

Part I: Analog Bioinstrumentation

Chapter 1: Introduction

The objective of Biomedical Engineering is to apply engineering principles to physiology and medicine. Because the human body is an extraordinarily complicated system, it includes aspects of mechanical, electrical, and chemical engineering, which requires undergraduate Biomedical Engineering programs to produce generalists. These generalists must understand the different disciplines not as separate silos, but as interconnected to one another. One of the goals of this book is therefore to highlight connections between instrumentation and other engineering disciplines so that a unified approach can be taken to this mélange of engineering fields. A second, more direct goal, is to provide an understanding of the basic concepts of instrumentation that are needed to take measurements from the body, design biomedical instruments, and to model, theoretically, certain aspects of the body. A third goal is to highlight aspects of electrical engineering concepts, such as positive and negative feedback, that can be used to understand and quantitatively model physiological phenomena, such as oscillatory cycles.

Some background in circuit and signal analysis is assumed. The most basic concepts related to circuits, signals, and linear systems are reviewed first. Next, because the theoretical content of this book is intended to be accompanied by a laboratory in which students gain experience with the construction and testing of instrumentation, a brief description is provided of the basic testing instruments, the signal generator, the power supply, the multimeter, the oscilloscope, and the breadboard. These components are the subject of the first laboratory exercise.

The basics of operational amplifiers are described in Chapter 4 because these devices provide a platform for the rapid construction of more complex circuits. They are also convenient microchips for the second laboratory where concepts from the previous chapters are illustrated. The most basic operational amplifier configurations, inverting and non-inverting, are introduced here and generalized from circuits with purely resistive elements to circuits with combined resistive and capacitive elements that provide easy control of circuit frequency response to create, for example, low pass, high pass, and band pass filters.

Chapter 5 provides some guidelines for circuit design, construction and debugging. Of these stages, debugging is generally the most time consuming, and the amount of time required to diagnose and fix problems can be substantially reduced through rational design and systematic construction and debugging. These guidelines become more important as the circuits being constructed become more complicated. *Although they are presented here in terms of electronic instruments, they can easily be extrapolated to the design of other complicated systems.*

Chapter 6 covers hypothesis testing, which is important to the characterization of biomedical sensors and to the interpretation of data collected from these sensors. Some concepts, such as the definitions of mean, median, mode, variance, and linear regression, are likely to be familiar to the student, but they are included for completeness. The section on probability may be new to many students. The heart of the chapter is the discussion of the Student's t test in its various

forms, of the F test, and of binary decisions. The chapter concludes with a discussion of the function available to easily perform statistical analysis in Excel.

Chapter 7 addresses electrical safety, including the effects of electricity on the body, a description of electrical systems within the hospital, and some aspects of food and drug administration (FDA) regulation.

Chapter 8 describes the various types of sensors and how the signal from the entity being measured is converted to a useable form. The original signal is not necessarily electrical in nature, but may be thermal, chemical, optical, or mechanical. While the useable form is often electrical, it may take other forms. A classical example is the stethoscope, where the useable form is the acoustic signal that is focused onto the examiner's ear.

Chapter 9 covers operational amplifiers in more detail, describing features of these devices that deviate from the ideal operational amplifier model. Some of these limitations will already have been observed by students in the laboratory.

Chapter 10 describes some specific filters that are more complicated than those that were constructed from the inverting and non-inverting operational amplifier configurations.

Chapter 11 covers biopotentials and their origins. Students will probably already be familiar with the Nernst and Goldman-Hodgkin-Katz equations, so the intent of that section is to review and provide improved insight into the concepts. The chapter concludes with a discussion of neural function and action potentials.

Chapter 12 on nonlinear circuits is included not only because a wide variety of bioinstruments are based around nonlinearity, but also because these circuits can be used as model for much of the unique behavior of biological systems themselves. For example, the voltage-controlled oscillator can be considered to be a model of oscillatory behavior throughout the human body, such as in autonomous breathing, heart beating, and circadian rhythms.*

Chapter 13 describes a specific device for measurement of capillary blood flow and blood oxygenation. The chapter forms the background for the plethysmography laboratory experiment.

Chapter 14 describes the basics of device design. This chapter can be considered to be the student's introduction to a process that they are likely to learn more about in their major design course.

Chapter 15 introduces the student to some of the more complicated (and expensive) major instruments found in a hospital. The intent is for the student to understand the physics of the original signal source and the process required to convert that signal to a usable format. This chapter is the end of Part I on analog bioinstrumentation.

* The author confesses that this is his favorite chapter.

Chapter 2: Review

The student is expected to have a knowledge of basic circuit theory, including an understanding of Kirchoff's laws, current-voltage relationships in resistors, capacitors, and inductors, behavior of elements in parallel and series, linearity, input and output impedance, the difference between AC and DC signals, transfer functions, Bode plots, magnitude and phase, and Euler's rule. These concepts are briefly reviewed here.

2.1 Kirchoff's Laws

Kirchoff's laws are best described with an example circuit. Consider the circuit in Figure 2-1

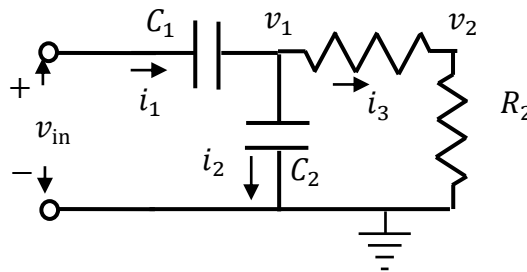


Figure 2-1: Simple circuit to illustrate Kirchoff's laws.

Kirchoff's current law states that charge cannot build up at a node, so the current that passes through C_1 and enters the node marked v_1 must be equal to the sum of currents leaving the node through R_1 and through C_2 .

$$i_1 = i_2 + i_3$$

$$C_1 \frac{d}{dt} (v_{in} - v_1) = C_2 \frac{dv_1}{dt} + \frac{v_1}{R_1 + R_2}.$$

By convention, the dependent variable (v_1) is placed on the left-hand side of the equation, and the independent variable (v_{in}) is placed on the right-hand side of the equation to obtain

$$(C_1 + C_2) \frac{dv_1}{dt} + \frac{v_1}{R_1 + R_2} = C_1 \frac{dv_{in}}{dt}.$$

This result is an ordinary differentia equation that relates the output variable v_1 to the input variable v_{in} .

Kirchoff's voltage law states that voltage can have only one value at a given location, so if voltage changes across elements in a loop, the sum of the changes must be zero. The example circuits has two loops, one that traverses v_{in} and the two capacitors, and one that traverses R_1 , R_2 , and C_2 . Thus,


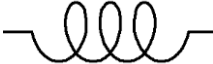

$$v_{in} - \frac{1}{C_1} \int i_1 dt - \frac{1}{C_2} \int i_2 dt = 0.$$

The sign of the v_{in} term is opposite from the signs of the other two terms because v_{in} is a voltage gain, whereas $\frac{1}{c_1} \int i_1 dt$ and $\frac{1}{c_2} \int i_2 dt$ are voltage losses.*

2.2 Resistance, Inductance, Capacitance

Resistors, inductors, and capacitors are elements that follow basic relationships between voltage and current, as show in Table 2-1. Here, v is voltage, i is current, R is resistance, L is inductance, and C is capacitance. Voltage is referred to as the “across” variable (measured at two points across the element) and current is referred to as the “through” variable (passing through the element).

Table 2-1: Resistors, inductors, and capacitors

Element	Equation	Symbol
Resistor	$v = iR$	
Inductor	$v = L \frac{di}{dt}$	
Capacitor	$i = C \frac{dv}{dt}$	

While we tend to think of these equations as describing the elements, it may be more accurate to consider the elements as approximations of the equations. The equations are used because they are the simplest ones that one can work with as an engineer. Therefore, the physical elements needed to be designed to match the equations.† It is thus no coincidence that other engineering elements follow the same equations. For example, in a mass spring damper system (relevant to muscle mechanics), with force (F) as the across variable and velocity (v) as the through variable, the spring is governed by $dF/dt = kv$, the damper is governed by $F = Cv$, and the mass is governed by $F = m dv/dt$. The mass is thus analogous to an inductance, which explains why inductance is sometimes referred to as inertance. The analogy extends to other behavior of the systems. For example, the damper is analogous to a resistor, and these are the two elements in each system that dissipate energy (generate heat).

Example: The windkessel model describes blood flow in an artery through equations like those that describe electrical circuit elements. Relevant variables are flow (Q) and pressure (P), and relevant parameters are resistance (R), caused by wall shear stress, compliance (c) describing the expansion of the arterial walls with pressure, and inertia (L) related to the mass of the fluid. Obtain equations resembling those in Table 2-1 for this analogy.

* The purpose here is to review Kirchoff’s laws. Nodal and mesh analysis (also based on Kirchoff’s laws) are generally recommended for complicated circuits because their application is straightforward, but these methods will not be reviewed here.

† While these commercially available elements follow the equations exceptionally well, none of them follow their governing equations perfectly. For example, any inductor will have some (small) resistance, which can affect a circuit’s behavior when frequencies are low so that the impedance caused by the inductance is negligible.

Answer: First consider the case where compliance and inertia can be ignored. Poiseuille's law then relates the flow and pressure in a cylindrical tube.

$$\Delta P = \frac{8\mu L}{\pi R_0^4} Q,$$

Where μ is dynamic viscosity, L is the length of the tube, and R_0 is the tube radius. This is equivalent to the equation for a resistor, with resistance replaced by

$$R = \frac{8\mu L}{\pi R_0^4}$$

Next consider only the effect of a change in flow caused by wall motion. Make the (somewhat justified) assumption that the change in volume of the artery is proportional to the change in pressure inside the artery.

$$dV = c dP$$

The change in volume over a differential time dt of a tube of length L is

$$dV = Q_{\text{in}} dt$$

From this and the previous equation

$$c dP = Q_{\text{in}} dt \Rightarrow Q_{\text{in}} = c \frac{dP}{dt}$$

The equation is analogous to capacitance, with Q_{in} in place of current, P in place of voltage, and c in place of capacitance. Finally, consider the effect of mass (inertia). Since force is mass times acceleration, mass is density times volume, and acceleration is rate of change of velocity (which is rate of change of Q times cross-sectional area),

$$P(\pi R_0^2) = \rho \frac{d}{dt} (\pi R_0^2 Q) \Rightarrow P = \rho \frac{dQ}{dt}$$

The equation is analogous to inductance, with Q in place of current, P in place of voltage, and ρ in place of inductance.*

None of these three elements is exact for blood flow, just as they are not exact for real circuit elements. However, they are reasonable approximations that can lead to useful deductions about the behavior of blood flow.

* This analogy explains why the term “inertance” is sometimes used in place of “inductance” in circuit analysis.

2.3 Elements in Series and Parallel

The overall resistance (R_{series}) caused by two resistors *in series* is the sum of the resistances, The overall inductance L_{series} caused by two inductors in series is the sum of the inductances, The overall capacitance C_{parallel} caused by two capacitors *in parallel*, is the sum of the capacitances. For resistors R_1 and R_2 in parallel.

$$\frac{1}{R_{\text{parallel}}} = \frac{1}{R_1} + \frac{1}{R_2}.$$

With some simple algebra,

$$R_{\text{parallel}} = \frac{R_1 R_2}{R_1 + R_2}.$$

The formula is easy to remember. First, note that it is the product divided by the sum. Second, if you have difficulty remembering whether the product or the sum belongs in the numerator, recognize that the dimensions of the formula must be a resistance, which means that the product belongs in the numerator to provide units of ohms. Otherwise, it would lead to the dimensions ohms^{-1} .

The same form applies for inductances in parallel and for capacitances in series.

$$L_{\text{parallel}} = \frac{L_1 L_2}{L_1 + L_2}.$$

$$C_{\text{series}} = \frac{C_1 C_2}{C_1 + C_2}$$

2.4 Time Constant

The output of a first order linear system to a step function is an exponential of the form

$$v_{\text{out}} = V_{\text{max}}(1 - e^{-t/\tau}) u_s(t),$$

where τ is the time constant, V_{max} is the maximum value that v_{out} will attain, and $u_s(t)$ is the unit step function. When $t = \tau$, the value of this output is

$$v_{\text{out}} = V_{\text{max}}(1 - e^{-1}) = V_{\text{max}}\left(1 - \frac{1}{e}\right) \approx 0.632 V_{\text{max}}.$$

To estimate the time constant for such a system, an step function is used as the input, and the amount of time required for the output to reach within 0.632 of its final value is measured. The time constant is sometimes also referred to as the “one over e time.”

For an RC low pass filter, τ is equal to the product RC , and the -3 db cutoff frequency of the filter is $\frac{1}{2\pi RC}$ Hz ($\frac{1}{RC}$ radians/s). The transfer function can then be written alternatively as

$$H(j\omega) = \frac{\omega_{-3db}}{j\omega + \omega_{-3db}} \quad \text{or} \quad \frac{1}{\frac{j\omega}{\omega_{-3db}} + 1} \quad \text{or} \quad \frac{1}{j\omega\tau + 1}$$

2.5 Impedance

Impedance (typically designated as Z) is derived from the Fourier transform of each equation in Table 2-1. Resistance does not depend on frequency, whereas the derivatives cause inductance and capacitance to depend on frequency,* so if $I(\omega)$ is the transform of $i(t)$ and $V(\omega)$ is the transform of $v(t)$, then

$$\begin{aligned} v &= iR \Rightarrow V(\omega) = R I(\omega) \\ v &= L \frac{di}{dt} \Rightarrow V(\omega) = j\omega L I(\omega) \\ i &= C \frac{dv}{dt} \Rightarrow I(\omega) = j\omega C V(\omega) \Rightarrow V(\omega) = \frac{I(\omega)}{j\omega C} \end{aligned}$$

It follows that the impedance for the resistor (transform of voltage divided by transform of current) is $Z_R = R$, the impedance of the inductor is $Z_L = j\omega L$, and the impedance of the capacitor is $Z_C = 1/j\omega C$.

Impedances combine in the same way that resistances combine. For two impedances in series

$$Z_{\text{series}} = Z_1 + Z_2$$

and for two impedances in parallel

$$Z_{\text{parallel}} = \frac{Z_1 Z_2}{Z_1 + Z_2}.$$

2.6 Diodes

An ideal diode is a device that conducts current (acts as a short circuit) if the voltage drop across it ($v_1 - v_2$) is positive, and as an open circuit if the voltage across it is negative. The symbol for this device is shown in Figure 2-2.

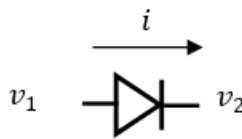


Figure 2-2: Diode symbol. Current i is zero if $v_2 > v_1$.

For real diodes, the relationship between the current and the voltage across the device is

$$i = I_s (e^{e_0(v_1 - v_2) \eta k_B T} - 1)$$

where I_s is the reverse saturation current, e_0 is the charge of an electron (in Coulombs), η is a nonlinearity factor, k_B is Boltzman's constant, and T is absolute temperature. This characteristic is plotted in Figure 2-3, along with the characteristic that would correspond to an ideal diode.

* Recall the Fourier transform theorem that $\mathcal{F}\{df(t)/dt\} = j\omega \mathcal{F}\{f(t)\}$.

For negative voltages, the real diode exhibits a small backward current, I_s . This current is miniscule and hence difficult to measure directly. For positive voltages, the diode can support a large current, which is typically determined by the voltage-current relationships in the rest of the circuit, so that the voltage drop $v_1 - v_2$ is small. However, the diode will not allow large currents until the voltage drop exceeds about 0.6 volts, so each diode will usually operate with a voltage drop of that size. This voltage difference is often referred to as “one diode drop.” The physiological analogy to a diode is a heart valve, where blood flow relates to current and pressure relates to voltage. Like a diode, the heart valve will not open (conduct) until the pressure (voltage) is sufficiently large.

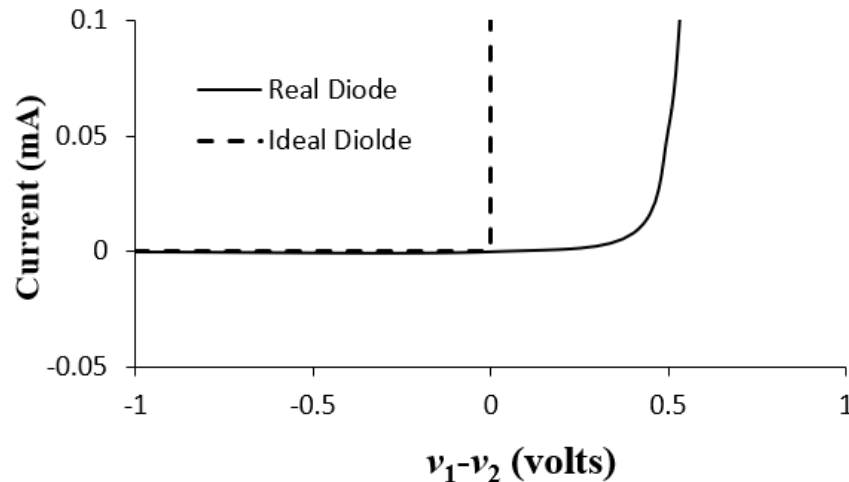


Figure 2-3: Current as a function of voltage drop for an ideal and real diode.

An example of the diode drop is shown in Figure 2-4. Assume initially that V_{in} is positive and the diode is ideal. For a 2 volt input signal, the current is $V_{in}/R = (2 \text{ volts})/(10 \text{ k}\Omega) = 0.1 \text{ mA}$. To allow that much current, the diode must have a drop across it of about 0.6 volts. That voltage drop decreases the current to $(1.4 \text{ volts})/(10 \text{ k}\Omega)$, but the smaller current is still sufficient to maintain the voltage across the diode near 0.6 volts. As a result, once V_{in} exceeds about 0.6 volts, the voltage across the diode will be approximately 0.6 volts.

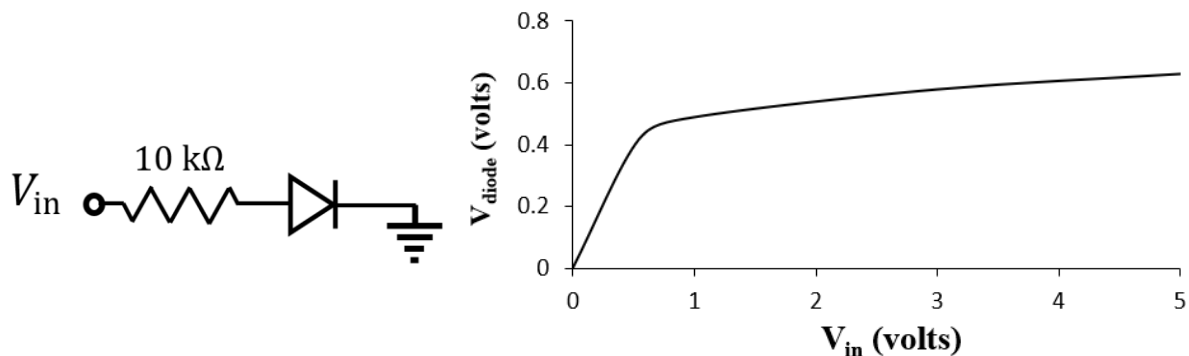


Figure 2-4: Simple example of the diode drop. Left: Circuit used to measure the current. Right: Voltage across the diode as the input voltage changes from 0 to 5 volts.

2.7 Linearity

2.7.1. Definition of Linearity

Linearity has a formal mathematical definition and a simple physical interpretation. A system that is linear has important practical consequences.

Mathematically, a linear function, $f(x)$, is one such that

$$f(a_1x_1 + a_2x_2) = a_1f(x_1) + a_2f(x_2) .$$

An example is the derivative. If $f(x) = \frac{dx}{dt}$
then

$$\frac{d}{dt}(a_1x_1 + a_2x_2) = a_1\frac{dx_1}{dt} + a_2\frac{dx_2}{dt} .$$

In contrast, the process of squaring a variable is not linear. If

$$f(x) = x^2$$

then

$$(a_1x_1 + a_2x_2)^2 = a_1^2x_1^2 + 2a_1x_1a_2x_2 + a_2^2x_2^2 \neq a_1^2x_1^2 + a_2^2x_2^2$$

Where the expression on the left is $f(a_1x_1 + a_2x_2)$ and the expression to the right of \neq is $f(a_1x_1) + f(a_2x_2)$. Notably, the general equation for a straight line, $f(x) = mx + b$, is not linear unless $b = 0$.

$$f(a_1x_1 + a_2x_2) = m(a_1x_1 + a_2x_2) + b$$

$$f(a_1x_1) + f(a_2x_2) = (m(a_1x_1) + b) + (m(a_2x_2) + b) = m(a_1x_1 + a_2x_2) + 2b .$$

Since $b \neq 2b$ unless $b = 0$, the mapping of a variable to a straight line is not linear unless the line passes through the origin.

2.7.2. Physical Interpretation of Linearity

Physically, if a system is linear, the sum of two inputs can be fed into the system and the output is the same as if the two signals were fed into two identical systems and then summed, as illustrated in Figure 2-5.

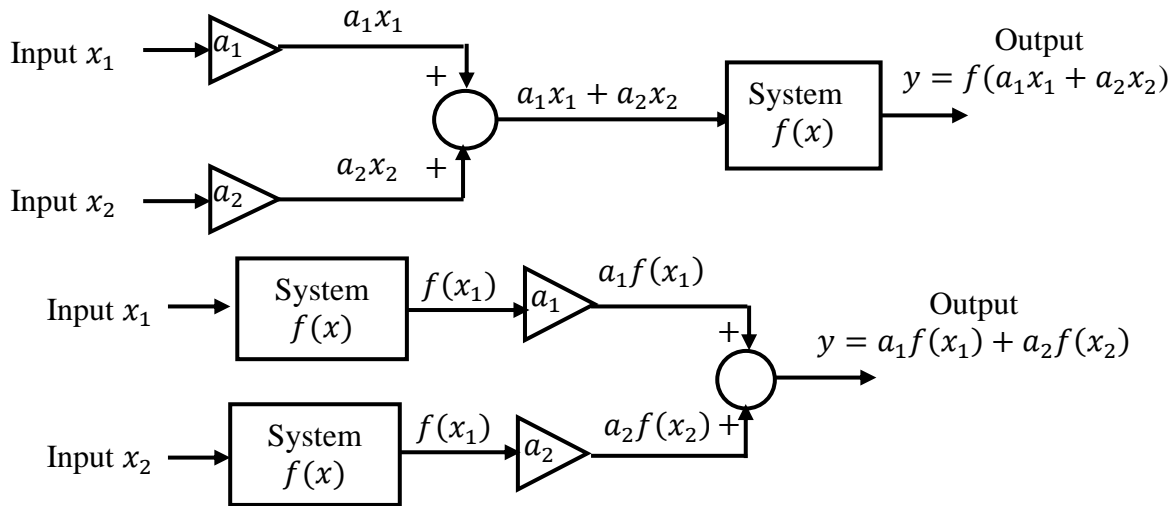


Figure 2-5: Physical interpretation of a linear system. The order through which x_1 and x_2 are fed through the system differ, but the resulting y is the same because $f(a_1x_1 + a_2x_2) = a_1f(x_1) + a_2f(x_2)$.

2.7.3. Superposition

If every component of a system is linear, then the response to two inputs is equal to the sum of responses to the inputs individually, even if the two inputs are applied to different locations in the system. With two inputs, for example, the response to the two inputs together can be obtained by setting the first to zero and solving for the response to the second, then setting the second input to zero and solving for the response to the first, then adding the two responses together. Also, for a single input that has DC and AC components, the response to each component can be obtained separately, and those responses can be added to get the complete response. If the system is nonlinear, however, where the DC signal is much larger than the AC signal, one approach is to solve for the response of the DC signal and then use that response to derive a linearized version of the response to the AC signal. This approach is frequently used in the analysis of transistors, which are inherently nonlinear.

2.8 Input and Output Impedance

Any real voltage source has a finite output impedance, which is modeled by the configuration shown in Figure 2-6.

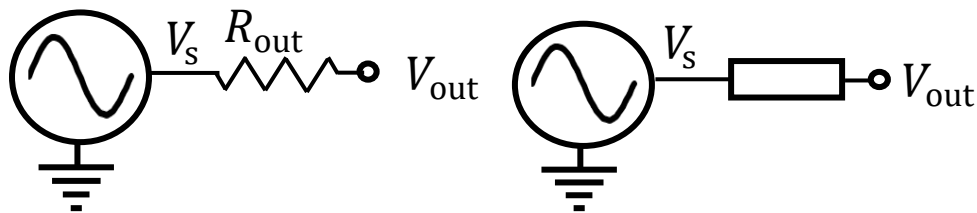


Figure 2-6: Model of a voltage source with an output impedance. Left: Resistive impedance. Right: Complex impedance.

In each case, the source generates a voltage V_s , but some of that voltage is lost through the impedance, and the amount lost depends on what is connected to V_{out} . Assume that $R_{out} = 50 \Omega$ and that V_{out} drives a load resistance R_L . Then R_{out} and R_L form a voltage divider. The current through these two resistors is $i = V_s / (R_{out} + R_L)$, and the voltage across R_L is $V_{out} = iR_L$, so that $V_{out} = V_s R_L / (R_{out} + R_L)$. If $R_L = 50 \Omega$, for example, then half of V_s is lost across R_{out} . If, however, $R_L \gg R_{out}$, a negligible amount of voltage is lost.

2.9 AC vs. DC Signals

2.9.1. Definition of AC and DC

Signals are broadly divided into those that vary in time (alternating current, or AC) and those that do not vary in time (direct current, or DC). The distinction is made because signals are generally easier to analyze if they do not vary in time. In a DC circuit, all capacitors act as open circuits, and all inductors work as short circuit, so the effective topology of the circuit can be simplified.

A simple example is

$$s(t) = \underbrace{8}_{\text{DC Component}} + \underbrace{2 \sin(2\pi(50)t)}_{\text{AC Component}}. \quad \text{Equation 2-1}$$

The AC behavior of a circuit is typically tested with inputs that are either square waves or sine waves. The square waves are used to examine the time response of a circuit, which is the amount of time required for the output to reach a constant value after a step change is made to the input of a circuit. Thus, the period, T , of the square wave is made large enough that the output approaches a constant value in $\frac{1}{2}T$ seconds. Sine waves are used to determine a circuit's frequency response. For a linear circuit, an input $A \sin(\omega t)$ will lead to an output $B \sin(\omega t - \varphi)$, where the ratio B/A is the magnitude of the circuit's transfer function, and φ is the phase of that transfer function.

The AC and DC components of the signal depicted by Equation 2-1 is shown in Figure 2-7. The AC component spans the peak-to-peak voltage. The amplitude is half of the peak-to-peak voltage. The offset is the difference between the time mean of the signal and zero.

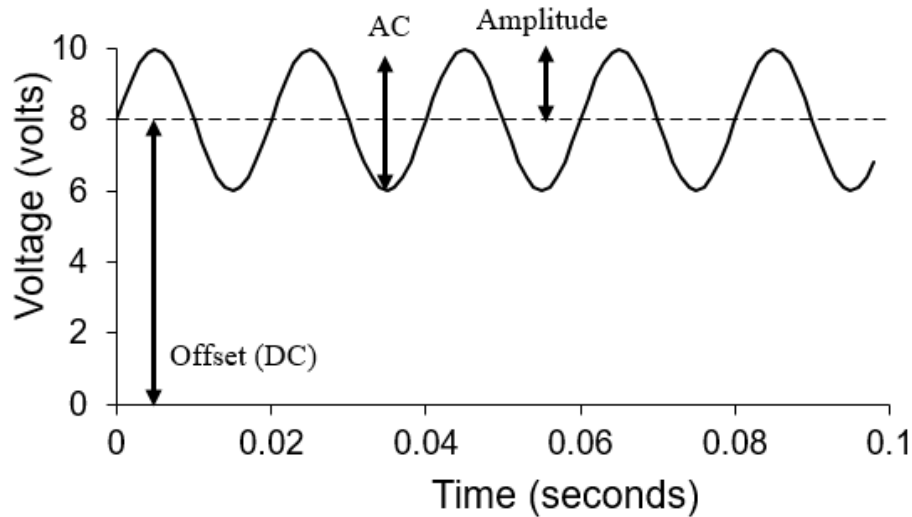


Figure 2-7: Plot of the signal from Equation 2-1, illustrating the DC and AC components of the signal.

By strict definition, the DC component does not change with time. In engineering practice, DC can change, but over a time that is too long to be relevant. Thus, whether a signal is AC or DC depends on the time frame of interest. For example, to an orthopedic surgeon implanting an artificial knee, the elastic modulus of the femur is a constant value (DC signal), but to a researcher studying the etiology of osteoporosis elastic modulus changes over time (AC signal).

2.9.2. DC as $\lim_{\omega \rightarrow 0} \{\cos(\omega t)\}$

In one sense, the distinction between an AC and a DC signal is arbitrary in that no signal is truly DC. Every signal changes over some period of time. The voltage from a power supply is considered to be DC, but it varies in time, at least in the sense that the power supply output was zero and changed to the output value when it was turned on and adjusted and was turned off after it was needed. Thus, a DC signal is one that changes only over a time frame that is longer than we care to worry about. This long time frame signal is conceptually similar to one with a low frequency. Consider that if a signal is given by $s(t) = A_0 \cos(\omega t)$ and if we let $\omega \rightarrow 0$, the signal goes to $s(t) = A_0 \cos(0 t) = A_0$. Thus, DC is the result of taking the limit of an AC signal as frequency goes to zero, as illustrated in **Figure 2-8**. The cosine with frequency 1 Hz has the appearance of a sinusoid. The cosine with frequency of 0.1 Hz varies less, and the cosine with a frequency of 0.01 appears to be unchanging within the time range of the plot.

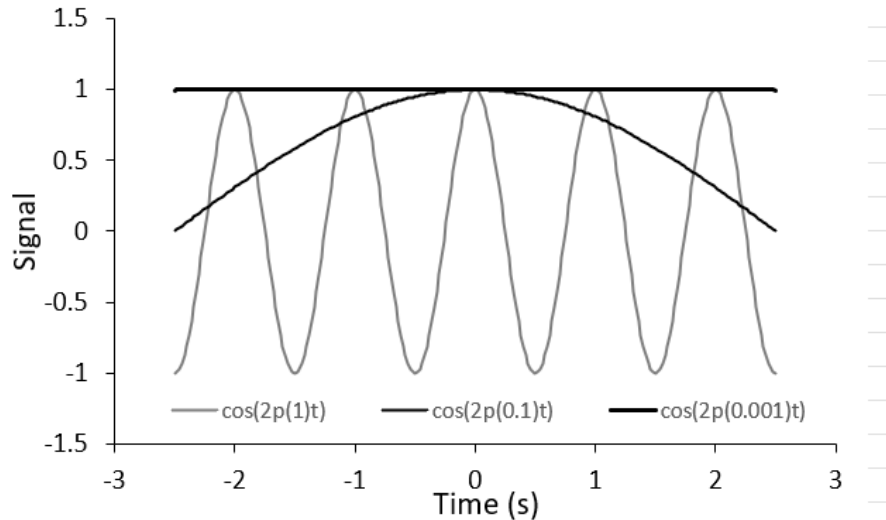


Figure 2-8: Illustration that as the frequency of a cosine wave decreases, it becomes a DC signal over the time range of interest.

This concept becomes important in the context of Fourier spectra, where $\omega = 0$ corresponds to the DC offset of a signal.

2.9.3. Rectification

The simplest method to determine the DC value of a signal is to take its time average. However, if the amplitude of an AC signal is needed, the time average leads to the signal's offset, regardless of the signal's amplitude. Therefore, a technique is needed to obtain a DC signal that is proportional to the amplitude of the AC signal. This technique is rectification. The DC offset of the signal is first subtracted, and the absolute value of the resulting AC signal is obtained and time averaged. The time average is a value that is linearly proportional to the original signal amplitude. Rectification will be described in more detail in Section 8.2.2.

2.9.4. Biasing

Nonlinear elements tend to have desirable behavior when the DC voltage is within a certain range. Consider a nonlinear resistor with the current-voltage characteristic shown in Figure 2-9. At low values of voltage, the current is zero over a wide range of voltages, so this resistor cannot be used to produce a change in current with voltage. Current also changes little with voltage for large voltage values. However, for medium sized voltages, current changes monotonically with voltage and if the change in voltage is small, the element is nearly linear. Therefore, it is useful to add a constant voltage to the input signal to bring the input signal value into a range where the element functions as a resistor. This technique is referred to as biasing, and the constant voltage is the bias voltage (here around 4 volts). The bias then leads to a constant current requirement for the element, and this constant current will be of importance in the study of the inputs to operational amplifiers.

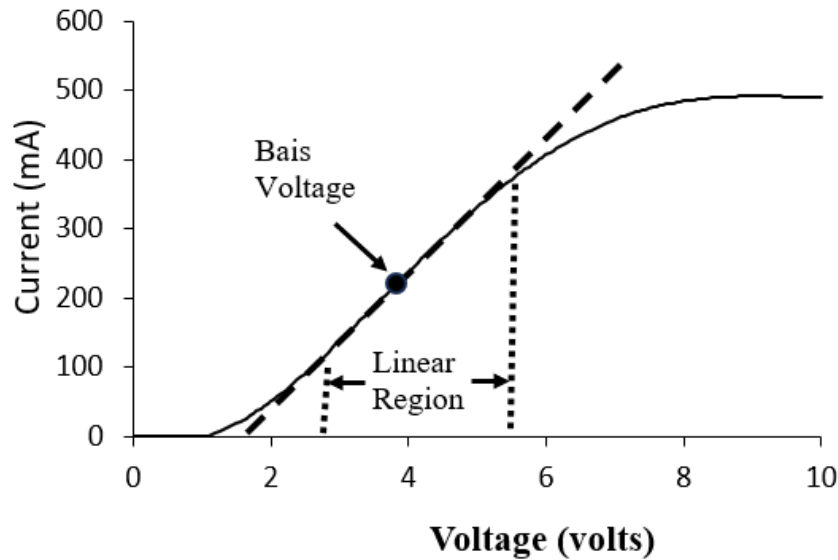


Figure 2-9: Current-voltage characteristic of a nonlinear resistor.

2.9.5. The difference between ω and f

Frequency can be quantified as either f , which is cycles per second, or ω , which is radians/s, where $\omega = 2\pi f$. Use of f is more intuitive, as it describes how many times something goes through a complete cycle in one second. Use of ω leads to simpler mathematical notation ($\sin(\omega t)$ is easier to write than $\sin(2\pi f t)$). Attention needs to be paid to which variable is used in a given situation. It is helpful to recall that ω will be larger than f by about a factor of six. In terms of ω , the -3 db frequency for an RC circuit is $\omega_{-3\text{db}} = \frac{1}{RC}$, whereas in terms of f it is $f_{-3\text{db}} = \frac{1}{2\pi RC}$ (smaller by a factor of 2π). Since $\tau = RC$ is the time constant for this circuit, if $\tau = 0.1$, the -3 db frequency will be either $\omega_{-3\text{db}} = \frac{1}{0.1} = 10$ radians/s or $f_{-3\text{db}} = \frac{1}{2\pi(0.1)} \approx 1.59$ Hz.

2.1 Euler's Rule

Euler's rule is a basic formula that relates sinewaves to complex exponentials. It is particularly useful in the analysis of circuits (and in Fourier analysis in general) because the use of transfer function often requires integrals and derivatives of sines and cosines. These operations are much more complicated when the sines and cosines are used directly, but they are simple when they are written as complex exponentials (because $d(e^{j\omega t})/dt$ is simple $j\omega e^{j\omega t}$). Euler's rule is*

$$e^{j\omega t} = \cos(\omega t) + j \sin(\omega t).$$

This expression also leads to

* A simple proof of Euler's rule is to form the function $f(x) = \frac{e^{jx}}{\cos(x) + j \sin(x)}$ and note first that $f(x) = 1$ if $x = 0$.

Then use the quotient rule to show that the derivative of $f(x)$ is zero. Therefore $f(x)$ is always 1, meaning that the numerator and denominator must always be equal.

$$\cos(\omega t) = \frac{(e^{j\omega t} + e^{-j\omega t})}{2}$$

and

$$\sin(\omega t) = \frac{(e^{j\omega t} - e^{-j\omega t})}{2j}.$$

As an example of the use of Euler's rule, consider the Laplace transform of a sine wave.

$$\mathcal{L}\{\sin(\omega_0 t)\} = \int_0^{\infty} e^{-st} \sin(\omega_0 t) dt$$

Evaluation of the integral involves a two-step process of integration by parts. In contrast, consider the complex exponential.

$$\mathcal{L}\{e^{j\omega_0 t}\} = \int_0^{\infty} e^{-st} e^{j\omega_0 t} dt = \int_0^{\infty} e^{(j\omega_0 - s)t} dt$$

The integral is easily evaluated as

$$\int_0^{\infty} e^{(j\omega_0 - s)t} dt = \left. \frac{e^{(j\omega_0 - s)t}}{j\omega_0 - s} \right|_0^{\infty}.$$

For positive values of s , the expression evaluated at ∞ will be zero. The expression at 0 gives

$$\mathcal{L}\{e^{j\omega_0 t}\} = \frac{1}{j\omega_0 - s}$$

Multiply the top and bottom by the complex conjugate of the denominator to express this transform as a real and imaginary part.

$$\frac{1}{j\omega_0 - s} \left(\frac{j\omega_0 + s}{j\omega_0 + s} \right) = \frac{j\omega_0 + s}{s^2 + \omega_0^2}.$$

But since $e^{j\omega_0 t} = \cos(\omega_0 t) + j \sin(\omega_0 t)$

$$\mathcal{L}\{e^{j\omega_0 t}\} = \mathcal{L}\{\cos(\omega_0 t) + j \sin(\omega_0 t)\} = \mathcal{L}\{\cos(\omega_0 t)\} + j \mathcal{L}\{\sin(\omega_0 t)\},$$

where the last expression is a consequence of the linearity of the Laplace transform. It follows that $\mathcal{L}\{\cos(\omega_0 t)\}$ is the real part of $\mathcal{L}\{e^{j\omega_0 t}\}$ and $\mathcal{L}\{\sin(\omega_0 t)\}$ is the imaginary part of $\mathcal{L}\{e^{j\omega_0 t}\}$.

$$\mathcal{L}\{\cos(\omega_0 t)\} = \Re \left\{ \frac{j\omega_0 + s}{s^2 + \omega_0^2} \right\} = \left\{ \frac{s}{s^2 + \omega_0^2} \right\}$$

$$\mathcal{L}\{\sin(\omega_0 t)\} = \Im \left\{ \frac{j\omega_0 + s}{s^2 + \omega_0^2} \right\} = \left\{ \frac{\omega_0}{s^2 + \omega_0^2} \right\}$$

Euler's rule has therefore greatly simplified the derivation of these two transforms.

2.2 Transfer Functions

Linear systems are described in the time domain by their impulse response, $h(\tau)$, which is the output of the system when a delta function is used as the input.* Then, the output to an arbitrary input is the convolution of that input with the impulse response.

$$y(t) = s(t) * h(t) = \int_{-\infty}^{\infty} s(t)h(t - \tau) d\tau \quad \text{Equation 2-2}$$

Given that the convolution is mathematically complicated, an easier technique is to transform both the input signal (sending $s(t)$ to $S(j\omega)$ and $h(\tau)$ to $H(j\omega)$, with the capital letters indicating the Fourier transformed variable). Then the Fourier transform of the output is $S(j\omega)H(j\omega)$, a simple multiplication (see Figure 2-10).

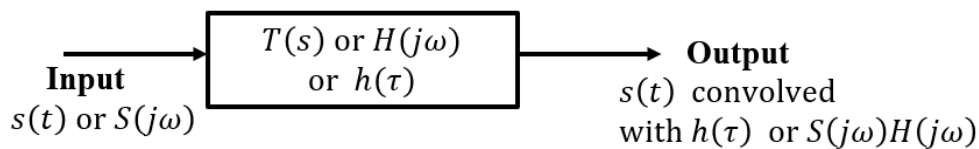


Figure 2-10: Linear system input and output in the time and frequency domains.

The simplest interpretation of the transfer function is that any signal $A \sin(\omega t)$, when used as an input to a system with a transfer function that has magnitude $|H(j\omega)|$ and phase φ at frequency ω will lead to an output of $|H(j\omega)|A \sin(\omega t + \varphi)$. The input and output signals are illustrated in Figure 2-11. Here, $|H(j\omega)|=0.7$ and $\varphi = \angle H(j\omega) = -30^\circ$. If the figure is interpreted as an oscilloscope trace of input and output, then the magnitude of the transfer function at this frequency can be calculated as

$$|H(j\omega)| = \frac{B}{A}$$

and the phase can be calculated from the marked time lag as

$$\varphi = -360 \frac{\ell}{T}.$$

where $T = 1/f$ is the period of the sine wave (f is the frequency $\omega/(2\pi)$). The negative sign is needed because the output lags the input.

* Recall that the delta function, $\delta(t)$, is zero everywhere except at $t = 0$, where it is ∞ , and it is defined such that $\int_{-\infty}^{\infty} \delta(t) dt = 1$. While the idea of inputting an infinite signal over an infinitesimal amount of time is counterintuitive, the physical interpretation is that the input is large over a small amount of time. The specific value of the input is less important than the value of its integral. For a delta function, the integral must be one. A good analogy is the striking of a bell with a hammer. The duration of the hammer blow is short, and the force is large, but what matters is the amount of energy that the hammer imparts.

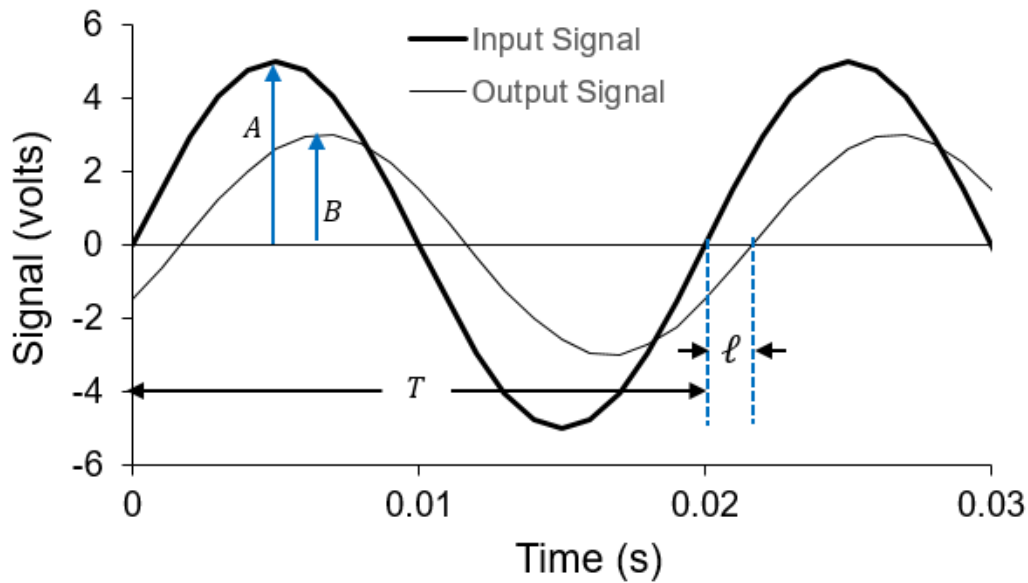


Figure 2-11: Interpretation of magnitude and phase.

2.2.1. Calculation of Magnitude and Phase from the Complex Transfer Function

Transfer functions for linear systems are complex functions of frequency, where the magnitude and phase have the interpretation of the previous section. For a complex number, $a + jb$ the magnitude is

$$|a + jb| = \sqrt{a^2 + b^2}$$

and the phase is

$$\angle(a + jb) = \text{atan}\left(\frac{b}{a}\right).$$

Transfer functions are not generally given in this simple form, where the real part, a , and imaginary part, b , are written explicitly. Instead, they are generally products and quotients of terms with the form $j\omega + \alpha$, where α may be real or complex. However, the calculation of magnitude and phase can still be obtained easily. For an expression constructed as products and quotients of complex numbers we can multiply the magnitudes of the individual terms to get magnitude and add/subtract phases to get the overall phase. Consider the transfer function

$$H(j\omega) = \frac{(j\omega b_1 + a_1)(j\omega b_2 + a_2)}{(j\omega b_3 + a_3)(j\omega b_4 + a_4)(j\omega b_5 + a_5)}$$

$$|H(j\omega)| = \frac{|j\omega b_1 + a_1| |j\omega b_2 + a_2|}{|j\omega b_3 + a_3| |j\omega b_4 + a_4| |j\omega b_5 + a_5|}$$

$$\angle H(j\omega) = (\angle j\omega b_1 + a_1) + \angle(j\omega b_2 + a_2) - \angle(j\omega b_3 + a_3) - \angle(j\omega b_4 + a_4) - \angle(j\omega b_5 + a_5)$$

If the a and b values are real, then

$$|H(j\omega)| = \frac{\sqrt{\omega^2 b_1^2 + a_1^2} \sqrt{\omega^2 b_2^2 + a_2^2}}{\sqrt{\omega^2 b_3^2 + a_3^2} \sqrt{\omega^2 b_3^2 + a_3^2} \sqrt{\omega^2 b_3^2 + a_3^2}}$$

$$\angle H(j\omega) = \text{atan}\left(\frac{\omega b_1}{a_1}\right) + \text{atan}\left(\frac{\omega b_2}{a_2}\right) - \text{atan}\left(\frac{\omega b_3}{a_3}\right) - \text{atan}\left(\frac{\omega b_4}{a_4}\right) - \text{atan}\left(\frac{\omega b_5}{a_5}\right)$$

If we then want to reconstruct $H(j\omega)$ from its magnitude and phase, we can use

$$H(j\omega) = |H(j\omega)|e^{j\angle H(j\omega)}.$$

Consequently, it is never necessary to decompose the transfer function into its real and imaginary parts.

