Instrumentation

November 26, 2010

Contents

1 Sensors and Transducers

Generally a **Sensor** is a device which allows us to take a measurement of some physical observable. A mercury thermometer is a sensor. Typically we want the output of the sensor to be an electrical signal so that we can do further manipulation and processing of the signal using electronic circuits and ultimately get a representation of the signal into a computer-readable form. Also we may want to go in the other direction: i.e. converting an electrical signal into some kind of physical observable (such as light, heat, sound, movement etc.) Here, the **Transducer** model is useful. An **Input Transducer** converts some physical observable into an electrical signal, and an **Output Transducer** goes in the other direction. Figure 1.1 illustrates the idea with a simple "public address" system. The microphone is an input transducer which converts speech (sound pressure variations) into a voltage signal. Then we have an amplifier, where we can also do more fancy things such as adjusting the tone (frequency content) of the signal. The amplifier drives a speaker which is the output transducer, whose job is to convert the amplified electrical signal back into a mechanical movement (of the speaker diaphragm) which in turn gives us sound. This is a good example since we'll be studying a (more complicated) system of sound/electrical transducers in the lab sessions.

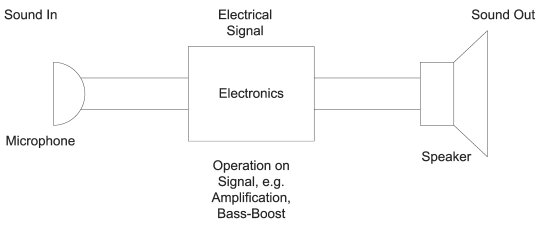


Figure 1.1: Input/Output Transducers

There are thousands of different types of transducers and we'll only look at a few in this course. This is not a course on detector physics so we don't have the time to study the details of individual transducers, however we will see that there are many similarities in how transducers behave so we can start to see how to use them in a more general (that is to say device-independent) way. In this handout we will look briefly at two different types of transducer.

1.1 Temperature Transducers

Thermocouple This is the simplest temperature transducer since it directly generates a voltage. A thermoelectric voltage is generated across any metal when the two ends are at different temperatures. Using two junctions between dissimilar metals - one held at a known temperature, the other at the point we wish to measure - the voltage generated is proportional to the temperature difference between junctions. The voltage is small - of the order a few μV/°C - so precision voltage measurement is required. The technique requires some care, however the thermocouple is robust, small, and operates over a very wide temperature range (-200 to 1500°C).

Platinum Resistance Temperature Device (PRTD) This is the standard for high measurement over the range -200 to 600°C. The PRTD shows a nicely linear variation in resistance as a function of temperature. Consequently we just need to measure the resistance (e.g. with a meter) then apply a conversion factor to get the temperature. On closer inspection it is found that the relationship is not quite linear, so for highest accuracy we need to use a polynomial fit. This non-linearity makes the conversion from resistance back to temperature more complex, but do-able.

Thermistor This is a semiconductor device where resistance *decreases* rapidly with temperature. The temperature coefficient of resistance is of the order 4%/°C, which is an order of magnitude better than the PRTD. The thermistor is also small, cheap and robust, and is since the requirement on accurate resistance measurement is relaxed compared to the PRTD, this makes the thermistor the natural choice for quick, easy temperature measurement over

the range -75 to 150°C. They are available in all sorts of packages for example the glass-encapsulated ones are good for harsh or chemical environments (though since glass is an insulator this slows down the response time to temperature changes). One drawback is that the temperature-resistance relationship looks like

$$\frac{1}{T} = A + B \ln R + C(\ln R)^3 \qquad (1.1)$$

To go from R to T we need either a calculation or a look-up table, both of which are possible with electronics and/or a micro-processor, but this does mean that we can't simply hook the thermistor up to a meter in the lab and take a quick reading of the temperature.

The point of this discussion is that none of the above are perfect and we have to make a choice according to

- the temperature range we need to cover and the accuracy required
- the harshness of the environment
- the ease of conversion into a temperature value

This is a general rule for transducers: the perfect device does not exist and we need to compromise.

AD590 Solid State Temperature Transducer Where such a situation exists, the semiconductor industry is always ready to sell us a solution. The AD590 from National Semiconductor is an integrated circuit device which produces a highly linear *current* output of 1μA/K. It works over a reasonable temperature range of -55 to 150°C. It comes in a range of packages (Figure 1.2) and can be bought for under $10 (more expensive than a thermistor but cheaper than a good PRTD).

A simple circuit of resistors converts the current output to a voltage, then a standard lab multi-meter gives us a very quick and easy direct reading of the temperature. This is the

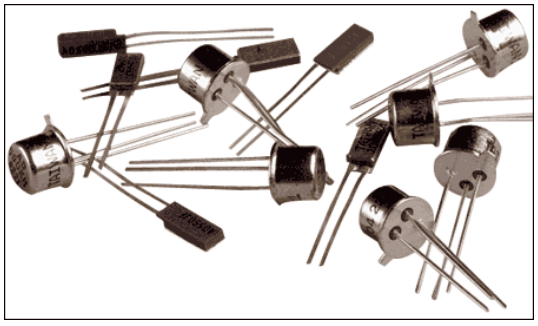


Figure 1.2: National Semiconductor AD590

big advantage of using the AD590. The complexity of linearising the output as a function of temperature is all handled within the circuitry of the device (this is what we are paying for). We don't generally need to be concerned with the physics of what goes on inside the device as long as we have a good data-sheet to describe its behaviour. The front-page of the AD590 is reproduced in Figure 1.3. It is useful to read the complete data-sheet and you are recommended to download a copy from the manufacturer's website at http://www.analog.com. Browse through it to see the level of detail that the manufacturer thinks is important for the user to know. You also might want to identify the ways in which the performance of this transducer departs from 'ideal' behaviour.

1.2 Photomultiplier Tube

For the temperature transducers we are not, generally, expecting to operate at the limits of physics (unless we are trying to measure mili-Kelvin, but then we would use a different technique anyway). We often need to measure the temperature of our experiment or equipment and the above techniques will usually suffice. Next we'll look at something more cutting-edge (recall the dark-matter experiment mentioned in Lecture 1). The Photomultiplier Tube (PMT) is a type of light intensity transducer. It is based on vacuum-tube technology. More modern semiconductor devices exist for measuring light but the PMT is superior for noise and sensitivity at

ANALOG DEVICES

2-Terminal IC Temperature Transducer

AD590

FEATURES

Linear current output: 1 µA/K
Wide temperature range: −55°C to +150°C
Probe-compatible ceramic sensor package
2-terminal device: voltage in/current out
Laser trimmed to ±0.5°C calibration accuracy (AD590M)
Excellent linearity: ±0.3°C over full range (AD590M)
Wide power supply range: 4 V to 30 V
Sensor isolation from case
Low cost

GENERAL DESCRIPTION

The AD590 is a 2-terminal integrated circuit temperature transducer that produces an output current proportional to absolute temperature. For supply voltages between 4 V and 30 V, the device acts as a high impedance, constant current regulator passing 1 µA/K. Laser trimming of the chip's thin-film resistors is used to calibrate the device to 298.2 µA output at 298.2 K (25°C).

The AD590 should be used in any temperature-sensing application below 150°C in which conventional electrical temperature sensors are currently employed. The inherent low cost of a monolithic integrated circuit combined with the elimination of support circuitry makes the AD590 an attractive alternative for many temperature measurement situations. Linearization circuitry, precision voltage amplifiers, resistance measuring circuitry, and cold junction compensation are not needed in applying the AD590.

In addition to temperature measurement, applications include temperature compensation or correction of discrete components, biasing proportional to absolute temperature, flow rate measurement, level detection of fluids and anemometry. The AD590 is available in chip form, making it suitable for hybrid circuits and fast temperature measurements in protected environments.

The AD590 is particularly useful in remote sensing applications. The device is insensitive to voltage drops over long lines due to its high impedance current output. Any well-insulated twisted pair is sufficient for operation at hundreds of feet from the receiving circuitry. The output characteristics also make the AD590 easy to multiplex: the current can be switched by a CMOS multiplexer, or the supply voltage can be switched by a logic gate output.

PIN CONFIGURATIONS

Figure 1. 2-Lead CQFP

Figure 2. 8-Lead SOIC

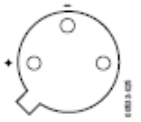


Figure 3. 3-Pin TO-52

PRODUCT HIGHLIGHTS

1. The AD590 is a calibrated, 2-terminal temperature sensor requiring only a dc voltage supply (4 V to 30 V). Costly transmitters, filters, lead wire compensation, and linearization circuits are all unnecessary in applying the device.
2. State-of-the-art laser trimming at the wafer level in conjunction with extensive final testing ensures that AD590 units are easily interchangeable.
3. Superior interface rejection occurs because the output is a current rather than a voltage. In addition, power requirements are low (1.5 mW @ 5 V @ 25°C). These features make the AD590 easy to apply as a remote sensor.
4. The high output impedance (>10 MΩ) provides excellent rejection of supply voltage drift and ripple. For instance, changing the power supply from 5 V to 10 V results in only a 1 µA maximum current change, or 1°C equivalent error.
5. The AD590 is electrically durable: it withstands a forward voltage of up to 44 V and a reverse voltage of 20 V. Therefore, supply irregularities or pin reversal does not damage the device.

One Technology Way, P.O. Box 9106, Norwood, MA 02062-9106, U.S.A.
Tel: 781.329.4700 www.analog.com
Fax: 781.461.3113

Figure 1.3: AD590 Data Sheet. See http://www.analog.com/static/imported-files/data_sheets/AD590.pdf

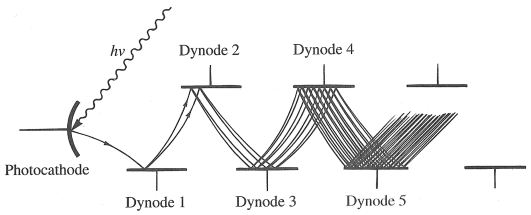


Figure 1.4: Principle of the PMT (from Diefenderfer and Holton)

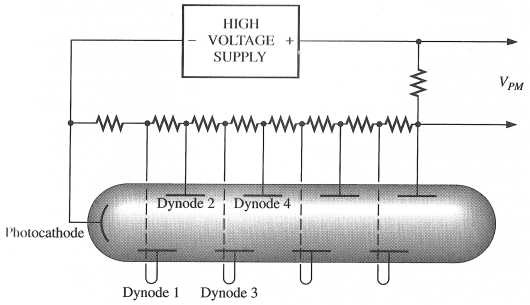


Figure 1.5: Schematic of a PMT Setup (from Diefenderfer and Holton)

Figure 1.6: A Selection of Photomultiplier Tubes available from Hamamatsu

extreme low-light levels (down to almost the single photon). Manufacturers such as Hamamatsu (http://www.hamamatsu.com/) have developed devices specifically for the physics community. At the extremes of measurement science we need to understand the physics of how the detector works in order to get the best out of it, which is why we'll look into the principle of the PMT.

The operation of the device is illustrated in Figure 1.4. A photon striking the photo-cathode liberates an electron. The first "dynode" is held more positive than the cathode so the liberated electron is accelerated towards it and on striking the dynode 2 electrons are released. Each subsequent dynode is progressively more positive so (in this example) we get an electron multiplication of 2^n where n is the number of dynodes.

Figure 1.5 shows a typical setup for the PMT. at the end of the tube the electrons are collected at a final anode and, flowing back to the power supply through a resistor, generate a voltage signal, the amplitude of which is proportional to the number of photons detected.

PMTs are not without their difficulties. A large (and dangerous) high-voltage supply is needed (several kV can be applied across the tube). Even when there is no incident light, thermal emission of electrons from the photo-cathode means that a current always flows in the tube. This called the **Dark Current** and we want to minimise it since it adds noise and reduces the dynamic range of our measurement (these are both topics which will be discussed in more detail later on). One way to do this is to cool down the PMT to reduce the thermal emission.

The **Quantum Efficiency** of the photocathode is important. This is the percentage of incident photons which generate an electron emission. 25% is a typical number but specialist tubes can have much higher figures. A selection of the tubes available from Hamamatsu is shown in figure 1.6

Further Reading Diefenderfer and Holton chapter 7 is good on the general topic of transducers and has more details of the temperature transducers and the PMT. Horowitz and Hill Chapter 15 is also very good though it will make more sense once we've covered some of the electronics involved. For the PMT there is a lot of very good background and technical information on the Hamamatsu website at http://www.hamamatsu.com/.

2 Signals

Generally, a signal is any time or spatially-varying quantity. We are used to time-varying quantities such as voltage, current, light intensity and sound pressure. A spatially varying signal might be, for example, altitude measured as a function of distance. Typically a signal represents something physical.

2.1 Continuous and Discrete Signals

Both the sound pressure and its electrical representation are signals in their own right, and both these types of signals are continuous functions of time. Generally, any physical quantity we may wish to deal with is represented by a continuous parameter such as mass, length, voltage, current etc. However, the very act of measurement creates a discrete signal. Each time we make a measurement of a quantity we are generating a discrete signal. For example, the height of the sea measured every hour, or the temperature in a reaction-chamber measured every second. These are examples of sampled signals. In physics and engineering we encounter mainly signals which have been both sampled and digitised; a subject we will come in to later. For now it is sufficient to know that a discrete signal is one where the continuous quantity has been sampled at regular intervals to create a discrete set or list of values. In the rest of this section we will be looking at continuous signals, however it is important to always bear in mind that a discrete equivalent exists.

2.2 Periodic Continuous Signals

We will need mathematical descriptions of continuous signals and to see how this is done it is helpful to consider some of the 'standard' signals prevalent in physics. Figure 2.1 shows six common waveforms. The simplest is the last (f) which is the sinusoid. This might represent for example the position of a mass oscillating with simple harmonic motion. Figure 2.1 shows the time and frequency-domain representations. For the latter, the relative amplitudes of the first six Fourier coefficients are given (Note that a separate handout on Fourier will come along in due course). By convention f is the fundamental frequency, and $2f$, $3f$ are the 2nd and 3rd harmonics, and so-on. Our sinusoid of course has only the fundamental, but the rectified version (e) has both a 'DC' or constant component at zero Hz, and also lots of higher harmonics which are due to the sharp corners in the waveform. Sawtooth (d), Triangle (c) and Square (b) waveforms are frequently seen in electronics. The triangle and square-waves have the characteristic that all the even harmonics are zero. However, all three have the characteristic that their Fourier series is infinite, i.e. an infinite number of harmonics is required to accurately represent the waveform.

2.2.1 Representation by Piecewise Continuous Functions

In the time-domain, we would like to have convenient analytic functions to describe our signals. An analytic function is one where both the function and its derivative are finite over the region of interest. The sinusoid is of course easy, but all of the others are non-analytic due to the discontinuities. The solution is to represent the signal as a piecewise continuous function. The first step is to state that for any periodic function

$$f(t+T) = f(T) \tag{2.1}$$

where T is the period. Then for the square wave

$$f(t) = \begin{cases} \frac{A}{2}, & |t| < \frac{T}{2} \\ \frac{-A}{2}, & \frac{T}{2} < |t| < T \end{cases} \tag{2.2}$$

This method provides a representation for any of these waveforms; they are non-analytic, but functionally useful and can be handled mathematically as we shall see.

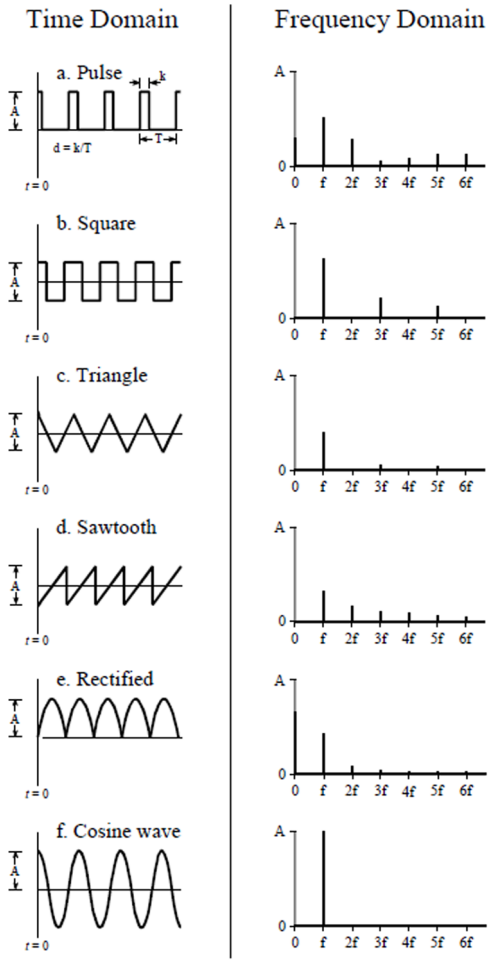


Figure 2.1: Waveforms frequently encountered in physics or engineering

2.3 Non-periodic Continuous Signals

There are a number of non-periodic functions which are particularly relevant to instrumentation due to their special properties. We'll review these here.

2.3.1 Unit Step Function

This is important both analytically and practically. Whenever we close a switch a step-voltage is applied; we can have a step change in applied force, temperature, etc.

$$u(t) = \begin{cases} 1, & t > 0 \\ 0, & t < 0 \end{cases} \tag{2.3}$$

Mathematically it is unsatisfactory as it is undefined at $t = 0$, however in reality we know that the change in voltage, force, temperature etc does not truly occur instantaneously. The unit step is the ideal, mathematical representation, and the discontinuity can be handled within the mathematical framework we will use. We can scale and time-shift the unit step. As an exercise try drawing $7u(t-6)$[1].

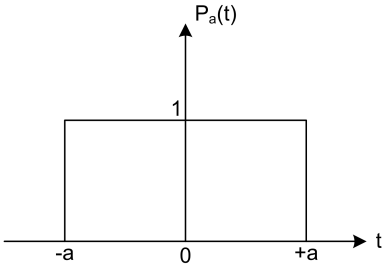


Figure 2.2: Rectangular Pulse Function

2.3.2 Rectangular Pulse Function

The unit pulse function (Figure 2.2) has width $2a$ and height 1. It can be scaled e.g. $hP_a(t)$ which may represent a voltage h applied for time $2a$.

$$P_a(t) = \begin{cases} 0, & |t| > a \\ 1, & |t| < a \end{cases} \tag{2.4}$$

One of the recurring themes in this course will be the idea that we can understand the behaviour of complex systems by looking at the response to 'simple' waveforms (such as the unit step) and then deduce what the system would do in response to combinations of simple waveforms. With this in mind, consider that the pulse can be built from a superposition of step functions

[1] See Poularikas Figure 1.5

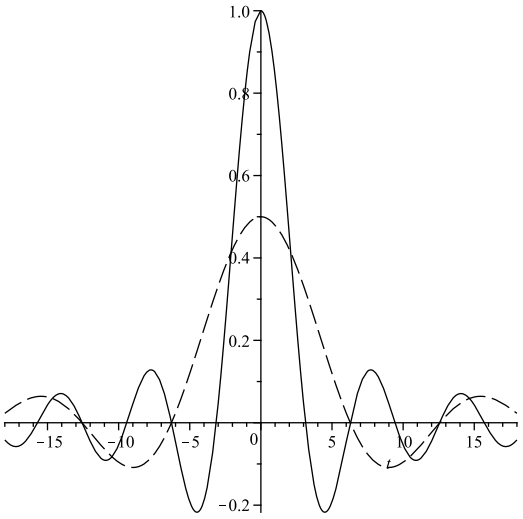


Figure 2.3: The Sinc function for $a = 1$ (solid line) and $a = \frac{1}{2}$ (dashed line)

$$P_a(t) = u(t+a) - u(t-a)$$

Also, we can multiply functions together. We can make a function zero for all time $t < 0$ by multiplying with the unit step. For example we can specify a non-symmetric pulse of width a by writing $u(t)P_a(t)$

2.3.3 Sinc Function

$$\text{sinc}_a(t) = \frac{\sin at}{t} \tag{2.5}$$

The sinc function is shown in Figure 2.3 for two values of a. It is important as this shape is (in the frequency domain) the Fourier transform of the Pulse function.

2.3.4 Delta Function

Occupies a central place in physics and signal analysis, the delta function can represent point sources, point charges, a concentrated force, or a voltage/current acting for a very short time. We'll define the delta function by its behaviour under integration

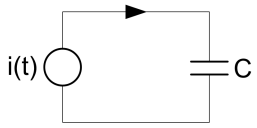


Figure 2.4: Charging a Capacitor

$$\delta(t) = 0 \qquad t \neq 0$$

$$\int_{-\infty}^{\infty} \delta(t) = 1$$

The delta function 'samples' the value of a function at any chosen time t_0

$$\int_{-\infty}^{\infty} f(t)\delta(t - t_0) = f(t_0)$$

The delta function can be considered the limiting case (or mathematical ideal) of many physical processes. For example, consider the circuit in Figure 2.4 where the current source can be programmed to generate short current pulses of duration ϵ and amplitude $1/\epsilon$

$$i(t) = \frac{1}{\epsilon} \quad 0 < t < \epsilon$$

As an exercise, plot the current for $\epsilon = 1, 1/2, 1/4$

From the relations $q = Cv$ and $i = \frac{dq}{dt}$ we can get the voltage as a function of time

$$v(t) = \frac{1}{C}\int_0^{\epsilon} \frac{1}{\epsilon} dt$$

Let's assume $C = 1\text{F}$ and ϵ is in seconds then

$$v(t) = \int_0^{\epsilon} \frac{1}{\epsilon} dt = \int_0^{\frac{\epsilon}{2}} \frac{2}{\epsilon} dt = \int_0^{\frac{\epsilon}{4}} \frac{4}{\epsilon} dt = 1$$

The area under the curve will always be 1. Further, as $\epsilon \to 0$ and the current pulse approaches a delta function, the voltage on the capacitor approaches a 1V step function[2]. There are two important lessons from this:

[2] See Poularikas Figure 1.9

Electrical	Mechanical	Comment
Resistor	Damper	Dissipates energy
Capacitor	Mass	Stores energy
Inductor	Spring	Stores energy
Current	Force	
Voltage	Velocity	

Table 2.1: Analogy between mechanical and electrical systems

1. An impulse of current on a capacitor produces a step change in voltage
2. A capacitor is a current integrator

Note that we could have modelled the same system as a force acting on a mass. A unit impulse force acting on a mass m results in a velocity $1/m$. Here we have the first analogy between electrical and mechanical systems. Capacitance is to the electrical system what mass is to a mechanical system; both might be said to provide 'inertia' to the system, and are capable of storing energy. The mass stores potential energy and releases it as kinetic energy, the capacitor stores energy due to the electric field across the device, and releases it as a flow of current. We'll see these kind of analogies many times. For now, it is sufficient to know that electrical circuits can be used to mimic, or model, the behaviour of mechanical, thermal, fluid, and other physical systems (see Table 2.1). This is useful in its own right, and since much of instrumentation is based on electronics it makes sense to study electrical systems in some detail.

2.3.5 Gaussian

The familiar 'bell-shaped' Gaussian is a continuous function generated from a negative squared exponent $f(t) = Ae^{-at^2}$ (figure 2.5).

Adding a parameter μ allows us to slide the curve along the x-axis, and normalising the area under the curve to 1, gives us the familiar Gaussian form

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \qquad (2.6)$$

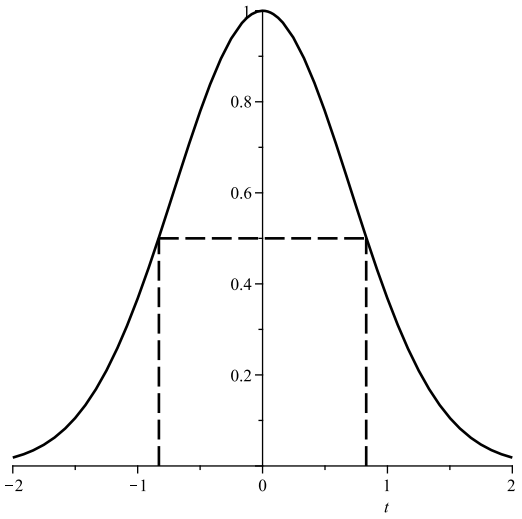


Figure 2.5: Gaussian pulse $f(t) = e^{-t^2}$, with pulse width identified as 'Full-Width at Half-Maximum' (FWHM)

The pulse width (FWHM) is always $2\sqrt{2\ln 2}\sigma$. The shape of fast pulses travelling in a medium such as a cable or an optical fibre is often described as being 'approximately Gaussian'. $f(t)$ may represent the voltage of a pulse travelling down a cable. Alternatively $f(t)$ could represent the electric field component of a light-pulse in an optical fibre, in which case we would typically use the pulse *intensity* $f(t)^2$ as this is the physical observable. As well as being physically realistic , this is also quite convenient:

Gaussian-squared. The square of a Gaussian is another Gaussian, so if we say that the signal amplitude is Gaussian-shaped then so is the signal power

Fourier Transform. The FT of a Gaussian is another Gaussian. This makes it quite a convenient signal shape to study.

3 DC Circuits

DC is short for Direct Current, which refers to a static or constant current flowing in a circuit. AC for Alternating Current refers to a (usually sinusoidal) varying current in a circuit. A DC current is what you get from a battery, while the mains provides AC at 50 Hz. In physics, and certainly in engineering, these terms are used rather loosely to mean any signal which is either constant or varying. For example we might talk about a DC voltage of 1.5V (from an AA battery), or $240V_{AC}$ (mains voltage). We might even describe the intensity of radiation from a Pulsar as an 'AC signal', or the weight of a kilogram of lead as a 'DC signal'.

In this section we'll review some of the basic concepts in DC circuits, i.e. ones where a constant current flows. We'll go back to some very basic concepts; this should be

- a refresher (in that you will have seen all of this before in the first year electronics course)
- an aide-memoire (summarising the basic concepts and equations)

3.1 Current

Current is rate of flow of charge

$$i = \frac{dq}{dt}$$

The SI unit is the amp, with one ampere (A) being equivalent to 1 coulomb (C) per second.

$$1\,A \equiv 1\,C/s$$

Note that by convention we use the letter i to represent current, but it can appear as either lower or upper-case. Our charge carriers are of course electrons, and since there are plenty of free electrons in metals it makes sense to use metal (usually Copper) to connect together the parts of our circuits, but generally the wires play no part of in the circuit other than as conduits for electrons, or more specifically to connect different parts of the circuit to the same potential. Refer to Figure 3.1, about the simplest circuit possible, it may represent a torch. A bulb is effectively a resistor; when current flows it gets so hot it glows incandescently. The battery provides a potential difference across its terminals. By convention the positive terminal is the larger 'plate'.

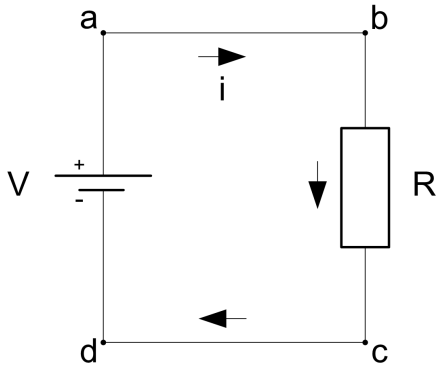


Figure 3.1: Conventional Current

3.1.1 Conventional Current

Electrons are attracted to the positive terminal and hence flow anti-clockwise, however conventional current is taken to flow in the opposite direction to actual electron current, and hence is defined as flowing from positive to negative potential. We might think of conventional current as flow of positive charge or 'holes'. We - like most textbooks - will always use conventional current, even when we are talking about beams of electrons flowing in a vacuum tube.

3.1.2 Conservation of Charge

The total number of electrons (or holes) in the circuit is constant. This is a statement of conservation of charge. Consequently, the current i at point a is the same as that at b, the same as that flowing through the resistor, to c, d and back through the battery. This is an important concept, always true, and for AC circuits too.

3.2 Resistance

The sheer abundance of conduction electrons in the Copper wires means that appreciable currents will flow with a net electron flow speed of the order 10^{-6}m/s. This is the *drift velocity* of the electrons. Contrast this with the average thermal speed of the electrons at room temperature which is about 10^{6}m/s. The overall behaviour of the electrons is therefore rather random and collision-dominated, and we must see current as a statistical average drift of electrons rather than an orderly procession. Because the electrons are in constant collision with each other and the atomic lattice of the conductor, the mobility of the electrons is key to their behaviour. Mobility is high in metals, but in Carbon it is about a thousand times lower. Clearly then, if we make a resistor from an insulating tube covered with a thin film of carbon then mobility will be much reduced and current will be severely restricted. Ohm's law gives

$$i = \frac{V}{R} \tag{3.1}$$

3.3 Potential

A difference in electrical potential causes current to flow, and of course when a charge moves through a potential differemce work is done. The SI unit of potential difference is the volt, being one joule per coulomb

$$1\,\mathrm{V} \equiv 1\,\mathrm{J/C}$$

Here, the battery does the work, converting stored chemical energy into current. When measuring PD it is important to understand the *sense* of the measurement. If our battery is a single AA then $V_{ad} = 1.5\,\mathrm{V}$. We would measure this with a meter by placing the red probe at a and the black probe at d. If we reversed the probe we would measure $V_{da} = -1.5\,\mathrm{V}$. $V_{ab} = V_{cd} = 0$ because the wires have zero resistance (ideally). Therefore $V_{bc} = V_{ad}$. Since we know the PD *across* the resistor we can calculate the current *through* it using Equation 3.1.

