Law.

## Selected topics from the following (time permitting)

Relationship of information theory to statistical mechanics.  
Quantum computing  
Quantum teleportation  
Reversible computation  
Maxwell's demon and the validity of statistical mechanics  
Information in the genetic code

## Time and frequency representations of a signal.

First we need to consider some mathematical properties of a signal  $g(t)$ . Thanks to Fourier, we know that in many cases you can choose to study the signal either in the time domain or in the frequency domain. This is really one of the fundamental techniques of physics, which can be applied to all linear systems (described by linear equations, such as the wave equation). You have met it before, so we won't dwell on details.

### 2 Fourier series

#### 2.1 Fourier's theorem

Consider a signal  $g(t)$  which is periodic in time  $t$ , with period  $T$ , so that  $g(t+T) = g(t)$  for  $-\infty < t < \infty$ .

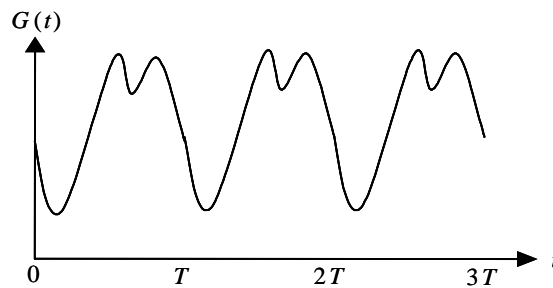


Figure 2.1

If  $g(t)$  is not pathological, Fourier's theorem says it can be expanded as a Fourier sum over components at frequencies  $2\pi n/T$ , where  $n$  is a positive integer:

$$g(t) = a_0 + 2 \sum_{n=1}^{\infty} \left[ a_n \cos\left(\frac{2\pi n t}{T}\right) + b_n \sin\left(\frac{2\pi n t}{T}\right) \right] \quad (2.1)$$

The Fourier coefficients  $a_n$  and  $b_n$  are given by

$$a_n = \frac{1}{T} \int_{-T/2}^{T/2} g(t) \cos\left(\frac{2\pi n t}{T}\right) dt \quad (2.2)$$

$$b_n = \frac{1}{T} \int_{-T/2}^{T/2} g(t) \sin\left(\frac{2\pi n t}{T}\right) dt. \quad (2.3)$$

To prove this, we multiply Eq. (2.1) by  $\cos(2\pi n t/T)$  or  $\sin(2\pi n t/T)$  and integrate from  $-T/2$  to  $T/2$ , then use orthogonality relationships such as

$$\frac{2}{T} \int_{-T/2}^{T/2} \cos\left(\frac{2\pi m t}{T}\right) \cos\left(\frac{2\pi n t}{T}\right) dt = d_{m,n} \quad (2.4)$$

The frequency of the waveform is  $1/T$ , and  $2\pi/T$  is called the *angular* frequency. We will stick to using the real frequency in these notes.

#### 2.2 Interpretation of the Fourier coefficients

The relative sizes of all the  $a_n$  and  $b_n$  determine the shape of the waveform. If all the  $c_n$ 's and  $T$  are known,  $g(t)$  is completely specified.  $a_n$  is the *correlation* of  $g(t)$  with a cosine wave at frequency  $2\pi n/T$  (see later for correlation functions).  $a_0$  is the correlation with  $1/T$ , ie, the mean value of the signal.

### 2.3 Complex Fourier series

The same thing can be written more compactly in complex notation. This is an example of how useful complex numbers are in dealing with oscillating, and therefore periodic, phenomena. If we define complex coefficients

$$c_n = a_n - ib_n = \frac{1}{T} \int_{-T/2}^{T/2} g(t) \exp\left(\frac{-2\pi i n t}{T}\right) dt \quad (2.5)$$

and

$$c_{-n} = a_n + ib_n = c_n^* \quad (2.6)$$

then we find that

$$g(t) = \sum_{n=-\infty}^{\infty} c_n \exp\left(\frac{2\pi i n t}{T}\right) \quad (2.7)$$

Eqs. (2.5) and (2.7) work together because an orthogonality relationship for complex exponentials:

$$\frac{1}{T} \int_{-T/2}^{T/2} \exp\left(\frac{2\pi i m t}{T}\right) \exp\left(\frac{-2\pi i n t}{T}\right) dt = \delta_{m,n} \quad (2.8)$$

The phase of  $c_n$  gives the phase, or time shift, of the frequency component at  $2\pi n/T$ . Having  $c_{-n} = c_n^*$  makes  $g(t)$  real and keeps the number of unknown coefficients the same.

### 2.4 Negative frequencies

Notice that in Eq. (2.6) we have introduced negative frequencies, with  $n < 0$ . These go inevitably with complex numbers and are there only for convenience. Consider the real signal

$$g(t) = \cos\left(\frac{2\pi t}{T}\right) = \frac{1}{2} \exp\left(\frac{2\pi i t}{T}\right) + \frac{1}{2} \exp\left(\frac{-2\pi i t}{T}\right) \quad (2.9)$$

Comparing this with Eq. (2.7) we see that  $c_1 = 1/2$ ,  $c_{-1} = 1/2$ , and  $c_n = 0$  for all other  $n$ . Thus the signal can be thought of as the sum of two equal-amplitude, counter-rotating components. The phase of the  $c_1$  component decreases with time.

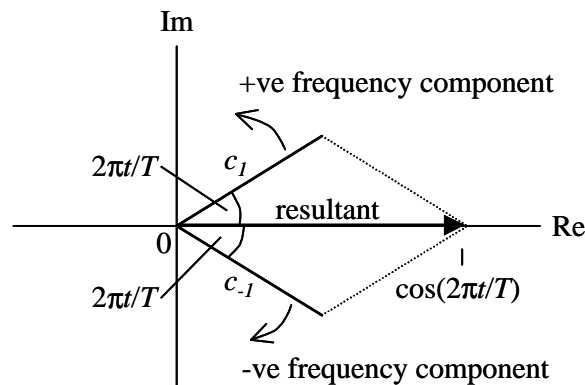


Figure 2.2

We will always use the complex notation, and therefore remember that frequency spectra always contain both negative and positive components. Any real signal has equal amounts of each.

## 2.5 Power and Parseval's theorem

What actually is  $g(t)$ ? It may be many things, but very often it is voltage  $V$  or current  $I$ . In that case, we can define the *power* as the electrical power the signal can dissipate in a  $1 \Omega$  resistor, which conveniently is just  $V^2$  or  $I^2$ , and so we don't need to specify whether we're talking about voltage or current.

We therefore define the normalized power, averaged over time, to be

$$P = \frac{1}{T} \int_{-T/2}^{T/2} |g(t)|^2 dt. \quad (2.10)$$

The modulus of  $g(t)$  is taken so that we can use the complex representation for voltages or currents.

Thus

$$P = \frac{1}{T} \int_{-T/2}^{T/2} g^*(t)g(t)dt, \quad (2.11)$$

Replace  $g(t)$  by its Fourier series,

$$P = \frac{1}{T} \int_{-T/2}^{T/2} g^*(t) \sum_{n=-\infty}^{\infty} c_n \exp\left(\frac{2\pi i n t}{T}\right) dt,$$

interchange the order of summation and integration,

$$P = \frac{1}{T} \sum_{n=-\infty}^{\infty} c_n \int_{-T/2}^{T/2} g^*(t) \exp\left(\frac{2\pi i n t}{T}\right) dt,$$

and use Eq. (2.5) to get

$$P = \sum_{n=-\infty}^{\infty} c_n c_n^* = \sum_{n=-\infty}^{\infty} |c_n|^2 \quad (2.12)$$

This is Parseval's power theorem. It is clear from this expression that the power associated the  $n$ 'th Fourier component is  $P_n = |c_n|^2$ , the square of the modulus of the coefficient for that component. The powers of the different components just add.

## 3 Fourier transforms

### 3.1 Signals of finite power and finite duration.

In real life we need to deal with nonperiodic signals. First, consider a signal  $g(t)$  which has finite time duration, going to zero at large  $|t|$ .

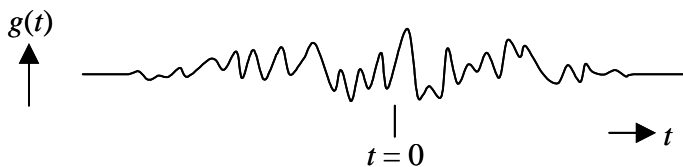


Figure 3.1. A finite duration signal.

We can deal with this by imagining it repeats at large  $T$ , working out the coefficients in its Fourier series, and then taking the limit  $T \rightarrow \infty$ . Starting with Eqs. (2.5) and (2.7), we substitute in

$$\Delta f = 1/T \quad f_n = n/T \quad G(f_n) = c_n T.$$

Then letting  $T \rightarrow \infty$ , so  $\Delta f \rightarrow df$ , we get

$$g(t) = \int_{-\infty}^{\infty} G(f) \exp(2\pi i f t) df \quad (3.1)$$

$$G(f) = \int_{-\infty}^{\infty} g(t) \exp(-2\pi i f t) dt \quad (3.2)$$

$G(f)$  is called the spectrum of  $g(t)$ . It is the continuous counterpart  $c_n$  in the Fourier expansion.  $g(t)$  and  $G(f)$  are a Fourier transform pair. Eq. (3.2) is a Fourier transform; Eq. (3.1) is an inverse Fourier transform, which looks the same but with the opposite sign inside the exponent. We abbreviate Eqs. (3.1) and (3.2) by

$$g(t) = \text{F.T.}^{-1}\{G(f)\}$$

$$G(f) = \text{F.T.}\{g(t)\}$$

All the same *information* is contained in  $G(f)$  as in  $g(t)$ , as each can be obtained from the other. We can either look at this signal as a collection of points in time or a collection of points in frequency space. The F.T. decomposes  $g(t)$  into sinusoidal components at different frequencies with amplitudes given by  $G(f)$ . The decomposition is only possible because of the *orthogonality* of any two of these components. This is all because the product of  $\cos(2\pi f_1 t)$  and  $\cos(2\pi f_2 t)$  averages to zero whenever  $f_1 \neq f_2$ , because the two waves are out of phase just as often as they are in phase.

Finally, since  $G(f_{-n}) = c_{-n} T = c_n^* T = G^*(f_n)$  we have

$$G(-f) = G^*(f). \quad (3.3)$$

## 3.2 Properties of the Fourier transform

Here are three simple theorems involving F.T.'s which can greatly simplify many calculations.

### 3.2.1 Time scaling theorem

$$\text{F.T.}\{g(at)\} = \int_{-\infty}^{\infty} g(at) \exp(-2\pi i f t) dt$$

Put  $\tau = at$ , so  $dt = a^{-1} d\tau$ ,

$$= \frac{1}{a} \int_{-\infty}^{\infty} g(\tau) \exp(-2\pi i f \tau / a) d\tau = \frac{1}{a} G(f/a)$$

So

$$\text{F.T.}\{g(at)\} = \frac{1}{a} G(f/a) \quad (3.4)$$

### 3.2.2 Time shifting theorem

$$\text{F.T.}\{g(t-t_0)\} = \exp(-2\pi i f t_0) G(f) \quad (3.5)$$

### 3.2.3 Frequency shifting theorem

$$\text{F.T.}\{\exp(-2\pi i f_0 t) g(t)\} = G(f + f_0) \quad (3.6)$$

### 3.3 Example: Fourier transform of a square pulse ('top-hat function')

Consider the signal  $g(t) = A$  for  $-T/2 < t < T/2$  and  $g(t) = 0$  otherwise.

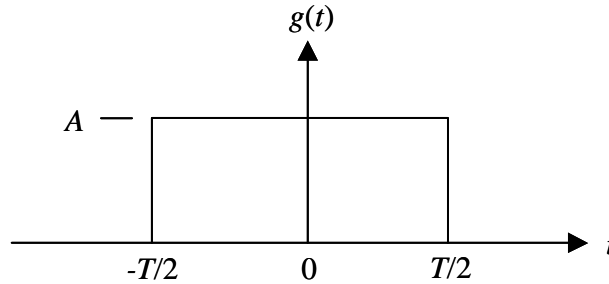


Figure 3.2.  
Square pulse, or  
'top-hat'

Its Fourier transform is

$$\begin{aligned}
 G(f) &= \int_{-T/2}^{T/2} A \exp(-2\pi i f t) dt \\
 &= A \left[ \frac{\exp(-2\pi i f t)}{-2\pi i f} \right]_{-T/2}^{T/2} \\
 &= \frac{A}{\pi f} \left( -\frac{1}{2i} \right) [\exp(-\pi i f T) - \exp(\pi i f T)] \\
 &= \frac{A}{\pi f} \sin \pi f T = AT \frac{\sin \pi f T}{\pi f T}
 \end{aligned}$$

By definition  $\text{sinc}(x) \equiv \sin(x)/x$ , so  $G(f) = AT \text{sinc}(\pi f T)$ . (3.7)

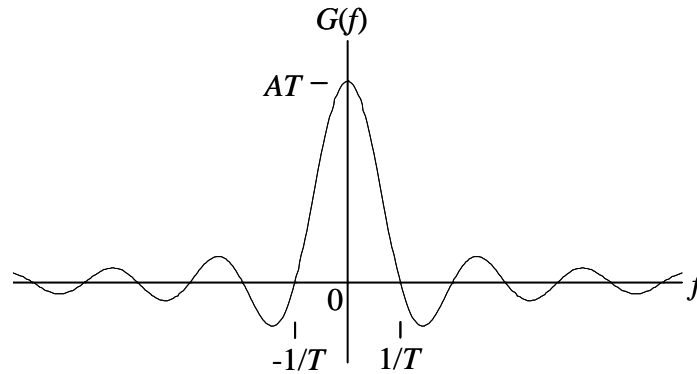


Figure 3.3.  
Spectrum of a  
square pulse.