Example:

For a transfer function of the form

$$H(j\omega) = \frac{10^7(j\omega + 40)}{(j\omega + 1)(j\omega + 200)(j\omega + 20,000)}$$

find expressions for the magnitude and phase, and use those expressions to generate a Bode plot of the transfer function over a range of 0.1 rad/s to 1,000,000 rad/s.

Solution:

Here, all of the b values are 1.

$$|H(j\omega)| = \frac{10^7 \sqrt{\omega^2 + 40^2}}{\sqrt{\omega^2 + 1^2} \sqrt{\omega^2 + 200^2} \sqrt{\omega^2 + 20,000^2}}$$

$$\angle H(j\omega) = \text{atan}\left(\frac{\omega}{40}\right) + \text{atan}\left(\frac{\omega}{1}\right) - \text{atan}\left(\frac{\omega}{200}\right) - \text{atan}\left(\frac{\omega}{20,000}\right)$$

The Bode plot (Figure 2-12) was generated in Excel.

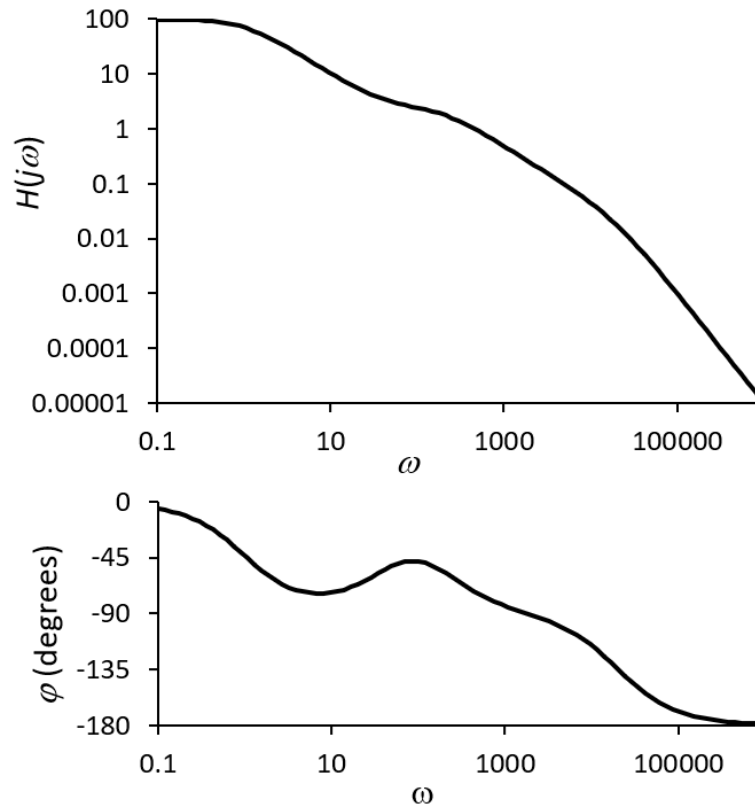


Figure 2-12: Bode plot of the transfer function

Calculation of the magnitude and phase are only slightly more complicated when the numerator or denominator of the transfer function has complex poles. Then, instead of factors of the form $j\omega + a$, a factor of the form $(j\omega)^2 + \alpha_1 j\omega + \alpha_2$ occurs. The quadratic term becomes $-\omega^2$, so the real part of this factor is $(\alpha_2 - \omega^2)$ and the imaginary part is $\alpha_1 \omega$. If the complex pole is in the denominator, then the system is referred to as underdamped; as ω^2 approaches α , the consequent smaller denominator causes a peak in the transfer function magnitude. If the complex pole is in the numerator, it causes a local trough in the transfer function magnitude which can be used to create a notch filter. An example is

$$H(j\omega) = \frac{(j\omega)^2 + j\omega + 20}{(j\omega)^2 + 20,000j\omega + 50,000}.$$

The magnitude plot is shown in Figure 2-13. The notch occurs at $\omega = \sqrt{20} \approx 4.47$, and the peak occurs at approximately $\sqrt{50,000} \approx 224$. For the case where $\alpha_1 = 0$, the transfer function at the notch frequency becomes zero.

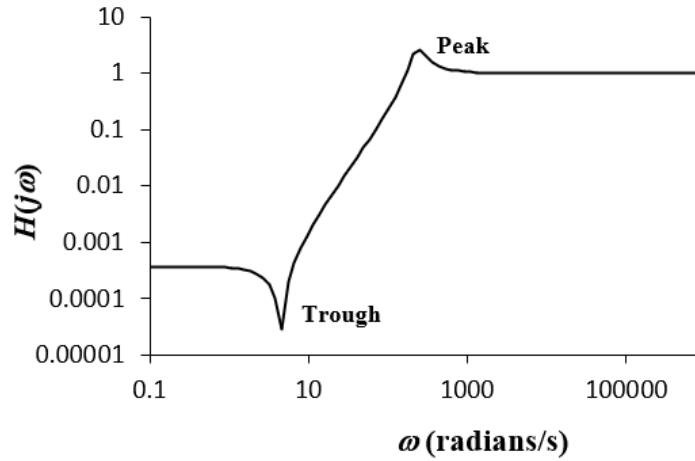


Figure 2-13: Bode plot with second-order factors in the numerator and denominator.

2.2.2. Poles and Zeros

Zeros of a transfer function are the values of $s = j\omega$ that cause the transfer function to become zero, where we consider s in the Laplace domain and $j\omega$ in the Fourier domain. Poles are the values of $s = j\omega$ that cause the denominator of the transfer function to become zero. Because physical values of ω are real, making $j\omega$ imaginary, and poles and zeros are also generally real, driving a system at a frequency equal to the zero or pole value does not typically mean that the output will be zero or infinite. Rather, the pole and zero values dictate the shape of the Bode plot. The transfer function

$$H(j\omega) = 100 \frac{j\omega + 100}{(j\omega + 1)(j\omega + 10,000)}$$

has one zero at $j\omega = -100$ and two poles at $j\omega = -1$ and $j\omega = -10,000$. Its magnitude is shown in Figure 2-14. If $\omega \ll 1$, $j\omega + 100 \approx 100$, $j\omega + 1 \approx 1$, and $j\omega + 10,000 \approx 10,000$, so the magnitude has a nearly constant value of $100(100)/((1)(10,000)) = 1$, which is shown as the upper horizontal dotted line in the figure. If $1 \ll \omega \ll 100$, then $j\omega + 100 \approx 100$, $j\omega + 1 \approx j\omega$, and $j\omega + 10,000 \approx 10,000$, so the magnitude is approximately

$$|H(j\omega)| \approx 100 \frac{100}{(\omega)(10,000)} = \frac{1}{\omega},$$

which decreases in proportion to $1/\omega$, as indicated by the diagonal dotted line in the figure labeled $1/\omega$. If $100 \ll \omega \ll 10,000$, then $j\omega + 100 \approx j\omega$, $j\omega + 1 \approx j\omega$, and $j\omega + 10,000 \approx 10,000$, so the magnitude is approximately

$$|H(j\omega)| \approx 100 \frac{\omega}{(\omega)(10,000)} = 0.01.$$

which is again nearly constant with respect to ω , as indicated in the figure. Finally, for $\omega \gg 10,000$, the transfer function reduces to $0.01/\omega$, as indicated on the plot. In general, when ω increases past a zero, the straight line that follows the curve is multiplied by ω , and when it increases past a pole, the approximating straight line is divided by ω . On a log-log plot, each

zero increases the slope by 1 and each pole decreases it by one. Specifically, to the left of $\omega = 1$, the line is a constant. After $\omega = 1$, $|H(j\omega)| \approx 1/\omega$, so $\log|H(j\omega)| \approx \log(\omega^{-1}) = -1 \log(\omega)$, with slope -1 . After $\omega = 100$ the slope increases by 1 (back to zero). After $\omega = 10,000$, the slope decreases to -1 . Another pole at a larger ω would decrease the slope by one more to -2 .

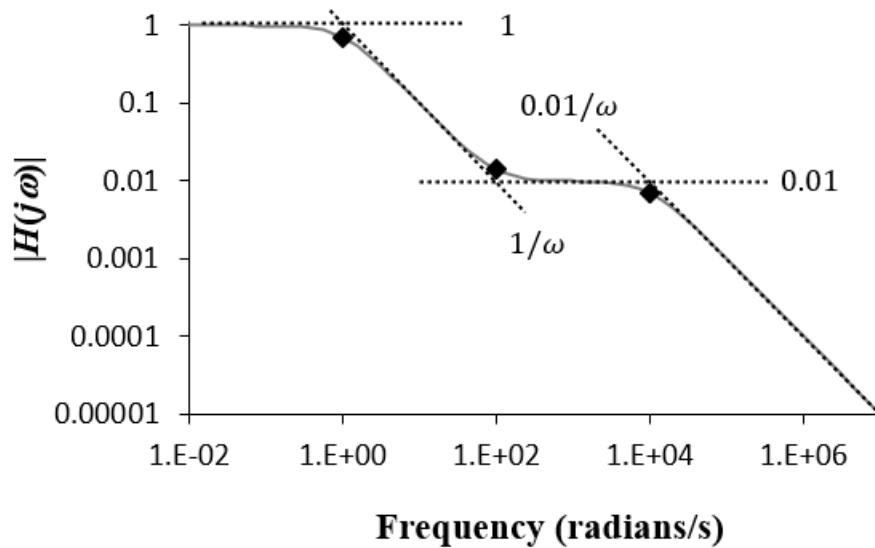


Figure 2-14: Magnitude of the transfer function $H(j\omega) = 100 \frac{j\omega+100}{(j\omega+1)(j\omega+10,000)}$.

The approximating lines intersect with one another at the pole and zero values. The exact curve is lower (higher) by 3 db from the intersections for the poles (zeros).

2.2.3. Time and Delays and Phase Lags

In physiological systems, time delays often occur, for example, as a result of the time required for a nerve signal to travel from its origin in the lower body, through the spine, and then to the brain. In the time domain, a delay of signal $s(t)$ by τ seconds leads to the signal $s(t - \tau)$. In the Fourier domain, it leads to the transformed signal multiplied by $e^{-j\omega\tau}$ (or by $e^{-s\tau}$ in the Laplace domain). Time delays are related to phase delays but differ mathematically. If a sine wave of frequency ω has a phase delay of φ , it becomes $\sin(\omega t - \varphi)$. If it is instead delayed by time τ , it becomes $\sin(\omega(t - \tau))$. Thus, the time delay is related to the phase delay by $\varphi = \omega\tau$. If a system that has transfer function $H_0(j\omega)$ delayed by τ , the delayed transfer function is

$$H(j\omega) = H_0(j\omega)e^{-j\omega\tau}.$$

The magnitude is

$$|H(j\omega)| = |H_0(j\omega)e^{-j\omega\tau}| = |H_0(j\omega)| |e^{-j\omega\tau}|.$$

Because $|e^{-j\omega\tau}| = 1^*$, the magnitude of the delayed transfer function is identical to that of the undelayed transfer function. The phase of $H(j\omega)$ is $\angle H_0(j\omega) - \omega\tau$ because phases of products add and because the phase of $e^{i\theta}$ is simply θ .

Example:

A low pass filter with transfer function

$$H_0(j\omega) = \frac{10}{j\omega + 10}$$

is part of a system that includes a time delay of 50 msec. Find the overall transfer function of the system.

Solution:

The time delay is 50 msec, or 0.05 seconds. The overall transfer function is then

$$H(j\omega) = H_0(j\omega)e^{-j\omega(0.05)} = \frac{10e^{-j\omega(0.05)}}{j\omega + 10}$$

2.2.4. Definitions of Low Pass, High Pass, and Band Pass

In general, a time invariant linear system will enhance signals at some frequencies and attenuate signals at others, as illustrated in the magnitude plot of Figure 2-12. A filter is a system that preferentially passes signal components with specific frequency ranges from its input to its output. For a low pass filter, the gain at low frequency is significantly larger than the gain at high frequency. It is said to pass low frequency components to the output. Such a filter is useful when the input signal is contaminated by high frequency noise or as a filter applied to the signal before digitization to prevent aliasing.[†] Similarly, a high pass filter allows the high frequency components to pass through it, and it would be used to remove low frequency noise. Different configurations of the high pass filter can also be used to take the derivative of a signal or to pre-whiten a signal. Pre-whitening is the process of taking a signal that has strong low frequency components and accentuating the high frequency components before the signal is digitized, and it is used when all frequencies of the signal are important but the weaker high frequency components would be otherwise degraded by the digitization process.[‡] A third filter, the band pass filter, preferentially passes frequencies in a medium range and attenuates signal components at both lower and higher frequencies outside of this range. It is used for signals, such as the electrocardiogram, that include both low frequency drift noise and high frequency noise.

2.3 Semiconductors

Modern electronic devices are constructed from semiconductors. A semiconductor is a material like silicon or germanium. In crystalline form the number of valence electrons of these atoms is equal to the number of atoms to which they are covalently bound, so few electrons, those that

* From Euler's rule, $e^{-j\omega\tau} = \cos(\omega\tau) + j \sin(\omega\tau)$, so the magnitude is $\sqrt{\cos^2(\omega\tau) + \sin^2(\omega\tau)} = 1$.

† Referred to as an anti-aliasing filter.

‡ Pre-whitening will be described later in the section on digital bioinstrumentation.

become dislodged by random fluctuations, are available to conduct electricity. If, however, an impurity, such as phosphorus, with five valence electrons is added to the silicon, the local bonding structure is altered, and the extra electron will conduct electricity. This altered material is referred to as n-type (n for negative). Alternatively, if the impurity has three valence electrons (e.g., boron or aluminum) it causes a location, referred to as a hole, in the semiconductor that can accept an electron and again increase the conductivity. This type of material is referred to as p-type. When n-type and p-type materials are abutted together to form a junction, the free electrons from the n-type material are drawn to the holes in the p-type material, which causes a net negative charge on the p-type material and a net positive charge on the n-type material. This junction is the basis for a semiconductor diode. If a voltage is applied across this junction, with the positive terminal connected to the p-type material and the negative voltage connected to the n-type material, a current flows. If the voltage is applied with the opposite polarity, the addition of more electrons to the p-type material and more holes to the n-type material increases the voltage barrier and impedes the conduction of electricity. Only a small amount of current, known as the reverse saturation current, flows. A transistor consists of two such junctions, with a p-type material sandwiched between two n-type materials (npn transistor) or an n-type material sandwiched between two p-type materials.

Chapter 3: Measurement Equipment

An instrumentation laboratory will have four basic pieces of equipment, a signal generator to produce test signals for an instrument, a power supply to provide power to active elements (e.g. operational amplifiers, transistors, timers, and gates), a multimeter to measure values of resistance, voltage, current, capacitance, and inductance, and an oscilloscope to produce a graphical representation of voltages. Companies produce a wide variety of each of these instruments, with different accuracies and functions, but they all have a number of features in common, and once proficiency is gained with variety, the migration to another variety is straightforward.

3.1 Signal Generator


The signal generator produces a repetitive waveform, such as a sine wave, a square wave, a pulse, a triangle wave (ramp), or a sawtooth wave. Those that we will use also generate a simulated electrocardiogram waveform. The front panel of the Agilent 33220A signal generator is shown in Figure 3-1. The buttons in the middle row on the left allow the user to select the desired waveform (sine, square, ramp, pulse, noise, or arbitrary). The upper row (blue buttons) allows you to input the desired parameters for your waveform. The function of each button changes with the selected waveform. For example, for the sine and square waves, three parameters are needed, the amplitude, frequency, and offset of the signal. For the pulse waveform, an additional parameter is needed to specify the duration of the pulse. Similarly, for the ramp waveform, you can specify separately the duration of the rising part of the ramp and the falling part of the ramp. To input the parameters, press the blue button under the parameter name and use the numerical pad on the upper left to type in the parameter value. You must specify the units of the value with the appropriate blue button before it is accepted by the device.

As an alternative to typing the waveform parameters into the keypad, you can use the knob at the upper left of the front panel. This knob will increase or decrease the given parameter, depending on whether it is turned clockwise or counterclockwise.



Figure 3-1: Agilent 33220A signal generator.

3.1.1. Power-Up

To obtain a signal from the device, use the button marked  to turn it on*, and press the button labeled “output,”† then take the signal from the BNC connection labeled “output.” To change these parameters, select each one with the blue soft buttons and use the keypad to type in the appropriate value. The soft buttons will then change so that you can select the units of the number that you type in (e.g., for frequency, Hz, kHz, MHz). After the units are selected, the parameter will be entered. An option is also provided with the soft buttons to cancel the change, reverting to the original parameter.

3.1.2. Waveforms and Waveform Parameters

The signal generator shown in Figure 3-1 provides six possible waveform shapes, sine wave, square wave, ramp, pulse, noise, and arbitrary. The sine and square wave functions require three parameters, frequency, amplitude, and offset.

The ramp function is used to generate triangular shaped waveforms. In addition to the frequency, amplitude, and offset options, it provides a “symmetry” option that controls the percent of each cycle for which the waveform is rising. If symmetry is set to 50%, a symmetric triangle waveform is generated. If symmetry is set to 100%, a sawtooth waveform is generated (Figure 3-2) a symmetry of 25% generates an asymmetric triangle.

* These devices tend to work much better when they are turned on.

† The device will not provide a signal until this button is pushed.

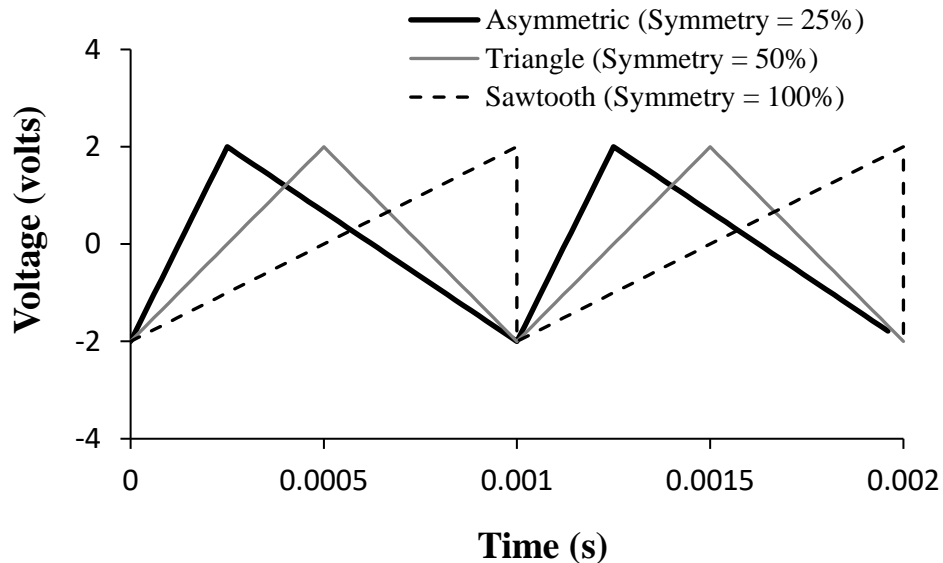


Figure 3-2: Example waveforms generated by the ramp waveform button, with frequency set to 1 kHz and amplitude set to 4 volts peak-to-peak.

The pulse function generates a repeated pulse, where the duty cycle is one of the soft button inputs. A duty cycle of 50% will generate a square wave, while a duty cycle of 1% will generate something that resembles a train of delta functions. For this waveform, the offset is halfway between the maximum and minimum voltage of the waveform.*

The noise waveform is Gaussian white noise with the selected amplitude and offset. Its bandwidth is 10 MHz,[†] independent of the frequency value selected by the user. The frequency parameter has no effect on this waveform.

The arbitrary waveform can be used to generate specialized signals. Of particular interest to biomedical engineering is the cardiac waveform, which is obtained from the Arb button and then the cardiac soft button. A physiological frequency should be selected (e.g., 1 Hz).

3.1.3. Output Impedance

An attribute of the device that may not be initially obvious is that it can account for the impedance of the device that it is driving. If the correct impedance setting is not selected, the voltage output indicated by the device's display will not correspond to the voltage value at the output. In almost all cases in this class, the signal generator will be connected to a circuit or device with high input impedance (such as the combination of the oscilloscope and 10x probe, which has an input impedance of 10 M Ω). The default for the signal generator is a device with an input impedance of 50 Ω . This value matches the signal generator's output impedance and would maximize the amount of power driving the device. If the signal generator is set to 2 volts, for example, it generates a 4-volt signal so that the value of the voltage at the device input is 2 volts; the 50 Ω output resistance and the assumed 50 Ω input impedance form a simple bridge

* This definition of "offset" seems obvious from the user's perspective, but it should be noted that it differs from the mean value of the signal if the duty cycle differs from 50%.

[†] True white noise would have an infinite bandwidth, but 10 MHz is sufficiently high for most applications.

network that divides the source voltage by 2. To select the needed impedance setting, press the “utility” button, followed by “output setup” (which will appear above one of the blue buttons) and select “High Z.”

3.1.4. Modulation

The device can generate signals that are amplitude modulated or frequency modulated. The parameters for the carrier signal (the signal to be modulated) are set through the same process that was used with the unmodulated signal. The modulating signal (the signal that gradually varies the signal’s amplitude or frequency) can come from either the signal generator itself or an external input port.

3.1.5. Synchronization

In various applications, it is useful to have a signal that is separate from the generated waveform and that can be used to synchronize some aspect of the device being driven with the waveform. The signal has a constant minimum and maximum values of 0 and 5 volts, which is standard for transistor-transistor logic (TTL). It can thus be used, for example, to turn another device on or off in a pattern that follows the signal frequency. This signal is available from the “Sync” port on the front panel. If the main output signal is a sine wave, a ramp, a pulses train, or an arbitrary waveform, the sync signal is a square wave with a 50% duty cycle. The sync signal has the same duty cycle as the main signal when the main signal is a square wave. For a modulated waveform based on an internally generated signal, the sync signal has the same frequency as the modulating waveform but for a modulated waveform based on an externally generated signal, the sync signal is based on the carrier frequency.

3.2 Power Supply

Figure 3-3 shows the front panel of an Agilent 3630A double-ended power supply, which will provide the DC supplies for operational amplifier circuits. The device can output three DC voltages simultaneously, a positive voltage that ranges from 0 to 20 volts from the terminal marked +20V, a negative voltage that ranges from –20 to zero volts from the terminal marked –20V, and a positive voltage that ranges from 0 to 6 volts from the terminal that is marked +6V. In all cases, the voltages are relative to the ground terminal voltage marked COM. The three voltages can be adjusted individually. The knob marked $\pm 20V$ adjusts the magnitude of the –20V and +20V terminals, and the tracking knob can be used to increase one of these terminals with respect to the other. The knob marked +6V adjusts the +6V terminal voltage. The two values on the meter at the upper left indicate voltage and current at the outputs. Voltage and current from only one output can be displayed at a given time. To specify which output is connected to the meter, use the buttons labeled “Meter.” These buttons have no effect on the output itself. For example, if you set the +20V output to +15 volts and then depress the +6V meter button, the +20V output will continue to output +15 volts. Yellow lights to the right of the meter display indicate when the needed output current is larger than the amount of current that the device can generate, as when the output voltage is set to a large value and the input impedance of the device being supplied is low.*

* These lights might indicate that you have shorted the output to ground, or that the circuit requires more power than can be delivered by the supply, as happens in some cases with an operational amplifier such as the LM675T.



Figure 3-3: Agilent E3630A dual-ended power supply.

3.3 Multimeter

The Agilent 34410A digital multimeter is shown in Figure 3-4. This device can be used to measure values of voltage and current in a circuit or to measure the values of resistances and capacitances for circuit components.*



Figure 3-4: Agilent 34410A Digital Multimeter

For most measurements (e.g., voltage, capacitance, and 2-wire resistance), the device to be measured is connected to the upper and middle inputs to the upper right of the front panel, with the ground (for voltage measurements) being the middle input.

* Do not attempt to measure the values of resistance and capacitances when the components are connected to the circuit. Doing so will lead to incorrect values for these parameters.

3.3.1. Voltage

To measure voltage, connect the red test lead to the voltage to be measured and the black test lead to ground. To measure the mean value of voltage (the offset), select “DC V.” To measure the RMS voltage, select “AC V.”

3.3.2. Current

Current is a “through” variable. To be measured, it must pass through the device or probe that senses it. In this case, the current must pass through the digital multimeter. A simple method for a device to do this measurement is to have the current pass through a small resistor (inside the multimeter) and to measure the voltage across that resistor. The measurement requires the port marked “I” on the lower right to be used instead of the port on the upper right. The measurement configuration is shown in Figure 3-5. Here, the value of current through resistor R_2 is desired (left). Resistors R_1 and resistor R_2 are disconnected from each other so that the multimeter can access the current. The current then flows out of resistor R_1 , into the multimeter through the I port, then back out of the multimeter through the LO port, through R_2 and then to ground.

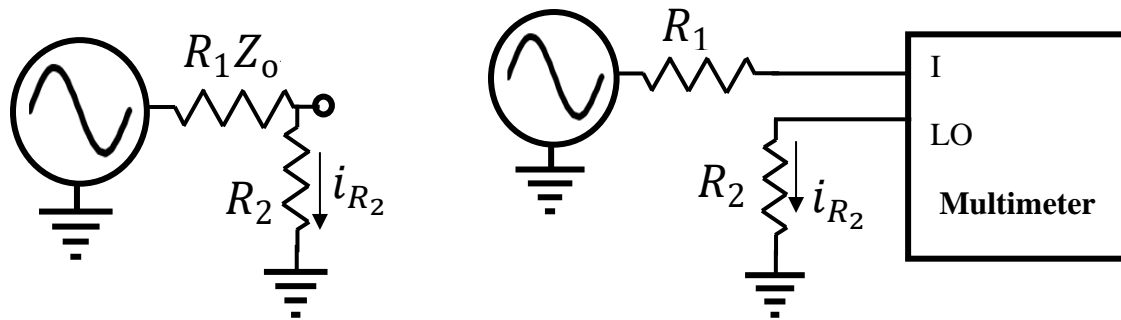


Figure 3-5: Measurement configuration for the current measurement. To measure the current through resistor R_2 , on the left, the multimeter is configured as on the right.

3.3.3. Capacitance

To measure capacitance, connect the two input leads across the capacitor and select “shift” followed by the button marked “ $\text{—}||\text{—}$.” Some multimeters will not have this function.

3.3.4. Two-Wire and Four-Wire Resistance

The two-wire resistance measurement is nearly always sufficient for measurements performed in this laboratory. To make this measurement, connect the two input leads to either side of the resistor and select the Ω 2W button. In this mode, the multimeter injects a small current (i_{test}) through the two leads and measures the resulting voltage (V_{meas}) between the two input ports. It then calculates resistance (R) as $R = V_{\text{meas}}/i_{\text{test}}$. The Ω 4W mode is used if large resistances occur within one or both test leads. In that case, the lead resistance would be included in the two-wire measurement. In the four-wire mode (selected as “shift” followed by the “ Ω 4W” button), the test leads are connected as before, but leads from the two output ports to the left of the test lead ports are also connected to these locations (red leads connected to the same location, and black leads connected to the same location). The two ports to the left emit the constant current, while the two ports to the right measure the voltage.

3.4 Oscilloscope

The oscilloscope provides a graphical representation of the input signal. Most oscilloscopes have two input channels so that two signals can be displayed at the same time. You will generally choose to use Channel 1 for the input signal and Channel 2 for an output signal. Controls on the oscilloscope allow you to adjust the size of the signal on the screen and the amount of time shown on the display. While these controls are somewhat intuitive, the oscilloscope has more advanced functions that are described in Laboratory 1, the tutorial on oscilloscopes.

3.5 Bread Board

Circuit development would be tedious if all elements needed to be directly soldered together before they were tested. The bread board* is a platform that allows components to be easily connected during prototyping. An example bread board is shown in Figure 3-6.

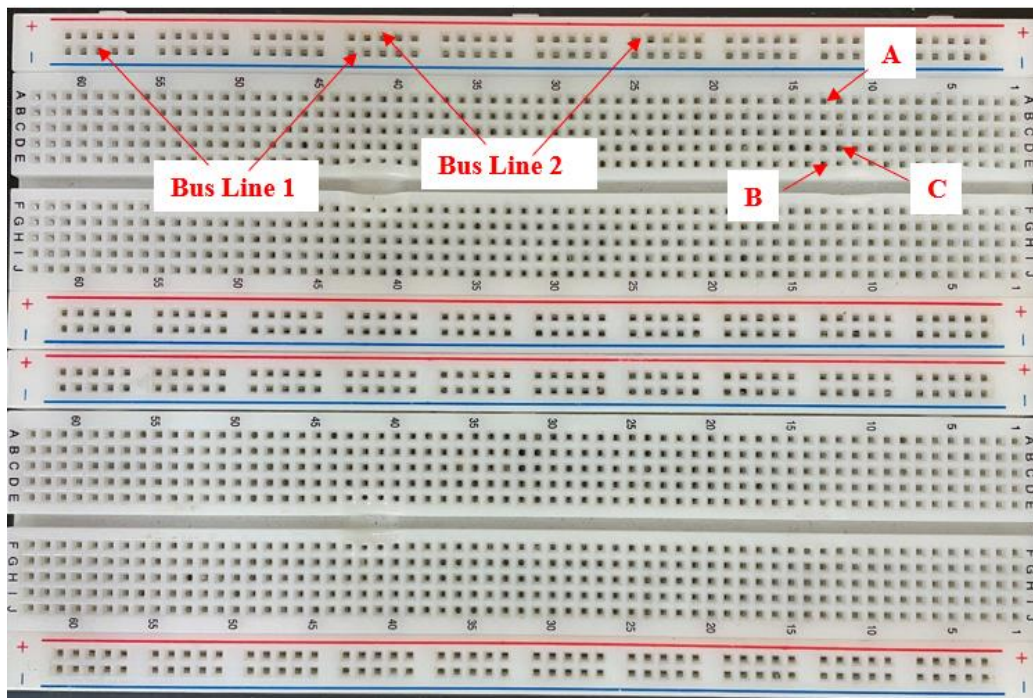


Figure 3-6: Example bread board.

Holes along a row, whether long or short, are connected. For example, a wire inserted into the hole marked “A” will be connected through the board to a wire inserted into the hole marked “B.” but will not be connected to the hole marked “C.” The long rows of holes are intended to be used as bus lines, which carry supply voltages and ground. Bus Line 1 and Bus Line 2 are not connected to each other. They are long because, in general, these voltages need to be connected at multiple places in the circuit. For example, if a circuit uses four operational amplifiers, each one requires a positive and negative power supply. (The supplies could be daisy chained, but

* The story is that before commercial bread boards were available, designers would drive nails into a literal bread board (or other flat wooden base) to provide support for components that would then be soldered onto the nails.

topologically it is easier to debug a circuit when all supplies of a given value are connected to the same bus. The shorter rows are used to interconnect the components. Integrated circuit chips, such as operational amplifiers and 555 timers, are placed between short rows so that the set of pins on one side connect to the row of (usually 5) holes on that side and the pins on the other side of the chip connect to the row of holes on the other side.

Chapter 4: Ideal Operational Amplifiers

4.1 Characteristics of Ideal Operational Amplifiers

An operational amplifier is a differential amplifier with specific characteristics. A differential amplifier is a device that takes voltages from two inputs, subtracts one from the other, and then amplifies the result by a factor A . Thus, if v_1 is the voltage at the negative input and v_2 is the voltage at the positive input, the voltage at the output (v_{out}) is

$$v_{out} = A(v_2 - v_1).$$

In general, a differential amplifier will have a finite input impedance, Z_{in} , which means that it will absorb some current from its two inputs. The specific characteristics of an ideal operational amplifier are that the input impedance is infinite (it does not absorb current from its input) and its gain is infinite. In practice, these two characteristics cannot be achieved, so real operational amplifiers are designed with large finite gain (typically above 10^5) and large input impedance (typically above 1 M Ω for the 741, and at least 15 M Ω for the OP07). If the device were used directly to amplify the difference between two signals, it would not be practical. However, it is configured with negative feedback such that the overall circuit provides a tunable finite gain.

4.2 Inverting Amplifier

The most basic operational amplifier configuration is the inverting amplifier, shown in Figure 4-1. A feedback resistor, R_f , is connected from the output to the inverting operational amplifier input, v^- , and an input resistor, R_i , is connected from the circuit input, v_{in} , to the inverting operational amplifier input. The positive operational amplifier input, v^+ , is connected to ground.

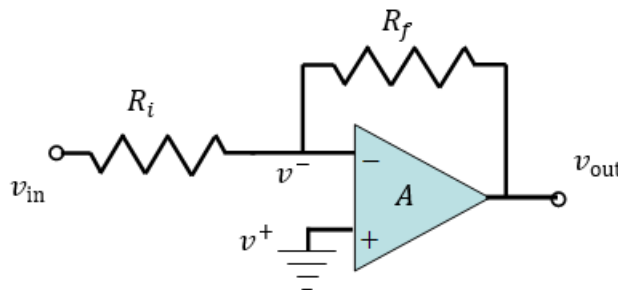


Figure 4-1: Inverting amplifier configuration for an operational amplifier.

The objective is to find the relationship between v_{out} and v_{in} for this circuit. It will be assumed initially that A is finite, after which the limit of the result will be taken as $A \rightarrow \infty$. The assumption of infinite input resistance at the operational amplifier terminals is made throughout, so that the currents into the operational amplifier terminals are zero amps. Three equations are used. First, the current through the input resistor is

$$i = \frac{v_{\text{in}} - v^-}{R_i}. \quad \text{Equation 4-1}$$

Next, all of this current must pass through R_f because the operational amplifier input takes no current.

$$i = \frac{v^- - v_{\text{out}}}{R_f}. \quad \text{Equation 4-2}$$

Finally, the equation for the operational amplifier is used to relate v_{out} to v^- and v^+ . Since the positive input is grounded, $v^+ = 0$ and

$$v_{\text{out}} = A(v^+ - v^-) = -Av^-. \quad \text{Equation 4-3}$$

Eliminate i from equations Equation 4-1 and Equation 4-2.

$$\frac{v_{\text{in}} - v^-}{R_i} = \frac{v^- - v_{\text{out}}}{R_f}. \quad \text{Equation 4-4}$$

Use Equation 4-3 to eliminate v^- .

$$\frac{v_{\text{in}} + v_{\text{out}}/A}{R_i} = -\frac{v_{\text{out}}/A + v_{\text{out}}}{R_f}. \quad \text{Equation 4-5}$$

Separate v_{out} on the left and v_{in} on the right.

$$v_{\text{out}} \left(\frac{1}{R_i A} + \frac{1}{R_f A} + \frac{1}{R_f} \right) = -\frac{v_{\text{in}}}{R_i}$$

Solve for the gain, $v_{\text{out}}/v_{\text{in}}$.

$$\frac{v_{\text{out}}}{v_{\text{in}}} = -\frac{1}{R_i} \left(\frac{1}{\frac{1}{R_i A} + \frac{1}{R_f A} + \frac{1}{R_f}} \right)$$

Multiply the right-hand side by $R_i R_f A$.

$$\frac{v_{\text{out}}}{v_{\text{in}}} = -\frac{1}{R_i} \left(\frac{R_i R_f A}{R_f + R_i + A R_i} \right).$$

Now consider the case where $A \rightarrow \infty$. The terms R_f and R_i in the denominator on the right are then much smaller than $A R_i$ and can be ignored, so that

$$\frac{v_{\text{out}}}{v_{\text{in}}} = -\frac{1}{R_i} \left(\frac{R_i R_f A}{A R_i} \right) = -\frac{R_f}{R_i}. \quad \text{Equation 4-6}$$

It follows that the gain of the circuit is $-R_f/R_i$, which can easily be changed with different values of the resistances. It will be shown that derivations of device gains can be more readily obtained, without the need to include the gain factor A in the analysis.

An important result from this analysis can be deduced from Equation 4-3. Because $v_{out} = A(v^+ - v^-)$, it follows that $(v^+ - v^-) = v_{out}/A$, and because $A \rightarrow \infty$, $(v^+ - v^-) \rightarrow 0$, so that the two operational amplifier inputs have the same voltage on them. For this reason, even though the negative input is not directly grounded, it must have the same voltage as ground. It is therefore referred to as a “virtual ground.”

4.3 Generalized Inverting Amplifier (Z Replaces R)

The gains (transfer functions) of more complex operational amplifiers that are based on the inverting amplifier configuration can be obtained from Equation 4-6 if resistances R_f and R_i are replaced with impedances Z_f and Z_i .

$$H(j\omega) = \frac{v_{out}}{v_{in}} = -\frac{Z_f}{Z_i}. \quad \text{Equation 4-7}$$

Consider the circuit shown in Figure 4-2.

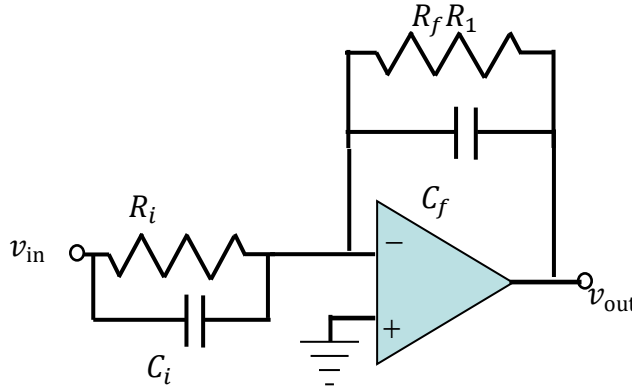


Figure 4-2: Operational amplifier circuit based on the inverting configuration, but with more complicated feedback and input networks.

The impedance for the feedback segment is the parallel combination of R_f and C_f , where the impedance of the capacitor is $1/(j\omega C_f)$. From the equation for parallel impedance,

$$Z_f = \frac{Z_R Z_C}{Z_R + Z_C} = \frac{R_f \left(\frac{1}{j\omega C_f} \right)}{R_f + \frac{1}{j\omega C_f}}$$

Multiply the right-hand side by $j\omega C_f$.

$$Z_f = \frac{j\omega R_f C_f}{j\omega R_f C_f + 1}.$$

The input segment follows the same pattern (resistor R_i in parallel with capacitor C_i), so

$$Z_i = \frac{j\omega R_i C_i}{j\omega R_i C_i + 1}.$$

Now use Equation 4-6, but replace R_f and R_i are with Z_f and Z_i .

$$\frac{v_{out}}{v_{in}} = -\frac{Z_f}{Z_i} = -\frac{\frac{j\omega R_f C_f}{j\omega R_f C_f + 1}}{\frac{j\omega R_i C_i}{j\omega R_i C_i + 1}} = -\frac{R_f C_f}{j\omega R_f C_f + 1} \left(\frac{j\omega R_i C_i + 1}{R_i C_i} \right)$$

The result is a frequency-dependent transfer function with one pole and one zero,

4.4 Filters Based on the Generalized Inverting Configuration

4.4.1. Low Pass

A simple low pass filter is shown in Figure 4-3, with a capacitor and resistor in the feedback path and a resistor in the input path. From the generalized inverting amplifier equation,

$$H(j\omega) = -\frac{Z_f}{Z_i}$$

where

$$Z_f = \frac{R_f \left(\frac{1}{j\omega C_f} \right)}{R_f + \frac{1}{j\omega C_f}} = \frac{R_f}{j\omega R_f C_f + 1}; \quad Z_i = R_i$$

So

$$\frac{v_o}{v_i} = \frac{R_f}{R_i} \left(\frac{1}{j\omega R_f C_f + 1} \right).$$

For low frequency ($\omega \ll \frac{1}{R_f C_f}$), the gain magnitude is $\frac{R_f}{R_i}$. For high frequency, ($\omega \gg \frac{1}{R_f C_f}$), the gain is $\frac{1}{\omega R_i C_f}$, so that high frequencies in the input signal are eliminated at the output.

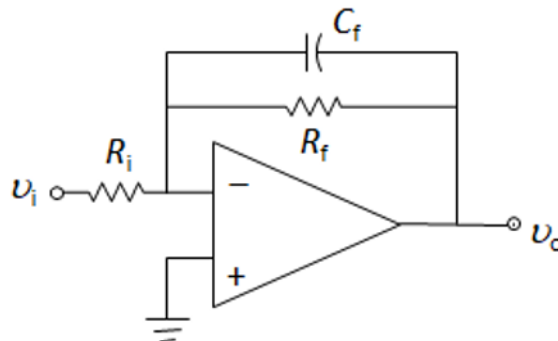


Figure 4-3: Simple single-pole low pass filter

4.4.2. High Pass

The single-pole high pass filter is shown in Figure 4-4. The feedback impedance is $Z_f = R_f$, and the input impedance is $R_i + \frac{1}{j\omega C_i}$, so the transfer function is

$$H(j\omega) = \frac{R_f}{R_i + \frac{1}{j\omega C_i}} = \frac{j\omega C_i R_f}{j\omega C_i R_i + 1} = \frac{R_f}{R_i} \frac{j\omega C_i R_i}{j\omega C_i R_i + 1}.$$

For $\omega \ll \frac{1}{C_i R_i}$, $H(j\omega) \approx j\omega C_i R_f$, which decreases to zero as ω decreases to zero, and for $\omega \gg \frac{1}{C_i R_i}$, $H(j\omega) \approx \frac{R_f}{R_i}$. Thus, the filter attenuates low frequency signals and keeps high frequency signals. In the above equation, the $\frac{R_f}{R_i}$ was factored from the rest of the transfer function because $j\omega C_i R_i / (j\omega C_i R_i + 1) \rightarrow 1$ for large values of ω , so that form emphasizes that the gain in the pass band is $\frac{R_f}{R_i}$.

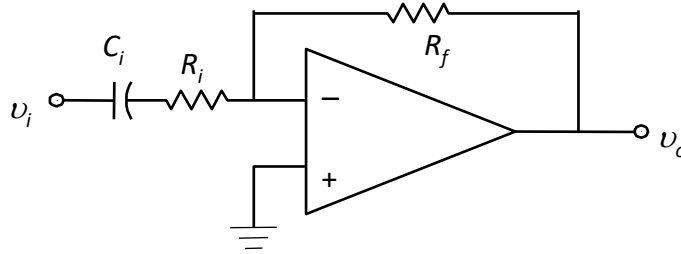


Figure 4-4: Single-pole high pass filter.

4.4.3. Band Pass

The single pole band pass filter is shown in Figure 4-5. The impedances are

$$Z_f = \frac{R_f \left(\frac{1}{j\omega C_f} \right)}{R_f + \frac{1}{j\omega C_f}} = \frac{R_f}{j\omega R_f C_f + 1}; \quad Z_i = R_i + \frac{1}{j\omega C_i} = \frac{j\omega R_i C_i + 1}{j\omega C_i}$$

The transfer function is

$$H(j\omega) = \frac{j\omega R_f C_i}{(j\omega R_f C_f + 1)(j\omega R_i C_i + 1)} = \frac{R_f}{R_i} \frac{j\omega R_i C_i}{(j\omega R_f C_f + 1)(j\omega R_i C_i + 1)}$$

Generally, $\frac{1}{R_i C_i} < \frac{1}{R_f C_f}$. The transfer function is the product of the low pass and high pass filter transfer function (without the gain factors) multiplied by the gain factor. Thus, it grows in proportion to ω at frequencies below $\frac{1}{R_i C_i}$ and decays in proportion to $\frac{1}{\omega}$ for frequencies above $\frac{1}{R_f C_f}$, with a gain in the pass band of $\frac{R_f}{R_i}$.

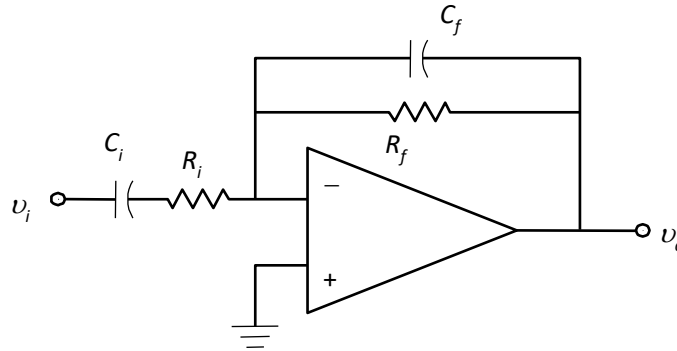


Figure 4-5: Single-pole band pass filter.

Example

An electrocardiogram amplifier is desired that removes baseline drift and high frequency noise. The desired low frequency cutoff frequency is 0.1 Hz, and the desired high frequency cutoff frequency is 100 Hz. Furthermore, a gain of 10 is desired within the pass band. Design a bandpass filter based on the configuration in Figure 4-5 to fill this need.

Solution

Three design criteria were specified, but four parameters are available to fulfill them (R_f , C_f , R_i , and C_i). One of these values can be selected without consideration of the design criteria. For reasons that will be explained later, a value of 100 k Ω will be selected for R_i . Then a value of $R_f = 100$ k Ω will provide the gain of 10 in the pass band. Since $\frac{1}{R_f C_f}$ is the cutoff frequency for the low pass portion of the filter, $\frac{1}{R_f C_f} = 2\pi(100 \text{ Hz})$, so $C_f = 1/((100,000 \text{ } \Omega)(2\pi)(100 \text{ Hz})) = 15.9 \text{ nF}$. Similarly, $\frac{1}{R_i C_i}$ is the low frequency cutoff for the high pass portion of the filter, so $C_i = 1/((100,000 \text{ } \Omega)(2\pi)(0.1 \text{ Hz})) = 15.9 \text{ } \mu\text{F}$.

4.5 Simplified Analysis Based on Two Fundamental Rules

While the derivation of the inverting amplifier configuration in Section 4.2 required several lines of text, the derivation provided some information that can be used to simplify analysis of other circuits. The simplified analysis is based on two simple rules.

4.5.1. The Two Rules

If the input voltages differ by any amount with an infinite gain, the output voltage will be infinite (or, in practical terms, it will go to saturation). Thus, if the output is not in saturation, the two input voltages must be identical. Furthermore, the assumed infinite impedance of the operational amplifier inputs no current must pass into these inputs. The two basic rules are then

Rule 1:

When the op-amp output is in its linear range, the two input terminals are at the same voltage.