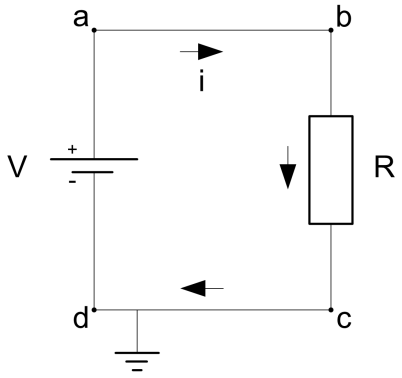


Figure 3.2: Grounded circuit establishes zero reference for potential

3.4 Kirchoff's Voltage Law

States that the sum of the PDs around any loop is zero

$$\sum v = 0 \tag{3.2}$$

Imagine sitting in the centre of the circuit with the meter probes and measuring each of the four sides in turn, going round clockwise and taking car to keep the sense of the measurement the same as you go.

$$V_{ad} + V_{ba} + V_{cb} + V_{dc} = 0$$

This is of course trivial, but there could be any number of components filling up the top and bottom sides. Since we know the current is the same in each, we can calculate the PD across each. This works for all components, i.e. including inductors, capacitors etc, and is suitable for AC circuits too.

3.5 Ground

PD is a difference measurement. It gives us the difference in potential between two points. What is the actual potential at a? or d? The answer is that we don't know as there is no reference potential in this circuit. It is said to be

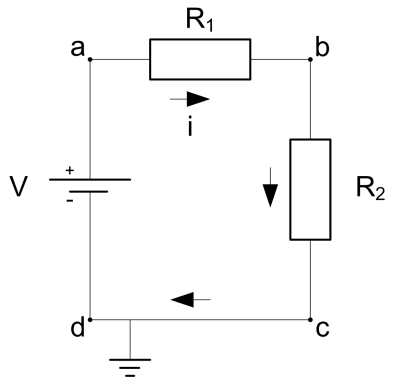


Figure 3.3: Series resistance

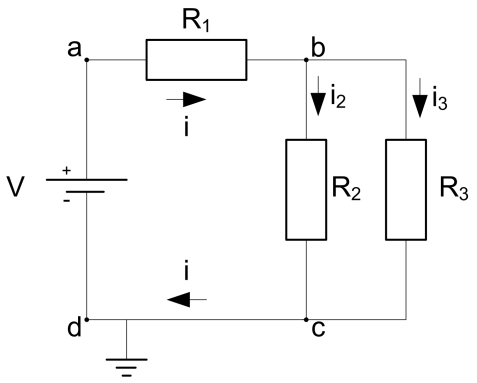


Figure 3.4: Parallel Resistors

floating. Any battery-powered kit from a multimeter to a laptop is floating. We could connect the circuit to ground as in Figure 3.2 then we would know that the potential at d (and c) is zero volts. Ground (the Earth pin on a plug) is genuinely connected physically to a large copper spike in the basement of the building driven into the earth. This is taken by convention as the zero reference for electrical potential. Now

$$V_a = 1.5\,\mathrm{V}$$

Scientists and engineers use voltage and potential difference interchangeably. This is wrong but in such common use as to be unavoidable. We might often say "the voltage across R is 1.5V" when we mean "the potential difference..."

3.6 Series Resistors and the Voltage Divider

Resistors in series add, i.e. $R = R_1 + R_2$

In Figure 3.3 we can use Ohm's law to state $V_{ab} = iR_1$. Note the sense of the subscript. Ohm's law gives us the voltage drop from a to b when a→b is the direction of conventional current. So a is at a higher potential than b (which Kirchoff confirms) and hence

$$V_b = V_a - iR_1$$

We find the current flowing in both resistors as

$$i = \frac{V_{ac}}{R_1 + R_2}$$

Combining the two above equations and noting that $V_{ac} = V_a$ we get the Voltage Divider equation

$$V_b = \frac{V_a R_1}{R_1 + R_2} \tag{3.3}$$

3.7 Parallel Resistors and the Current Divider

For resistors in parallel such as Figure 3.4 use

$$\frac{1}{R_{||}} = \frac{1}{R_2} + \frac{1}{R_3}$$

We can calculate the current as

$$i = \frac{V_a}{R_1 + R_{||}}$$

We can calculate V_b since R_1 and $R_{||}$ are a voltage divider. Then

$$i_2 = \frac{V_b}{R_2}, \qquad i_3 = \frac{V_b}{R_3} \tag{3.4}$$

We would find that the Current Divider rule is

$$i_2 = \frac{iR_3}{R_2 + R_3}$$

Note the similarities and differences with the Voltage Divider rule. Generally this can get confusing and it's easier to just remember the voltage divider rule and then use the method of equations 3.4 to get the currents in the branches.

3.8 Kirchoff's Current Law

Conservation of charge tells us that the current going into point b must equal the current coming out, so $i = i_1 + i_2$. This is enshrined in Kirchoff's Current Law which states that the sum of all the currents *into* a node (junction) is zero

$$\sum i = 0$$

To draw this we would change the direction of the i_2 and i_3 arrows *which also changes their algebraic sign* and then indeed the sum *into* b is zero.

3.9 General Method for Circuit Analysis

The above rules give us everything we need for the vast majority of circuit analysis. Usually we are given as a starting point some knowledge of the current or some potentials. As a general approach

- If you know the current in any branch of a circuit then calculate the PD across any components and then determine the potentials (using Ohm and KVL) of any branching points in the circuit.
- Conversely if you are given the potential relative to ground at any point then try to calculate the current in that branch and use a combination of Ohm and KCL to deduce any currents.

3.10 Voltage and Current Sources

An ideal voltage source provides a fixed voltage across its terminals which is independent of the load resistance to which it is connected (and hence the amount of current drawn). It appears to have zero resistance, in that it dissipates no power. In reality, there is a limit to how much current a voltage source can supply, and there is always some internal resistance. A battery is a good voltage source example: a 9V battery connected to a 3 Ohm load will supply a full 3A (for a few minutes). The battery will get hot due to it's non-zero internal resistance (less than an Ohm). The load resistor will get very hot! On circuit diagrams you'll often see the battery symbol as used here or sometimes just a circle with the "V" symbol.

An ideal current source provides a fixed current which is indepenedent of the load resistance to which it is connected (and the amount of voltage it must generate). In reality there is a limit to how much voltage a current source can generate. There is no simple physical example of a current source; often they are quite complicated and expensive pieces of lab equipment. The standard circuit symbol is two part-overlapping circles, though quite often you'll see just a single circle marked "I", so take care to distinguish these from voltage sources.

3.11 Thevenin Equivalent Circuit

As a final remark on DC circuits, Thevenin's Theorem is frequently used in instrumentation since the Thevenin Equivalent Circuit is a convenient way of encapsulating the behaviour of more complex circuits which may represent, for example, a sensor. Thevenin states that

Any network of many sources and impedances can be replaced by a single source in series with a single impedance

Sources are voltage sources such as in Figure 3.5, or current sources. Here we consider resistances but the method works also for generalised impedances as we shall see later.

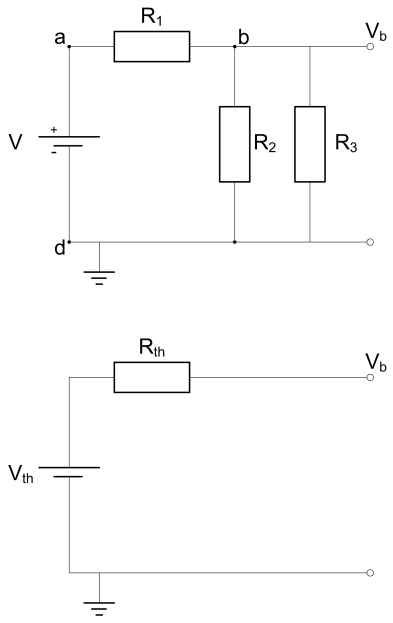


Figure 3.5: Thevenin Equivalent Circuit

Thevenin's Theorem tells us that the top and bottom circuits in figure 3.5 are exactly equivalent. If the two circuits were inside a "black-box" with just the two terminals on the outside then nothing we could do from the outside could tell the difference. To get the Thevenin equivalent circuit we just need to determine - by calculation or measurement - the values of V_{th} and R_{th}.

Thevenin Voltage V_{th} is the open-circuit PD across the terminals. Open circuit means that no current is drawn down the terminal branches, i.e. when there is no 'load' attached to the circuit. For example if we measure the voltage at b with a multi-meter then it has an input impedance of MΩ so, as long as $R_n \sim k\Omega$ we can confidently assert that the measured voltage is very close to the open-circuit value. The meter hardly 'loads' the circuit at all. In this case, we have already calculated V_{th} since it is equal to V_b

Thevenin Resistance R_{th} is the combined impedance (resistance) when

1. Any voltage sources are short-circuited and
2. Any current sources are open-circuited

We short-circuit the battery by removing it and connecting across points a and d. Then all three resistors are in parallel so

$$\frac{1}{R_{th}} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}$$

This is the resistance we would measure if we connected a resistance meter across the terminals of the circuit.

The point of the Thevenin circuit is that it behaves *exactly* like the original, however it is easier to see how it behaves under different load conditions. Connecting a high impedance meter across the terminals we measure V_{th}, since no current flows in R_{th}, however were we to connect a low-impedance load (say a light bulb) then it is easy to see how the current flowing means there is a voltage drop across R_{th}. Issues of input and output impedance are very important and we will come back to this later.

In the Thevenin equivalent circuit, V_{th} is an idealised voltage source, and R_{th} is the apparent series (internal) resistance of the circuit. As an example, the Thevenin equivalent of the battery mentioned in section 3.10 would be a voltage source of 9V in series with a resistance of about $\frac{1}{4}\Omega$

Further Reading There are very readable sections on DC circuits in Diefenderfer (Chapter 1) and Horowitz and Hill (Chapter 1)

4 AC Circuits

As previously mentioned, electronic circuits are good analogues for other physical systems (see Table 2.1). Dynamical systems are represented by circuits processing time-varying signals. Further, we need to understand how AC circuits can manipulate electrical signals from sensor systems. We introduced the capacitor in 2.3.4, and will now look at AC components and signals in general.

4.1 Capacitor

Essentially two parallel metal plates with an insulating dielectric in-between, the capacitor has ideally infitite ohmic resistance to DC current. When a PD is applied it stores energy in the electric field. By definition the capacitance is measured in farad (coulomb per volt)

$$C = \frac{q}{V} \tag{4.1}$$

From which we obtain the key relations

$$V(t) = \frac{1}{C}\int_{-\infty}^{t} I(\tau)d\tau \tag{4.2}$$

$$I(t) = C\frac{dV(t)}{dt} \tag{4.3}$$

See also Figure 2.4. Note that for AC signals the current, integrated over one cycle, leaves no net charge on a capacitor, and hence $\langle V(t)\rangle = 0$. See Poularikas Example 2.1. Capacitors in parallel add together (i.e. opposite to resistor behaviour) which is logical when we consider that 2 caps in parallel are the same as a single cap with plates twice as large[3].

As a consequence of the above equations, when we apply a sinusoidal voltage to a capacitor, the current through the capacitor leads the voltage across it by 90°.

[3] For further information see Diefenderfer 2.2 & 2.3

4.2 Inductor

An inductor is essentially a coil of wire (or a solenoid), usually wound around a core of high magnetic permeability such as iron or ferrite to increase the inductance (in henry, or weber per amp)

$$L = \frac{\psi}{I} \tag{4.4}$$

Using Faraday's law $V(t) = d\psi/dt$ we obtain the relations

$$I(t) = \frac{1}{L}\int_{-\infty}^{t} V(\tau)d\tau \tag{4.5}$$

$$V(t) = L\frac{dI(t)}{dt} \tag{4.6}$$

As a consequence, when we apply a sinusoidal voltage to an inductor, the current through the inductor lags the voltage across it by 90°.

4.3 Complex Representation of Signals

Consider a voltage $V(t) = V_0 \cos\omega t$ driving a capacitor C. Then the current which flows in the capacitor is

$$I(t) = C\frac{dV(t)}{dt} = -C\omega V_0 \sin\omega t \tag{4.7}$$

Clearly there is a voltage/current phase difference of $\pi/2$. It's frequently more convenient to use a complex representation. A voltage $V(t) = V_0\cos(\omega t + \phi)$ is represented by the complex number $V = V_0 e^{j\phi}$. This represents the voltage as a vector in the complex plane. There are a couple of points to note:

1. The signal (voltage) is of course a real quantity and is represented by the real part of the complex quantity
2. Engineers typically use j instead of i to avoid confusion with current

3. In most situations we are just concerned with the relative amplitude and phase of the various voltages in the circuit so the time-varying part of the representation $e^{j\omega t}$ is taken for granted.

To get the actual voltage signal, multiply by $e^{j\omega t}$and take the real part

$$V(t) = \mathrm{Re}\{Ve^{j\omega t}\} = V_0\cos(\omega t + \phi) \qquad (4.8)$$

4.4 Phasor Representation

Consider the vector $\underline{P} = (a + jb)$ plotted on the Argand diagram (figure 4.1). $a = P\cos\theta$, $b = P\sin\theta$ and

$$P = \sqrt{a^2 + b^2} = \sqrt{\underline{P}\underline{P}^*}$$

Using the Euler relations we see that the vector is represented as $Pe^{j\theta}$. If we advance the vector by a phase of $\pi/2$ we get $Pe^{j\theta+\pi/2}$ and if we advance by π then we get $Pe^{j\theta+\pi/2} = -Pe^{j\theta}$ from which we can see that $e^{j\pi} = -1$. Taking the square-root of both sides $e^{j\pi/2} = \sqrt{-1}$. Consequently, multiplication by j represents an anti-clockwise rotation by $\pi/2$. The relevance of this to AC circuits comes from the fact that a multiplication by $e^{j\theta}$ represents a rotation θ in the complex plane. In general then any sinusoidal physical quantity (voltage, current etc.) can be thought of as a vector rotating anti-clockwise. The Argand diagram is a 'snapshot' e.g. at time $t = 0$.

4.5 Reactance

Returning to the capacitor example of 4.3 we can see

$$V(t) = V_0\cos(\omega t + \phi) = \mathrm{Re}\{Ve^{j\omega t}\}$$

$$I(t) = C\frac{dV(t)}{dt} \quad = -C\omega V_0\sin\omega t \quad = \mathrm{Re}\left\{\frac{Ve^{j\omega t}}{-\frac{j}{\omega C}}\right\}$$

$$= \mathrm{Re}\left\{\frac{Ve^{j\omega t}}{X_c}\right\}$$

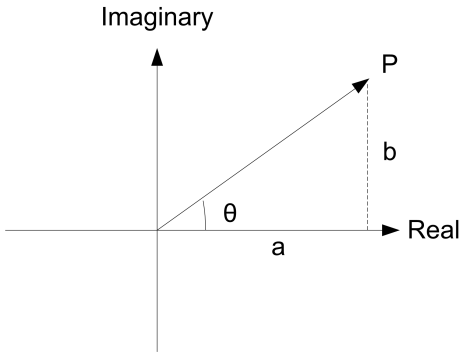


Figure 4.1: Phasor Representation on the Argand Diagram

Where X_c is the Reactance of the capacitor. Similarly we define the reactance of the inductor X_L. The reactance is the AC impedance of the circuit element.

$$X_C = \frac{-j}{\omega C} \qquad (4.9)$$

$$X_L = j\omega L \qquad (4.10)$$

Clearly, inductors and capacitors have purely imaginary impedance, which gives them the quality that they induce a $\pi/2$ phase shift between voltage and current and hence dissipate no power. The resistor is the opposite: it has purely real resistance, dissipates power, and the voltage is always in phase with the current.

4.6 Complex Impedance

The complex impedance Z of any circuit element or combination of elements is the vector combination of the resistive and reactive components according to the same rules as pure resistances

$$Z_{series} = Z_1 + Z_2 \qquad (4.11)$$

$$\frac{1}{Z_{||}} = \frac{1}{Z_1} + \frac{1}{Z_2} \qquad (4.12)$$

We can visualise this with the Argand diagram. If we have a resistor in series with an inductor

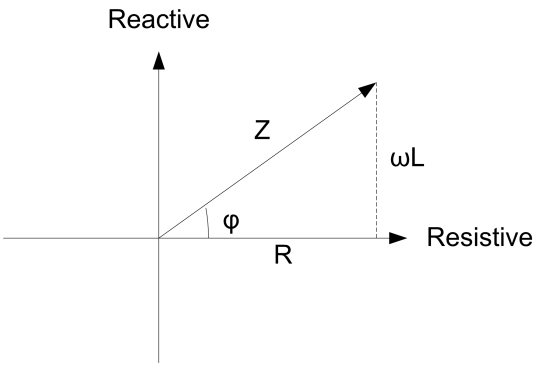


Figure 4.2: Complex impedance on the argand diagram

then $Z = R + j\omega L$ where $|Z| = \sqrt{R^2 + \omega^2 L^2}$ and $\phi = \tan^{-1}\left(\frac{\omega L}{R}\right)$

Figure 4.2 requires some interpretation. The Argand diagram gives us the relative phases of the voltages in our circuit. The voltage across the inductor (positive imaginary axis) leads the voltage across the resistor (real axis) by $\pi/2$ (Note: this is consistent with the "V leads I" rule for inductors; V is in phase with I for the resistor). Let us say that we apply a voltage $V(t) = V_0 \cos(\omega t + \phi) = \mathrm{Re}\{Ve^{j\omega t}\}$. Then the Argand diagram gives us a 'snapshot' of the voltages at time $t = 0$. Actual voltages (as would be measured with the 'scope) are projections onto the real axis. The voltage across the resistor is at its peak value (phase=0). The voltage across the inductor is zero (phase=$\pi/2$), and the applied voltage (across both) is $V_0 \cos\phi$.

Note that at all times Kirchoff's Voltage Law applies: the instantaneous applied voltage always equals the sum of the instantaneous voltages on the resistor and inductor.

4.7 Ohm's Law Generalised

The capacitive and inductive reactances have units of Ohms (Ω). Ohm's law is applicable to impedances generally.

$$V = IZ \tag{4.13}$$

Where all the quantities are complex. Note however that only V and I can be represented as phasors (time-varying complex quantities). The impedance Z is a complex *constant*. This simultaneously illustrates the usefulness and the limtation of the phasor representation: If we know the phasor current I we can multiply it by the by the complex impedance Z to get the phasor voltage V

$$V_0 e^{j\phi} = I_0 e^{j\psi} Z_0 e^{j\theta}$$

where the amplitudes obey $V_0 = I_0 Z_0$ and the phases $\phi = \psi + \theta$. This is much easier than solving problems using the triginometic forms. However we can never multiply two phasors together - that would represent the product of two sinusoids which would give us new frequency components! In practice, the use of phasors is limited to situations like equation 4.13 where a complex constant modifies the amplitude and phase of a phasor. This does turn out to be incredibly useful and will be the basis later on for much of our work on *linear systems*. This also means that phasor notation can only represent systems at a given frequency (that is to say, while Z is a function of frequency, for any I we get a V of the *same* frequency).

Reactive components have two important properties which we will see the use of in future sections:

1. Phase shifting (as discussed above)

2. Frequency dependence. At DC (zero Hz) the reactance of the inductor is zero. It is a short-circuit. At high-frequencies, the reactance tends to ∞ (open circuit). The capacitor is the opposite way round. as we have seen before, it is open-circuit at DC.

Note that if we just want to work with the magnitudes of the voltages and currents (i.e. we don't care about the detailed phase relationships) then we can use $|Z| = |V|/|I|$, so for a capacitor we can calculate the *amplitude* of the current as $I_0 = V_0 \omega C$

4.8 Power in AC Circuits

If we have a sinusoidal voltage $V = V_0 \sin \omega t$ then, even though the average value of V is zero, the signal has the capacity to dissipate power when it is applied across a resistor R according to

$$P = VI = \frac{V^2}{R} = \frac{V_0^2 \sin^2 \omega t}{R} \tag{4.14}$$

This is the *instantaneous* power at any time t. More useful (see boxed example) is the *Average Power*

$$\langle P \rangle = \frac{V_0^2}{R} \langle \sin^2 \omega t \rangle = \frac{V_0^2}{2R} \tag{4.15}$$

The average[4] is taken over a cycle of ω.

4.8.1 Root Mean Square

The RMS or *Root Mean Square* of a Signal is a measure of the power in the signal. For our voltage V it is

$$V_{RMS} = \sqrt{\langle V^2 \rangle} \tag{4.16}$$

(*it does exactly what the name says!*) Hence for a sinusoidal signal

$$V_{RMS} = \frac{V_0}{\sqrt{2}} \tag{4.17}$$

The average power dissipated in a resistor is

$$\langle P \rangle = \frac{V_{RMS}^2}{R} = V_{RMS} I_{RMS} \tag{4.18}$$

The above is always true, regardless of the shape of the waveform. RMS is a useful measure since it gives the equivalent power (or heating effect) as the constant 'DC' voltage of the same value.

Example: The UK mains voltage is 240 volts RMS, the amplitude of the mains signal is about 340V, and the peak to peak voltage is 680V. A 60W light bulb (old-fashioned tungsten-type) will have a resistance of 960Ω

[4] To see why the average value of a squared-sine function is $\frac{1}{2}$ consider the identity $2\sin^2\theta = 1 - \cos 2\theta$. Over a cycle, $\langle \cos 2\theta \rangle = 0$

4.8.2 Power in Reactive Circuits

For circuits with inductance and/or capacitance the average power becomes

$$\langle P \rangle = V_{RMS} I_{RMS} \cos \phi \tag{4.19}$$

Where ϕ is the phase difference between V and I and $\cos\phi$ is often called the 'power factor'. To understand this consider again the capacitor. The voltage is $\pi/2$ out of phase with the current. The instantaneous power delivered to the capacitor is $P = VI$ (sketch this to visualise it). Positive power is when the capacitor is charging and negative power is discharging. Clearly, over a cycle the average power is zero. This is true for any circuit with purely imaginary impedance. Circuits with some real (resistive) component of impedance will dissipate power. In general we could determine average power by integrating over a cycle and dividing by the time for one cycle. This gives

$$\begin{aligned} \langle P \rangle &= \frac{\mathrm{Re}\{VI^*\}}{2} \\ &= \frac{\mathrm{Re}\{V^*I\}}{2} \\ &= V_{RMS} I_{RMS} \cos \phi \end{aligned}$$

5 Sampling and Digitization

Most signals we encounter in physics are continuous functions of time, such as voltage or light intensity. If we want a computer to interact with these variables we have to first sample and then digitize them. The key point to take from this is that the process of sampling quantises time while digitization quantises the parameter being measured. Ultimately, this means we lose information about the original signal, and we have to be very careful in how we interpret digitized signals. This section will outline some of the basics of this process.

Note that sampling and digitization is not entirely a phenomenon of the computer-age. Taking a temperature reading (every hour, or 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 *digitization*.

Note also that in the following text we will follow the approach given in Chapter 3 of the book "The Scientist and Engineer's Guide to Digital Signal Processing" by Steven Smith. This book is excellent, very readable and provides a good descriptive text though not always with as much mathematical backing as we might like. It is also available online[5].

5.1 Sampled Signals

As suggested above, sampling and digitization is a two-stage process, as illustrated in the central "block 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. Now, usually we want a series of samples with uniform spacing in time. If the time between samples is T_s then we can *define* a sampled version of the function as

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

[5] www.dspguide.com

This says we take the function $f(t)$ 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 $t = nT_s$.

5.2 Sample and Hold

In practice this mathematical representation is very far from reality, since

1. 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 finite time to take a sample.

2. As per figure 5.1, we want to pass the sampled value into the next block which performs the digitization. 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 circuit for the sample and hold is relatively straightforward, but we'll 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.

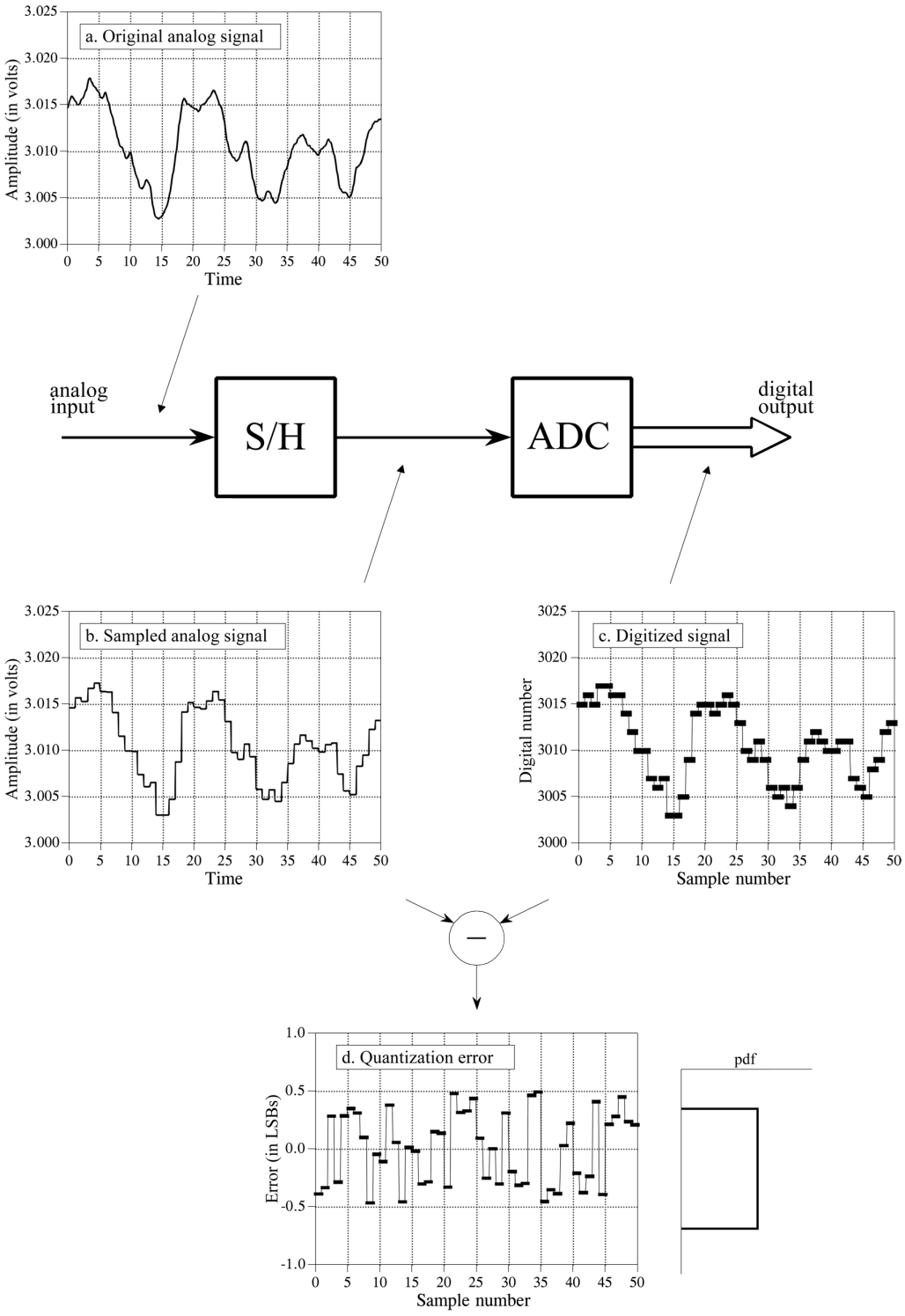


Figure 5.1: Waveforms illustrating the two-stage process of Sampling and Digitization (from Smith, www.dspguide.com)

5.3 Digitization

In figure 5.1 we can assume that the input signal can vary from 0 to 4.095 volts which the **Analogue to Digital Converter** (ADC) will convert 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'll come onto soon. The ADC is a rather more complicated circuit than the sample and hold, and again we'll look at this later.

5.4 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 examples. Digital electronics uses binary representation of numbers since this fits nicely with the fundamentally two-state "ON=+5V" and "OFF=0V" nature of digital logic[6].

To some extent the "number of bits" is a measure of precision or at least 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 (rightmost binary digit) is 0.25 mV.

[6] See Horowitz and Hill chapter 8

Binary	Decimal
000000000000	0000
000000000001	0001
000000001110	0014
111111111111	4095

Table 5.1: Some 12-bit Binary Numbers and their Decimal Equivalents

5.5 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 uniformly distributed between ± 0.5 LSB. This means we can define the noise statistically according to its mean μ and standard deviation σ

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

Now, for distributions with zero mean, the RMS is the same as the standard deviation[7]. If we want to reduce the RMS noise in our digitized signal, we need to "*improve the resolution of the measurement by increasing the number of bits*"

Example: Quantisation Noise as Percentage of Full-Scale Input Passing an analogue signal through an 8-bit ADC with a adds an RMS noise of $1/\sqrt{12}$LSB. Since the full-range (maximum) input is 2^8, 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 digitization simply adds in quadrature with any noise already existing in the analogue signal (and as we shall see later, there always is *some* noise in any signal).

[7] Look up the definitions to check this if you are unsure

Example: Quantisation Noise Compared to 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 $1/2$ mV then this translates to $1/2$ LSB in digital numbers and the total noise is $\sqrt{1/4 + 1/12} \cong 0.6$LSB. Note how the digitization 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 subject, recall that the ADC returns only integer values but we are talking here about noise being a fraction of 1LSB. We will only see this noise in the statistics of a large number of samples, each of which is an integer, but taken together have non-integer statistics.