#### Notes:

- (1) This is almost the only F.T. we will actually ever calculate by integration. In most cases we can use the various F.T. theorems (such as 3.2.1, 2 and 3) to avoid doing integrals.
- (2) The width of the spectrum is  $\sim 1/T$ , so that as usual, there is an inverse relation between the time and frequency representations of the signal:

$$\text{(duration of pulse)} \times \text{(width of spectrum)} \approx 1$$

- (3) In optics, the sinc function is the Fraunhofer diffraction pattern for light passing through a rectangular slit. In that case, the image on the screen is a *spatial* Fourier transform of the shape of the slit, but the maths is just the same as above.

### 3.4 Finite power, infinite duration signals

Most often we have to deal with signals that continue indefinitely, such as electrical noise. These may not be periodic, but they are **stationary**, meaning roughly that they continue to behave in the same sort of way as  $t \rightarrow \infty$ .



Figure 3.4. A stationary signal

We can apply Fourier theory to such signals, as long as we introduce the weird and wonderful Dirac  $\delta$ -function.

### 3.5 The $\delta$ -function

The  $\delta$ -function is effectively an infinitely sharp spike with a well defined area (integral). We need it here because if we allow time durations to go to infinity we have to allow the widths of spectral features to go to zero. For example, if the top-hat function above becomes very, very wide, the sinc function becomes very, very narrow – but its area  $A$  remains constant. In fact in many circumstances the sinc function with  $A = 1$  can be used as a  $\delta$ -function. But the  $\delta$ -function is really defined by its properties:

$$\delta(t) = 0 \quad \text{for } t \neq 0 \quad (3.8)$$

but 
$$\int_{-\infty}^{\infty} \delta(t) dt = 1 \quad (3.9)$$

and 
$$\int_{-\infty}^{\infty} g(t) \delta(t) dt = g(0) . \quad (3.10)$$

From (3.10) it follows that 
$$\int_{-\infty}^{\infty} g(t) \delta(t - t_0) dt = g(t_0) . \quad (3.11)$$

This means that when it appears inside an integral,  $\delta(t - t_0)$  picks out the value of the rest of the integrand at  $t = t_0$ .

Also from (3.10), 
$$\text{F.T.}\{\delta(t)\} = 1 . \quad (3.12)$$

This means the spectrum of a delta function is completely flat, between  $\pm\infty$  in frequency!

Then from the time-shift theorem, 
$$\text{F.T.}\{\delta(t - t_0)\} = \exp(-2\pi i f t_0) . \quad (3.13)$$

From the frequency-shift theorem, 
$$\text{F.T.}\{\exp(-2\pi i f_0 t)\} = \delta(f - f_0) . \quad (3.14)$$

### 3.6 Example: Fourier transform of a cosine wave

Consider 
$$g(t) = \cos(2\pi f_0 t) = \frac{1}{2} [\exp(2\pi i f_0 t) + \exp(-2\pi i f_0 t)]$$

From Eq. (3.14), its spectrum is 
$$G(f) = \frac{1}{2} \delta(f - f_0) + \frac{1}{2} \delta(f + f_0) . \quad (3.15)$$

We see the same two components as in Eq. (2.9). However now the spectrum  $G(f)$  is continuous and we can plot it as a function. We can add any other frequency we like to this spectrum, which we couldn't do when we were talking about Fourier series because then the only allowed frequencies were multiples of  $\pi/T$ .

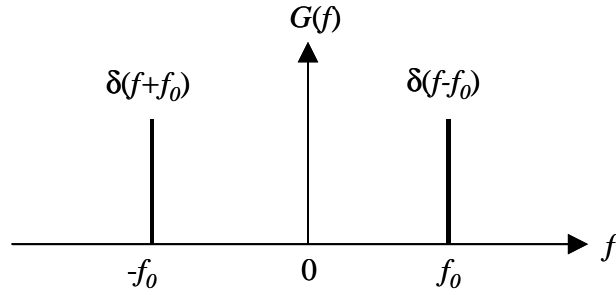


Figure 3.5

## 4 Transmission of signals through linear systems

By a system we mean something (perhaps a device, perhaps a communications channel) that produces an output signal  $y(t)$  in response to an input signal  $x(t)$ . The action of the system can be represented by an operator (mathematically, a functional)  $L$  such that  $y(t) = L\{x(t)\}$ .

### 4.1 Linear systems

A ‘linear’ system is one that obeys the principle of superposition, ie,

$$L\{ax_1(t) + bx_2(t)\} = aL\{x_1(t)\} + bL\{x_2(t)\} = ay_1(t) + by_2(t) \quad (4.1)$$

This means that if two signals  $x_1(t)$  and  $x_2(t)$  are superimposed at the input, the output is what you’d get if you fed the signals through the system separately and then superimposed them afterwards. The two signals are not *mixed* by the system.

### 4.2 Impulse response function

Consider the output of a linear system when the input is a very sharp pulse, ie, a  $\delta$ -function.

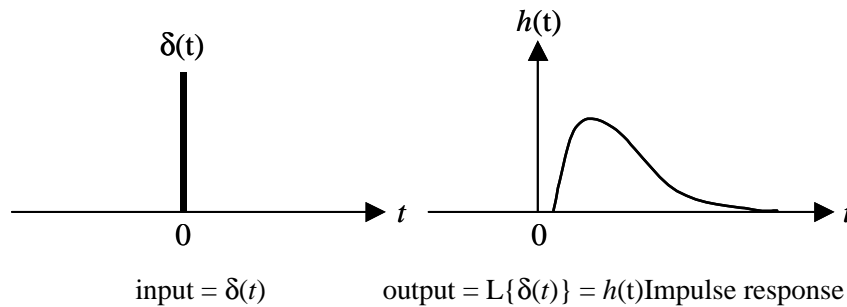


Figure 4.1

The output  $h(t) = L\{\delta(t)\}$  is called the impulse response function. Its shape reflects the physical delays and response mechanisms within the system. The behaviour of a *linear* system is completely characterised by  $h(t)$ . For a general input  $g_1(t)$ , we can write the output in terms of the impulse response function using Eq. (3.11):

$$\begin{aligned} g_2(t) &= L\{g_1(t)\} = L\left\{\int_{-\infty}^{\infty} g_1(\tau) \delta(\tau - t) d\tau\right\} \\ &= \int_{-\infty}^{\infty} g_1(\tau) L\{\delta(\tau - t)\} d\tau \end{aligned} \quad (4.3)$$

using the fact that  $L$  doesn’t operate on dummy variable  $\tau$ .  $L\{\delta(\tau - t)\}$  is the response to a  $\delta$ -function spike at  $t = \tau$ , which is  $h(t - \tau)$ . Therefore,

$$g_2(t) = \int_{-\infty}^{\infty} g_1(\tau) h(t - \tau) d\tau. \quad (4.4)$$

### 4.3 Convolution

Eq. (4.3) defines the function  $g_2(t)$  as the *convolution* of the two functions  $g_1(t)$  and  $h(t)$ . It's very important to remember that it's  $t-\tau$ , not  $\tau-t$ , in the second function in the integrand for a convolution. Get it the wrong way round and you have a *correlation* instead of a convolution (see Section 8.x). A shorthand notation for the convolution integral is

$$g_2(t) = g_1(t) \otimes h(t) \quad (4.5)$$

Convolution mixes two functions together to make a single hybrid function. We can understand it as follows, in terms of the above derivation. Eqs. (4.3) and (3.11) say that  $g_1(t)$  can be represented as a train of back-to-back  $\delta$ -functions with amplitude  $g_1(\tau)$ . Eq. (4.4) says that the output is the sum of the responses  $h(t-\tau)$  of the system to all the pulses in this train. Take for example  $g_1(t)$  to be a top-hat function. Then  $g_2(t)$  is a smeared version of it, as shown below. The broader is the impulse response function  $h(t)$ , the more smeared is the output  $g_2(t)$ .

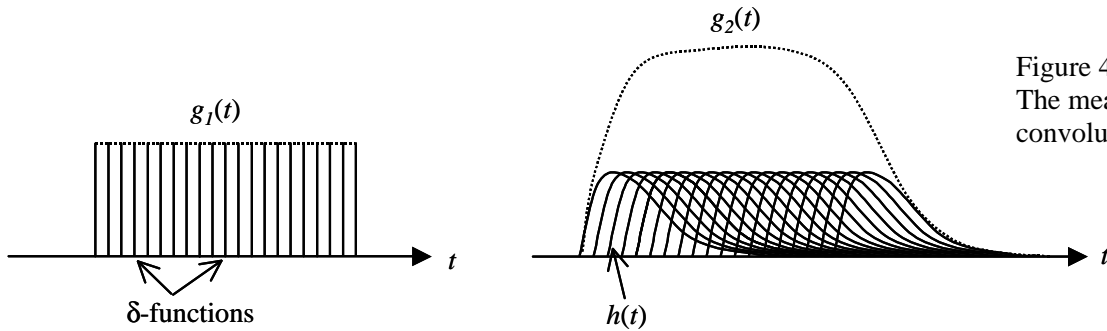


Figure 4.2.  
The meaning of convolution.

You can show that  $g_1(t) \otimes h(t) = h(t) \otimes g_1(t)$ , i.e. convolution commutes. This means that we get the same result if we chop  $h(t)$  into  $\delta$ -functions and pretend that  $g_1(t)$  is the impulse response function.

### 4.4 The convolution theorem

Taking the Fourier transform of the convolution function yields what is probably the most powerful theorem in Fourier theory. For any two functions  $g(t)$  and  $h(t)$ , we have

$$F.T.\{g(t) \otimes h(t)\} = \int_{-\infty}^{\infty} \left[ \int_{-\infty}^{\infty} g(\mathbf{t})h(t-\mathbf{t})d\mathbf{t} \right] \exp(-2\mathbf{pif}t)dt . \quad (4.6)$$

Reverse the order of integration,

$$= \int_{-\infty}^{\infty} g(\mathbf{t}) \left[ \int_{-\infty}^{\infty} h(t-\mathbf{t})\exp(-2\mathbf{pif}t)dt \right] d\mathbf{t} ,$$

make the substitution  $t = \mathbf{a} + \mathbf{t}$ ,

$$= \int_{-\infty}^{\infty} g(\mathbf{t}) \left[ \int_{-\infty}^{\infty} h(\mathbf{a})\exp\{-2\mathbf{pif}(\mathbf{a} + \mathbf{t})\}d\mathbf{a} \right] d\mathbf{t}$$

$$= \int_{-\infty}^{\infty} g(\mathbf{t})\exp(-2\mathbf{pif}t)dt \left[ \int_{-\infty}^{\infty} h(\mathbf{a})\exp(-2\mathbf{pif}t)da \right] ,$$

and rename the dummy variables,

$$= \left[ \int_{-\infty}^{\infty} g(\mathbf{t})\exp(-2\mathbf{pif}t)dt \right] \left[ \int_{-\infty}^{\infty} h(\mathbf{t})\exp(-2\mathbf{pif}t)dt \right]$$

and you get the **convolution theorem**,  $F.T.\{g(t) \otimes h(t)\} = F.T.\{g(t)\}F.T.\{h(t)\} . \quad (4.7)$

A useful corollary of this is  $F.T.\{g(t)\} \otimes F.T.\{h(t)\} = F.T.\{g(t)h(t)\} . \quad (4.8)$

Together these are summarized by remembering that **Fourier transforming converts a convolution to a product, and a product to a convolution.**

## 4.5 The transfer function

For the linear system, we define the spectrum of the input  $G_2(f) = F.T.\{g_2(t)\}$  and the output,  $G_1(f) = F.T.\{g_1(t)\}$ , and define the transfer function  $H(f) = F.T.\{h(t)\}$ , Eqs. (4.5) and (4.7) tell us that

$$G_2(f) = G_1(f)H(f) . \quad (4.9)$$

In words, the output of a linear system can be obtained by multiplying each spectral component of the input by the complex number which is the value of the transfer function  $H(f)$  at that frequency. Just as in the time domain, the system is completely characterised by the impulse response function  $h(t)$ , in the frequency representation it is completely characterised by the transfer function  $H(f)$  which is the Fourier transform of  $h(t)$ .

## 5 Nyquist's sampling theorem

We've now done enough maths to be able to prove possibly the oldest result in information theory, first derived by Nyquist (1924).

### 5.1 Statement

The sampling theorem states that **a signal of bandwidth  $B$  can be perfectly reconstructed from samples taken with period  $T$  providing  $1/T > 2B$** . In this case the discrete set of sampled points contains exactly the same *information* as the (continuous) original signal. (Note: having bandwidth  $B$  means the spectrum is nonzero only for frequencies  $f \leq B$ .) This means that if  $1/T > 2B$  there is only one, unique way of drawing the original signal through the black dots (sample points) in the following sketch, as indicated by the solid line. If  $1/T < 2B$  there are other ways which involve extra wiggles between the dots, such as the dotted line.

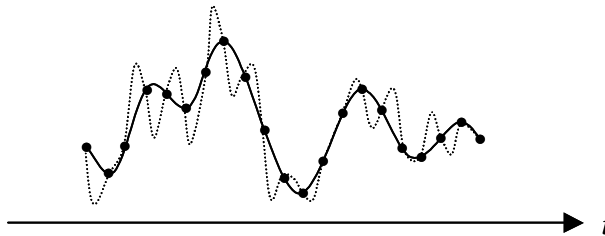


Figure 5.1. Two possible signals that fit one set of sampled data points

### 5.2 Proof

The original signal  $g_1(t)$ , whose spectrum is  $G_1(f) = F.T.\{g_1(t)\}$ , is sampled at times  $t = nT$  for all integer  $n$ . Let us define the function  $g_s(t)$  to be the sum of a set of  $\delta$ -functions at the sampled times, each weighted by  $g_1(t)$ .

$$\begin{aligned} g_s(t) &= \sum_{n=-\infty}^{\infty} \mathbf{d}(t - nT) g_1(t) \\ &= g_1(t) \left[ \sum_{n=-\infty}^{\infty} \mathbf{d}(t - nT) \right] \end{aligned} \quad (5.1)$$



Figure 5.2. The function  $g_s(t)$ .

