4. Transmission lines

In normal circuit analysis (e.g. Appendix I on Circuit Theory) we consider that voltages and currents propagate instantaneously. For many practical purposes this assumption is valid, i.e. we can assume for any piece of wire that the voltage at one end will be the same as the voltage at the other end. However:

- The wire will have some resistance, so if it is very long we may have a small voltage drop due to simple ohmic loss. In general, silver-plated copper wire is common-place and ohmic loss is negligible (and usually can be ignored);
- The voltage and current signal will actually propagate down a cable at about two-thirds of the speed of light. Since modern electronics operates at very high frequencies, and often we want to transmit information over quite long distances, this turns out to be a much more important consideration.

For example, consider the wired ethernet link which connects through Blackett (dimension $\sim$100 m). The datarate is up to 1 Gb/s. It takes about 500 ns for information to travel from one end of the building to the other, in which time a "sender" will have sent about 500 bits (62.5 bytes) of data before any information is received at the far end. So, 500 bits of data are travelling simultaneously down the wire at any given time – in the form of coded pulses. For high-frequency signals, the speed of light is actually rather sluggish and something of a limiting factor. Consequently, for circuits where the signal must run over a considerable cable length, we must we must consider the signal to be a wave propagating along the wire. A rule-of-thumb says we must worry about this for wire lengths L such that

$$L \approx \frac{\lambda}{10}\,.$$

There is a further complication: when L is comparable to λ, then what we have built is an aerial: the wire will very effectively couple electromagnetic energy into the environment. Our signal source and wire are a transmitter! This also works the other way around: any electromagnetic radiation in the environment will couple into the wire, and we will pick this up as noise or interference.

For all the above reasons and more we often choose to send high frequency signals down coaxial cables rather than single wires.

4.1. Coaxial cable

This is a type of wired connection where the signal is carried on a central conductor of radius r surrounded by a sleeve of dielectric material with relative permittivity ϵ. This is wrapped with an outer conductor (usually a braid) of radius R; this outer conductor is grounded. Since the inside-to-outside conductors form a capacitor, it is easy to see that the cable will have a capacitance which is dependent on the length of the cable. This is usually specified as a capacitance per unit length C' where

$$C' = \frac{2\pi\epsilon_0\epsilon}{\ln(R/r)}\,.$$

Similarly, there is an inductance per unit length

$$L' = \frac{\ln(R/r)\mu_0\mu}{2\pi}\,,$$

where μ is the relative permeability of the dielectric, which is always 1 for non-magnetic materials.

As we might expect, C' and L' depend entirely on the geometry and materials of the cable. The cable acts like a waveguide: signals will propagate via the transverse electromagnetic mode, with a radial E-field and a circumferential B-field. This means that the E-field is entirely trapped within the cable, and will not radiate away like it might do from a plain wire. Also, the outer grounded shield conductor 'shields' the signal from any pick-up of external interference.

The actual capacitance and inductance of a cable of length l are $C'l$ and $L'l$, which both tend to ∞ as $l \to \infty$. However, a propagating wave does not see the whole cable: this is analogous to a mechanical wave propagating along an infinite chain with a certain mass per unit length.

4.2. Characteristic impedance

Because it is modelled as a capacitance and an inductance, the cable is a linear system, in that the applied complex voltage results in a proportional complex current, the ratio of the two being the *characteristic impedance* of the cable:

$$Z_0 = \frac{\text{applied sinusoidal voltage}}{\text{resultant sinusoidal current}}\,.$$

Z_0 is characteristic of the medium through which the signal propagates. This is analogous to the acoustic and electromagnetic impedances of materials which can also be considered as the ratio of some applied-to-resultant sinusoid. This understanding of characteristic impedance is only relevant for cables significantly greater in spatial scale than the wavelength of the signal propagating. You will have covered the propagation of waves on a transmission line in Year 2 E&M, and all textbooks cover this well. The voltage and current on an infinite coaxial cable can both be represented by the wave equation with the solution that any arbitrary-shaped signal will propagate down the line at speed

$$v = \frac{1}{\sqrt{L'C'}}\,, \tag{4.1}$$

and the characteristic impedance is given by

$$Z_0 = \sqrt{\frac{L'}{C'}}\,. \tag{4.2}$$

Consider an infinite length of coaxial cable. If we introduce a voltage wave into the end of the cable it will propagate down the cable at speed v, never to return. Simultaneously, the voltage wave will draw a current wave, and the ratio of the two, at any point, is the characteristic impedance Z_0.

You can easily conclude that, assuming unity permeability, $v = c/\sqrt{\epsilon_r}$, which is around $2/3\,c$ for a typical cable dielectric. So, signals travel at approximately 1 foot/ns in free space, and at around 2/3 of that speed in a typical coax cable. For a small circuit this may not be a problem, but if your fast sensor is, say, 10 m away from its front-end amplifier, this can be a significant time.