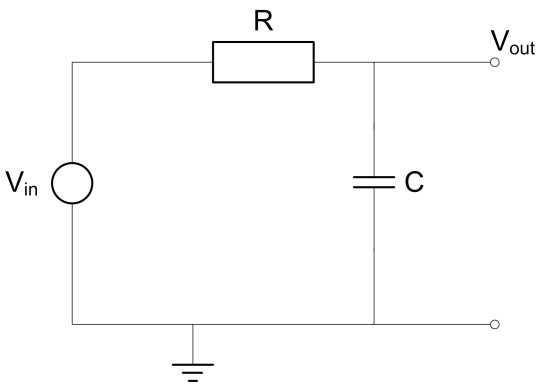


Figure 6.1: Simple RC circuit - The Low-Pass Filter

6 The RC Circuit

To see how to use the AC circuit analysis discussed in section 4 we'll study the series RC circuit in some detail. This also provides a simple illustration of several important concepts for the course. Because the capacitor has an impedance which is frequency-dependent, the RC circuit is an example of the class of systems known as filters, which will be very important for our studies.

As shown in figure 6.1, if we apply an input voltage across resistor and capacitor we can choose to take the output across either the resistor of the capacitor. Here we'll take the output across the capacitor. Consider the limiting cases of frequency: as $\omega \to \infty$ then the impedance of the capacitor tends to zero and is effectively short-circuit, so the entire input voltage appears across the resistor (by KVL) and $V_C \to 0$. Conversely at DC (e.g. if we apply a constant voltage) then the impedance of the capacitor tends to ∞, the current flowing tends to zero, hence $V_R \to 0$, and the entire applied voltage appears across V_C. We'll study two intermediate behaviours in more detail: first the behaviour under steady-state AC conditions, and secondly the transient response.

6.1 Steady-State Behaviour

In the folowing discussion, voltages and impedances are complex quantities. In the steady-state condition, we are working with periodic signals which have been present at the input for sufficiently long time that there is no remaining 'transient behaviour' (which we will cover in the next section. V_{in} is an AC sinusoidal waveform and the voltage across the capacitor we label as V_{out}. Complex Ohm's law gives

$$I = \frac{V_{in}}{Z} = \frac{V_{in}}{R - \frac{j}{\omega C}}$$

$$V_{out} = -\frac{Ij}{\omega C}$$

We want to find the output in terms of the input which we define as the Gain

$$Gain = \frac{V_{out}}{V_{in}} = \frac{-j}{\omega CR - j}$$

Multiplying by the complex conjugate of the denominator and simplifying gets

$$Gain = \frac{1 - j\omega CR}{\omega^2 C^2 R^2 + 1} \tag{6.1}$$

As anticipated the Gain is both frequency-dependent and complex (implying a phase change at the output). To visualise this we need to plot the magnitude and phase of the Gain

$$|Gain| = \sqrt{Gain \times Gain^*} = \frac{1}{\sqrt{\omega^2 C^2 R^2 + 1}} \tag{6.2}$$

$$\begin{aligned} \phi &= \tan^{-1}\left(\frac{\mathrm{Im}\{Gain\}}{\mathrm{Re}\{Gain\}}\right) \\ &= \tan^{-1}(-\omega CR) \\ &= -\tan^{-1}(\omega CR) \end{aligned} \tag{6.3}$$

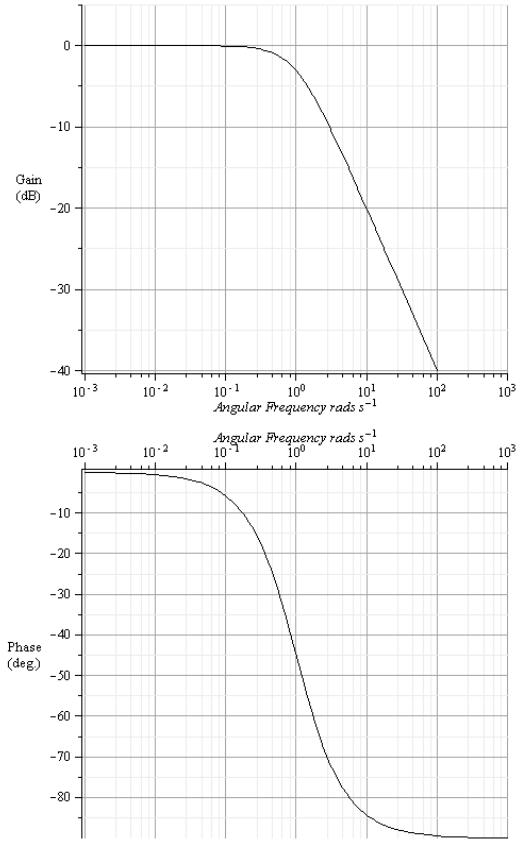


Figure 6.2: Bode Plot of Low-Pass Filter

6.2 Bode Plot

The Gain and phase plots together constitute the **Bode Plot** for the RC filter system. In Figure 6.2 the frequency axis has been normalised to units of $\omega_c = \frac{1}{RC} = 1\,\mathrm{rad/s}$.

For the Bode plot the gain magnitude is traditionally plotted log/log in dB, i.e. $20\log_{10} Gain$. The phase is usually given in degrees.

6.3 Phase Interpretation

The interpretation of the Gain magnitude is clear: it is the ratio of the amplitudes of the output voltage (that across the capacitor) to that of the input voltage (applied to both the capacitor and the resistor in series). The interpretation of the phase is more difficult. We understand that there is a phase difference, but between what and what? And does it lead or lag?

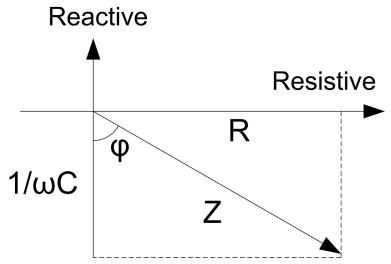


Figure 6.3: Argand Diagram for the Low-Pass Filter

6.3.1 By Argand Diagram

To understand this, we can use the Argand diagram (Figure 6.3). At this instant in time, we know that the voltage V_R across the resistor (and the current through it) is along the real axis i.e. it has a phase angle zero. The voltage V_C across the capacitor is always out of phase with the current and we can see that it has a phase angle -$\pi/2$. The voltage on the input is along Z, the combined impedance. The phase of the Gain is defined as the angle by which the output voltage leads the input voltage and here this is negative, i.e. $-\phi = -\tan^{-1}(\omega CR)$.

6.3.2 By Phasor Representation

A simpler but maybe less intuitive method is to consider the phasor representations of the quantities involved and use

$$V_{out} = Gain \times V_{in}$$

We know V_{in} lies along Z at $t = 0$ so we can write that the phase angle is (clockwise from the real axis) $\theta = {}^{3\pi}/_{2} + \phi$ and consequently

$$\begin{aligned} V_{out} &= |Gain|\, e^{j\phi}\, |V_{in}|\, e^{j\theta} \\ &= |Gain||V_{in}|e^{j(\theta+\phi)} \end{aligned}$$

As we know from above, ϕ is negative, so it is clear that V_{out} lags V_{in} by ϕ.

6.4 Low-Pass Filter

Frequencies $\omega < \omega_c$ pass from input to output with little attentuation (or phase change). ω_c is the **Cut-off Frequency**, i.e. the frequency at which the Gain has fallen by 3dB (a factor of $1/\sqrt{2}$) from the maximum. If we imagine that the filter is driving (i.e. connected across the output terminals) a load resistance R_{load} then the cut-off frequency is where the power delivered into R_{load} has fallen by half compared to the flat or 'pass-band' range of frequencies. For $\omega \gg \omega_c$ we can see

$$|Gain| \propto \frac{1}{\omega} \tag{6.4}$$

On the Bode Plot we see this as a slope of -20dB per decade of frequency (a decade is a factor of 10). This is equivalent to -6dB per octave (an octave is a factor of two in frequency). Both these figures are often quoted as the characteristic 'roll-off' frequency behaviour for a first-order (see section 6.5) low-pass filter.

High frequencies are severely attenuated, hence the understanding of this circuit as a **Low-Pass Filter**. As an example, were we to build a filter with $R = 1\Omega$, $C = 1$F and apply a 1 volt signal at $\omega = 1$ rad/s then the waveforms would be as per Figure 6.4.

Note that, by KVL, the voltage across the resistor has the opposite behaviour, so if we swap resistor and capacitor we have a **High-Pass Filter.**

6.5 Transient Response of the High-Pass Filter

For situations where we do not have a 'steady-state' we need to consider the transient response of the circuit. This is quite common in instrumentation applications; something happens at time $t = 0$ and we need to observe how the system evolves for times $t > 0$. Consider

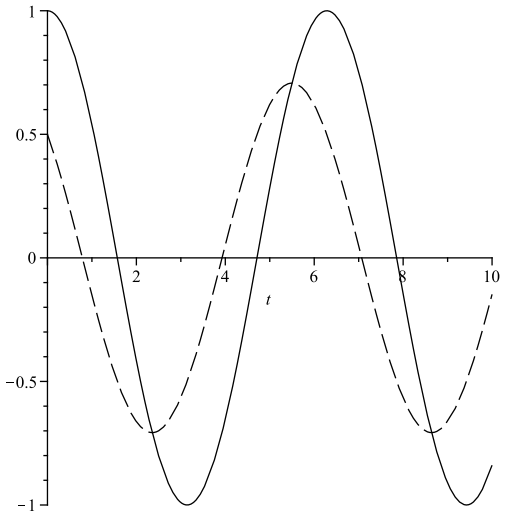


Figure 6.4: Low-Pass Filter Waveforms (solid line is input, dashed line is output)

the modified circuit in Figure 6.5. Note that in this figure we've swapped the positions of the resistor and capacitor. This is just because we want to observe the voltage across the resistor, but in terms of frequency-response this is a high-pass filter. If we set the applied voltage to be a constant V and then close the switch at time $t = 0$ then $V(t) = u(t)V$. From now on we'll just consider $t > 0$ so

$$\begin{aligned} V(t) &= V \\ V_R(t) &= i(t)R \\ V_c(t) &= \frac{q(t)}{C} \end{aligned}$$

Using KVL we find

$$V = V_R(t) + V_C(t)$$

Taking the time derivative we get

$$0 = RC\frac{di}{dt} + i \tag{6.5}$$

Note that we have set up a first-order ordinary differential equation. As a fundamental rule,

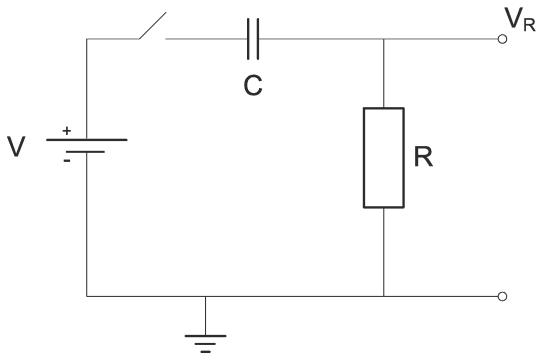


Figure 6.5: Switched RC Circuit

there must be as many arbitrary constants in the solution to the equation as the order of the equation. This means we need additional information if we are to find a complete solution, and this information comes from a knowledge of the *state* of the system, usually taken to be the initial time $t = 0$. You will have come across several methods for solving ODEs and the form of Equation 6.5, where the function and its derivative must add to zero, suggests we attempt a trial solution of the form

$$i(t) = Ae^{st} \tag{6.6}$$

Substituting into Equation 6.5

$$RCsAe^{st} + Ae^{st} = 0$$

Yields

$$s = -\frac{1}{RC}$$

To find the unknown A we apply knowledge of the initial condition of the circuit. This is assumed to be relaxed. That is to say the capacitor is uncharged, therefore $V_C(0) = 0$, and we can take $V_R(0) = V$. A physical understanding of this is that, the instant the switch is closed, there is no charge on the capacitor so an instantaneous current $i(0) = V/R$ will flow. Substituting into Equation 6.6

$$i(0) = A = \frac{V}{R}$$

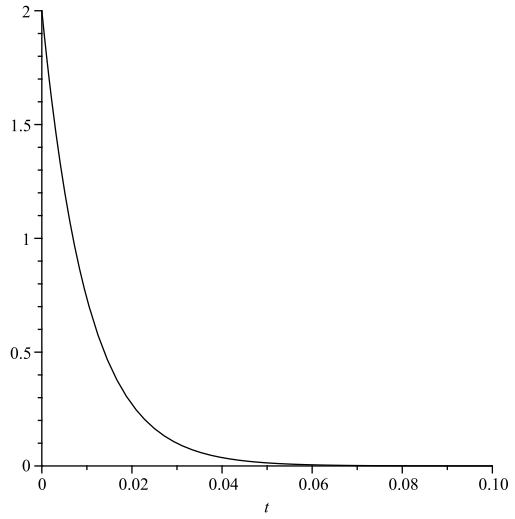


Figure 6.6: Time-domain response of the high-pass filter to step-input

So the full solution is

$$i(t) = \frac{V}{R}e^{-t/RC} \tag{6.7}$$

From which

$$V_R(t) = i(t)R = Ve^{-t/RC} \tag{6.8}$$

The exponential is characteristic of first-order systems. Figure 6.6 gives the response of a high-pass filter with $R = 10\text{k}\Omega$, $C = 1\mu\text{F}$ to a step input of 2 V.

The **Time Constant** of the system is RC. As $t \to \infty$ the capacitor achieves full-charge and hence $V_R \to 0$ as no current flows. Note that had we chosen a different initial condition then the solution is different.

6.6 Initial Conditions

There are three factors which determine how the voltage in our circuit evolves.

1. The differential equations which govern the circuit. These are entirely fixed and governed by the physics of the circuit.

2. The applied signal, in our case the voltage $V(t)$. This is frequently referred to as the **Forcing Function**; it is the externally applied signal which causes our system to respond.

3. The initial conditions of the system. In section 6.5 we specified that the system was **Initially Relaxed**, that is to say the capacitor carried no initial charge. This is not always the case.

Imagine now that at some time after the capacitor is fully charged we open the switch. No current can flow round the circuit therefore the capacitor remains charged. $V_C = V$. If we now close the switch again, nothing happens - no current flows and V_R stays zero. Imagine that we now instantaneouly change our forcing function to zero, i.e. $V(t) = 0$. Then following the same methods as above we will find

$$V_R(t) = -Ve^{-\frac{t}{RC}}$$

6.7 Integration and Differentiation

If we now repeat this on/off sequence every T seconds where $T \gg RC$ then our forcing function is a square wave between zero and V with period T. Plotted on a scale of several input cycles, the output looks like a sequence of spikes (positive and negative; it is worth sketching this). The ouptput is then *approximately* the time-derivative of the input signal. Mathematically, we can show this by

$$V_R = iR = RC\frac{dV_c}{dt} \approx RC\frac{dV_{in}}{dt}$$

For $\omega \gg \omega_c$ we find the impedance of the capacitor dominates over that of the resistor therefore $V_{in} \approx V_C$

Conversely, were we to go back to the low-pass configuration, but increase the values of R and C to make the time constant $RC \gg T$, We would find the circuit integrates the square wave to give a triangle output. Note that this relationship between the square wave and the triangle wave makes sense from a Fourier point of view, as we saw in section 7.2. The Fourier coefficients of the triangle-wave are similar to those of the square-wave but scaled by $1/n$. This makes sense when we consider the action of the low-pass filter on the square wave as it has a frequency response of $1/f$. Likewise the high-pass filter differentiates by scaling the Fourier coefficients by n. This turns the square wave into a sequence of delta-functions.

We'll come back to this later on but for now it is important to have a physical understanding of these RC circuits and how they behave in both the time and frequency domains. The principal characteristics are summarised in Table 6.1.

6.8 Physical Analogues

The RC circuit is a good analogue for physical systems where the rate of change of a parameter is proportional to the magnitude of the parameter itself (see Equation 6.5). For example, many thermal calculations apply the knowledge that heat-flow (and hence rate of change of temperature) is proportional to temperature difference. In mechanics we might have a situation where a braking-force (hence rate of change of velocity) is proportional to velocity. Such systems can be modelled by RC circuits, and indeed in the past before the use of numerical methods and computational models, it was common practice to model physical systems with such 'analogue computers'.

6.9 Linearity of the RC Filter

The reason we have spent so much effort studying the RC filter is that it belongs to a class of systems known as *Linear Systems*. A future handout is dedicated to the properties of linear systems. We will see how we can predict the behaviour of linear systems not just under sinusoidal inputs but with any arbitrary input. This will become a key theme in our studies, and in later parts of the course we will show how to use Fourier and Laplace transforms to

Behaviour	Low-Pass	High-Pass
Take voltage across	C	R
Time Constant	RC	RC
$\omega \ll \omega_c$	flat Gain 0 dB	Gain slope +20 dB/decade. Differentiates
Cut-off frequency	$\omega_c = 1/RC$	$\omega_c = 1/RC$
$\omega \gg \omega_c$	Gain slope -20 dB/decade. Integrates	flat Gain 0 dB

Table 6.1: Summary of RC Circuit Characteristics

gain easy solutions to some seemingly very complex problems in **systems analysis**, with the important proviso that all the component parts of the system *are themselves linear.*

To summarise what we know about the RC filter we have the Gain (here for the low-pass filter) given by

$$Gain = \frac{V_{out}}{V_{in}} = \frac{1 - j\omega CR}{\omega^2 C^2 R^2 + 1}$$

- Gain is a complex quantity and also a function of frequency. The plot of Gain vs. frequency is the Bode Plot
- For any frequency ω the Gain gives us the output sinusoid V_{out} amplitude and phase compared to the input V_{in}
- Output and input sinusoid have the same frequency ω but an adjusted amplitude and phase
- The behaviour of the RC filter is governed by first order ordinary differential equations (for example equation 6.5) hence it is a *first-order linear-system*
- The time domain response to a step input is the familiar exponential form with a characteristic time-constant.

Here is where our understanding of signals as complex quantities (in electronics, Phasors, see sections 4.3 & 4.4) becomes useful. If we write the input as a sinusoid

$$V_{in} = Ae^{j\theta}e^{j\omega t}$$

and the gain (evaluated at the frequency ω) as

$$Gain = Ge^{j\phi}$$

then

$$V_{out} = Gain \times V_{in} = Ge^{j\phi}Ae^{j\theta}e^{j\omega t}$$

As was stated in section 4.3 the time dependent part of the expression is frequently omitted since it is common to input and output. *We can do this because the linear system always preserves the frequency of the input.*

$$V_{out} = GAe^{j(\theta+\phi)}$$

That is to say, the output is the input scaled by a factor G (the magnitude of the Gain) and rotated by an angle ϕ (the phase angle of the Gain). Now comes the most important point which will be essential for our future studies. For a linear system, the output is a linear superposition of the input. If our input is composed a several sinusoids, we can apply the above calculation for each input component and sum the results to get the output. Since Fourier tells us *any* input can be decomposed into sinusoids, this means that the Gain tells us what the output will be for *any arbitrary input waveform.* This is profoundly important, as it allows use to use Fourier (and related) techniques to solve problems.

7 Fourier Series of Periodic Waveforms

We touched on this in Figure 2.1. Wherever there exists periodicity, we should seek a Fourier representation, and we will see how a Fourier understanding of signals is essential in the field of instrumentation. In practice, all physically realistic periodic signals obey the Dirichlet Conditions[8] and are therefore transformable into a **Fourier Series**. Hence any signal $f(t)$ with period T can be expressed by

$$\begin{aligned} f(t) &= \sum_{n=-\infty}^{\infty} \alpha_n e^{jn\omega_0 t} \\ &= \sum_{n=-\infty}^{\infty} |\alpha_n| \, e^{j(n\omega_0 t + \phi_n)} \end{aligned} \quad (7.1)$$

Where the α_n are complex constants given by

$$\begin{aligned} \alpha_n &= \frac{1}{T}\int_{t_0}^{t_0+T} f(t) e^{-jn\omega_0 t} dt & (7.2) \\ &= |\alpha_n| \, e^{j\phi_n} \\ &= |\alpha_n| \cos\phi_n + j\,|\alpha_n| \sin\phi_n & (7.3) \\ \omega_0 &= \frac{2\pi}{T} \end{aligned}$$

Note that ω_0 is a *constant* which is determined by the repeat-period of our function.

Since the α_n are determined from $f(t)$, and *vice versa*, they are known as a **Fourier Transform Pair**. At any time t_0 the Fourier expansion of the function converges to $f(t_0)$ as long as the function is continuous at t_0. If the function discontinuous then the Fourier expansion converges to a point mid-way between the discontinuity. If $f(t)$ is real then

$$\begin{aligned} \alpha_{-n} &= \alpha_n^* \\ f(t) &= \alpha_0 + \sum_{n=1}^{\infty} [(\alpha_n + \alpha_n^*)\cos n\omega_0 t \\ &\quad + j(\alpha_n - \alpha_n^*)\sin n\omega_0 t] \end{aligned} \quad (7.4)$$

[8] See Poularikas section 3.2

Which can be written in trigonometric form as[9]

$$\begin{aligned} f(t) &= \frac{A_0}{2} + \sum_{n=1}^{\infty} (A_n \cos n\omega_0 t + B_n \sin n\omega_0 t) \\ A_0 &= \frac{2}{T}\int_{t_0}^{t_0+T} f(t)dt \\ A_n &= \frac{2}{T}\int_{t_0}^{t_0+T} f(t)\cos n\omega_0 t \, dt \\ B_n &= \frac{2}{T}\int_{t_0}^{t_0+T} f(t)\sin n\omega_0 t \, dt \end{aligned} \quad (7.5)$$

Or

$$\begin{aligned} f(t) &= \frac{A_0}{2} + \sum_{n=1}^{\infty} C_n \cos(n\omega_0 t + \phi_n) \\ C_n &= \sqrt{A_n^2 + B_n^2} \\ \phi_n &= \tan^{-1}\left(\frac{B_n}{A_n}\right) \end{aligned} \quad (7.6)$$

7.1 The Square Wave

The Fourier expansion of a square wave of amplitude 1 and period 1 is

$$f(t) = \frac{4}{\pi}\left(\sin t + \frac{\sin 3t}{3} + \frac{\sin 5t}{5} + \frac{\sin 7t}{7} + \cdots\right) \quad (7.7)$$

i.e. all the A's are zero, as are all even B_n. Figure 7.1 illustrates the first nine coefficients and Figure 7.2 shows the square wave reproduced from the first 2, 3 and 9 non-zero coefficients. As n increases so does the fidelity, though we are always left with an overshoot of about 10% at the edges (Gibbs' phenomenon). Fidelity is worst at the edges, and this improves rapidly with n. From this we know that sharp-edges are represented by high-frequencies in the expansion. We see this if we run a square wave through a low-pass filter: the edges are 'rounded off' (see section 6.5).

[9] Note that you would not be expected to reproduce these formulae for an exam!

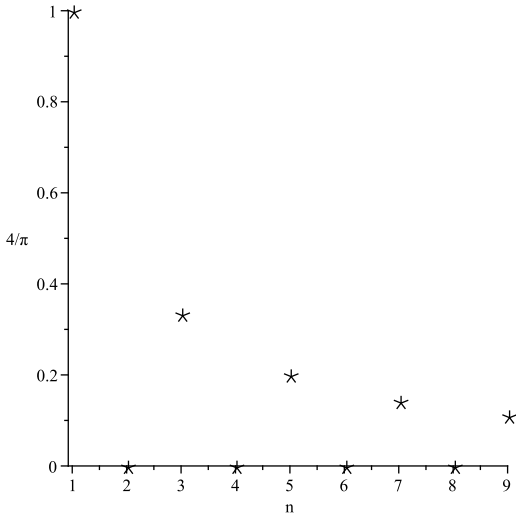


Figure 7.1: Square Wave amplitude coefficients

Related is the fact that the higher harmonics in the expansion contribute only to the detail of the waveform. We can see in Figure 7.2 where the centre plot shows a passable square wave from just the fundamental plus the 3rd and 5th harmonics. Figure 7.3 a plot of the power-spectrum (amplitude coefficients squared, normalised to power of the fundamental). The higher harmonics contribute a tiny fraction of the overall signal power.

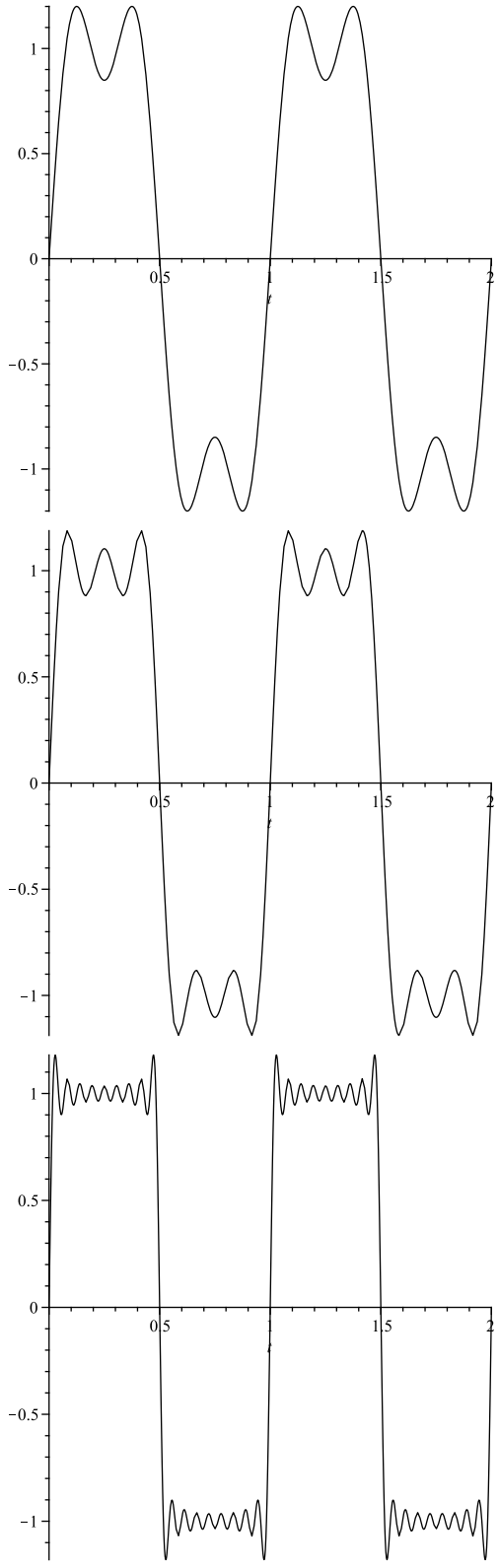
7.2 The Triangle Wave

The Fourier expansion of a triangle wave of amplitude 1 and period 1 is

$$f(t) = \frac{8}{\pi^2}\left(\sin t - \frac{\sin 3t}{9} + \frac{\sin 5t}{25} - \frac{\sin 7t}{49} + \cdots\right) \tag{7.8}$$

Figure 7.4 shows the first 9 coefficients and subsequent expansion. Note that

1. There are negative coefficients
2. The coefficients are similar to those of the square-wave but scaled by $1/n$. This makes sense when we consider the action



32Figure 7.2: Square wave reproduced from the first 2, 3 and 9 non-zero Fourier coefficients

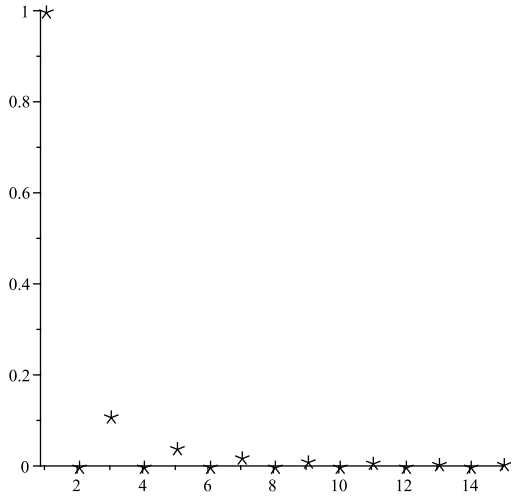


Figure 7.3: Power Spectrum for the square wave

of the low-pass filter on the square wave (section 6.7)

7.3 Symmetry

An even-symmetric function has symmetry about the $t = 0$ axis (i.e. the same under reflection about the vertical axis, e.g. $\cos(t)$), while an odd-symmetric function has rotational symmetry about the origin (unchanged after rotation 180 degrees around the origin, e.g. $\sin(t)$). The functions considered above are odd which simplifies the expansion somewhat according to Table 7.1 .

7.4 Spectra

Typically, we might specify a spectrum of a signal by its coefficients C_n (known as the amplitude spectrum) and their corresponding phases ϕ_n (known as the phase spectrum). The spectra studied in the previous sections are special cases with zero phase spectra. It is worth noting that choosing an appropriate origin for the representation of a signal can significantly simplify the expansion in this way[10].