Asking whether  $g_1(t)$  can be reconstructed from the samples is equivalent to asking whether it can be obtained from  $g_s(t)$ . To do this, we first Fourier transform Eq. (5.1) and use the convolution theorem, Eq. (4.8), to obtain the spectrum of  $g_s(t)$ :

$$\begin{aligned} G_s(f) &= F.T.\{g_s(t)\} = F.T.\left\{g_1(t) \sum_{n=-\infty}^{\infty} \mathbf{d}(t-nT)\right\} \\ &= F.T.\{g_1(t)\} \otimes F.T.\left\{\sum_{n=-\infty}^{\infty} \mathbf{d}(t-nT)\right\} \\ &= G_1(f) \otimes F.T.\left\{\sum_{n=-\infty}^{\infty} \mathbf{d}(t-nT)\right\} \end{aligned} \quad (5.2)$$

To deal with the sum of  $\delta$ -functions on the right we Fourier expand it:

$$\sum_{n=-\infty}^{\infty} \mathbf{d}(t-nT) = \sum_m c_m \exp \frac{2\mathbf{p}imt}{T}.$$

Now 
$$c_m = \frac{1}{T} \int_{-T/2}^{T/2} \mathbf{d}(t) \exp\left(\frac{-2\mathbf{p}imt}{T}\right) dt = \frac{1}{T},$$

so 
$$\sum_{n=-\infty}^{\infty} \mathbf{d}(t-nT) = \frac{1}{T} \sum_{m=-\infty}^{\infty} \exp \frac{2\mathbf{p}imt}{T}. \quad (5.3)$$

Therefore 
$$F.T.\left\{\sum_{n=-\infty}^{\infty} \mathbf{d}(t-nT)\right\} = \frac{1}{T} F.T.\left\{\sum_{m=-\infty}^{\infty} \exp \frac{2\mathbf{p}imt}{T}\right\},$$

and using Eq. (3.14), 
$$= \frac{1}{T} \sum_{m=-\infty}^{\infty} \mathbf{d}\left(f - \frac{m}{T}\right). \quad (5.4)$$

Hence the spectrum of a ‘fence’ of  $\delta$ -functions is another fence of  $\delta$ -functions.

Putting this into Eq. (5.2) gives 
$$G_s(f) = G_1(f) \otimes \left[\frac{1}{T} \sum_{m=-\infty}^{\infty} \mathbf{d}\left(f - \frac{m}{T}\right)\right]. \quad (5.5)$$

We thus see that the spectrum  $G_s(f)$  of the sampled signal is a convolution of the spectrum  $G_1(f)$  of the original signal and this fence of  $\delta$ -functions at frequencies  $m/T$ . The result is the shape of  $G_1(f)$  centered on each  $\delta$ -function position:

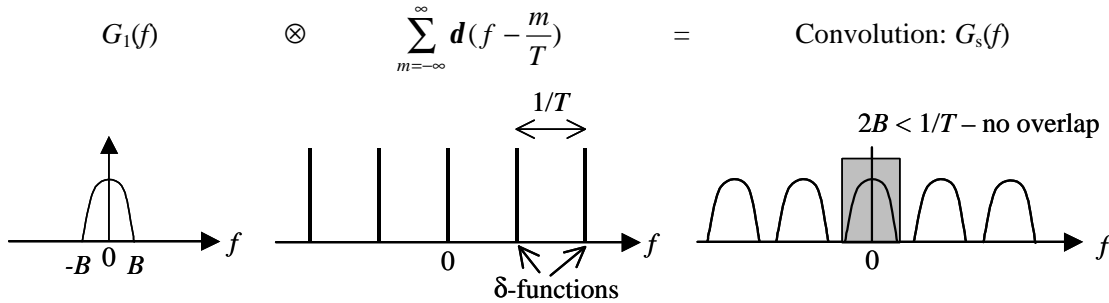


Figure 5.3

If  $1/T > 2B$ , the images of  $G_1(f)$  in the convolution don't overlap. In this case, if one filters out from  $G_s(f)$  all components with  $f > 2B$ , the result is  $G_1(f)$ , which we know contains exactly the same information as its Fourier transform  $g_1(t)$ . We conclude that in this case the sampled values contain all the information needed to reconstruct  $g_1(t)$ .

If  $1/T < 2B$ , spectral components from different parts of  $G_1(f)$  are mixed together in  $G_s(f)$ .

They can no longer be separated by filtering, and the result is an unavoidable mixing of frequency components with  $f > 1/T$  down to lower frequencies in the reconstructed signal, which is known as *aliasing*.

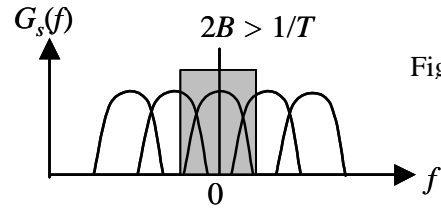


Figure 5.4

### 5.3 Signal reconstruction

How should we actually reconstruct a continuous signal  $g_{rec}(t)$  from the sampled data points that will equal  $g_1(t)$  in the case  $1/T > 2B$ ? According to the above we obtain the spectrum of  $g_{rec}(t)$  from the sampled values by generating  $g_s(t)$ , Fourier transforming it, and remove spectral components with  $|f| > B$  while leaving those with  $|f| < B$  unchanged. Thus

$$F.T.\{g_{rec}(t)\} = F.T.\{g_s(t)\}H(f)$$

where  $H(f)$  is the transfer function of the filter used, as in Section 4.5. This means that  $g_{rec}(t)$  is obtained by filtering  $g_s(t)$ . As we showed in Section 4 then,

$$g_{rec}(t) = g_s(t) \otimes F.T.^{-1}\{H(f)\}. \quad (5.6)$$

The best possible filter has a top-hat transfer function,  $H(f) = 1/2B$  for  $|f| < B$  and  $H(f) = 0$  otherwise. Using the same integration as in Section 3.3, we find that

$$F.T.^{-1}\{H(f)\} = \text{sinc}(2\pi Bt),$$

so

$$\begin{aligned} g_{rec}(t) &= \left[ \sum_{n=-\infty}^{\infty} g_1(t) \mathbf{d}(t - nT) \right] \otimes \text{sinc}(2\pi Bt), \\ &= \sum_{n=-\infty}^{\infty} \left[ \int_{-\infty}^{\infty} g_1(t) \mathbf{d}(t - nT) \text{sinc}[2\pi B(t - t)] dt \right] \\ &= \sum_{n=-\infty}^{\infty} g_1(nT) \text{sinc}[2\pi B(t - nT)]. \end{aligned} \quad (5.7)$$

Hence the signal can be reconstructed by adding together a series of sinc functions centered at the sampling times and multiplied by the sampled values. At any sampling point, say  $t = mT$ , all the sinc functions vanish except the one with  $n = m$ ,  $\text{sinc}[2\pi B(t - mT)]$ .

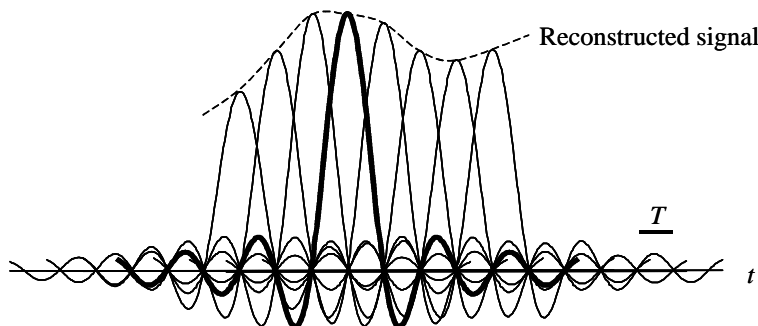


Figure 5.5

If we use a different filter, where  $H(f)$  doesn't drop to zero immediately above  $f = B$  but only does so at  $f_{max} > B$ , then we need  $1/T > 2f_{max}$  to avoid catching the tails of the next image of  $G_1(t)$  in  $G_s(f)$  (see Fig. (5.3)).

## 5.4 Causality

The ideal top-hat filter has an impulse response function  $h(t) = F.T.^{-1}\{H(f)\} = \text{sinc}(2\pi Bt)$ . This is fine if you start with the sampled data and use a computer to do the above reconstruction. However, any *analog* filter must be *causal*, which means its impulse response function  $h(t)$  can be nonzero only for  $t > 0$ . That is, the output can't be earlier than the input. The time-symmetric sinc function is therefore not realizable in an analog circuit. In fact, the amplitude and phase of  $H(f)$  for a real filter are connected by a "dispersion" equation, and compromises must be made which inevitably mean that  $f_{\text{max}} > B$ . Fortunately, we now live in the digital age and rarely need to worry about such problems any more!

## 6 Random variables

### 6.1 Discrete and continuous random variables

A random variable  $X$  is a function which assigns numbers to the possible outcomes of an experiment. These numbers are chosen expediently by the experimenter.

- Examples:
- (a) Throw a die:  $X(\text{outcome is even}) = 1$   
 $X(\text{outcome is odd}) = 0$
  - (b) Throw a die:  $X(\text{die has } k \text{ dots showing}) = k.$
  - (c) Measure a voltage:  $X(\text{voltage is } V \text{ volts}) = V$

In example (a) and (b) the random variable is **discrete**. Associated with a discrete random variable  $X$  are the probabilities  $P_X(x_i)$  that  $X$  takes each possible value  $x_i$ . In example (a),  $P_X(1) = 1/2$  and  $P_X(0) = 1/2$ . From the definition of probability we must have (summing over all possible  $x_i$ ),

$$\sum_{x_i} P_X(x_i) = 1. \quad (6.1)$$

In example (c) the random variable is **continuous**. Associated with a continuous random variable  $X$  is a *probability density*  $p_X(x)$  such that  $p_X(x)dx$  is the probability that  $X$  lies between  $x$  and  $x+dx$ . Now from the definition of probability,

$$\int_{-\infty}^{\infty} p_X(x)dx = 1. \quad (6.2)$$

Note that an absolute probability, as for a discrete random variable, is written in upper case –  $P_X(x_i)$ , while a probability density, as for a continuous random variable, is written in lower case –  $p_X(x)$ .

### 6.2 Multivariate probabilities

Multivariate probabilities are the extensions of the above to deal with multiple variables simultaneously.

For a discrete random variable, the bivariate (joint) probability  $P_{X_1, X_2}(x_1, x_2)$  is the probability that  $X_1 = x_1$  and  $X_2 = x_2$  in the same measurement. Bivariate probabilities must by definition obey the following:

$$\sum_{x_1} \sum_{x_2} P_{X_1, X_2}(x_1, x_2) = 1, \quad (6.3)$$

$$\sum_{x_1} P_{X_1, X_2}(x_1, x_2) = P_{X_2}(x_2), \quad (6.4)$$

and 
$$\sum_{x_2} P_{X_1, X_2}(x_1, x_2) = P_{X_1}(x_1). \quad (6.5)$$

Note: we don't usually bother to write commas between the  $x_1$  and  $x_2$  – but it doesn't mean they're multiplied! Also note that there are many alternative notation systems! For example,  $P_{X_1, X_2}(x_1, x_2)$  is sometimes written  $P(A, B)$ . We will stick to one self-consistent notation scheme throughout this course. If you don't like this one, please blame Dr R. Pettifer, who created this course. Or blame me for not writing  $x$ 's and  $X$ 's differently enough.

For a continuous random variable the sums become integrals:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p_{X_1, X_2}(x_1, x_2) dx_1 dx_2 = 1 \quad (6.6)$$

$$\int_{-\infty}^{\infty} p_{X_1, X_2}(x_1, x_2) dx_1 = p_{X_2}(x_2) \quad (6.7)$$

$$\int_{-\infty}^{\infty} p_{X_1, X_2}(x_1, x_2) dx_2 = p_{X_1}(x_1) \quad (6.8)$$

### 6.3 Combining probabilities:

#### 6.3.1 Statistically independent case

If  $X_1$  and  $X_2$  are statistically independent, their probabilities just multiply:

$$P_{X_1 X_2}(x_1 x_2) = P_{X_1}(x_1)P_{X_2}(x_2) \quad (6.9)$$

$$P_{X_1 X_2 X_3}(x_1 x_2 x_3) = P_{X_1}(x_1)P_{X_2}(x_2)P_{X_3}(x_3)$$

and so on. In the continuous case it's exactly the same, e.g.,

$$p_{X_1 X_2}(x_1 x_2) = p_{X_1}(x_1)p_{X_2}(x_2). \quad (6.10)$$

Here's a very simple example: Flip two coins independently at the same time. We define one random variable associated with each coin, and list the possible values for each measurement and their probabilities:

coin 1, $X_1$	$X_1(\text{coin 1 comes up heads}) = 1$	$P_{X_1}(1) = 1/2$
	$X_1(\text{coin 1 comes up tails}) = 0$	$P_{X_1}(0) = 1/2$
coin 2, $X_2$	$X_2(\text{coin 2 comes up heads}) = 1$	$P_{X_2}(1) = 1/2$
	$X_2(\text{coin 2 comes up tails}) = 0$	$P_{X_2}(0) = 1/2$

The probability that both come up heads =  $P_{X_1 X_2}(11) = P_{X_1}(1)P_{X_2}(1) = (1/2)(1/2) = 1/4$ .

#### 6.3.2 Statistically dependent case

If  $X_1$  and  $X_2$  are statistically dependent, we need to introduce conditional probabilities. The conditional probability  $P_{X_1|X_2}(x_1 | x_2)$  is the probability that  $X_1 = x_1$  given  $X_2 = x_2$ . Conditional and absolute probabilities are related in the following way: for discrete random variables,

$$P_{X_1 X_2}(x_1 x_2) = P_{X_1|X_2}(x_1 | x_2)P_{X_2}(x_2) = P_{X_2|X_1}(x_2 | x_1)P_{X_1}(x_1), \quad (6.11)$$

or for continuous random variables, the same thing with small  $p$ 's:

$$p_{X_1 X_2}(x_1 x_2) = p_{X_1|X_2}(x_1 | x_2)p_{X_2}(x_2) = p_{X_2|X_1}(x_2 | x_1)p_{X_1}(x_1). \quad (6.12)$$

For independent variables,  $p_{X_1|X_2}(x_1 | x_2) = p_{X_1}(x_1)$ .