4.3. Input impedance of a matched cable

First imagine a coaxial cable of infinite length. Seen from the outside, the infinite coax appears to 'swallow' the voltage signal (it disappears into the cable, never to return), so it acts the same as a resistor of value Z_0 in that it appears to dissipate all of the energy in the signal. Now imagine that the infinite cable is cut down to a finite length and terminated at the far-end with a resistor of value Z_0. When the signal reaches the far end of the cable, it encounters the resistor and is dissipated. No signal returns back up the cable. Therefore, a coax *terminated* with a resistor of value Z_0 (said to be a *matched cable*) has an input impedance which appears to be entirely real and have the value Z_0. Further, we cannot tell the difference between an infinite coax and a short one terminated with Z_0: from the sending end, they appear identical.

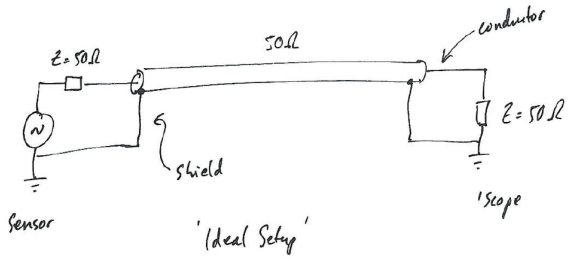

Figure 4.1: A matched transmission line delivers the power from the source to the load impedance with no reflections.

Note that the (ideal) coax cable is purely capacitive and inductive, so it does not dissipate any of the signal power. Also, the interpretation of Z_0 as being a real constant means that it is not frequency-dependent. So, the correctly terminated cable is "transparent", in that the signal passes down it with no reflection or absorption. The only effect of the cable is the time delay it introduces as the signal propagates down its length.

4.4. Mismatched cables

If we join two cables with impedances Z_1 and Z_2 then, as expected, we will get reflections at the boundary according to

$$k_r = \frac{Z_2 - Z_1}{Z_2 + Z_1}\,, \tag{4.3}$$

where k_r is the amplitude reflection coefficient (the ratio of incident wave amplitude to reflected wave amplitude). This corresponds to the Fresnel reflection of a light wave at normal incidence in optics: the square of this quantity gives the reflected intensity, and $1 - k_r^2$ is the transmitted intensity (and there is also an acoustic analogue).

The above equation can be derived by consideration of the boundary conditions at the junction. If the incident voltage amplitude is v_i and incident current amplitude is i_i, then we have also a transmitted wave (v_t, i_t) and a reflected wave (v_r, i_r). At the boundary, we have

$$v_i + v_r = v_t\,,$$

since the boundary is conductive the potentials are the same either side of the junction. Also, by conservation of current at the junction:

$$i_i - i_r = i_t\,.$$

The reflected current wave is travelling in the opposite direction to the incident wave. We also have everywhere:

$$v_i = Z_0\, i_i\,. \tag{4.4}$$

Since Z_0 is real, voltage and current are in phase; the solution yields equation (4.3) for the reflected amplitude (and k_r^2 for reflected power). Note that we can only apply this rather simplified analysis to obtain equation (4.3) since the v and i waves are in phase, i.e. because Z_0 is real.

4.5. Transforming impedance

Since the coax cable is truly a macroscopic medium through which an electromagnetic signal may propagate we can use exactly the same tricks which we saw applied in optics, where coated elements and interference coatings are often used.

Intermediate impedance. If we have a source with impedance Z_1 which we cannot change – such as an aerial (see example below) – which we need to connect to a non-matching load impedance Z_2, then we can insert a length of cable with $Z_0 = \sqrt{Z_1 Z_2}$ which will have the effect of minimising the overall signal reflection and therefore maximising transmission. An example from optics: magnesium fluoride (MgF_2) coatings have about the right refractive index to better match air (or vacuum) to glass.

Quarter-wave section. If we make the intermediate cable section exactly $\lambda/4$ in length, then we will obtain interference cancellation and hence we can approach a maximum of 100% transmission, albeit at a single frequency only. The same is true in optics: if the above MgF_2 coating has thickness $\lambda/4$, we can make the two reflections interfere destructively with each other and couple 100% of the intensity: we have thus made an anti-reflection coating at a particular wavelength λ.

4.6. Application example: the TV aerial cable

This is a simple example which serves to illustrate many advantages of using coax cables. The traditional aerial mounted on the roof is connected to the receiver using coax cable of 75Ω characteristic impedance. This impedance matches that of the aerial quite well. The aerial – at its simplest just a length of wire – is tuned to pick-up frequencies in the desired range. We want to transfer maximum signal power from the aerial into the receiver, with minimum reflections which would distort and diminish the received signal. The aerial is tuned so as to reject pick-up of unwanted frequencies, e.g. shorter wavelengths such as mobile phones or longer wavelengths such as FM radio. The screened/grounded coax cable prevents these signals from being picked-up en-route to the receiver. Since the TV signal is entirely contained within the cable (like a waveguide) it does not couple capacitively to any nearby metal objects, such as the drain-pipe down which the cable may have been routed.