[10] See Poularikas section 3.3

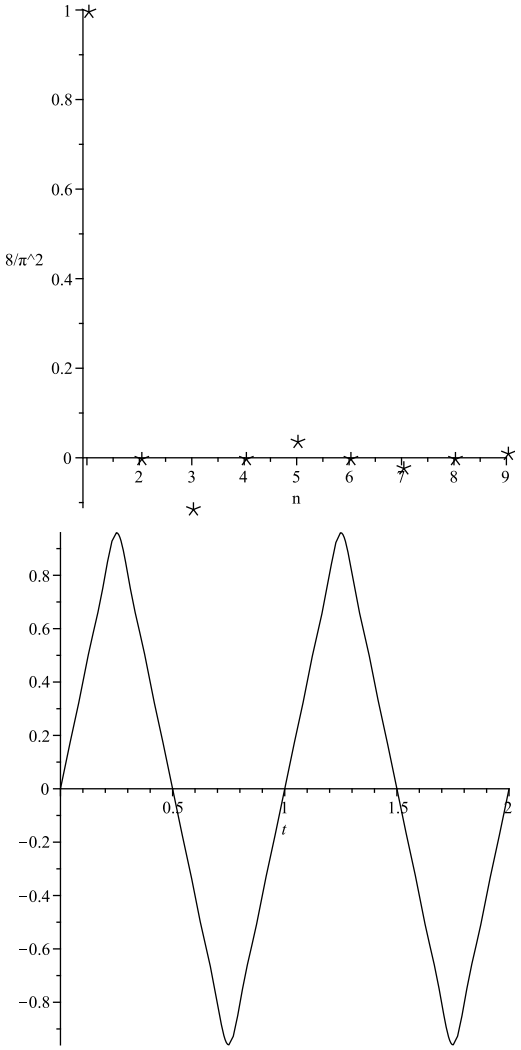


Figure 7.4: Triangle Wave Coefficients and Waveform

Symmetry	Functional Form	A_0	A_n	B_n
Even	$f(t) = f(-t)$	Exist	Exist	0
Odd	$f(t) = -f(-t)$	0	0	Exist

Table 7.1: Symmetries and Fourier Coefficients

7.5 Finite Signals

Finite signals, such as a single arbitrary-shaped pulse within the interval $t_0 < t < t_0+T$, can be represented by assuming that the pulse repeats infinitely with period T. The expansion is done in the normal way but the final range of applicability is limited to the range of the original function[11].

7.6 The Energy Relation

Parseval's theorem applied to a signal $f(t)$ expanded as a Fourier series gives

$$\frac{1}{T}\int_{t_0}^{t_0+T} f(t)^2 \mathrm{dt} = \frac{A_0^2}{4} + \sum_{n=1}^{\infty} \frac{C_n^2}{2} \qquad (7.9)$$

If we consider $f(t)^2$ to be the power in the signal then the LHS of equation 7.9 is the average power in the signal over the period of 1 cycle. The RHS is the power of the signal in the frequency-domain, expressed as a sum of the power of the individual frequency components.

Equation 7.9 is known as the **Energy Relation** since it tells us that energy measured in the time domain is the same as the energy measured in the frequency domain, a consequence of conservation of energy.

7.7 Applicability, or use in Representing Real Signals

The Fourier series is excellent for describing the periodic, repetitive signals frequently encountered in physics and engineering, such as the square, triangle and sawtooth waveforms and indeed any arbitrary waveform *so long as it has a repeat period T*. We see in the Fourier series of the square and triangle waveforms (sections 7.1 and 7.2) a fundamental signal at the repeat period T which we can recognise in both time and frequency representations, and we can see that the fundamental carries most of the signal power. The Fourier series is very useful where we have a 'steady-state' periodic signal input into a system. By **Steady State** we mean to imply that the signal has existed long enough that it can be reasonably described as a Fourier series. Mathematically this would mean that the signal should have existed for all time, which is of course not realistic, however if the signal has existed long enough for any **Transient Effects** to have died away then this is a reasonable assumption. By transient effects we mean the initial response of the system to a sudden start-up of the signal, for example of we take a system and suddenly apply a square wave input then the initial response, in say the first few cycles of the square wave, may be rather different from the behaviour after a thousand cycles. This is because many systems have some 'memory' or inertia. As a simple example imagine pushing someone on a playground swing; the periodic push is the input but it takes some number of cycles to build up to the full amplitude. We will come back to these ideas of transient effects later on and see how to deal with them, however for now it is important to understand that the Fourier series is an excellent way of describing *periodic* signals that have existed for a time much longer than any *time-constant* (or memory) of the system being studied.

[11] See Poularikas figure 3.13

8 Aliasing and the Sampling Theorem

As suggested in section 5.1, we throw away a lot of information about a signal when we digitize it because we quantise both time *and* the parameter being measured. This is probably the single most important point to understand about digitized signals: *information lost in the digitization process is lost forever* - there is no way we can get back 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 digitization 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**.

8.1 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's 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'll 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 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 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. This is bad; it is often referred to as *improper sampling*.

8.2 Aliasing

This is best illustrated graphically, in the first instance. Figure 8.1 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 (kilo-samples-per-second), 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. The brain still tells us this is a roughly-drawn sinusoid at the correct frequency. It turns out that the brain is very good at this sort of thing. Also, 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 brain 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

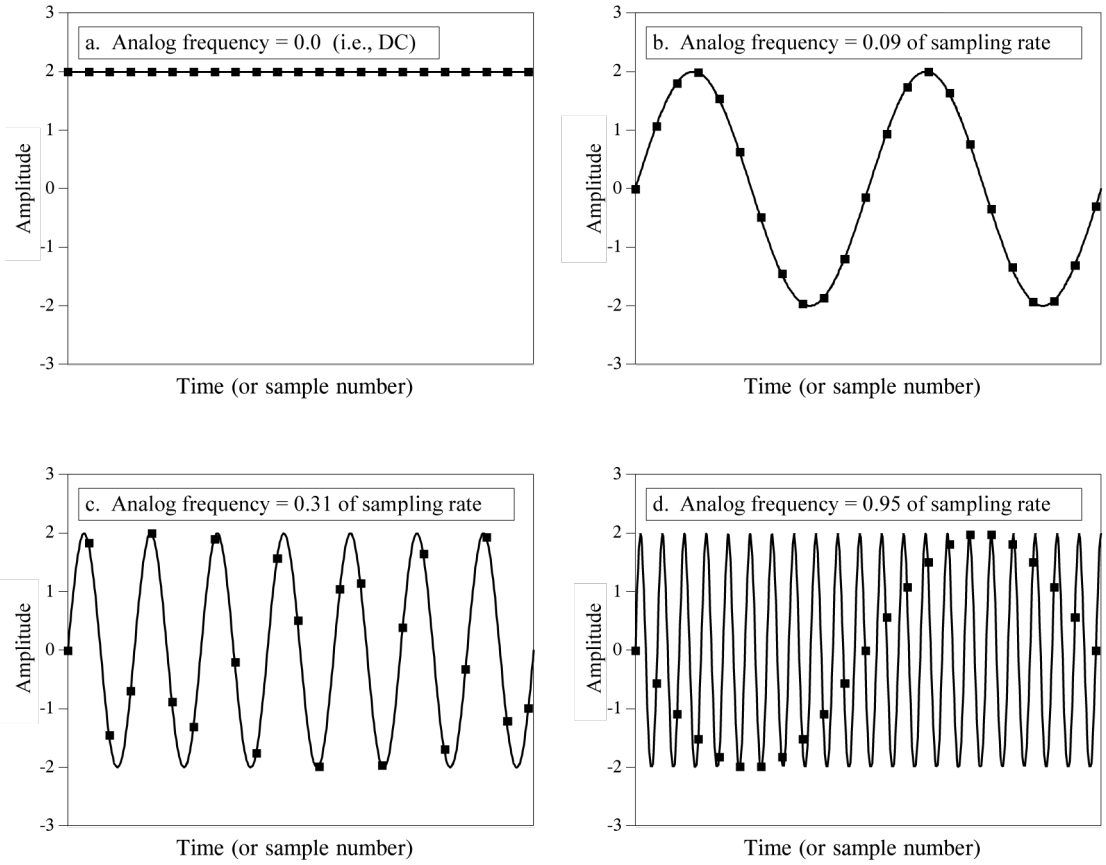


Figure 8.1: Graphical Representation of Proper and Improper Sampling (from Smith, www.dspguide.com)

$$f_a = f_s - f$$

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

$$f_a = |nf_s - f| \tag{8.1}$$

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'll 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 3 are *aliases*.

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 fre-

quencies, 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.

8.3 The Frequency Characteristics of Sampled Signals

To understand sampling properly, we need to take a more mathematical approach. This requires a Fourier understanding of our signals, which is covered in more detail in sections 7 and 12, though you should have done this already in the 2nd year Fourier course. This is not hard however, and should lead us to a more complete understanding of sampled signals which is essential in modern-day experimental physics. Here again is the equation for the sampled signal.

$$f_s(t) = \sum_{n=-\infty}^{\infty} f(nT_s)\delta(t - nT_s) \qquad (8.2)$$

One important point to start with is that a sampled signal is fundamentally unlike any other kind of continuous signal you will have come across before. According to equation 8.2, 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 equation 8.2 to get[12]

$$\begin{aligned} F_s(\omega) &= \frac{1}{T_s}\sum_{n=-\infty}^{\infty} F(\omega + n\omega_s) \qquad (8.3) \\ \omega_s &= \frac{2\pi}{T_s} = 2\pi f_s \end{aligned}$$

This shows that the spectrum of the sampled signal is the same as the original signal, but repeated infinitely along the frequency axis. Figure 8.2 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 it illustrates a 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 does provide a rather satisfying explanation for aliasing. In equation 8.2 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[13]. 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. 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.

[12] We will cover the Fourier transform in more detail in section 12. 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 derive this.

[13] See the table of Fourier transforms in Poularikas

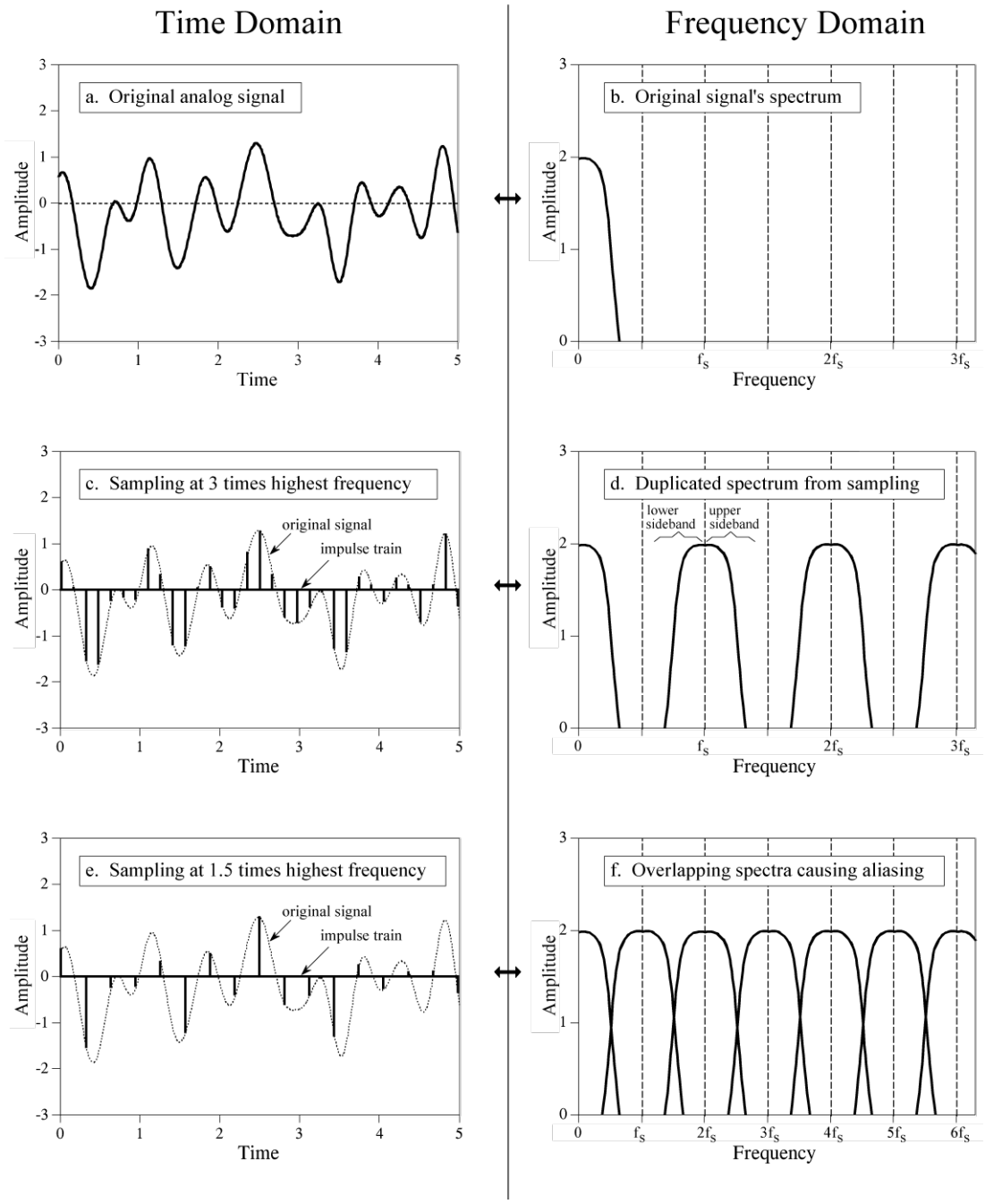


Figure 8.2: Sampled Signals in the Frequency Domain (from Smith, www.dspguide.com)

As a final word on the topic, take another look at the spectrum of the properly sampled signal, 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 2 profound consequences:

1. For proper-sampling, we have not lost *any* 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 get back to the original signal.

The latter can be done by the use of a low-pass filter (section 6.4). This technique is the basis of Digital to Analogue Conversion, as we will see later.

8.4 Anti-Aliasing Filters

We are not finished with aliasing yet. It is such an important topic for digital signals that a significant part of the discipline of **Digital Signal Processing** (DSP) is devoted to dealing with it. As seen in the previous section, if our signal is contained within a narrow band of frequencies below the Nyquist, then all is OK. However in practice this is rarely the case for two reasons:

1. Real signals (either pulses or repetitive waveforms) tend to have mathematically infinite waveforms (for example the square-wave, section 7.1). While no real signal is mathematically perfect, any signal with rapid changes (what engineers call "edges") has a very wide spectrum
2. All signals contain some noise, and noise is usually *broadband*, i.e. existing right across the frequency spectrum

In order to prevent these high frequency components aliasing into our digitized signal, it us 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 problem of aliasing is so bad that this is usually a price worth paying. Figure 8.3 shows a 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 digitized by the ADC and can then be stored and processed by the computer (central box).

> Note that once the signal is in a digital form we can do all sorts of things to it such as "Equalisation" which generally means adjusting the relative amplitudes of different frequency components. "Bass-boost" is a common process designed to counteract the fact that cheap headphones typically have a very poor bass-response. We will see this for real in later lab-sessions. Another popular equalisation is "loudness". This is intended for quiet listening where the human ear, under quiet conditions, has poor response at the very low and high ends of the frequency range. The loudness feature boosts these extremes of frequency but leaves everything in the middle flat.

Everything to the left of the diagram is the recording system. The right side is "playback". As mentioned in the previous section we need a Digital to Analogue Converter and another filter to properly reconstruct the original analogue signal, which then drives the speakers. It's worth noting that all of this, and of course much more, is contained within the modern iPod.

Entire books have been written on the subject of anti-alias filters, and electrical engineers receive complete lecture courses on filter design. For our purposes, including the lab-sessions, we will use the simple RC low-pass filter discussed in some detail in section 6. The ideal filter would have a very sharp cut-off, that is to say it would allow through all frequencies below the Nyquist and completely block anything above.

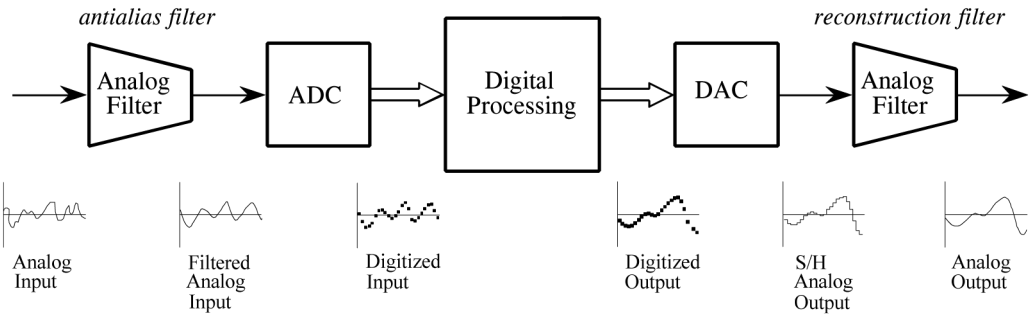


Figure 8.3: Correctly Setup Digital Signal Processing System (e.g. iPod) (from Smith, www.dspguide.com)

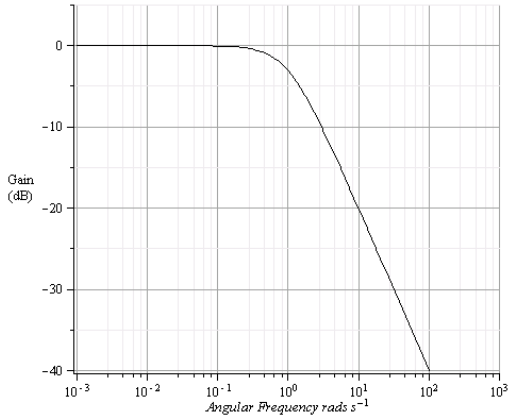


Figure 8.4: Response of the RC Low-Pass Filter

Unfortunately this is impossible. In the case of the RC low-pass filter we usually and choose the -3 dB frequency of the filter to be the same as the Nyquist frequency of the sampling. The details of this will be covered in section 6 and the lab-sessions. This is however a compromise, as can be seen in figure 8.4. This filter would be suitable for a sampling with Nyquist of 1 rad/s. Some frequencies below the Nyquist are attenuated (undesirable) and also some frequencies above the Nyquist will still exist (also undesirable). Filters with a sharper "cut-off" are possible, but beyond the scope of this course[14].

[14]For more information, see Smith chapter 3 at www.dspguide.com and also Horowitz & Hill chapter 4

9 Differential Signals

9.1 Single-Ended Signals

In all of the discussions so far to do with electrical signals we have considered our signal to be a voltage relative to ground potential (zero volts). This means that we would measure the signals - e.g. with a Digital Multi-Meter (DMM) or oscilloscope - by connecting the black lead of the meter to ground and the red lead to the measurement point. These signals are known as **Single-Ended** signals. In principle, a single-ended signal can be carried from A to B on a single piece of wire, that is to say if A and B are two different units (such as experiment and oscilloscope) then we only need a single wire to transmit the signal. This assumes that the zero-volts ground at A and B is the same. There are reasons why this might not be exactly true, which we will look at later in the course when we study noise. It is for this reason that it is standard practice, when probing a circuit, to connect the black lead of the scope to a ground-point as close as possible to the point where you are probing. Furthermore, for practical purposes when we want to transmit a real signal from A to B we might use something like a piece of co-axial cable whereby the outer shield (the outer conductor in the co-ax, usually a braid of fine wires woven together) of the cable is connected to ground *at both ends*. A drawing of this kind of setup was given in lectures (the section on co-axial cables). This kind of configuration ensures that "ground" has the same potential at both ends of the line, so that the signal we are measuring has the same amplitude at both ends.

9.2 Differential Signals

While the single-ended signal is of course always specified as being relative to ground potential, the **Differential Signal** has the subtle difference that it is relative to *some other potential* (which is itself usually not ground). As an example of this consider a sensor which has two-terminals and produces an output across

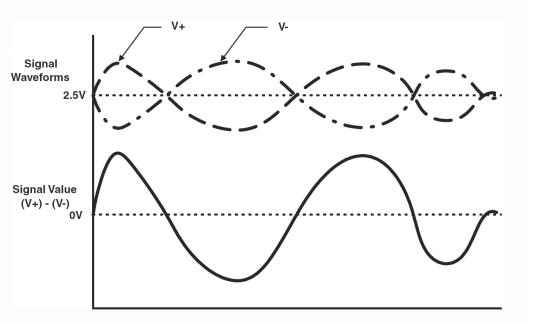


Figure 9.1: Differential Signal

these terminals which is a small time-varying voltage. An example of such a sensor might be an accelerometer or a strain gauge. We are only interested in the *difference* between the two terminals. We don't care what the actual potential of either terminal is relative to ground; the sensor information is a Differential Signal. This is illustrated in figure 9.1. Here the potentials of each line appear to be varying about a level of 2.5V (upper traces). The differential signal is determined by taking the difference of the two.

We can define the differential signal as

$$V_D = V_+ - V_-$$

9.3 Common-Mode Signal

The Common-Mode Signal is defined as follows

$$V_{CM} = \frac{V_+ + V_-}{2}$$

This is the common level about which both terminals vary, here 2.5V. In practice, V_{CM} is the average value of the potentials at the two terminals, at any time.

The main point to consider here, is that V_D represents the signal that we want to measure, while V_{CM} represents some kind of background which we are not interested in. In order to buffer and amplify our small differential signal we will make use of a Difference Amplifier

9.4 Difference Amplifiers

Recall that a standard op-amp is itself a differential amplifier since the output obeys the equation $V_{out} = A(V_{in+} - V_{in-})$. However we have a very high gain and use our op-amp in closed-loop mode so we operate with feedback such that (in the IGA approximation) $V_{in+} - V_{in-} \approx 0$. Further, one of our inputs is usually connected to ground (see the inverting amplifier design for example). It may not be desirable to connect one of our sensor terminals to ground. In this case it would short our common-mode voltage to ground which could have undesirable consequences for the operation of the sensor. Another reason is that this could add noise to our sensitive measurement (as we will see later, ground is not necessarily a "clean" zero volts; there can be significant amounts of low and high frequency noise on the ground). In order to preserve the isolation of our sensor terminals from ground we need a special sort of amplifier called the **Difference Amplifier**, or ultimately the **Instrumentation Amplifier**. The circuits for these will be covered in lectures and also in section 10.

9.5 Common-Mode Rejection

The main points to consider when choosing a difference amplifier are

- The isolation from ground at the 2 inputs (ideally very high)
- The amplification of the differential voltage (we would like to choose this)
- The amplification of the common-mode voltage (ideally zero)

Since no amplifier is perfect there will always be some common-mode voltage which gets through the amplifier and appears on the output. In practice the output voltage is given by

$$V_o = A_D(V_+ - V_-) + \frac{1}{2}A_{CM}(V_+ - V_-)$$

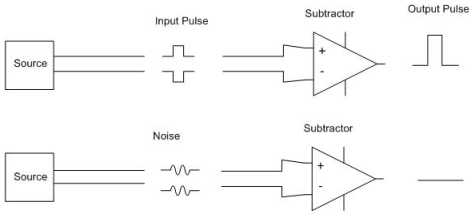


Figure 9.2: Noise Rejection in Differential Signals

A_d is the **Differential Gain** and we want this to be large. A_{cm} is the **Common-Mode Gain** and we want this to be small. For any particular amplifier design which we might choose to build we can calculate or measure the **Common-Mode Rejection Ratio** (CMRR) which is usually given in dB and defined as

$$\text{CMRR} = 20\log_{10}\frac{A_D}{A_{CM}}$$

CMRR is therefore a figure-of-merit for a difference amplifier design. The AD620 Instrumentation amplifier from Analog Devices has a CMRR of 100dB, which is very high performance. More specialist devices can achieve 120dB or more.

9.6 Noise Rejection

An advantage of working with differential signals is illustrated in figure 9.2. Here the signal is represented by a differential pulse. We know that wires can act as aerials to pick-up noise from radio waves or other signals elsewhere in the same circuit. However, if the two wires which carry the differential signal are reasonable close to each other then it is reasonable to assume that the amount of noise picked-up will be the same on both. That is to say, the noise is a common-mode signal. This means that the amplifier will remove this noise.

10 Instrumentation Amplifier

The **Instrumentation Amplifier** is widely used in scientific apparatus because of its flexibility and performance. We'll study it in some detail because it summarises in a single circuit more-or-less everything that we need to know about amplifiers and their practical applications. This is easily the most complex op-amp circuit that we'll cover in this course, and while it looks taxing at first glance we'll see that by breaking the design down into its component parts and applying the simple rules for op-amp circuits, the analysis of the instrumentation amplifier is quite simple.

Firstly we'll summarise the properties of the instrumentation amplifier:

- High **Differential-Mode Gain** (A_D) and very low **Common-Mode Gain** (A_{CM}), which means it amplifies the *difference* between two inputs and doesn't respond to identical (*common*) voltages.
- Consequently it has a very high **Common-Mode Rejection Ratio** (typically 100 dB)
- High input impedance on both inputs, so we can connect it across any 2 points in the circuit where we might wish to measure the potential difference, without affecting the operation of the circuit.

The starting point for the instrumentation amplifier is the simple difference amplifier which we covered in lectures, so we'll recap this in the next section

10.1 Difference Amplifier

Note first that the op-amp is fundamentally a difference amplifier since it obeys the rule $V_{out} = A(V_{in+} - V_{in-})$ and in lectures we have seen how to use this to make practical amplifiers such as the inverting op-amp circuit. This works fine for many applications, however you will notice that the non-inverting input is always connected to ground. This means we are always amplifying a voltage *relative to ground potential*. In some circumstances this is undesirable. Consider for example a complex experimental setup within which we want to measure the voltage across a single resistor (maybe because we want to be able to calculate the current through it). Connecting the inverting op-amp circuit across the resistor would connect one side to ground, probably stopping our circuit from working properly. The solution to this may be to attach some more resistors to the non-inverting input. The result is the **Difference Amplifier** as shown in figure 10.1.

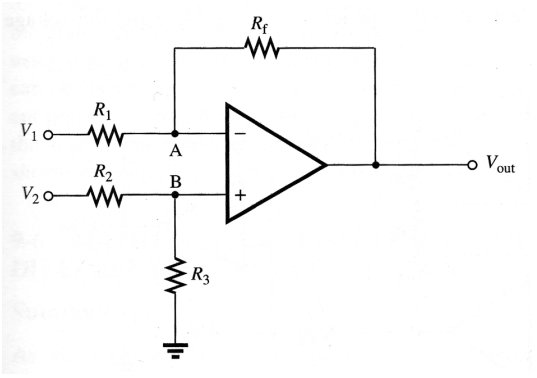


Figure 10.1: The Difference Amplifier (Diefenderfer chapter 9)

10.1.1 Circuit Analysis

To analyse this circuit we need Kirchhoff's Current Law which states that the sum of all the currents into any node in the circuit is zero, $\Sigma i = 0$. Separating the top and bottom branches from the op-amp – as done in the lectures for the non-inverting amplifier – we get (figure 10.2)

$$\begin{aligned} I_1 &= \frac{V_1 - V_A}{R_1} \\ &= \frac{V_A - V_{out}}{R_f} \end{aligned}$$

Solving for V_A

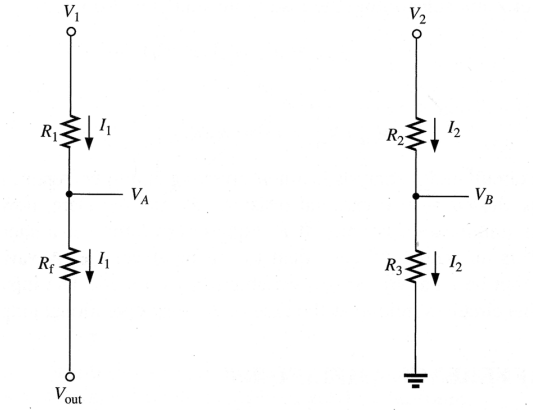


Figure 10.2: Circuit Analysis for the Difference Amplifier (Diefenderfer chapter 9)

$$V_A = \frac{V_1 R_f + V_{out} R_1}{R_1 + R_f}$$

For the other branch, we can do the same, or since the bottom of the circuit is connected to ground we can just treat it as a voltage divider and write

$$V_B = \frac{V_2 R_3}{R_2 + R_3}$$

Since op-amp rule 1 requires that $V_A = V_B$ we can solve for V_{out} to yield

$$V_{out} = V_2 \left(\frac{R_3}{R_1}\right)\left(\frac{R_1 + R_f}{R_2 + R_3}\right) - V_1 \left(\frac{R_f}{R_1}\right)$$

Now if we choose to set $R_2 = R_1$ and $R_3 = R_f$ then we get

$$V_{out} = (V_2 - V_1)\left(\frac{R_f}{R_1}\right) \tag{10.1}$$

This simple equation describes the difference amplifier, which has a number of useful applications in situations where we can guarantee to match the requirement $R_2 = R_1$ and $R_3 = R_f$. However there are a couple of drawbacks with this design as we shall see in the next section.

10.1.2 Problems with the Difference Amplifier

1. If we don't match $R_1 = R_2$ and $R_3 = R_f$ *exactly* then we get a **Common Mode Signal** at V_{out} as well as the **Differential Signal** that we want (see section 9 and also the box below). Any common mode signal is an error signal, since it depends on how carefully you build the amplifier circuit.

2. It has a relatively low input impedance (at V_2 this is equal to $R_2 + R_3$). This can cause us to 'load' the signal we are trying to measure by drawing current from it. You may be wondering why we don't just make the values of the resistors very large, say some MΩ? The problem with this is that the inputs of the op-amp inevitably have some stray-capacitance, so we would have a time-constant effect which would slow-down voltage swings at the input. We need to keep $R_2 + R_3$ to be some kΩ.