Rule 2:

No current flows into either input terminal of the op-amp.

With these two rules, we easily obtain the gain of the inverting amplifier from stating that the current into the input resistor is equal to the current through the feedback resistor.

$$i = \frac{v_{in} - 0}{R_{in}} = \frac{(0 - v_{out})}{R_f} \Rightarrow \frac{v_{out}}{v_{in}} = \frac{R_f}{R_i}.$$

4.5.2. Simplified Non-Inverting Amplifier Derivation

With these two rules, we easily obtain the gain of the inverting amplifier from stating that the current into the input resistor $((v_{in} - v^-)/R_{in} = (v_{in} - v^+)/R_{in} = (v_{in} - 0)/R_{in})$ is equal to the current through the feedback resistor $((0 - v_{out})/R_f)$.

$$i = \frac{v_{in} - 0}{R_{in}} = \frac{(0 - v_{out})}{R_f} \Rightarrow \frac{v_{out}}{v_{in}} = \frac{R_f}{R_i}.$$

4.5.3. Voltage Follower

The voltage follower is shown in Figure 4-6. The output voltage is directly connected to the inverting input. Because the two inputs must be identical if the output is finite, the negative input must be equal to v_{in} . Therefore, the output voltage must also be equal to v_{in} , so that the gain of the circuit is

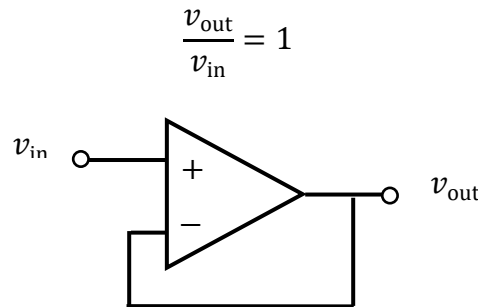


Figure 4-6: Voltage follower.

The circuit therefore has no advantage in terms of gain, but it has a substantial benefit in that its input impedance is large, while the output impedance is small. Consequently, it can be placed between an source with high output impedance and a device with low input impedance so that the discrepancy in impedance does not cause loss of signal. For example, if one were to directly drive a DC motor, which has an input resistance of $8\ \Omega$ with a signal generator that has an output resistance of $50\ \Omega$, the two resistances would act as a bridge and reduce the driving voltage by a factor of

$$\frac{R_{motor}}{R_{source} + R_{motor}} = \frac{8}{50\ \Omega + 8\ \Omega} \approx 0.14.$$

Thus, if 2 volts between the input leads of the motor is sufficient to drive it, a signal generator set to 2 volts would not drive the motor because the voltage divider effect would lead to a voltage of only $2(0.14) = 0.28$ volts across the motor. To compensate for this loss, one can place an LM675 power operational amplifier in the voltage follower configuration between the source and motor. The large input resistance of the voltage follower prevents the source output impedance

from leading to voltage loss, and the small output resistance of the voltage follower prevents loss between it and the motor,

4.5.4. Summing Amplifier

The inverting amplifier can be modified as shown in Figure 4-7 to include multiple input resistors, each of which is connected to a different input voltage. The inverting input acts as a virtual ground. The combined current through R_a , R_b , and R_c must pass through R_f . Therefore,

$$i_{R_f} = \frac{v_1}{R_a} + \frac{v_2}{R_b} + \frac{v_3}{R_c}.$$

Because

$$\underbrace{0}_{\text{Virtual Ground Voltage}} - \underbrace{i_{R_f} R_f}_{\text{Voltage Drop Across } R_f} = v_{\text{out}},$$

$$v_{\text{out}} = - \left(\frac{v_1}{R_a} + \frac{v_2}{R_b} + \frac{v_3}{R_c} \right) R_f.$$

If $R_a = R_b = R_c \equiv R_i$, then the output is

$$v_{\text{out}} = - \frac{R_f}{R_i} (v_1 + v_2 + v_3).$$

So that the configuration sums the inputs with a gain of $-R_f/R_i$. The output can also be made to weight the three input voltages differently, if the values of R_a , R_b , and R_c are selected accordingly.

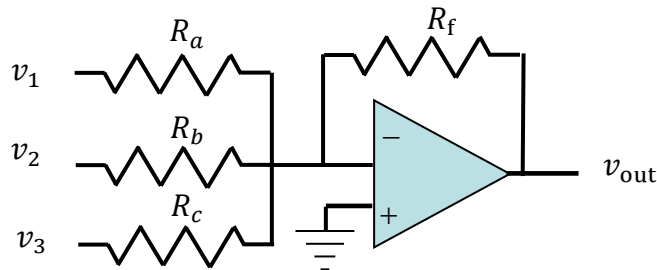


Figure 4-7: Summing Amplifier

Feedback control applications generally require feedback and feed-forward signal paths that are added to and subtracted from the main signal. The summing amplifier and its analog, the summing integrator are essential to this task.

4.5.5. Differentiator

The output of the circuit in Figure 4-8 is the negative of the derivative of the input voltage, as is best illustrated in the time domain, rather than the frequency domain. Because the inverting input is a virtual ground, the current through the input capacitor is

$$i_{\text{in}} = C_i \frac{dv_{\text{in}}}{dt}.$$

This current passes through R_f so that

$$0 - v_{\text{out}} = R_f C_i \frac{dv_{\text{in}}}{dt} \Rightarrow v_{\text{out}} = -R_f C_i \frac{dv_{\text{in}}}{dt}.$$

The factor $R_f C_i$ specifies the gain.

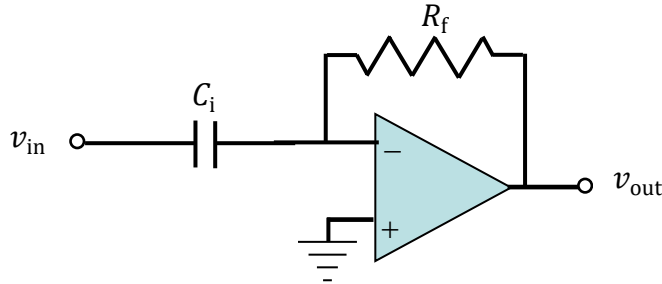


Figure 4-8: Differentiator

The result can also be obtained in the frequency domain through the impedance relationship of Equation 4-7.

$$Z_f = R_f; \quad Z_i = \frac{1}{j\omega C_i}$$

$$H(j\omega) = -\frac{Z_f}{Z_i} = \frac{R_f}{1/j\omega C_i} = j\omega R_f C_i.$$

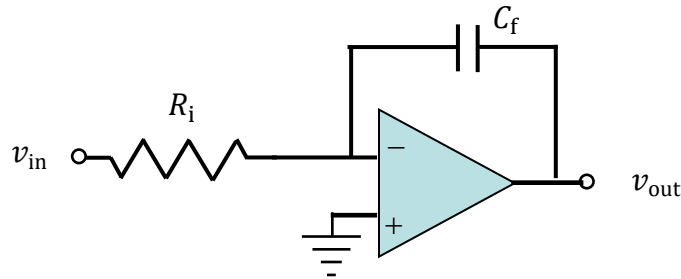
Recall here that multiplication by ω in the frequency domain is equivalent to differentiation in the time domain.

In practice a differentiator will cease to increase beyond a given frequency, and such behavior is not desirable because it would ultimately lead to excessively large gain for large frequencies. The limitations of operational amplifier bandwidth (to be discussed in Section 4.7) limit the high frequency gain, but it may be necessary to also build that limitation in the circuit. One can add a resistor in series with C_i (which converts the circuit into the high pass filter described in Section 4.4.2) or place a capacitor in parallel with the feedback resistor (which is equivalent to passing the signal through a low pass filter).

4.5.6. Integrator

The integrator circuit is shown in Figure 4-9. The current through R_i is v_{in}/R_i , which is equal to $-C_f dv_{\text{out}}/dt$. The output voltage is then

$$v_{\text{out}} = \frac{1}{R_i C_f} \int v_{\text{in}} dt.$$

**Figure 4-9: Integrator**

Again, the result can be obtained in the frequency domain through the impedance relationship of Equation 4-7.

$$Z_f = \frac{1}{j\omega C_f}; \quad Z_i = R_i$$

$$H(j\omega) = -\frac{Z_f}{Z_i} = \frac{1/(j\omega C_f)}{R_i} = \frac{1}{j\omega R_i C_f}.$$

This circuit, when used by itself, is marginally stable in that any DC offset voltage on v_{in} will eventually lead to an infinite output. In practice, the voltage will increase (for a negative offset) or decrease (for a positive offset) until it nears the positive or negative supply voltage. However, the circuit is of practical value in the design of feedback control systems.

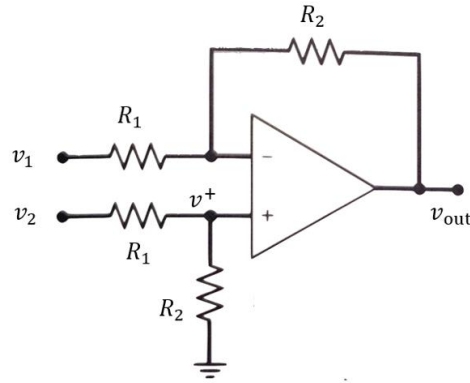
4.5.7. Summing Integrator

As with the summing amplifier (Section 4.5.4) multiple input resistors, (e.g., R_1 , R_2 , and R_3) can be attached in parallel with different voltages, (v_1 , v_2 , and v_3 , respectively), applied to them. The device then becomes a summing integrator, where the output is

$$v_{out} = \frac{1}{C_f} \int \left(\frac{v_1}{R_1} + \frac{v_2}{R_2} + \frac{v_3}{R_3} \right) dt$$

4.5.8. Differential Amplifier

The circuit in Figure 4-10 is a differential amplifier which amplifies the difference between the voltages at the positive and negative inputs, i.e., $v_{out} = \frac{R_2}{R_1}(v_2 - v_1)$

**Figure 4-10: Differential Amplifier**

The voltage at the positive input can be easily found because the R_1 and R_2 attached to that node form a voltage divider.

$$v^+ = v_2 \frac{R_2}{R_1 + R_2} \text{ (voltage divider)} \quad \text{Equation 4-8}$$

On the negative input side, the current through R_1 must be equal to the current through the feedback resistor R_2 .

$$\begin{aligned} \frac{v_1 - v^-}{R_1} &= \frac{(v^- - v_{\text{out}})}{R_2} \\ \Rightarrow v_{\text{out}} &= -\frac{v_1 R_2}{R_1} + \frac{R_2 v^-}{R_1} + v^- \end{aligned}$$

But because $v^- = v^+$ if v_{out} is finite, Equation 4-8 can be used to replace v^- .

$$v_{\text{out}} = -\frac{v_1 R_2}{R_1} + \frac{v_2 R_2^2}{R_1(R_1 + R_2)} + \frac{v_2 R_2}{R_1 + R_2}$$

Factor out $-\frac{R_2}{R_1}$.

$$v_{\text{out}} = -\frac{R_2}{R_1} \left(v_1 - \frac{v_2 R_2}{R_1 + R_2} - \frac{v_2 R_1}{R_1 + R_2} \right)$$

Combine the last two terms.

$$v_{\text{out}} = -\frac{R_2}{R_1} \left(v_1 - \frac{v_2 (R_1 + R_2)}{R_1 + R_2} \right)$$

The factor $R_1 + R_2$ cancels $R_1 + R_2$ in the denominator.

$$v_{\text{out}} = \frac{R_2}{R_1} (v_2 - v_1)$$

The two R_1 must be well-matched, as must the two R_2 resistors. A disadvantage of this circuit is that its input impedance can be too low for some applications, such as the amplification of

biopotentials from the skin, which have output impedances on the order of 10 k Ω . It is therefore often coupled with a section with high input impedance discussed in the next section.

An important measure for a differential amplifier is the common mode rejection ratio. Ideally, the output of the amplifier is equal to the gain multiplied by the difference between the two inputs, $v_{\text{out}} = G(v_2 - v_1)$. However, if both v_2 and v_1 are changed equally, it will still affect v_{out} , so that a more accurate equation is $v_{\text{out}} = G_d(v_2 - v_1) + G_c(v_2 + v_1)/2$. G_c is the common mode gain that describes the effect of the average of v_2 and v_1 on the output. For a good differential amplifier, G_c should be much less than G_d , the differential gain. Various factors affect G_c , including mismatch between resistors on the positive and negative sides of the constructed amplifier. The common mode rejection ratio (CMRR) describes the degree to which the amplifier's output is immune to the common mode gain. Specifically,

$$\text{CMRR} = \frac{G_d}{G_c}.$$

A large value of CMRR indicates a good differential amplifier. CMRR may also be expressed in db.

$$\text{CMRR}_{\text{db}} = 20 \log \left(\frac{G_d}{G_c} \right).$$

4.5.9. Differential Amplifier Impedance Stage

To increase the input impedance of the differential amplifier, it is preceded by the input stage shown in Figure 4-11. By Rule 1, the voltage at node a is v_1 , and the voltage at node b is v_2 . The current through R_1 is then $i = (v_1 - v_2)/R_1$. By Rule 2, i is also the current through both of the R_2 resistors. Thus,

$$\begin{aligned} v_3 - v_1 &= iR_2 = \frac{R_2}{R_1}(v_1 - v_2) \\ v_2 - v_4 &= iR_2 = \frac{R_2}{R_1}(v_1 - v_2) \end{aligned}$$

Add these two equations.

$$\begin{aligned} v_3 - v_4 &= \frac{2R_2}{R_1}(v_1 - v_2) + (v_1 - v_2) \\ \frac{(v_3 - v_4)}{v_1 - v_2} &= \frac{2R_2}{R_1} + 1 = \frac{2R_2 + R_1}{R_1} \end{aligned}$$

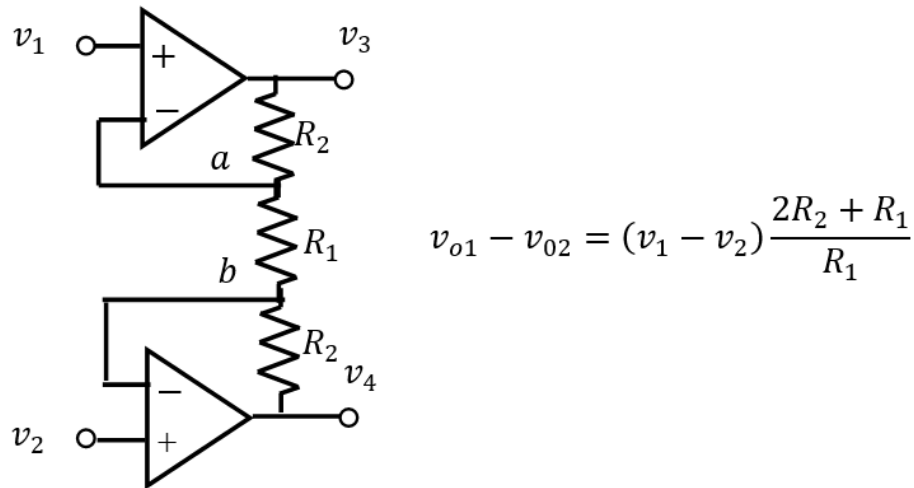


Figure 4-11: Impedance stage of the differential amplifier.

4.5.10. Instrumentation Amplifier (ina118 or LT1168 in LTSpice)

The stage in Figure 4-11 does not provide a single-ended output, so it must be combined with the stage in Figure 4-10. It is possible to purchase integrated circuits that combine these two stages into a single instrumentation amplifier, which will be used later in this course. Examples of this integrated circuit are the ina118 and the LT1168. The LT1168 is available as a component in LTSpice.

4.5.11. Sallen and Key Configuration

The Sallen and Key configuration* leads to a second-order transfer function with a single operational amplifier. From Rule 1, $V^+ = V^- = V_{out}$. The current through R_1 must equal the current through C_2 .

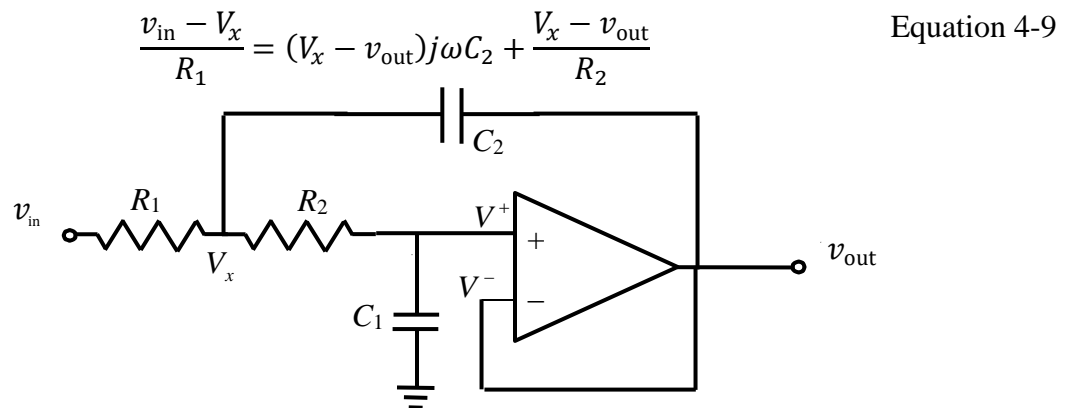


Figure 4-12: Sallen and Key configuration.

* The configuration shown here is simplified from a more general Sallen and Key circuit. The purpose is to consider the simplicity of circuit derivation with the two basic assumptions rather than to provide detail about the configuration itself.

The current through R_2 is equal to the current through C_1 .

$$\frac{V_x - v_{\text{out}}}{R_2} = v_{\text{out}} j\omega C_2 \quad \text{Equation 4-10}$$

Use Equation 4-10 in both terms on the right hand side of Equation 4-9.

$$\frac{v_{\text{in}} - V_x}{R_1} = v_{\text{out}} j\omega C_2 R_2 (j\omega C_2) + v_{\text{out}} j\omega C_2 \quad \text{Equation 4-11}$$

Solve Equation 4-10 for V_x .

$$V_x = v_{\text{out}} j\omega R_2 C_2 + v_{\text{out}} \quad \text{Equation 4-12}$$

Use this expression in Equation 4-11.

$$\frac{v_{\text{in}} - v_{\text{out}} j\omega R_2 C_2 + v_{\text{out}}}{R_1} = v_{\text{out}} j\omega C_2 R_2 (j\omega C_2) + v_{\text{out}} j\omega C_2$$

Now gather the v_{out} terms

$$\frac{v_{\text{in}}}{R_1} = v_{\text{out}} \left(-\omega^2 C_2^2 R_2 + j\omega C_2 \left(1 + \frac{R_2}{R_1} \right) + \frac{1}{R_1} \right)$$

Then

$$\frac{v_{\text{out}}}{v_{\text{in}}} = \frac{1}{(-\omega^2 C_2^2 R_1 R_2 + j\omega C_2 (R_1 + R_2) + 1)}$$

4.6 Practical Resistor and Capacitor Values

In practical circuits, it is practical to have resistor values at least on the order of 1 k Ω . Consider that 10 volts into a 1 k Ω resistor leads to an easily achievable 10 mA current (100 mW power requirement). Ten volts into 1 Ω , however, would require 10 Watts of power, and would quickly heat up the circuit. Even a circuit that has only 10 such resistors would require 100 Watts.

On the other end, resistors larger than 10 M Ω are uncommon. If the resistance becomes too high, then oils from your hand during handling can alter the resistance.

Capacitances will typically be in the 100 pF to 10 μ F range. Capacitors smaller than 100 pF are on the order of the capacitance inherent to the bread boards that are used in prototyping, and larger capacitors are bulky.

4.7 Bandwidth

Although the simple analysis of operational amplifier circuits assumes a large frequency-independent gain, for practical amplifiers the gain decreases as frequency increases. In general, the open-loop gain A begins at a value between 10^5 and 10^6 , but decreases in proportion to $1/\omega$, beginning at a low frequency. Additional poles exist near 1 MHz. The datasheet for an OP07,

for example, shows a DC gain of about 7×10^5 and a -3 db frequency of about 2 Hz, so that the gain drops to 1 at a frequency of about 800 kHz. The decrease in open-loop gain does not strongly affect the closed loop gain until the signal frequency is increased to a value near the frequency where the closed loop gain and open-loop gain are similar. The magnitude plot is sketched in Figure 4-13. If an amplifier circuit is created from this operational amplifier with a gain of 10 (20 db) (horizontal dotted line), the closed-loop gain of the circuit will decrease for frequencies near and above 100,000 Hz.

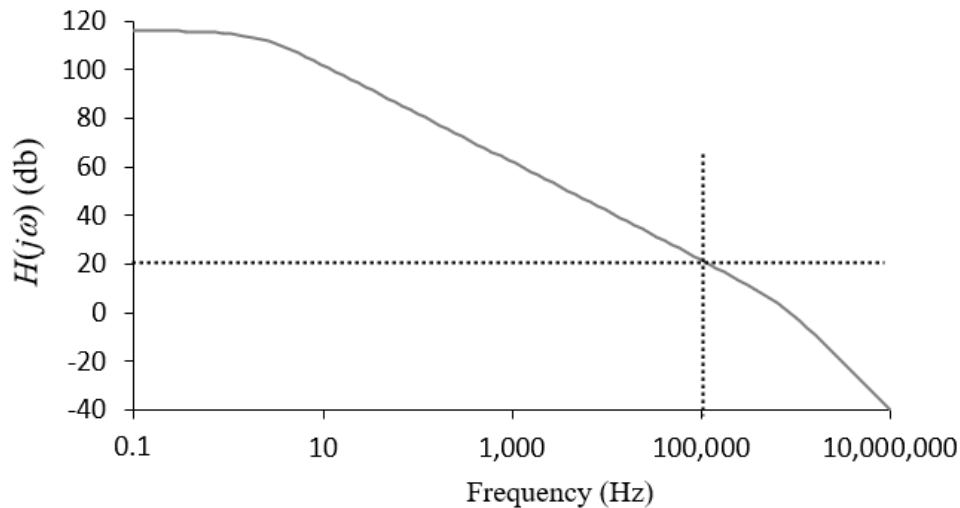


Figure 4-13: Approximate magnitude plot for the open-loop gain of an OP07.

4.8 Comparators

A comparator (Figure 4-14) resembles an operational amplifier in that it has high gain and input impedance. It is designed to be used without the feedback that linearizes an operational amplifier and to therefore provide only two output voltages, a maximum value when the positive input voltage (V_{in}) is greater than the negative input voltage (V_{ref}), and a minimum value when the opposite is true. It can serve as an interface between an analog signal at the input side and a digital signal at the output side. If it sends its output to a TTL system, the minimum voltage will be 0 volts (logic value 0) and the maximum voltage will be 5 volts (logic value 1). For analog to digital conversion, several comparators are combined to compare the analog signal to a set of fixed voltage values, and the logic value of each comparison is interpreted as the digital number.

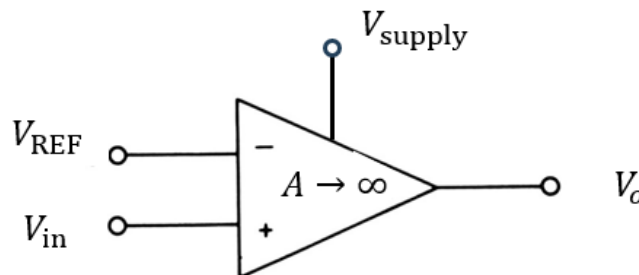


Figure 4-14: Comparator Schematic

Comparators are significantly faster than operational amplifiers. The bandwidth of operational amplifiers is generally limited (compensated) so that the circuit maintains stability in high gain applications. Comparators are intended to operate as unstable elements, operating in the nonlinear region of the device, so they do not require compensation.

A comparator is the hardware equivalent of the if statement in MATLAB.

```
if( $V_{in} > V_{ref}$ )
     $V_o = 5$ ;
else
     $V_o = 0$ ;
end
```

The similarity with an operational amplifier is such that it is possible to use an operational amplifier in place of a comparator in some applications. However, integrated circuits that are sold as comparators are optimized to perform the comparison function, with features such as a substantially larger slew rate. Many comparators also operate differently from operational amplifiers. When ($V_{in} < V_{ref}$), the output is connected to the negative to the negative supply voltage, but when $V_{in} > V_{ref}$, the output is disconnected from that supply and left floating. To ensure that the output rises to the positive supply voltage when $V_{in} > V_{ref}$, a resistor must be connected from the output to the voltage on the positive power supply. This component, typically 1 k Ω , is referred to as a pull-up resistor.

4.9 Relation to Sensors

Sensors themselves can be thought of as filters in the sense that the voltage out is a filtered version of the physical quantity in. For example, a thermocouple will not instantaneously change its output voltage when the temperature is instantaneously changed (e.g. by moving the sensor from room temperature to an ice bath). Hence the physical signal is being low pass filtered.

Thus, like operational amplifier circuits, sensors have a transfer function, an output impedance, and an analog to input impedance, in the sense that the sensor must not alter the physical quantity being measured.* Biomedical sensors must be selected such that their frequency response is appropriate for the measurement of interest. An understanding of operational amplifier filters thus relates to the ability to work with biosensors.

4.10 Software to Generate the Bode Plot

4.10.1. Excel

Students may initially be daunted when asked to use Excel to generate and plot the magnitude and phase data in Excel. However, a few simple tricks can greatly simplify the process. Consider the transfer function of Equation 4-13.

* To complete the analogy, if a circuit has low input impedance, it alters the physical quantity (voltage) at its input.

$$H(j\omega) = 42 \frac{j\omega + 100}{(j\omega + 20)(j\omega + 800)} \quad \text{Equation 4-13}$$

To obtain the magnitude and phase, first break the expression into three parts, the first being $j\omega + 100$ in the numerator, the second being $j\omega + 20$ in the denominator, and the third being $j\omega + 800$ in the denominator. To obtain the magnitude, divide the magnitude of the numerator but the product of the magnitudes of each term in the denominator.*

$$|H(j\omega)| = 42 \frac{|j\omega + 100|}{|j\omega + 20||j\omega + 800|} = \frac{\sqrt{\omega^2 + 100^2}}{\sqrt{\omega^2 + 20^2}\sqrt{\omega^2 + 800^2}} \quad \text{Equation 4-14}$$

Similarly, break Equation 4-14 into the same terms to calculate the phase.†

$$\angle H(j\omega) = \angle \left(42 \frac{|j\omega + 100|}{|j\omega + 20||j\omega + 800|} \right) = \angle 42 + \angle(j\omega + 100) - \angle(j\omega + 20) - \angle(j\omega + 800)$$

The angle of each term is the arctangent of the imaginary part divided by the real part. Since the imaginary part of 42 is zero, $\angle 42 = 0$, so any real coefficients are irrelevant to the phase calculation.

$$\angle H(j\omega) = \arctan\left(\frac{\omega}{100}\right) - \arctan\left(\frac{\omega}{20}\right) - \arctan\left(\frac{\omega}{800}\right) \quad \text{Equation 4-15}$$

You now want to create a frequency axis. Assume that the bode plot needs to span from 0.1 Hz to 100 kHz. You would like the frequency values to be logarithmically spaced from one another. Spacing them linearly would cause you to have only a few frequencies in the lower decades and a ridiculously large number of frequencies in the higher decades. I.e. if you decide to space the frequencies 0.01 Hz from one another, then you would end up with about 1 million points between 10 kHz and 100 kHz. Therefore, generate each frequency so that it is some factor larger than the previous value. If you choose 1.2 as that factor, then the first few frequencies will be approximately 0.1, 0.12, 0.144, 0.173, 0.207, 0.248, 0.299 Hz. In Excel, the axis is set up as shown in Figure 4-15.

* More generally, it can easily be shown that $\left| \frac{N_1 N_2 \cdots N_n}{D_1 D_2 \cdots D_m} \right| = \frac{|N_1| |N_2| \cdots |N_n|}{|D_1| |D_2| \cdots |D_m|}$

† More generally for phases, $\angle \left(\frac{N_1 N_2 \cdots N_n}{D_1 D_2 \cdots D_m} \right) = \angle N_1 + \angle N_2 + \cdots + \angle N_n - \angle D_1 - \angle D_2 - \cdots - \angle D_m$, where \angle is read as “the phase of.”

<div> <div>wn</div> <div>✕ ✓ \mathcal{f}_x</div> <div>100</div> </div>							
	A	B	C	D	E	F	G
1	Omega (rad/s)	Frequency (Hz)	Magnitude	Phase (degrees)			
2	=B2*2*PI()	0.1				wn=	100 rad/s
3	=B3*2*PI()	=B2*1.2				wda=	20 rad/s
4	=B4*2*PI()	=B3*1.2				wdb=	800 rad/s
5	=B5*2*PI()	=B4*1.2					
6	=B6*2*PI()	=B5*1.2					
7	=B7*2*PI()	=B6*1.2					
8	=B8*2*PI()	=B7*1.2					
9	=B9*2*PI()	=B8*1.2					
10	=B10*2*PI()	=B9*1.2					

Figure 4-15: Frequency axis for the transfer function calculation.

Figure 1 shows three additional labeled columns besides the frequency in Hz. Column A is the frequency ω in radians/s, which is convenient because the transfer function equation is written in terms of ω . Column C is where the magnitude will be calculated, and Column D is where the phase will be calculated. We *could* now use Equations 2 and 3 directly to calculate the magnitude and phase. However, for additional flexibility, the cutoff frequencies 100, 20, and 800 can be defined as variables so that, if one or more of these values changes, we will need to change only their defined values, as opposed to making changes to the equation. The practice also improves the readability of the worksheet.

Find a location on the worksheet that is not being used (and that you otherwise do not anticipate using in the future) and type in the information shown in the range \$F2:\$G4. Column G has the values that are to be defined. To define the value in cell \$G\$2 as “wn,” first click on the cell, then click on the window above Row A and type “wn”. Next click on cell \$G\$3 and type wda in the window above Column A. Do the same for Cell \$G\$4, naming it wdb.*

Now the equations can be easily calculated, as shown in Figure 4-16.

	A	B	C	D
1	Omega (rad/s)	Frequency (Hz)	Magnitude	Phase (degrees)
2	=B2*2*PI()	0.1	=42*SQRT(A2^2+wn^2)/(SQRT(A2^2+wda^2)*SQRT(A2^2+wdb^2))	=(ATAN(A2/wn)-ATAN(A2/wda)-ATAN(A2/wdb))*180/PI()
3	=B3*2*PI()	=B2*1.2	=42*SQRT(A3^2+wn^2)/(SQRT(A3^2+wda^2)*SQRT(A3^2+wdb^2))	=(ATAN(A3/wn)-ATAN(A3/wda)-ATAN(A3/wdb))*180/PI()
4	=B4*2*PI()	=B3*1.2	=42*SQRT(A4^2+wn^2)/(SQRT(A4^2+wda^2)*SQRT(A4^2+wdb^2))	=(ATAN(A4/wn)-ATAN(A4/wda)-ATAN(A4/wdb))*180/PI()
5	=B5*2*PI()	=B4*1.2	=42*SQRT(A5^2+wn^2)/(SQRT(A5^2+wda^2)*SQRT(A5^2+wdb^2))	=(ATAN(A5/wn)-ATAN(A5/wda)-ATAN(A5/wdb))*180/PI()
6	=B6*2*PI()	=B5*1.2	=42*SQRT(A6^2+wn^2)/(SQRT(A6^2+wda^2)*SQRT(A6^2+wdb^2))	=(ATAN(A6/wn)-ATAN(A6/wda)-ATAN(A6/wdb))*180/PI()
7	=B7*2*PI()	=B6*1.2	=42*SQRT(A7^2+wn^2)/(SQRT(A7^2+wda^2)*SQRT(A7^2+wdb^2))	=(ATAN(A7/wn)-ATAN(A7/wda)-ATAN(A7/wdb))*180/PI()
8	=B8*2*PI()	=B7*1.2	=42*SQRT(A8^2+wn^2)/(SQRT(A8^2+wda^2)*SQRT(A8^2+wdb^2))	=(ATAN(A8/wn)-ATAN(A8/wda)-ATAN(A8/wdb))*180/PI()

Figure 4-16: Formulas for the magnitude and phase calculations.

* These variable names are chosen to be mnemonic. The variable name wn stands for “the omega in the numerator.” Likewise, wda and wdb are “numerator omega a” and “numerator omega b.” Note that you cannot use names like w1 and w2 because Excel interprets them as cell addresses (e.g., Column W, Row 2).

The resulting Bode plot is shown in Figure 4-17.

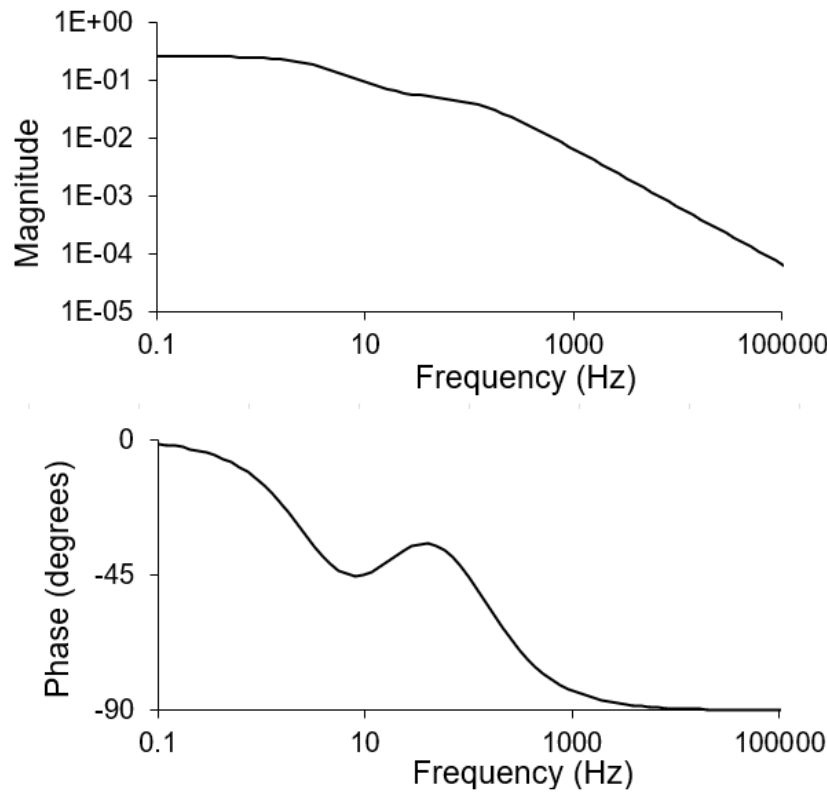


Figure 4-17: Excel-generated Bode plot

4.10.2. MATLAB

Generation of a Bode plot in MATLAB is relatively simple and can be approached in various ways. One method, shown in the code below, is to first define a frequency axis, w , that ranges from 0.1 to 100,000. First, define an array of numbers from -1 to 5 in steps of 0.2. Then raise 10 to those numbers so that the w values are evenly spaced, logarithmically. Then calculate the complex transfer function H , where $1j$ is used to indicate multiplication by $\sqrt{-1}$. Then plot the magnitude of H on one plot and the phase of H ($\text{angle}(H)$) on another.

```
pwr = -1:0.2:5;
w = 10.^pwr; % Define a frequency array from 10^-1 to 10^5.
H = 42*(1j*w+100)./((1j*w+20).*(1j*w+800)); % Calculate H for each w.
figure("name","Magnitde");
loglog(w,abs(H)); % Plot the magnitude on a log-log plot.
xlabel("Frequency (Hz)");
ylabel("Magnititude");
figure("name","Phase");
semilogx(w,angle(H)*180/pi); % Plot the phase on a semilog plot.
xlabel("Frequency (Hz)");
ylabel("Phase (degrees)");
```

The plot is shown in Figure 4-18.

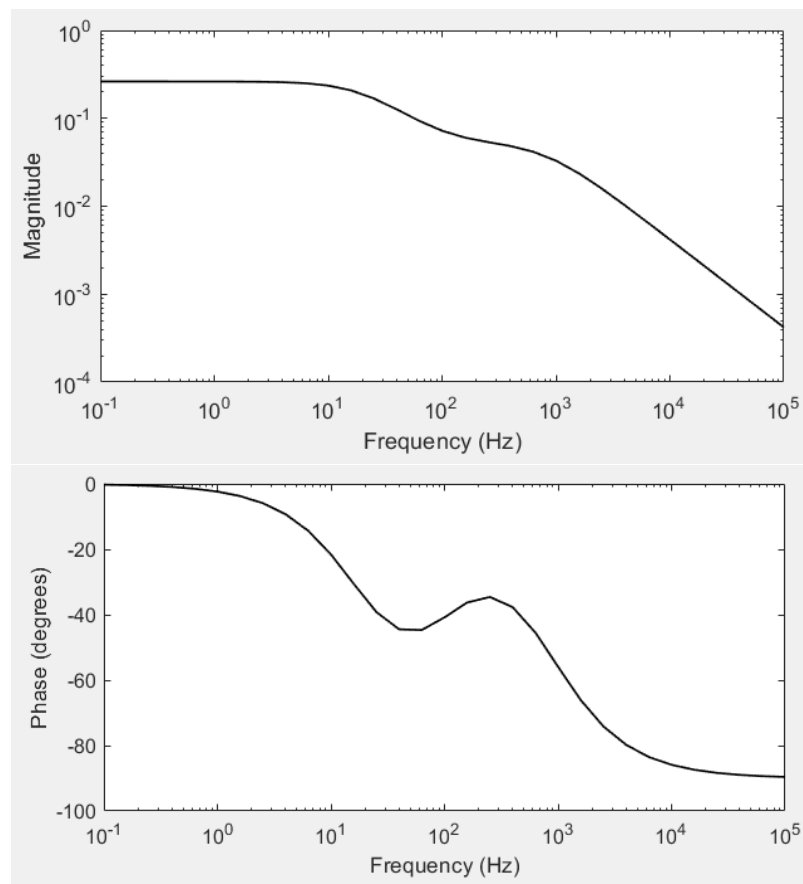


Figure 4-18: MATLAB-generated Bode plot

Alternatively, the `bode()` command can be used. The transfer function is defined from its amplitude, A , its zeros, z , and its poles, p , with the `zp2tf` command. The beginning and end of the frequency range are specified within the `bode()` command. The resulting plot is shown in Figure 4-19.

```
A=42;
z = [-100];
p = [-20, -800];
[num,den] = zp2tf(z,p,A);
bode(tf(num,den),{0.1,100000})
```

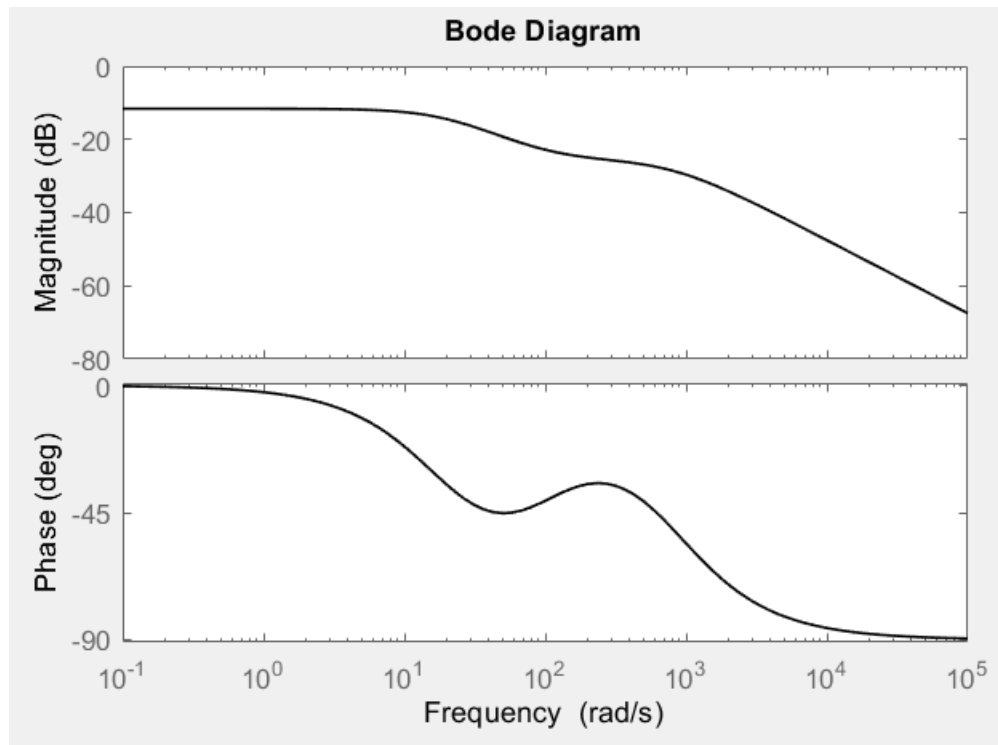


Figure 4-19: MATLAB Bode plot generated with the `bode()` command.

4.10.3. LTSpice

The simulation LTSpice can also be used to generate a Bode plot. For this software, the mathematical form of the transfer function does not need to be known because the software obtains the result through a simulation of the constructed circuit. Some instruction is needed to design circuits in LTSpice, so it is left as a laboratory exercise.

Chapter 5: Circuit Testing and Debugging

While circuit debugging can be time consuming, several practices can decrease the time considerably. These policies are described here.*

5.1 Design the system to facilitate debugging.

Use of a modular circuit design simplifies the debugging process because each stage of the circuit can be tested individually. For each stage, the output to a set of test signals should be known in advance,[†] and individual stages should be tested before they are interconnected. If possible, the system should be simulated in software before it is assembled to ensure that it functions correctly in principle and to determine what the correct voltage waveforms should be at the inputs and outputs of each module.

A consistent set of units needs to be used throughout the design. While it may be obvious that units such as volts, amps, ohms, and farads will be used in the circuitry, a circuit that interfaces with a mechanical, thermal, or chemical system may include off-the-shelf components whose specifications are communicated in less obvious units.

5.2 Use a layout that facilitates debugging.

You must have a sketch of the circuit before it is constructed, which can be either a hand sketch or a printout from simulation software such as LTSpice. To the extent possible, your physical circuit should follow the layout of the drawing so that you can easily identify where each module lies on the breadboard.

It is also important to use wire color-coding. Because ground wires are ubiquitous on circuits, you can reserve black wire, for example, explicitly for this purpose. My personal approach is to use red for the +12 volt DC supply, yellow for the −12 volt DC supply, blue for the +5 volt supply, and white for connections within circuits and green for connections between modules.

Use the bus lines on the bread board judiciously. The ground from your supply voltage should be connected to one of the bus lines on the bread board so that the individual grounds for each module can be short connections to the bus, rather than a tangle leading to some location remote from the bread board. The same advice applies to the power supplies.

5.3 Construct the system in verifiable stages.

A common mistake is to assemble the entire system with the hope that it will work correctly as soon as it is turned on. However, when it does not work correctly, it is not possible to know what needs to be changed to fix the problem. It is better, therefore, to assemble and test each module individually before they are connected together, making good use of the modular structure identified in Section 5.1. If each module works correctly and the system still does not

* These practices are generalizable to debugging in other projects besides circuitry, such as computer programming or experimental design, but will be described specifically with respect to circuits here.

[†] The set of known outputs for given inputs are referred to as positive controls.

work, it is most likely because they are not connected correctly. If multiple modules are being used, the connections between them can be tested in pairs to further locate which connection is problematic. After all modules have been connected, input a test signal into the first module and then monitor the voltages at the input and output of each of the other modules. Frequently, when the overall circuit is tested, the signal at the input of each module differs in amplitude or another aspect from that with which the module was originally tested.

In testing each module, be consistent in the use of the oscilloscope. With few exceptions, you should use Channel 1 of the oscilloscope for the test signal that is being used as an input and Channel 2 for the output. Always use the oscilloscope to verify that the input signal is correct. Do not rely on the signal generator parameters alone.

5.4 Debug Modules Systematically

In many cases, the circuit does not work properly because of a frequently occurring problem that can be easily diagnosed and fixed. These frequent problems should be sought out first.

5.4.1. Verify that the test leads are not broken.

No circuit can be adequately tested if one of the test leads has a broken connection. Before doing anything else, check that every lead has continuity. Specifically, check the voltage and ground leads from the power supply and the leads from the digital multimeter. Set the multimeter in ohms mode, connect the multimeter leads across the lead to be tested, and confirm that the resistance is on the order of one ohm.

5.4.2. Verify that the circuit components are correct.

Use the multimeter to double-check the value for each capacitor, resistor, or inductor. Read the number on each chip that you use to make sure it is correct. (Do not assume that it is an OP07, simply because it was in the OP07 drawer).

5.4.3. Double-check the pinout of each chip

Pinouts vary from chip to chip. A common error is to expect an amplifier chip to have the same pinout as an operational amplifier, such as the OP07, but differences can be subtle and important. The ina118 instrumentation amplifier, for example, has the positive input, negative input, positive supply voltage, negative supply voltage, and output on the same pins as the OP07, but requires a reference signal on Pin 5. Alternately, one might be given an LM358 to replace an OP07, without recognizing that the LM358 has a different pinout because it contains two operational amplifiers in the same eight pin package. Good practice is to always read the number on the chip, look up its datasheet, read the pinout, and check the “Typical Application” section of the datasheet to ensure that the chip is well understood.

5.4.4. Make sure that your measuring instruments function correctly.

As with the circuit, you can perform positive control tests on the measuring instruments. For example, to test the resistance function on the multimeter, use it to measure resistance across a known resistor. Then, to test the DC voltage function, measure a voltage from the power supply.

Connect the power supply directly to the oscilloscope before you connect it to the circuit. If the oscilloscope does not register a voltage, it typically indicates that (1) one (or both) of the power supply leads is bad (most likely scenario), (2) the oscilloscope cable is bad, (3) you are not using the correct ground on the power supply, or (4) the oscilloscope is set on AC coupling instead of DC coupling.