### 6.4 Expectation

Expectation can be written as an operator  $E$ , which means take the average value of the thing it's operating on – usually a function of a random variable.

Discrete case: 
$$E\{f(X)\} = \sum_{x_i} f(x_i)P_X(x_i) \quad (6.13)$$

Continuous case: 
$$E\{f(X)\} = \int_{-\infty}^{\infty} f(x)p_X(x)dx \quad (6.14)$$

The expectation operator always has the following properties:

$$E\{f_1(X) + f_2(X)\} = E\{f_1(X)\} + E\{f_2(X)\}, \quad (6.15)$$

$$E\{cf(X)\} = cE\{f(X)\}. \quad (6.16)$$

For statistically independent variables only, there is another result:

$$E\{f_1(X_1)f_2(X_2)\} = E\{f_1(X_1)\}E\{f_2(X_2)\}. \quad (6.17)$$

This is because  $E\{f_1(X_1)f_2(X_2)\} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_1(x_1)f_2(x_2)p_{X_1X_2}(x_1x_2)dx_1dx_2$  (taking the continuous case),

and using (6.12) this is equal to  $\int_{-\infty}^{\infty} f_1(x_1)p_{X_1}(x_1)dx_1 \int_{-\infty}^{\infty} f_2(x_2)p_{X_2}(x_2)dx_2 = E\{f_1(X_1)\}E\{f_2(X_2)\}$ .

### 6.5 Simple example: two logic levels

Consider a trivial example: a two-level logic signal is conveyed by two voltage levels  $V = 0$  and  $-8$  V on a wire, and the 0 V level is 3 times more likely than the 0 V level. What's the average voltage?

First, define a discrete random variable which has two levels corresponding to the two logic states:

$$\begin{aligned} X(\text{logic level } 0) &= x_0 = 0 & P_X(0) &= 0.75 \\ X(\text{logic level } 1) &= x_1 = 1 & P_X(1) &= 0.25 \end{aligned}$$

Note: we choose  $x_0 = 0, x_1 = 1$  to illustrate that  $X$  doesn't have to equal the voltage.

Next, define a function  $f(X)$  which maps the logic state  $X$  to a voltage level:  $f(0) = 0$  V and  $f(1) = -8$  V.

$$\begin{aligned} \text{Then the average voltage is } E\{f(x)\} &= \sum_{x_i} f(x_i)P_X(x_i) \text{ using Eq. (6.13)} \\ &= f(0)P_X(0) + f(1)P_X(1) \\ &= (0 \text{ V}) \times 0.75 + (-8 \text{ V}) \times 0.25 = -2 \text{ V} . \end{aligned}$$

### 6.6 Mean value

For a continuous random variable, 
$$\bar{X} = E\{X\} = \int_{-\infty}^{\infty} xp_X(x)dx \tag{6.18}$$

Similarly for a discrete one, 
$$\bar{X} = E\{X\} = \sum_{x_i} x_iP_X(x_i) .$$

### 6.7 Mean squared value

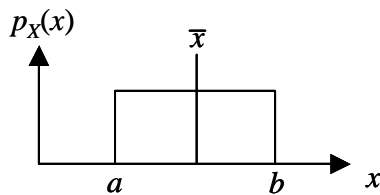
$$\overline{X^2} = E\{X^2\} = \int_{-\infty}^{\infty} x^2p_X(x)dx . \tag{6.19}$$

### 6.8 Variance

$$\begin{aligned} s_X^2 &= E\{(X - \bar{X})^2\} = \int_{-\infty}^{\infty} (x - \bar{X})^2 p_X(x)dx \\ &= E\{X^2 - 2X\bar{X} + \bar{X}^2\} = E\{X^2\} - 2\bar{X}E\{X\} + \bar{X}^2 \\ &= E\{X^2\} - E\{X\}^2 \\ &= \overline{X^2} - \bar{X}^2 \end{aligned} \tag{6.20}$$

### 6.9 Example: uniform probability density.

$p_X(x) = \frac{1}{b-a}$  for  $a < x < b$  and  $p_X(x) = 0$  otherwise.



$$\begin{aligned} \bar{x} &= \int_a^b x \frac{1}{b-a} dx = \left[ \frac{x^2}{2(b-a)} \right]_a^b = \frac{a+b}{2} \\ \overline{x^2} &= \int_a^b x^2 \frac{1}{b-a} dx = \left[ \frac{x^3}{3(b-a)} \right]_a^b = \frac{a^2 + ab - b^2}{3} \\ s_x^2 &= \overline{x^2} - \bar{x}^2 = \frac{(a-b)^2}{12} \end{aligned}$$

## 6.10 Case of electrical noise

An electrical noise signal might be represented by  $X$  = voltage or current (see Section 2.5). In that case,

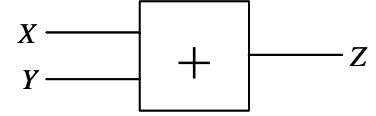
$$\overline{X}^2 = \text{mean normalized power in signal}$$

$$\overline{X}^2 = \text{dc normalized power}$$

$$\sigma_X^2 = \text{noise power in signal}$$

## 7 Adding random variables

Often we are interested in the sum of two or more random variables, eg  $Z = X + Y$ . For instance, several electrical noise signals might be added together to contribute to the total noise in a circuit.



### 7.1 Adding means:

$$\overline{Z} = E\{Z\} = E\{X + Y\} = E\{X\} + E\{Y\} = \overline{X} + \overline{Y} \quad (7.1)$$

Thus dc power adds, as it surely should.

### 7.2 Adding variance (noise power)

$$\begin{aligned} \mathbf{s}_Z^2 &= E\{(Z - \overline{Z})^2\} = E\{(X + Y - \overline{X} - \overline{Y})^2\} = E\{(X - \overline{X} + Y - \overline{Y})^2\} \\ &= E\{(X - \overline{X})^2 + 2(X - \overline{X})(Y - \overline{Y}) + (Y - \overline{Y})^2\} \\ &= \mathbf{s}_X^2 + 2E\{(X - \overline{X})(Y - \overline{Y})\} + \mathbf{s}_Y^2 \end{aligned} \quad (7.2)$$

If  $X$  and  $Y$  are *statistically independent*, Eq. (6.17) tells us that  $E\{(X - \overline{X})(Y - \overline{Y})\} = E\{X - \overline{X}\}E\{Y - \overline{Y}\}$  which is zero, and hence the middle term in Eq. (6.22) vanishes and the variances add:

$$\mathbf{s}_Z^2 = \mathbf{s}_X^2 + \mathbf{s}_Y^2 \quad (7.3)$$

Hence if the outputs of independent noise sources are added, their noise powers add too. One can repeat the process for many noise sources, adding them in one at a time, so that

$$\mathbf{s}_{total}^2 = \sum_{\text{all sources}} \mathbf{s}_i^2 \quad (7.4)$$

Note however that if there is any correlation between  $X$  and  $Y$  then the cross-term in Eq. (6.22) becomes important.

### 7.3 Adding probability density

We want the probability density  $p_Z(z)$  for the sum  $Z$  in terms of the densities  $p_X(x)$  and  $p_Y(y)$  of  $X$  and  $Y$ . For statistically independent variables the result is a convolution:

$$p_Z(z) = \int_{-\infty}^{\infty} p_X(x)p_Y(z-x)dx = p_X(x) \otimes p_Y(y) \quad (7.5)$$

The justification of this is a little awkward, involving differentiating the cumulative distribution function, but roughly it means that for a given  $z$  you can choose any  $x$  as long as  $y = z - x$ , so you have to integrate probability over all possible  $x$  values.

The good old convolution theorem then gives:

$$F.T.\{p_Z(z)\} = F.T.\{p_X(x)\}F.T.\{p_Y(y)\} \quad (7.6)$$

### 7.4 Characteristic functions

The Fourier transform of the probability density,  $F.T.\{p_X(x)\}$ , is called the characteristic function of the variable  $X$ . It is useful because of Eq. (6.27) – when you add independent variables their characteristic functions multiply.

### 7.5 Example: Gaussian distributions

The most common probability distribution in the world is the Gaussian (or *normal*) distribution:

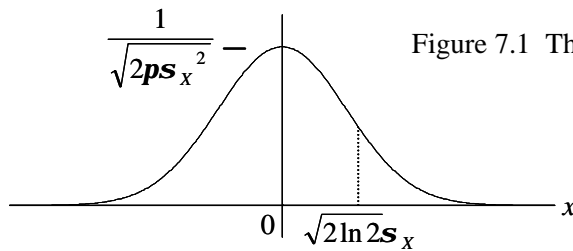


Figure 7.1 The Gaussian distribution.

$$p_X(x) = \frac{1}{\sqrt{2ps_x^2}} \exp\left(-\frac{x^2}{2s_x^2}\right), \quad (7.7)$$

$$\text{half-width half-max} = \sqrt{2 \ln 2} s_x \quad (7.8)$$

The characteristic function of a Gaussian distributed variable is

$$\begin{aligned} F.T.\{p_X(x)\} &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2ps_x^2}} \exp\left(-\frac{x^2}{2s_x^2}\right) \exp(-2pifx) dx \\ &= \exp(-2p^2s^2f^2) \end{aligned} \quad (7.9)$$

Notes:  $f \leftrightarrow 1/x$  is conjugate to  $x$ . We used a standard integral  $\int_{-\infty}^{\infty} \exp(-a^2x^2) \cos(bx) dx = \sqrt{\pi}/a \exp(-b^2/4a^2)$ . The characteristic function of a Gaussian is another Gaussian. Finally, although we have assumed a distribution centered at  $x = 0$ , shifting the origin makes only cosmetic differences to all the results.

### 7.6 Adding Gaussian variables

Assume both  $X$  and  $Y$  are Gaussian distributed. Thus from Eq. (7.6),

$$\begin{aligned} F.T.\{p_Z(z)\} &= F.T.\{p_Y(y)\} F.T.\{p_X(x)\} \\ &= \exp(-2p^2s_Y^2f^2) \exp(-2p^2s_X^2f^2) \\ &= \exp(-2p^2(s_X^2 + s_Y^2)f^2) \end{aligned}$$

Inverting this,

$$p_z(z) = \frac{1}{\sqrt{2p(s_X^2 + s_Y^2)}} \exp\left(-\frac{z^2}{2(s_X^2 + s_Y^2)}\right) \quad (7.10)$$

Thus  $Z$ , the sum of  $X$  and  $Y$ , is also Gaussian distributed. It is also apparent that  $s_Z^2 = s_X^2 + s_Y^2$ , as expected from Eq. (7.3).

### 7.7 The central limit theorem

Statement: given a set of random variables  $X_1, X_2, \dots, X_n$ , having the same *arbitrary* probability density  $p_{X_1}(x) = p_{X_2}(x) = \dots = p_{X_n}(x)$ , their sum  $Y = \sum_i X_i$  has a *Gaussian* probability density as  $n \rightarrow \infty$ .

Unfortunately this is very hard to prove, but it is important so we need to remember it. It is the underlying reason why Gaussian distributions are so common, as we will see in the case of electrical noise.

Example: take the top-hat probability distribution with  $p_{X_i}(x_i) = 1$  for  $-1/2 < x < 1/2$  and 0 otherwise. Each time a variable is added the original  $p_X(x)$  is convolved in once more. The result quickly approaches a Gaussian:

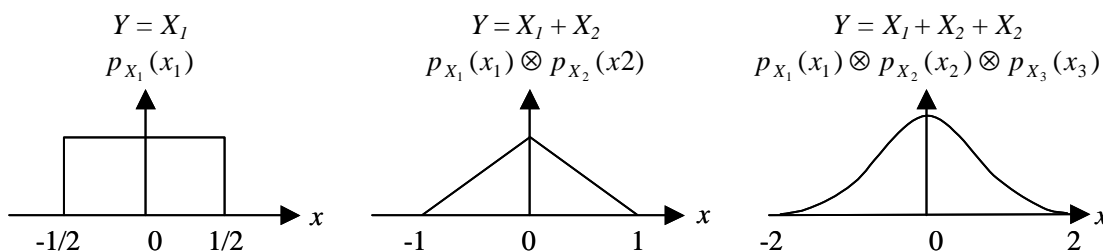


Figure 7.2

## 8 Noise.

### 8.1 Properties of stationary processes (in other words, noise)

A stationary process is one whose statistical properties are invariant in time (as in Section 3.4):



Figure 7.3.

An evolutionary process is not, and its properties vary in time:

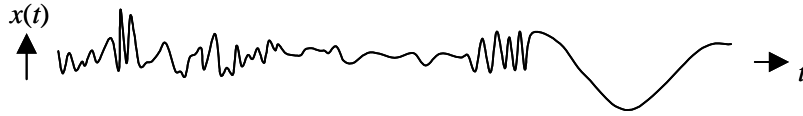


Figure 7.4.

The boundary between these two cases is not sharp in the real world. Evolutionary processes are very complex, so we only deal here with processes which are effectively stationary on the longest timescale in the measurement. Note that  $x(t)$  is usually taken as real for such processes.

### 8.2 The autocorrelation function

This measures the degree of correlation between a signal and a time shifted version of itself. It can be defined in two ways:

(a) By time averaging: 
$$R(\mathbf{t}) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t)x(t+\mathbf{t})dt \quad (8.1)$$

(b) By ensemble averaging: 
$$R(\mathbf{t}) = E\{X_t X_{t+\mathbf{t}}\} \quad (8.2)$$

In the ensemble average (it's a bit subtle, and comes from statistical mechanics), you average over all possible forms that the signal could take which are consistent with what you know about its statistical properties. The meaning of the variables is as follows. You pick a time  $t$  at which to evaluate the expression. Then you define  $X_t$  to be the random variable corresponding to measurement at that particular time  $t$ , and  $X_{t+\tau}$  to be another random variable corresponding to measurement at a later time  $t+\tau$ . You then average over all possible values of  $X_t$  and  $X_{t+\tau}$ , with the averaging done over all possible processes, ie all members of the ensemble. If you don't get why it's called an ensemble, don't worry about it until you do statistical mechanics. Note that the choice of  $t$  is irrelevant on the right-hand side because all the probability measures involved in the calculation of the expectation must be independent of  $t$  for a stationary process.