Common Mode Gain As an exercise, consider what happens if we set $\Delta V = 0$, i.e. $V_1 = V_2 = v$ and also $R_2 = \alpha R_1$, $R_3 = \beta R_f$. You should find that for any non-zero common-mode input voltage v you get a non-zero output voltage, which is highly undesirable. Further, you can find the common mode gain

$$A_{CM} = \frac{V_{out}}{v}$$

You should discover that this is a linear function of $(\beta/\alpha - 1)$. The ideal for the difference amplifier is to ensure $\alpha = \beta = 1$ hence $A_{CM} = 0$

Recalling that op-amps are small and cheap, we can make use of the high-input impedance **Buffer** or **Voltage Follower** circuit. The simplest thing to do would be to just attach a buffer to each of the inputs at V_1 and V_2. This completely solves the input impedance problem – the points in the circuit we are trying to measure are now isolated from our difference

amplifier by the buffer impedances, which can be $\sim 10^9\Omega$. However, the common-mode problem (problem (1) above) remains. The reason is that the difference amplifier is trying to do two jobs: both reject common-mode signals and amplify differential mode signals. We can improve this situation by moving some of the gain into the input stage, which relaxes a bit the requirement to match the resistors. This design is known as the **Instrumentation Amplifier**.

10.2 Classic Instrumentation Amplifier Design

The design of the classic Instrumentation Amplifier is shown in figure 10.3. We can split the analysis into the input stage and the output stage, and note that the output stage is already done as it is essentially the difference amplifier we have just seen.

10.2.1 Input Stage

This is everything to the left of R_3. We can analyse this in the usual way. Rule 1 tells us $V_A = V_1$ and $V_B = V_2$. So the current in the three resistors

$$\begin{aligned} I &= \frac{V_A - V_B}{R_2} \\ &= \frac{V_1 - V_2}{R_2} \end{aligned}$$

Hence

$$V_C = V_1 + IR_1 = V_1 + \left(\frac{R_1}{R_2}\right)(V_1 - V_2)$$

$$V_D = V_2 - IR_1 = V_2 - \left(\frac{R_1}{R_2}\right)(V_1 - V_2)$$

What we can see here is that each buffer op-amp produces an output equal to its input plus an amplified version of the difference between the two inputs. So, we can say that the input stage gives us a common-mode gain $A_{CM} = 1$, and a differential mode gain $A_D \gg 1$. This is a good start.

10.2.2 Output Stage

Since the output stage is the difference amplifier we looked at earlier, we can use equation 10.1 to write

$$V_{out} = (V_D - V_C)\left(\frac{R_4}{R_3}\right)$$

Substituting in the equations for V_C and V_D

$$\begin{aligned} V_{out} &= \left(\frac{R_4}{R_3}\right)\left(V_2 - V_1 - 2(V_1 - V_2)\left(\frac{R_1}{R_2}\right)\right) \\ &= (V_2 - V_1)\left(\frac{R_4}{R_3}\right)\left(1 + 2\left(\frac{R_1}{R_2}\right)\right) \end{aligned}$$

So, the amplifier produces a voltage output which is proportional to the difference between the two input voltages. In this stage, the common-mode gain is zero. We haven't eliminated the need to choose our resistors carefully, but we have spread the amplification over the two stages of the circuit, and at each stage we maximise the differential gain, and minimise the common mode gain. It is worth going back to the definition of the instrumentation amplifier at the beginning of section 10 to check that you understand why it has all of the stated properties.

10.3 Applications

The instrumentation amplifier can be used wherever we want to measure the voltage across any circuit element without connecting either side to ground. For example we might imagine that the input of a multi-meter looks like the instrumentation amplifier because we can connect across any 2 points in a circuit without

1. drawing any current from the circuit

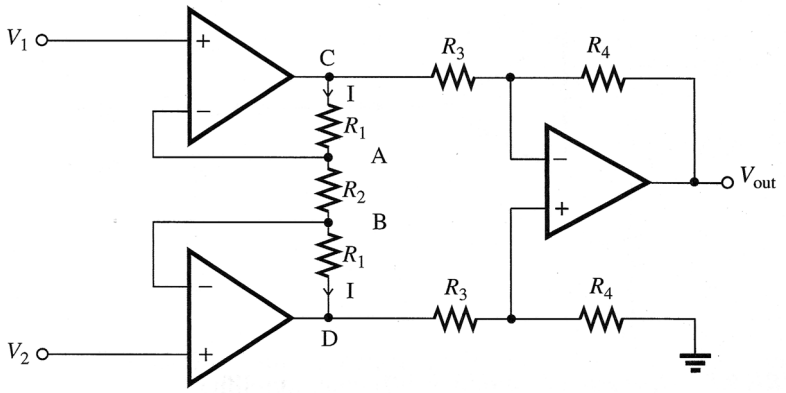


Figure 10.3: Instrumentation Amplifier (Diefenderfer chapter 9)

2. shorting the connection point to ground

The instrumentation amplifier is most useful in applications such as a strain gauge where we have a small differential signal (representing the measurement) superimposed on a large DC level on each terminal (the common-mode signal). The DC level on the terminals can fluctuate; it doesn't matter since the instrumentation amplifier only responds to the difference signal.

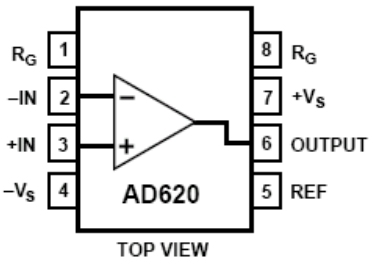


Figure 10.4: Pin Connections for the Analogue Devices AD620 Instrumentation Amplifier

10.4 Integrated Circuit Instrumentation Amplifiers

These days the semiconductor industry provides a 'ready-made' implementation on a single chip. One such device is the Analog Devices AD620 (figure 10.4). This has a CMRR (Common Mode Rejection Ratio) of 100dB and a differential gain adjustable up to 1000 set by an external gain resistor connected across pins 1 and 8 (this is R_2 in our circuit). All the other resistors are inside. The advantage of these devices is that the internal resistors have been matched to be as close as possible. This is done by laser-trimming the individual resistors for each part. Because they are manufactured in quantity, this process becomes cost effective and the AD620 retails for less than $10.

Nevertheless, especially for high-end physical science applications, sometimes you may need to build your own amplifier directly from op-amps, as described above. In measurement science, it is also useful to know a bit about what goes on inside the 'black-box'. Also, if you can understand how the instrumentation amplifier works then you have all the basic skills need to understand more or less any of the commonly-used op-amp circuits!

11 Linear Systems

We can characterise a physical system by its ability to accept an input parameter and produce an output in response. Our understanding of a sensor, or transducer, is a perfect example. we will now examine this relationship between input and output in more detail.

Systems can be linear or non-linear. A **Linear System** is one for which a linear relation exists between cause and effect. An obvious example is a spring for which force is directly proportional to extension. However, for many systems such a casual inspection will not tell us whether it is linear or **Non-linear**. Take for example the RC filter. We can satisfy ourselves, after consideration, that it is linear since at any given frequency the gain is independent of the input. So, doubling the input voltage doubles the output voltage. To determine whether a system is linear or not we need to examine the mathematical formulation of the input/output relationship. The reason we are interested in this is because we can use quite simple mathematics to understand and predict the behaviour of really very complex systems.

11.1 Linear Time-Invariant Systems

The time response to an arbitrary forcing function can be described by a differential equation of the form

$$a_0 x + a_1 \frac{dx}{dt} + a_2 \frac{d^2x}{dt^2} + \cdots a_n \frac{d^n x}{dt^n} =$$

$$b_0 y + b_1 \frac{dy}{dt} + b_2 \frac{d^2y}{dt^2} + \cdots b_n \frac{d^n y}{dt^n} \qquad (11.1)$$

where a_n and b_n are constants, $x(t)$ the output of the system and $y(t)$ a forcing function. Systems which are described by such a differential equation are often termed **Linear Time Invariant** or LTI systems. The **Time Invariance** comes from the fact that the behaviour of the system (its differential equations) does not vary with time, that is to say the behaviour, under a given input, is the same today as it was yesterday. LTI systems have two important properties.

Frequency Preservation If the input to an LTI system is a sinusoid then the output is also a sinusoid of the same frequency, but it may have its amplitude and phase modified.

Superposition If the input to an LTI system is the sum of several sinusoids then the output of the system is the sum of the system response to each individual sinusoid in isolation.

11.2 Use of Fourier Techniques

The properties of superposition and frequency preservation allow us to apply useful Fourier techniques to real world problems involving LTI systems. Any real world signal can be Fourier decomposed into a sum of sinusoids of suitable frequency, phase and amplitude, and we can follow the propagation of each sinusoid through an LTI system (each component will have its phase and amplitude modified) and then sum the output sinusoids to recover the response of the system to the original signal. We can see from these properties that if we add several LTI systems together in series then the resultant system is also LTI.

We will see how to apply these ideas later on.

The most common systems we come across in instrumentation are quite well described by rather simplified versions of equation 11.1. It is worth briefly revising how simple zero, first and second order systems respond to a simple forcing function such as a step $u(t)$ or an impulse $\delta(t)$. The aim of this section is to briefly run through some of the background mathematics, particularly for the second order system. Note that the second order system is just the damped simple harmonic oscillator, looked at from the instrumentation perspective. As

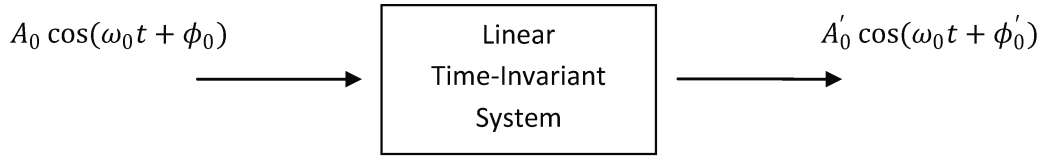


Figure 11.1: Frequency Preservation of LTI Systems

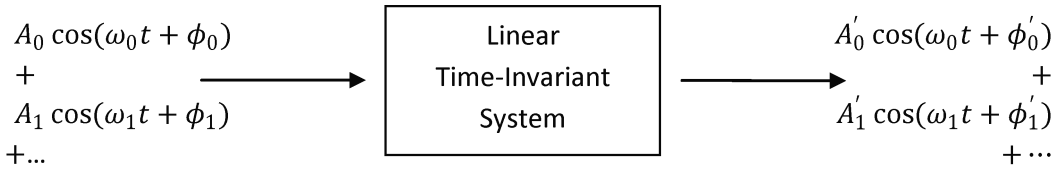


Figure 11.2: Superposition Property of LTI Systems

such you will already have covered it a number of times before in other courses. What we typically find is that whilst the differential equations describing isolated zero, first and second order systems are generally quite tractable mathematically, it can be extremely difficult to both formulate and solve a differential equation describing the behaviour of a more complex instrument composed of several distinct elements. However, as suggested above there are solutions to this based on Fourier techniques, which we will come on to later.

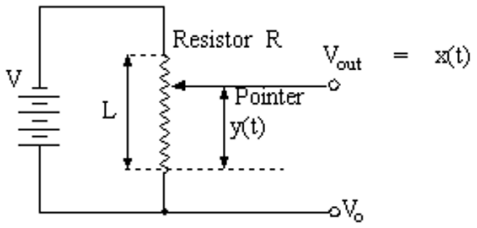


Figure 11.3: Variable Resistor (or potential divider) used to measure displacement

11.3 Zero-Order Systems

Zero order systems are described by an equation of the form

$$a_0 x = b_0 y$$

where a_o and b_o are constants, $y(t)$ is the input or forcing function applied to the system, and $x(t)$ its output. They thus represent the most trivial of LTI system. This type of system should respond immediately to a changing input in time, and has no "memory" as such. Zero order systems don't care what has happened to them in the past, only what the state of the input is now. This type of behaviour would be ideal for a sensor, we would get infinitely fast time response, and no problems with oscillation, overshoot, settling times and so forth. Unfortunately such systems are extremely rare in real life instruments, but we will look briefly at them to provide us with some scene setting examples. A good example of a zero order system would be a variable resistor (Figure 11.3). The input to the system $y(t)$ is the position of a pointer along the resistor, and the output $x(t)$ is a voltage V_{out} that appears at the end of the pointer arm. For the zero order case V_{out} exactly follows $y(t)$, and at any time,

$$V_{out} = \frac{V y(t)}{L}$$

where V is the voltage across the full length L of the resistor. Any change in $y(t)$ gives an immediate corresponding change in V_{out}. As soon

as $y(t)$ stops changing, so does V_{out}. This however is unphysical for a real world system, and in our example effects such as the slight flexing of a mechanical pointing arm as it moves or stray inductive or capacitive effects would stop V_{out} *exactly* tracking $y(t)$. More generally we will have to formulate and solve a differential equation to find the time response of a system to any given input. However under certain conditions there are some real physical systems which are zero-order. For example an idealised amplifier is zero order: the output is simply (and always) an amplified version of the input.

The zero order system is characterised by the Gain, or more generally **Static Sensitivity** (also known as **Steady State Gain**)

$$K_s = \frac{b_0}{a_0} = \frac{V}{L} \text{ Volts/m}$$

11.4 First Order Systems

First order systems are described by a differential equation of the form

$$a_1 \frac{dx}{dt} + a_0 x = b_0 y$$

The rate of change of the output $x(t)$ is thus proportional to the *difference* between the systems current state (at a given point in time) and the value of the forcing function. To solve a differential equation of this type we generally try a test solution of the form $x(t) = Ae^{st} + c$ where A, $s$and c are constants. We then apply boundary conditions to find the values of these constants. We applied this method to analyse the behaviour of the RC circuit in section 6.5. Here we'll apply it to another simple system, a thermometer plunged into a heat-bath . The rate of heat flow into the thermometer is proportional to the difference between the temperature of the thermometer $\theta(t)$ and the heat reservoir θ_R. Our boundary conditions are that at $t = 0$ the thermometer is at its starting temperature θ_0 and for $t \to \infty$ the thermometer reaches the same temperature as the heat reservoir. The solution is

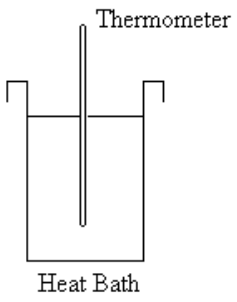


Figure 11.4: Thermometer in Heat-Bath

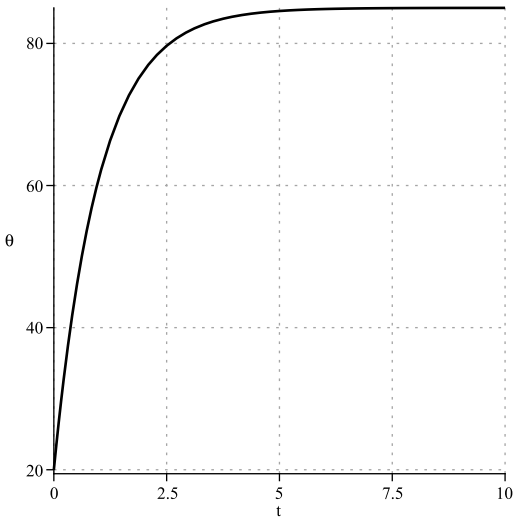


Figure 11.5: Room temperature thermometer plunged into a nice hot cup of tea

$$\theta(t) = \theta_0 + (\theta_R - \theta_0)(1 - e^{-t/\tau})$$

where τ is a time constant dependent on the conductivity of the thermometer, the heat bath, and the thermal mass of the thermometer. τ is the time taken for the system to reach 63.2% of its steady state response, after the application of a step input. This type of time response is very common in instrumentation, see the hand out on the AD950 temperature sensor for an example. We often refer to the sort of plot shown in figure 11.5 as an RC curve, as it follows the form of the voltage across a capacitor as we

charge it up from a DC source through a resistor. The system initially responds rapidly to a step input, as the difference between its starting point and the steady state value is large. As time progresses, the difference between steady state and actual value becomes smaller, and the rate of change falls off correspondingly. After a sufficiently long time the system reaches its steady state value, though for a real sensor, the time taken to reach 95% of its steady state value, or just the time constant τ will be given on the data sheet. Referring back to our defining differential equation for a first order system, we see that the steady state response (static sensitivity) is as before given by

$$K_s = \frac{b_0}{a_0}$$

On a final practical note, remember that all the elements of an instrument will each have some individual time response, and these have to be matched to each other for best results. There is no point using an ultra fast sensor if the rest of the system simply cannot keep up with it. The time response of the whole system is a convolution of all the individual time responses.

11.5 Second Order Systems

Second order systems are described by a differential equation of the form

$$a_2\frac{d^2x}{dt^2} + a_1\frac{dx}{dt} + a_0x = b_0y \qquad (11.2)$$

This as the defining equation of a damped simple harmonic oscillator. A good example of this type of system in instrumentation is a spring balance (figure 11.6, although there are many non-mechanical physical analogues). Second order systems can show the type of delayed time response to a step input that we saw in first order systems, but they can also *oscillate* after a step $u(t)$ or impulse $\delta(t)$ input. We can begin by summing all the forces that act on the mass m. There is a gravitational force mg, a force due to the extension of the spring $kx(t)$, a velocity dependent

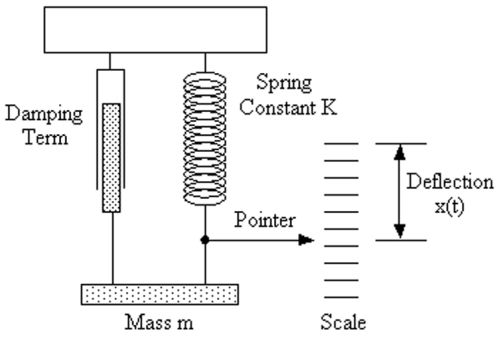


Figure 11.6: Spring Balance

damping term cdx/dt and an acceleration term md^2x/dt^2. Summing all the forces on the mass gives ΣForces = Mass . Acceleration

$$m\frac{d^2x}{dt^2} + c\frac{dx}{dt} + kx = mg$$

There are several limiting cases of the behaviour of this system that are easy to find. For the steady state case there are no changes as a function of time and $d^2x/dt^2 = dx/dt = 0$, the response is the same as a zero order system. The static response of the system is thus $mg = kx$ therefore the static sensitivity is

$$K_s = \frac{b_0}{a_0} = \frac{g}{k}$$

and therefore in the steady-state case $x = mK_s$ Note that here we take our forcing function to be m, the mass hung on the balance.

11.5.1 Solving the Second-Order Differential Equation

We want to find the solution of equation 11.2. To find a general solution of this equation, we try a test solution of the form $x(t) = Ae^{st}$, and for simplicity we will take $y = 0$ (i.e. no input, or zero forcing function). Substituting into equation 11.2 gives

$$a_2As^2e^{st} + a_1Ase^{st} + a_0Ae^{st} = 0$$

Here we use the fact that zero is a valid solution. Solving for s we find

$$\begin{aligned} s &= \frac{-a_1 \pm \sqrt{a_1^2 - 4a_2a_0}}{2a_2} \\ &= \frac{-a_1}{2a_2} \pm \sqrt{\frac{a_1^2}{4a_2a_0} - \frac{a_0}{a_2}} \end{aligned}$$

Whether the roots of this equation are zero, real or complex will have a major effect on how the system responds to a changing input. We simplify this equation by setting

$$\begin{aligned} \omega_0^2 &= \frac{a_0}{a_2} \\ \xi &= \frac{a_1}{2\sqrt{a_0a_2}} \end{aligned}$$

Substituting these two expressions into the solution gives

$$s = -\xi\omega_0 \pm \sqrt{\xi^2 - 1}$$

with two roots $s_1 = -\xi\omega_0 + \sqrt{\xi^2 - 1}$ and $s_2 = -\xi\omega_0 - \sqrt{\xi^2 - 1}$. As we have two roots the general solution for the second order differential equation is

$$x(t) = Ae^{s_1 t} + Be^{s_2 t}$$

where A and B are constants. For physically realistic values of ω_0 and ξ the solution is typically an exponentially-decaying sinusoid, as we will see in the following section. Note that here, to simplify the solution, we took $y = 0$. This is the **Zero Input Solution**, but it is not trivial since it just specifies that there is no input to the system. However the system could already be in motion ($\frac{dx}{dt} \neq 0$) due to some previous excitation. Alternatively, we could have specified that the system was stationary at some non-zero x position and then 'let-go' at $t = 0$ (this would be equivalent to pulling-down the spring-balance then releasing it at $t = 0$). We will come back to this later but for now it is important to understand that the initial conditions of the system are essential to understand future behaviour.

11.5.2 Behaviour of the Damped Harmonic Oscillator

Returning to our spring balance example, we have

$$\begin{aligned} \omega_0^2 &= \frac{a_0}{a_2} = \frac{k}{m} \\ \xi &= \frac{a_1}{2\sqrt{a_0a_2}} = \frac{c}{2\sqrt{mk}} \end{aligned}$$

ω_0 is the **Natural Frequency** of the system and ξ is the **Damping Ratio**. For the zero damping ($c = 0 \implies \xi = 0$) case any disturbance of the system will result in free running oscillations at a frequency ω_0. This type of free running behaviour is almost always a bad thing for an instrument. After all we want to use it to measure the forcing function. We will however see that allowing a little oscillation to occur can actually be a good thing for an instrument, improving its time response at the cost of some small and rapidly damped oscillations after an abrupt change to the input. The behaviour is summarised in figure 11.7 which gives the response of the balance to a step change input, i.e. suddenly placing the mass m on the balance at time $t = 0$. The plots are normalised to show a unit-response and give 6 different values of $\xi = 0, 0.01, 0.1, 0.5, 1$ and 3 (in the order left-to-right and top-to-bottom). For $\xi = 0$ we have free running at ω_0. This would be disastrous in a real balance since we want to use the **Static Response** to determine the mass, i.e. we want the balance to settle. For $\xi = 0.01$ and 0.1 the system is **Under-damped**. $\xi = 0.6$ is the kind of behaviour we might expect from a real spring-balance as it is probably the optimum: some oscillation but rapidly reaching equilibrium. A car's suspension also has this sort of behaviour. $\xi = 1$ is **Critically Damped** - there is no oscillation but the system takes quite a time to reach the equilibrium. The spring closers for doors might be designed for this sort of behaviour. $\xi = 3$ is **Over-damped**; the behaviour is reminiscent of the first-order system.

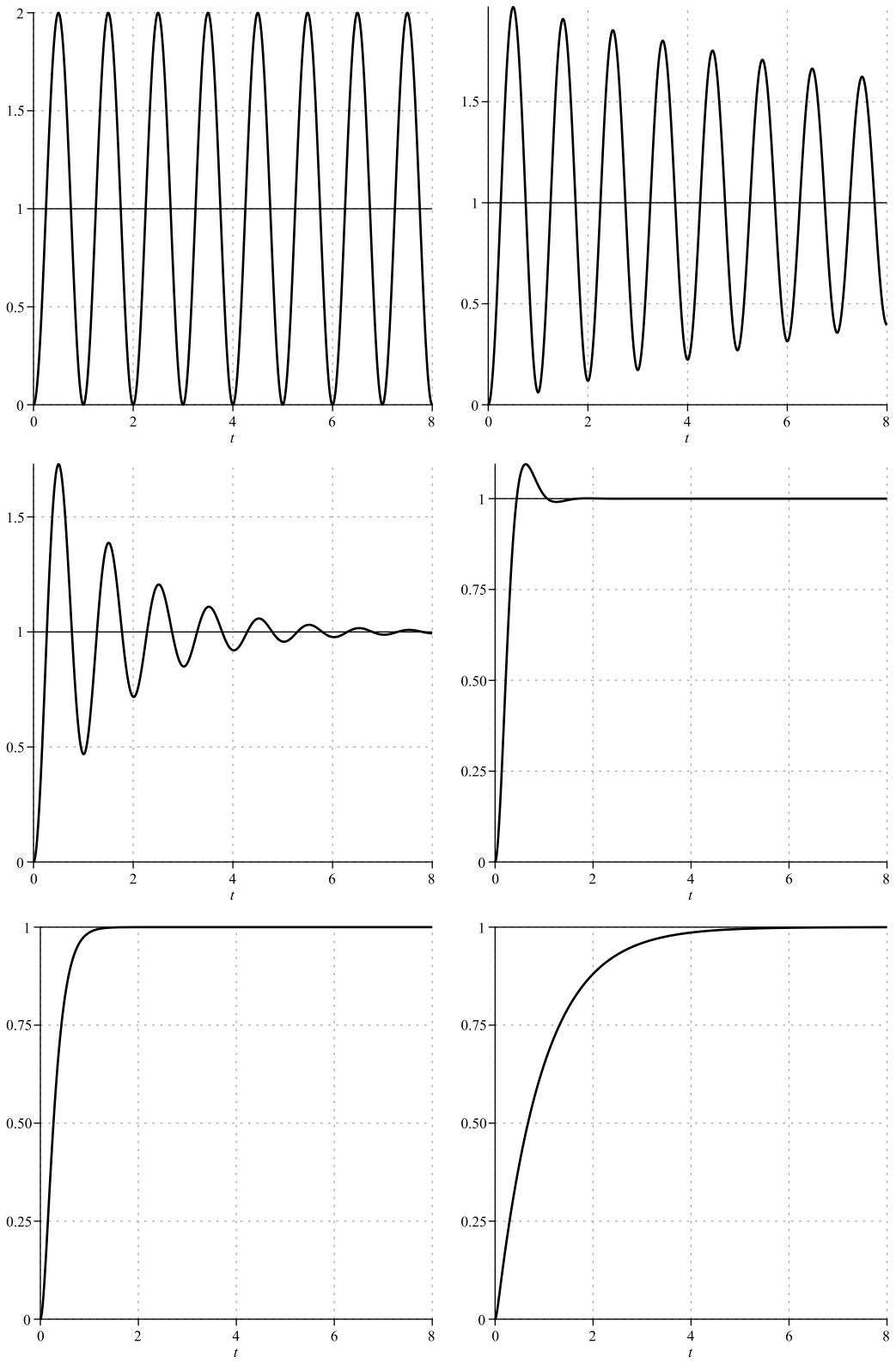


Figure 11.7: Response of the spring-balance to a unit step-change input for $\xi = 0$, 0.01, 0.1, 0.6, 1 and 3. The natural frequency is 1 Hz.

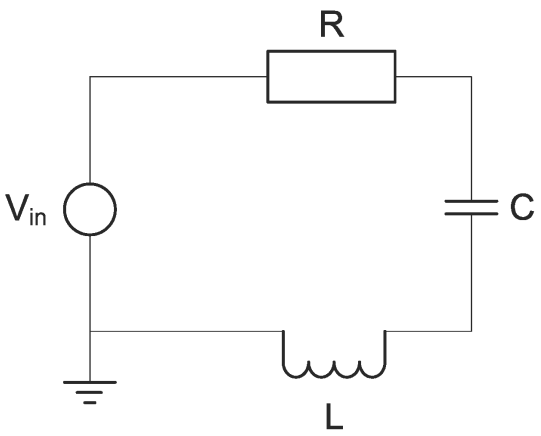


Figure 11.8: The RLC Circuit

11.5.3 The RLC Circuit

The **RLC Circuit** is the electrical analogue of the damped mechanical oscillator. In fact we can use an analogue electronic system to simulate a complex mechanical system because of this equivalence. Whilst this is quite a powerful technique it has to a great extent been superseded by numerical computer simulations these days. In the RLC circuit we have the equivalence with the damped spring balance or damped harmonic oscillator as per table 2.1

As with the mechanical spring-balance we can determine the differential equations which govern the behaviour through use of some physical laws of conservation; in this case we use Kirchhoff's Voltage Law

$$V_R + V_C + V_L = V_{in}$$

which we can re-write as

$$iR + L\frac{di}{dt} + \frac{1}{C}\int_{-\infty}^{t} i(\tau)d\tau = V_{in}$$

If we take V_{in} to be a step input at $t = 0$ then for all times $t > 0$ we can write

$$L\frac{d^2i}{dt^2} + R\frac{di}{dt} + \frac{1}{C}i = 0$$

Therefore the series RLC circuit shows the same sort of transient response as the mechanical oscillator.

$$\begin{aligned} \omega_0^2 &= \frac{a_0}{a_2} = \frac{1}{LC} \\ \xi &= \frac{a_1}{2\sqrt{a_0 a_2}} = \frac{R}{2}\sqrt{\frac{C}{L}} \end{aligned} \tag{11.3}$$

ω_0 is again the natural frequency of the circuit. We can understand the frequency-domain behaviour of the circuit best using the complex impedance approach given in section 4.5. The total impedance of the circuit is given by

$$Z = R + j\omega L - \frac{j}{\omega C}$$

Since R is fixed, the circuit has a minimum impedance for

$$j\omega L = \frac{j}{\omega C}$$

Which is the same result as equation 11.3. Further since Z is a function of frequency it is interesting to plot the response of the system as a function of frequency. Figure 11.9 shows the current flowing in the circuit as a function of frequency. For this figure, we take $L = R = C = 1$ and the applied voltage is 1 in amplitude and swept in the range $\omega = 0.01 \ldots 100$. The **Resonance** at $\omega_0 = 1$rad/s is clearly visible[15]. At the resonance, the reactive part of the impedance is zero, so we have only resistive impedance, hence the current is 1 amp. At low frequencies the impedance $Z \to \infty$ because of the capacitor and at high frequencies $Z \to \infty$ because of the inductor. Therefore the overall shape of the curve is physically understandable. The mechanical oscillator would also show a similar behaviour in

[15] Note that the Natural Frequency is the *un-driven* oscillation frequency, as we saw for the spring-balance where $y = constant$. For the RLC circuit we found the Resonant Frequency under *driven* conditions (swept sinusoid input). In general for second-order systems the natural frequency and resonant frequency are close but not identical. But the RLC circuit is one where they *are* identical.