5.4.5. Make sure your measuring instruments are in the correct mode.

Often circuit will appear to output the wrong value because the multimeter is not set in the correct mode for the measurement, as when, for example, a signals AC component needs to be measured, but the multimeter is set for DC volts. Alternatively, some multimeters require the user to select the expected range of the voltage (e.g., 0 to 200 mV, 0 to 2 mV, 0 to 20 V) if the range maximum is set too high, the resolution will be poor, and if it is set too low, the multimeter will indicate an overload.

Settings on the oscilloscope require more attention. Common errors include

- i. A wrong coupling setting (AC vs. DC).
- ii. A channel set on “invert.”
- iii. An incorrect attenuation setting for the probe being used (1X, 10X).
- iv. A sweep time that is too slow for the signal. With digital oscilloscopes, the sample rate for digitization changes with the sweep time, so a sweep time that causes sample rate to be less than half the frequency of the signal will lead to aliasing.

5.4.6. Check the power supplies.

Perhaps the most common error, and one that leads to significant wasted laboratory time, is having a chip that is not connected to the power supply. The reasons for this error vary from the simplest (the power supply is not turned on or is not set to the right voltage) to the more subtle (the pin for the chip is not set correctly in the breadboard). Thus, it is not sufficient to simply check the meter on the power supply to ensure that it reads the correct voltage. Instead, the power supply pins on the chip need to be tested directly. To check them, touch the oscilloscope probe (or multimeter probe) directly to the pin, not to the wire (allegedly) connected to the pin or to the adjacent location on the breadboard. Where voltage supplies are concerned, the only voltage that matters is the one on the pin itself.

5.4.7. Check the ground connections.

Grounds need to be double-checked, especially when the circuit output seems noisier than it should be. Systematically check the ground connected to (1) Each of the microchips, (2) the multimeter, (2) the two oscilloscope inputs, (3) the power supply, and ensure that they are all connected to one another.

5.4.8. Check issues with connectivity.

If the sub-modules function correctly but the overall circuit does not, three problems are likely. The first is that the connection from one module to the next is incorrect (or that the two modules are referenced to different ground levels). The second is that a module with high output impedance is driving one with low input impedance, so that a unity buffer is required between the two. The third is that the test signal is causing one of the modules to exceed its capabilities.

For example, a module may exceed the supply voltage, the maximum current output, the frequency band, or the operational amplifier's slew rate.

5.4.9. Suspect a bad chip.

A chip should be replaced only if the other basic checks do not cause the circuit to function. IC chips do burn out, but an incorrect connection is more likely. Once you have verified that a chip is bad dispose of it so that it does not cause the same frustration for someone else. However, have a laboratory instructor confirm the state of the chip.

5.5 Troubleshooting

If all steps in Section 5.4 have been followed and the completed system does not function, further troubleshooting steps can be taken.

Recheck the outputs of the individual modules to ensure that they are still correct. It may be useful to test them with signals other than the positive control that you initially used. I.e., if the outputs are correct for the positive control, a more complicated signal may provide different results that could be instructive. If they are not correct for the positive control, a simpler test signal may be constructive. The objective is to find the simplest signal that provides incorrect results.

Change one aspect of the circuit at a time. If changing more than one aspect causes the circuit to work, you will never know which change was needed.

It is often helpful to describe the problem to a co-worker, a process that improves your own understanding of the malfunction.

In some cases, it may be helpful to rebuild the circuit from the beginning. If resources allow, leave the original circuit in place and construct a new circuit on a different breadboard (or on a different part of the breadboard) with different components. If the new circuit works, compare it to the original to decipher the original problem.

Finally, ask for help. Check other people's results to discover how they solved specific problems. Speak to co-workers, lab mates, classmates who have experience or have made more progress than you have. You can also check the internet for solutions to the specific problem.

Chapter 6: Hypothesis Testing

Experiments generally have a degree of randomness to them, especially in biomedical engineering applications, where a person's response to a treatment or prevention strategy depends on genetic makeup, environment, previous experiences, and other variables that cannot be completely controlled. Any measurement taken from the body will have some scatter (variance) within a population. With clinical trials, we may be interested in whether an intervention improves some health parameter, and that parameter may improve in some experiments and degrade in others.

Some insight can be gained from the mean measured value of the parameter of interest, but this mean value will change with each additional experiment. If enough experiments are performed, that value generally tends to converge to a stable value which we will refer to as the true mean, but if the variation is large from experiment to experiment, many experiments must be performed before that convergence occurs. Furthermore, if the difference between the true mean before intervention is small, the experimental means will not be distinguishable unless the true means are estimated closely. A fundamental question in hypothesis testing is therefore to determine whether the results of enough experiments have been collected to distinguish the two means.*

The need for hypothesis testing can be illustrated with a simple example. It is well known that the outcome of a coin toss is as likely to be heads as tails. If the coin is tossed n times, it is expected that an equal number of heads and tails will result. Clearly if $n = 1$ or $n = 3$ the number of heads and tails will not be equal. The case of $n = 4$ is interesting. The outcome of this experiment has 16 possible outcomes (TTTT, HTTT, THTT, HHTT, etc.). Of these outcomes, only 6 result in two heads and two tails. The result is thus less likely to be evenly split between heads and tails than it is to be skewed towards either heads or tails (evenly split only $100(6/16) = 37.5\%$ of the time).

6.1 Terminology of Probability

6.1.1. Distribution

A distribution maps each possible outcome of an experiment with the probability that that outcome will occur, which is usually plotted as probability as a function of outcomes. For the single coin toss, the outcomes are heads and tails. The distribution plot has a value of $\frac{1}{2}$ for the value of heads and $\frac{1}{2}$ for the value of tails. For the result of two coin tosses, the possible outcomes are HH, HT, TH, and TT, so the distribution has a value of $\frac{1}{4}$ for an outcome of two heads, $\frac{1}{2}$ for a value of one heads, one tail, and $\frac{1}{4}$ for a value of two tails. This simple example illustrates that some possible outcomes are more likely than others. Distributions can also be continuous, such as the distribution of healthy human body temperatures, which may demonstrate a peak at $97.7^\circ\text{F}^{[1]}$, but can take on a continuous set of values. For a discrete probability distribution, the sum of all probabilities must be 1, and for a continuous distribution,

* The discussion here begins with a focus of distinguishing mean values, but hypothesis tests can be performed on any statistical value (e.g., mode, variance, kurtosis).

the integral of all probabilities must be 1. That is, regardless of which outcome is more or less likely, one of the outcomes must occur.

The normal (Gaussian), Rayleigh, and uniform distributions are frequently encountered. These are illustrated in Figure 6-1. The uniform distribution cannot take a value lower than its lower limit or larger than its upper limit. The Gaussian distribution can take any value, but the probability that it will take a value farther from the mean value by about three standard deviations from its mean is extremely low. The Rayleigh distribution cannot take negative values, and is therefore relevant for parameters that must be positive, such as squared quantities. It can be shown that if x_1 and x_2 are Gaussian distributed variables, then the new variable $y = \sqrt{x_1^2 + x_2^2}$ is Rayleigh distributed, where the positive root is taken.

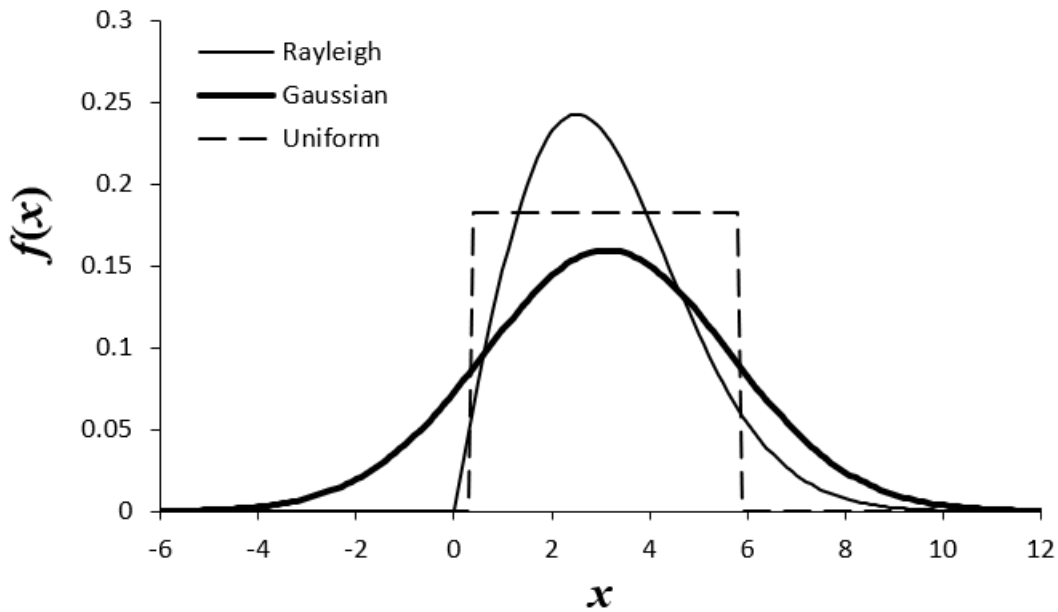


Figure 6-1: Commonly encountered distributions. Each distribution shown in this plot has the same mean (3.13) and standard deviation (2.5).

The formulas for these distributions are

$$\text{Uniform } f(x) = \begin{cases} 0 & \text{if } x < a \text{ or } x > b \\ \frac{1}{b-a} & \text{if } a \leq x \leq b \end{cases} \quad \text{Equation 6-1}$$

$$\text{Gaussian } f(x) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} \quad \text{Equation 6-2}$$

$$\text{Rayleigh } f(x) = \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}} \quad \text{Equation 6-3}$$

6.1.2. Statistic

A statistic is a numeric value derived from a data set. The mean value is a statistic, as is the standard deviation. A statistic can be a parameter derived from a probability distribution.

6.1.3. Population vs. Sample

A population is the collection of all specimens to which a given set of criteria apply. E.g., “all people who are 30 years old, blue-eyed and red-headed” represents a population based on the four criteria underlined.

A sample is a subset of the population that is selected for measurement. Generally, it is not practical to measure all of the specimens.

An underlying statistic for a population will typically differ from the statistic for the sample. $\bar{x}_{\text{sample}} \neq \bar{x}_{\text{population}}$ because \bar{x}_{sample} is based on a small collection of the specimens, and $\bar{x}_{\text{population}}$ is based on every specimen. For example, population mean for all possible coin toss experiments (number of heads divided by the number of coin tosses) is $\frac{1}{2}$. The sample mean for a given 4-toss experiment is, as discussed above, likely to be different from $\frac{1}{2}$.

6.1.4. Hypothesis

A hypothesis is the proposal of the effect, testable through scientific investigation, of one intervention, phenomenon, or variable on an outcome. Hypothesis testing begins with a null hypothesis, which is a statement that the intervention (x) has no effect on the outcome (y) and proposes an alternative hypothesis, which is a statement of how the outcome is expected to change with the intervention. Typical alternative hypotheses are (1) a change in x changes y . (2) an increase in x increases y , and (3) an increase in x decreases y .

6.2 Basic Statistics

Before the method for hypothesis testing is described, it is useful to review some of the basic statistical parameters.

6.2.1. Expected Value

The expected value of a continuous random variable x is closely related to the mean, and is defined in terms of its probability distribution, $f(x)$, as

$$E\{x\} = \int x f(x) dx .$$

For a discrete random variable, it is similarly defined as

$$E\{x\} = \sum_{i=1}^N x_i f_i(x_i)$$

where the sum is over all possible outcomes x_i . The expected value can be thought of as a weighted average of all possible outcomes, where the probability density function provides the weighting factor for each outcome.

As a simple example, the throw of a dice has six possible outcomes of equal probability, so the probability density is $1/6$ for each outcome. The expected value is

$$E\{x\} = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right) = \frac{21}{6} = 3.5.$$

This example demonstrates that, despite its name, the expected value of a random variable is not necessarily expected at all; one would be highly surprised if you rolled a dice and the outcome was 3.5. More reasonably, it represents the mean value that would be obtained from a large number of die rolls.*

6.2.2. Arithmetic Mean

True Mean

The true mean, μ , is the **expected value** of a variable, and it does not depend on how many samples of the variable have been taken. It is a property of the underlying process.

Sample Mean

The sample mean is the mean value of a variable calculated from a finite number of samples of that variable. The definition is

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i,$$

where each x_i is a sample, and the sum is over the n samples.

6.2.3. Geometric Mean

The geometric mean is defined as

$$\bar{x}_{\text{geom}} = \left(\prod_{i=1}^n x_i \right)^{\frac{1}{n}}.$$

The geometric mean is used for measurements that tend to be exponentially related to one another, such as in cell growth, where each new generation of cells depends on the number of cells in the previous generation.

6.2.4. Mode

The mode of a distribution is the outcome with the highest probability. For measured data, it is the data value that occurs most frequently or, for distributed data, it is the range of values in a histogram within which most of the data values occur.

* Not all random processes work in such a way that the mean of a large number of trials converges to the expected value, but it will be assumed to be true here. Random processes that have this property are referred to as “ergodic.”

6.2.5. Median

The median, x_m , of a distribution is the value in the distribution for which half of the area under the curve $f(x)$ lies at x values less than x_m and the other half lies at values greater than x_m . For collected data, the sample median is the value where half of the measurements are less than x_m and half are greater than x_m .

Example:

Find the median value of the Rayleigh distribution.

Solution:

The integral of the probability distribution up to the median must equal 0.5. With $f(x)$ as defined in Equation 6-3,

$$\int_0^{x_m} f(x) dx = \int_0^{x_m} \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}} dx = 0.5.$$

Use a u substitution with $u = x^2$ so that $x = \sqrt{u}$ and $du = 2x dx$.

$$\int_0^{u_m} \frac{1}{2\sigma^2} e^{-\frac{u}{2\sigma^2}} du = \frac{1}{2\sigma^2} \int_0^{u_m} e^{-\frac{u}{2\sigma^2}} du = 0.5$$

But $\int e^{-\frac{u}{2\sigma^2}} du = -2\sigma^2 e^{-\frac{u}{2\sigma^2}}$, so

$$\begin{aligned} -e^{-\frac{u}{2\sigma^2}} \Big|_0^{u_m} &= 0.5 \\ -\left(e^{-\frac{u_m}{2\sigma^2}} - 1\right) &= 0.5 \Rightarrow -e^{-\frac{u_m}{2\sigma^2}} = -0.5 \Rightarrow -\frac{u_m}{2\sigma^2} = \ln(0.5) = -0.69315 \\ u_m &= 2\sigma^2(\ln(2)) \\ x_m &= \sqrt{u_m} = \sqrt{2\ln(2)} \sigma \end{aligned}$$

6.2.6. Variance

In probability, the variance is defined as the expected value of $(x - \bar{x})^2$.

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx. \quad \text{Equation 6-4}$$

In terms of data, the variance is

$$\sigma^2 = \frac{1}{n} \sum (x_i - \mu)^2. \quad \text{Equation 6-5}$$

The relationship between these two equations is clearer when one realizes that $f(x)$ describes all possible values that x can take on and the frequency with which each value is expected to occur. Therefore, if enough values are obtained in the time series, they should do so with the

frequencies dictated by the probability density function, so that averaging over time is equivalent to averaging over probabilities.

Use of Equation 6-5 assumes that the true mean, μ , is known. In practice, the mean must be estimated from the data (sample mean), which leads to an underestimate of σ^2 . The underestimate is corrected if the coefficient $1/n$ is replaced with $1/(n - 1)$. The sample variance is calculated from

$$s^2 = \frac{1}{n - 1} \sum (x_i - \bar{x})^2.$$

The relationship between this definition

6.2.1. Standard Deviation

Standard deviation is defined as the square root of the variance. The parameter is more intuitive than variance because it has the same dimensions as the variable itself, and it can be quickly (though roughly) estimated from a glance at the probability distribution.

6.2.2. Least Squares Parameters

The parameters obtained from a linear regression (slope and intercept) are themselves statistics. More generally, parameters from any regression model are statistics.

6.3 The Student's T Test

6.3.1. The t Statistic

The Student's t test* is used to determine whether, given the measured difference in means between two data sets, along with the standard deviation and the number of data points collected, the difference in the calculated means is likely to have resulted from an underlying difference in the experiment under study (i.e., a difference in the means of the probability distributions that describe the two cases). The evidence for an underlying trend becomes stronger with a larger ratio of the difference in the means divided by the standard deviation and with a greater amount of collected data. Therefore, the t test examines a statistic of the form

$$t = \sqrt{\tilde{n}} \frac{\bar{x}_1 - \bar{x}_2}{\tilde{\sigma}},$$

* The t test was developed by William Sealy Gosset in 1908. He was a brewer at Arthur Guinness Son & Co, Ltd in Dublin, Ireland and was interested in sampling for quality control. To protect intellectual property, Guinness did not allow employees to publish papers, so Gosset published under the pseudonym "Student." Consequently, the word "Student" is capitalized as a proper name.

According to the American Psychological Association (APA) style format, t test (and others, e.g. F test) is not hyphenated when it is used as a noun but hyphenated when used as an adjective.

"We performed a t test." (Not hyphenated)

"The t -test results were statistically significant." (Hyphenated)

where \bar{x}_1 and \bar{x}_2 are the calculated means of the two samples. Here, the tilde over the variables is shorthand for “a measure of.” Specific measures of n and σ will be defined according to the type of t test applied (equal variance, unequal variance, or paired).

6.3.2. The t Statistic Probability Distribution

The probability for the t statistic is

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

Where ν is the number of degrees of freedom for the statistical test and Γ is the gamma function.* This probability distribution is shown in Figure 6-2. These curves show that t is most likely to occur in a range between about -2 and $+2$ if the data are related only by random chance.† A t value of 8, for example, would be highly unlikely and would lead to the conclusion that the difference in the means for the two data sets were was not a consequence of random chance (i.e., that the two means were fundamentally different).

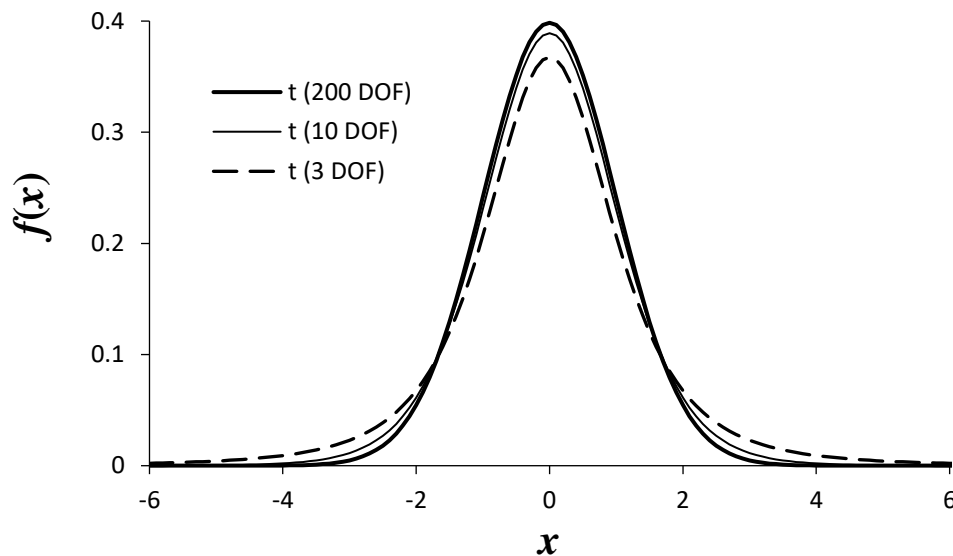


Figure 6-2: Probability distribution for the t statistic for 200, 10, and 3 degrees of freedom (ν).

The t test implicitly assumes that the data x_i are Gaussian distributed. Once the t value is calculated, it is compared to a t distribution to determine the probability that the calculated Δx would be obtained from two samples that have identical underlying mean values. This

* The gamma function is defined as $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x} dx$. For integral values of z , this integral reduces to z factorial, so the gamma function is considered to be the extension of the factorial operator for non-integer arguments.

† More specifically, for 10 degrees of freedom, the total probability (area under the $f(x)$ curve) between values of -2.2 and $+2.2$ is about 0.952.

probability is referred to as the p value. In medicine, one is frequently interested in interventions that improve an aspect of health. A low p value indicates low probability that the difference between the mean value for the intervention and non-intervention cases is random, hence that the intervention was the likely cause of the change. (If you formulate a hypothesis that, “Drug A will lower glucose levels in Type 1 diabetic patients,” then the p value can be considered as the probability that you are wrong.)

6.3.3. Degrees of Freedom

The number of degrees of freedom is equal to the number of independent pieces of information minus the number of restrictions on that information. For the t test (assuming equal variances), with n measurements in one category being measured and m measurements in the other, $n + m$ measurements (pieces of information) are available, but in each category a mean (restriction) must be calculated. Therefore, the number of degrees of freedom is $n + m - 2$.

6.3.4. Equal Variances

The estimated values in the t statistic depend on whether the underlying variances for the two samples are identical or different. If they are identical, a t test “assuming equal variances” is used, where the t statistic is

$$t = \frac{\sqrt{\frac{n_1 n_2}{n_1 + n_2}} (\bar{x}_1 - \bar{x}_2)}{s_{\text{pooled}}},$$

where s_{pooled} is a weighted average of the sample standard deviations s_1 and s_2 of the two data sets.

$$s_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

6.3.5. Unequal Variances

If the two data sets have different variances, the number of degrees of freedom will be decreased. If the number of data points collected is not equal for the two data sets, the t statistic will also change. These formulas will not be given here, but they are coded into statistical analysis programs. (The number of degrees of freedom can be a non-integer.) The t test is then referred to as Welch’s t test.

In many cases, the p values for equal and unequal variance are similar. One technique that can be used to decide which test to use is to perform an F test to compare the variances of the two data sets and use Welch’s test if the difference is statistically different. That method is discouraged, but it does provide a basis for the decision. The preferred method is to decide whether to use the equal variances or unequal variances test before the data are collected, based on the experimenter’s understanding of the underlying problem. Another option is to always use the Welch’s test, which will always lead to a larger p value and reduce the number of false positive results.

6.3.6. Paired test

The paired t test is used when some aspect of each item in the first data set relates it specifically to an item in the second data set. For example, to determine whether a 6 month exercise routine led to a decrease in blood pressure, one could use an unpaired test, in which the average blood pressure of all participants prior to the routine was compared to the average blood pressure at the end of the 6th months, or one could use a paired test, in which each individual's blood pressure at the beginning was compared to that individual's blood pressure at the end. The paired test is more likely to lead to statistical significance because the variance for each individual is expected to be much less than that for the group. The paired test would not be an option if the experiment were performed differently in that one group of subjects was used as the non-exercise pool and another was used as the exercise pool.

6.3.7. One-sample t test

A t test can be performed on one set of data to determine if the mean of that data set is different from (or greater than, or less than) a constant value. If a teacher wanted to know if a group of students was simply guessing on a true or false test, the one-sample test could be run to compare the set of test scores to the expected value for guessing of 50%. The t statistic is

$$t = \frac{\bar{x} - x_0}{s/\sqrt{n}},$$

where s is the sample standard deviation and n is the number of sample points. The number of degrees of freedom is $n - 1$.

6.4 F Test

The F test is used to determine whether the standard deviations obtained from two data sets are significantly different. The F statistic is

$$F = \frac{\sigma_2^2}{\sigma_1^2},$$

where σ_2^2 is the larger of the two variances. The F probability distribution is a function of F and of (separately) the number of degrees of freedom (v_2) of σ_2^2 and the number of degrees of freedom (v_1) of σ_1^2 , according to*

$$f(F, v_1, v_2) = \frac{1}{x B\left(\frac{v_1}{2}, \frac{v_2}{2}\right)} \left(\frac{(v_1 F)^{v_1} v_2^{v_2}}{(v_1 F + v_2)^{v_1 + v_2}} \right)^{\frac{1}{2}}$$

Where $B\left(\frac{v_1}{2}, \frac{v_2}{2}\right)$ is the beta function. Because this probability function depends on both v_1 and v_2 , both values are needed. v_1 is the number of samples used to calculate σ_1^2 minus 1, and v_2 is the number of samples used to calculate σ_2^2 .

* Have no fear. You will likely never need to perform this calculation.

6.5 Bonferroni Correction

A p value less than the stated α threshold does not constitute a proof that the difference in means between two data sets reflects an underlying difference in the processes. For example, if $p < 0.05$, it means that one would expect to find a difference in the mean values from two normally distributed random processes with the same expected value only 5% of the time. For an illustration, use Excel to generate 1000 pairs of normally distributed random numbers* with the same mean and standard deviation in Columns A and B. Break these pairs into 100 groups of 10 pairs each. Because the 1000 numbers in Column A were generated with the same probability distribution as the 1000 numbers in Column B, it is known that the difference in the true means between the two columns must be zero. However, if you run 100 t tests, with each test applied to the 10 pairs in each group, on average you should find five p values less than 0.05, indicating significance where none exists; the t test will indicate a false positive 5% of the time.

Consider a less-than-scrupulous company that decides to test a supplement for 20 possible effects on the body (effects such as lowered blood pressure, cholesterol, weight, or glucose, improved vision or hearing). Based on the previous paragraph, one of the 20 effects is almost certain to show significance that does not result from the supplement, but rather is the result of random chance. The Bonferroni correction readjusts the α value in cases such as this where multiple comparisons are being made. The α value is divided by the number of comparisons. For example, if four comparisons are being made and it is desired to be certain at the 5% level that an effect is the result of the intervention under study, α is set to $0.05/4$, or 0.0125.†

6.6 Linear Regression

Much of biostatistics reduces to finding relationships between data sets. With the t test and F test, the data sets are modeled as collections with a given mean and standard deviation. With the linear regression, the interrelationship between two variables is examined and modeled with a slope and intercept. Typically, the slope is the parameter of most interest; one would like to know if a change in one variable translates to a change in the other. For example, one would like to know if an increase in the dose of a blood pressure medicine corresponds to a further reduction of blood pressure. The linear regression method is briefly reviewed here, followed by a discussion of the statistics derived from it and the associated hypothesis tests.

6.6.1. Derivation

Assume y is linearly related to x . The least-squares regression line minimizes the sum of squared distances of the sample points (x_i, y_i) from the line (Figure 6-3).

* While Excel has the RAND() function to generate uniform random numbers, it does not have an analogous function for normal random numbers, but these can be generated either through the Data Analysis add-in or through the command NORMINV(RAND(),xmean,xdev). This command generates a uniformly distributed random number with RAND() and then maps it to a normally distributed random number with mean xmean and standard deviation xdev.

† This rule is simplified, although it works relatively well for small numbers of comparisons. A more accurate rule is to $(1 - (1 - \alpha)^n)/\alpha$, where n is the number of tests being made. That rule, for example, leads to a correction factor of 8 for 10 comparisons rather than a factor of 10.

$$\sum_{i=1}^n d_i^2 = \sum_{i=1}^n (y_i - (mx_i + b))^2 \quad \text{Equation 6-6}$$

where y_i is the value of y measured at x_i and \hat{y}_i is the value of y estimated from the value x_i through the linear regression line. The square ensures that each term is positive and it weighs distances farther from the line more than distances near the line.

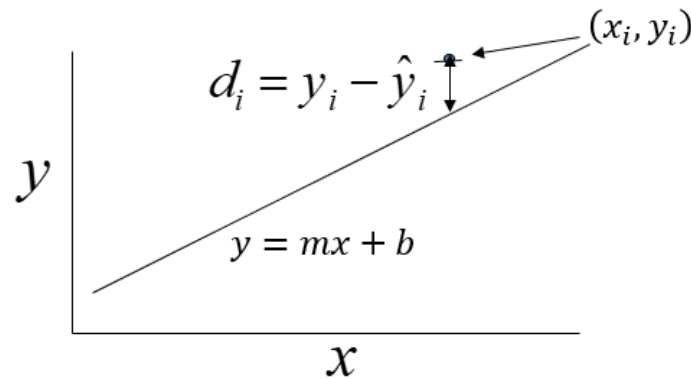


Figure 6-3: Linear regression for y as a function of x , where d_i is the distance of the data point at x_i from the value of the line at x_i .

To find m and b , minimize this summation with respect to these two parameters. The partial derivative with respect to m is

$$\begin{aligned} \frac{\partial}{\partial m} \left(\sum_{i=1}^n (y_i - (mx_i + b))^2 \right) &= 0 \\ \sum_{i=1}^n 2(y_i - mx_i - b)(-x_i) &= 0 \\ \sum_{i=1}^n -2y_i x_i + \sum_{i=1}^n 2mx_i^2 + \sum_{i=1}^n 2bx_i &= 0 \end{aligned}$$

m and b are constant, so they can be taken out of the summations, and we can divide by 2.

$$-\sum_{i=1}^n y_i x_i + m \sum_{i=1}^n x_i^2 + b \sum_{i=1}^n x_i = 0$$

Since $\frac{1}{n} \sum_{i=1}^n x_i$ is the average of the x values (i.e. \bar{x}), the last term is $bn\bar{x}$. The other summations can easily be calculated from the data values (x_i, y_i) .

Similarly, the partial derivative with respect to b is

$$\frac{\partial}{\partial b} \left(\sum_{i=1}^n (y_i - (mx_i + b))^2 \right) = 0$$

$$\sum_{i=1}^n 2(y_i - mx_i - b)(-1) = 0$$

$$\sum_{i=1}^n -2y_i + \sum_{i=1}^n 2mx_i + \sum_{i=1}^n 2b = 0$$

Again, m and b are constant and we can divide by 2.

$$-\sum_{i=1}^n y_i + m \sum_{i=1}^n x_i + \sum_{i=1}^n b = 0$$

Since b is constant, $\sum_{i=1}^n b = nb$. Also, $\sum_{i=1}^n x_i$ is again $n\bar{x}$, and $\sum_{i=1}^n y_i = n\bar{y}$.

$$-n\bar{y} + mn\bar{x} + nb = 0$$

Solve this equation simultaneously with the equation on the previous slide or the two unknowns m and b . With some algebra, these equations reduce to

$$m = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$b = \bar{y} - m\bar{x}$$

6.6.2. Correlation Coefficient

The correlation coefficient, r , is measure of the variation of the data from this linear model is

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

If $r = \pm 1$, the fit is perfect. If $r = 0$, the two variables are uncorrelated. This statistic can also be written as

$$r = \sqrt{m_x m_y},$$

where m_y is the slope obtained by fitting y as a function of x (as above) and m_x is the slope obtained by fitting x as a function of y . Typically, one reports the value of r^2 rather than r .

6.6.3. Higher Order Regression

The same method can be used to derive equations for higher order polynomial fit models. An n^{th} order fit, for example, involves n coefficients. The equation for the squared difference between the data set and the model is obtained, in the manner of Equation 6-6, the partial derivatives with respect to each coefficient are taken, and the equations are again solved simultaneously. In principle, the same technique can be applied to derive regression equations for any other parametric equation.

6.6.4. Multi-Variable Regression

It is frequently valuable to find the influence of multiple independent variables, x_i on a dependent variable, y . For example, cholesterol levels might be modeled as a function of diet, exercise, and a genetic marker. The model equation is

$$\hat{y} = a_1x_1 + a_2x_2 + a_3x_3 + \cdots + a_nx_n + b$$

The derivation again involves formation of the equivalent sum of squared distances (Equation 6-6), the partial derivatives with respect the a and b coefficients, and simultaneous solution of the resulting $n + 1$ equations.

6.6.5. p Value

Linear regression leads to the t statistic

$$t = r \sqrt{\frac{n-2}{1-r^2}}.$$

Where n is the number of data pairs, (x_i, y_i) in the regression. The number of degrees of freedom is $n - 2$ because two parameters, slope and intercept, are used in the model. From the t statistic, a p value can be derived to determine the probability that the slope of the regression line is non-zero, i.e., that the y variable changes with the x variable so that the slope is not a consequence of random chance.

6.7 Binary Decisions

Diagnostic tests are not perfect, but they must be used to make critical decisions about an individual's health. Those decisions can include whether to perform an intervention, such as a drug, surgery, or physical therapy, or simply to undergo an additional test. The failure to perform surgery to remove a malignant tumor, for example, can have severe and deadly consequences, but the performance of a surgery that is not needed can be equally dangerous. As a result, tests must be evaluated for their performance in terms of truth tables, which are 2 by 2 matrices (Figure 6-4). The left column represents all cases where the patient had the given pathology (labeled as "Pathology") and the right column represents all cases where the patient did not have the given pathology ("No Pathology"). The top row indicates all cases where the test indicated the pathology (Positive), and the bottom row indicates all cases where the test did not indicate the pathology. For the case shown, 675 patients had the pathology and tested as positive, 59 had the pathology and tested as negative, 102 did not have the pathology and tested as positive, and 203 did not have the pathology and tested as negative. Thus, the test was correct in $675 + 203$ cases and incorrect in $102 + 59$ cases.

	Pathology	No Pathology
Positive Test	675	102
Negative Test	59	203

Figure 6-4: A truth table for a diagnosis.

In a case where the test correctly identifies the pathology, it is referred to as a true positive (TP), and where it correctly identifies the absence of the pathology, it is referred to as a true negative (TN). Similarly, a negative test in the presence of the pathology is a false negative (FN), and a positive test in the absence of the pathology is a false positive (FP). It is more important to identify as many cases as possible if the pathology is serious and the consequences of treatment are relatively benign, whereas it is more important to identify the absence of the pathology if its consequences are less serious and the treatment is risky. Here we present the terminology that describes the various ways to evaluate the effectiveness of the test.

6.7.1. Accuracy

Accuracy is the number of true positives plus true negatives divided by the total number of tests performed.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

For the case shown in Figure 6-4

$$\text{Accuracy} = \frac{675 + 203}{675 + 203 + 102 + 59} = 0.845, \text{ or } 84.5\% .$$

The test is then incorrect a little over 15% of the time.

6.7.2. Prevalence

Prevalence describes how frequently the pathology arises in the population under study. It is the sum of the first column of the truth table divided by the total number of cases tested.

$$\text{Accuracy} = \frac{\text{TP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

For the case shown in Figure 6-4

$$\text{Accuracy} = \frac{675 + 59}{675 + 203 + 102 + 59} = 0.706, \text{ or } 70.6\% .$$

6.7.3. Sensitivity

The sensitivity, or true positive fraction (TPF) is

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP} + \text{FN})}$$

For the case shown in Figure 6-4

$$\text{Sensitivity} = \frac{675}{675 + 59} = 0.869, \text{ or } 86.9\% .$$

Sensitivity reflects the ability of the test to find positive cases, but it does not consider false negatives. A test can have good sensitivity and still be a horrible test. For example, a “test” where the measure is that “everyone who has ever breathed has diabetes” would have 100% sensitivity for diabetes (all people with diabetes would be identified), but would not be of any practical use.

6.7.4. Specificity

The specificity, or true negative fraction, is

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

For the case shown in Figure 6-4

$$\text{Specificity} = \frac{203}{102 + 203} = 0.775, \text{ or } 77.5\%.$$

This statistic quantifies the ability of the test to correctly identify negative cases, regardless of whether the test can identify true cases.

6.7.5. Dependence of Accuracy on Prevalence

Accuracy is a property not only of the capability of the test, but also of the prevalence of the pathology within the population. Consider the two cases of Figure 6-5, where 2003 tests have been performed. The test correctly identifies 2 out of three cases and provides a false negative one time out of 9. In the case on the left, the prevalence of the pathology is about 0.15%. The accuracy is about 10.1%. In the case on the right, the prevalence is 15.1%, and the accuracy is 18.6%, nearly double the other case.

	Pathology	No Pathology		Pathology	No Pathology
Positive Test	2	1800	Positive Test	200	1600
Negative Test	1	200	Negative Test	100	178

Figure 6-5: A truth table for a diagnosis.

6.7.6. Receiver-Operator Curves

The data upon which a binary decision is made generally takes on a range of values, where a given threshold of that value is used to decide whether the pathology is present or requires an intervention. The value could be, for example, the size of a spot on an image, or a measure of its shape (irregularity), or it could be the level of a compound in a blood test. While a yes or no evaluation must be made, tests on some patients will produce a value above the threshold when no pathology exists, while others will produce a value below the threshold when the pathology does exist. The threshold value must therefore be chosen to minimize these errors. The problem is illustrated in Figure 6-6. If the threshold (vertical dashed line) is taken as in the figure on the left (measured value 4.6), the false positives (region shaded with vertical lines) are approximately equal to the false negatives (region shaded with diagonal lines). If a lower threshold is taken, as in the figure on the right (measured value 3.9), fewer false negatives occur, but more false positives occur. Similarly, if the threshold is increased, less false positives occur at the expense of an increase in the number of false negatives. The choice of threshold value depends on the relative consequences of false positives versus false negatives.

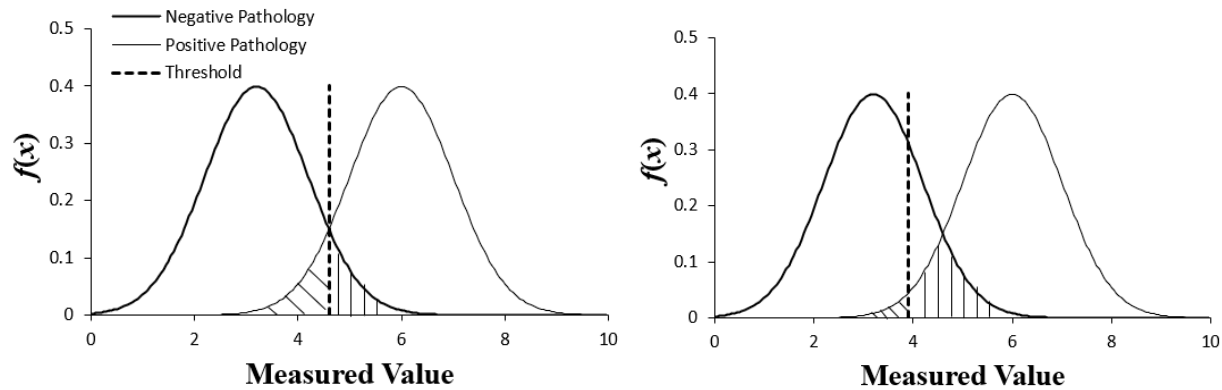


Figure 6-6: Probability distribution of true positives and true negatives as a function of the measured value.

A receiver-operator curve (ROC) is used to assist in threshold selection. It is a plot that compares the sensitivity to the specificity for different values of the threshold. Specifically, the vertical axis is sensitivity, and the horizontal axis is $1 - \text{specificity}$. Examples are shown in Figure 6-7. The dashed line indicates that the test has no predictive power and is no better than a random guess. The upper curve (thick black line) is the most predictive, in that it has large values of sensitivity with large values of specificity (low values of $1 - \text{sensitivity}$). An optimal threshold is one that maximizes sensitivity and specificity, which for this curve is at a value of $(1 - \text{specificity})$ of about 0.25 and a value of sensitivity of about 0.92. The thinner black curve is also predictive, but less so than the thick black curve. The gray curve is a case where the test is worse than a random guess, leading to an incorrect answer more frequently than a correct answer. Such a test could be made predictive, however, if one were to always select the opposite decision from its result.*

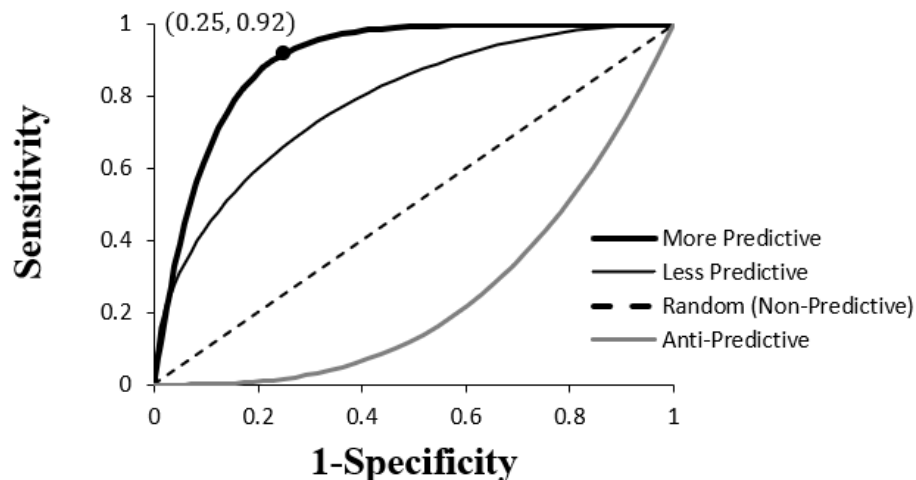


Figure 6-7: Receiver-operator curves.

* This strategy is frequently followed by Rocky and Bullwinkle in seeking the advice of Captain Peter “Wrongway” Peachfuzz.

6.8 Statistics in Excel

Although Excel is not specifically designed for statistical analysis (in contrast to programs like R, JMP, SPSS, or Minitab), it nevertheless has features that facilitate basic hypothesis testing. These will not be detailed here, but if a test is needed, the student can perform a web search to identify the function name and syntax used by Excel. Two basic examples are T.TEST() and F.TEST().

For a more visual option, the Excel data analysis add-in can be used. This add-in needs to be added to the command ribbon, as it does not appear there by default. From Excel, select “File,” then “Options,” then “Add-ins.” At the bottom of the resulting pop-up menu is a window labeled “Manage.” Select “Excel Add-ins” and left-click on “Go...” Then select the analysis tool pack and left-click on “OK.”

The user should be careful with the F test in Excel. If this test is run through the data analysis add-in, the returned p value is that for a one-sided test, where the alternative hypothesis specifies one standard deviation that is larger than the other. If the test is run through the F.TEST() command, the returned p value is that for a two-sided test, where the alternative hypothesis states only that the two standard deviations differ. The p value from the data analysis add-in is labeled “one tail,” but the p value from F.TEST() is unlabeled.

Chapter 7: Electrical Safety

7.1 General Hospital Safety

While hospitals are generally thought of as places of healing, they can nevertheless be a dangerous environment. Medical procedures expose the patient to risks that exceed what would be encountered in the home (Table 7-1). Heflin et al.,[2] for example, reported 10,395 device-related adverse events in 2004. These events are generally caused by improper use of a device. Consequently, the designer must think not only about the quality and safety of the device itself, but also about modes in which it might be improperly used. Hospital safety standards include International Electrotechnical Commission (IEC) 601-1 standard, Underwriters Laboratories (UL) Standard, and ANSI/AAMI ES1-1993 standards.

Table 7-1: Hazards encountered in the hospital.

Fire	Waste	Ultrasound
Harsh chemicals	Loud noise	Magnets
Drugs	Electrical shock	Ultraviolet light
Microorganisms	X-rays	Microwaves
Lasers		

7.2 Effects of Electricity on the Body

The effect of electricity on the body is typically characterized by the amount of current that is delivered and the sensitivity of the organ(s) to which it is delivered. (The heart, for example, is particularly sensitive to electrical shock.) Because current is given by $i = V/R$, if the entry point of the electrical signal has high resistance (R), the current is low, and because the power delivered is iV , the smaller the current at a given voltage, the smaller the amount of electrical power delivered to the body. The skin generally has high resistance, so a voltage applied to its surface will have a significantly smaller effect than a voltage applied under the skin, through a catheter, for example.

7.2.1. Description of the Effects

The effects of a given amount of current varies substantially from person to person. However, general current ranges can be identified within which specific effects occur. Potential effects are:

Perception: The subject feels the electrical current.

Tetany: The current causes muscles to contract continuously.

Fatigue: The current exhausts the muscle of its ability to function.

Pain: The current is sensed as pain.

Respiratory Paralysis: The subject loses the ability to breathe.

Ventricular Fibrillation: The heart contracts repeatedly and rapidly in an uncoordinated manner.

Sustained heart contraction: The heart undergoes a constant contraction.

Burns: Tissue is damaged by the heat generated by the current (power).

If a subject grabs onto a conduit and the current is large enough to cause tetany of the hand muscles, the subject will not be able to let go of the conduit. The current above which this condition occurs is referred to as the **let go current**.

The current ranges at which these effects occur for a 60 Hz signal are shown in Figure 7-1.

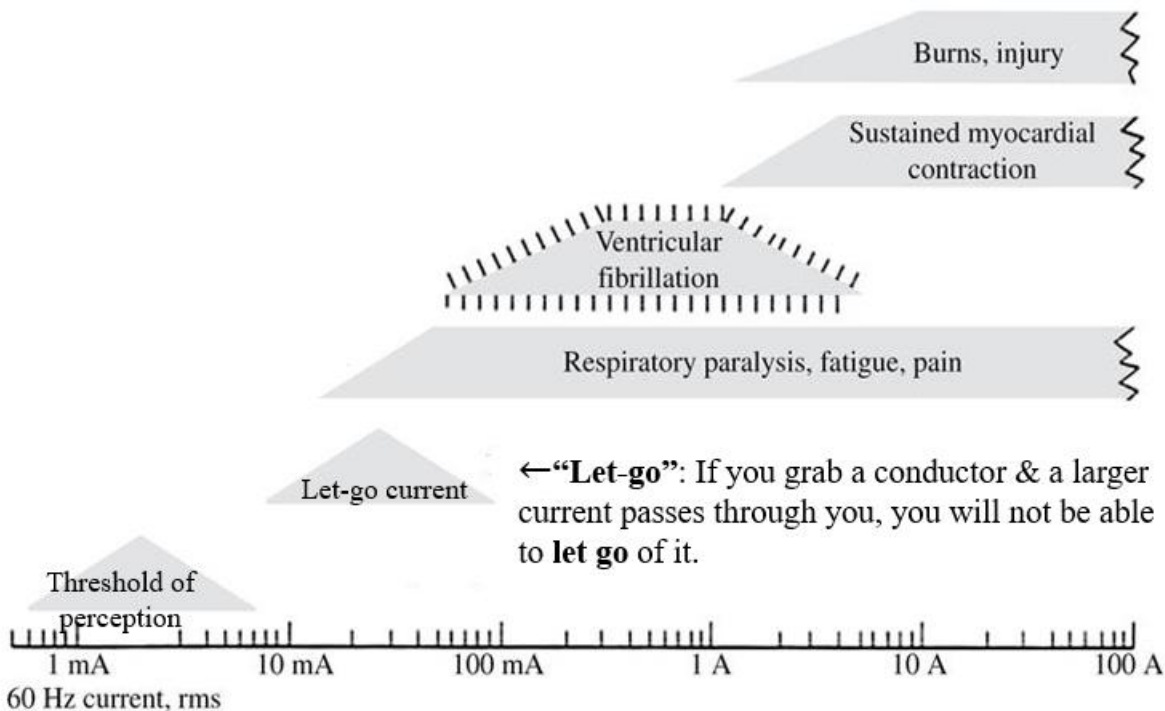


Figure 7-1: Ranges at which physiological effects occur for a 60 Hz current.