In the case that  $x(t)$  has only discrete levels  $x_i$  between which it jumps as a function of time, we can rewrite Eq. (8.2) as follows:

$$R(\mathbf{t}) = \sum_{x_i} \sum_{x_j} x_i x_j P_{X_t X_{t+\mathbf{t}}}(x_i x_j) \quad (8.3)$$

Here the first sum is over all possible values of  $X_t$  and the second sum is the same for for  $X_{t+\mathbf{t}}$ . The possible values are of courses the same for both. Consistently with our earlier terminology,  $P_{X_t} P_{X_{t+\mathbf{t}}}(x_i x_j)$  is the joint probability that  $X_t = x_i$  and  $X_{t+\mathbf{t}} = x_j$ . According to the discussion above,  $P_{X_t} P_{X_{t+\mathbf{t}}}(x_i x_j)$  must be independent of  $t$ .

If ensemble and time averaging always give the same result, the process is said to be **ergodic**. Most cases of interest to us obey ergodicity.

### 8.3 Properties of the autocorrelation function

(i)  $R(\mathbf{t}) = R(-\mathbf{t})$  – it's a real symmetric function. (8.4)

(ii)  $R(\tau)$  has a maximum at  $\tau = 0$ .

The width of  $R(t)$  reflects is a measure of the fastest timescale of the fluctuations in  $x(t)$ :

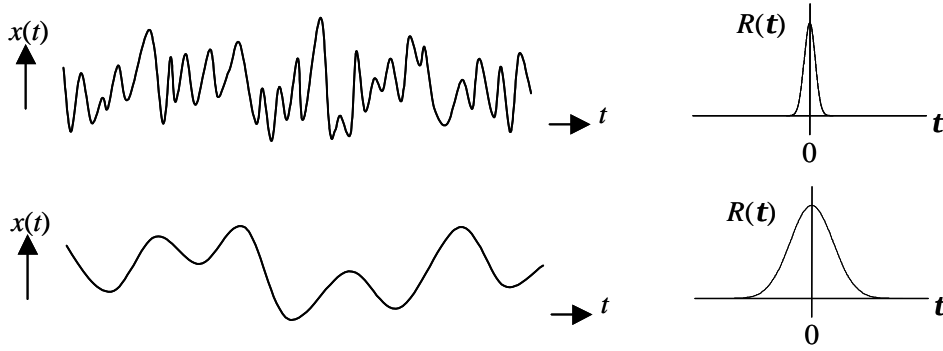


Figure 7.5. Rapidly (top) and slowly (bottom) fluctuating signals, and their respective autocorrelation functions.

(iii)  $R(0) = \text{normalised power in the process (signal)} = \overline{X^2}$  (8.5)

(iv) The Fourier transform of  $R(t)$  is the power spectrum (see below.)

### 8.4 Power spectrum of a stationary process

From section 3.1, the Fourier spectrum of  $x(t)$  is  $G(f) = \int_{-\infty}^{\infty} x(t) \exp(-2\pi i f t) dt$ .

Note that here the limit  $T \rightarrow \infty$  has already been taken in defining the F.T. By reasoning analogous to that in section 2.5, the total normalised power in the process is

$$\text{power } P = \overline{X^2} = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} [x(t)]^2 dt$$

(substitute for  $x(t)$  using Eq. 3.1) 
$$= \int_{-\infty}^{\infty} \lim_{T \rightarrow \infty} \frac{|G(f)|^2}{T} df .$$
 (8.6)

If we define 
$$S(f) = \lim_{T \rightarrow \infty} \frac{|G(f)|^2}{T},$$
 (8.7)

the power becomes 
$$P = \int_{-\infty}^{\infty} S(f) df .$$
 (8.8)

From Eq. (8.8) we deduce that  $S(f)$  must be the power spectrum of the process, ie  $S(f)df = \text{noise power in the frequency range } f \text{ to } f+df$ .

Since  $G^*(f) = G(-f)$  from Eq. (3.3), we find that  $S(f)$  is a real, symmetric function.

Note:  $S(f)$  contains no phase information. Thus the power spectrum of a signal is not sufficient to reconstruct it, because the phase information in  $x(t)$  has been discarded. In other words, different signals can have the same power spectrum.

### 8.5 The Wiener-Khintchine theorem

This says that the power spectrum is the Fourier transform of the autocorrelation function:

$$F.T.\{R(t)\} = \int_{-\infty}^{\infty} \left[ \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t)x(t+t) dt \right] \exp(-2\pi i f t) dt$$

Reverse the integration order, 
$$= \lim_{T \rightarrow \infty} \frac{1}{T} \int_{t=-T/2}^{T/2} \int_{t=-\infty}^{\infty} x(t)x(t+t) \exp(-2\pi i f t) dt dt$$

put  $t = s - t$  in the  $t$  integral, 
$$= \lim_{T \rightarrow \infty} \frac{1}{T} \int_{t=-T/2}^{T/2} \int_{t=-\infty}^{\infty} x(t)x(s) \exp[-2\pi i f (s-t)] ds dt$$

rearrange, 
$$= \lim_{T \rightarrow \infty} \frac{1}{T} \int_{t=-T/2}^{T/2} x(t) \exp(2\pi i f t) dt \int_{t=-\infty}^{\infty} x(s) \exp(-2\pi i f s) ds$$

and be a bit sloppy (sorry maths guys),  $= \lim_{T \rightarrow \infty} \frac{1}{T} G^*(f)G(f)$ ,

and finally we get the eponymous theorem:

$$F.T.\{R(\mathbf{t})\} = S(f). \quad (8.9)$$

Both  $R(\mathbf{t})$  and  $S(f)$  are real symmetric functions, so the forward and inverse F.T.'s are identical, and

$$F.T.\{S(f)\} = R(\mathbf{t}). \quad (8.10)$$

Since they form a Fourier transform pair, they contain exactly the same information. This important theorem says that we can find the power spectrum of a process if we know its autocorrelation function. We will see examples of its use in section 9, in finding the power spectra of different types of noise.

### 8.6 Corollary: power transmission through a linear system

For a linear system (section 4.9) such as a filter, you can show that the power spectra of input  $X$  and output  $Y$  are simply related by the modulus squared of the transfer function  $H(f)$ :

$$S_Y(f) = S_X(f) |H(f)|^2. \quad (8.11)$$

### 8.7 White noise

This is defined as a process with a flat power spectrum,  $S(f) = C$  (a constant), up to some high cutoff frequency  $f_{max}$ . Much real noise is approximately white. Any real signal has a finite  $f_{max}$ , because  $f_{max} = \infty$  implies from Eq. (8.8) infinite power  $P$ , which is nonsensical. Still, for the most common noise processes  $f_{max}$  is very high, and we can usually treat it as infinite. Then from Eqs. (8.10) and (3.12),

$$R(\mathbf{t}) = F.T.\{C\} = C\mathbf{d}(\mathbf{t}). \quad (8.12)$$

Hence for white noise there is no correlation between points at nearby times – if you sample it, every point will be randomly picked according to the distribution function.

### 8.8 Filtering white noise

Consider putting white noise through a filter (a linear system) with a top-hat function for  $H(f)$  and therefore also for its power transfer,  $|H(f)|^2 = 1$  for  $|f| < B$  and  $|H(f)|^2 = 0$  otherwise. From Eq. (8.11), the power spectrum of the output is  $S_Y(f) = C|H(f)|^2$ . The total normalized power in the output is therefore

$$P = \int_{-\infty}^{\infty} S_Y(f) df = \int_{-B}^B C df = 2BC, \quad (8.13)$$

and its autocorrelation function is

$$R_Y(\mathbf{t}) = F.T.\{S_Y(f)\} = F.T.\{C|H(f)|^2\} = 2BC \text{sinc}(\mathbf{p} B \mathbf{t}). \quad (8.14)$$

Thus the autocorrelation function of filtered white noise has a finite width of about  $B^{-1}$ . This means that the filter has smoothed the noise process, whose fastest wiggles now have a typical timescale of  $B^{-1}$ , and that values at nearby times have become correlated.

### 8.9 Signal-to-noise ratio

Any analog signal carrying useful information has noise superimposed on it. Later, information theory will tell us how much information can be carried by a signal in the presence of noise. The situation is quantified by the signal-to-noise ratio, known as ‘‘S/N’’ and defined for the convenience of engineers as

$$\text{“S/N” (in dB)} = 10 \log_{10} \left( \frac{\text{Total signal power}}{\text{Total noise power}} \right). \quad (8.15)$$

If the signal has a bandwidth  $B$  and the noise has a bandwidth  $f_{max} > B$ , then S/N can be improved simply by filtering the total, signal+noise, to remove all components with  $f > B$ , without harming the signal.

## 9 Types of electrical noise

Electrical noise (excluding pickup from external circuits) results from the random motion of electrons in circuit elements. The measured current and voltage are both the sum of contributions from all moving electrons in the circuit. Noise comes in a variety of forms, the most important of which are (1) shot noise, (2) thermal noise, and (3) flicker noise. Each of these noise types has its own characteristic dependence on current, frequency and temperature. Thermal noise is the most universal, being present in anything with a resistance, while shot noise and flicker noise only occur under certain circumstances. Shot and thermal noise are white, while flicker noise is ‘pink’. In large devices, all these noise types are found to have Gaussian distributions, consistent with the central limit theorem.

### 9.1 Shot noise

This is white noise in the *current* resulting from the *granularity of charge* (i.e., charge comes only in multiples of the charge  $e$  on an electron). It is thus very fundamental, but only shows up in circuit components in which current is carried by electrons having to cross a potential barrier, such as in going through the base in a transistor or across the vacuum between cathode and anode in a thermionic valve. In such devices, each time an electron traverses the barrier there is a pulse of current. On average all the pulses add up to give a dc current, but the pulse-like nature also gives rise to noise. If you fed the signal to a loudspeaker you’d hear the noise of popcorn popping, or an audience clapping.

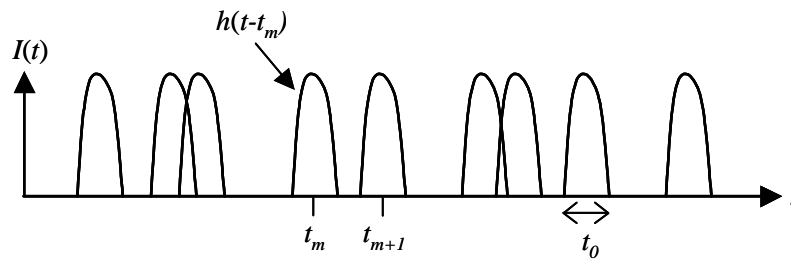


Figure 9.1.  
Decomposition  
of current into  
pulses.

We can work out its power spectrum by deducing the autocorrelation function and using the Wiener-Khinchine theorem. The total current  $I(t)$  is built of pulses with shape  $h(t)$  of width  $\sim t_0$ . Each pulse results from an individual electron passing through the device. The area under one pulse is one electron charge, so after normalizing  $h(t)$ ,

$$\int_{-\infty}^{\infty} h(t) dt = 1, \quad (9.1)$$

the current is

$$I(t) = \sum_m e h(t - t_m), \quad (9.2)$$

where  $t_m$  is the arrival time of the  $m$ 'th pulse. First, let's assume that  $h(t) = d(t)$ , so

$$I(t) = \sum_m e d(t - t_m). \quad (9.3)$$

According to Eq. (8.1) the autocorrelation function is then

$$\begin{aligned} R(\mathbf{t}) &= \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} I(t) I(t + \mathbf{t}) dt \\ &= e^2 \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} \left[ \sum_m d(t - t_m) \right] \left[ \sum_n d(t - t_n + \mathbf{t}) \right] dt \end{aligned}$$

$$\begin{aligned}
&= e^2 \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} \sum_{m,n} \mathbf{d}(t-t_m) \mathbf{d}(t-t_n + t) dt \\
&= e^2 \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{m,n} \mathbf{d}(t_m - t_n + t) . \\
&= e^2 \lim_{T \rightarrow \infty} \frac{1}{T} \left[ \sum_m \mathbf{d}(t) + \sum_{m,n \neq m} \mathbf{d}(t_m - t_n + t) \right] . \tag{9.4}
\end{aligned}$$

Picking out the case  $n=m$ ,

If there are  $N = \bar{I}T/e$  pulses in time  $T$ , the first term becomes  $e^2 \lim_{T \rightarrow \infty} \frac{N}{T} \mathbf{d}(t) = e\bar{I}\mathbf{d}(t)$ .

The second term is a “grass” of  $N^2$   $\mathbf{d}$ -functions of height  $e^2/T$ . As  $T \rightarrow \infty$ , as long as there are no correlations between the different  $t_m$ , this term becomes a flat background of height  $N^2 e^2/T^2 = \bar{I}^2$ . Hence we get

$$R(t) = e\bar{I}\mathbf{d}(t) + \bar{I}^2 . \tag{9.5}$$

By the Wiener-Khintchine theorem (Eq. 8.9), the spectrum is therefore

$$\begin{aligned}
S_I(f) &= \int_{-\infty}^{\infty} [e\bar{I}\mathbf{d}(t) + \bar{I}^2] \exp(-2\pi i f t) dt \\
&= e\bar{I} + \bar{I}^2 \mathbf{d}(f) . \tag{9.6}
\end{aligned}$$

The term  $\bar{I}^2 \mathbf{d}(f)$  is the d.c. power, and the term  $e\bar{I}$  is the (frequency-independent) shot noise. Most often the shot noise is quoted using a positive-frequency only convention, in which case the result (ignoring the dc term) is

$$S_I(f)_{f>0 \text{ only}} = 2e\bar{I} . \tag{9.7}$$

Finally, to deal with the fact that real pulses have finite width, we return to Eq. (9.2), and note that it is exactly what you’d get if you passed the series of  $\mathbf{d}$ -functions of Eq. (9.3) through a filter with impulse response function  $h(t)$  (see sections 4.2-4.5). Thus according to Eq. (8.11), the power spectrum with finite-width pulses is

$$\begin{aligned}
S_I(f) &= |H(f)|^2 [e\bar{I} + \bar{I}^2 \mathbf{d}(f)] \\
&= |H(f)|^2 e\bar{I} + \bar{I}^2 \mathbf{d}(f) , \tag{9.8}
\end{aligned}$$

where as usual  $H(f) = F.T.\{h(f)\}$ , and we used the fact that  $H(0) = 1$ . The real spectrum is thus only flat when  $H(f)$  is flat, ie for  $f < f_{\max} \approx 1/t_0$ . Here  $t_0$  is the width of  $h(f)$ , which is roughly the transit time of an electron through the barrier. This is typically  $10^{-10}$  s, so  $f_{\max} \sim 10^{10}$  Hz.