4.7. Real-world effects

The coax cables described above were assumed to be ideal or *lossless*. In practice, any substantial length of cable will have some ohmic resistance which will contribute to an overall reduction of the signal amplitude or power. This is called *ohmic loss* and is independent of frequency. For very high-frequency signals the dielectric material is not fully efficient, and hence the E-field will dissipate some power into the dielectric. This is called *dielectric loss* and is frequency-dependent. The net effect is usually quoted as an overall power attenuation factor per unit length. For example the typical type of RG-58 cable commonly found in labs has a characteristic impedance of 50Ω and an attenuation factor varying between 0.11 dB/m at 50 MHz and 1.4 dB/m at 2 GHz.

4.8. Key points

These are the main points from this section:

- The characteristic impedance Z_0 is only a function of material and geometric properties of the cable;
- Z_0 is real and independent of frequency, yet the ideal line is lossless;
- The voltage and current on the line are everywhere in-phase, and their ratio is Z_0;
- A line terminated with a resistance $R = Z_0$ appears to be totally transparent: its input impedance is R;
- A mismatched line will result in reflections at the boundary with an ensuing reduction in signal power. The reflected power will be proportional to k_r^2;
- A line left open (or short-circuited) at the far end will result in 100% reflection of the signal; the returning signal will be inverted for the short-circuit case.

5. Digital signals

Most signals we encounter in physics are continuous functions of time, such as light intensity, temperature, pressure, etc. If we want a computer to interact with these variables we have to first sample and then digitise them. The process of *sampling* quantises time, while *digitisation* quantises the parameter being measured. Ultimately, this means we lose some information about the original signal, and we have to be careful in how we interpret digitised signals. This section will outline some of the basics; we postpone a discussion of the electronics implementation of some key circuits for analogue-to-digital conversion until Section 12.

Note that sampling (and digitisation) are not entirely phenomena of the computer age. Taking a temperature reading (every hour, every day, ...) is a process of sampling. Converting the height of mercury in the thermometer (a continuous, analogue parameter) into a number written down in a notebook is digitisation.

This section is based on Chapter 3 of the book "The Scientist and Engineer's Guide to Digital Signal Processing" by Steven Smith. This book is very readable and provides a good descriptive text though not always with as much mathematical backing as we might like. It is available online.[2]

Learn it in lab
The first lab session is devoted to the topics discussed in this section – you will see how signals can be properly sampled and digitised, and the effects of aliasing in detail. Conversely, you should read this section again before that particular lab.

5.1. Sampling

As suggested above, sampling-and-digitisation is a two-stage process, as illustrated in the central diagram of Figure 5.1. Taking an analogue signal (typically a voltage) as an input, we must first sample it, which is mathematically equivalent to taking an instantaneous value of the function at some point in time. Usually, we want a series of samples with uniform time spacing. If the time between samples is T_s, then we can define a sampled version of the function as

$$f_s(t) = \sum_{n=-\infty}^{\infty} f(nT_s)\,\delta(t - nT_s)\,. \tag{5.1}$$

This says we take the function $f_s(t)$ and we multiply it with an infinite sequence of delta functions (mathematically this is the 'comb' function). What we get is an infinite sequence of delta functions, each one weighted by the instantaneous value of the function. The main point to take from this is that a *sampled signal* is mathematically very different from a continuous signal in that it is non-zero only for the values of time $t = nT_s$.

[2] www.dspguide.com

Sample and hold. In practice this mathematical representation is very far from reality, since:

- We know that it is actually very difficult to get an instantaneous value of a function since any real-world hardware will take a short but non-zero time to take a sample;
- As per Figure 5.1, we want to pass the sampled value into the next block which performs the digitisation. This can take some considerable length of time to do, so we want to value of the sample to persist.

The solution is the *sample-and-hold* circuit, whose job is to take a snapshot sample of the waveform and then hold that value on its output. Note that the output is still an analogue voltage. The circuit takes a sample when commanded to do so by an external *clock signal*. The clock is usually a square wave with period T_s. The sample typically occurs on each *rising edge* of the clock signal, so if we have a nicely stable clock signal we will have very regular sampling, which is very important for the quality of the sampled signal. The variability in the sampling period is called "jitter", and is usually very small. The circuit for the sample and hold is relatively straightforward, but we will look at this later.

As shown in Figure 5.1, changes at the input that occur between sample times are ignored, that is to say, sampling converts time (the independent variable) from continuous to discrete.

5.2. Digitisation

In Figure 5.1 we can assume that the input signal can vary from 0 to 4.095 V which the *Analogue to Digital Converter (ADC)* will translate into the digital numbers 0 to 4095. The ADC produces an integer value for each of the samples. Only integer values are allowed, which means that the ADC has a *digital resolution* of 1 mV (the smallest change on the input which will guarantee a change in the output number). Therefore, the ADC performs a *quantisation*: it converts the voltage (the dependent variable) from continuous to discrete. The bottom panel of the figure shows the error that this process introduces called the *quantisation error*, defined as the difference between the sampled signal and its digital representation. It it presented in terms of digital numbers so the probability density function (PDF) shows that it is uniformly distributed between ± 0.5. The fundamental digital number is frequently referred to as the *Least Significant Bit (LSB)*, for reasons we will discuss soon. The ADC is a rather more complicated circuit than the sample-and-hold, and again we will examine this later.