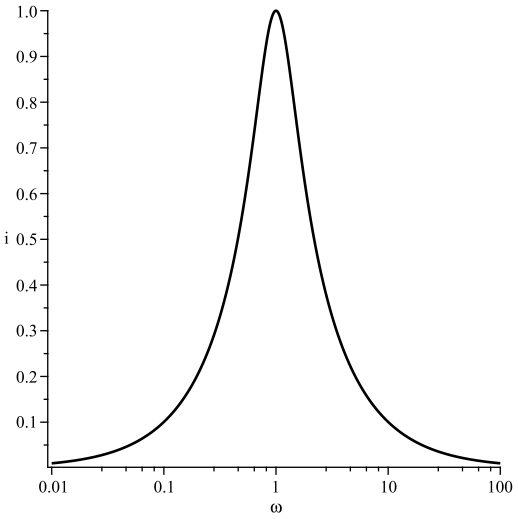


Figure 11.9: Current flowing in the RLC circuit as a function of frequency

response to a swept-sinusoidal forcing function (force). Moreover, the RLC circuit will show similar time-domain response to a step-input.

12 Fourier Transform of Non-periodic Signals

Often we want to represent a single discrete pulse such as the rectangular pulse $P_a(t)$ (see section 2.3.2). We said in section 7.5 that this is possible by specifying the pulse as a *periodic function* but only applicable over a limited range (for example a square wave with period $T = 2a$ but defined only on the interval $0 < t < T$). This is frequently inconvenient, and in any case we may want a mathematical representation that is genuinely zero for all time other than the period when the function acts. For example the statement that "a pulse $P_a(t)$ acts on a system" requires that the input is zero for all $|t| > a$. Of course we could calculate the Fourier series for the pulse, but we would find (in contrast to periodic functions) that we need a seriously large number of terms in the series. The solution then is the **Fourier Integral**, which is after all just the limiting case of the Fourier Series as the period $T \to \infty$.

12.1 Fourier Transform of a Continuous-Time Function

The **Fourier Transform** of a continuous function $f(t)$ is written $\mathcal{F}\{f(t)\} = F(\omega)$ and the inverse is $f(t) = \mathcal{F}^{-1}\{F(\omega)\}$. So we have the **Forward Transform**:

$$\mathcal{F}\{f(t)\} = F(\omega) = \int_{-\infty}^{\infty} f(t)e^{-j\omega t}dt \quad (12.1)$$

and the **Inverse Transform**:

$$f(t) = \mathcal{F}^{-1}\{F(\omega)\} = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)e^{j\omega t}d\omega \quad (12.2)$$

$F(\omega)$ is the **Spectrum Function** of $f(t)$. If $F(\omega)$ is complex we would usually represent it as a plot of the real and imaginary parts versus ω. Sometimes it is more useful to plot the absolute value $|F(\omega)| = \sqrt{\Re\{F(\omega)\}^2 + \Im\{F(\omega)\}^2}$ and the argument $\operatorname{Arg} F(\omega) = \tan^{-1}(\Im\{F(\omega)\}/\Re\{F(\omega)\})$ in which case $|F(\omega)|$ is called the **Amplitude Spectrum** and $\operatorname{Arg} F(\omega)$ is the **Phase Spectrum**.

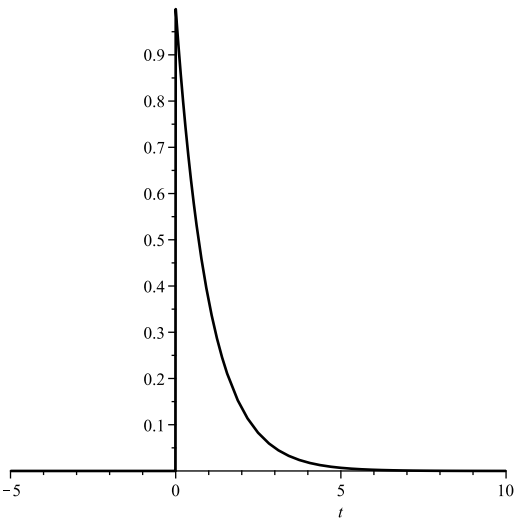


Figure 12.1: $f(t) = u(t)e^{-t}$

12.2 Fourier Transform of Real Functions

Physical signals are real continuous functions of time. Consider the function $f(t) = u(t)e^{-t}$ (figure 12.1) which has the Fourier transform

$$F(\omega) = \frac{1}{1 + j\omega}$$

The real, imaginary, magnitude and argument representations of the spectrum are given in figure 12.2.

There are some rules for real functions which are useful for our studies

Real Functions The reflected form of the spectrum (i.e. reflected about the $\omega = 0$ axis) is the complex conjugate of the spectrum

$$F(-\omega) = F^*(\omega)$$

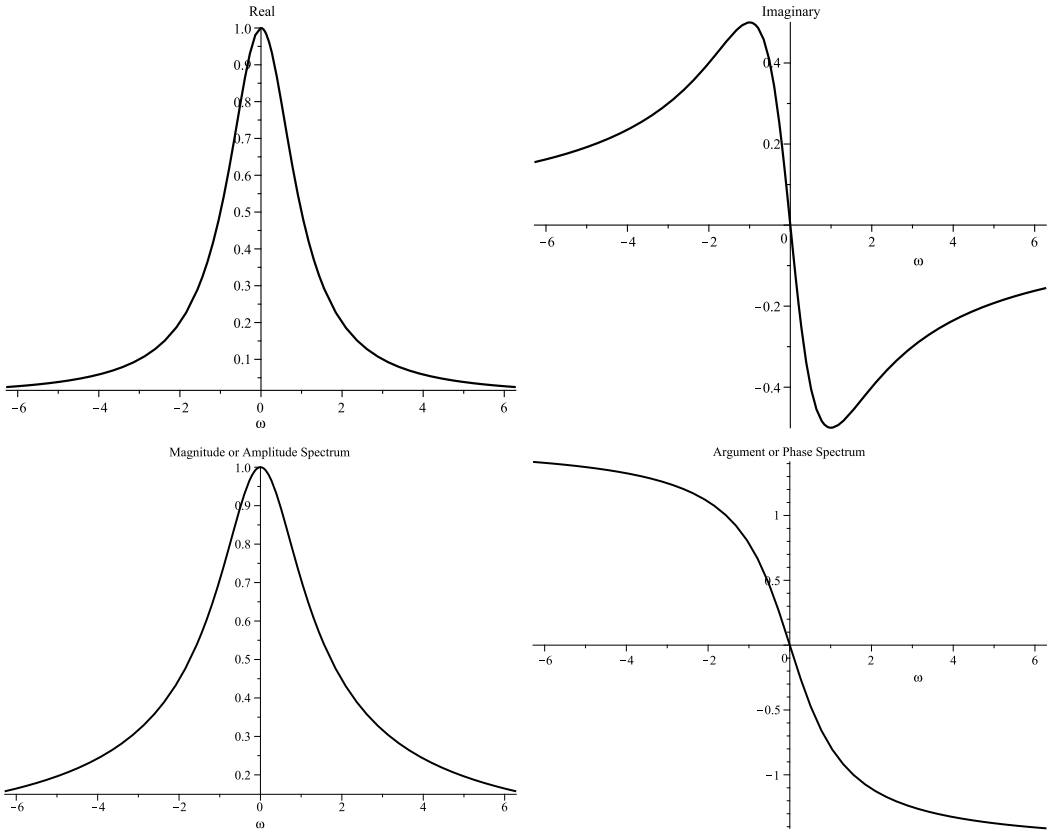


Figure 12.2: $F(\omega)$ representations in terms of $\Re\{F(\omega)\}$ and $\Im\{F(\omega)\}$ (top two panels) and Magnitude and Phase spectrum (bottom two panels)

Real and Even Functions The imaginary part of the spectrum is zero, and we can calculate the real part by

$$\begin{aligned} \Re\{F(\omega)\} &= 2\int_0^\infty f(t)\cos(\omega t)\,dt \\ \Im\{F(\omega)\} &= 0 \end{aligned}$$

Real and Odd Functions Conversely

$$\begin{aligned} \Re\{F(\omega)\} &= 0 \\ \Im\{F(\omega)\} &= -j\int_{-\infty}^\infty f(t)\sin(\omega t)\,dt \end{aligned}$$

Since, for most practical purposes, we can *choose* where to take the origin, these rules can simplify calculation of the Fourier transform (see also the Time Shift .

property below)

12.3 Properties of the Fourier Transform

Linearity

$$\mathcal{F}\{af_1(t) + bf_2(t)\} = aF_1(\omega) + bF_2(\omega)$$

Time Shift

$$\mathcal{F}\{f(t \pm t_0)\} = e^{\pm j\omega t_0}F(\omega)$$

Frequency Shift

$$\mathcal{F}\{e^{\pm j\omega_0 t}f(t)\} = F(\omega \mp \omega_0)$$

Scaling

$$\mathcal{F}\{f(at)\} = \frac{1}{|a|}F\left(\frac{\omega}{a}\right)$$

Derivative

$$\mathcal{F}\left\{\frac{d^n f(t)}{dt^n}\right\} = (j\omega)^n F(\omega)$$

12.4 Fourier Transform of some Common Signals

The Pulse function and the delta function are both important signals for instrumentation applications. We may be trying to measure the height and/or width of a pulse generated by in an experiment. In order to understand how our instrument will react to the pulse, we need an understanding of the pulse's frequency content. The delta function, because of its special frequency content, is frequently used as a 'test' input into a system in order to determine a system's *impulse response*. We will deal with this later, for now it is important to appreciate the frequency content of these signals.

Pulse Function The Fourier transform of the pulse function (time-domain) is the sinc function (frequency domain), and the Fourier transform of the sinc function (time domain) is the pulse function (frequency domain)

$$\begin{aligned}\mathcal{F}\{P_a(t)\} &= sinc_a(\omega)\\ \mathcal{F}\{sinc_a(t)\} &= P_a(\omega)\end{aligned}$$

This result means that the pulse function, which is finite in time, has an infinite range of frequencies associated with it. Conversely, the sinc function in the time domain, which exists for all time $-\infty < t < \infty$ has only a finite range of frequencies. Since all instrumentation has frequency-dependent behaviour, the implications of this are profound, as we'll see later. It is well worth sketching these for a few different values of a.

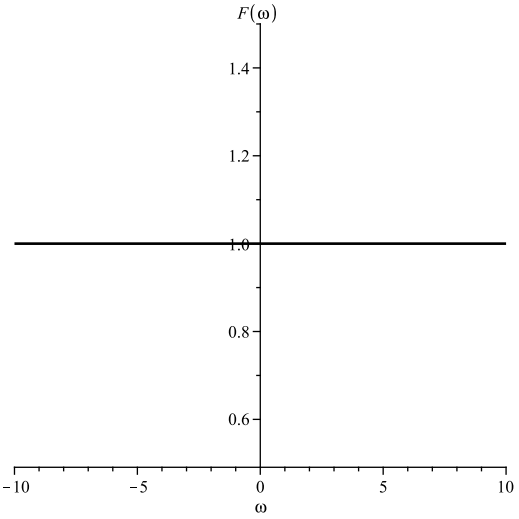


Figure 12.3: Fourier Transform of the Delta Function $\delta(t)$

Delta Function The Fourier transform of the delta-function is a constant, and *vice-versa*

$$\begin{aligned}\mathcal{F}\{\delta(t)\} &= 1\\ \mathcal{F}\{A\delta(t)\} &= A\\ \mathcal{F}\{A\} &= 2\pi\delta(\omega)\end{aligned}$$

Again the implications are profound. The delta-function (figure 12.3) contains *all frequencies in equal measure*. A physical example of a delta function input to a system is hitting a mass with a hammer. The hammer transfers a defined amount of energy to the mass in a very short period of time (ideally applying infinite force for an infinitely short period of time such that the total momentum transferred is 1). This means that the hammer blow excites all frequencies *simultaneously*. We can use this to test a system's frequency response. Mechanical engineers do this in practice: strike the object to be tested with a hammer and record the spectrum of frequencies which results. The peaks in the spectrum are where the object resonates.

The physical meaning of $\mathcal{F}\{A\} = 2\pi\delta(\omega)$ is also important. A constant (or DC) input to a system is a delta function at $\omega = 0$ in the frequency

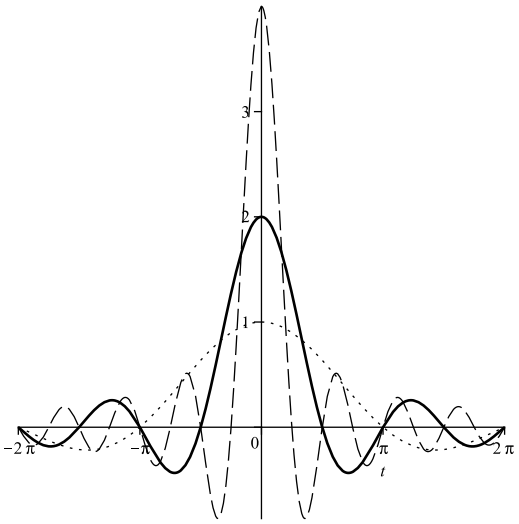


Figure 12.4: Fourier Transform of the Pulse Function $P_a(t)$ for $a = 1$ (solid line), $a = 2$ (dashed line) and $a = 1/2$ (dotted line) [Note: horizontal axis should be labelled ω]

domain. This means it is a constant frequency and that frequency is zero.

Sinusoid

$$\mathcal{F}\{A\cos(\omega_0 t)\} = \pi A[\delta(\omega-\omega_0)+\delta(\omega+\omega_0)]$$

The sinusoid is of course a single frequency which is represented by the delta-function. Mathematically it exists at both $\pm\omega_0$, though of course only positive frequencies exist in reality. It is useful to sketch the above[16].

12.5 Frequency Content of Pulsed Signals

Figure 12.4 shows the Fourier transform of the pulse function $P_a(t)$ for various values of a. The spectrum is the sinc function. Observe that the lowest frequency at which the function crosses the time axis is at $\omega = \pi/2a$. This means that as a increases (the pulse gets wider) then the spectrum of the pulse gets narrower. In the limit as $a \to \infty$ then as we expect from the previous section, the frequency function becomes infinitely narrow and tends to a delta function.

We can state this as a general rule: very short pulses (in the time domain) have very wide spectra, whilst very wide, flat, pulses have nicely compact spectra. We can formalise this general rule using fundamental physical considerations deriving from the uncertainty principle. Ultimately this leads to a dimensionless quantity called the **Time-Bandwidth Product**. We will cover this in lectures.

Note that the maximum height of the spectrum is at zero frequency and is given by

$$\begin{aligned} F(\omega) &= 2\frac{\sin a\omega}{\omega} \\ \lim_{\omega\to 0} F(\omega) &= 2a \end{aligned}$$

So as the pulse gets narrower, its spectrum gets both wider and less tall. This is also physically intuitive; as the pulse gets shorter in time the energy it has gets less, therefore in the frequency domain the total energy must also be conserved by the spectrum becoming wide but low in amplitude (see also section 12.6 following).

As a final point, it is worth considering what would happen if, as the pulse gets narrower, we let it get taller such that the total *area* of the pulse remains 1. Then we would find that as $a \to 0$ then the pulse height $\to \infty$, so we have a delta function in the time domain. In the frequency domain we will see that this is a constant (see figure 12.3) and from the above arguments and figure 12.4 it is clear that this is the limiting case of the sinc function as $a \to 0$[17].

12.6 Parseval's Theorem

To extend what we saw in section 7.9 for Fourier series to the Fourier integral, Parseval's theo-

[16] Figure 4.26 of Poularikas gives a very nice pictorial representation of the Fourier representation for many other common signals. It is well worth browsing this to get a feel for the behaviour of signals in the frequency domain.

[17] There is a nice worked example of this in Poularikas (Example 4.19 in section 4.3). It is well worth looking this up.

rem gives us the energy relation for the time and frequency domains

$$\begin{aligned} E &= \int_{-\infty}^{\infty} |f(t)|^2 dt \qquad (12.3) \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} |F(\omega)|^2 d\omega \end{aligned}$$

The energy in the time domain is the same as the energy in the frequency domain, as a consequence of conservation of energy. The quantity

$$\Phi(\omega) = \frac{1}{2\pi} |F(\omega)|^2$$

is known as the **Energy Spectral Density** of the signal such that the energy in a infinitesimal band $d\omega$ is $\Phi(\omega)d\omega$ and we can get a measure of the energy in a range or band of frequencies as

$$\Delta E = \int_{\omega_1}^{\omega_2} \Phi(\omega) d\omega$$

This is useful when, for example, an instrument limits our frequency range to $\omega_1 < \omega < \omega_2$ and we wish to know how much of the signal's energy is preserved.

Note that there is a problem with equation 12.3 in that any periodic signal such as a sinusoid exists for all time and hence has (mathematically) infinite energy. Consequently you will often see the term **Power Spectral Density** which has the meaning of signal power per unit frequency. It is calculated in a rather different way to energy spectral density, but it can be handled in a similar way, that is to say if $\Phi(\omega)$ is a **PSD** then we can get the power in the signal by integrating over the frequency range.

As a further note, the interpretation of energy and power in signals in this manner is possible because, if for example $f(t)$ is a voltage and we assume it acts on a load of 1Ω then

$$\text{Power} = \frac{V^2}{R} \propto f(t)^2$$

That is to say both $|f(t)|^2$ and $|F(\omega)|^2$ are *proportional to* power, but we have to be careful with the units if we are doing numerical calculations.

13 Laplace Transform

The integral definitions of the Fourier and Laplace transforms look rather similar at first glance. Both involve an integral of a function multiplied by an exponential term over time. However there are some fundamental differences between these two transforms, they tell us different things about physical systems, and we apply them in rather different ways.

13.1 Integral Definitions for the Fourier and Laplace Transforms

Fourier Transform

$$\mathcal{F}\{f(t)\} = F(\omega) = \int_{-\infty}^{\infty} f(t)e^{-j\omega t}dt$$

We have transformed a function of time $f(t)$ into a function $F(\omega)$ in the frequency domain. Note that the integral is symmetric in time and involves both positive and negative time.

Laplace Transform

$$\mathcal{L}\{f(t)\} = F(s) = \int_{0}^{\infty} f(t)e^{-st}dt$$

Here we say that we have transformed a function of time $f(t)$ into a function $F(s)$ in the s domain, where s is a complex variable

$$s = (\sigma + j\omega)$$

The integral is **Single-Sided**: it only involves values of time greater than zero.

13.2 Physical interpretation of the Transforms

In general when we Fourier transform a real physical observable the resulting function also represents a physical observable (though the raw transform might also contain negative frequencies that don't carry useful information). For example when we Fourier transform a pulse shape $f(t)$, a function of time, we find the spectrum $F(\omega)$ required to support that pulse shape. You will have met this concept before in other courses such as optics. In contrast, the Laplace transform does not generally represent a useful physical observable, although it is an effective tool for understanding how systems behave. As an example of this type of idea in other areas of physics, remember that a wave function in quantum mechanics might include a complex term, and that we cannot directly interpret. However we still use complex numbers as a useful mathematical device with which to tackle a wide range of problems involving paired variables (amplitude and phase for example), and we can recover physical observables at the end of the problem.

When we analyse a complex device in instrumentation, we can find ourselves having to deal with a series of linked differential equations and an arbitrary forcing function. The usefulness of the Laplace transform lies in its ability to change a differential equation in the time domain into an algebraic expression in the s-domain. After the transformation we find the problem rather easier to solve, and we can then transform back to the time domain to recover the real world behaviour of the system. We can also use the Laplace transform to examine how a system behaves after the application of an impulse input $\delta(t)$ or a step $u(t)$. This allows us to know whether the system is stable or not, a problem of importance in instrumentation.

13.3 Fourier Analysis of Systems

The use of the Fourier Transform in systems analysis requires the decomposition of the forcing function $y(t)$ into its spectrum function $Y(\omega)$. This requires an integral *over all time*. This means that in practice we need to know the value of the forcing function over the entire past history of the system and for all future time. If we do know this then, given that we know the transfer function of the system $G(\omega)$, we can get the spectrum of the system output $X(\omega)$ from the relation

$$G(\omega) = \frac{X(\omega)}{Y(\omega)}$$

This is the defining relation for the **Transfer Function**. We can get the system response in the time domain by inverse transforming

$$x(t) = \mathcal{F}^{-1}\{X(\omega)\}$$

This approach works perfectly well for systems where the input signal is a continuous repetitive input such as a sinusoid, which for all practical purposes we can take to be infinite in time. This is not just theoretical: in the lab sessions we will see how we can model the transfer functions of the electret microphone, the piezo-sounder, the filter etc in terms of their frequency response $G(\omega)$. This works because for practical applications we can assume that the input signal is infinite. We know that it is not, but it has existed for long enough that any transient or "start-up" response of the system has long since ceased to be apparent in the output of the system. That is to say, the system output is the same as it would have been had the input existed for all time.

13.4 Advantages of the Laplace Transform

In systems analysis we often want to understand the response of the system to a single pulse applied at the input at a time usually taken to be $t = 0$. This is known as **Transient Analysis** and the Fourier Transform has some limitations for this

1. The Fourier integral does not converge for some functions such as the unit step $y(t) = u(t)$
2. The transfer function can involve integrals which are difficult to evaluate
3. The system must be **Initially Relaxed**

The last of these means that, at time $t = 0$, the system has no motion (mechanical system) or no currents/voltages (electrical system). In many practical cases this is not a problem (when we switch on the Elvis prototyping board we know that our circuit is initially relaxed) however there are situations where we might want to understand how our circuit responds to a step input, and intuitively we know that the response will be very different if the capacitors are initially charged, or if currents are already flowing.

The solution to all these problems is the **Laplace Transform**. Because it is a single-side integral, we only need to evaluate it for $t > 0$. This means we don't care about the input to the system for $t < 0$. As long as we know

1. The input for $t > 0$ and
2. The initial conditions of the system

we can evaluate how the system will evolve for all time $t > 0$. This is a physically realistic situation. For a real system, we know the state of the system *now* and we apply an input *now*; we want to observe how the system will evolve.

As a final note, the Laplace Transform makes the mathematics of differential equations rather easy to solve. A differential equation in the time-domain becomes a simple algebraic expression in the s-domain.

13.5 Properties of the Laplace Transform

The most important properties of the Laplace Transform are its linearity and behaviour under differentiation and integration. These properties are described below.

Linearity Like the Fourier Transform, the Laplace transform is a linear process

$$\mathcal{L}\{K_1 f_1(t) + K_2 f_2(t)\} = K_1 F_1(s) + K_2 F_2(s)$$

First Time Derivative

$$\mathcal{L}\left\{\frac{d}{dt}f(t)\right\} = sF(s) - f(0+)$$

$f(0+)$ is the **initial condition** of the system. It is the value of $f(t)$ as $t \to 0$ from positive t. Recall that we don't know (or care) about f for $t < 0$.

Second Time Derivative

$$\mathcal{L}\left\{\frac{d^2}{dt^2}f(t)\right\} = s^2F(s) - sf(0+) - f^{(1)}(0+)$$

$f^{(1)}(0+)$ is the first derivative of $f(t)$ evaluated as $t \to 0+$

Integral with Zero Initial Conditions

$$\mathcal{L}\left\{\int_0^t f(\xi)d\xi\right\} = \frac{F(s)}{s}$$

Integral with Initial Conditions

$$\mathcal{L}\left\{\int_0^t f(\xi)d\xi\right\} = \frac{F(s)}{s} + \frac{f^{(-1)}(0+)}{s}$$

where

$$f^{(-1)}(0+) = \lim_{t\to 0+}\int_{-\infty}^{0} f(\xi)d\xi$$

Some further properties of the transform are

Frequency Shifting

$$\mathcal{L}\{f(t-\lambda)u(t-\lambda)\} = e^{-s\lambda}F(s)$$

The $u(t-\lambda)$ here is used to ensure f is zero for all $t < \lambda$

Scaling

$$\mathcal{L}\left\{f\left(\frac{t}{a}\right)\right\} = aF(as)$$

For $a > 0$

13.6 Elementary Laplace Transform Pairs

The forward transform is relatively straightforward to derive for many common functions. The inverse transform, by contrast, is given by

$$\mathcal{L}^{-1}\{F(s)\} = f(t) = \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty} F(s)e^{st}ds$$

and therefore requires integration in the complex plane. Don't try this at home! In order to perform most inverse and forward transforms, it is usually sufficient to simplify expressions and use tables of elementary transform pairs such as the one given in figure 13.1.

Entry No.	$f(t) = \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty} F(s)e^{st}\,ds$	$F(s) = \int_0^\infty f(t)e^{-st}\,dt$
1	$\delta(t)$	1
2	$u(t)$	$\frac{1}{s}$
3	$t^n \quad \text{for} \quad n > 0$	$\frac{n!}{s^{n+1}}$
4	e^{-at}	$\frac{1}{s+a}$
5	te^{-at}	$\frac{1}{(s+a)^2}$
6	$\frac{t^{n-1}e^{-at}}{(n-1)!}$	$\frac{1}{(s+a)^n}$
7	$\frac{1}{b-a}(e^{-at} - e^{-bt})\ a \neq b$	$\frac{1}{(s+a)(s+b)}$
8	$-\frac{1}{b-a}(ae^{-at} - be^{-bt})\ a \neq b$	$\frac{s}{(s+a)(s+b)}$
9	$\sin \omega t$	$\frac{\omega}{s^2+\omega^2}$
10	$\cos \omega t$	$\frac{s}{s^2+\omega^2}$
11	$e^{-at}\sin \omega t$	$\frac{\omega}{(s+a)^2+\omega^2}$
12	$e^{-at}\cos \omega t$	$\frac{s+a}{(s+a)^2+\omega^2}$
13	$\sinh \omega t$	$\frac{\omega}{s^2-\omega^2}$
14	$\cosh \omega t$	$\frac{s}{s^2-\omega^2}$
15	$\frac{\sqrt{a^2+\omega^2}}{\omega}\sin(\omega t + \phi),\ \phi = \tan^{-1}\frac{\omega}{a}$	$\frac{s+a}{s^2+\omega^2}$
16	$\frac{\omega_n}{\sqrt{1-\zeta^2}}e^{-\zeta\omega_n t}\sin(\omega_n\sqrt{1-\zeta^2})t \qquad \lvert\zeta\rvert < 1$	$\frac{\omega_n^2}{s^2+2\zeta\omega_n s+\omega_n^2}$
17	$\frac{1}{a^2+\omega^2} + \frac{1}{\omega\sqrt{a^2+\omega^2}}e^{-at}\sin(\omega t - \phi),$ $\phi = \tan^{-1}\left(\frac{\omega}{-a}\right)$	$\frac{1}{s[(s+a)^2+\omega^2]}$
18	$1 - \frac{1}{\sqrt{1-\zeta^2}}e^{-\zeta\omega_n t}\sin(\omega_n\sqrt{1-\zeta^2}\,t + \phi),$ $\phi = \cos^{-1}\zeta,\ \lvert\zeta\rvert < 1$	$\frac{\omega_n^2}{s(s^2+2\zeta\omega_n s+\omega_n^2)}$

Figure 13.1: Elementary Laplace Transform Pairs (from Poularikas and Seeley)

14 Solving Problems with the Laplace Transform

Laplace transforms allow us to find the time response of complex systems driven by an arbitrary forcing function, often without having to directly formulate and solve the differential equation. We use them for solving initial value problems that start at a time $t = 0$. This is in contrast to the Fourier transform that deals in integrals over both positive and negative time. This initial value problem however is what we actually deal with on most experiments. We set up some apparatus and then throw a switch at a starting time $t = 0$. We then want to know what happens for positive time. To formulate and solve Laplace transforms for simple systems we can just look up the relevant time domain functions and their s-domain transforms in a table (see figure 13.1). For a more complex system we will probably need to simplify the Laplace transform representation somewhat before we can do this. The technique we quite commonly use to do this involves partial fractions (see section 14.5).

Note that the Laplace transform of the delta function $\delta(t)$ is 1. This means that the time response of a system subject to an impulse at its input is given by the inverse Laplace transform of the transfer function. This is one of the reasons why the delta function is so commonly used in analysis of systems. Another way of understanding this is to recall that the Fourier Transform of a $\delta(t)$ function is an infinite flat spectrum, so a perfect impulse contains all frequencies in equal amounts. In other words, when we apply an impulse to a system we stimulate it with all frequencies *simultaneously*. In practical applications, such as the Elvis Bode Analyser, it is inefficient to stimulate the system with all frequencies at once (and indeed impossible to generate a mathematically *true* impulse, so the Bode Analyser sweeps the frequency input in order to build-up the Transfer Function.

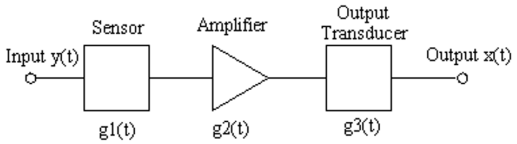


Figure 14.1: Instrument System Consisting of three Functional Blocks in Series

14.1 Step-Input Response for a Series System

> *An instrumentation system is constructed from several stages as shown in figure 14.1 Determine the response of the instrument to a unit-step applied to the input*

We begin by determining the individual response of the sensor, amplifier and transducer to the $\delta(t)$ function. This is the Impulse Response which can be determined either analytically or experimentally.