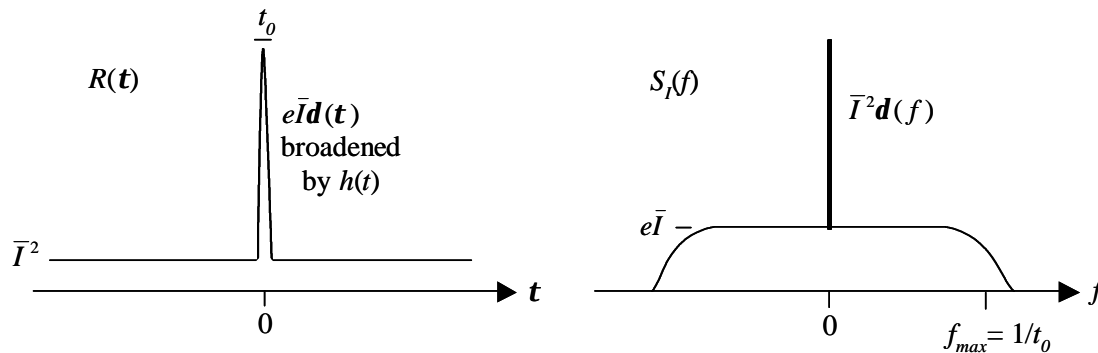


Figure 9.2. Autocorrelation function and spectrum of current with shot noise.

If there are on average many electrons crossing the barrier at once, so at any one time there are many overlapping pulses superimposed to make  $I(t)$ , then  $I(t)$  is Gaussian by the central limit theorem again.

## 9.2 Thermal noise (also known as Nyquist or Johnson noise)

This is white noise in the *current or voltage* resulting from the *thermal motion* of the electrons. It is even more fundamental than shot noise, in the sense that it would be there whatever the nature of the charge was. It is related to black body radiation, which involves the fluctuations of the electromagnetic field in thermal equilibrium. Its magnitude was first derived by Nyquist, in the following way.

We construct a thought experiment in order to work out the voltage associated with the electromagnetic radiation emitted and absorbed by a resistor  $R$  in thermal equilibrium. We take a single-mode transmission line of length  $L$  and impedance  $R$ , terminated at each end with a matching resistors  $R$  (see Figure 9.3). In this situation, all the noise power emitted by one resistor due to thermal fluctuations travels without reflection along the transmission line to be absorbed by the other resistor. We will find the power traveling along the line in thermal equilibrium and equate it to the power emitted by the resistor. In equilibrium this must equal the power absorbed, which depends on the resistance.

We can find the energy stored in the line by suddenly closing switches at each end to short out the resistors. After that, the short-terminated transmission line has allowed modes of frequency,  $f_n = nc/(2L)$ , spaced by  $df = c/(2L)$ , where  $n$  is an integer and  $c$  is the wave velocity. By the equipartition theorem there's an average energy  $k_B T$  in each mode, and the energy stored in a frequency band  $Df$  is thus

$$E = \frac{\Delta f}{df} k_B T = \frac{2L}{c} \Delta f . k_B T .$$

This energy must have been emitted by the two resistors during a time  $t = L/c$  before the switches were closed. Hence the power emitted by each resistor during that time was

$$P = \frac{E}{2t} = \left( \frac{2L}{c} \Delta f . k_B T \right) / \frac{2L}{c} ,$$

so

$$P = \Delta f . k_B T . \quad (9.9)$$

Notice this simple result: (power emitted)/(bandwidth) =  $k_B T$ .

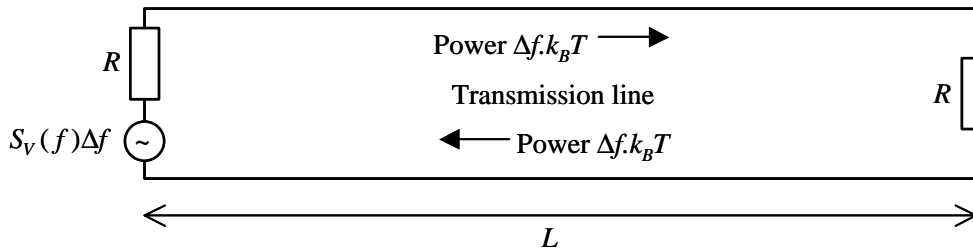


Figure 9.3. Configuration for Nyquist's derivation of thermal noise.

The resistor on the right absorbs all the incident power on it,  $\Delta f . k_B T$ . If that power is considered to come from a noise source in series with the resistor on the left, generating a voltage power spectrum  $S_V(f)$  as shown, the delivered power is  $S_V(f) \Delta f / 4R$ . Thus

$$\frac{S_V(f) \Delta f}{4R} = \Delta f . k_B T$$

and so the final result is that thermal noise has a white power spectrum whose magnitude depends only on the resistance and the temperature:

$$S_V(f) = 4 . k_B T R . \quad (9.10)$$

Note that it is completely independent of the nature of the resistor!

The cutoff frequency  $f_{\max}$  for thermal noise is determined by quantum mechanics. Modes of the transmission line with an energy quantum  $hf$  greater than  $k_B T$ , are 'frozen out' and so there are no thermal fluctuations if

$$f > f_{\max} \approx k_B T / h . \quad (9.11)$$

Since the thermal noise is the sum of the voltages induced by many electrons fluctuating, the central limit theorem predicts that it has a Gaussian distribution.

### 9.3 Flicker (1/f) noise

This consists of fluctuations of the *resistance* of a circuit element, and is *not* white! It leads to voltage fluctuations called flicker noise if a constant current is applied. It's not as fundamental as shot and thermal noise and has many different origins, depending on what factors determine the electrical resistance (and there are many possible.) The most common cause however is impurities trapping and releasing electrons: the resistance is different when there is an electron trapped from when there isn't. Flicker noise is not always present, and it is particularly large in devices which rely on action at surfaces and interfaces, such as field-effect transistors. This is because (a) impurities tend to occur at interfaces, and (b) in the operation of such devices the current tends to be forced near the impurities at the interface.

Now we must treat the resistance of the device,  $R(t)$ , as the stationary process. When the device is biased with a constant current  $I$ , the voltage across it is  $V(t) = IR(t)$ . Thus  $\overline{V^2} = I^2 \overline{R^2}$ , and so the voltage noise power spectrum is directly related to the power spectrum of  $R(t)$ :

$$S_V(f) = I^2 S_R(f). \quad (9.12)$$

This is a general result for flicker noise. Additionally, it is very often found experimentally in standard electronic devices that  $S_R(f) = A/f^a$ , where  $A$  is approximately independent of  $I$  and  $f$  (though it may be temperature dependent), and  $a$  is close to, though not exactly unity. This rule is obeyed over many orders of magnitude in  $f$  from kHz down to  $\mu$ Hz. The result is a typical "1/f" behaviour:

$$S_V(f) = \frac{AI^2}{f^a}. \quad (9.13)$$

Remember that the parameters  $A$  and  $a$  are system-dependent and not universal. However, in a large device there are usually many independent processes contributing to  $R(f)$ , and as usual by the central limit theorem their sum has a Gaussian distribution, so  $S_V(f)$  is usually Gaussian.

### 9.4 Flicker noise in small devices: random telegraph signals

Modern field-effect transistors are so small that there are only a few impurities in each one. A single impurity gives rise to a noise signal called a "random telegraph signal" which looks like this:

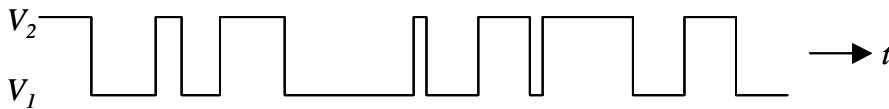


Figure 9.4.  
A "random telegraph signal".

It switches between two discrete levels at random points in time – one level corresponds to the the impurity holding a trapped electron, the other to it being empty. In large devices, the total noise is the sum of many of such signals superimposed. The distribution of their parameters results in the  $1/f$  dependence, and the (slightly generalized) central limit theorem tells us that their sum has a Gaussian distribution.

As with shot noise, we calculate the power spectrum of the noise process by finding the autocorrelation function and using the Wiener Khintchine theorem. This time we start with Eq. (8.3) for the autocorrelation function of a process with a discrete distribution:

$$R(t) = \sum_{x_i} \sum_{x_j} x_i x_j P_{X_i X_{i+t}}(x_i x_j).$$

This process has only two levels, which we are free to choose to have the computationally simple values  $x_0 = V_1 = 0$  and  $x_1 = V_2 - V_1 = a$ . We can then write out the double sum ( $i$  and  $j$  are either 0 or 1):

$$\begin{aligned} R(t) &= 0.0.P_{X_i X_{i+t}}(00) + 0.a.P_{X_i X_{i+t}}(0a) + a.0.P_{X_i X_{i+t}}(a0) + a.a.P_{X_i X_{i+t}}(aa) \\ &= a^2 P_{X_i X_{i+t}}(aa), \\ &= a^2 P_{X_i}(a) P_{X_{i+t}|X_i}(a|a). \end{aligned} \quad (9.14)$$

and using Eq. (6.11),

We will tackle here only the simple case where the probabilities of higher and lower levels are equal, so that  $P_{X_t}(a) = P_{X_t}(0) = 1/2$ . It remains then to calculate the  $P_{X_{t+\tau}|X_t}(a|a)$ , which is the (conditional) probability of finding  $x(t+\tau) = a$  given that  $x(t) = a$ . A little thought shows that this is equal to the probability of there being an even number of transitions in time  $|\tau|$ , which (see box 1) is  $\frac{1}{2}[1 + \exp(-2k|\tau|)]$ , where  $k$  is the average number of crossings per unit time.  $k^{-1} = T$  is the characteristic timescale of the transitions, and we had to take the modulus of  $\tau$  to deal with  $\tau < 0$ . Finally we get

$$R(\tau) = \frac{1}{4}a^2[1 + \exp(-2k|\tau|)]. \quad (9.15)$$

The power spectrum is therefore

$$\begin{aligned} S(f) &= \int_{-\infty}^{\infty} R(\tau) \exp(-2\pi i f \tau) d\tau = \int_{-\infty}^{\infty} \frac{A^2}{4} [1 + \exp(-2k|\tau|)] \exp(-2\pi i f \tau) d\tau \\ &= \frac{A^2}{4} \left\{ \int_{-\infty}^{\infty} \exp(-2\pi i f \tau) d\tau + \int_{-\infty}^0 \exp(2k\tau - 2\pi i f \tau) d\tau + \int_0^{\infty} \exp(-2k\tau - 2\pi i f \tau) d\tau \right\} \\ &= \frac{A^2}{4} \left\{ \mathbf{d}(f) + \left[ \frac{\exp(2k\tau - 2\pi i f \tau)}{2k - 2\pi i f} \right]_{-\infty}^0 + \left[ \frac{\exp(-2k\tau - 2\pi i f \tau)}{-2k - 2\pi i f} \right]_0^{\infty} \right\} \\ &= \frac{A^2}{4} \left\{ \mathbf{d}(f) + \frac{1}{2k - 2\pi i f} + \frac{1}{2k + 2\pi i f} \right\} \end{aligned}$$

and finally

$$S(f) = \frac{A^2}{4} \left\{ \mathbf{d}(f) + \frac{4k}{4k^2 + 4\pi^2 f^2} \right\} = \frac{A^2}{4} \left\{ \mathbf{d}(f) + \frac{T}{1 + (\pi T f)^2} \right\}. \quad (9.16)$$

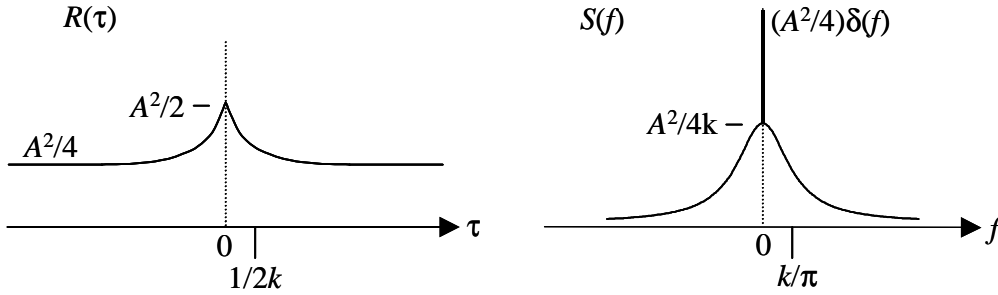


Figure 9.5. Autocorrelation function and power spectrum of a random telegraph signal

Hence the power spectrum is Lorentzian. The extra term  $(A^2/4)\delta(f)$  is the d.c. power, which is not very useful here because we subtracted an arbitrary d.c. background when we set  $x_0 = 0$ .