7.2.2. Dependence on Frequency (AC vs. DC)

The effect of a given amount of current depends strongly on the frequency. The largest effect occurs for an AC signal, where, for example, the let-go current varies by about 10% between frequencies of 10 and 200 Hz. This current varies for different subjects, but occurs in the range of 10 to 22 mA for most people. At 5 Hz, the upper range increases to about 40 mA. For a DC signal, the upper range is closer to 90 mA.

7.2.3. Dependence on Duration

The amount of damage caused by an electrical shock depends on the duration over which it is applied. A short duration shock may be painful, but not damaging. A shock on the order of 1 second can be fatal if 100 mA (the ventricular fibrillation threshold) reaches the heart. However, if the heart is not damaged, cessation of the current can lead to a spontaneous resumption of a normal cardiac rhythm. If the let-go current is exceeded such that the subject holds onto the current source for more than 5 seconds, death can ensue as a result of asphyxiation or heart failure.

7.2.4. Dependence on Sex

Studies indicate that women have a lower sensory threshold for electrical current than men and a lower threshold of pain [3]. Dalziel also found a lower let-go current for women than for men, with a median value of 11.4 for women and a median value of 16 for men [4].

7.2.5. Dependence on Point of Entry

The amount of current delivered to the body depends on the resistance between the body and the current source. Furthermore, because the heart is a vulnerable electroactive organ, the amount of current that passes through it largely determines the risk to the body. A relatively small voltage (microshock) applied directly to the heart through the blood stream, for example with a catheter, produces a large local current because the impedance between the source and the heart is small and because all of the current produced passes through the heart. The widely accepted safety limit to prevent microshocks is 10 mA. A larger voltage (macroshock) applied to dry skin is subject to an impedance on the order of 15 k Ω to 1 M Ω , and much of the resulting smaller current passes through the body around the heart.

Microshock usually results from leakage currents in equipment or differences in voltage between the grounded conductive surfaces due to large currents in grounding systems. Grounded equipment casings reduce macroshock. Properly grounded chassis reduce the risk of macroshock even when a ground fault from a hot line to the equipment case is created.

For an externally applied macroshock to cause fibrillation, a relatively large voltage and current is required. The locations of the two macroshock entry points (voltage and ground) also influences the danger; if the two points are on the same extremity, the risk of fibrillation is small, even for high currents because most of the current bypasses the heart. As an example, for dogs, the current needed for fibrillation is greater for ECG lead I (LA–RA) electrodes than for ECG leads II and III (LL–RA and LL–LA).

Conversely, for defibrillation, where a shock is imposed on the heart to synchronize and reset the muscular components during ventricular fibrillation, the two paddles are placed such that the heart is between them, with one to the left side of the chest and the other toward the right side.

7.3 Standard Electrical Power Supplies

Several steps are involved in the transfer of electricity from the power plant to the standard electrical socket. The power arrives at the building as a high AC voltage of 4800 volts RMS. A step-down transformer converts this voltage to 240 volts RMS. This transformer has a grounded center tap, so that it can supply two 110 volt RMS outlets where the voltages are taken between the two sides of the transformer and the ground tap (**Figure 7-2**). It may initially appear that the leftmost prong of the electrical socket (the neutral prong) is redundant to the third prong because they are attached to ground at the same location. However, when the device being powered uses a large amount of current, a voltage drop can occur between the neutral prong and the ground attachment location because the neutral wire (color-coded white) has a small but finite resistance. Thus, the neutral wire and the ground wire can have different voltages. Since the upper socket in a given receptacle is connected to the same wires as the lower socket, that voltage difference depends on the amount of current being provided by both sockets.

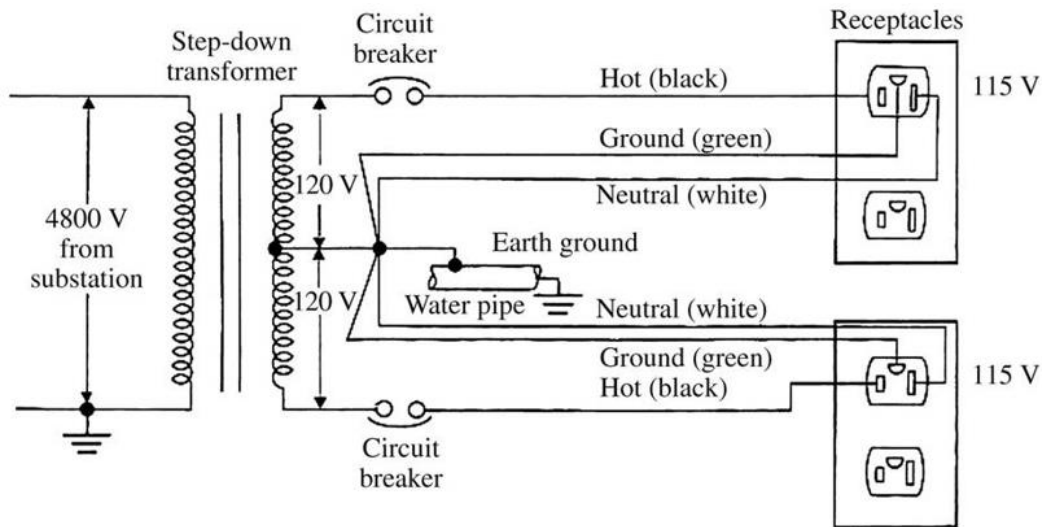


Figure 7-2: Conversion of the electrical power from the substation to the electrical socket.

Because the neutral wire on a 120 V circuit is connected to ground, a connection between the hot conductor and any grounded object in the patient's environment poses a macroshock hazard.

The third prong on the electrical socket provides additional protection for the patient in rare cases when the power supply for a device that is connected to the patient becomes shorted to the device's chassis. With the assumption that the patient is grounded, the current from the power supply has a low impedance pathway through the patient to ground. When the chassis is connected to the third (ground) prong, however, the pathway provided by this connection to ground is much less than the pathway through the patient, so that the current through the patient is limited. Additional protection is provided by the circuit breaker on the circuit, which detects the resulting unusually large current and disconnects the 110 volt RMS voltage source from the hot (black) prongs on the associated receptacles.

7.4 Hospital Requirements

Maximal potentials permitted between any two exposed conductive surfaces in the vicinity of the patient are specified by the 1996 NEC, Article 517-15. In general-care areas, voltage differences cannot exceed 500 mV under normal operation. In critical-care areas, voltage differences cannot exceed 40 mV under normal operation. Also, in critical-care areas, all exposed conductive surfaces in the vicinity of the patient must be grounded at a single patient-grounding point.

7.5 Medical Device Regulation by the FDA

The Food and Drug Administration (FDA) formulates and enforces regulations related to food, drugs, medical devices, radiation-emitting products, vaccines, blood, biologics, animal and veterinary medicines, cosmetics, and tobacco products. Medical devices must be approved by the FDA before they can be marketed.

7.5.1. Definition of a Medical Device

The FDA has a specific definition of a medical device (according to Section 201(h)(1) of the Food, Drug, and Cosmetic Act). A device is

An instrument, apparatus, implement, machine, contrivance, implant, in vitro reagent, or other similar or related article, including a component part, or accessory which is:

- *(A) recognized in the official National Formulary, or the United States Pharmacopoeia, or any supplement to them,*
- *(B) intended for use in the diagnosis of disease or other conditions, or in the cure, mitigation, treatment, or prevention of disease, in man or other animals, or*
- *(C) intended to affect the structure or any function of the body of man or other animals, and which does not achieve its primary intended purposes through chemical action within or on the body of man or other animals and which is not dependent upon being metabolized for the achievement of its primary intended purposes. The term "device" does not include software functions excluded pursuant to section 520(o).*

Software can be considered a device if it either controls the function of the device or is built into the device. Thus, a software update for a device is considered a device. Software not associated with a device, such as that used for accounting, patient records, or other administrative functions is not a device.

7.5.2. Good Manufacturing Practices

Manufacturers of all medical devices must follow a set of guidelines referred to as good manufacturing practices (GMP) to ensure quality of the product. The basic objectives are to provide a product that is free from contamination, that is consistent across all realization of the device, that was manufactured by well-trained personnel, that has been checked for quality during the manufacturing process, and whose manufacture has been well documented. When GMP is followed, it can lead to consistent high-quality products and reduced liability.

7.5.3. Premarket Approval and 510k

A device is approved for marketing through one of two mechanisms. If the device is considered to have a low potential to harm the patient, it can be filed through a 510k application. Devices that have a greater potential for harm must be reviewed through a premarket approval (PMA), which is a more lengthy process that requires the submission of data that demonstrates a products safety and effectiveness.

7.5.4. Classes of Medical Devices

Devices are categorized according to their impact on the patient and the potential consequences of failure as Class I (lowest impact), Class II (medium impact), and Class III (highest impact). Each class has a different set of requirements for FDA approval. Marketing of these devices requires adherence to basic standards, GMP, and proper record keeping. A premarket notification, registration, and device listing are required.

Class I devices are non-life sustaining. Their failure will affect the patient's health minimally or not at all. Examples include stethoscopes, bandages, crutches, and latex examination gloves.

Class II devices impose medium to high risk on the patient. Examples include blood pressure cuffs, contact lenses, subdural catheters, syringes, oximeters, CPAP, and surgical gloves. Some devices are given a higher designation as Class IIb. These devices include ventilators, infusion pumps, refillable insulin pens, and diagnostic x-ray machines. For approval, Class II devices frequently require only a 510k.

Class III devices can subject the patient to high risk. Examples are cochlear implants, vascular stents, surgical meshes, defibrillators, and pacemakers. For approval, Class III devices nearly always require a PMA.

Chapter 8: Sensors, Transducers, and Signal Conversion

In general, a measurement involves conversion of a physical quantity into a readable format, but in most cases, the readable format is an electrical signal* (voltage or current) because these signals are highly versatile. The sensor component of a measurement device is that which changes a property with respect to the physical quantity of interest. A thermistor, for example, changes its resistance with the physical property temperature. A transducer converts one form of energy into another. Generally, sensors and transducers are combined to create a measurement device. However, this chapter will not focus on the distinction between sensors and transducers, as common language tends to use these terms interchangeably.† The primary concern for this chapter is the conversion of a physical quantity to a voltage or current.

8.1 Change in Impedance

A large number of measurement devices depend on the variation of the sensing element's impedance. Circuitry is then needed to convert that impedance change to a voltage change, but the circuitry can be relatively simple.

8.1.1. Resistive sensors

Figure 8-1 shows a simple circuit for the conversion of resistance to voltage. A constant supply voltage (V_{supply}) is connected to one side of a constant resistor (R), and the varying sensor resistor (R_s) is attached between the other side and ground. The output (V_{out}) is taken between the two resistors.

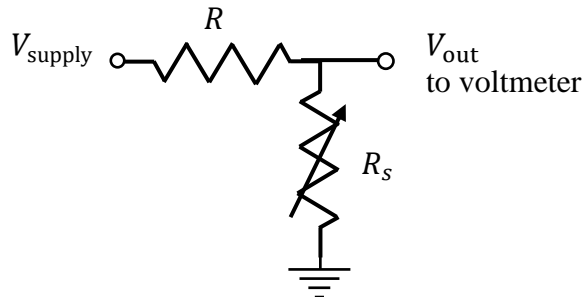


Figure 8-1: Bridge configuration to convert a change in resistance to a change in voltage.

The input impedance into the voltmeter is assumed to be high, so that all current through R must also pass through R_s .

$$i = \frac{V_{\text{supply}}}{R + R_s}$$

V_{out} is simply the current times R_s .

* Exceptions include, for example, a mercury thermometer or the overpressure valve on a water heater. Both are sensors, but neither of them involve voltage.

† If a surgeon asks a nurse to “Check the pressure transducer reading,” the nurse is not likely to respond, “Do you mean pressure transducer or pressure sensor?”

$$V_{\text{out}} = iR_s = \frac{V_{\text{supply}}R_s}{R + R_s}.$$

If R_s is small throughout its range so that $R_s \ll R$, the output voltage is linearly proportional to R_s . However, small ranges for R_s translate to small ranges in V_{out} , so it is not unusual to use this bridge in a nonlinear range. Also, sensing resistors typically do not go to zero, so the circuit will nearly always lead to a voltage offset. A more complicated bridge, the Wheatstone bridge, can be configured such that the voltage output is zero at a finite value of the resistance. This bridge will be described in Section 8.2. Some of the more common resistive devices are described here.

Thermistor

The thermistor changes its resistance with temperature. They are relatively small in size (on the order of a millimeter). The equation that relates resistance to temperature is

$$R(T) = R_0 e^{\beta\left(\frac{1}{T} - \frac{1}{T_0}\right)},$$

Where R_0 is the resistance at a fixed reference temperature T_0 , and all temperatures are absolute (e.g., Kelvin). The equation can be rewritten as $R = R_0 e^{-\beta/T_0} e^{\beta/T} = R_\infty e^{\beta/T}$, so that R_∞ is the theoretical resistance at extremely high temperature.* The relationship is nonlinear, but it can be approximated as linear if the variations in temperature are sufficiently small. The Taylor series around T_0 is

$$R(T) \approx R(T_0) + (T - T_0)R'(T - T_0)$$

$$R(T) \approx R_0 + (T - T_0) \left(R_0 \beta e^{\beta\left(\frac{1}{T} - \frac{1}{T_0}\right)} \right) \frac{1}{T^2} \Bigg|_{T=T_0} = R_0 + (T - T_0) \frac{\beta R_0}{T_0^2}.$$

The constant $\beta R_0/T_0^2$ is designated as K , the temperature coefficient, which can be positive or negative, depending on the thermistor material. Positive K is referred to as “positive temperature coefficient” (PTC), and negative K is referred to as “negative temperature coefficient.”

In using a thermistor, current must be kept low enough to prevent dissipated power from self-heating the device. If self-heating occurs, voltage can decrease with temperature instead of increasing like a normal resistor, so that a negative resistance effect occurs.

Photoresistor

The photoresistor changes its resistance with the intensity of light impinging upon it. The relationship between resistance and light intensity follows the power law

$$\frac{R}{R_0} = \left(\frac{I}{I_0} \right)^{-\gamma}.$$

* “Theoretical” because an infinite temperature would clearly destroy the thermistor.

Thus, resistance decreases with increasing light intensity.

Strain Gauge

The strain gauge is in essence a wire that looped back and forth and placed on a backing, as shown in Figure 8-2. It is glued to a material that is being strained. As strain is applied to that material, the wire stretches, which causes its length to increase and its cross-sectional area to decrease. The equation for resistance of a wire is

$$R = \frac{\rho L}{A},$$

where ρ is resistivity, L is the length of the wire, and A is the cross-sectional area of the wire. Thus, the increases in L and decrease in A both increase the wire's resistance, so increased strain leads to increased resistance. The gauge is connected through the leads into a bridge so that the change in resistance is converted to a change in voltage. The gauge is primarily sensitive to strain along the length of the long wires, but it will have some small sensitivity in the perpendicular direction.

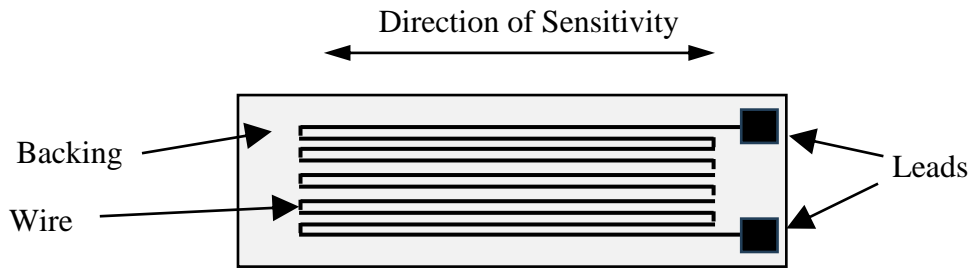


Figure 8-2: Strain gauge

One application of a strain gauge is in material testing, where a sample of the material to be tested is formed into a coupon (i.e., an appropriate geometry for the desired test), the gauge is glued to the coupon (Figure 8-3), and the coupon is placed into a material testing machine that stretches, compresses, and/or twists the coupon with a known load or torque.

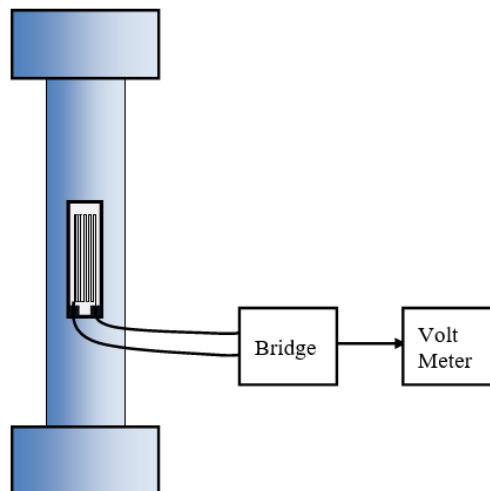


Figure 8-3: Strain gauge attached to a coupon for material testing.

Two gauges mounted perpendicular to one another can provide measurements of axial and lateral strain. A gauge for torsion has the wires arranged diagonally.

Strain gauges are also incorporated into transportation bridges to monitor the bridges' state of degradation and determine whether maintenance is necessary. In biomedical engineering, a strain gauge can be attached to bone to test its material properties or to a biomaterial for material testing. The strain gauge principle is also incorporated into transducers for force and pressure, as will be discussed in Section 8.2.1.

Piezoresistor

A piezoresistor can be used in applications in a manner similar to a resistor, however, whereas strain on a resistor changes length and area in the formula $R = \rho LA$, strain on a piezoresistor also changes the resistivity (ρ). Thus, the piezoresistor can have larger sensitivity to strain. A piezoresistor should not be confused with a piezoelectric element, for which the strain on the element determines the amount of charge that it generates.

8.1.2. Capacitive

The capacitance of a capacitor increases in proportion to the reciprocal of the spacing between the plates. Thus, any force applied to the capacitance that decreases this spacing will increase capacitance. This effect allows specially designed capacitors to be used as sensors for force, pressure, and displacement. Furthermore, because the capacitance depends on the properties of the dielectric material between the plate, changes in this material's properties through, for example, a change in humidity, can be exploited to change the capacitance. If the capacitor is placed in a bridge, the change in force will lead to a change in voltage. The bridge and its supplying voltage are more complicated than a resistor bridge, but based on the same principles, as will be described in Section 8.2.2.

8.1.3. Inductive

As an inductor is moved close to a conducting metal, its inductance changes. The consequent change in impedance can be exploited to detect metals and displacements (e.g., from the inductor to a metal). Section **Error! Reference source not found.** provides more detail in this type of sensor.

8.1.4. Acoustic: Heart sounds

Audio frequency sound can be generated by biological phenomena that include turbulent blood flow (generated by an arterial stenosis or stenosed heart valve), movement of material through the digestive system, and airflow in the lung. Lung sounds are divided into rales (clicking sounds evident during inhalation), rhonchi (e.g., snoring), stridor (high pitched sounds in the trachea or throat), and wheezing from the bronchi. Furthermore, a percussion test, in which the examiner raps on the patient's back and listens for reverberation, provides information about fluid in the lung. Several instruments are available to sense audio frequency sounds, such as the capacitive microphone (Section 8.6) and the piezoelectric sensor (Section 8.7.4). Of course, the classic acoustical instrument is the stethoscope, and despite the relative simplicity of this instrument, its practical utility should not be underestimated.

8.1.5. Blood flow

Blood flow rate is a useful measured parameter *in vivo*. During bypass surgery, for example, the surgeon uses it to verify that flow in the grafted vessel is adequate. Two primary methods are used to measure blood flow rate *in vivo*, the electromagnetic flow meter and the ultrasonic flow meter.

Electromagnetic flow meter

The basis for a simple electromagnetic flow meter is shown in Figure 8-4. When a conducting material moves with velocity \mathbf{u} through a stationary magnetic field of strength \mathbf{B} , an electrical field $\mathbf{u} \times \mathbf{B}$ is produced, where the direction of the field is obtained from the right-hand rule. The induced voltage between two points on the conductor is the integral of the electric field along a given path line. A voltage difference is then produced by this field between electrodes between locations marked + and - that can be measured. With the assumption that the magnetic field is perpendicular to the flow velocity and the velocity is uniform over the cross-section, the induced voltage will be $V = uBD$, where u and B are the magnitudes of vectors \mathbf{u} and \mathbf{B} , and D is the distance between the electrodes (diameter of the vessel). The device requires that the working fluid be conductive, which is not a problem *in vivo* because of the high saline content of blood. If it is used *in vitro*, salt can be added to the fluid.

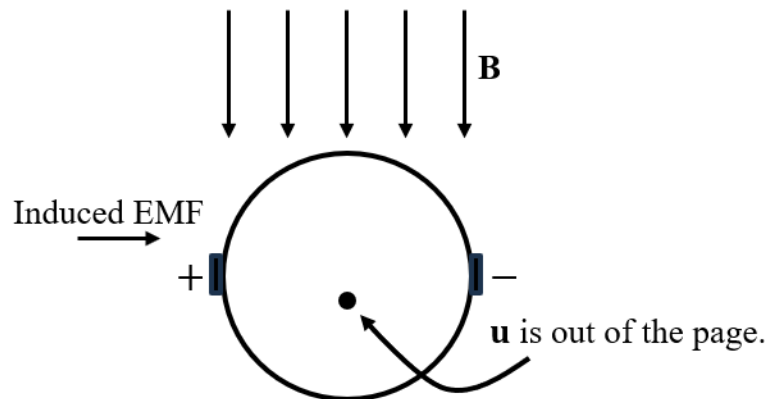


Figure 8-4: Simple electromagnetic flow meter. The movement of a conducting media through the magnetic field \mathbf{B} induces an electric field that leads to a voltage difference between the + electrode on the left and the - electrode on the right. The bullet (•) indicates the velocity vector oriented out of the page.

While this simple equation illustrates the basic principle of the device, several problems arise that require commercially available electromagnetic flow meters to be substantially more complicated. The first problem is that the velocity profile will generally not be uniform, so the induced voltage will depend to some extent on the nature of the flow, with fully turbulent flow and fully laminar (Poiseuille) flow leading to different results. Secondly, the generation of a completely uniform magnetic field is not practical, so adjustments must be made to compensate for the non-uniformity. The third problem is noise. The flow meter will pick up electrical signals from the heart (the EKG), the electrode/fluid interface. It will also pick up a drifting voltage at the electrode-fluid interface that is on the order of the velocity signal. In addition, it will pick up a common type of noise, known as $1/f$ noise, that increases with decreased frequency. The solution to these problems is to modulate the magnetic field so that the induced

voltage occurs at a frequency that is beyond that of the EKG, the drift, and the $1/f$ noise. The required electronics is then further complicated because the varying magnetic field induces an additional voltage component referred to as the transformer voltage that must be removed through a quadrature demodulation feedback mechanism. A more complete description can be found in [5].

Ultrasonic flow meter

The ultrasonic flow meter uses the transit time of an ultrasound pulse between two transceivers aligned at an angle θ along the flow direction to estimate flow rate. The principle is illustrated in Figure 8-5.

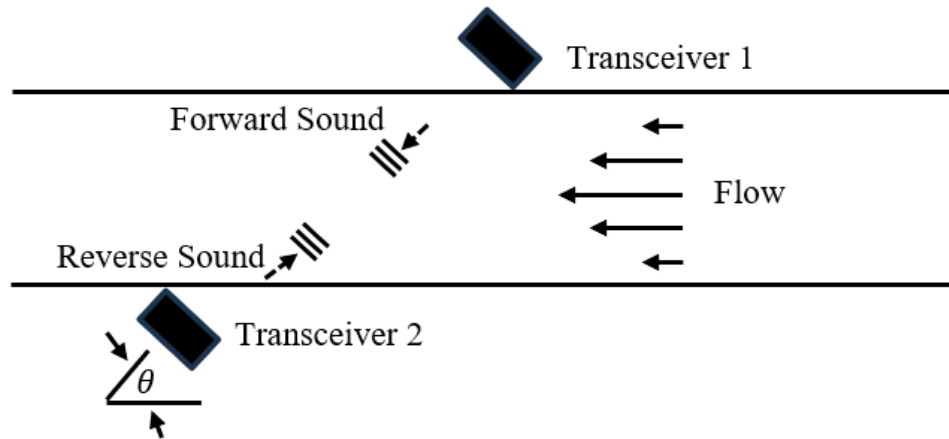


Figure 8-5: Concept of the electromagnetic flow meter.

The component of sound velocity in the direction of a line between Transceiver 1 and Transceiver 2 is $c + u \cos(\theta)$, where c is the speed of sound in the fluid and u is the axial velocity. The component of velocity in the opposite direction is $c - u \cos(\theta)$. If the distance between the two probes is L , then the difference between the time required for sound to pass from Transceiver 2 to Transceiver 1 and from Transceiver 1 to Transceiver 2 is

$$\Delta t = \frac{D}{c - u \cos(\theta)} - \frac{D}{c + u \cos(\theta)}.$$

Because $u \ll c$, the expressions can be approximated by the first two terms of the Taylor series expanded in $u \cos(\theta)$.

$$\Delta t \approx D \left(\frac{1}{c} + \frac{u \cos(\theta)}{c^2} \right) - D \left(\frac{1}{c} - \frac{u \cos(\theta)}{c^2} \right) = \frac{2Du \cos(\theta)}{c^2}$$

8.2 Wheatstone Bridge

A bridge is needed to convert a change in resistance to a voltage. The voltage divider bridge diagrammed in Section 8.1.1 has the disadvantage that the output cannot be zero or negative. The Wheatstone bridge configuration circumvents this problem.

8.2.1. Resistive

The resistive Wheatstone bridge configuration is shown in **Figure 8-6**. At least one of the resistors R_A , R_B , R_C , or R_D is variable (changing according to the property being measured). In some cases, multiple resistors may be variable, and all four are variable in some applications. The student may have been introduced to the Wheatstone bridge configuration, where a fifth resistor is included between Node a and Node b . Note, however, that the volt meter must have a finite input resistance and that this input resistance is topographically located in the same position as the fifth resistor. In many applications it is assumed that the input resistance is much larger than R_A through R_D , which simplifies the analysis of the circuit.

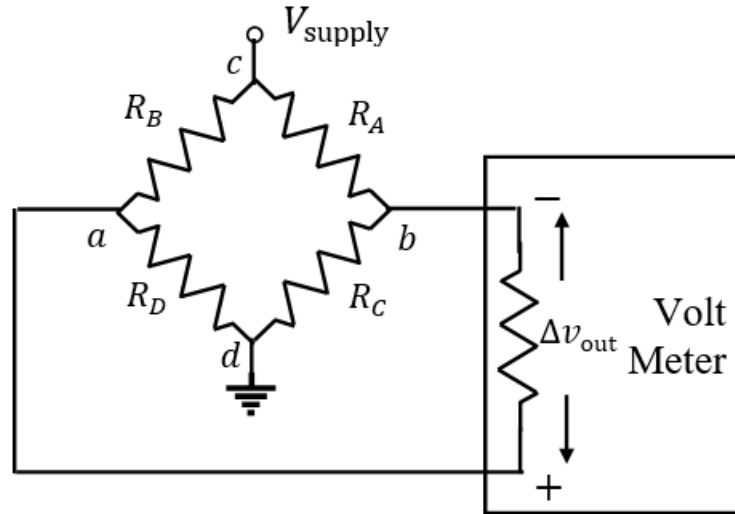


Figure 8-6: Wheatstone bridge as used to convert a change in resistance to a voltage.

One generally first finds a set of conditions for which the bridge is zeroed, meaning that the voltage read by the volt meter (the voltage between Node a and Node b) is zero. The voltage at Node a is the current through R_D multiplied by R_D . The current through R_D is equal to the current in R_B , and since no current is drawn by the volt meter as a result of its large input resistance, this current is $V_{\text{supply}}/(R_B + R_D)$. Hence, $v_a = V_{\text{supply}}R_D/(R_B + R_D)$. Similarly, $v_b = v_a = V_{\text{supply}}R_C/(R_A + R_C)$, so that

$$\Delta v_{\text{out}} \equiv v_a - v_b = V_s \left(\frac{R_D}{R_B + R_D} - \frac{R_C}{R_A + R_C} \right) \quad \text{Equation 8-1}$$

If the bridge is balanced, $v_a - v_b = 0$, so

$$\begin{aligned} \frac{R_D}{R_B + R_D} &= \frac{R_C}{R_A + R_C} \Rightarrow R_D R_A + R_D R_C = R_C R_B + R_C R_D \\ &\Rightarrow R_D R_A = R_C R_B \Rightarrow \frac{R_A}{R_C} = \frac{R_B}{R_D}. \end{aligned}$$

Specifically, the balance condition requires that the ratio between the upper and lower resistors on the left-hand side equal the ratio between the upper and lower resistors on the right-hand side.

Equation 8-1 is nonlinear. However, if the change in resistance is small, Δv_{out} is approximately proportional to that change and the result is linear. For example, assume that R_D changes by only 1% and that $R_B = R_D$. Then

$$\Delta v_{\text{out}} \equiv V_s \left(\frac{R_D + 0.01R_D}{R_B + R_D + 0.01R_D} - \frac{R_C}{R_A + R_C} \right).$$

Because $0.01R_D$ is two orders of magnitude smaller than R_D it can be ignored in the denominator. Because the bridge was initially balanced, $R_D/(R_B + R_D)$ cancels with $R_C/(R_A + R_C)$, and the result is the linear relationship $\Delta v_{\text{out}} \approx V_s(\Delta R_D)/(R_B + R_D)$. The analysis is more complicated if R_B or R_A changes, but a linear result is still obtained.

One type of pressure transducer uses the strain gauge principle to determine the displacement of a piston that is being forced in one direction by the pressure to be measured. The concept is shown in Figure 8-7. The piston (gray region) is displaced to the right by the applied pressure and is attached to stationary regions through wires with resistances R_A , R_B , R_C , and R_D . All four wires are stretched so that they are under tension when the pressure is zero. The displacement stretches wires R_B and R_C , which makes them longer and decreases their cross-sectional area, thus increasing their resistance. The same displacement relaxes wires R_A and R_D , decreasing their lengths and increasing their cross-sectional area, thus decreasing their resistance. The four resistances are then connected together as in **Figure 8-6**, and the output voltage is governed by Figure 8-1. Here, all four resistors contribute to an increase in the absolute value of Δv_{out} .

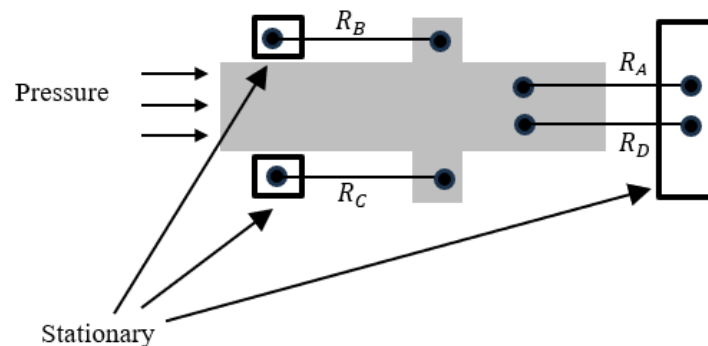


Figure 8-7: Schematic for a strain gauge type pressure transducer.

8.2.2. Capacitive/Inductive

For some sensors, it is the capacitance or the inductance of the sensing element that changes instead of the resistance. The Wheatstone bridge can still be used in these cases, but with some modification because one must use AC rather than DC signals. Impedance for inductors and capacitors is purely reactive and does not depend on capacitance or inductance at 0 frequency (DC). In a capacitance bridge, the impedance of each capacitor for DC is infinite, independent of the capacitance, and if it is an inductance bridge, the impedance of the inductors for DC is zero, independent of the inductance. The modified bridge is shown in Figure 8-8 for the inductance bridge. First, the driving signal is AC, rather than DC. Second, the detector is an RMS volt meter rather than a DC volt meter. The RMS volt meter is a combination of a rectifier, a low pass filter, and a DC volt meter. The cutoff frequency for the low pass filter is selected to

be substantially lower than the driving frequency so that enough cycles of the output are averaged over time to obtain a consistent signal.

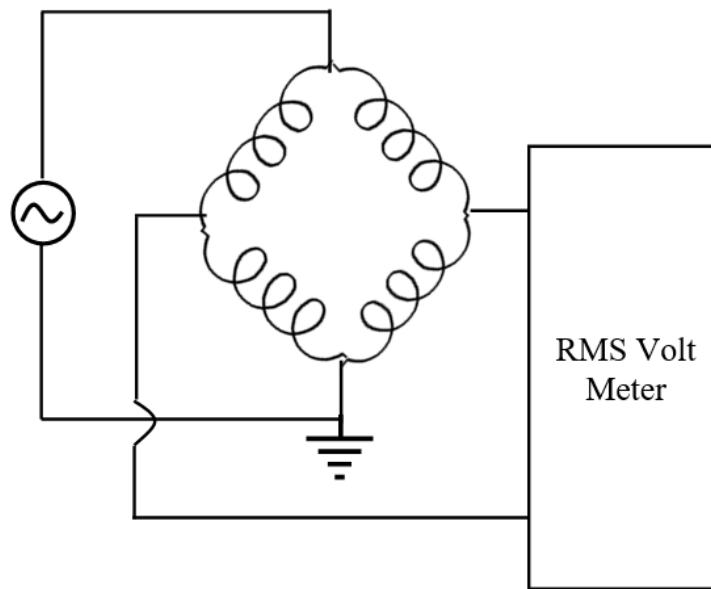


Figure 8-8: Wheatstone bridge for inductive elements.

Because the impedance of the active element(s) can be a function of time. It is therefore useful to have an output signal from the volt meter that is proportional to the impedance. Alternatively, one can feed the signal into a rectifier/low pass filter combination and take the output from that network. This process, converting a signal with varying amplitude to a varying signal that is proportional to the amplitude is referred to as demodulation.

8.3 Rectifier

8.3.1. Diode Bridge

One technique for rectification of an AC signal is through the diode Wheatstone bridge shown in Figure 8-9. Although this circuit has substantial disadvantages it is described here because its operation is simple to understand and forms the basis for understanding the more practical field effect transistor bridge. Assume that each diode is perfect, such that it acts as an open circuit when the voltage across it is positive and a short circuits when the voltage across it is negative. When the input signal (V_{signal}) is greater than ground, diode D_1 is reverse biased so that it does not conduct, and diode D_3 connects the output signal on the left to ground. Diode D_2 is forward biased, with D_4 reverse biased, so that V_{signal} is connected to the output node on the right. V_{rect} is then equal to V_{signal} . If V_{signal} is negative, diodes D_1 and D_4 conduct while diodes D_3 and D_2 do not, so that ground is connected to the output signal on the right and the negative signal is connected to the output signal on the left. The low side of V_{rect} is then negative, so that V_{rect} is positive. V_{rect} is then positive whether V_{signal} is positive or negative

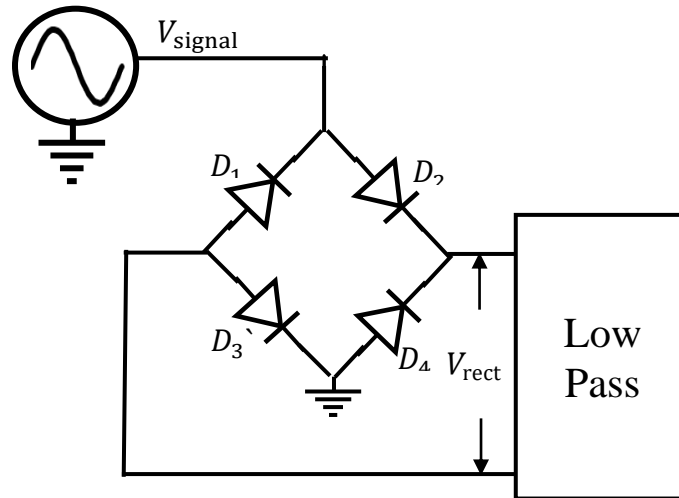


Figure 8-9: Diode Wheatstone bridge.

A problem with this circuit is that practical diodes are imperfect, and require a finite voltage across them before they conduct substantially. That voltage is generally taken as approximately 0.7 volts, so a substantial amount of V_{signal} is lost with this configuration. Figure 8-10 shows the output signal from such a rectifier. The circuit works well for input voltages above about 0.6 volts, but the output does not follow voltage below that value.

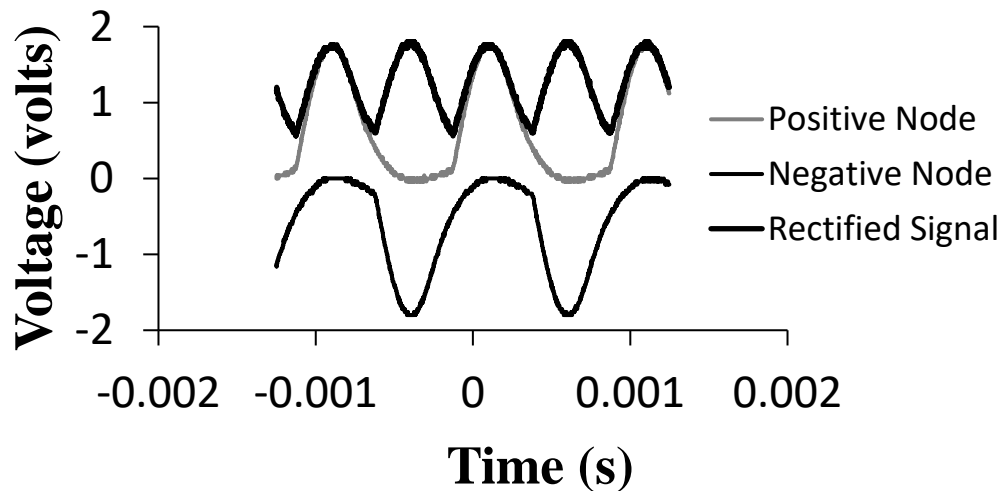


Figure 8-10: Signals from a diode bridge.

8.3.2. Field effect transistor bridge

A more effective rectifier bridge can be constructed from MOSFET transistors. The configuration uses two P-type transistors and two N-type transistors, as shown in Figure 8-11. The circuit will rectify voltages that are below the 0.5 volt limit encountered by the diode bridge.

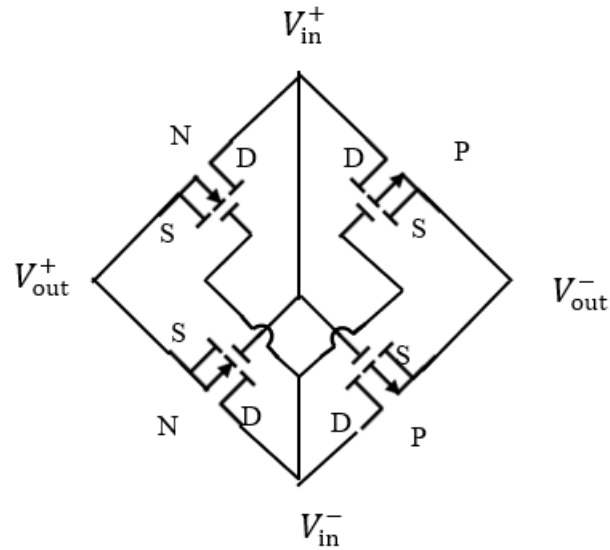


Figure 8-11: Rectifier bridge constructed from MOSFET transistors.

8.4 Coils

If an AC current passes through a coiled wire, it generates a magnetic field that induces an AC current in a coil placed nearby. The strength of the magnetic field, and hence magnitude of the induced current depends on the distance between the two coils, and it can also be modified by a conductive material placed between the two coils. This phenomenon is the basis of several instruments.

8.4.1. Displacement Transducer

Two configurations for a simple displacement transducer are sketched in Figure 8-12. The source generates a varying magnetic field that is picked up by the receiver (meter) and translated into a current in the coil. The magnetic field weakens with distance from the source. Without the iron core (left), the distance between the two coils determines the strength of the output signal. With the iron core (right), the core carries the magnetic field to the receiving coil, so the position of the core determines the strength of the received signal, and the coils themselves are typically kept stationary.

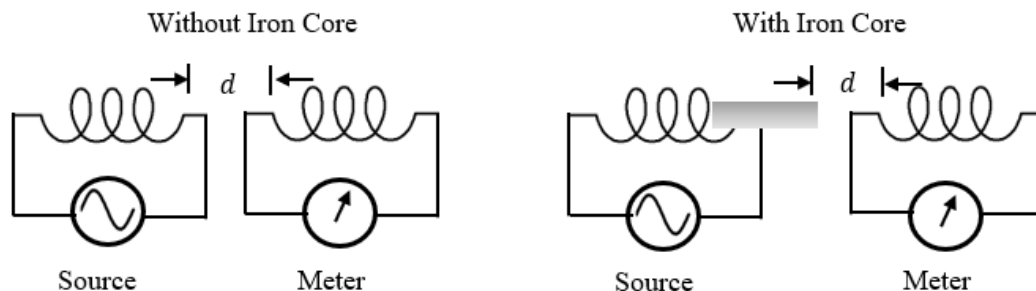


Figure 8-12: Simple displacement transducer. Without the iron core (Left), the detected signal is related to the distance between the two coils. With the iron core (Right), the detected signal is related to the position of the core, and the two coils are held stationary.

8.4.2. Transformer

A transformer uses two coils to change the amplitude of an AC voltage. The configuration is shown in Figure 8-13. In the figure, the V_{in} side has six loops ($N_{in} = 6$) and the V_{out} side has four loops ($N_{out} = 4$). The ratio of the input amplitude to the output amplitude is equal to the ratio of the loops, so $V_{out}/V_{in} = N_{out}/N_{in} = \frac{4}{6} = 2/3$. Thus, the output amplitude will be attenuated by a factor of 2/3rds. If the output has more coils than the input, then the output has a larger voltage amplitude. Conservation of energy dictates that the current at the output must be less than that at the input because power is voltage times current. Transformers are ubiquitous in commercial power supply lines, for example to convert the 4800 volt power from the power substation to the 250 volt power inside the home or business. They can also be used to increase voltage in cases where the resulting reduction in current is not a concern, for example, in driving a piezoelectric crystal.

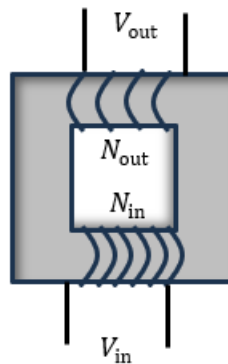


Figure 8-13: Transformer

8.4.3. Magnetic Pickup

Coils induce a current whenever a fluctuating magnetic field is present. A classic example is the pickup on an electric guitar. The pickup contains a coil wrapped around a permanent magnet. As the metal string vibrates through the magnetic field, it causes the field to fluctuate. The fluctuating field then induces a current in the coil, which is amplified and sent to the speaker to produce sound. The magnetic stylus on an audio turntable works similarly. As the needle is moved by the bumps in the record grooves, it moves a permanent magnet that causes fluctuations in the magnetic field that are picked up by coils. For stereo, two coils are used; one is sensitive to motion in a direction at 45 degrees from the record groove, and the other is sensitive to motion in a direction 135 degrees from the record groove.

8.5 Phase-Sensitive Demodulator

An alternative coil-based displacement transducer is shown in Figure 8-14. The signal from the source is carried through the iron rod to the other two coils. Thus, when the iron rod is closer to the rightmost bottom coil, the voltage across that coil is greater than the voltage across the leftmost coil. Because of the topology of the connections between the two lower coils, a voltage carried to the two coils with polarity as shown in the figure will be read on the meter as a positive component from the right coil minus the component from the left coil. Hence, if the

core is in the middle position such that equal voltage is carried to the two coils, the meter will read zero.

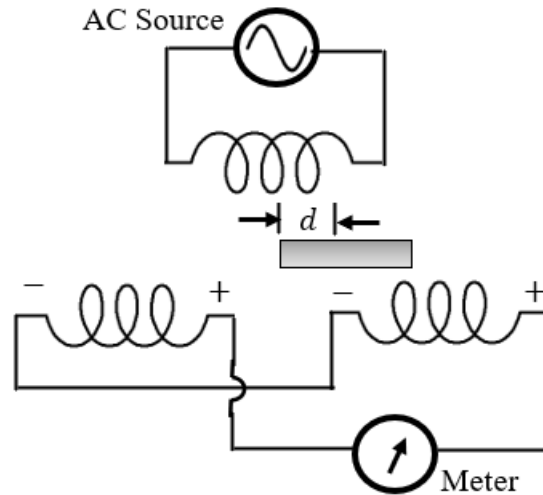


Figure 8-14: Coil-based displacement transducer designed to detect positive and negative displacements.