For the instrument in the figure let us say we determine the impulse responses in the time-domain, that is to say we have measured the input and output transducers' responses to impulse inputs. As mentioned previously, this is not always the most efficient method, but it is sometimes used in reality, especially for certain mechanical/acoustic transducers. In this case let us say we measure the impulse responses to be:

- Sensor $g_1(t) = e^{-t}$
- Output transducer $g_3(t) = e^{-2t}$

Let us further say that we have tested the amplifier and found it to have a voltage gain of 4. If we assume that it is an ideal amplifier then (as a zero-order system) it will have an impulse response given by:

- Amplifier $g_2(t) = 4\delta(t)$

What is the output of the system as a function of time for an input step $y(t) = u(t)$? First

we need to formulate a transfer function for the complete system. We know the time domain responses of the individual components and from these we can find their transfer functions. From our table of Laplace transforms (13.1)

$$\begin{aligned} G_1(s) &= \frac{1}{s+1} \\ G_2(s) &= 4 \\ G_3(s) &= \frac{1}{s+2} \end{aligned}$$

Now we can find G(s), the transfer function for the whole system.

$$\begin{aligned} G(s) &= G_1(s)\,G_2(s)\,G_3(s) \\ &= \frac{4}{(s+1)(s+2)} \end{aligned}$$

The input to the system is a unit step function $y(t) = u(t)$, and the Laplace transform of this is simply $1/s$. To find the time response of the system we recall that $X(s) = G(s)Y(s)$ where $X(s)$ and $Y(s)$ are the Laplace transforms of the input and output of the system, and $G(s)$ the transfer function.

$$\begin{aligned} x(t) &= \mathcal{L}^{-1}\{X(s)\} \\ &= \mathcal{L}^{-1}\{G(s)\,Y(s)\} \\ &= \mathcal{L}^{-1}\left\{\frac{4}{s(s+1)(s+2)}\right\} \end{aligned}$$

Our task now is to evaluate the inverse Laplace transform of this function, and we do this for our example using partial fractions. We write our s-domain function as a sum of partial fractions (see section 14.5)

$$X(s) = \frac{4}{s(s+1)(s+2)} = \frac{A_0}{s} + \frac{A_1}{s+1} + \frac{A_2}{s+2}$$

This is an identity and valid for all values of s therefore we can choose values of s which make some of the constants disappear. Multiplying through by s and evaluating for $s = 0$ gives

$$\left.\frac{4}{(s+1)(s+2)} = A_0 + \frac{A_1 s}{s+1} + \frac{A_2 s}{s+2}\right|_{s=0}$$

$$A_0 = 2$$

Similarly for A_1 we can multiply through by $s+1$ and evaluating for $s = -1$

$$\left.\frac{4}{s(s+2)} = \frac{A_0(s+1)}{s} + A_1 + \frac{A_2(s+1)}{s+2}\right|_{s=-1}$$

$$A_1 = -4$$

And for A_2 we can multiply through by $s+2$ and evaluating for $s = -2$

$$A_2 = 2$$

Putting these constants back into our identity for $X(s)$

$$X(s) = \frac{2}{s} - \frac{4}{s+1} + \frac{2}{s+2}$$

This is now in a form that we can simply write down a time response for, with reference to our Laplace transform tables. If we do this for the individual components of $X(s)$ we find

$$x(t) = 2 - 4e^{-t} + 2e^{-2t}$$

This function is plotted in figure 14.2 for $t = 0 \ldots 8$ seconds. Note that it shows the sort of time behaviour we expect from first order systems subject to a step input (or a heavily damped second order system). The response to the step is an initial rise, which then slows down, and tends to some steady state value for large t.

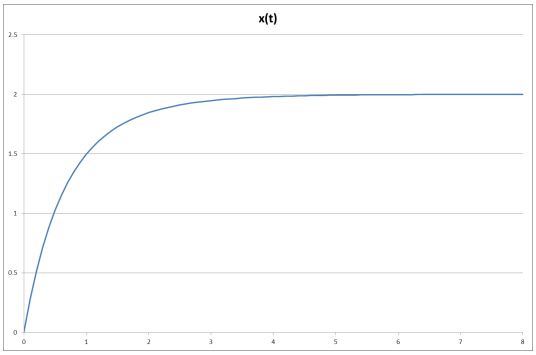


Figure 14.2: System Response to a step-input $y(t) = u(t)$

14.2 Solving for Impulse Response with a Known Transfer Function

Find the time-domain impulse-response of a system with a transfer function

$$G(s) = \frac{s+3}{(s+2)^2}$$

To find the time-domain response of the system we note $g(t) = \mathcal{L}^{-1}\{G(s)\} = \mathcal{L}^{-1}\{X(s)\}$ since $y(t) = \delta(t)$, $Y(s) = 1$ and therefore

$$g(t) = \mathcal{L}^{-1}\left\{\frac{s+3}{(s+2)^2}\right\}$$

As before we need a partial fraction expression to simplify the inverse transform. For denominators of order >1 we must write the PF as

$$\frac{s+3}{(s+2)^2} = \frac{A_2}{(s+2)^2} + \frac{A_1}{s+2}$$

To find A_2 we multiply $G(s)$ by $(s+2)^2$ and set $s = -2$ to get $A_2 = 1$

We now need to find A_1, however there is a problem with continuing the method. To find this coefficient we cannot use $s = -2$ again as we have already used this to find A_2. Instead we can differentiate $G(s)(s+2)^2$ (see section 14.5)

$$\frac{d}{ds}G(s)(s+2)^2 = \frac{d}{ds}(s+3) = \frac{d}{ds}(A_2 + A_1(s+2))$$

$$1 = A_1$$

Therefore

$$G(s) = \frac{1}{(s+2)^2} + \frac{1}{s+2}$$

Which can be easily inverse transformed into

$$g(t) = te^{-2t} + e^{-2t}$$

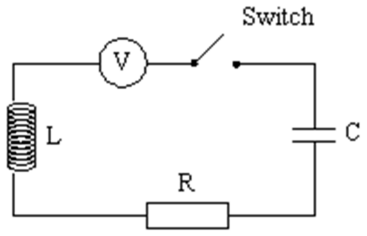


Figure 14.3: Switched LCR Circuit

14.3 Switched Electrical Circuit

The two previous examples were a little divorced from reality. We'll now use the Laplace transform to analyse a simple electrical circuit. One of the major uses of the Laplace transform in electrical engineering is for initial value problems where we throw a switch, or turn on a circuit. What we then want to know is the behaviour of the circuit for all times after we throw the switch.

Figure 14.3 shows a simple LCR series circuit with a voltage source V and a switch. If the voltage source was a sine wave generator, and the switch had been closed for a long time then

we could use our usual expressions for the complex impedance (essentially a Fourier approach) to find the current and voltage for various components in the circuit. However here the applied voltage is a constant V and is applied once at time $t = 0$. AC circuit analysis is not appropriate for this sort of transient input problem where we want to know "what happens immediately after we throw the switch?" We can use the Laplace transform to answer this.

> *For a DC voltage source* $V = 300\,\mathrm{V}$*, inductor* $L = 2\,\mathrm{H}$*, capacitor* $C = 0.02\,\mathrm{F}$ *and resistor* $R = 16\,\Omega$ *find an expression for the charge on the capacitor* $q(t)$ *and the current* $I(t)$ *in the circuit for all times* $t > 0$ *after the switch is closed. Assume that the circuit is initially relaxed.*

We begin by determining the differential equation which governs the circuit. We can write the voltage across each of the components as

- Voltage across the inductor $= L\frac{dI}{dt}$
- Voltage across the resistor $= IR$
- Voltage across the capacitor $= {}^{q}/_{C}$

Using Kirchhoff's voltage law to sum the voltages around the series circuit we can write (for time $t > 0$, i.e. *after* the switch is closed)

$$L\frac{dI}{dt} + IR + \frac{q}{C} = V \tag{14.1}$$

Since $I = {}^{dq}/_{dt}$ (we know that the instantaneous current is the same everywhere in the circuit, due to conservation of charge, or Kirchhoff's current law). Also as $V = 0$ for all time $t < 0$ we can write this as

$$Vu(t) = L\frac{d^2q}{dt^2} + R\frac{dq}{dt} + \frac{q}{C} \tag{14.2}$$

The differential equation is second order in time therefore we may expect that the system may be capable of oscillation as well as tending to some steady-state condition at some time t much after we close the switch. To solve the DE, we take the Laplace Transform

$$\begin{aligned}\frac{V}{s} &= L(s^2Q(s) - sq(0+) - \frac{dq}{dt}(0+)) \\ &+ R(sQ(s) - q(0+)) + \frac{Q(s)}{C}\end{aligned} \tag{14.3}$$

Here we see how the initial conditions of the system are automatically taken care of by the Laplace transform. Since the circuit is initially relaxed, we start with $q = 0$ and $\dot{q} = \ddot{q} = 0$, so we can further simplify the algebraic expression

$$\frac{300}{s} = 2s^2Q(s) + 16sQ(s) + 50Q(s)$$

$$Q(s) = \frac{150}{s(s^2 + 8s + 25)}$$

Again using partial fractions, here we must write

$$\frac{150}{s(s^2 + 8s + 25)} = \frac{A}{s} + \frac{Bs + C}{s^2 + 8s + 25}$$

Multiplying by s and setting $s = 0$ yields $A = 6$. Multiplying by $s(s^2 + 8s + 25)$, taking ${}^{d}/_{ds}$ and setting $s = 0$ yields $B = -6$. Again taking ${}^{d}/_{ds}$ and setting $s = 0$ yields $C = -48$.

$$Q(s) = \frac{150}{s(s^2 + 8s + 25)} = \frac{6}{s} - \frac{6s + 48}{s^2 + 8s + 25}$$

This requires further simplification before we can used the standard transforms

$$\begin{aligned}Q(s) &= \frac{6}{s} - \frac{6(s+4) + 24}{(s+4)^2 + 9} \\ &= \frac{6}{s} - \frac{6(s+4)}{(s+4)^2 + 9} - \frac{24}{(s+4)^2 + 9}\end{aligned}$$

We can now use entries 2, 11 and 12 in figure 13.1

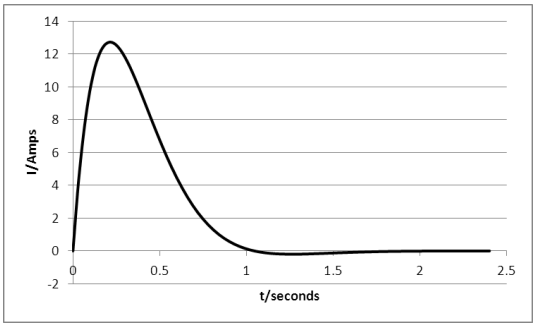


Figure 14.4: Time Response of the Switched LCR Circuit

$$q(t) = 6 - 6e^{-4t}\cos(3t) - 8e^{-4t}\sin(3t)$$

Now we have $q(t)$ we can differentiate to find the current

$$I(t) = \frac{dq}{dt} = 50e^{-4t}\sin(3t)$$

This is illustrated in figure 14.4. Note that this is the kind of response we expect from a damped harmonic oscillator. The result is physically meaningful: at the instant at which the switch is closed, the rate of change of voltage across the capacitor is very large so a large current will flow, with the rate of change of current limited by the inductor. Electrical energy is stored in the electric field across the capacitor and in the magnetic field of the inductor. The two exchange energy in an oscillatory fashion like a mechanical oscillator exchanging kinetic and potential energy. The $\sin 3t$ term reflects this. As the system evolves, electrical energy is converted to heat energy in the resistor, so the amplitude of the response decays (the exponential term). For large t the applied voltage (the forcing function V) is a constant and the voltage across the capacitor tends to a constant hence the impedance of the capacitor tends to ∞ and the current tends to zero.

In this example we started from the understanding that the system was "initially relaxed". This is usually the starting point for most systems. When we switch-on the Elvis prototyping board, we assume that the voltages and currents are zero, and this is usually right, however consider in this case that someone has snuck into the lab overnight and charged the capacitor up to 100V. The charge on the capacitor will stay put until the switch is closed, and we can imagine that the $I(t)$ response will be very different! The defining differential equation for the circuit is still valid (see equations 14.1 and 14.2) and we can see that the transformation into the s-domain (equation 14.3) automatically includes an initial value for q.

14.4 Other Applications

We've seen here how Laplace methods allow complex systems to be solved by transforming differential equations in the time-domain into algebraic expressions in the s-domain. The fact that the Laplace transform is single-sided and takes into account initial conditions means that it is very well suited for solving transient analysis problems in mechanical and electrical engineering. These methods are applicable across the physical and engineering disciplines. As mentioned in lectures, biological systems (including complex feedback loops) can be modelled. Other applications are in economics, where the parameter to be determined could be price and the constants relate to supply and demand. In general, any continuous-time system represented by ordinary differential equations can be studied, especially where we want to set the system up in some starting configuration and observe how it evolves with time.

As a final note, you should be aware that there is an equivalent to the Laplace transform for discrete (quantised) signals called the Z-transform. This is of central importance in the field of digital signal processing, though we won't be able to cover this in the course.

14.5 Partial Fractions

Consider

$$\frac{2}{5} - \frac{3}{4} + \frac{1}{2} = \frac{8 - 15 + 10}{20}$$

Where 20 is the lowest common multiple. Partial Fraction Decomposition is the reverse of this process. In general, any rational function

$$\frac{P(s)}{Q(s)}$$

can be re-written using partial fractions. First, it is essential to check that the degree of the numerator is less than the degree of the denominator. If not, it is necessary to first divide-out by long division. This is tedious so we will avoid it in this course, and in general the sort of real-world problems we will encounter naturally lead a higher degree of s in the denominator. Proceed as follows:

For each factor of $Q(s)$ in the form $(as + b)^m$ introduce terms

$$\frac{A_1}{as + b} + \frac{A_2}{(as + b)^2} + \cdots \frac{A_m}{(as + b)^m}$$

For factors of the form $(cs^2 + ds + e)^n$ introduce terms

$$\frac{C_1 s + D_1}{cs^2 + ds + e} + \frac{C_2 s + D_2}{(cs^2 + ds + e)^2} + \cdots \frac{C_n s + D_n}{(cs^2 + ds + e)^n}$$

There are various methods to solve for the constants but remember that

$$\frac{P(s)}{Q(s)} = \frac{A_1}{as + b} + \frac{C_1 s + D_1}{cs^2 + ds + e} + \cdots$$

is an identity valid for all values of s. We can re-arrange and substitute values of s which cause some unknowns to disappear. For example in

$$X(s) = \frac{s + 3}{(s + 2)^2} = \frac{A_1}{s + 2} + \frac{A_2}{(s + 2)^2}$$

we can multiply $X(s)$ by $(s + 2)^2$and set $s = -2$ to get A_2. To get A_1 we can differentiate $X(s)(s+2)^2$ with respect to s and then use $s = -2$ a second time. Alternatively we can gather terms in powers of s to get a set of simultaneous equations

$$X(s)(s + 2)^2 = s + 3 = A_1(s + 2) + A_2$$

Gather terms:

$$0 = s(A_1 - 1) + (A_2 + 2A_1 - 3)$$

From which

$$\begin{aligned} A_1 - 1 &= 0 \\ A_2 + 2A_1 - 3 &= 0 \end{aligned}$$

Whichever method is best depends on the form and complexity of the functions.

15 Stability in LTI Systems

We briefly looked at stability in the context of feedback amplifiers. We saw how a negative feedback system can go into unstable positive feedback when the phase change at the output tends to 180°. In an amplifier this is very bad, since the feedback then tends to *increase* the gain rather than decrease it. The output tends to oscillate, with the amplifier output switching between $\pm V_{ss}$ (the supply voltage) at the oscillation frequency.

Fortunately, if we have a mathematical model of our system then we can test for stability by exciting all frequencies at the input using a delta function. We find three cases

1. Stable: The output responds, comes to rest and stays where it is. In this case, the system responds to the impulse-input by moving to a new equilibrium state. This is typical of many first-order mechanical or thermal systems - the impulse puts a certain amount of energy into the system which moves to a new (stable) equilibrium.
2. Asymptotically Stable: The output returns to the original equilibrium state. An amplifier would do this.
3. Unstable: The output grows (or shows growing oscillation) until it reaches some physical limit

15.1 Poles and Zeros of the Transfer Function

Writing the transfer function as

$$G(s) = \frac{P(s)}{Q(s)}$$

- Values of s such that $P(s) = 0$ are **Zeros** i.e. $G(s) \to 0$ as $P(s) \to 0$
- Values of s such that $Q(s) = 0$ are **Poles** i.e. $G(s) \to \infty$ as $Q(s) \to 0$

In real (physical) systems the transfer function $G(s)$ is always a ratio of 2 polynomials with real coefficients. This results in the poles and zeros being either

- Real *or*
- Complex-conjugate pairs

We can check for stability by seeing where these values of s lie on the complex plane. Generally

- Roots in the $\sigma < 0$ side of the plane are stable
- Roots in the $\sigma > 0$ side of the plane are un-stable

15.2 Complex Roots

Sometimes roots will be complex, and as stated before in real physical systems these always appear as complex conjugate pairs Assuming we can expand the Transfer Function as

$$G(s) = \sum_{k=1}^{n} \frac{A_k}{s - s_k}$$

Then the the complex-conjugate roots will be of the form

$$s_k = \sigma_k \pm j\omega_k$$

So we get a set of solutions of the form

$$g_k(t) = A_k e^{s_k t} + A_k^* e^{s_k^* t}$$

Since $A_k = a + jb$ and $s_k = \sigma_k + j\omega_k$ we can simplify to

$$\begin{aligned} g_k(t) &= 2\sqrt{a^2 + b^2} e^{\sigma_k t} \cos(\omega_k t) + \beta_k \\ \beta_k &= \tan^{-1} \frac{b}{a} \end{aligned}$$

These results are illustrated graphically in figure 15.1. In the top panel, a real negative root

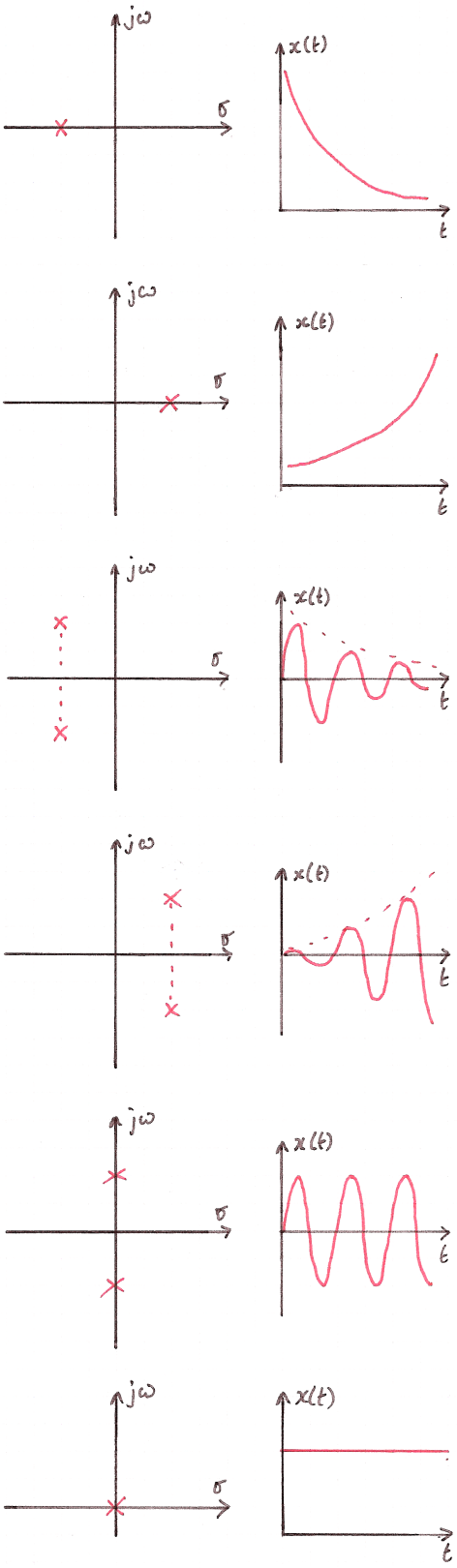


Figure 15.1: Graphical Illustration of Roots in the Complex Plane

exists and the system is asymptotically stable. In the second panel the system is unstable as the output grows without limit. In the third panel complex-conjugate roots exist to the left of the $\sigma = 0$ axis. The result is damped oscillation and the system is asymptotically stable. With complex roots on the RHS of the axis we have growing oscillation and the system is unstable. The last two panels show some special cases. With purely imaginary roots we have sustained oscillation, which is nonetheless stable. If the root is at $s = 0$ then the response is a constant, which is stable.

16 Bandwidth

We've seen in section 12 how there is an intimate relationship between the duration of a pulse in the time domain and its range of frequencies in the Fourier transform. The latter is generally known as the **Bandwidth** of the pulse. More generally, bandwidth is any range of frequencies and is consequently measured in Hz or rad/s. Bandwidth is the frequency-domain equivalent of 'duration' in the time domain. We have talked about continuous functions such as a voltage $V(t)$ having an instantaneous value at time t. While mathematically correct we recognise that any real measurement of V takes a finite time τ and that, typically, our measuring instrument might perform an averaging function on the signal, so we can write

$$V_\tau(t) = \frac{1}{\tau}\int_{t'=t}^{t'=t+\tau} V(t')dt'$$

Top-end equipment can sample signals on a timescale of ns, but there is always some finite measurement time. Consequently there is always a finite range of frequencies associated with any measurement. An average over time in the time-domain is equivalent to an average over frequencies in the frequency domain.

16.1 Pulse Width and Bandwidth

It is tempting to describe the width of a pulse in the time domain as its non-zero time range, however many signals we can describe mathematically such as the Gaussian-pulse would technically have infinite duration, even if this is not physical. An easier way to get a 'measure' of the pulse is to take the Full-Width at Half Maximum (FWHM), that is the width of the pulse measured at half its maximum height[18].

[18] See also See http://hyperphysics.phy-astr.gsu.edu/HBASE/math/gaufcn2.html

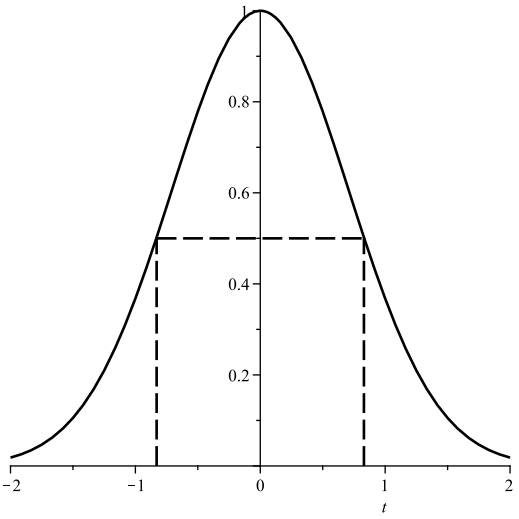


Figure 16.1: Full-Width at Half-Maximum for a Gaussian-shaped pulse

A point to note: in electronics, we measure the *amplitude* of parameters such as voltage and current. Therefore typically we measure the FWHM of the pulse amplitude. The Frequency-domain equivalent would be the FWHM of the amplitude spectrum (the Fourier transform). However in optics the amplitude of the electric field is not a physical observable, so by convention we take the FWHM of the pulse *intensity* or *power*, that is the square of the amplitude. Therefore in the frequency domain we define the bandwidth of the pulse as the FWHM of the power spectrum (the magnitude of the amplitude-transform squared).

16.2 Bandwidth for a Filter

If we are using a Bode plot (see section 6.2) to characterise a filter's frequency response then frequently we will summarise this numerically as the *bandwidth* of the filter. Here, this is the range of frequencies where the filter gain is within 3dB of the maximum value. For a low-pass filter (section 6.4) this means the bandwidth is the range of frequencies from 0 to ω_c. If we consider the RLC circuit (section 11.5.3) to be a Bandpass Filter then the bandwidth

is the range of frequencies within 3dB of the peak. Note that -3dB is a factor of $1/\sqrt{2}$ in amplitude. Effectively then, ω_c for a filter is the frequency at which the output power is half the input power (at the same frequency).

16.3 Time-Bandwidth Product

A central theme in the instrumentation course, communications, optics and many other branches of physics is that a signal that has a finite duration in time must therefore have associated with it some finite spread of frequencies. For any pulse shape we choose, and with a given duration (FWHM), we can calculate the spectrum by taking the Fourier transform and thus get the **Spectral Width** (FWHM). Therefore, for any given pulse-shape, the **Time-Bandwidth Product** is a constant that we can calculate.

It is important to note that the definition of $\Delta\nu$ and Δt we are using here is the FWHM definition. If we take a Gaussian shaped pulse (section 2.3.5) then mathematically it has infinite duration in the time-domain and also also an infinite range of frequencies. However if we use the FWHM definition then we can calculate the Time-Bandwidth product to be finite (and about 0.44, see example)

Example: Ultra-Fast Laser Pulse The duration of a laser-pulse is by convention taken to be the FWHM of the pulse's optical intensity (amplitude squared, $f(t)^2$), and its bandwidth is then the FWHM of the power spectrum (amplitude spectrum squared, $|F(\omega)|^2$). Using these definitions of duration and bandwidth, we can calculate the time-bandwidth product for any given pulse shape. For the Gaussian, this turns out to be about 0.44 (the proof of this is left as a problem sheet question). If we wish to produce a Gaussian-shaped pulse of duration 10 fs, we discover that we require a bandwidth of 44 THz!

16.4 Application to Very Short Pulses

An alternative way of thinking about this is from the point-of-view of an instrument with limited bandwidth (as all instruments have, fundamentally). This means that the instrument can only reproduce signals over a finite range of frequencies, and ultimately this limits how short the pulse can be in the time domain. This is a practical consideration of special importance in the field of ultra-fast laser physics.

The time-bandwidth product is dependent on the shape of the pulse. Physically, realisable pulses with values as low as 0.3 are possible, and some specialist lasers are able to generate these pulses with durations close to this limit. This is not just of academic interest (see box below)

Application: Optical Fibre Communications To get get the maximum data throughput on a optical fibre link it is necessary that the pulses (representing the data) are both short and very close together. Typically a limiting factor is chromatic dispersion (frequency-dependent propagation speed). This results in the pulse shape getting distorted as it travels down the fibre, as different frequency components travel at different speeds. Ultimately, the problem is that the pulses merge together in such a way that the receiver can not decode the original train of pulses. For a given pulse duration, *transform-limited* pulses (see next section) are those with the minimum possible spectral width. In optical communications, a transmitter emitting close to transform-limited pulses minimises the effect of chromatic dispersion, thus maximising the possible transmission distance.

Real pulses are of course only approximations to the mathematical ideal. The 'quality' of a very-short pulse is measured by how close we can get to the ideal time-bandwidth product.

16.5 Fourier Transform Limit

The **Fourier Transform Limit** gives us a useful tool that we can apply in the lab to test the validity of measurements such as the duration of an ultra-short laser pulse. We say that a pulse or signal is *transform limited* if it contains (in the frequency domain) exactly the minimum range of frequencies required to support the pulse shape. Pulses with more frequencies than are required by the transform limit are physically possible, but those with less are not. One important result that follows from the transform limit is that that any signal that contains a sudden change, (delta functions, step functions, square waves and so forth) has associated with it a large spread of frequencies. Real instruments cannot deal with infinite frequency ranges; they always have some finite bandwidth. This means that while perfect pulses, steps, square-waves and so forth provide us with useful mathematical tools for analysing instruments, we never actually get to see them in real life.

Physics imposes a fundamental limit on how small the time-bandwidth product can be. We can show that this in both a purely classical formulation and (rather pleasingly) by using quantum mechanics as well. If we apply a combination of a Fourier and an inverse Fourier transform to an arbitrary function of time $f(t)$ we find that there is a fixed relationship between the temporal width Δt and the bandwidth $\Delta\nu$ such that the time bandwidth product $\Delta\nu\Delta t$ must obey the inequality

$$\Delta\nu\Delta t \geq \frac{1}{4\pi}$$

If we look to quantum mechanics and the energy/time uncertainty principle

$$\Delta E\Delta t \geq \frac{\hbar}{2}$$

we see an identical result. Dividing the uncertainty principle result through by h we obtain the same transform limit which we can obtain classically (through the Fourier transform). The time bandwidth product of a real pulse is almost without exception $> 1/4\pi$. By Fourier transforming real world pulse shapes such as the Gaussian $Ae^{-a^2t^2}$ and calculating $\Delta\nu\Delta t$ we can obtain numerical values for the transform limit in such cases[19].

[19] There is a good amount of useful information about ultra-fast laser physics at http://www.rp-photonics.com/encyclopedia.html. Start by by using the site search for "Transform Limit"

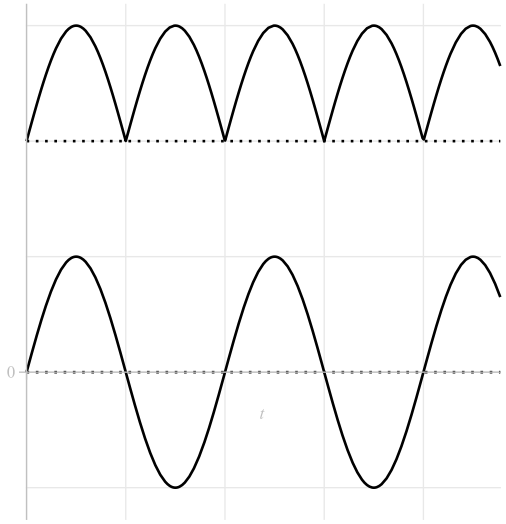


Figure 17.1: Sinusoid (lower trace) and its Rectified Version (upper trace). In both cases the zero-level is shown by the dotted line (the rectified trace has been shifted up for clarity)

17 Signal Rectification

Sometimes, we need to generate a **Rectified** version of a signal, that is to say a signal where the current only ever flows in one direction. Typically, we'll look at the voltage of the rectified signal and we find that any of the +ve going parts of the waveform are preserved and any -ve going parts of the signal have been reflected so they are now +ve. Figure 17.1 shows this for a sinusoid (lower trace), with the rectified version being consistently above the zero-level (mathematically this is just $|\sin(\omega t)|$). Rectified signals are useful in many places and we'll see one particularly clever application in the next section, but here we'll concentrate on the basic implementation.