Box 1. If the transitions occur at completely random times (with no correlations between them), the probability of finding  $n$  transitions in a time  $|\tau|$  is given by the Poisson distribution,

$$P(n) = \frac{e^{-k|\tau|} (k|\tau|)^n}{n!}. \quad (9.17)$$

Hence the total probability of having an even number of transitions is

$$\begin{aligned} P(\text{even}) &= P(0) + P(2) + P(4) + \dots = \exp(-k|\tau|) \left[ 1 + \frac{(k|\tau|)^2}{2!} + \frac{(k|\tau|)^4}{4!} + \dots \right] \\ &= \frac{1}{2} \exp(-k|\tau|) \left\{ \left[ 1 - \frac{(k|\tau|)^1}{1!} + \frac{(k|\tau|)^2}{2!} - \frac{(k|\tau|)^3}{3!} + \dots \right] + \left[ 1 + \frac{(k|\tau|)^1}{1!} + \frac{(k|\tau|)^2}{2!} + \frac{(k|\tau|)^3}{3!} + \dots \right] \right\} \\ &= \frac{1}{2} \exp(-k|\tau|) [\exp(-k|\tau|) + \exp(k|\tau|)] = \frac{1}{2} [1 + \exp(-2k|\tau|)]. \end{aligned} \quad (9.18)$$

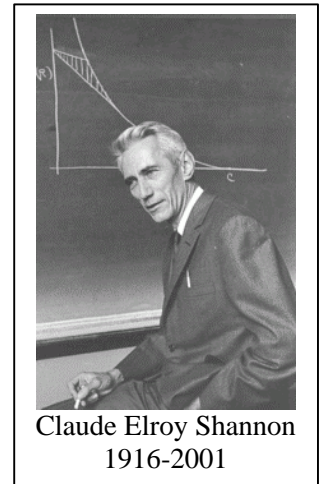
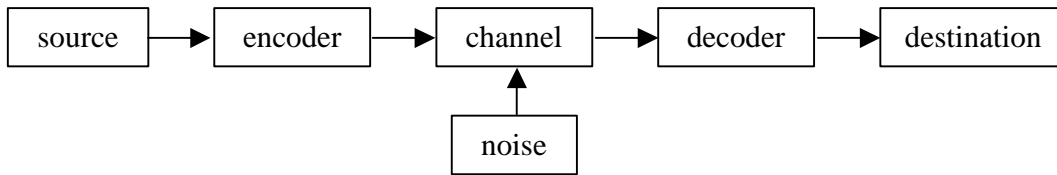
## 9.5 The generality of 1/f phenomena

A wide range of phenomena in the world are found to have an approximately  $1/f$  distribution, including both Beethoven and rock music. An even wider range of things has an  $f^x$  distribution where  $x$  is not near unity, such as the stock market indexes, earthquakes, and the sizes of cities. The key point about a power-law distribution is that it is scale invariant, meaning that there is no natural frequency. That is, no frequency is special. This is not true for instance with shot or thermal noise, where there is a characteristic frequency  $f_{max}$ . Scale invariance is frequently a sign of mathematical chaos. The field has been very active in the last decades, with toy earthquakes occurring irregularly in piles of sand in the basements of many physics departments (look up “self-organised criticality” if you’re interested).

## 10 Communication and Information

### 10.1 Communication theory

“The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point.” From “A Mathematical Theory of Communication”, Shannon, 1948.



Examples of the application of communication theory include:

- Two people talking
- Two computers communicating over an analog phone line or a satellite link
- Cells reproducing, where the parent DNA has to be reproduced in the daughter DNA
- Storage on a hard disk (in this case, data is communicated in time, not space)

In order to make sense of this block diagram, we have to know how to

- measure information content (note that “information” must not be confused with “meaning!”)
- encode data
- communicate (perfectly) over imperfect communication channels (having noise)

### 10.2 Messages, symbols and signals

The source selects messages from a set of possible ones. E.g., you (the source) select ‘yes’ from {yes, no, maybe}.

A message consists of a sequence of symbols (y-e-s) belonging to an alphabet (a,b,...z). There may be one or more symbols per message.

The source may be discrete: eg symbols are alphabet letters, Morse code, DNA sequence, bits or bytes  
or continuous: eg sound, voltage, light intensity

We concentrate on discrete sources. You can convert a continuous source to a discrete source by sampling (obeying the Nyquist criterion in  $t$ ) and binning (with resolution better than the uncertainty or noise in  $x$ ). We come back to continuous sources when we derive the Shannon-Hartley law (Sec. 13.x).

A sequence of messages is encoded into a signal for sending over the channel as a series of symbols. A never-ending signal is called a process.

### 10.3 Memoryless vs Markovian signals

Signals can be divided into two types:

- (1) Memoryless – no inter-symbol dependence, ie each successive symbol is a random choice from the alphabet – “bpdign cuwgm”
- (2) Markovian – symbols are correlated – “have a pleasant day”. The point is not that this string of symbols has meaning, but that up to a limit each symbol can be predicted from knowledge of the others using the rules of English. The same information is contained in “hv a plsnt dy”, as long as the receiver knows that the source has a habit of dropping vowels.

Markovian signals are much more common than memoryless ones.

## 10.4 Definition of information

The basic definition of information as a quantity, which we will gradually get a feeling for, is

$$\boxed{\text{Information} = \log(\text{number of possible choices})} \quad (10.1)$$

Consider a source having available a set of messages  $X = \{x_1, x_2, x_3 \dots x_n\}$  with corresponding probabilities  $P_X(x_i)$  of selecting message  $x_i$ .

For a *memoryless* process, we then define the information associated with the choice of  $x_i$  for one message as minus the log of the probability,

$$I(x_i) = -\log P_X(x_i) . \quad (10.2)$$

Similarly, the information associated with choosing two messages  $x_i$  and  $x_j$  in a row is (sticking to our earlier notorious probability notation) the log of the joint probability of messages  $x_i$  and  $x_j$ :

$$I(x_i x_j) = -\log P_{X_i X_j}(x_i x_j) , \quad (10.3)$$

and so on, using the multivariate probability for longer chains of messages.

Since the choices of messages  $x_i$  and  $x_j$  are completely independent for a memoryless process, Eq. (6.9) says that  $P_{X_i X_j}(x_i x_j) = P_X(x_i)P_X(x_j)$ , so that

$$I(x_i x_j) = I(x_i) + I(x_j) . \quad (10.4)$$

This quantity  $I$  is therefore additive as messages are added to the signal. This is just what we want for something that quantifies information. We expect there to be  $n$  times as much information in  $n$  messages as in one. The minus sign in Eq. (10.2) is needed because the *greater* is the probability of  $x_i$ , the *less* information is conveyed by choosing it – the number of choices in Eq. (10.1) is *smaller*.

For a Markovian process, Eqs. (10.2)-(10.4) are not valid, and the total information conveyed by  $n$  messages is always less than that for a memoryless process. This is because if there are correlations between the messages, the number of possible choices for the string of messages is reduced. However, Eq. (10.1) always remains correct.

## 10.5 Units of information

The (dimensionless) units of information depend on the base we take the logarithms in.

$$\log_2 \rightarrow \text{bits}$$

$$\log_e \rightarrow \text{nats}$$

$$\log_{10} \rightarrow \text{hartleys.}$$

We will always assume base 2, which makes sense because we are surrounded by binary logic signals. One can convert between them if necessary. For instance,  $I(\text{bits}) = I(\text{nats})/\log_e(2)$ .

## 10.6 Example: memoryless binary process

Consider the binary alphabet  $S = \{0, 1\}$ , with  $P_S(0) = P_S(1) = 1/2$ . What is the information associated with a message  $X$  consisting of a sequence of  $n$  symbols chosen from this alphabet?

If we choose  $n$  symbols, we have  $2^n$  possible messages – 001001...00, 101011...1, etc. These are all equiprobable because  $P_S(0) = P_S(1)$ , so that  $P_X(x_i) = 1/2^n$ .

From Eq. (10.2), the information associated with any message  $x_i$  is  $I(x_i) = -\log_2 P_X(x_i) = \log_2(2^n) = n$ .

Alternatively, from Eq. (10.1),  $I = \log_2(\text{number of choices}) = \log_2(2^n) = n$ .

In this case, the information is just the number of bits, as you'd expect.

## 11 Limits on efficiency and coding

The efficiency of communication depends on three things:-

- (i) the properties of the source,
- (ii) the choice of coding algorithm, and
- (iii) the rate of errors in transmission along the channel.

Coding involves choosing an algorithm to convert the output of the source into sequences of symbols suitable for sending along the channel, and decoding them at the other end. For instance “have a pleasant day” might be converted to a binary string “10011000101100111”. The most efficient code is one which requires the smallest number of symbols to be sent on average per message while guaranteeing less than a specified small rate of errors.

Shannon showed that:-

- (i) the properties of the source are measured in terms of its “entropy” and “redundancy” (see section 11.1),
- (ii) there is a limit to the efficiency of coding in the absence of errors, given by his “noiseless coding theorem” (see section 11.3); and
- (iii) even in the presence of transmission errors (noise), it is possible to find codes such that the received error rate is arbitrarily small, as specified by his “noisy coding theorem” (see section 12).

### 11.1 Entropy of a source

For a memoryless source producing symbols  $x_i$  from an alphabet  $X$ , the average information per symbol,  $H(X)$ , is given by

$$H(X) = E\{I(x_i)\} = -\sum_i P_X(x_i) \log P_X(x_i). \quad (11.1)$$

$H(X)$  is called the ‘entropy’ of the source, because in statistical mechanics the entropy  $S$  looks just like this, the probabilities there being those of the microstates in an ensemble. This points to the connection between information theory and statistical mechanics. (See the Resource Letter by W. T. Grandy for more.)

Entropy has units of bits per symbol if the log is base 2.

One can show that  $H(X) \geq 0$ , with the zero occurring if one symbol has probability unity and all others zero. This is because if the source produces the same message every time, each message conveys no information.

For a Markovian source, correlations reduce the information per symbol and thus by definition the entropy, and then more complicated expressions than Eq. (10.5) are needed for the entropy.

### 11.2 Maximum entropy of a source

If the memoryless source is limited to a given alphabet, the entropy has a maximum possible value  $H_{\max}(X) = \log N$  which is obtained when all  $N$  messages are equiprobable, i.e., when  $P_X(x_i) = 1/N$  for all  $x_i$ :

$$H_{\max}(X) = -\sum_{x_i} \frac{1}{N} \log \frac{1}{N} = \log N. \quad (11.2)$$

To prove this, consider  $H(X) - H_{\max}(X) = -\sum_i P_X(x_i) \log P_X(x_i) - \log N.$

$$\begin{aligned} \text{Using } \sum_i P_X(x_i) &= 1, \\ &= -\sum_i P_X(x_i) \log P_X(x_i) - \left[ \sum_i P_X(x_i) \right] \log N \\ &= \sum_i P_X(x_i) \log \frac{1}{NP_X(x_i)}, \end{aligned}$$

$$\text{and using the fact that } \log x \leq x - 1, \quad \leq \sum_i P_X(x_i) \left[ \frac{1}{NP_X(x_i)} - 1 \right] = \frac{1}{N} - 1 \leq 0.$$

$$\text{Thus } H(X) - H_{\max}(X) \leq 0. \quad (11.3)$$

Hence  $H(X)$  cannot be greater than  $H_{\max}(X)$ . The entropy  $H(X)$  is a property of the source. The most efficient source selects its messages with equal probability (and with no correlation between them).

### 11.3 Example: the game of Twenty Questions

**Question:** In this game, your friend thinks of an object. You are allowed to ask them questions to which they can answer only yes or no, like “is it vegetable or mineral?”. You win if you can find out what the object is in under 20 questions. What strategy should you adopt?

**Answer:** You want to maximize the information gained from the answer each question. For each question, your verbally restricted friend has to choose a message from an alphabet of only  $N = 2$  symbols,  $X = \{\text{“yes”}, \text{“no”}\}$ . The information per message,  $H(X)$ , has a maximum  $H_{\max}(X)$  of  $\log_2 N = 1$  bit, which occurs if each answer is equiprobable. Thus the optimum strategy is to ask questions where you estimate the probability of the answer being yes is  $1/2$  each time, based on everything you know (including the answers to the previous questions). This is just common sense.

#### 11.4 Relative entropy and redundancy of a source

The relative entropy of a source is defined generally to be  $\mathbf{a} = \frac{H(X)}{H_{\max}(X)} \leq 1$ . (11.4)

The redundancy of a source is defined to be  $\mathbf{b} = 1 - \mathbf{a} = 1 - \frac{H(X)}{H_{\max}(X)} \leq 1$ . (11.5)

These quantities measure how efficient the source is. For a memoryless source with an  $N$ -letter alphabet,  $H_{\max}(X) = \log N$  and so  $\mathbf{b} = 1 - H(X)/\log N$ .

#### 11.5 Example: English.

**Question:** what are the entropy and redundancy of a source of spoken English (e.g., me)?

**Answer:** Lower-case English has an alphabet of  $N = 27$  (26 letters + space) symbols,  $X = \{\text{“a”}, \text{“b”}, \dots, \text{“z”}, \text{“ ”}\}$ . Assuming English is a memoryless process with equal probabilities for all symbols, the entropy is

$$H(X) = H_{\max}(X) = \log_2 27 = 4.75 \text{ bits/letter.}$$

Allowing for the different probabilities of the letters (compare e.g. ‘e’ with ‘x’) but still assuming it’s memoryless,

$$\begin{aligned} H(X) &= -\sum_{x_i} P_X(x_i) \log P_X(x_i) \\ &= -P(\text{“a”}) \log \frac{1}{P(\text{“a”})} - P(\text{“b”}) \log \frac{1}{P(\text{“b”})} - \dots = 4.3 \text{ bits/letter.} \end{aligned}$$

However, taking into account the Markovian nature of English, the true entropy is only  $H_{\text{real}}(X) \approx 1$  bit/letter !

The relative entropy is therefore  $\mathbf{a} = H_{\text{real}}(X)/H_{\max}(X) \sim 1/\log_2 27 \sim 0.2$ .

The redundancy of English is  $\mathbf{h} = 1 - \mathbf{a} \sim 0.8$ .