The signals across the meter are shown as a function of time in Figure 8-15 for three core positions. When the core is closer to the right-hand coil, the meter reads a voltage signal that is in phase with the input signal. When the core is closer to the center, the voltages from the left and right coil cancel one another, so the output is attenuated. When the core is closer to the left-hand coil, the output is once again large, but out of phase with the input signal. Straightforward demodulation cannot distinguish left and right positions. To distinguish left from right, a detector is needed that provides a negative DC voltage when the input and output signals are out of phase and a positive signal when the input and output signals are in phase. This detector is referred to as a phase-sensitive demodulator.

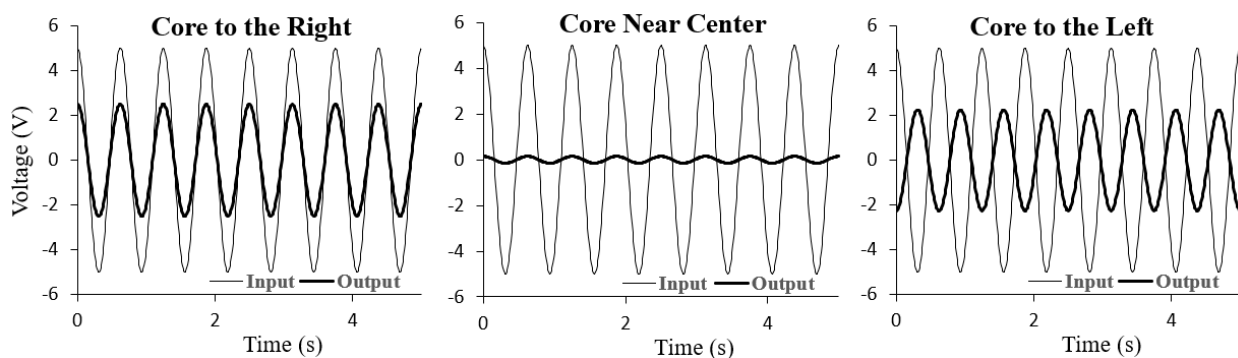


Figure 8-15: Signals across the meter for the displacement transducer shown in Figure 8-14.

The phase-sensitive demodulator is conceptually simple, and the signals are shown in Figure 8-16. First, the detector creates a signal that has a value of +1 when the input signal is positive and -1 when the input signal is negative. The component that creates such a signal is an operational amplifier with no feedback so that the output saturates positive when the input is positive and negative when the input is negative. Next, the detector multiplies the output signal

by this square wave. If the input and output signals are in phase, this product is positive. If they are out of phase, the product is negative. When the signal is low pass filtered, the final result is a positive value when the signals are in phase and a negative value when the signals are out of phase.

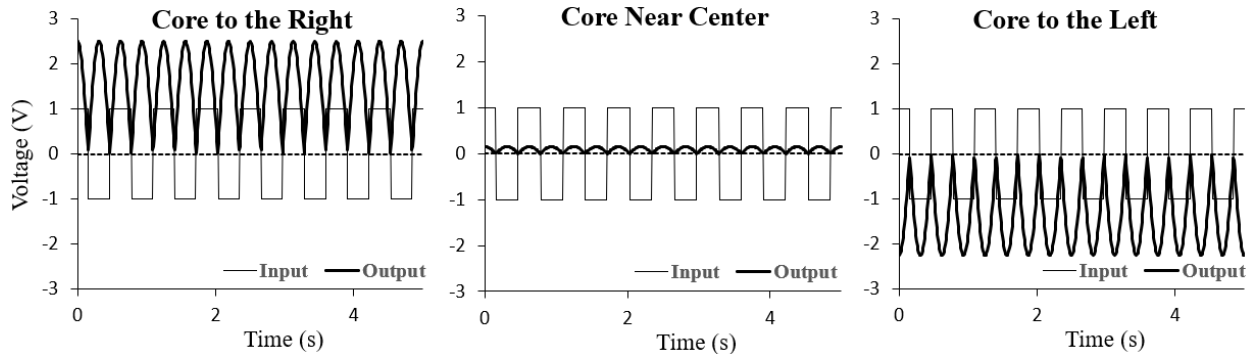


Figure 8-16: Intermediate signals for the phase-sensitive demodulator, as it is applied to the signals in Figure 8-15.

An implementation of the phase-sensitive demodulator is shown in Figure 8-17. The driving voltage of the transducer is converted to a square wave through the lower operational amplifier, and the square wave drives a voltage-controlled switch. If the control voltage is positive, the output of the switch is connected to Input 2, and if the voltage is negative, the output is connected to Input 1. The output from the displacement transducer is fed into both an inverting operational amplifier circuit, which leads to Input 1 to a voltage-controlled switch. The signal is also carried directly to Input 2 of the switch. The switch is controlled by the square wave. Thus, the operation of the switch is equivalent to multiplying the signal from the transducer output by $+1$ or -1 .

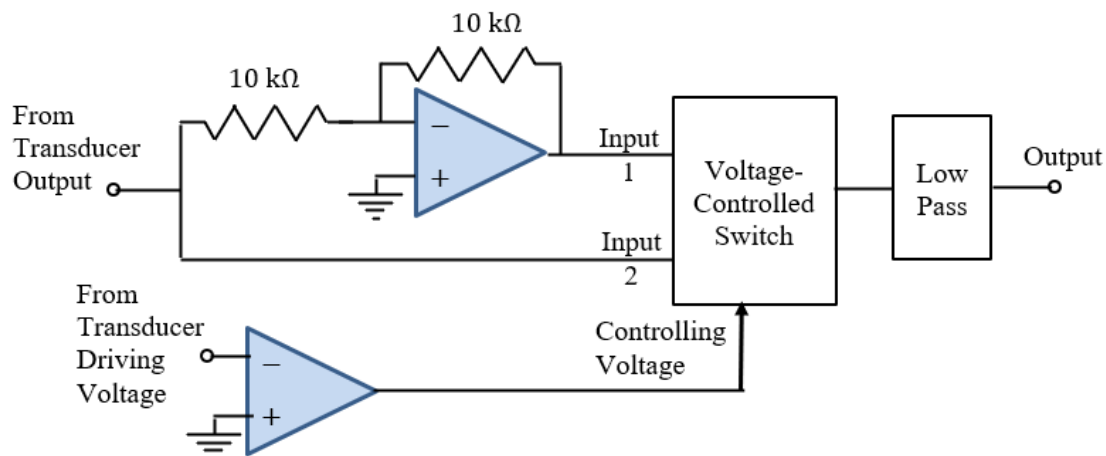


Figure 8-17: Implementation of the phase-sensitive demodulator.

8.6 Capacitive and Electret Microphone

One of the most common microphone configurations uses the change in capacitance of an element to detect sound. Instead of the Wheatstone bridge configuration, the capacitive microphone uses the circuit shown in Figure 8-18.

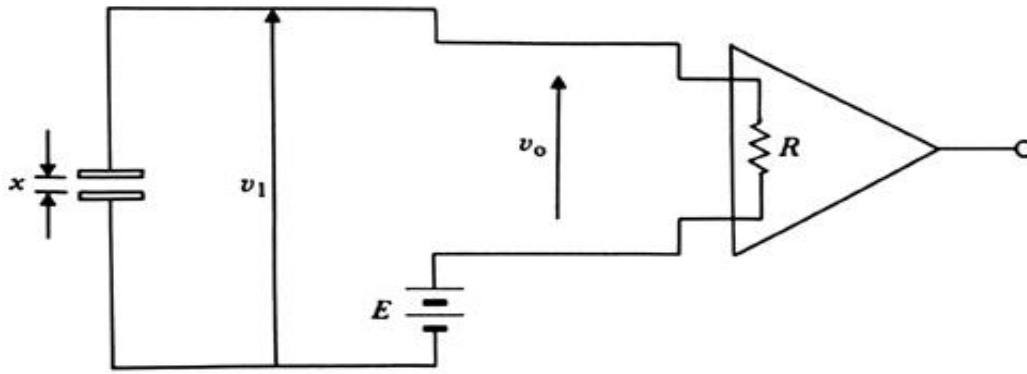


Figure 8-18: Capacitive microphone circuit.

The displacement between the plates of the capacitor changes in response to pressure. The resulting capacitance (C) is

$$C = \epsilon_0 \epsilon_r \frac{A}{x}. \quad \text{Equation 8-2}$$

where A is the cross-sectional area of the capacitor plates, x is the separation between the plates, ϵ_0 is the electric constant ($\approx 9.85 \times 10^{-12}$), and ϵ_r is the dielectric constant. An amplifier circuit for this type of sensor is shown in Figure 1.

Kirchov's voltage law applied to the circuit leads to Equation 8-3.

$$v_o = v_1 + E \quad \text{Equation 8-3}$$

Kirchov's current law states that the current through the capacitor is equal to the current through the resistor.

$$\frac{v_o}{R} = -\frac{d}{dt}(Cv_1) \quad \text{Equation 8-4}$$

Equation 8-3 and Equation 8-4 can be combined to eliminate v_1 .

$$\frac{v_o}{R} = -\frac{d}{dt}(C(v_o - E)) \quad \text{Equation 8-5}$$

The capacitor spacing, x , will be divided into a constant term, x_0 , and a term that fluctuates with the pressure on the sensor, Δx .

$$x = x_0 + \Delta x \quad \text{Equation 8-6}$$

Now x can be replaced by Equation 8-6 in Equation 8-2, and the resulting expression for C can be used in Equation 8-5.

$$\frac{v_o}{R} = -\frac{d}{dt} \left(\epsilon_0 \epsilon_r \frac{A}{x_0 + \Delta x} (v_o - E) \right) \quad \text{Equation 8-7}$$

Assume that $x_0 \gg \Delta x(t)$ so that $1/(x_0 + \Delta x(t))$ can be expanded in a Taylor series around x_0 . The expression is then approximated by the first two terms as in Equation 8-8.

$$\frac{1}{x_0 + \Delta x(t)} \approx \frac{1}{x_0} - \frac{\Delta x}{x_0^2} \quad \text{Equation 8-8}$$

The capacitance can therefore be written as

$$C = \epsilon_0 \epsilon_r \frac{A}{x} \approx \epsilon_0 \epsilon_r \frac{A}{x_0} \left(1 - \frac{\Delta x}{x_0} \right) = C_0 \left(1 - \frac{\Delta x}{x_0} \right), \quad \text{Equation 8-9}$$

where

$$C_0 \equiv \epsilon_0 \epsilon_r \frac{A}{x_0}. \quad \text{Equation 8-10}$$

Equation 8-7 is then

$$\frac{v_o}{R} = \frac{d}{dt} \left(C_0 \left(1 - \frac{\Delta x}{x_0} \right) (E - v_o) \right). \quad \text{Equation 8-11}$$

Expand the product of the two parentheses in Equation 8-11.

$$v_o = RC_0 \frac{d}{dt} \left(E - \frac{E\Delta x}{x_0} - v_o + \frac{v_o\Delta x}{x_0} \right). \quad \text{Equation 8-12}$$

Take the derivative of each of the terms in parentheses separately, noting that E is constant, so its derivative is zero.

$$v_o = RC_0 \left(-\frac{E}{x_0} \frac{d\Delta x}{dt} - \frac{dv_o}{dt} + \frac{1}{x_0} \frac{d(v_o\Delta x)}{dt} \right). \quad \text{Equation 8-13}$$

The derivative in the last term is

$$\frac{d(v_o\Delta x)}{dt} = v_o \frac{d\Delta x}{dt} + \Delta x \frac{dv_o}{dt}. \quad \text{Equation 8-14}$$

Therefore, Equation 8-13 can be regrouped as

$$v_o = \frac{RC_0}{x_0} \left((v_o - E) \frac{d\Delta x}{dt} + (\Delta x - x_0) \frac{dv_o}{dt} \right). \quad \text{Equation 8-15}$$

With $x_0 \gg \Delta x$ and $E \gg v_o$, Equation 8-15 becomes

$$v_o = -\frac{RC_0}{x_0} \left(E \frac{d\Delta x}{dt} + x_0 \frac{dv_o}{dt} \right). \quad \text{Equation 8-16}$$

Write this equation in a more standard form.

$$\frac{dv_o}{dt} + \frac{1}{RC_0} v_o = -\frac{E}{x_0} \frac{d\Delta x}{dt}, \quad \text{Equation 8-17}$$

Equation 8-17 is a first order ordinary differential equation in v_o , with input $-\frac{E}{x_0} \frac{d\Delta x}{dt}$. Take the Fourier transform of this equation.

$$j\omega V_o + \frac{1}{RC_0} V_o = -\frac{j\omega E}{x_0} X \quad \text{Equation 8-18}$$

The transfer function between displacement and voltage is then

$$\frac{V}{X} = -\frac{j\omega E/(x_0)}{j\omega + \frac{1}{RC_0}} = -\frac{E}{x_0} \frac{j\omega RC_0}{(j\omega RC_0 + 1)} \quad \text{Equation 8-19}$$

The sensor and circuit therefore act as a high-pass filter. Generally, one would like to select values of R and C_0 so that the low-frequency cutoff is lower than the low frequency limit of human hearing, or about 20 Hz.

The voltage source E for this circuit can come from a battery or can be provided by an audio amplifier. When it is supplied by the amplifier, it is referred to as phantom power. A variation on this circuit is the electret microphone, for which power is supplied by a charge built into the sensor and a JFET transistor is incorporated in the element to perform the task of the differential amplifier. Electret microphones are common in electronic devices such as cell phones and personal computers. A typical circuit is shown in Figure 8-19.

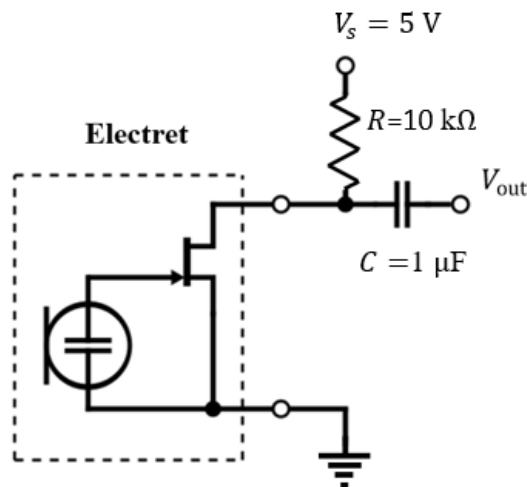


Figure 8-19: Electret microphone circuit.

8.7 Direct generation of voltage or current

Several transducer types generate an electrical signal directly, without the need for an exciting voltage or a bridge network. These transducers will generally require amplifiers with specific designs to enhance the signal to a readable or recordable level.

8.7.1. Thermocouple

When two wires made of dissimilar metals are connected, a voltage is generated by the Seebeck effect (a temperature-dependent voltage at the junction) and the Thomson effect (a voltage induced in the wire by a temperature gradient). Although the relationship between temperature and voltage is nonlinear, it can be approximated as linear for sufficiently small temperature ranges. A thermocouple consists of two such junctions, the temperature-sensing junction (hot junction) and the reference junction (cold junction), and the voltage is a function of the temperature difference between these two junctions. The equation for voltage as a function of temperature is $V = \alpha(T_H - T_C)$, where V is the voltage, T_H is the temperature at the hot junction, T_C is the temperature at the cold junction, and α is the Seebeck coefficient. Historically, the reference junction was kept at a temperature of 0 °C because this temperature was easily produced by an ice bath. Different metals are selected for the thermocouple, depending on the desired Seebeck coefficient and the temperature range of interest. Common metal combinations are iron-constantan, chromel-alumel, chromel-constantan, and copper-constantan. In practice, the total voltage measured by a volt meter depends on the voltage change at the hot junction, the voltage change at the cold junction, and the voltage changes at the volt meter contacts. In practical applications, thermocouples are calibrated for a 0 °C cold junction and with the thermocouple leads connected to copper leads at the voltmeter. When thermocouples are used for practical application, the water bath is eliminated, and the equivalent voltage is produced by an electronic cold junction. With an electronic cold junction, it is necessary that the two thermocouple leads both be at the same temperature as the electronic cold junction, which includes internal compensation for changes in temperature. The LT1025 is an example of a microchip that performs this function, and an example circuit for its use can be found in the device's datasheet.

8.7.2. Biopotential

Biomedical signals, including the electrocardiogram, electroencephalogram, electrooculogram, and electromyogram are generated naturally. They are picked up with special electrodes and amplified with a high gain, high impedance amplifier, such as that described in Section 4.5.10.

8.7.3. Chemical Sensors

8.7.4. Piezoelectric

A piezoelectric material induces a voltage in response to a change in stress in the material. The mechanism is illustrated in Figure 8-20. When the lattice is stressed, the two negative charges in the upper half are displaced, moving a net negative charge upward, while the two positive charges in the lower half move a net positive charge downward. Thus, the material has a larger negative charge on the upper half than on the lower half, which leads to a potential. The charge, q , induced by this potential is proportional to the force, F , on the crystal, according to

$$q = kF .$$

The parameter k is the piezoelectric constant in coulombs/newton. If the system acts like a parallel-plate capacitor, where the voltage V across the capacitor is charge q divided by capacitance C ,

$$V = \frac{q}{C} = \frac{kF}{C} = \frac{kxF}{\epsilon_0\epsilon_r A}, \quad \text{Equation 8-20}$$

where ϵ_0 is dielectric constant of free space, ϵ_r is relative dielectric constant, A is capacitor area, and x is plate separation. Typical values for k are 2.3 pC/N for quartz and 140 pC/N for barium titanate.

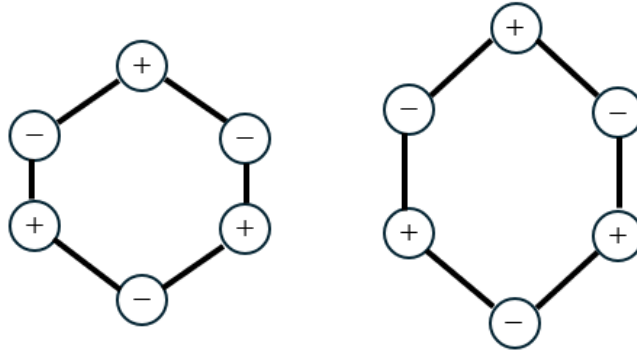


Figure 8-20: Crystal structure for a piezoelectric material under unstressed (left) and stressed (right) conditions.

The piezoelectric can also be analyzed in terms of K (Coulombs/meter) to convert displacement to charge.

Frequency Response

The frequency response of a piezoelectric crystal is shown in Figure 8-21. The curve has four regions. As frequency decreases below a lower cutoff (here, 1 Hz), the response decreases to zero. In an intermediate region, the response is flat. At a frequency above this flat region, a resonance occurs (here 1 kHz), and as frequency increases above this value, the response decreases. For applications, such as acoustic or displacement measurements, the flat region is used. For applications where the signal frequency is fixed or nearly fixed around a given value, such as buzzers or ultrasonic measurements, the resonant frequency can be exploited to provide strong signals.

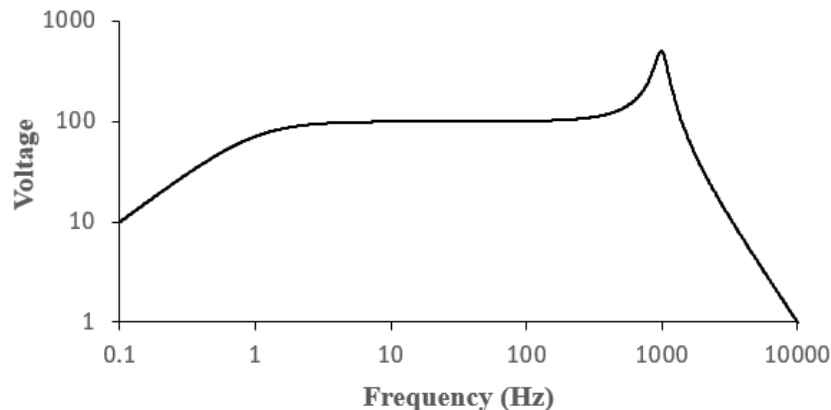


Figure 8-21: Frequency response for a piezoelectric crystal.

According to the reciprocity theorem of acoustics, the transfer function to convert pressure to voltage is the same as the transfer function to convert voltage to pressure. Thus, the curve in Figure 8-21 applies whether a signal is being transmitted from the crystal (as in the transmitter of an ultrasound probe) or received from the sound field (as in the receiver of an ultrasound probe).

An equivalent circuit for the piezoelectric crystal is shown in Figure 8-22. The current source is equal to dq/dt . The elements with subscript m are mechanical, indicating damping (R_m), spring (C_m), and inertia (L_m). These elements are responsible for the resonance at higher frequency. Element C_s holds the charge generated by the crystal. A leakage current can drain from the sensor through resistor R_s , which is extremely large, on the order of 200 G Ω .

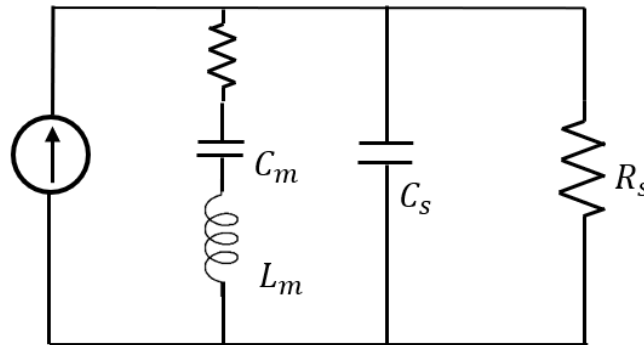


Figure 8-22: Equivalent circuit for the piezoelectric crystal.

Applications

Piezoelectric crystals can be used as both sensors and actuators. Because of the roll off at low frequency shown in Figure 8-21, they are not suited to DC applications. However, the flat region in the transfer function provides them with a good response for acoustic measurements, and the resonance is well-suited to Doppler and imaging ultrasound. They can also be used as vibration sensors for prosthetics, and they have been investigated for biological energy harvesting. When they are produced as nanoparticles, they can provide electrical stimuli to cells and tissues in response to an applied acoustic field.

Displacement-Charge Characteristic

A displacement applied to the crystal will induce a charge that slowly drains from the device with an exponential decay curve. Once all or part of the charge is drained, a release of the displacement induces a charge of equal and opposite value. I.e., if a charge of 1 nC is produced by a displacement of 1 μm and the actuation is released after the charge drains, a charge of -1 nC will be induced across the crystal.

Amplification

Two types of amplifiers can be used with piezoelectric sensors, the voltage amplifier and the charge amplifier. The voltage amplifier is simply an amplifier in the non-inverting configuration. For the charge amplifier, the crystal behaves as a current source, $i = dq/dt$. Because the inverting input of the operational amplifier is a virtual ground, the voltage across C_{eq}

and R_s must be zero, so the current i must pass through the feedback elements. If R_b is large (which it will be by design) then all of the current will pass into C_f so that

$$0 - v_{\text{out}} = \frac{1}{C_f} \int i \, dt \Rightarrow v_{\text{out}} = -\frac{1}{C_f} \int i \, dt = -iq.$$

Thus, the output voltage is proportional to the charge (hence the term “charge amplifier.” The resistor R_b is needed to feed the bias current of the operational amplifier. For optimal low frequency cutoff, this resistor will be large, and for optimal gain, C_f will be small. If, for example, $C_f = 1 \text{ nF}$ and $R_b = 10 \text{ M}\Omega$, then the low frequency cutoff will be $f_c = 1/(2\pi R_b C_f) \approx 16 \text{ Hz}$, which is adequate for the measurement of audible sound. Resistors larger than $10 \text{ M}\Omega$ are not common. Also, large values of R_b will lead to voltage offsets that are large, so it is necessary that the operational amplifier that is used have low bias current, such as one with JFET input transistors. An example is the TL062A, which has a bias current of 200 pA .

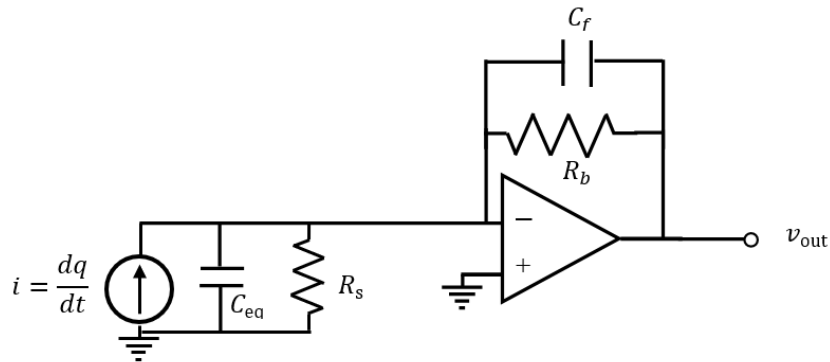


Figure 8-23: Charge amplifier for a piezoelectric sensor.

Perhaps surprisingly, if a charge amplifier is constructed on a breadboard with C_f in the tens of picofarads, removal of C_f completely will have little effect on the operation of the circuit because the connections on the breadboard itself have an internal capacitance on the order of 100 pF .

8.8 Optical Measurement

8.8.1. Photodiode and Phototransistor

Photodiodes and phototransistors are similar in function, but have distinct advantages and disadvantages. Both are used to generate a current that is controlled by a light source.

Like a standard diode, a photodiode is formed from a junction of p-type and n-type materials (refer to Section 2.3). If zero voltage is applied to the two terminals of the device, a small reverse current (from the n side to the p side) is generated, and an increased light intensity increases the magnitude of this current. The diode functions in a manner similar to a solar cell. If a voltage is applied across the device, with the n-type side positive and the p-side negative (reverse biased), the amount of current will increase with the amount of light incident on the

junction. This current is relatively insensitive to the applied voltage. The characteristic is shown in Figure 8-24.

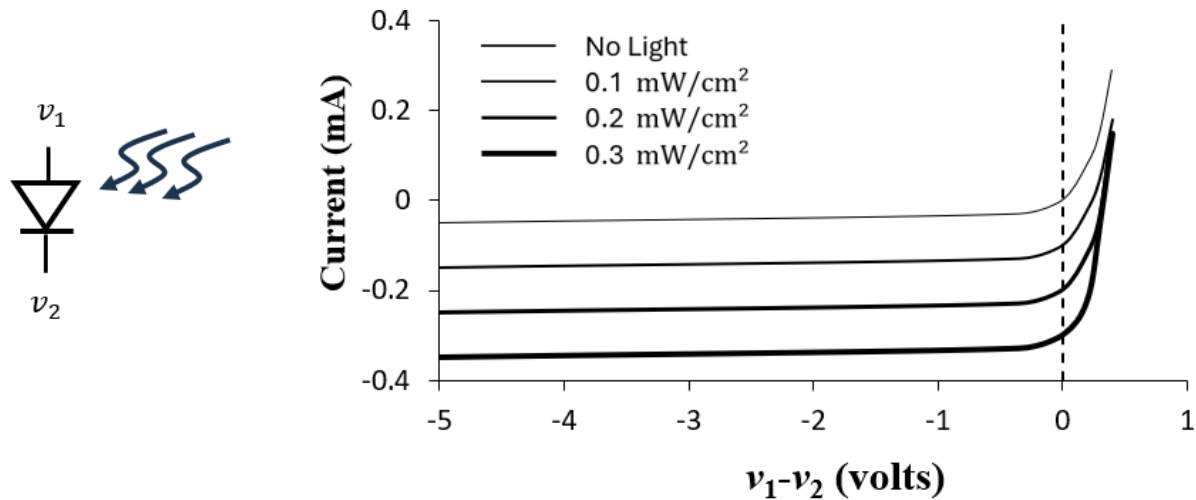


Figure 8-24: Characteristic of a photodiode.

A phototransistor has a geometry similar to that of a transistor, with either n-type sandwiched between two p-type semiconductors or p-type sandwiched between two n-type semiconductors. The light sensitive region is considered to be the center semiconductor (the transistor base).

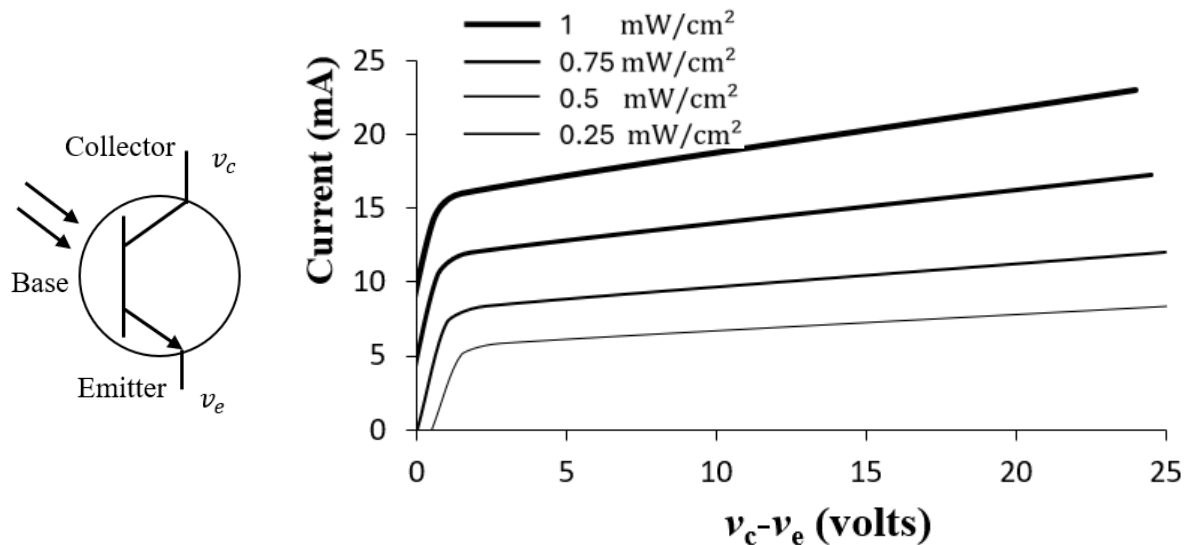


Figure 8-25: Characteristic of a phototransistor.

In general, the photodiode has a faster response time than the phototransistor, but is less sensitive. Photodiodes also have a more linear relationship between light and current, and their operating parameters are less sensitive to temperature.

8.9 Tachometers

Tachometers measure the rotation rate of an object. In a simple implementation, a location is marked on the rotating object, and a method is devised to generate a pulse whenever that location

passes by a sensor. For example, the marker can be a spot of reflective tape or paint placed on the object and a light source can be directed to the object such that light is reflected back to a sensor (e.g., a phototransistor or photodiode). The time between pulses, τ_p , is clocked, and the rotation rate is calculated as $1/\tau_p$. Alternatively, the center of the rotating object can be attached to a rotating potentiometer configured in a simple bridge (The potentiometer is a variable resistance that changes linearly with the angle of the rotating object. The constant supply voltage, V_s , is applied across a reference resistor, R , and the output voltage, V_{out} , is taken between R and the potentiometer. An edge detector is then used to determine the times at which the sharp downward transitions occur.

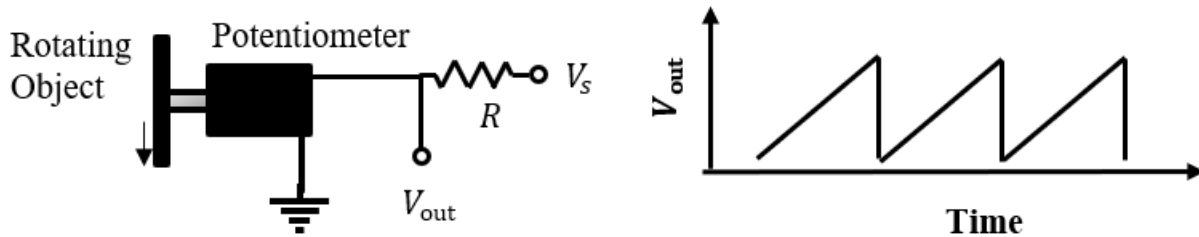


Figure 8-26). The potentiometer is a variable resistance that changes linearly with the angle of the rotating object. The constant supply voltage, V_s , is applied across a reference resistor, R , and the output voltage, V_{out} , is taken between R and the potentiometer. An edge detector is then used to determine the times at which the sharp downward transitions occur.

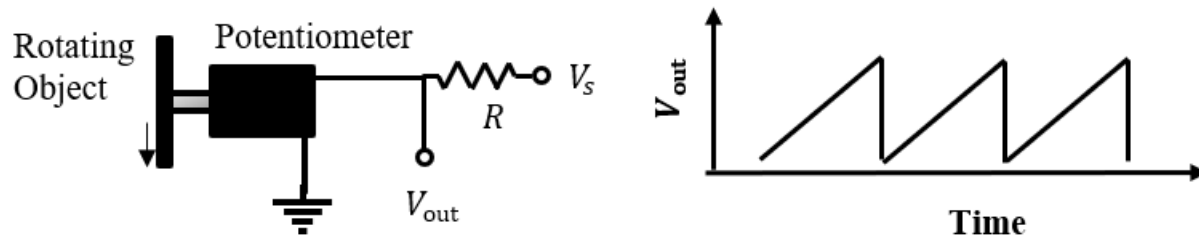


Figure 8-26: Example tachometer implementation.

8.10 Combined Concepts

The ideas in the above sections can be combined to produce application-specific sensors. If a sensor does not exist to directly convert the signal of interest to a voltage, that signal can often be converted to another energy form for which a sensor does exist or is readily available. The tachometer is a simple example, where the signal of interest, rotation rate, is converted to a light signal or to a resistance change. The activity being monitored must change some aspect of a direct signal (voltage, light intensity, heat) or an impedance (inductance, capacitance, resistance), and that change must be detected. The impedance changes will require a bridge network to convert them to a voltage. Additional circuitry will be needed to convert the changed quantity to the desired measurement.

Chapter 9: Sensor Dynamic Characteristics

Sensors, like circuits, are governed by transfer functions. These transfer functions can be zeroth order, first order, second order, or higher order. Generally, only orders zero through two are discussed because the characteristics of higher order systems are analogous to those of the lower order systems. Zeroth order systems are those that do not depend on frequency. First order systems, such as the RC circuit or thermistor, exhibit exponential decay behavior of the form $e^{-\alpha t}$. Second order systems, such as the LRC circuit or accelerometer, have both inertial (e.g., mass) and storage elements (e.g., spring) that lead to decaying oscillations of the form $e^{-\alpha t} \cos(\omega t)$.

9.1.1. Sensitivity and Specificity

The sensitivity of a sensor is the slope of the calibration curve (Figure 9-1). If the sensor is linear, sensitivity is constant. If the sensor response is nonlinear, the concept of sensitivity assumes that the variable being measured varies by a small amount, Δu , from a fixed value, u_0 . The idea should be familiar. The output is expanded as a Taylor series around the fixed point, where only the linear term is kept, so that the model for the output y is $y = y_0 + \alpha(\Delta u)$, where y_0 is the value of the output at input value u_0 , and α is the slope of the calibration curve at input value u_0 . *Sensitivity is dimensional*, converting the units of the measurand to (usually) volts. In cases where the range of the measurement is so large that this linearization will lead to excessive error, the calibration curve will need to be modeled with a nonlinear function. For example, the calibration on the right in Figure 9-1 might be modeled as $y = a_0\sqrt{u}$, with a_0 a constant. This sensor exhibits lower sensitivity as the input value increases.

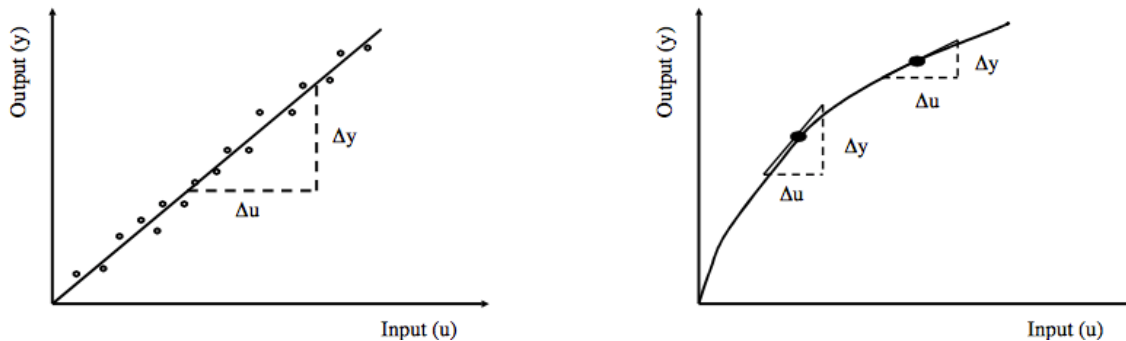


Figure 9-1: Sensitivity of a sensor. Left: A linear sensor. Right: A nonlinear sensor.

9.1.2. Drift

Drift is a change in the calibration curve that occurs slowly over time. The curve can change in both the intercept and the slope, as shown in Figure 9-2. A change in the intercept is referred to as zero drift, where all output values change by the same amount, so that the slope does not change. Sensitivity drift is a change in the slope of the calibration with time. A common cause of drift is a change in temperature of the sensor or its associated electronics, which in turn can be caused by a change in the ambient conditions or by self-heating of the device. Strategies to account for drift are to (1) let the system come to an equilibrium temperature before taking measurements or (2) recalibrate the sensor frequently and keep careful track of which calibration is associated with each set of measurements,

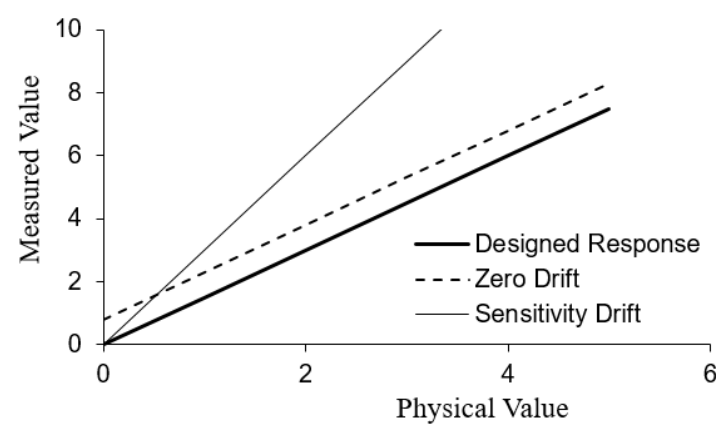


Figure 9-2: Zero drift and sensitivity drift.

A sensor's sensitivity is the smallest change in the measurand that can be detected. This value will be largely determined by the signal to noise level of the measurement. A sensor's specificity is its ability to distinguish the measurement of interest from changes in other variables. For example, a serotonin sensor with good sensitivity to that agent that responds to thromboxane may not be appropriate for a study of platelet activation, where both serotonin and thromboxane are released.

9.1.3. Signal to Noise Ratio

In analogy to the signal to noise ratio of an amplifier, the signal to noise ratio for a sensor is the ratio between the signal generated by the quantity being measured and all other factors that cause a change in the output signal. It can be expressed as a non-dimensional number, $R_{sn} = V_{measured}/V_{other}$, or as a quantity in decibels, where $R_{sn(db)} = 20 \log(V_{measured}/V_{other})$. The definition is slightly different when the quantity being measured is power or intensity, where the constant 20 is replaced by 10. As an example, if a sensor is directly measuring acoustic pressure, p_s , with noise p_n , then $R_{sn(db)} = 20 \log(p_s/p_n)$, but if it is measuring acoustic power, P_s , with power noise P_n , then $R_{sn(db)} = 10 \log(P_s/P_n)$. The reason is that power is proportional to the square of pressure, so that $10 \log(P_s/P_n) = 10 \log((p_s/p_n)^2) = 20 \log(p_s/p_n)$.

9.1.4. Linearity

The linearity of a sensor follows the discussion in Section 2.7. Specifically, if y_1 is the response of the sensor to physical variable x_1 and y_2 is the response of the system to physical variable x_2 , then the response to physical variable $x_1 + x_2$ is $y_1 + y_2$. The consequence is that the sensor has the characteristic of a straight line, as shown in the left-hand panel of Figure 9-1. If the calibration curve does not pass through the coordinates $(u = 0, y = 0)$, it is technically not linear, but the variables can be redefined easily to provide a linear relationship. For example, if the y-intercept is y_{int} , the output variable can be redefined as $\hat{y} = y - y_{int}$ so that \hat{y} is linear with u . No instrument has a perfect linear response. For example, even an operational amplifier, which we consider to be linear, will saturate when the output reaches the supply voltage, so it is linear only within a range determined by the supply voltage.

9.2 First and Second Order Behavior

The calibration curve describes the output of a system for a constant value of the measured signal. Because measured signals vary in time, one must know the frequency response of the sensor to ensure that it is sufficiently fast to capture the most rapidly changing values of the measured quantity that are of interest. As with amplifiers, the transfer function ($T(s)$) describes the dynamic behavior of the system. It is the Laplace transform of the system.

The magnitude, $|T(s)|$, is the ratio of the sensor output to the sensor input for a sinusoidal change in the input. Direct measurement of $|T(s)|$ can be relatively easy for some instruments and less straightforward for others. For example, for a microphone, it is easy to transmit a sinusoidally changing pressure at a specific frequency (e.g., by playing a note on a musical instrument), but for a thermistor, creation of a sinusoidally varying temperature is difficult. In the case of the thermistor, an easier approach is to measure the step response when the sensor is moved from a water bath at one temperature to a water bath at another temperature. A sensor's magnitude is dimensional. For the microphone, the dimensions are volts/pascal.*

The phase, defined as $\varphi = \text{atan}(\Im\{T(s)\}/\Re\{T(s)\})$, is the difference between the phase of the output and the phase of the input at a given frequency. If the output is delayed by a time t_{lag} from the input, the phase shift is negative. Phase shift is related to the time lag and the frequency as:

$$\varphi = 360^\circ \left(\frac{t_{\text{lag}}}{T} \right) = 360^\circ (t_{\text{lag}} f)$$

Again, this idea is well-familiar to you for amplifiers, but the concept also applies to sensors.

The transfer function for a sensor comes from the differential equation that describes the system. An example (for a 2nd order system) is

$$\frac{d^2 y_{\text{out}}}{dt^2} + b \frac{dy_{\text{out}}}{dt} + c y_{\text{out}} = x_{\text{in}},$$

where y is the output (usually voltage) and x_{in} is the quantity being measured. Take the Laplace transform of the equation

$$\begin{aligned} s^2 Y_{\text{out}} + b s Y_{\text{out}} + c Y_{\text{out}} &= X_{\text{in}} \\ \frac{Y_{\text{out}}}{X_{\text{in}}} &= \frac{1}{s^2 + b s + c} = T(s) \end{aligned}$$

Here we have used the well-known result from Laplace transform analysis that

$$\mathcal{L}\left\{\frac{dy}{dt}\right\} = s\mathcal{L}\{y\} - y(0); \mathcal{L}\left\{\frac{d^2 y}{dt^2}\right\} = s^2 \mathcal{L}\{y\} - s y(0) - y'(0)$$

For the transfer function, we set all of the initial conditions ($y(0)$, $y'(0)$, etc.) to zero.

If the output of one linear system is fed into another (the systems are cascaded) the transfer function of the combined system is the product of the transfer functions of the individual systems.

* However, sound is usually expressed in decibels with respect to 2×10^{-4} pascals. The reference 2×10^{-4} pascals is considered to be the threshold of human hearing, so a sound of 0 db is barely audible. The sound of a jackhammer at 1 meter distance is 2 pascals, which, expressed in decibels is $20 \log(2/2 \times 10^5) = 100$ db.

9.2.1. Zero Order

For a zero order sensor, the output does not depend on time, only on the input value. No system is truly zero order, but if the system is relatively fast compared to any possible change in the measurand, it behaves as if it is zero order. Even a component as simple as a resistor has a characteristic that changes with respect to frequency, as a result of stray capacitances and inductances that cannot be eliminated in physical devices. However, the effect of frequency is not apparent for typical resistors until the frequency reaches into the 10s of MHz.

An example of a zero-order instrument (for sufficiently low frequencies) is a linear potentiometer in which the output voltage is directly proportional to the input displacement (Figure 9-3). Here, $V_{out} = Ex/L$, where x is the distance from the negative (–) line. The measurand is L , and the output is V_{out} . The slope K is equal to E/L . No phase shift is introduced from this device, in theory, but in practice, stray inductances and capacitances will introduce a time dependence at high frequencies.

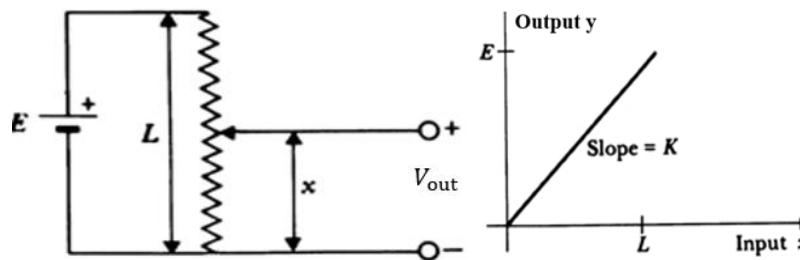


Figure 9-3: (Left) Potentiometer configured to measure position and (Right) the associated calibration curve.

9.2.2. First Order Sensors

If you move a thermocouple from a bath at T_1 to one at T_2 , the output will not change immediately because time is required for heat to warm the thermocouple. The response is similar to the charging of an RC circuit (Figure 9-4).

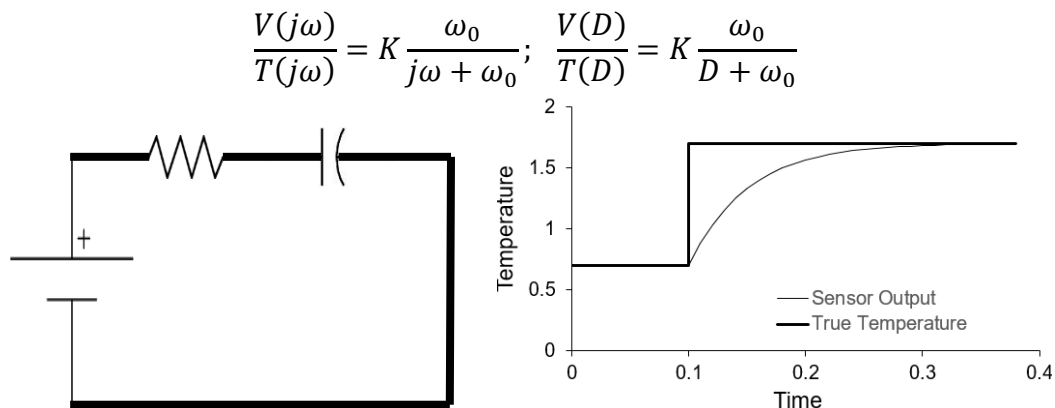


Figure 9-4: (Left) Electrical model for a thermistor. (Right) Response of the thermistor to a step change in temperature.