Binary digits. A 'bit' is a binary digit which consequently has only two values: '1' or '0'. In our ADC example we can have any digital number between 0 and 4095, that is to say 4096 distinct values, which are represented by a 12-bit binary number (since $2^{12} = 4096$). Table 5.1 gives some

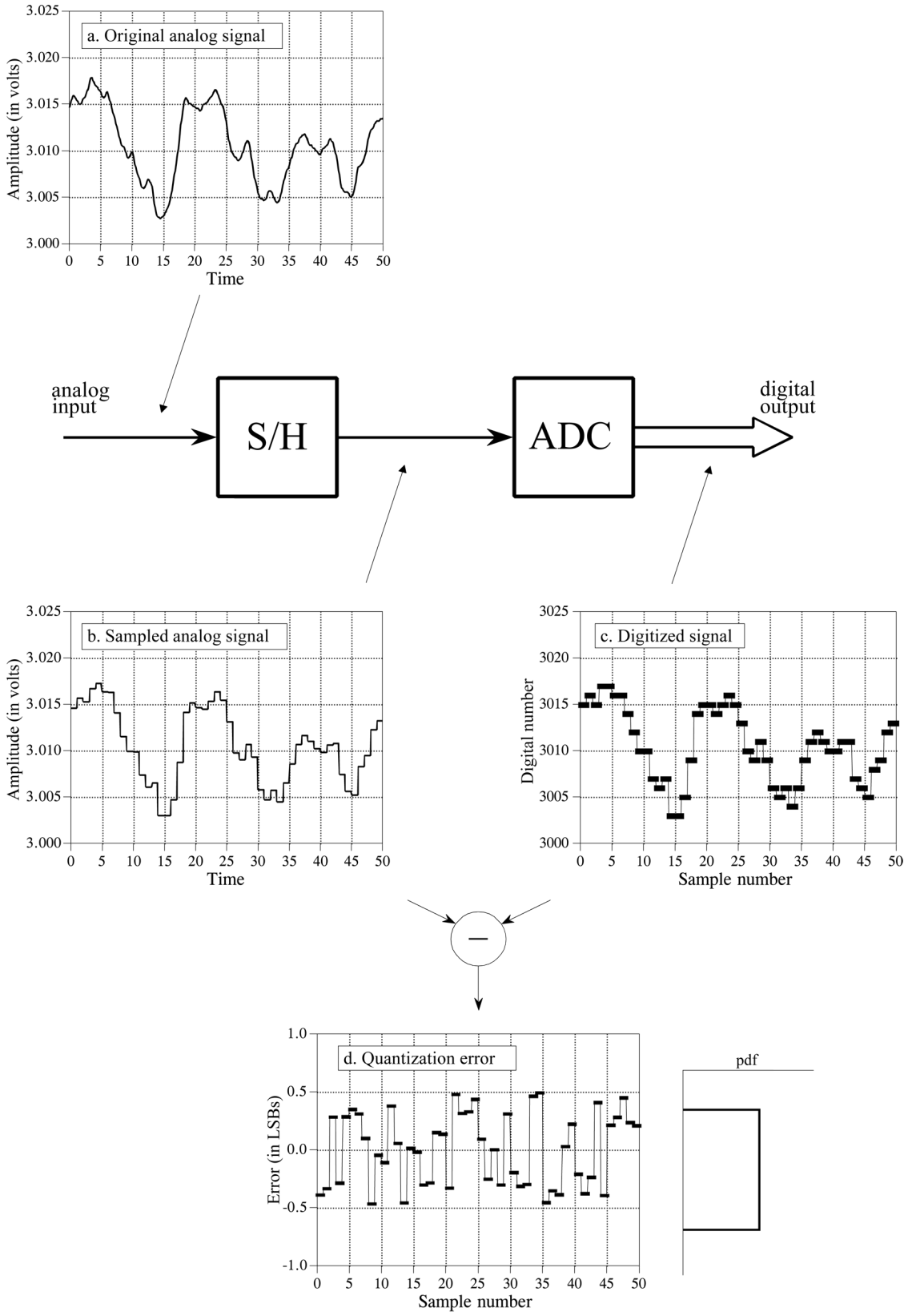

Figure 5.1: Waveforms illustrating the two-stage process of sampling and digitisation (from Smith, www.dspguide.com).

Table 5.1: Some 12-bit binary numbers and their decimal equivalents.

binary	decimal
000000000000	0000
000000000001	0001
000000001110	0014
111111111111	4095

examples. Digital electronics uses binary representation of numbers since this fits nicely with the fundamentally two-state 'ON=+5 V' and 'OFF=0 V' nature of digital logic.[3] The 'number of bits' is a measure of *digital resolution*. If we were to use a 14-bit converter over the same input range then the resolution would increase four-fold (each digital number now represents 0.25 mV). In engineering-speak, the LSB (right-most binary digit) is 0.25 mV.

Quantisation error. Quantisation results in nothing more than the addition of some random noise to our signal. As mentioned above and in Figure 5.1, this extra noise is usually uniformly distributed between ±0.5 LSB. This means we can define the noise statistically according to its mean μ and standard deviation σ:

$$\begin{aligned} \mu &= 0 \\ \sigma &= \frac{1}{\sqrt{12}}\,\mathrm{LSB}\,. \end{aligned}$$

Now, for distributions with zero mean, the RMS is the same as the standard deviation (check the definitions if you are unsure). If we want to reduce the RMS noise in our digitised signal, we need to "improve the resolution of the measurement by increasing the number of bits".