The redundancy is high because successive letters and words are very strongly correlated. One consequence is that email can be compressed by a large factor. Another is that predictive text inputs in mobile phones work very well.

#### 11.6 Shannon’s noiseless coding theorem; code length

Once the source entropy  $H(X)$  is given, Shannon’s noiseless coding theorem (which we won’t prove) tells us that the code length  $L$ , which is the *average number of bits* (not necessarily an integer) needed to represent each symbol from the source, cannot be less than the source entropy, but it can be made arbitrarily close to it by clever coding:

$$L \geq H(X). \tag{11.6}$$

That is, the best any code can achieve is  $L = H(X)$ . Usually real codes don’t manage this, especially if extra bits need to be added to enable correction of errors due to noise, and the result is *redundancy of the coding*, meaning that more symbols need to be sent than would be absolutely necessary if there were no noise.

#### 11.7 Efficiency and redundancy of a code

We define the efficiency  $\eta$  of a code by  $\mathbf{h} = \frac{H(X)}{L} \leq 1$ , (11.7)

and its redundancy by 
$$R = 1 - h = 1 - \frac{H(X)}{L} \leq 1 . \quad (11.8)$$

Note that these are distinct from the relative entropy and redundancy of the source! From Eqs. (11.5) and (11.8) we see that the redundancy  $R$  of the encoding equals the redundancy  $h$  of the source if  $L = H_{\max}(X) = \log N$ , but  $L$  can be either larger or smaller than this.

When  $R \neq 0$  the coded signal can be compressed to have a shorter  $L$  and still contain the same information, though achieving nearly perfect compression is generally hard.

### 11.8 Example: ASCII text

What is the redundancy of encoding (a) a random string of Latin letters, and (b) English prose, into ASCII text? In this encoding, each English letter (symbol) is converted to a string of 8 bits (a byte), specified by the value (1–256) of the ASCII character assigned to that letter. The code length is therefore  $L = 8$  for every letter.

In case (a), from section 11.5,  $H(X) = \log_2 27$ , so by Eq. (11.8) the redundancy is

$$R = 1 - \frac{H(X)}{L} = 1 - \frac{\log_2 27}{8} \approx 0.41$$

In case (b),  $H(X) = H_{\text{real}}(X) \approx 1$  bit/symbol, and the redundancy is much bigger:

$$R = 1 - \frac{H(X)}{L} \approx 1 - \frac{1}{8} \approx 0.88 .$$

This implies that ASCII-coded English prose can be compressed by a factor of up to 8.

### 11.9 Reducing redundancy for a given alphabet - Huffman codes

Coding is a complex subject. We will just mention two ways in which the redundancy of a code can be reduced (ie, efficiency increased).

The most efficient code for the noiseless transmission of a memoryless source is a Huffman code. This assigns a unique binary string to each symbol from the source, using the fewest possible bits for each symbol while obeying the following rules:

- (1) the higher the probability of a symbol, the fewer bits are used to represent it, and
- (2) the division between bit strings representing consecutive sent symbols is always unambiguous.

The reasons for these two rules are common sense. We don't have time to go through this in detail, though you can find it in any of the books. The result is that  $L$  is typically only a few % bigger than  $H(X)$  for a memoryless source, so that the efficiency  $h > 0.9$  and the redundancy  $R < 0.1$  (see Eqs. 11.6-11.8) – very much better than the ASCII English coding above!

### 11.10 Reducing redundancy by combining symbols

If a source is memoryless, the redundancy can be reduced by encoding the symbols from the source in groups.

**Example:** How could we reduce the redundancy of the ASCII coding of a memoryless sequence of Latin characters in part (a) of section 11.8, if we have no choice but to use standard  $L = 8$  codes (perhaps because we have an 8-bit computer memory)?

Consider encoding strings of  $N$  Latin characters as strings of  $M$  ASCII symbols. The source entropy is then  $N \log_2 27$  and the code length is  $L = 8M$ , so

$$R = 1 - \frac{N \log_2 27}{8M} \approx 1 - \frac{N}{M} \times 0.594 .$$

We can choose  $N/M$  so that  $R$  is close to zero; for instance, if  $N = 5$  and  $M = 3$ ,  $R = 0.009$ .

This simple technique is less helpful if the source is Markovian, as you can verify. In that case however, encoding the symbols from the source in suitable groups can still be very effective. For example, when encoding English, one could take the source signal word by word, (so that the source “alphabet” consists of all the words in the dictionary), and replace each word by a binary string assigned by the Huffman technique. This completely takes account of correlations between the letters within each word, but not of the correlations between words resulting from grammar and context.

## 12. Errors and limits on information flow

In all real systems there is noise, and as a result the signal received at the output of the channel is not always exactly the same as that sent to the input – it contains errors, and the channel is said to be unreliable. Shannon showed that the errors can be corrected for very effectively by using appropriate coding. Thanks to this, we find we do not have to worry at all about errors even in 100 Gbytes of data stored on a hard disk.

### 12.1 Communicating via a channel: the channel matrix

Let  $X = \{x_1, x_2, \dots, x_M\}$  be the alphabet of symbols sent into the channel, and  $Y = \{y_1, y_2, \dots, y_N\}$  be the alphabet for the output. These are often the same, but need not be.

The error rates are then characterized statistically by the channel matrix:

$$M = \begin{bmatrix} P_{Y|X}(y_1 | x_1) & P_{Y|X}(y_2 | x_1) & \dots & \dots & \dots \\ P_{Y|X}(y_1 | x_2) & P_{Y|X}(y_2 | x_2) & & & \\ \dots & & \dots & & \\ \dots & & & \dots & \\ \dots & & & & P_{Y|X}(y_N | x_M) \end{bmatrix}. \quad (12.1)$$

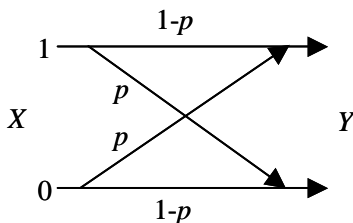
Here  $P_{Y|X}(y_i | x_j) =$  probability of receiving symbol  $y_i$  given that symbol  $x_j$  was sent.

The elements in each row sum to unity: 
$$\sum_{y_i} P_{Y|X}(y_i | x_j) = 1. \quad (12.2)$$

This assumes that some output signal must be received whenever something is sent.

If there are no errors,  $P_{Y|X}(y_i | x_j)$  is 1 for  $i = j$  and 0 for  $i \neq j$ .

### 12.2 The binary symmetric channel



The simplest example of an unreliable channel is the “binary symmetric channel”, with  $X = Y = \{0,1\}$ , and with a single “crossover” probability  $p$  that an error occurs, ie that a 0 get changed to a 1 or vice versa. The channel matrix is

$$\begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix}. \quad (12.3)$$

The binary symmetric channel is very useful for illustrating concepts.

### 12.3 Error correction – parity checks, Hamming codes, etc.

There are many ways of coding which allow errors to be detected and corrected, at the expense of increased redundancy. Examples are:

- (1) Repetition codes, where each symbol in the encoded signal is repeated a fixed number  $M$  times. If one of the repetitions is different from the others, the error is obvious and can be corrected.
- (2) Parity check bits, where an extra bit is added to each string of  $M$  bits in a binary signal. This parity is set to “1” only if there are an odd number of ‘1’s in the string. If any one-bit error occurs in the string on transmission, the parity check no longer works and an error is known to have occurred, although it cannot be corrected.
- (3) Hamming codes, where several parity bits are added to each string of  $M$  bits, in such a way that if a single error occurs it can be unambiguously located and corrected.

## 12.4 Conditional entropy, Mutual information, Equivocation and Irrelevance

There are a set of quantities called the *system entropies* which one can define for a communication channel. They are all measured in bits per symbol.

First are the input entropy

$$H(X) = -\sum_{x_i} P_X(x_i) \log P_X(x_i) \quad (12.4)$$

and the output entropy

$$H(Y) = -\sum_{y_i} P_Y(y_i) \log P_Y(y_i) \quad (12.5).$$

These are the average information per symbol going into and coming out of the channel.

We also define the equivocations:

$$H(X|Y) = -\sum_{x_i} \sum_{y_j} P_{XY}(x_i, y_j) \log P_{X|Y}(x_i | y_j) \quad (12.6)$$

and

$$H(Y|X) = -\sum_{x_i} \sum_{y_j} P_{XY}(x_i, y_j) \log P_{Y|X}(y_j | x_i). \quad (12.7)$$

$H(X|Y)$  measures the average information per symbol lost along the channel, and  $H(Y|X)$  measures the average information contributed by corruption.

Finally, we define something called the mutual information:

$$I(X;Y) = \sum_{x_i} \sum_{y_j} P_{XY}(x_i, y_j) \log \frac{P_{XY}(x_i, y_j)}{P_X(x_i)P_Y(y_j)}. \quad (12.8)$$

Using  $P_{XY}(x_i, y_j) = P_{Y|X}(y_j | x_i)P_X(x_i)$ , we find

$$\begin{aligned} I(X;Y) &= \sum_{x_i} \sum_{y_j} P_{XY}(x_i, y_j) \log \frac{P_{Y|X}(y_j | x_i)}{P_Y(y_j)} \\ &= \sum_{x_i} \sum_{y_j} P_{XY}(x_i, y_j) \log P_{Y|X}(y_j | x_i) - \sum_{x_i} \sum_{y_j} P_{XY}(x_i, y_j) \log P_Y(y_j) \\ &= -H(Y|X) - \sum_{y_j} P_Y(y_j) \log P_Y(y_j) \\ &= -H(Y|X) + H(Y) \end{aligned}$$

Thus

$$I(X;Y) = H(X) - H(X|Y) \quad (12.9)$$

By similar reasoning, we also find that

$$I(X;Y) = H(Y) - H(Y|X) \quad (12.10)$$

The mutual information measures the useful information carried by the channel.

Schematically, the various terms can be drawn in a flow diagram for the channel as in Fig. 12.1.

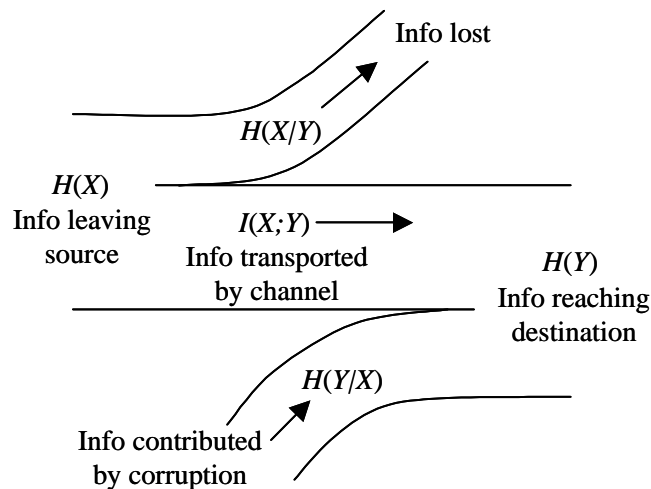


Figure 12.1. information flow through an unreliable channel.

## 12.6 Channel capacity

Shannon defined the channel capacity  $C$  (in bits per symbol) as the maximum value of  $I(X;Y)$  subject to varying the set of probabilities  $P_X(x_i)$  of the input.

For the binary symmetric channel you can show that

$$C = 1 + (1 - p) \log_2(1 - p) + p \log_2 p \quad (12.11)$$

## 12.8 Shannon's noisy coding theorem

**Statement:** It is possible to transmit signals with an arbitrarily low error rate (in bits per symbol) along an unreliable channel provided the rate of transmission does not exceed the channel capacity  $C$ .

The proof of this is cool, but it would require two lectures and some new maths, so it's not included in this course. It relies on grouping symbols from the source into very long strings (see section 11.10), and the fact that you can in principle produce a code where all the coded messages are more distant from each other in something called 'message space' than any given amount.

## 12.9 Continuous channels

A continuous channel is one that carries a continuous random variable, such as voltage. To deal with this, in the equations above we replace all the sums by integrals and the probabilities by probability densities, as in section 6.2. Strictly mathematically there is a problem in doing this, because the amount of information in a continuous variable measured to infinite precision is infinite. In any physical situation however the signal can only be measured to a certain accuracy, and can be placed in bins of size  $D$ , say, without loss of information. Thus what we really do is replace  $P_X(x_i)$  by  $p_X(x)D$  etc, and just approximate the result by an integral. Then, as long as the bin size is the same at the source and destination, you can show that the term dependent on  $D$  is identical at both ends and drops out, so that it has no consequence and can be omitted from the equations.

For instance,

Eq. (12.5) becomes 
$$H(Y) = - \int_{-\infty}^{\infty} p_Y(y) \log p_Y(y) dy,$$

Eq. (12.6) becomes 
$$H(X|Y) = - \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p_{XY}(xy) \log p_{X|Y}(x|y) dx dy,$$

and Eq. (12.8) becomes 
$$I(X;Y) = - \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p_{XY}(xy) \log \frac{p_{XY}(xy)}{p_X(x)p_Y(y)} dx dy$$

Eqs. (12.9) and (12.10) and all other similar relations remain valid.

## 12.10 The Shannon-Hartley law

This law tells you the channel capacity for band-limited continuous signals disturbed by white gaussian additive noise. The result is

$$C = \frac{1}{2} \log \left( 1 + \frac{\mathbf{s}_X^2}{\mathbf{s}_N^2} \right) = \frac{1}{2} \log \left( 1 + \frac{S}{N} \right)$$

where  $\mathbf{s}_N^2$  is the variance of the gaussian noise and  $\mathbf{s}_X^2$  is the variance of the input from the source, and the signal-to-noise ratio  $S/N$  is *not* measured logarithmically. The source is taken as gaussian, because one can show that, for a specified average power level, this maximizes the mutual information, as required in the definition in section 12.6. The proof is wholly based on concepts which we have covered in this course, but unfortunately we don't have time to include it this year!

## 13 Hopefully to be incorporated in the future:

Information theory and statistical mechanics, Maxwell's demon, Algorithmic information and complexity  
Jaynes's theory of probability, Quantum computing and encryption.