If an instrument contains a single energy-storage element, then it is a first order system. The equation has the form

$$\frac{dy(t)}{dt} + \alpha y(t) = x(t)$$

with initial condition $y(0) = y_0$. With zero initial condition and constant input ($x(t) = x_0$), the solution is

$$y(t) = \frac{x_0}{\alpha} (1 - e^{-\alpha t})$$

We then have a time constant of $\tau = 1/\alpha$, as indicated by the argument of the exponential.

9.2.3. Second Order Sensors

A classical example of a second order system is the mass-spring-damper system (Figure 9-5), which is mathematically equivalent to an LRC circuit. The governing differential equation for a second-order system contains second derivative terms. The differential equation is derived from a balance of forces on the mass. If the mass' position is x , increasing to the right, then inertial force (mass times acceleration) must balance the spring force, kx , and the damper force, $C dx/dt$, where k is the spring constant, and C is the damper coefficient. The function $f(t)$ is an externally applied force on the mass that may vary in time.

$$m \frac{d^2x}{dt^2} = -C \frac{dx}{dt} - kx + f(t) \Rightarrow m \frac{d^2x}{dt^2} + C \frac{dx}{dt} + kx = f(t).$$

The negative signs in the first form of this equation arise because an increase in x stretches the spring, which accelerates the mass in the negative x direction, as does motion of the damper in the positive direction.

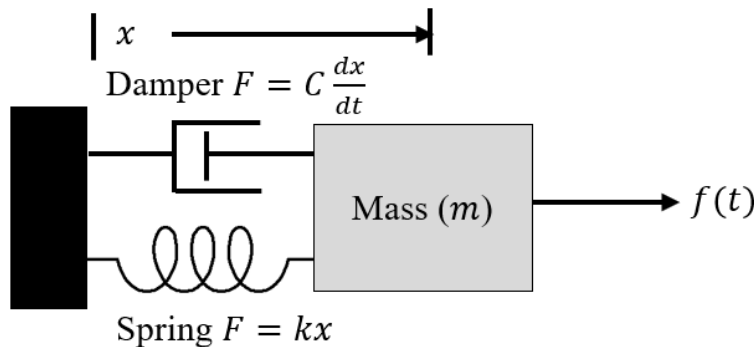


Figure 9-5: Mass-spring-damper system.

The Laplace transform approach leads to the transformed equation

$$ms^2X(s) + CsX(s) + kX(s) = F(s)$$

where $X(s)$ and $F(s)$ are the Laplace transforms of $x(t)$ and $f(t)$, respectively. The transfer function, $T(s) \equiv X(s)/F(s)$ is

$$T(s) = \frac{1}{ms^2 + Cs + k} = \frac{1/m}{s^2 + \frac{C}{m}s + \frac{k}{m}}$$

Comparison of this transfer function to the standard second order transfer function

$$T(s) = \frac{\alpha \omega_n^2}{s^2 + 2\zeta \omega_n s + \omega_n^2}$$

Demonstrates that the natural frequency is ω_n is $\sqrt{k/m}$ (based on the last term in the denominator) and the damping factor is $C/(2\omega_n m) = C/(2\sqrt{km})$. To obtain the frequency response in terms of the Fourier transform, substitute $s = j\omega$.

$$H(j\omega) = \frac{\alpha \omega_n^2}{-\omega^2 + 2\zeta \omega_n j\omega + \omega_n^2} = \frac{\alpha \omega_n^2}{(\omega_n^2 - \omega^2) + 2\zeta \omega_n j\omega}$$

The consequences of this form, specifically the peak that arises in the Bode plot magnitude from the term $\omega_n^2 - \omega^2$, was discussed in Section 2.2.1.

To find the response of this system to a unit step function $x(t) = u_s(t)$, multiply the transfer function by $\frac{1}{s}$ (the transform of $u_s(t)$) and take the inverse transform.

$$y(t) = \mathcal{L}^{-1} \left\{ \frac{1}{s} \frac{\alpha \omega_n^2}{s^2 + 2\zeta \omega_n s + \omega_n^2} \right\} = \mathcal{L}^{-1} \left\{ \frac{A_1}{s} + \frac{A_2}{s^2 + 2\zeta \omega_n s + \omega_n^2} \right\}$$

The inverse transform is evaluated through the partial fractions method, which depends on the poles of $s^2 + 2\zeta \omega_n s + \omega_n^2$.

The inverse transform of A_1/s simply gives back the step function. The inverse transform of the A_2 term is a transient that delays the output from the step function value. This inverse transform depends on the poles of $\frac{A_2}{s^2 + 2\zeta \omega_n s + \omega_n^2}$. The poles of $\frac{A_2}{(s^2 + 2\zeta \omega_n s + \omega_n^2)}$ are

$$p_{1,2} = \frac{(-b \pm \sqrt{b^2 - 4ac})}{2a} \quad (\text{binomial theorem})$$

$$p_{1,2} = -\zeta \omega_n \pm \omega_n \sqrt{\zeta^2 - 1}$$

Thus, if $\zeta > 1$, the poles are real, but if $\zeta < 1$, the poles are imaginary, which leads to an oscillatory behavior. For the mass spring system with $\zeta = C/(2\sqrt{km})$, the requirement is then $C < 2\sqrt{km}$. Physically, decreased damping and increased spring and mass lead to complex poles and underdamped behavior.

The responses of this system to a unit step function $x(t) = u_s(t)$ are

Overdamped ($\zeta > 1$)

$$y(t) = K \left[1 - e^{-\frac{\omega_n t}{2\zeta}} \right] u_s(t)$$

Critically Damped ($\zeta = 1$)

$$y(t) = K[1 - (1 + \omega_n t)e^{-\omega_n t}]u_s(t)$$

Underdamped ($\zeta < 1$)

$$y(t) = K \left[1 - \frac{1}{\sqrt{1 - \zeta^2}} e^{-\zeta \omega_n t} \cos \left(\sqrt{1 - \zeta^2} \omega_n t - \varphi \right) \right] u_s(t)$$

A common example of a second order sensor is a spring type force transducer, where force is detected from the compression of a spring, similar to a bathroom scale.

9.3 Test Signals

Regardless of the order of the instrument, standard input waveforms are used to determine its response, as shown in .

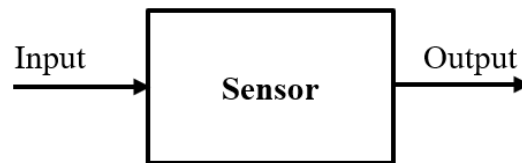


Figure 9-6: Generic setup to test a sensor.

The system characteristics of interest include the system's order (zero, first, or second) and the parameters (time constant, natural frequency, and damping factor). Standard input signals (Figure 9-7) are a step, a ramp, a pulse, and a sinusoid. A pulse input with sufficiently small duration ϵ approximates a delta function and reveals the impulse response, $h(\tau)$ of the system. You have already used the sinusoidal and step signals repeatedly in the laboratory.

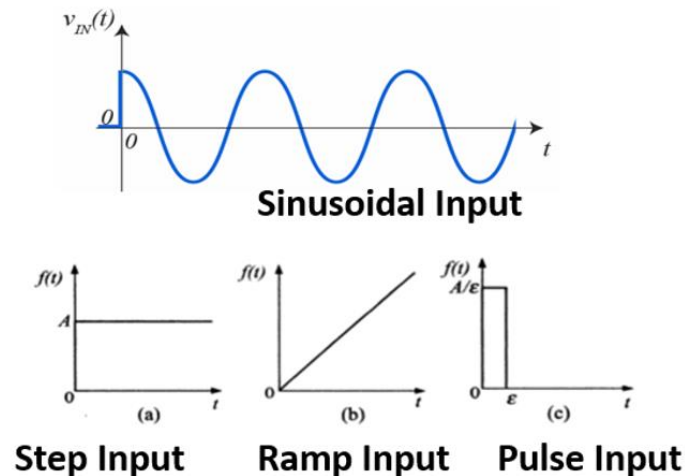


Figure 9-7: Standard test signals.

Chapter 10: Non-Ideal Operational Amplifier Characteristics

The basic derivations of operational amplifier circuits, assuming infinite input resistance and infinite gain, were described for several configurations in Chapter 4: Ideal Operational Amplifiers. These derivations provide excellent models that can be used to obtain Bode plots for filters constructed from these elements. In practice, the operational amplifiers have properties that deviate from their ideal characteristics, including finite (rather than infinite) open loop gain, finite bandwidth, non-zero output impedance, slewing (inability to change the output instantaneously), an offset voltage (a small, but finite output voltage for zero input voltage), bias current (a small constant current absorbed by the inputs), output current limitation, common mode gain, and temperature. These effects will be discussed in this chapter, along with their implications.

10.1 Finite Gain and Bandwidth

The finite gain and bandwidth of an operational amplifier were addressed earlier. A typical operational amplifier gain is on the order of 10^5 , which means that the construction of closed-loop circuits with larger gain is not possible. Similarly, most operational amplifiers are designed so that their gain decreases in proportion to $1/\omega$. Closed-loop circuits with larger gain, therefore have a narrower bandwidth because the open loop gain approaches the closed-loop gain at a lower frequency.

10.2 Stability Considerations

Any course in feedback control will discuss stability in detail. If a system has *asymptotic* stable, then the change in output caused by a finite pulse input to the system will eventually decay to zero. If it is unstable, then the output will grow exponentially until some limit is reached (such as the voltage output of an operational amplifier reaching the power supply). In between these two cases is the marginally stable system, where the output caused by a pulse input neither decays nor grows. For a system to be asymptotically stable, the real parts of every pole must be less than zero, i.e., if the transfer function is represented as

$$T(s) = \frac{(s + a_1)(s + a_2) \cdots (s + a_n)}{(s + b_1)(s + b_2) \cdots (s + b_m)},$$

the real parts of all b_k values are positive. If the real part of one or more of the b_k values is zero, but none of the real parts are negative, then the system is marginally stable. The explanation for this constraint on the b_k values is that $T(s)$ can be rewritten mathematically through a partial fraction expansion as

$$T(s) = \frac{A_1}{s + b_1} + \frac{A_2}{s + b_2} + \cdots + \frac{A_m}{s + b_m},$$

and each of these terms leads to a transient (through the inverse Laplace transform) of $A_k e^{-b_k t}$, which decays to zero if the real part of b_k is positive and increases exponentially if b_k is negative.

A separate type of stability is Bounded Input-Bounded Output (BIBO) stability, which means that whenever the input is finite, the output must be finite. A system can be asymptotically stable, but not BIBO stable, but if the system is BIBO stable, it must also be asymptotically stable.

10.2.1. Integrator

The integrator is an example of a marginally stable circuit, where the transfer function is $T(s) = 1/(sRC)$ so that the single pole is at $s = 0$. A pulse delivered to its input will cause a change in the output to a new value that will remain constant if the input returns to zero volts. (This behavior is expected from the definition of integration.) An important consequence of the marginal stability in this case is that any constant voltage, regardless of its magnitude, applied to the input will eventually cause the output magnitude to grow until the device goes to saturation. This behavior is problematic in that it complicates the testing of an integrator by itself. Regardless of how much care is taken, it is not possible to ensure that the input signal has an offset that is absolutely zero volts.

10.2.2. Differentiator

The transfer function $T(s) = sRC$ has a zero at $s = 0$ and no poles, so it is asymptotically stable. It is nonetheless BIBO unstable given that a step function input, which is bounded, leads to a delta function output, which is unbounded. One view of the differentiator is to note that the magnitude of the output increases without bound as frequency increases ($j\omega RC \rightarrow \infty$ as $\omega \rightarrow \infty$). For practical operational amplifiers, given that the open-loop bandwidth is finite, the gain must ultimately be limited. Furthermore, it is not physically possible for a frequency generator to provide a true step function; the step from one level to the other must happen over a finite amount of time. Consequently, the problem can resolve itself, but the large gain at the middle-range frequencies can nonetheless lead to saturation of the output.

10.3 Bias Current

10.3.1. Origin

Operational amplifiers are constructed from transistor elements, and these elements are nonlinear. A certain amount of current at the input is required to place them into a linear region. Therefore, operational amplifiers are designed to have a small but finite current, the bias current, pass into their input terminals, even when the voltages at these terminals are zero. This bias current is incidental for a large number of applications, but it can be problematic in some cases. For example, it is the reason why it is not possible to design an inverting amplifier where the input and feedback resistors are replaced with capacitors (Figure 10-1). Because DC current cannot pass through the capacitor, the circuit would not be able to provide bias current to the inverting input.

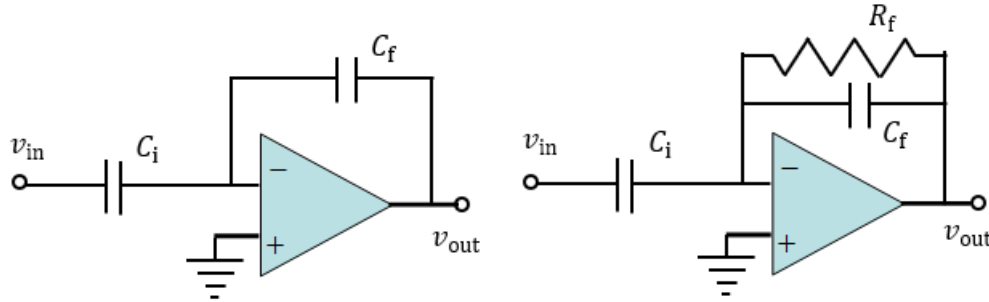


Figure 10-1: (Left) Hypothetical inverting amplifier with resistors replaced by capacitors. If the bias current were not problematic, the gain would be C_i/C_f . (Right) Inverting amplifier modified to allow bias current through R_f . The modified circuit will act as a low pass filter with a frequency cutoff of $f = 1/(2\pi R_f C_f)$, so larger values of R_f will provide wider bandwidths.

10.3.2. Compensation

The bias current will cause a voltage offset at the output of the operational amplifier. Consider an inverting amplifier configuration where the bias currents into the negative and positive inputs are both i_b . The positive input connected to ground through a compensation resistor R_c , rather than through a direct connection (Figure 10-2).

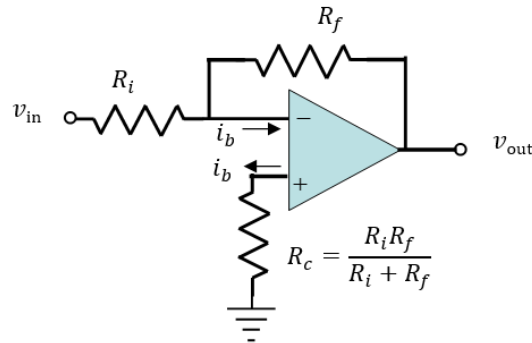


Figure 10-2: Inverting amplifier configuration with a compensation resistor connected to the positive input.

With zero input, and the positive input grounded, the open-loop gain leads to

$$v_{out} = A(v^+ - v^-) = A(i_b R_c - v^-)$$

From Kirchov's current law at the connection between R_i and R_f ,

$$\frac{0 - v^-}{R_i} - \frac{v^- - v_{out}}{R_f} = i_b \Rightarrow v^- \left(\frac{1}{R_i} + \frac{1}{R_f} \right) = -\frac{v_{out}}{R_f} + i_b \Rightarrow v^- = -\left(\frac{v_{out} R_i}{R_f + R_i} \right) + i_b \left(\frac{R_f R_i}{R_f + R_i} \right)$$

Use this result for v^- in the previous equation.

$$v_{out} = A \left(i_b R_c + \frac{v_{out} R_i}{R_f + R_i} - i_b \frac{R_f R_i}{R_f + R_i} \right)$$

$$v_{\text{out}} \left(1 - \frac{AR_i}{R_f + R_i} \right) = Ai_b \left(R_c - \frac{R_f R_i}{R_f + R_i} \right).$$

For v_{out} to be zero, the expression on the right must be zero, which can happen only if R_c is the parallel combination of R_f and R_i .

Without compensation ($R_c = 0$), the output voltage caused by the bias current is

$$v_{\text{out}} = - \frac{Ai_b R_f R_i}{R_f + R_i - AR_i}.$$

Since AR_i is typically much larger than $R_f + R_i$ (in the denominator), this expression simplifies to

$$v_{\text{out}} = i_b R_f$$

10.4 Slew Rate

If the input of an operational amplifier changes instantaneously such that the output must change from $-V_{\text{supply}}$ to $+V_{\text{supply}}$, the output change will not occur instantaneously, but over a finite amount of time. The slew rate is the change in output voltage divided by the time required for that change to happen. The datasheet for the operational amplifier will report the slew rate in volts/ μsec .

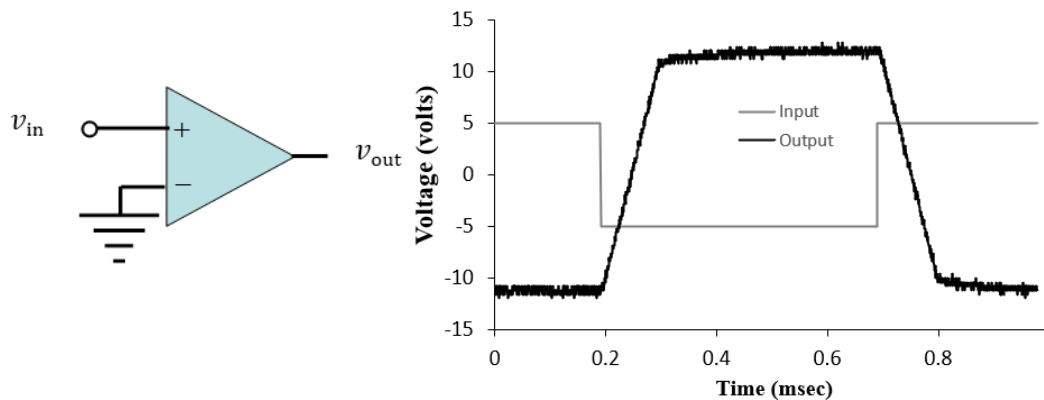


Figure 10-3: An OP07 in open-loop mode (left) and its output as a function of time (right) when the input changes from negative to positive. About 1 msec is required for the output to slew from -11 volts to +11 volts, so that the slew rate is (22 volts)/(1 msec), or 0.22 volts/ μsec .

10.5 Output Offset

Ideally if both inputs to an operational amplifier are grounded, the output should be 0 V. However, in practice, a 0 volt input causes an offset on the output. Similarly, if a finite voltage is applied to the input, the output will be equal to the gain times the input plus gain times a generally small V_{offset} . This voltage is in addition to the offset voltage caused by the bias current, as discussed in Section 10.3. The offset is caused by the op-amp's asymmetry between

the positive and negative input components, and it differs from chip to chip. The offset is most bothersome when the input voltage is only on the order of few mV. If the operational amplifier is used to generate a circuit with gain G , the circuit will have a constant value added to the output equal to $(V_{\text{offset}})G$. The AC gain is not affected.

Some operational amplifiers provide an offset adjustment. For example, to configure the OP07 for offset adjustment, connect a potentiometer (trim pot) between Pins 1 and 5. Then apply the supply voltage to the wiper arm of the potentiometer and adjust the setting of the wiper arm to eliminate the offset. The OP07 adjustment range is typically ± 4 mV.

10.6 Drift

As with sensors, operational amplifiers are subject to drift caused by temperature fluctuations. Temperature fluctuations are caused by both environmental fluctuations and the internal generation of heat by the amplifier. Low drift amplifiers are more expensive, but can be purchased if required by the application.

10.7 Input Resistance

Although the ideal model for the operational amplifier assumes that the input resistance is infinite, it has a finite value. For the OP07, it is on the order of 50 M Ω , which is adequate unless the source of the voltage to be amplified (or filtered) is has an extremely large output resistance. When the operational amplifier is placed in a closed-loop system, the overall input resistance of that system can be much greater than the input resistance of the operational amplifier chip. For a non-inverting amplifier with a resistor to ground R_g , the input impedance is

$$R_{\text{in}} = \frac{AR_g}{G},$$

where A is the operational amplifier open-loop gain and G is the non-inverting amplifier closed-loop gain.

10.8 Output Resistance

All electrical components have a finite output resistance, which is modeled as a resistor in series with the voltage source of the element, as shown in Figure 10-4. The datasheet for the OP07 lists a typical open-loop output resistance of 75 Ω . However, when the operational amplifier is placed in a closed-loop system, the output impedance of the overall circuit is typically much less. For the non-inverting amplifier configuration, for example, the closed-loop output resistance is

$$R_{\text{cl}} = \frac{R_{\text{out}}(R_f + R_g)}{AR_g} = \frac{R_{\text{out}}G}{A},$$

where A is the operational amplifier gain and G is the closed-loop gain. Thus, if the closed-loop gain is 100 and A is 10^5 , the closed-loop output resistance is only 0.075 Ω , so that output resistance is typically not a limitation of an operational amplifier.

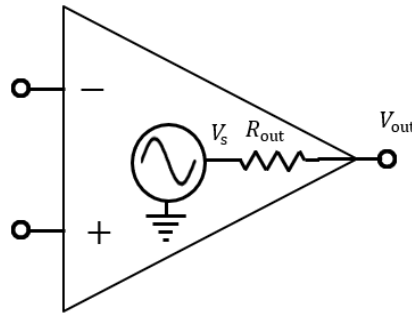


Figure 10-4: Schematic model of an operational amplifier output impedance.

10.9 Current Limitation

Operational amplifiers are limited by the amount of current they can supply, and this current limitation differs from the limitations caused by output impedance. The maximum amount of current that can be supplied by an OP07 is on the order of 20 mA. Consider a non-inverting amplifier with a gain of 100 driving a $10\ \Omega$ load resistor. The load resistor is much larger than the output impedance of the closed loop system, as seen from the calculation in Section 10.8. The expected current through the load resistor for a 1 volt output voltage is then $(1\text{ volt})/(10\ \Omega) = 0.1\text{ A}$. In practice, that current would be limited to about 20 mA.

Specialized operational amplifiers are available that can output much larger currents. An example is the LM657T, which has a maximum current output of 3 A

10.10 Common Mode Rejection Ratio

Common mode rejection ratio was described in terms of a differential amplifier in Section 4.5.8. It is noted here that the individual operational amplifier chip will also have a common mode rejection ratio. This parameter will generally be large. The data sheet for an OP07 specifies that the common mode rejection ratio is at least 97 db and is typically 120 db. Thus, the gain to the difference between the positive and negative inputs will be about five orders of magnitude larger than the gain to the average input.

10.11 Specialized Operational Amplifiers

While all operational amplifiers have the limitations described in this chapter, specialized amplifiers exist that have capabilities beyond those of the OP07. Such capabilities include wide bandwidth, extending into the 10s of MHz, high output power, extremely low input bias current (field effect-based devices), and low offset. If a specific performance requirement is needed for one of these devices, it will probably not be difficult to find, either through a web search or through direct contact with an electronics supply company representative.

Chapter 11: More Complex Analog Circuits

11.1 Linear Filters

11.1.1. Notch

The notch filter removes signal at a specific frequency. A practical notch filter will also attenuate signal at frequencies near the notch frequency. The simplest notch filter has the transfer function

$$H(j\omega) = \frac{(j\omega)^2 + \omega_n^2}{(j\omega)^2 + 2j\zeta\omega_0\omega + \omega_0^2}.$$

The j^2 components in the numerator and denominator become -1 , so that the transfer function simplifies to

$$H(j\omega) = \frac{\omega_n^2 - \omega^2}{\omega_0^2 - \omega^2 + 2j\zeta\omega_0\omega}.$$

The notch occurs because an ω value of ω_n causes the numerator to become zero. The magnitude of this transfer function is shown in Figure 11-1. In the figure, the notch frequency is reduced by only a factor of about 100 because the transfer function is not calculated at exactly 60 Hz; the zero value at exactly 60 Hz could not be shown on a log-log plot. The -3 db limits for the notch are 51.2 and 70.3 Hz, so signal within this range will be significantly distorted.

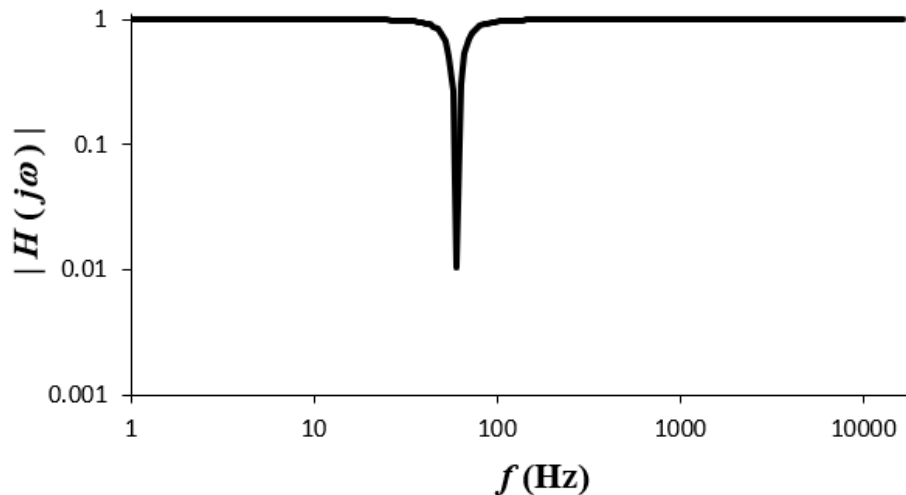


Figure 11-1: Magnitude of the notch filter transfer function with a notch at 60 Hz.

11.1.1.2. Pre-Whitening

Frequently, the Fourier power of a signal tends to decrease as frequency increases, a trend that is referred to as pink noise.³⁸ This effect is observed in the measurement of turbulent flow fluctuations (fluctuations in velocity with time). The lower frequency fluctuations are caused by the larger eddies; because the eddies are large and pass by the measuring probe with the mean velocity of the flow, they require a longer time to pass the probe and hence correspond to smaller frequencies. Because they are large, their signal is large. The trend of decreasing amplitude with frequency is also observed physiologically with the electromyogram, heart sounds, bruits (turbulence generated by an arterial stenosis, and the electroencephalogram).

If a pink noise signal is digitized, the larger low frequencies are recorded with good resolution, but the smaller high frequencies are not. A solution to this problem is to pass the signal through a filter that accentuates the higher frequency amplitudes before the signal is digitized. The signal is then digitally filtered in the opposite direction to restore its original shape. Because the filter output is a signal with a spectral power that is relatively constant with frequency, it is referred to as a pre-whitening filter.

A differentiator, as described in Section 4.5.5, is a form of pre-whitening filter that increases amplitude in proportion to ω . However, it leads to a gain of 0 for the DC part of the signal, and it overly amplifies signal components at large frequencies, particularly those beyond the Nyquist frequency. Therefore, a pre-whitening filter is more likely to have the transfer function

$$H(j\omega) = \frac{j\omega + \omega_1}{j\omega + \omega_2},$$

where $\omega_1 < \omega_2$. This transfer function is relatively flat for $\omega < \omega_1$ so that low frequency components are not removed, rises in proportion to ω between ω_1 and ω_2 , and becomes flat again after $\omega = \omega_2$. It must be coupled with an anti-aliasing filter to remove components beyond the Nyquist frequency.

³⁸ The terminology arises in analogy to the visual spectrum, where red corresponds to low frequencies and violet corresponds to higher frequencies. White light includes all colors, so it has a flat spectrum. Pink contains mostly red, but includes lesser amounts of the higher frequency colors that lead to a color that is more pastel red than bright red.

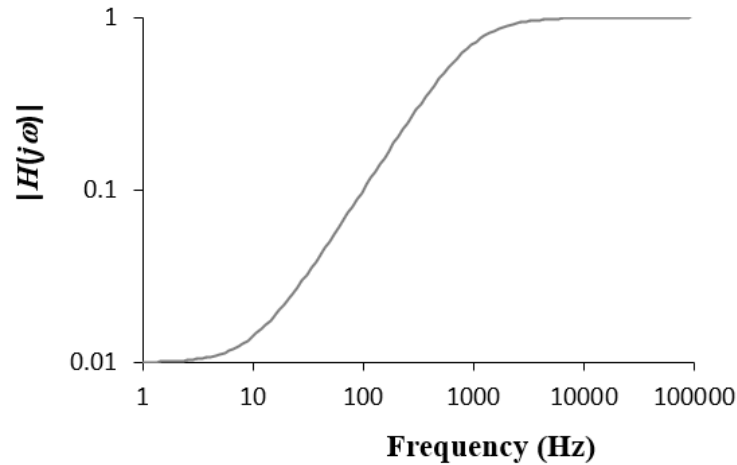


Figure 11-2: Transfer function for a single pole pre-whitening filter. Between 10 Hz and 1000 Hz, the filter enhances the higher frequency components of the signal.

11.1.3. Butterworth Filters

Transfer function

The Butterworth filter is particularly useful because it has a highly flat pass band and a rapid rolloff. Thus, it approximates the ideal bandpass filter (brick wall filter). The Butterworth formulation can be used to generate low pass, high pass, and bandpass filters, but the discussion here focuses on the low pass version. (To obtain the high pass version, replace ω/ω_0 with ω_0/ω .) The transfer function for the low pass filter is defined by the magnitude squared of its transfer function.

$$|H(j\omega)|^2 = \frac{1}{1 + \left(\frac{\omega}{\omega_0}\right)^{2n}}, \quad \text{Equation 11-1}$$

where n is the order of the filter. Since $|H(j\omega)|^2 = H(j\omega)H^*(j\omega)$, with the $*$ indicating complex conjugate, one can deduce that the transfer function

$$H(j\omega) = \frac{1}{1 + j\left(\frac{\omega}{\omega_0}\right)^n},$$

will have the required magnitude, for

$$H^*(j\omega) = \frac{1}{1 - j\left(\frac{\omega}{\omega_0}\right)^n},$$

so that

$$H(j\omega)H^*(j\omega) = \left(\frac{1}{1 + j\left(\frac{\omega}{\omega_0}\right)^n}\right) \left(\frac{1}{1 - j\left(\frac{\omega}{\omega_0}\right)^n}\right) = \frac{1}{1 + \left(\frac{\omega}{\omega_0}\right)^{2n}}$$

as required. Since the denominator of this expression is a polynomial, it can be rewritten in terms of its roots. Therefore, it is possible to write

$$|H(j\omega)|^2 = \frac{1}{(w - p_1)(w - p_2)(w - p_3) \cdots (w - p_{2n})}, \quad \text{Equation 11-2}$$

where $w \equiv \omega/\omega_0$, and the p_i are the values of w at which the transfer function would go to infinity. (Note: We are not concerned that the transfer function goes to infinity at these values of w because the values will be complex and w will always be real for a physical filter).

The poles, p_i , are the roots of $1 + w^{2n} = 0$, so they are the values of w such that $w^{2n} = -1$, i.e. of $w = \sqrt[2n]{-1}$. In other words $p_i = \sqrt[2n]{-1}$. To determine these roots, we can rewrite -1 as $e^{j(1+2k)\pi}$. The validity of this representation can be verified from Euler's rule:

$$e^{j(1+2k)\pi} = \cos((1+2k)\pi) + j \sin((1+2k)\pi) = -1 + 0j = -1.$$

The $2n^{\text{th}}$ roots of -1 are therefore

$$\sqrt[2n]{e^{j(1+2k)\pi}} = (e^{j(1+2k)\pi})^{\frac{1}{2n}} = e^{\frac{j(1+2k)\pi}{2n}} \quad \text{Equation 11-3}$$

For example, if $n = 2$, then $2n = 4$ and the roots are:

$$\begin{aligned} \sqrt[4]{-1} &= e^{\frac{j(1+2k)\pi}{4}} = \left(e^{\frac{j\pi}{4}}, e^{\frac{3j\pi}{4}}, e^{\frac{5j\pi}{4}}, e^{\frac{7j\pi}{4}} \right) \\ &= \left(\frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}j, -\frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}j, -\frac{1}{\sqrt{2}} - \frac{1}{\sqrt{2}}j, \frac{1}{\sqrt{2}} - \frac{1}{\sqrt{2}}j \right) \end{aligned}$$

Therefore, the square of the transfer function can be written as:

$$|H(jw)|^2 = \frac{1}{\left(w - \left(\frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}j \right) \right) \left(w - \left(-\frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}j \right) \right) \left(w - \left(-\frac{1}{\sqrt{2}} - \frac{1}{\sqrt{2}}j \right) \right) \left(w - \left(\frac{1}{\sqrt{2}} - \frac{1}{\sqrt{2}}j \right) \right)} \quad \text{Equation 11-4}$$

Equation 11-4 can be regrouped slightly as

$$H(jw)H^*(jw) = \frac{1}{\underbrace{\left(w - \left(\frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}j \right) \right) \left(w - \left(-\frac{1}{\sqrt{2}} - \frac{1}{\sqrt{2}}j \right) \right)}_{\text{Term 1}} \underbrace{\left(w - \left(-\frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}j \right) \right) \left(w - \left(\frac{1}{\sqrt{2}} - \frac{1}{\sqrt{2}}j \right) \right)}_{\text{Term 2}}}, \quad \text{Equation 11-5}$$

where $H^*(jw)$ is the complex conjugate of $H(jw)$. But notice that Term 2 in Equation 5 is the complex conjugate of Term 1. Therefore, $H(jw)$ itself can be written as:

$$H(j\omega) = \frac{1}{\left(w - \left(\frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}j\right)\right)\left(w - \left(-\frac{1}{\sqrt{2}} - \frac{1}{\sqrt{2}}j\right)\right)} \quad \text{Equation 11-6}$$

The decomposition of Equation 11-6 will be possible for all orders n . Specifically, the first n roots from Equation 3 can be selected to completely define the transfer function. These roots will be the ones that have a positive imaginary part. Equation 6 can also be written in terms of the Laplace parameter s , where $s = j\omega$, i.e. $\omega = -js$.

$$H(j\omega) = \frac{-1}{\left(s - \left(\frac{1}{\sqrt{2}}j - \frac{1}{\sqrt{2}}\right)\right)\left(s - \left(-\frac{1}{\sqrt{2}}j - \frac{1}{\sqrt{2}}\right)\right)}. \quad \text{Equation 11-7}$$

The poles will also always be either complex conjugate pairs, or the real value $s = -1$ (which will appear whenever n is odd). Those poles that are complex conjugate pairs can be grouped together as follows. Let $p_i = -a + jb$ and $p_{i+1} = -a - jb$ (in Equation 11-7, $p_1 = -\frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}j$, and $p_2 = -\frac{1}{\sqrt{2}} - \frac{1}{\sqrt{2}}j$). Then

$$(s - p_i)(s - p_{i+1}) = (s - (-a + jb))(s - (-a - jb)) \quad \text{Equation 11-8}$$

$$\begin{aligned} &= s^2 + 2as + a^2 + b^2 \\ &= s^2 + 2\zeta\omega_0s + \omega_0^2 \end{aligned} \quad \text{Equation 11-9}$$

where

$$\begin{aligned} \omega_0^2 &\equiv a^2 + b^2 \\ a &\equiv \zeta\omega_0 \Rightarrow \zeta = \frac{a}{\omega_0} = \frac{a}{\sqrt{a^2 + b^2}} \end{aligned}$$

In terms of $\omega = js$, the result of Equation 11-9 is $-\omega^2 + 2\zeta\omega_0\omega + a^2 + b^2$. This form of a denominator represents a second order system with characteristic frequency $\omega_0 = \sqrt{a^2 + b^2}$ and damping factor $\zeta = a/\sqrt{a^2 + b^2}$. However, because $a + jb$ is a $2n^{\text{th}}$ root of -1 , the magnitude of $a + jb$ is $|a + jb| = \sqrt{a^2 + b^2} = |e^{j\pi(1+2k)}| = 1$. Therefore, $\omega_0 = 1$ and $\zeta = a$. It follows that, given the real and imaginary parts of the poles, the n^{th} order Butterworth filter transfer function can be constructed from n second order transfer functions, each of which has $\omega_0 = 1$ and ζ the negative real part of the pole.

Translating the Transfer Function into a Circuit

If an n^{th} order Butterworth filter is to be designed, the first n $2n^{\text{th}}$ roots of -1 must first be obtained. These roots will have the form

$$\begin{array}{lll}
w_{1,2} = \pm\alpha_1 + j\beta_1 & \Rightarrow & s_{1,2} = -\beta_1 \pm j\alpha_1 \\
w_{3,4} = \pm\alpha_2 + j\beta_2 & \Rightarrow & s_{3,4} = -\beta_2 \pm j\alpha_2 \\
\vdots & \vdots & \vdots \\
w_{m-1,m} = \pm\alpha_{n/2} + j\beta_{m/2} & \Rightarrow & s_{m-1,m} = -\beta_{m/2} \pm j\alpha_{m/2} \\
(w_n = j) & \Rightarrow & (s_n = -1)
\end{array}$$

where the last root, j , will be present when n is odd, in which case $m = n - 1$ (otherwise $m = n$). Each pair of poles can be used to form a transfer function of the form:

$$H(jw) = \frac{1}{s^2 + 2\zeta s + 1}$$

The transfer function for the Sallen and Key circuit is:

$$H(s) = \frac{\omega_0^2}{s^2 + 2\zeta\omega_0 s + \omega_0^2}, \text{ where } \omega_0 = \frac{1}{\sqrt{R_1 R_2 C_1 C_2}} \text{ and } \zeta = \frac{1}{2} \frac{C_1(R_1 + R_2)}{\sqrt{R_1 R_2 C_1 C_2}}$$

If $w = \omega/\omega_0$, then

$$H(s) = \frac{1}{s^2 + 2\zeta s + 1} \quad \text{Equation 11-10}$$

and

$$\zeta = \frac{1}{2} C_1(R_1 + R_2) \quad \text{Equation 11-11}$$

Therefore, we can find two resistors and one capacitor that provide the correct value of ζ . Next, choose the value for C_2 that provides the resulting frequency $\omega_0 = \frac{1}{\sqrt{R_1 R_2 C_1 C_2}} = 1$. This frequency will need to be scaled to obtain the correct cutoff frequency for the filter. You can multiply the values of either both capacitors or both resistors to obtain a new cutoff frequency, and you will not change the value of ζ .

Example:

Consider a 4th order Butterworth filter with a cutoff frequency of 12 kHz. The first four 8th roots of -1 can be found with the following Matlab code:

First, define the 8th order polynomial $p = \omega^8 + 0\omega^7 + 0\omega^6 + 0\omega^5 + 0\omega^4 + 0\omega^3 + 0\omega^2 + 0\omega + 1$

```
p = [1 0 0 0 0 0 0 0 1]; % Defines the 8th order polynomial
```

```
% Find the roots of this polynomial
r = roots(p);
```

Matlab will find the following roots:

$$\begin{aligned}
& -0.924 + 0.383i \\
& -0.924 - 0.3837i \\
& -0.383 + 0.924i \\
& -0.383 - 0.924i \\
& 0.383 + 0.924i \\
& 0.383 - 0.924i \\
& 0.924 + 0.383i \\
& 0.924 - 0.383i
\end{aligned}$$

We will select the four of these roots with positive imaginary parts to assure stability. Each root will contribute a term $\omega - r_i$ to the denominator of the transfer function, so in terms of ω , the transfer function is:

$$H(j\omega) = \frac{1}{(\omega - (0.924 + 0.383j))(\omega - (-0.924 + 0.383j))(\omega - (0.383 + 0.924j))(\omega - (-0.383 + 0.924j))}$$

We can multiply the denominator by $1 = j \times j \times j \times j$ (i.e., multiply each pole by j , and recognize that $s = j\omega$).

$$H(s) = \frac{1}{\underbrace{(s + 0.924j + 0.383)(s - 0.924j + 0.383)}_{\text{Complex Conjugate Pair 1}} \underbrace{(s + 0.383j + 0.924)(s - 0.383j + 0.924)}_{\text{Complex Conjugate Pair 2}}}$$

Collect the poles that are now complex conjugate pairs into second order polynomials in s . For example, Complex Conjugate Pair 1 becomes:

$$\begin{aligned}
& (s + 0.9239j + 0.3827)(s - 0.9239j + 0.3827) \\
& = s^2 + (0.9239j + 0.3827)s + (-0.9239j + 0.3827)s + (0.9239^2 + 0.3827^2) \\
& = s^2 + 2(0.3827)s + 1 \\
& = s^2 + 0.7654s + 1
\end{aligned}$$

and complex Conjugate Pair 2 becomes $s^2 + 1.8478s + 1$. Therefore, we are interested in implementing the following transfer function with the Sallen and Key circuit.

$$H(s) = \frac{1}{(s^2 + 0.7654s + 1)(s^2 + 1.848s + 1)} \quad \text{Equation 11-12}$$

For the first second-order polynomial, we need $\zeta = 0.7654/2 = 0.3821$. As a start, assume that the two resistors, R_1 and R_2 are 10 k Ω . Then from Eq. 8:

$$\zeta = \frac{1}{2} C_1 (R_1 + R_2) \Rightarrow C_1 = \frac{2\zeta}{R_1 + R_2} = \frac{2\zeta}{2 \times 10^4} = \zeta \times 10^{-4} = 3.821 \times 10^{-5}$$

To find the correct value of C_2 , we must first find that value which gives an ω_0 of 1. Therefore:

$$\begin{aligned} \omega_0 &= \frac{1}{\sqrt{R_1 R_2 C_1 C_2}} = \frac{1}{\sqrt{10^4 \times 10^4 \times 3.821 \times 10^{-5} \times C_2}} = 1 \\ \Rightarrow C_2 &= \frac{1}{10^4 \times 10^4 \times 3.821 \times 10^{-5}} = 2.618 \times 10^{-4} \end{aligned}$$

Now the 12 kHz cutoff frequency can be implemented by rescaling ω_0 . If we multiply both capacitors by a the same factor, c , we will not change the damping factor, of the Sallen in Key circuit because:

$$\zeta_{s\&k} = \frac{C_1(R_1 + R_2)}{2\sqrt{R_1 R_2 C_1 C_2}} \frac{c}{c} = \frac{(cC_1)(R_1 + R_2)}{2\sqrt{R_1 R_2 (cC_1)(cC_2)}}$$

A 12 kHz cutoff frequency corresponds to $\omega_0 = 2\pi(12000) = 75,398 \text{ rad/s}$. Therefore, we need to divide both capacitor values by 75,398. Consequently, the design parameters for the Sallen and Key circuit are:

$$R_1 = R_2 = 10 \text{ k}\Omega$$

$$C_1 = 3.821 \times 10^{-5} / 75,398 = 5.07 \times 10^{-10} = 0.507 \text{ nF}$$

$$C_2 = 2.618 \times 10^{-4} / 75,398 = 3.47 \times 10^{-9} = 3.47 \text{ nF}$$

Exercise: Find appropriate resistance and capacitance values for the other 2nd order polynomial of Eq. 10.

Chapter 12: Biopotentials

12.1 Cell Membranes

Cell membranes are made of a lipid bilayer. A lipid is a construct that consists of a hydrophilic head and a hydrophobic tail (Figure 12-1). The structure is further complicated by proteins, channels, and other molecules that act to sense the environment (receptors) and transport agents into and out of the cell. The head is made hydrophilic (water loving) through its negative charge, and the hydrophobic region has a neutral charge. Uncharged molecules easily pass through the membrane.

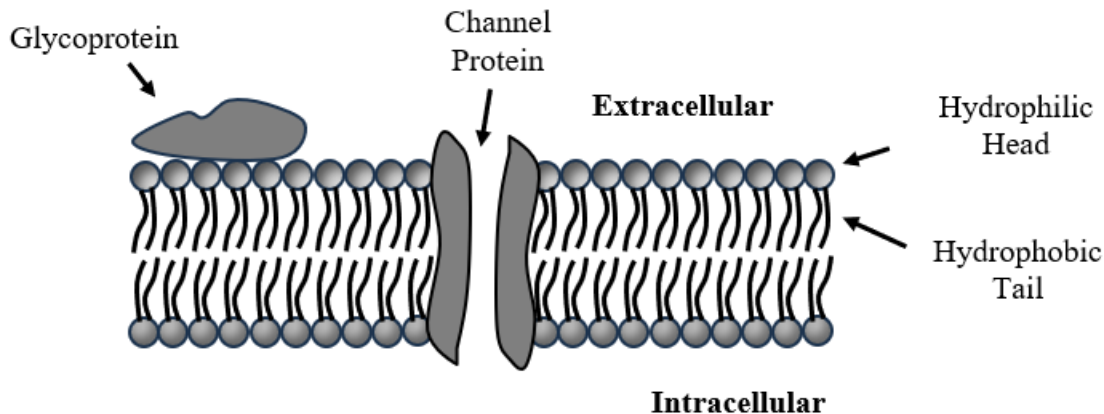


Figure 12-1: A lipid bilayer.

12.1.1. Transmembrane Transport

Transport of materials through the membrane can be either passive, requiring no energy source, or active, requiring energy in a form such as ATP, an electropotential, or light. The simplest transport mechanism is diffusion, which is driven principally by concentration and voltage differences, where the voltage differences depend on ion concentrations and valences. At equilibrium, the concentration gradients and electrical potentials balance each other.

The lipid component of the cell membrane acts as a barrier to diffusion of charged materials and water. The transport of these materials is therefore controlled by membrane structures in the form of channel proteins or carrier proteins. This transport allows concentrations of these materials to differ substantially between the extracellular and intracellular regions. The transporter proteins can operate either actively or passively.