Example: Quantisation noise and full-scale
Passing an analogue signal through an 8-bit ADC adds an RMS noise of $1/\sqrt{12}$ LSB. Since the full-range (maximum) input is $2^8 = 256$, this is about 0.1% of the full-range value. If we switch to a 12-bit ADC this is reduced to 0.007% (much more acceptable, generally).

This understanding of quantisation noise is extremely powerful, because the RMS noise generated during digitisation simply adds in quadrature with any noise already existing in the analogue signal (and as we shall see later, there is always some noise in any signal).

[3] See Horowitz & Hill chapter 8.

Example: Quantisation noise and signal noise
Returning to our original example of the 12-bit converter with a resolution of 1 mV. Say the RMS noise measured in the analogue signal is 0.5 mV then this translates to 1/2 LSB in digital numbers and the total noise is $\sqrt{1/4 + 1/12} \approx 0.6$ LSB. Note how the digitisation increases the overall noise by a small amount. We might be tempted to upgrade to a 14-bit converter, but actually there is little point, since the total noise will always be dominated by that in the original signal.

As a final note on the topic, recall that the ADC returns only integer values but we are discussing here noise being a fraction of 1 LSB. We will only see this noise in the statistics of a large number of samples, each of which is an integer, but taken together they have non-integer statistics.

5.3. Aliasing and the sampling theorem

We throw away a lot of information about a signal when we digitise it because we quantise both time and the parameter being measured. This is probably the single most important point to understand about digitised signals: information lost in the digitisation process is lost forever – there is no way we can recover the information about what the signal was doing between samples, nor can we say anything about the small amplitude details which are below the resolution of our ADC. In the case of the latter, we have seen how the digitisation of the signal can be characterised as extra noise added to the signal, which we can understand statistically. Now we will turn our attention to the effect of the former, that is to say the sampling of the signal and the quantisation of time. First we will look at the *Sampling Theorem*, which is actually a rule about how to ensure *proper sampling*, and then we will look at *aliasing*, which is the effect we get if we break the sampling theorem rule, resulting in *improper sampling*.

The Sampling Theorem. Put simply, a digital signal can only properly represent frequencies up to *one half of the sample rate ($1/T_s$, where T_s is the sample period)*. This is somewhat intuitive. What is the minimum number of points we need to (very crudely) represent a sinusoid? It is worth trying to sketch this on a piece of paper, with maybe 10, 5, 3 or 2 points per cycle of the sinusoid. Join the points with a straight line. It should be clear that 2 is the absolute limit; it looks more like a triangle wave than a sinusoid, but the key point is that it represents the correct frequency. We will look at this in more detail in the next section. The formal statement of the sampling theorem (also known as the Shannon sampling theorem, or the Nyquist theorem) is that:

A continuous signal can only be properly sampled if it does not contain frequency components above one half of the sampling rate.

For example, a sample rate of 2000 samples/s (S/s) requires that the original signal only contain frequencies below 1000 Hz. Half the sample rate (here 1000 Hz) is called the *Nyquist frequency*. If signals above the Nyquist frequency are present they will be aliased, which means they will appear as new signals at frequencies below the Nyquist frequency. This is bad; it is often referred to as *improper sampling*.

Aliasing. This is best illustrated graphically, in the first instance. Figure 5.2 shows some some sinusoids before and after sampling. The continuous line is the input and the square markers are the sampled values. The question is: for each waveform input can we unambiguously recreate this from the samples? For the DC input (panel (a), a cosine of zero frequency!) we can certainly say "yes". For panel (b) we might have a 90 Hz signal sampled at 1 kS/s, so there are more than 11 samples per cycle of the signal. This is clearly proper sampling, according to Nyquist. For panel (c) we have only 3 samples per cycle. Imagine you take away the input (solid line) and join adjacent dots. Intuition tells us this is a roughly-drawn sinusoid at the correct frequency. It turns out that the human brain is very good at this sort of pattern recognition. More importantly, if we were to put these samples into a Fast-Fourier Transform program, it would also correctly report the frequency of the signal. This is still *proper sampling*. In (d) we push this further to just over 1 sample per cycle. Disaster: both the human and the FFT report a sinusoid at the wrong frequency. This is *improper sampling*. In fact, if the sample rate is 1 kS/s and the input frequency is 950 Hz, then we get a sinusoid of frequency 50 Hz in the digital data. In general, for an input frequency f and a sample rate f_s, then if $f > f_s/2$ we get an alias frequency f_a such that

$$f_a = f_s - f\,.$$

Now, the problem arises that our input signal may contain frequencies going to values many times f_s, so we need to generalise this to

$$f_a = |nf_s - f|\,, \tag{5.2}$$

where n is an integer chosen to give a value of nf_s as close as possible to f. This may seem a little confusing but we will see why this is in the next section. As an example of this, assume $f_s = 100$ Hz and the input signal contains all of the frequencies 25, 70, 160 and 510 Hz added together. The spectrum of the analogue signal would show all these frequencies. The spectrum of the digital signal shows the frequencies 25, 30, 40 and 10 Hz. The first is correct, but the last three are *aliases*.