17.1 Diode Rectifier

Since the diode only conducts current in one direction it is ideally suited for the job. We can make a simple **Full-Wave Rectifier** (which produces an output as per figure 17.1) using a simple circuit of 4 diodes[20].

Example: Power Supply We might want to build a power-supply to charge-up a battery-operated device. This means converting 240 V AC mains to 5 V DC. This is done in three steps. Firstly, a transformer steps the voltage down from 240 V AC to 5 V AC. Secondly, a rectifier gives us a signal per figure 17.1. Thirdly the output is **Smoothed** using a capacitor. The output is more-or-less a constant voltage, with some **Ripple**[a]. Here, we're using the capacitor as a charge-storage device.

[a] See Horowitz and Hill section 1.27

17.2 Synchronous Rectifier

There are problems with diode rectifiers, for example we get a voltage drop across each diode of about 0.6 V, which means that if we are trying to rectify small AC voltages then we can get a significant reduction in the output amplitude[21]. As previously mentioned, op-amps provide a simple, cheap and effective alternative to many traditional techniques, and rectification is no exception. The principle can be seen by looking again at figure 17.1; the input waveform is fine when it is +ve but must be inverted when it swings -ve. That is to say, we need to run it through an amplifier with Gain=1 for +ve voltages and Gain=-1 for -ve voltages. Another way of looking at this is shown in figure 17.2; here we see that the input signal (top trace) needs to be *multiplied* by a square wave *at the same frequency* (middle trace) to achieve the rectified signal (bottom trace). It is tempting to try and design a circuit to directly multiply two signals together, but this turns out to be quite hard. However we've seen that op-amps can be used to multiply by a constant Gain, so the technique here is to use the square wave to switch the Gain between ± 1. A schematic

[20] See Horowitz and Hill section 1.26

[21] Figure 17.1 shows idealised rectification, however if you look in detail at Horowitz and Hill figure 1.71 you will see that they have taken the small voltage drop into account.

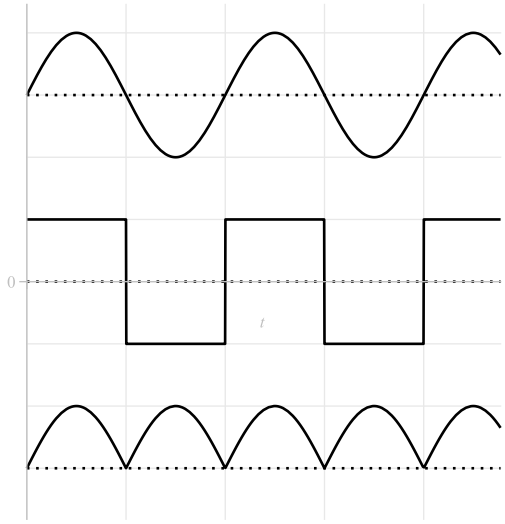


Figure 17.2: Sinusoid Multiplied by a Square Wave of the same Frequency (and with no phase difference)

arrangement for achieving this is given in figure 17.3. The circuit is covered in lectures, and you can also see this in Horowitz and Hill figure 15.37[22].

In figure 17.3 the square wave is called the *reference signal* (we'll see why later) and it is used to control a switch, which switches the signal from inverted to non-inverted and back again each time the reference signal changes sign. We'd describe this as an *edge-triggered* switch, since the transition happens on each rising or falling-edge of the reference signal. This kind of arrangement is quite common, for example a mechanical relay would be fine, though these days there are ready-made electronic switches which will do the job much faster. We'll take it as read that the switch is a ready-made component and not be concerned any further with it. The main point to take from this is that *if the reference and input are of the same frequency (and zero phase difference) then the output is a rectified version of the input.*

This requirement that the two inputs be at the same frequency seems quite difficult. However in many real-life applications such as switched-mode power converters[23] where the two signals are derived from a common source so you know they are going to be the same.

The fact that this sort of rectifier takes two inputs leads to an application called the *Phase Detector* which is a key technique in instrumentation and finds application in more-or-less every field of science.

[22] Though we don't need the output RC filter yet, this will come later when we look at phase detectors

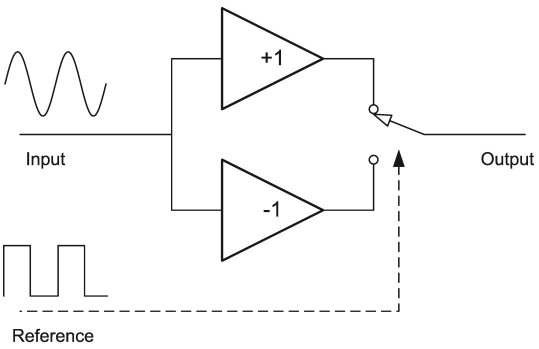


Figure 17.3: Schematic of the Synchronous Rectifier

[23] You don't need to know any details about this, it's just an example

18 Phase Detector

To quote Horowitz and Hill:

"This is a method of considerable subtlety"

Which is their way of saying that while it looks simple it is actually quite complicated to fully understand. Because of its importance, though, we will go through the topic in some detail.

The **Phase Detector** simply consists of a synchronous rectifier with a low-pass filter at the output. The circuit for this is given in lectures. The importance of the phase detector will be seen when we consider what happens if the two input signals are either out-of-phase, different frequencies (or-both). First, we need to take a look at the function of the low-pass filter

18.1 Averaging Circuit

Here we use the low-pass filter to perform an *average* of the rectified signal. We have seen how, for repetitive input signals of high frequency ($\omega \gg 1/RC$) the low-pass filter *approximates* to an integrator according to

$$V_{out} = \frac{1}{RC}\int V_{in}dt$$

Now this breaks down for our rectified signal since it is always positive and hence the integral tends to infinity, which is clearly not physical since our output amplitude can never be larger than the input (we have no amplifier here!). However, intuitively we can see that integrating a signal over some time τ then dividing the result by τ yields the average. If we integrate over one time constant ($\tau = RC$) then

$$V_{out} = \frac{1}{\tau}\int_0^{\tau} V_{in}dt = \langle V_{in}\rangle$$

This makes sense from a physical point of view[24]: when the input voltage to the filter is greater than the voltage on the capacitor then the capacitor is charging up; when the input is less then the capacitor is discharging. This charge/discharge cycle will be seen as ripple on the capacitor voltage, but for high frequency inputs ($\omega \gg 1/RC$) the ripple is very small and the capacitor voltage will approximate the average input, i.e. $V_{out} \approxeq \langle V_{in}\rangle$.

Returning now to the phase detector, we can treat this as a system with two adjustable input variables:

1. The phase difference between the two inputs
2. The (relative) frequencies of the two inputs

We'll now proceed to investigate the behaviour of these in turn

18.2 Behaviour with a Phase Difference

This is best understood graphically, as presented in figure 18.1. Note that in all the cases presented here, the two input have *exactly the same frequency*. In the first panel, there is zero phase difference. Note that in all these figures, the waveforms from top to bottom are: input signal, reference signal, rectified signal and averaged (output) signal[25]. For zero phase difference we get a constant positive output which turns out to be $2/\pi$ times the amplitude of the input signal. Inverting the input signal ($\phi = \pi$) gives us a negative output of the same size. For

[24] It may be more intuitive to think about this in the frequency domain: consider the Fourier components of the input signal and recall that the filter only lets through the very low frequency components. Note that the rectified signal has a DC component (referring back to 7 we see that this means there is a non-zero value of A_0) which represents the long-term average of the signal.

[25] For information, these figures have been generated using Maple. The input signal is defined as $V_{in} = \sin(t+\phi)$, the reference signal is defined as a piecewise continuous function representing a square wave and the rectified signal is $V_{rect} = V_{in} \times V_{Ref}$ The output of the filter is computed by integrating V_{rect} over 10 cycles of the input signal. The number 10 is chosen more or less arbitrarily here, however in a real system this would define the required value of RC to be used in our circuit since the integration time equals RC.

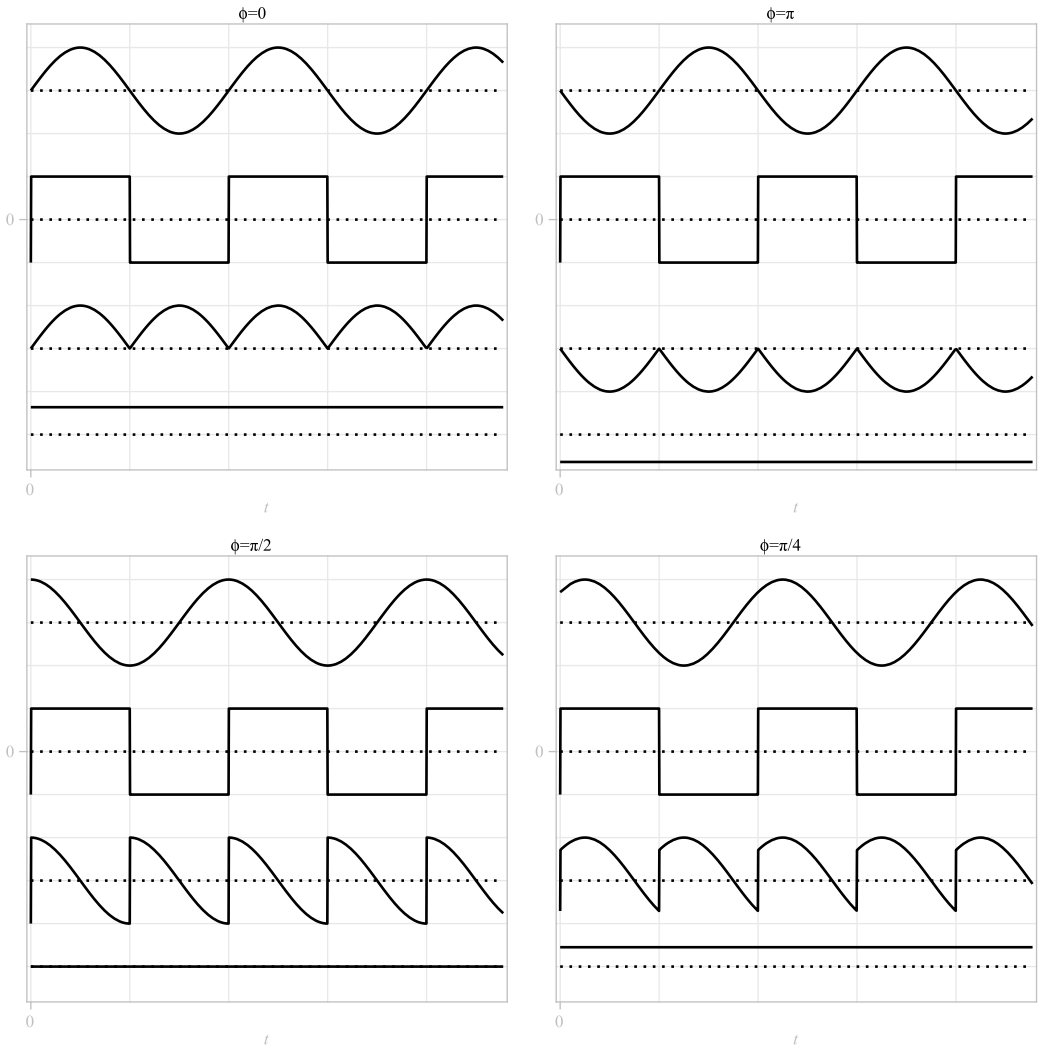


Figure 18.1: Phase Detector behaviour with 4 different values of phase difference ϕ between input and reference signal. Note that in all these figures, the waveforms from top to bottom are: input signal, reference signal, rectified signal and averaged (output) signal.

the $\phi = \pi/2$ case we get a zero output and for the "in-between" case of $\phi = \pi/4$ we get an in-between output. Clearly then the phase detector (sometimes called the Phase *Sensitive* Detector) is highly sensitive to differences in phase between the two signals. Note visually how the averaging circuit gives the appropriate output.

For an input signal $V_{in} = A\sin(\omega t + \phi)$ the output of the phase detector is

$$V_{out} = \frac{2A}{\pi}\cos\phi \tag{18.1}$$

This can be checked by performing the rectification/integration directly as

$$V_{out} = \int_0^{\pi/\omega} V_{in}dt - \int_{\pi/\omega}^{2\pi/\omega} V_{in}dt$$

In summary then, when the two inputs have

the same frequency, the phase detector gives a voltage output which is a function of the phase difference between the two. This in itself is very important and has applications which we won't go into here, because the most useful function of the phase detector becomes apparent when we consider how the circuit behaves as we vary the *relative frequency* of the two inputs.

18.3 Behaviour with a Frequency Difference

In this situation we define two different frequencies ω_{input} for the input signal and ω_{ref} for the reference. This is illustrated with the four plots in figure 18.2. In the first panel we see the case $\omega_{input} = \omega_{ref}$ and $\phi = 0$ (which is the same as the first panel in figure 18.1). For $\omega_{input} = 2\omega_{ref}$ we can clearly see that the output will always be zero, whatever the phase difference between the two signals. If we choose an irrational ratio such as $\omega_{input} = 1.3\omega_{ref}$ (there is no fixed phase difference) then as long as we average over a sufficiently large number of cycles we always get zero output[26]. We need to reduce the ratio to $\omega_{input} = 1.01\omega_{ref}$ before we start to see any kind of output from the detector. It looks like a constant output, but if we plot the signal over a longer time period (bottom-right panel) we can see that it is sinusoidally varying. In fact the output varies as $\cos(t\Delta\omega)$ where $\Delta\omega$ is the *difference* between the two input signals, *as long as $\Delta\omega$ is a low enough frequency to pass through the filter.*

We can understand this behaviour by looking at the frequency content of the signals. Let's say the reference frequency is ω then the input frequency is $\omega + \Delta\omega$. We can write the two signals as

$$\begin{aligned} V_{in} &= A\sin(\omega + \Delta\omega)t \\ V_{ref} &= \frac{4}{\pi}\left[\sin\omega t + \frac{\sin 3\omega t}{3} + \frac{\sin 5\omega t}{5}\cdots\right] \end{aligned}$$

[26] Again here, these plots have been generated by integrating over 10 cycles.

Here we use a Fourier expansion to represent the square wave. We just need the first three terms of the Fourier series to understand the behaviour. Multiplying the two together we find the rectified signal (i.e. *before* the filter) is given by[27]

$$V_{rect} = \frac{4A}{\pi}[] \tag{18.2}$$

Now we know that $\omega \gg \omega_c$ so frequencies ω and above will not pass through the filter. For $\Delta\omega \ll \omega_c$ the first term in equation18.2 will pass through the filter un-attenuated[28]. We can define the output as follows[29]

$$V_{out} = \begin{cases} \frac{2A}{\pi}\cos(t\Delta\omega) & \Delta\omega \ll \omega_c \\ 0 & \Delta\omega \gg \omega_c \end{cases} \tag{18.3}$$

18.4 Lock-In Amplifier

One of the most important implications of equation 18.3 is that the difference between the two input frequencies has to be very small if we are to get any output from the phase detector. We can set the value of ω_c by choosing the values of R and C in the output filter, and therefore we can make the range of frequencies that the circuit responds to very narrow indeed (values as low as 1 Hz are possible). Since the reference frequency may be kHz or MHz, this means the circuit is very good at "picking-out" frequency components of the input signal with very high *frequency resolution.* Generally our input signal will consist of many frequencies (and noise) so the phase detector is good for asking questions such as "*What is the amplitude of the input signal at 1 kHz measured in a bandwidth of 1 Hz?*".

A typical commercially-available implementation of the phase detector is known as the

[27] Note that we need to make use of the trig identity $\sin A \sin B = \frac{1}{2}[\cos(A-B) - \cos(A+B)]$ to get this result. This tells us that the product of two sinusoids is (co-)sinusoids at the difference and sum frequencies.

[28] For low frequencies the filter is *not* an integrator!

[29] Note that we can use a similar argument to demonstrate equation 18.1 too

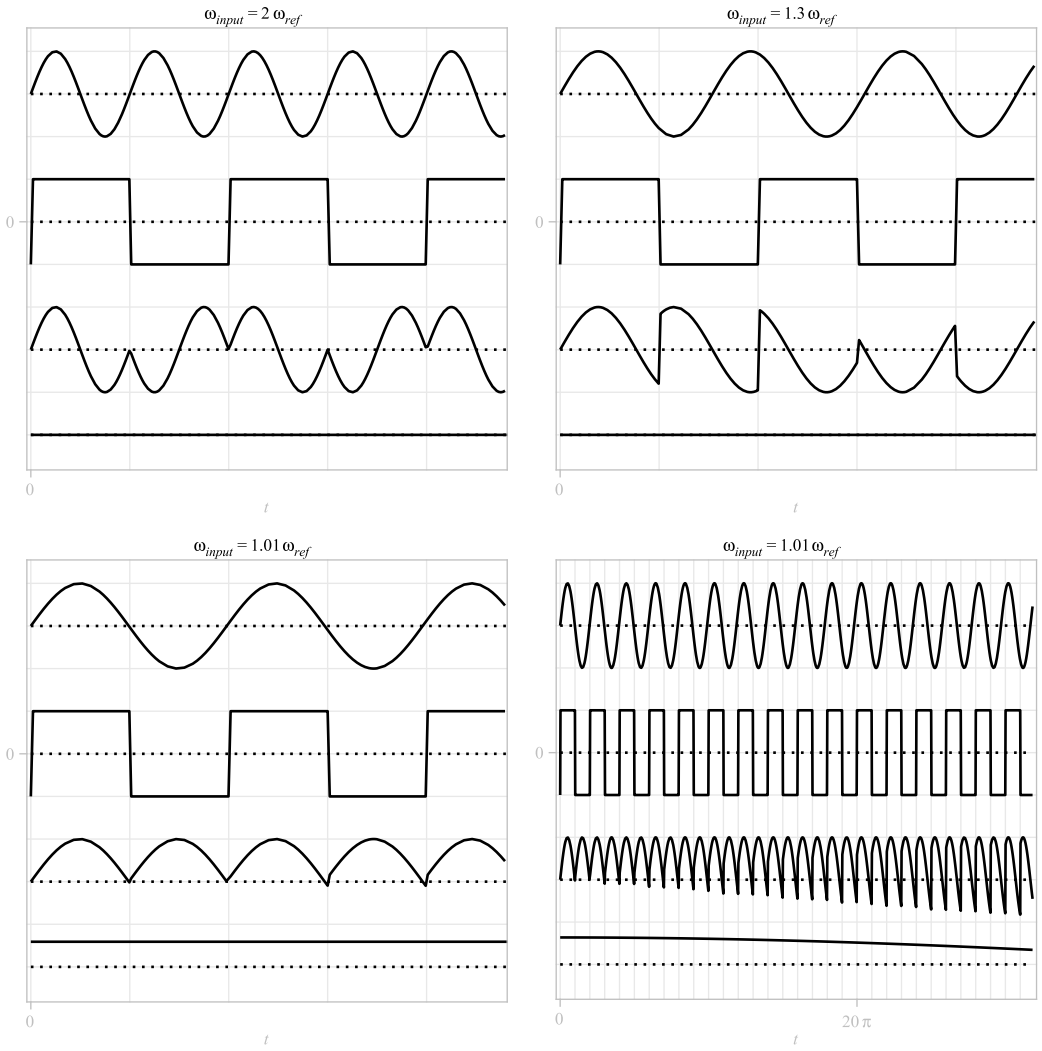


Figure 18.2: Phase Detector behaviour where the input signal has a different frequency to the reference frequency. Note that in all these figures, the waveforms from top to bottom are: input signal, reference signal, rectified signal and averaged (output) signal.

Lock-In Amplifier and in its simplest implementation this is a box with an input, two knobs and an output. The first knob is used to set the reference frequency (an internally generated square wave) and the second knob allows the phase between the two signals to be adjusted. The procedure to be used goes as follows: the input is connected to the signal we wish to study and the output V_{out} is usually sent to an oscilloscope. The reference frequency is set to 1 kHz and if there is any component of the input signal at this frequency then we will see an output on the 'scope. According to equation 18.3 we may find that the output varies a little so we fine-adjust the frequency knob to get a fix (or "lock") with a steady output. At this point $\Delta\omega = 0$ so now equation 18.1 comes into play. We now twiddle the phase knob to get a maximum output, and from this we can determine the value of the amplitude A at the

given frequency.

Practical Lock-In Amplifiers also have complicated extra features but the only one we need to be concerned with is the ability to take the reference frequency as an external input so we can generate this signal ourselves[30].

18.5 Signal Analyser

The lock-in amplifier gives us the capability to measure the amplitude of a signal at a specific frequency and with a very narrow bandwidth. This is the basic requirement of a **Signal Analyser**, a device which gives as output the amplitude of a signal as a function of frequency. We can use a signal generator to sweep the reference frequency over the required range and we can measure the amplitude at the output as a function of frequency. This will give us the amplitude spectrum of the signal. Note that in practice, signal analysers such as the one you have used with the ELVIS equipment, use a different technique, however it is important to understand the concept, and in fact some high-end *spectrum analysers* do use a variation on the method described above.

18.6 Extracting Signal From Noise

One of the most powerful uses of the phase-detector/lock-in amplifier technique is when we have a signal we wish to measure at a known frequency, but heavily contaminated with noise. For example we may have a sensor which we know produces a signal at 1 kHz but with a very small amplitude, say tens of μV. We may find that our signal has noise of some volts RMS. The signal to noise ratio is then some -100dB. The situation seems hopeless: were we to plot the data in the time-domain we would just see the noise; the signal at 1 kHz would be buried. However, recalling that noise is usually *broad-band* we understand that we can reduce the amount of noise in a measurement by reducing the bandwidth of the measurement. For example for thermally generated noise we found that

$$V_{noise,RMS} \propto \sqrt{B}$$

The lock-in amplifier can still measure the signal because

1. We expect a well defined signal frequency which is repetitive and can be synchronously detected (in comparison with the noise which is random and ultimately averages to zero)
2. The lock-in technique is highly selective in frequency, therefore it reduces the bandwidth of the measurement.

In practice a SNR of -100 dB or worse is manageable with this technique.

Example: Measurement Bandwidth of an Oscilloscope Imagine we have a measurement where the observed noise is $10^6\times$the signal. If the noise is principally thermal in origin then the noise per unit Hz is a constant so we can reduce the total noise measured by reducing the bandwidth of the measurement. Measuring with an oscilloscope (typical bandwidth 100 MHz) we would be unable to see the signal. However, if we first filter the signal with a band-pass filter of width 100 Hz then the SNR improves by a factor of a million. Note that the part of the signal we are interested in must of course be within the filter range.

In summary, we should think of this as a frequency-domain technique. If the signal of interest has a definite and fixed frequency (i.e. it is **Narrow-Band**), and if the noise is **Broad-Band**, (i.e. across the whole frequency range of the measurement) then we can see that in the frequency-domain the two sources look quite different, and can be separated by the lock-in technique.

[30] One practical consequence is that we can measure the ref frequency directly on a 'scope rather than having to rely on a value read from the dial on the front of the box

18.7 Low-Frequency or DC Signals

The approach outlined in section 18.6 above works fine for AC signals with a frequency that it is sufficiently high that it is above the regime where the $1/f$ noise dominates (recall the composite noise spectrum figure given in lectures). Under these circumstances, the noise - measured in the same band as the signal - is usually tolerable. However as the signal frequency reduces, and ultimately as we go towards a "DC" measurement then the $1/f$ noise dominates and eventually overwhelms the signal.

We may have a detector which gives a constant voltage of a few μV with some volts RMS of noise. Another example would be measuring the intensity of light from a distant source, when the detector output is swamped by ambient light. Or we could be trying to detect faint radio-emission from a star. These measurements are all small, constant values subject to noise. The solution here is to shift the frequency of the signal to be measured from near DC to a higher frequency (well above the regime where the $1/f$ dominates).

This can be done by a variety of methods which we will look at next, though they are all essentially **Modulation** techniques, or as it is more typically known for this application, **Chopping**

18.8 Modulation Techniques

We can move a DC or very low-frequency signal to a higher frequency by modulating or chopping the signal. The name derives from optical experiments where a 'chopper' wheel (a rotating blade) physically interrupts the light beam in front of the detector. This modulates the light beam from a constant to an on/off signal *at a well defined frequency* (given by the rotation rate of the chopper wheel). In Fourier terms, the signal changes from a delta function at zero frequency to a delta function at the modulation frequency. This has the added advantage that we can use the same frequency source for the modulation and the reference of the lock-in, so we are guaranteed to have the same frequency. This is best illustrated with an example[31].

18.8.1 Demonstration Experiment using Light Intensity

Imagine we have an LED powered by a battery, and on the other side of the room we have a photodetector[32]. The question is: can the detector tell whether the light is on or not? In a very dark room, with not too large a distance between the two, it is possible. With the LED off, the trace on the 'scope is a noisy straight line, and when we switch the LED on then the line goes up just a little. Note here that the detector is acting as an intensity meter. The LED light is a constant output "DC" signal. The detector responds to light of all visible wavelengths so when we switch the room lights on or open the curtains the detector is flooded with ambient light millions of times brighter than the LED. We have to drop down the scale of the 'scope by several orders of magnitude to get a new reading of the intensity. Now, if we switch the LED off/on we see no difference in the trace because it is simply buried by the ambient light.

Now imagine we discard the battery and instead power the LED with a sine wave from a signal generator. This modulates the light intensity from DC to the frequency of the signal generator

$$I = I_0 \sin \omega t$$

Further, we also connect the signal generator output to the reference input of a lock-in amplifier, and the output of the detector the signal input of the lock-in. Now we can synchronously detect the LED intensity at the modulation frequency. Because modulation and reference frequencies are the same, we just need to adjust

[31] In fact this is a classic demonstration experiment described in many texts, that is to say this is an actual real, working example

[32] This might consist of a photo-diode connected to an op-amp amplifier with the output sent to a 'scope

the phase knob to get a maximum output. We will see a nice strong constant signal out of the lock-in. If we disconnect the LED it goes to zero. This works the same both in the dark *and* in bright ambient light.

This is a striking example of extracting a tiny signal from noise. To be clear about what's happening here, we need to recall that

1. The modulated signal has a very narrow, well-defined frequency
2. The noise has lots of sources: mostly the more-or-less constant ambient light, plus some 50 Hz from mains lights, plus $1/f$ and thermal contributions from the detector system
3. Modulation moves the signal above the low-frequency noise, to where we have just the broad thermal noise etc
4. Consequently the noise, in the 1 Hz bandwidth of the lock-in, is actually quite small

18.8.2 Chopping

The demonstration experiment in the previous section is not very realistic, since it is usually[33] impractical to electrically modulate the signal. Usually the quantity to be measured needs to be mechanically or optically "chopped". The name derives from optical experiments where a rotating wheel or blade physically interrupts a light beam and turns it from a constant to a modulated quantity. Since we know the frequency of the "chopper" wheel we can use this as the reference for the lock-in. Typically the arrangement is similar to that described above except here the signal generator drives the chopper. Typically our light signal derives from some experiment, usually involving the measurement of some low-level laser light.

[33] Though not always; there are some cases where the electrical modulation method can be used though we won't go into this here

18.8.3 Radio Astronomy

This is a very pleasing example of the chopping technique taken from the world of radio astronomy. A radio-telescope is a dish antenna used to measure radio-frequency emission from distant stars[34]. The signal is very faint, while noise sources abound: we have the general cosmological background, terrestrial (man-made) radio-sources, plus all the usual electrical noise sources in the receiver. To make this worse, fluctuations in the atmosphere and ionosphere add extra low-frequency noise to the signal. Radio-astronomers use the chopping/synchronous-detection scheme as described previously. This can be done by *physically* rocking the dish side-to-side thus sweeping it across the target star. This directly modulates the signal intensity, while all the noise sources remain the same. With large dishes this is impractical, so often there is a secondary reflector at the focus of the dish (Cassegrain type), and it suffices to rock the secondary. A further refinement, which avoids any mechanical action (and can be run much faster) is a setup which uses a single dish with two signal-receiving waveguide/antennae. One is exactly on-axis and receives the target signal, the other is slightly off-axis so it receives all the noise sources but none of the signal from the star. The radiometer (power-meter for radio-frequency signals[35]) is electrically switched between the two sources at the modulation frequency. Radio-astronomers refer to this as "Dicke-Switching" after Robert Dicke who invented the technique during the early days of radar.

[34] Read more about radio astronomy and radiometers at www.nrao.edu/index.php/learn/radioastronomy/radiotelescopes and www.cv.nrao.edu/course/astr534/Radiometers.html

[35] Note that the most sensitive radiometers use another noise minimisation technique we have already discussed: the receiver/amplifier stages are frequently cooled to cryogenic temperatures in order to minimise thermal noise.