But it gets worse. If we see a signal at 10 Hz in the digital data we have no means of knowing if the original analogue signal was 10 Hz, 90 Hz, 110 Hz, 190 Hz, etc. It could be that all of these signals were present, so all of the aliases would add on top of the real 10 Hz signal, thus destroying any knowledge we might need about the amplitude of the original 10 Hz signal. This illustrates a key point about aliasing: not only does it generate new false frequencies, it can also destroy information about the correct, lower frequencies. Because it is so important, we will study this in more detail in the first lab session.

5.4. The frequency characteristics of sampled signals

One important point to begin with is that a sampled signal is fundamentally unlike any other kind of continuous signal you will have come across before. According to equation (5.1), the original signal has been multiplied by a series of delta functions to create what we might call an 'impulse train'. It is non-zero for values of nT_s, but zero in-between. This gives it a very complicated spectrum, which we can see by taking the Fourier Transform of (5.1) to get[4]

$$F_s(\omega) = \frac{1}{T_s}\sum_{n=-\infty}^{\infty} F(\omega + n\omega_s)\,, \tag{5.3}$$

$$\omega_s = \frac{2\pi}{T_s} = 2\pi f_s\,.$$

This shows that the spectrum of the sampled signal is the same as the original signal, but repeated infinitely along the frequency axis. Figure 5.3 gives a graphical illustration of this. Panels (a) and (b) show the original signal and its spectrum. We can see that the signal is limited to a band of frequencies below the Nyquist frequency, so we should be able to sample it properly. In fact, we are sampling comfortably above this at about 3 times the highest frequency in the analogue signal. Note that this signal and its spectrum is highly stylised; real signals rarely have such neatly compact spectra, as we shall see later; however, this illustrates the principle here.

Panel (c) shows the sampled version of the signal, in the form of an impulse train, and in the spectrum (d) we can see the new repeating frequencies generated by the sampling process. The reason for this is not straightforward but once understood it does provide a rather satisfying explanation for aliasing. In equation (5.1) we can see that the original signal was multiplied by an infinite sequence of delta-functions (the so-called comb function). Now, the Fourier transform of the comb function happens to be another comb function.[5] Further, multiplication in the time-domain is equivalent to convolution in the frequency domain. Therefore, in the frequency-domain we expect that the spectrum of the sampled signal is the spectrum of the original signal convolved with a comb function.

[4] We will cover the integral Fourier transform in more detail in the next section. In general, the mathematics for sampled signals and the equivalent transforms into the frequency domain are rather involved and beyond the scope of this course. The result is quoted here to give an understanding of the behaviour of the sampled signal under the Fourier transform, but you would not be expected to know or derive this.

[5] See the table of Fourier transforms in Poularikas.

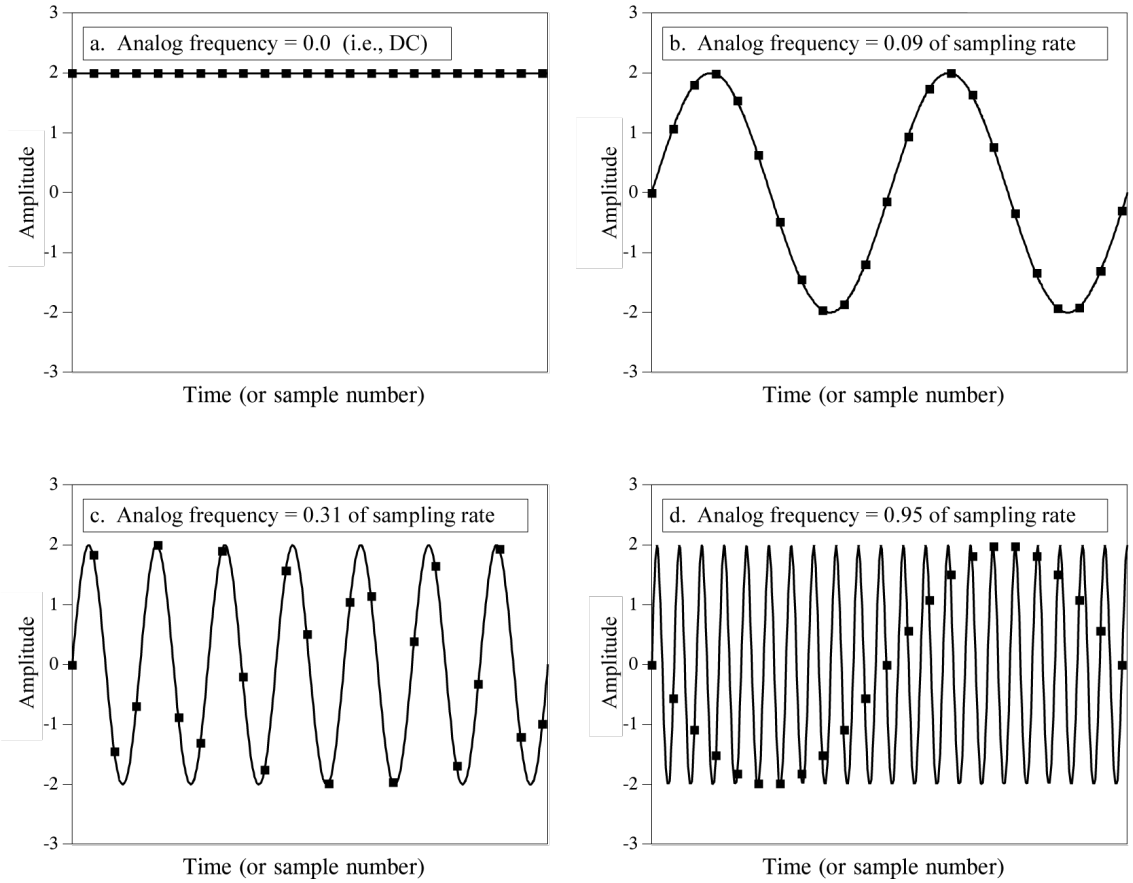

Figure 5.2: Graphical representation of proper and improper sampling (from Smith, www.dspguide.com).

This is why the spectrum repeats infinitely. Note that in the figure only the positive frequency range is shown. We know that the Fourier transform generates negative frequencies as well; this is what accounts for the part of the spectrum labeled "lower-sideband". The spectra "copies" repeat each multiple of the sample frequency. If we reduce the sample rate by a factor of two – panel (e) – then the spectra get closer together – panel (f) – and cross-over into each other. Recall that in the sampled signal we only properly represent signals up to the Nyquist frequency ($f_s/2$) then we see here how frequencies from the first repeated spectrum intrude into this range – this is aliasing.

As a final word on the topic, take another look at the spectrum of the properly sampled signal in panel (d). We can see that the sampled spectrum contains in the range 0 to $f_s/2$ *an exact copy of the original signal's spectrum.* This has two profound consequences:

1. For proper sampling, we have not lost any frequency information as a result of the sampling process.
2. If we take the the sampled signal and remove all the frequencies above $f_s/2$ then we will recover the original signal.

The latter can be done by the use of a low-pass filter. This technique is the basis of *digital-to-analogue conversion*, as we will see later.

5.5. Anti-alias filters

We are not finished with aliasing yet. As seen in the previous section, if our signal is contained within a narrow band of frequencies below the Nyquist frequency, then all is OK. However, in practice this is rarely the case for two reasons:

1. Many real signals (either pulses or repetitive waveforms) can approach mathematically infinite waveforms (for example, the square wave). While no real signal is mathematically perfect, any signal with rapid changes ("edges") has a very wide frequency spectrum;
2. All signals contain noise, and noise is usually broadband, i.e. existing across the frequency spectrum.

In order to prevent these high frequency components aliasing into our digitised signal, it is usually preferable to remove them from the signal *before* sampling. Note that this is a compromise solution: if we filter out high frequency parts of our signal we certainly degrade it; however, the

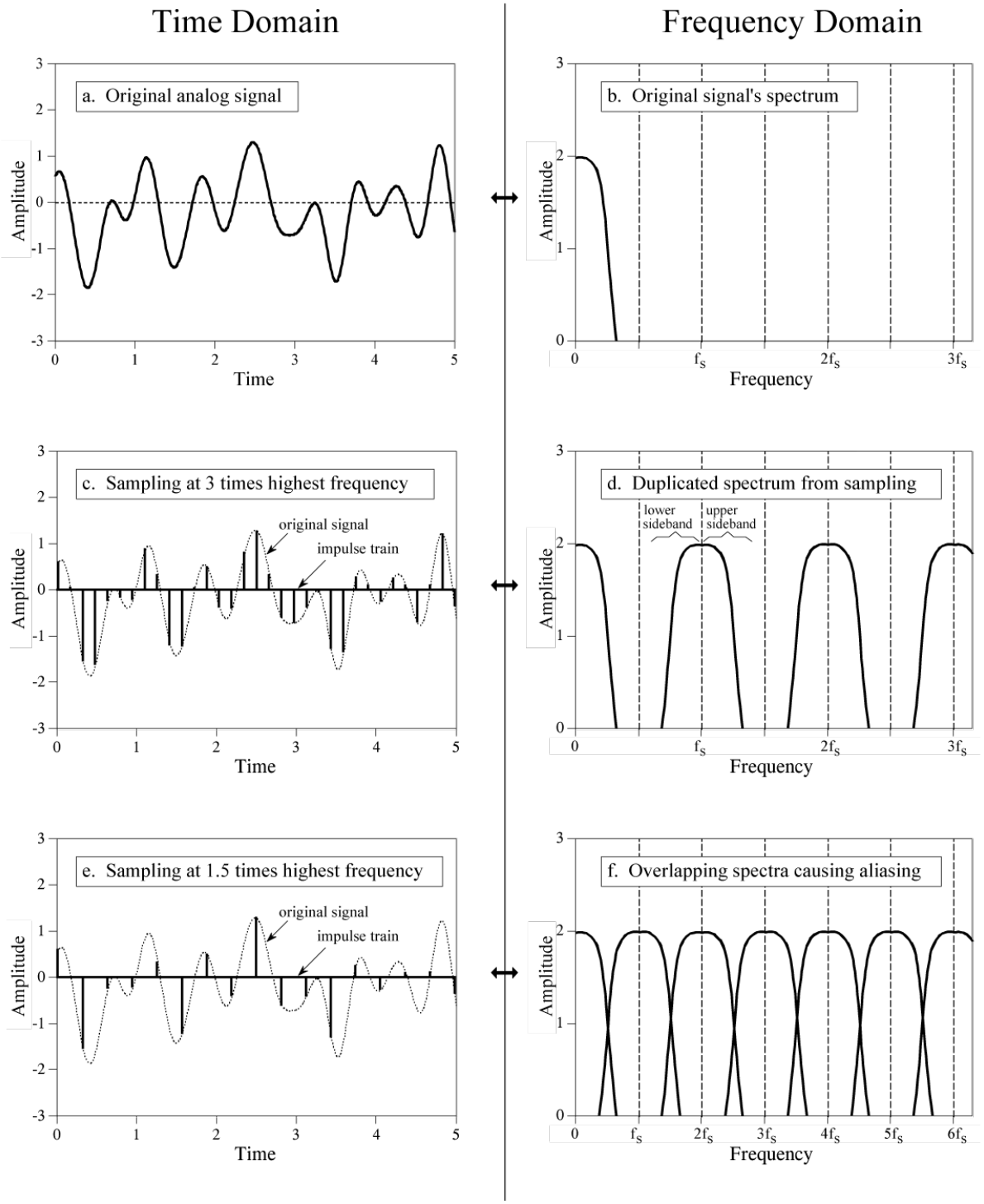

Figure 5.3: Sampled signals in the frequency domain (from Smith, www.dspguide.com).

problem of aliasing is so bad that this is usually a price worth paying. Figure 5.4 shows a Digital Signal Processing (DSP) system as it should be setup. Imagine this is an audio system; then the analogue input on the left is a voltage signal coming from a microphone. The first stage is a filter designed to remove frequency components above the Nyquist frequency. This is called the *anti-alias filter*. The signal is then sampled and digitised by the ADC and can then be stored and processed by the computer (central box). Everything to the left of the diagram is

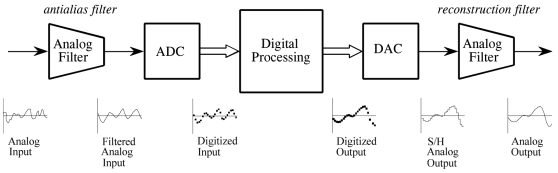

Figure 5.4: Correctly set up Digital Signal Processing system (from Smith, www.dspguide.com).

the recording system. The right side is for "playback". As mentioned above, we need a *Digital to Analogue Converter (DAC)* and another filter to properly reconstruct the original analogue signal, which then drives the speakers. It is worth noting that all of this, and of course much more, is contained within the modern mobile phone...

Entire books have been written on the subject of anti-alias filters, and electrical engineers receive entire lecture courses on filter design. For our purposes, including the lab sessions, we will use the simple RC low-pass filter discussed in Section 7. The ideal filter would have a very sharp cut-off, that is to say, it would allow through all frequencies below the Nyquist frequency and completely block anything above. Unfortunately, this is impossible. In the case of the RC low-pass filter we usually and choose the -3 dB cut-off frequency of the filter to be the same as the Nyquist frequency of the sampling. This is, however, a compromise, as can be seen in Figure 5.5. This filter would be suitable for sampling with Nyquist of 1 rad/s. Some frequencies below the Nyquist value are attenuated (undesirable) and also some frequencies above Nyquist will still exist (also undesirable). Filters with a sharper "cut-off" are possible; we will study this briefly when we consider active filters built with op-amps, later in the course.[6]

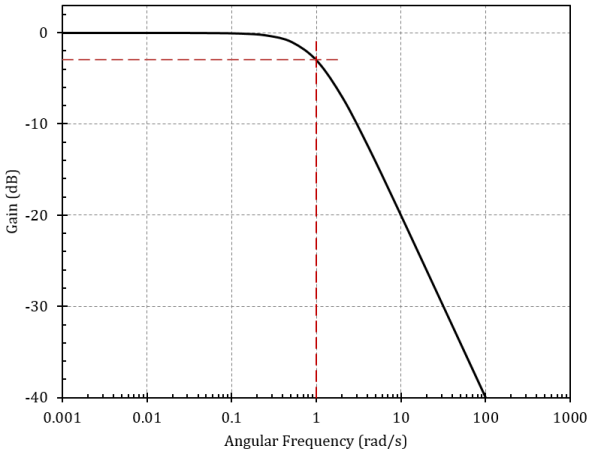

Figure 5.5: Response of an RC low-pass filter with a (-3 dB) cut-off frequency of 1 rad/s.

[6] See Smith chapter 3 at www.dspguide.com and also Horowitz & Hill chapter 4.