The mechanism by which transporter proteins operate passively is referred to as facilitated diffusion. Transport must be along the diffusion gradient, from high concentration to low concentration. Facilitated diffusion circumvents the barrier to large and charged (polar) substances, thus allowing them to pass across the cell membrane. It is achieved through the operation of one of three protein types. Channel proteins span the membrane to form pores with a configuration that is specific to the properties of the diffusing substance. An example is aquaporin, which allows the passage of water molecules. Gated channel proteins open or close in

response to signals that may include the binding of the carried substance to a receptor or may be controlled by a third signal, such as a voltage change. Carrier proteins undergo a conformation change upon binding to the carried substance, and the conformational change carries the substance to the other side of the membrane.

Active transport uses carrier proteins that work against a concentration gradient. They require energy in a form such as ATP, an electropotential, or light. The proteins hydrolyze ATP (or other sources of chemical energy) to transport ions.

12.1.2. Membrane Electrical Properties

The different ion concentrations inside and outside of the cell induce a voltage difference between the intracellular and extracellular regions. The effects of this voltage depend on the electrical properties of the membrane, and these properties are summarized by a simple electrical circuit model (Figure 12-2). The resistor labeled g_{eq} represents the pores and other features that allow the passage of charged substances across the membrane. It is described in terms of its conductance instead of its resistance simply because these features are in parallel, meaning that an increase in the number of these features increases the conductance; it is easier to add conductances rather than resistances in parallel. The voltage potential, E_{eq} , is caused by the ion concentrations. The capacitance, C_m , exists because the membrane acts as a thin gap between the voltages inside and outside of the cell.

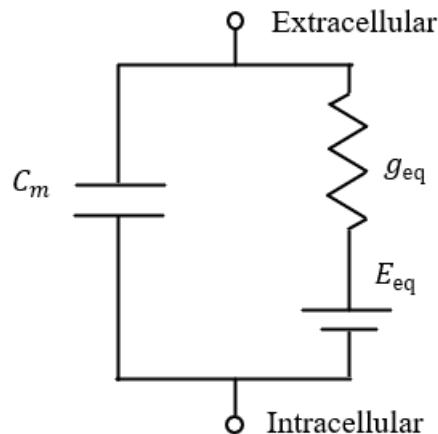


Figure 12-2: Electrical circuit model of the cell membrane

Generally, a small current flows through the equivalent conductance. The current is composed of multiple ions, such as potassium, sodium, and chloride, each of which has its own conductance. Furthermore, these conductances can change with time, controlled by the opening and closing of ion-specific channels. The most dramatic example of such a change is during an action potential (Section 0).

12.2 The Nernst Equation

The Nernst equation describes the voltage required to maintain the concentration difference across a membrane for a single ion, where the membrane has some finite permeability to that ion. The equation is

$$V = \frac{RT}{zF} \ln \left(\frac{C_o}{C_i} \right) \text{ (in volts)}$$

Where V is the voltage difference, C_o is the extracellular concentration, C_i is the intracellular concentration, R is the ideal gas constant (2 calories/mole/Kelvin), T is the temperature in Kelvins, F is Faraday's constant (2.3×10^4 calories/volt/mole), and z is the charge of the ion. At 37° C for a monovalent cation the equation reduces to

$$V = 61.5 \log \left(\frac{C_o}{C_i} \right) \text{ (in mV) .}$$

Some manipulation was used to obtain the coefficient of 61.5. First, the logarithm was changed from base e to base 10, and second, the units of voltage were changed to mV instead of volts.

The Nernst potential is not the potential **caused by** the difference in concentration outside and inside. It is the potential **required** to maintain that concentration difference, a potential generally caused by another ion that is not permeable to the cell membrane. E.g., potassium concentration is larger inside the cell than outside the cell. You would think, therefore, that all the + ions inside the cell would lead to a positive intracellular voltage (which would be true if we were dealing with the potential **caused by** the concentration difference). However, the Nernst equation is negative. The interpretation is that *maintenance of this concentration difference for potassium requires a negative intercellular potential equal to the Nernst potential.*

12.3 Goldman-Hodgkin-Katz Equation

The voltage required to maintain a given concentration profile across a membrane depends on the cross-membrane concentrations of all ions present and on the relative permeabilities of each ion. This voltage is described by the Goldman-Hodgkin-Katz Equation. If all of the ions are monovalent, the equation is

$$V_{\text{GHK}} = \frac{RT}{F} \ln \left(\frac{(\sum_m P_m [C_m])_{\text{extracellular}} + (\sum_n P_n [A_n])_{\text{intracellular}}}{(\sum_m P_m [C_m])_{\text{intracellular}} + (\sum_n P_n [A_n])_{\text{extracellular}}} \right).$$

Here, $[C_m]$ represents the concentration of the m^{th} cation and $[A_n]$ represents the concentration of the n^{th} anion while P_m and P_n represent the corresponding relative permeabilities. As an example, if sodium and potassium are the only two relevant ions, the equation is

$$V_{\text{GHK}} = \frac{RT}{F} \ln \left(\frac{P_{\text{Na}^+} [\text{Na}^+]_{\text{extracellular}} + P_{\text{K}^+} [\text{K}^+]_{\text{extracellular}}}{P_{\text{Na}^+} [\text{Na}^+]_{\text{intracellular}} + P_{\text{K}^+} [\text{K}^+]_{\text{intracellular}}} \right).$$

This form of the equation is valid only for monovalent ions. For ions with greater valence, it becomes more complicated.

The equation shows that ions with greater permeability require a larger voltage potential to maintain their concentration gradient. The value of the resting membrane potential must fall between the Nernst potentials.

It is instructive to understand why permeability is an important parameter in the Goldman-Hodgkin-Katz equation, but not in the Nernst equation. Both equations are used to determine the potential that balances the ion concentrations at steady state. For the Nernst equation, only one ion is considered. The membrane's permeability to that ion affects only the total amount of time required to reach equilibrium. Larger permeability means that equilibrium will be attained more quickly, but because the concentration difference of interest is the steady value, the amount of time required to reach steady state is irrelevant. For the Goldman-Hodgkin-Katz equation, multiple ions are passing across the membrane. Steady state is a condition where the net flux of all ions is zero. Ions with large permeability have a larger flux than ions with a smaller permeability. Therefore, when the sum of all fluxes is set to zero, these ions will contribute more strongly to the voltage required to balance fluxes, and hence to the Goldman-Hodgkin-Katz voltage.

The Goldman-Hodgkin-Katz voltage is caused by ions that are not included in the equation. These ions (such as charged proteins) are considered to have zero membrane permeability.

Important ions that affect the voltage potential between the intracellular and extracellular space are sodium, potassium, chloride, and calcium. Typical resting concentrations of these ions are listed in Table 12-1. Potassium is the ion with the largest intracellular concentration. The other ions have their largest concentrations outside of the cell.

Table 12-1: Parameters for selected ion types.

Ion	Extracellular Concentration (mM)	Intracellular Concentration (mM)	$\frac{[\text{Ion}]_o}{[\text{Ion}]_i}$	Nernst Potential at 37 °C (mV)	Relative Permeability
Na ⁺	145	12	12	+67	1
K ⁺	4	155	0.026	−98	40
Ca ²⁺	1.5	0.0001	15,000	+129	
Cl [−]	123	4.2	29	−90	20

12.4 Cellular Time Constants

A technique to examine the properties of a cell is to use a micro probe that can inject a current directly into the cell. Because the membrane can be modeled as an RC circuit, as in Figure 12-3, the intracellular voltage will take a finite amount of time to build up and will then drain from the cell. With ΔV the difference between the intracellular and extracellular voltage,

$$C_m \frac{d\Delta V}{dt} + \frac{\Delta V}{R_{eq}} = I(t).$$

Assume that a constant current of strength I_0 is injected between time $t = 0$ and $t = t_{\max}$. Then $I(t) = I_0(u_s(t) - u_s(t - t_0))$, where $u_s(t)$ is the unit step function. The solution to the

equation is composed of the familiar charging and discharging exponentials, with time constant $\tau = R_{eq}C_m$, where the charging portion increases asymptotically to a voltage difference of $\Delta V = I_0 R_{eq}/C_m$.

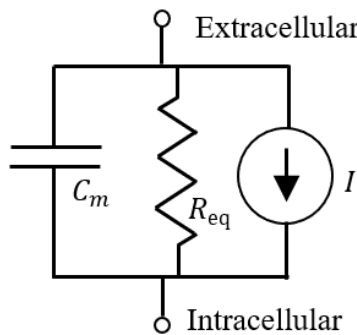


Figure 12-3: Equivalent circuit for the membrane with current injection.

When a sensory receptor on the cell membrane is activated, it generates a potential. It is generally a depolarization caused by inward current flow.

12.5 Neural Communication

Neurons transfer signals to one another through both electrical and chemical signals. Chemical communication is more common, but direct electrical signaling does occur.

12.5.1. Direct Electrical Connections

For direct electrical connections, the cell membranes are close enough to one another to allow passage of ions from one cell to the other through gap junctions. These pores can also allow other agents to pass from one cell to another. Direct electrical connections are faster than neural transmitter-based connections, but they are less specific in their effects, and they are bidirectional.

12.5.2. Neural Transmitters

Most neural communication occurs through neural transmitters that are released at the axon terminal of the presynaptic neuron and bind to receptors on the postsynaptic neuron to open or close ion channels. Neural transmitters can be excitatory (tending to increase the next cell's activity), inhibitory (tending to decrease the next cell's activity), or modulatory, which alters neuronal behavior in a general region. Modulatory neural transmitters include gamma-aminobutyric acid (GABA), which is the primary inhibitory agent, dopamine, which can be inhibitory or excitatory, depending on the receptors it binds to on the target nerves, and serotonin, which is inhibitory. Serotonin is well known in psychiatry as a modulator of mood, where selective serotonin reuptake inhibitors (SSRIs) are often prescribed to reduce the rate at which serotonin is removed from the system and hence increase the amount present. Common SSRIs include Prozac (fluoxetine), Celexa (citalopram), and Zoloft (sertraline).

12.6 Action Potentials

An action potential is a depolarization of the neuron that can trigger the release of neurotransmitters at the synapse. In the resting state, the voltage difference between the

intracellular and extracellular regions is about -70 mV. As suggested by Table 12-1, The extracellular space mostly contains sodium and chloride, while the extracellular space contains mostly potassium. If a stimulus causes this voltage to increase to about -50 mV, an action potential is initiated. The sequence of events for an action potential are shown in Figure 12-4.

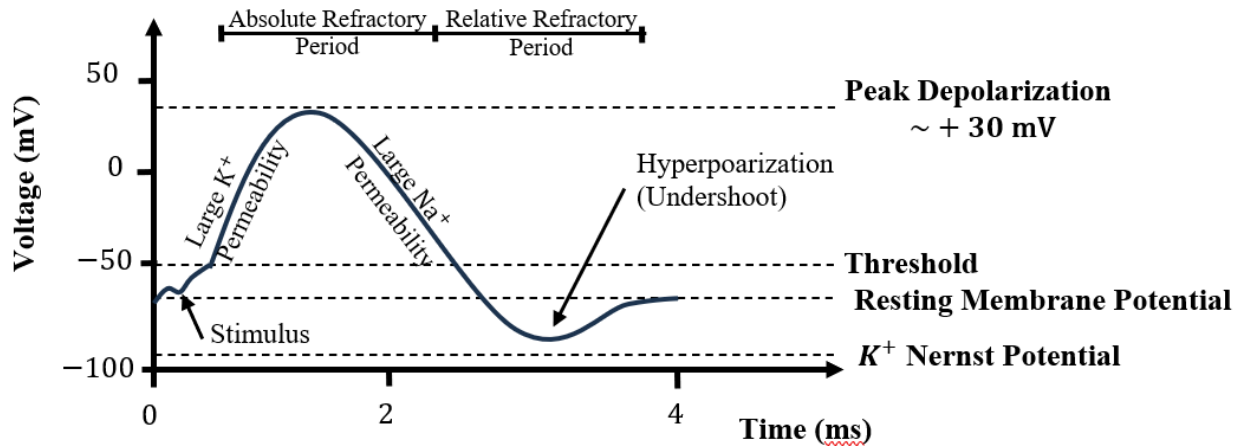


Figure 12-4: Sequence of events for an action potential.

The action potential begins when a stimulus reduces the polarization (increases the intracellular voltage). If the transmembrane potential increases to the threshold level, it causes the K^+ channels to open, which increases permeability to K^+ ions and allows them to enter the cell. Because positive ions are entering the cell, the intracellular region becomes more positive, ultimately reaching about 30 mV above the extracellular potential. At the peak, the K^+ channels begin to close, and opening of the Na^+ channels allows them to exit the cell, which repolarizes the neuron. The amount of Na^+ efflux is sufficient to hyperpolarize the cell. The total voltage change for this efflux is therefore larger than the Nernst potential for Na^+ , which seems peculiar. However, during this time, the sodium/potassium pumps are still in operation, so even before the end of the action potential, K^+ is also being removed from the cell. After hyperpolarization, the sodium/potassium pump continues the repolarization, but some leakage of K^+ occurs so that, while the pump is tending to further polarize the cell, the net result is an increase in intracellular potential back to the resting potential.

The action potential mechanism, where a small increase in intracellular potential leads to a greater increase in intracellular potential (through opening of the K^+ channels), is a classic example of positive feedback. It is combined with a negative feedback effect, where the maximum depolarization is limited by the finite concentrations of the ions and subsequent closing of the K^+ channels, which is a negative feedback effect that results in the one-shot behavior,

Two refractory periods are shown in Figure 12-4. During the absolute refractory period, no level of stimulus can generate an action potential. During the relative refractory period, an action potential can be generated but the required stimulus level is elevated.

12.7 Propagation of the Action Potential

The action potential begins at the cell body and propagates down the axon from the hillock (attachment to the cell body) to the synapse. When the cell body depolarizes, the hillock is affected first. The ionic concentration changes at that location diffuse axially, leading to a depolarization across the membrane at the adjacent axial region. This depolarization, being greater than the threshold, induces another local action potential. This process must travel in the direction from the cell body to the synapse because the refractory period will not allow it to propagate in the opposite direction.

Chapter 13: Nonlinear Circuits

Nonlinear elements are generally positive feedback elements, where output values (states) are not linearly related to the input values. A common paradigm in physiology is the combination of positive feedback with negative feedback. For these systems, not only is the relationship between input and output nonlinear, but it is also not monotonic, i.e., multiple outputs states can exist for a given set of input conditions. These latter systems generally have three possible behaviors, one-shot, change of state, and oscillation.

13.1 Monotonic Nonlinear Devices

The discussion of nonlinear circuits will begin with devices that, while nonlinear, nevertheless have a unique output for every given input.

13.1.1. Rectifier

Rectifier circuits based on either diodes or the field effect transistor have already been discussed. It is evident that these devices are nonlinear, based on the definition of linearity, where $f(a) + f(b)$ must equal $f(a + b)$. Given that the rectifier takes the absolute value of its input, we would need to have $|a| + |b| = |a + b|$, which is true if both a and b are either positive or negative, but is false if one of them is positive and the other is negative.

13.1.2. Log & Exponential

The simple diode is uniquely suited to be used in a device that takes the logarithm or exponentiates the input. Consider the operational amplifier circuit shown in Figure 13-1. Recall that the equation for current through a diode is

$$i = I_s(e^{e_0(v_1 - v_2) \eta k_B T} - 1)$$

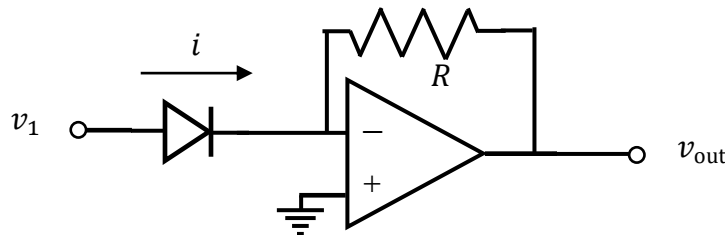


Figure 13-1: Operational amplifier configuration for an exponential functional operator.

Given that the inverting input of the operational amplifier is a virtual ground, the current through the diode is

$$i = I_s(e^{e_0 v_1 \eta k_B T} - 1) .$$

This current must pass through the feedback resistor, so that the output voltage is

$$v_{\text{out}} = I_s(e^{e_0 v_1 \eta k_B T} - 1)R = I_s R e^{e_0 v_1 \eta k_B T} - I_s R$$

The output is then the exponential of the input voltage with an offset of $-I_s R$. The term $-I_s R$ is a constant that can be subtracted easily from the output to yield only the exponential.

Because the logarithm is the inverse of the exponential, a logarithm circuit is designed similarly, with the diode placed in the feedback loop rather than the input (). In this case, the input current is v_1/R , and the feedback current is

$$i = v_1/R = I_s(e^{e_0(-v_{\text{out}}) \eta k_B T} - 1).$$

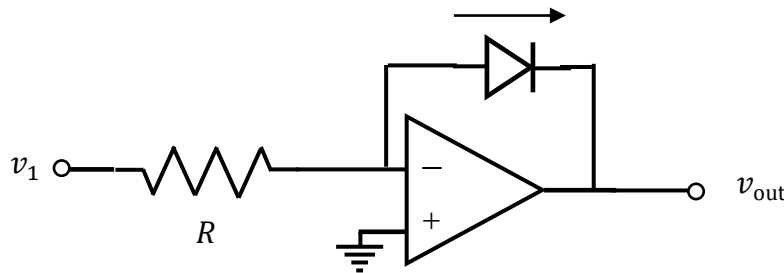


Figure 13-2: Operational amplifier configuration for a logarithmic functional operator.

Divide through by I_s and add 1 to both sides

$$\frac{v_1}{I_s R} + 1 = e^{e_0(-v_{\text{out}}) \eta k_B T}$$

Take the logarithm of both sides

$$\ln\left(\frac{v_1}{I_s R} + 1\right) = e_0(-v_{\text{out}}) \eta k_B T$$

So

$$v_{\text{out}} = -\frac{1}{e_0 \eta k_B T} \ln\left(\frac{v_1 + I_s R}{I_s}\right)$$

Here, one would need to subtract $I_s R$ from v_1 before it is presented to the input terminal.

13.1.3. Signal Multiplier

The logarithm and exponentiation circuits can be combined to design a system that multiplies two signals, as shown in Figure 13-3.

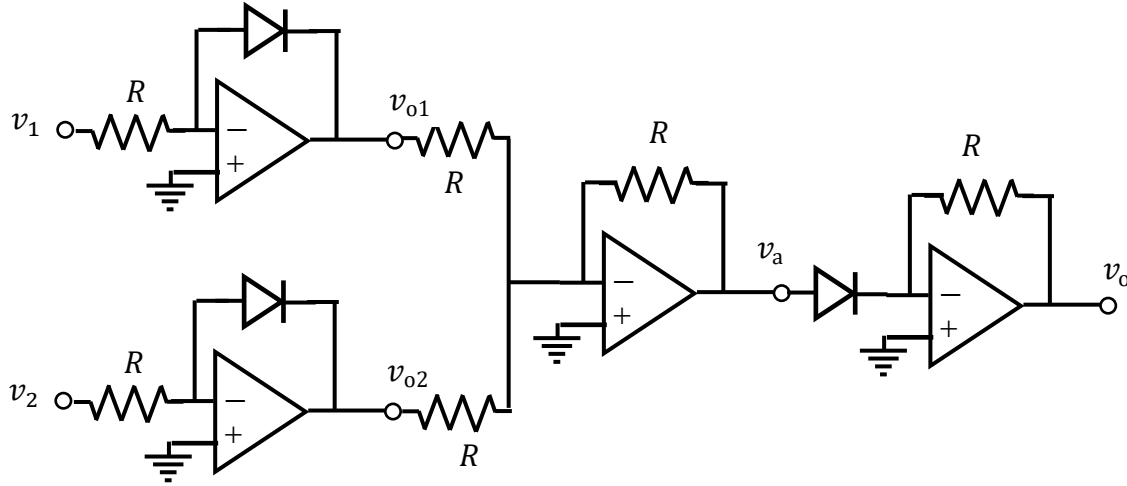


Figure 13-3: Analog signal multiplier.

The basic idea is to take the logarithm of two voltages, add the logarithms, and then exponentiate, or, in equation form,

$$xy = e^{\ln(x) + \ln(y)} = e^{\ln(xy)} = xy.$$

The outputs of the first stage (natural logarithm circuits) are

$$v_{o1} = -\frac{1}{\alpha} \ln \left(\frac{v_1}{I_s R} + 1 \right)$$

$$v_{o2} = -\frac{1}{\alpha} \ln \left(\frac{v_2}{I_s R} + 1 \right),$$

where $\alpha \equiv e_0 \eta k_B T$. The second stage sums V_{o1} and V_{o2} and multiplies by -1 .

$$v_a = - \left(-\frac{1}{\alpha} \ln \left(\frac{v_1}{I_s R} + 1 \right) + -\frac{1}{\alpha} \ln \left(\frac{v_2}{I_s R} + 1 \right) \right)$$

From the summing property of logarithms

$$v_a = \frac{1}{\alpha} \ln \left(\left(\frac{v_1}{I_s R} + 1 \right) \left(\frac{v_2}{I_s R} + 1 \right) \right)$$

The third stage (with the diode now in the input and the resistor in the feedback) multiplies its input by α and exponentiates. Thus,

$$v_o = -R I_s \left(e^{\ln \left(\left(\frac{v_1}{I_s R} + 1 \right) \left(\frac{v_2}{I_s R} + 1 \right) \right)} - 1 \right)$$

$$v_o = -R I_s \left(\left(\frac{v_1}{I_s R} + 1 \right) \left(\frac{v_2}{I_s R} + 1 \right) - 1 \right)$$

$$v_o = -\left(\frac{v_1 v_2}{I_s R} + v_1 + v_2\right)$$

Therefore, the two inputs, v_1 and v_2 , are multiplied together with a weighting of $I_s R$. The terms v_1 and v_2 are extra, but they can be easily subtracted from v_o with a pair of summing amplifiers. The first summing amplifier generates $-(v_1 + v_2)$, and the second generates $-\left(v_o + (-(v_1 + v_2))\right)$, leading to $+\frac{v_1 v_2}{I_s R}$, a scaled product. If I_s for the selected diode is 5 μA , and an R of 200 $\text{k}\Omega$ is selected, then the scale factor will be 1. Otherwise, the rescaling can be done with the final summing inverter.

13.2 Combined Positive and Negative Feedback Systems

Systems that combine positive and negative feedback generally have one of three behaviors.

1. *Change of state*: When the input changes above a threshold, the output changes to a new, constant state. Both states are stable under identical conditions, but the state that exists at any time depends on the history of the system. A simple electronic example of this type of system is the flip-flop.
2. *One-shot*: A given behavior of the input causes the output to change for a fixed amount of time and then return to its original state. E. g. the nerve action potential.
3. *Oscillation*: The interplay between positive and negative feedback causes the output to change with a periodic waveform. E.g., heartbeat.

13.2.1. One-Shot

One-shot behavior means that the system has a single stable state and whereas it can be triggered to change its state radically, any new state will be unstable and must eventually return to the stable state. The classic physiological example of one-shot behavior is the nerve action potential. After the intracellular potential increases to a threshold level, it leads to the opening of sodium channels, which allow positive ions into the cell, increasing the intracellular level further, which in turn opens more of the sodium channels. The process continues until a sodium equilibrium is reached, and the subsequent opening of potassium channels allows potassium ions to exit the cell and restores the negative potential. The nerve's original state is not reached, however, until the sodium/potassium pumps restore the original concentrations of these two ions.

13.2.2. Transition to a New Stable Level

For state transition, the system must have two or more stable states at which it can exist for a single set of input conditions. The change from living to dead is perhaps the most dramatic state change. In both the living and dead state, the environment may not have changed, but the system state is dramatically different. The lack of environmental change is key to this concept. For example, pupil contraction to light cannot be considered a state transition in this sense because it occurs only as long as the light stimulus is present.

13.2.3. Oscillation

Oscillations occur in multiple areas of physiology with oscillation frequencies that are rapid, as in the case of epilepsy, and oscillation frequencies that are much slower, such as circadian

rhythms, the menstrual cycle, and the mania-depression cyclic pattern of bipolar disorder. These patterns may be modulated by external events, but the oscillation itself is a feature that occurs without the need for a specific environmental change. Systems that oscillate freely with no driving input change have analogs in simple electronic circuits that will be described in this section.

13.3 Nonlinear System Examples

13.3.1. Schmitt Trigger

A Schmitt trigger is a single-input, single-output device that demonstrates hysteresis. The output takes on two levels, which will be designated here as V^+ and V^- , and a range of inputs exists such that the output can be either value. An example of the input-output relationship is shown in **Figure 13-4**, where V^+ is 12 volts and V^- is -12 volts. If the output is initially V^+ , it will not switch to V^- until the input is a negative value ($-V_{\text{thresh}}$), whereas if the output is initially V^- , it will not switch to V^+ until the input is a positive value (V_{thresh}). Much like the flip-flop, the output's current value depends on its history.

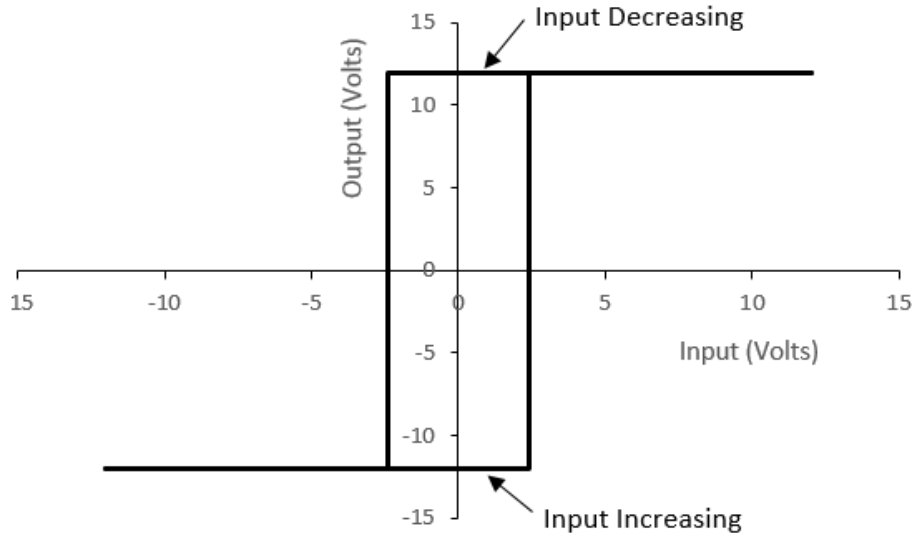


Figure 13-4: Input-output relationship for a Schmitt trigger.

An operational amplifier implementation of the Schmitt trigger is shown in **Figure 13-5**. It superficially resembles the inverting amplifier, but in contrast to that configuration, the feedback resistor connects the output with the positive input rather than to the negative input. Therefore, if the voltage (v_x) at the positive input is positive, the output will be positively saturated, and if the voltage at the positive input is negative, the output will be negatively saturated. With the assumption that saturation occurs at ± 12 volts, the positive feedback effect is readily described as follows. Assume that V_{out} and V_{in} are initially at a value that leads to $V_+ = 0$. Now let the input increase. The amplifier will cause V_{out} to also increase, which will cause V_+ to increase. This increase in V_+ will further increase V_{out} , and the process must continue until V_{out} reaches its maximum value of $+12$ volts. A similar effect occurs if V_{in} decreases from the initial point, but it leads to $V_{\text{out}} = -12$ volts. Therefore, V_{out} has only two stable values, ± 12 volts. A surprising result is that for a range of input values, the output value can be either $+12$ volts or -12 volts, depending on the history of the inputs.

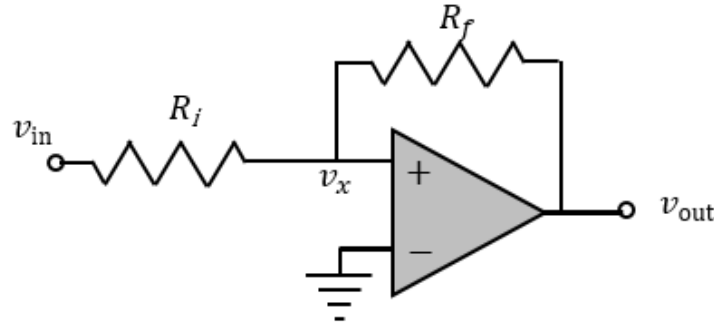


Figure 13-5: An operational amplifier-based version of a Schmitt trigger.

Assume first that both v_{out} and v_{in} are +12 volts. This configuration is consistent because v_x will also need to be +12 volts, so the positive input is larger than the negative input and the output will consequently be saturated to +12 volts. Now gradually reduce v_{in} . The value of v_x can be calculated from the knowledge that the current must be the same in the two resistors.

$$\begin{aligned}\frac{v_{in} - v_x}{R_i} &= \frac{v_x - v_{out}}{R_f} \\ v_x \left(\frac{1}{R_f} + \frac{1}{R_i} \right) &= \frac{v_{in}}{R_i} + \frac{12 \text{ volts}}{R_f} \\ \frac{v_x (R_i + R_f)}{R_f R_i} &= \frac{v_{in}}{R_i} + \frac{12 \text{ volts}}{R_f} \\ v_x &= \frac{\left(\frac{v_{in}}{R_i} + \frac{12 \text{ volts}}{R_f} \right) (R_f R_i)}{R_i + R_f}.\end{aligned}$$

The output voltage will change to -12 volts when v_x crosses 0, i.e. where

$$\left(\frac{v_{in}}{R_i} + \frac{12 \text{ volts}}{R_f} \right) = 0 \Rightarrow v_{in} = -(12 \text{ volts}) \frac{R_i}{R_f}$$

This point is stable for the system; because both the input and output are now negative, v_x must also be negative, and therefore the positive input is less than the negative input and the output must be -12 volts. From the same equations, but with v_{out} now -12 volts, the output will not switch back to positive until

$$v_{in} = +(12 \text{ volts}) \frac{R_i}{R_f}.$$

Thus, for an input between these two values, the output will be either positive or negative, depending on earlier values of v_{in} . The system therefore has the hysteresis loop that was seen in **Figure 13-4**.

13.3.2. Voltage-Controlled Oscillator

One of the three characteristic behaviors of combined positive-negative feedback systems is oscillation. The positive feedback Schmitt trigger element can be combined with negative feedback to create a device that spontaneously undergoes such oscillation. The circuit is shown in Figure 13-6.

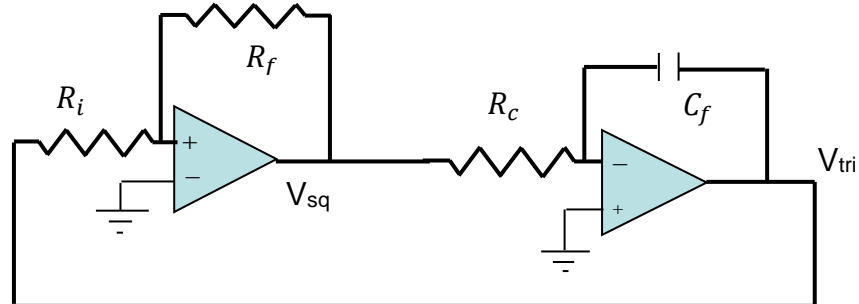


Figure 13-6: Oscillator based on a Schmitt trigger and inverting integrator.

The output V_{sq} is a square wave that switches from the negative saturation voltage, $-V_{sat}$, to the positive saturation voltage and back. V_{tri} is a triangle wave. Assume that V_{sq} is $+V_{sat}$. The output of the integrator will be $=\frac{1}{RC} \int V_{sat} dt$, and since V_{sat} is constant, the integral is $-\frac{V_{sat}}{RC} t$, which is a downward straight line with respect to time. When this voltage reaches the negative trigger point of the Schmitt trigger, V_{sq} will become $-V_{sat}$, and the integrator will begin to integrate upwards. When its output reaches the positive trigger point of the Schmitt trigger, V_{sq} becomes $+V_{sat}$, and the cycle repeats.

13.3.3. Comparator

Comparators were described in Section 4.8 because of their similarity to operational amplifiers. They are included in this section as a reminder that they are nonlinear devices that have only two possible output levels. Comparators are often used as an interface between analog and digital signals, as their inputs typically span a continuum of levels and their outputs are binary.

13.3.4. Inverter

The inverter takes a single logical input and converts it to its opposite. If the input is True, the output is False, and if the input is False, the output is True. The symbol for an inverter is shown in Figure 13-7. The circle at the output of the inverter indicates “not,” so that the inverter output can be read as “not the input.”

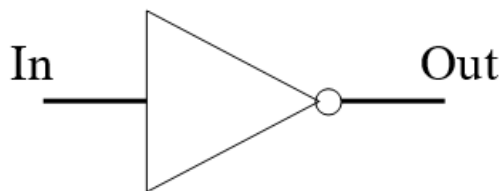


Figure 13-7: Symbol for an inverter.

13.3.5. And, Nand, Or, Nor, Xor, and Nxor Gates

Logical gates generally take two input signals and perform the indicated logical operation on them. An And gate is shown in Figure 13-8. The output is true only if both A and B are true. Otherwise it is False.

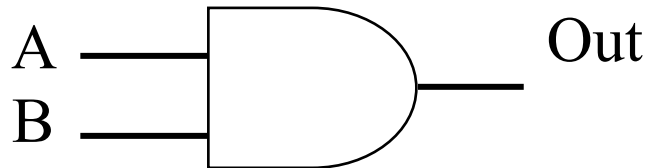


Figure 13-8: And gate

The basic gates are described by a truth table (Table 13-1: Truth table for an And gate, where the values for the inputs A and B are along the top and left-hand side of the table, respectively.). Thus, an And gate's output is True if both A and B are true and False otherwise. (For simplicity, logical values in the tables to follow will be indicated as 1 or 0.)

Table 13-1: Truth table for an And gate, where the values for the inputs A and B are along the top and left-hand side of the table, respectively.

B \ A	True (1)	False (0)
True (1)	True (1)	False (0)
False (0)	False (0)	False (0)

The Nand gate has the opposite truth table from the And gate. The output is False if A and B are True and True otherwise. The Nand gate is equivalent to the combination of an And gate followed by an inverter.

The Or gate output is True if either A is True or B is True., whereas the Nor gate, again, equivalent to an Or gate followed by an inverter, is False if either A True or B is true. Logically, then, the output of the Or gate is False if both inputs are False and True if both inputs are True. In some cases, the required function is to discern whether A and B have opposite values, i.e., one would wish the output of an Or gate to be zero if both A and B are 1. The Xor (exclusive Or) gate provides this function. The output is True if either A is True and B is False or A is False and B is True, but is zero if both A and B are False or if both A and B are True. The Nxor gate has the opposite truth table from the Xor gate. The symbols for these gates and their truth tables are provided in Figure 13-9.

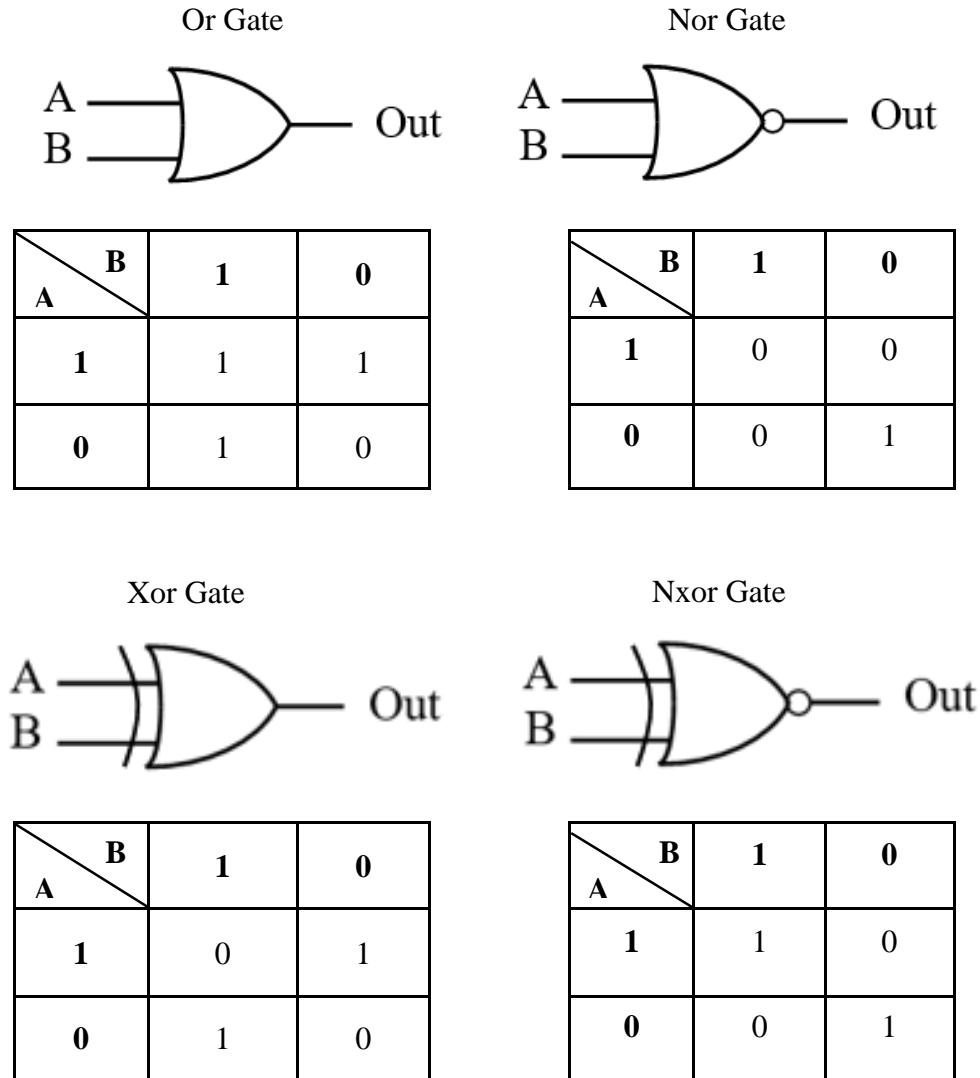


Figure 13-9: Gate symbols and truth tables

13.3.6. Flip-Flops

One of the characteristics of nonlinear devices is that a given set of inputs can lead to different outputs, depending on the history of the inputs. The flip-flop is a classical illustration of this phenomenon. A flip-flop has two inputs, but knowledge of the current value of these inputs is not sufficient to know the output level.

Figure 13-10 shows a flip-flop constructed from two Nor gates. This device is referred to as a SR flipflop, with ‘S’ for “set” and ‘R’ for “reset.” A quick analysis will show that if both *S* and *R* are 0, then *Q* can be either 1 or 0.

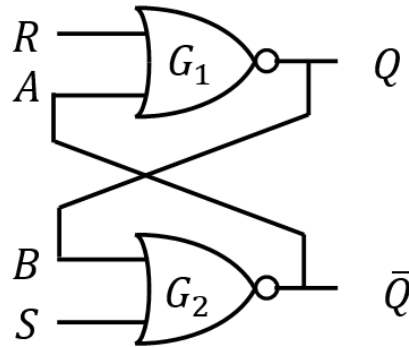


Figure 13-10: SR flip flop constructed from two Nor gates.

The truth table for this device is given in Table 13-2. Assume that $R = 1$ and $S = 0$. Because the upper device is a Nor gate, Q must be 0. (If either of the inputs to a Nor gate is 1, the output is 0.) With $B = Q = 0$, and $S = 0$, $\bar{Q} = 1$, so $B = 1$ as well. This condition is described by the third row of the truth table.

Similarly, assume that $R = 0$ and $S = 1$. Because the lower device is a Nor gate, \bar{Q} must be 0. (If either of the inputs to a Nor gate is 1, the output is 0.) With $A = \bar{Q} = 0$, and $R = 0$, $\bar{Q} = 1$, so $B = 1$ as well. This condition is described by the second row of the truth table.

Table 13-2: Logic table for a SR flip-flop.

R	S	Q_{n+1}	\bar{Q}_{n+1}
0	0	Q_n	\bar{Q}_n
0	1	1	0
1	0	0	1
1	1	Not Used	Not Used

The case where both R and S are zero behaves differently. If R and S are initially 0 and 1, respectively, the output (Q) will be 1, and if S switches to 0, the output will remain 0 because G_2 still has one input with a value of 1, and both outputs will not change. However, if R and S are 1 and 0, respectively, \bar{Q} will be 1, and when R switches to 0, the output of G_1 will still be 1, so the two outputs will not change. The output can therefore be either 0 or 1 when both R and S are zero, depending on the past input history.

Like the Schmitt trigger, the flip-flop has an output that depends not only on the current value of the inputs but also on the previous values of the inputs (i.e., it has hysteresis or memory).

The terminology “Set-Reset” is a mnemonic. If the flip-flop starts in the ambiguous state $R = S = 0$ and S is changed to 1, it “sets” the output Q to zero. If it starts in the $R = S = 0$ state and R is changed to 1, it “resets” the output Q to 1.

The condition where R and S are both 1 is not used for the following reasons. If $R = 1$ and $S = 1$, both outputs are 0, which means that the two outputs do not have the desired complementary relationship. More importantly, if both R and S are 1 and both switch to 0 simultaneously, the output state will depend on which of the two switches first. Although they switch ostensibly at the same time, one of them will do so slightly earlier than the other. If S switches first, then Q will be 1, while if R switches first, Q will be 0. Thus, because $R = 1$ and $S = 1$ leads to unpredictable results, it is not used.

Physiological Examples

Some physiological examples of responses caused by positive feedback systems are listed in Table 13-3. The responses are categorized in terms of one-shot, state transition, and oscillatory behavior. In some cases, the categorization is debatable. For example, a single beat of the heart can be considered a one-shot behavior given that, once the signal is sent to the sinoatrial node, a single ejection is triggered. However, the process is also oscillatory because it is controlled by an oscillatory system outside of the heart itself. This situation is analogous to the electronic systems that have been discussed. The Schmitt trigger, for example, is in itself a device characterized by state transition, but it becomes an oscillator when combined with the integrator.

Table 13-3: Examples of physiological responses related to combined positive and negative feedback.

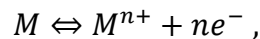
Behavior	Examples	
One-shot	Action Potential Sneezing Coughing Knee Jerk Reflex Blinking Turret Syndrome Yawning	Vomiting Peristalsis Hiccupping Blood Coagulation Ovulation Eructation Laughing
State transition	Childbirth Death Sleep/Wake Eating Ear Popping Inflammation (Autoimmune Disorder)	Cell Division Defecation Micturition Apoptosis Fainting
Oscillation	Circadian Rhythms Heartbeat Autonomous Breathing Menstruation The Cell Cycle	Stuttering Cycle of Bipolar Syndrome Parkinson's Disease Glucose-Insulin Oscillations[6]

Chapter 14: Biopotential Electrodes

Biopotential electrodes are used to collect the electrical potentials from the body related to the electrocardiogram (EKG), electroencephalogram (EEG), electromyogram (EMG), and electrooculogram (EOG). They can be considered to be transducers in that they convert current within the body in the form of ions to current in a conducting wire in the form of electrons.

14.1 Electrode interface

A chemical-to-electrical transition occurs at the electrode interface. The electrode is composed of a neutral atom, M , and it is immersed in an electrolyte that contains cations of that metal and enough anions, A^- , to maintain a neutral charge. Figure 14-1a shows the situation immediately after the electrode is placed in the electrolyte, for the case where $n = 1$. The electrode material will then oxidize according to the reaction



which is shown in Figure 14-1b. When the electrode and electrolyte materials reach equilibrium, extra cations, M^+ , are released into the electrolyte which leads to an electrical potential, which is the cause of the half-cell potential.

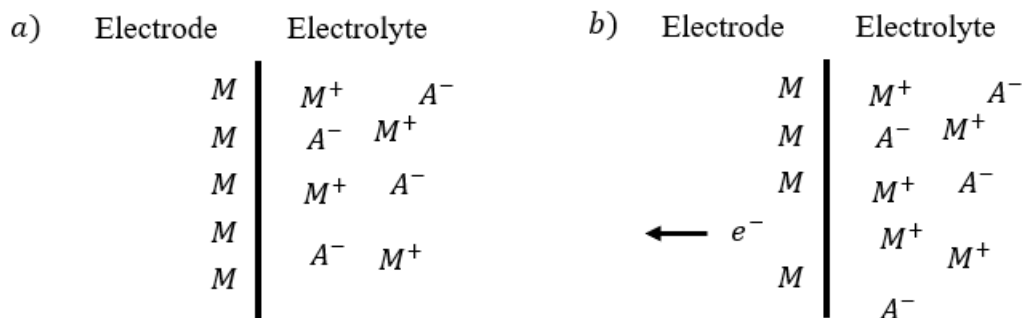


Figure 14-1: The interface between the chemical environment of the body and the electrical environment of the electrode. (a) At initial contact between the electrode and electrolyte. (b) After equilibrium.

14.2 Half-cell potential

The half-cell potential must be measured with respect to a reference electrode. The reference is arbitrarily taken to be a platinum electrode over which H_2 gas is bubbled. That is, the half-cell potential for pt/H_2 is considered to be zero volts. The half-cell potential, E , depends on ion concentration according to the Nernst equation

$$E = E^0 + \frac{RT}{nF} \ln a_{c^{n+}},$$

where E^0 is the standard half cell potential, n is the valence of the ion, $a_{c^{n+}}$ is the activity of cation c^{n+} , F is Faraday's constant, and R is the ideal gas constant. Activity is the concentration

of the reactant that is available for reaction, which is essentially the reactant's concentration for dilute solutions.

Biopotential electrodes are fed into amplifiers with large impedance, so the flow of current is minimal. Nonetheless, current flow will cause an overpotential through three mechanisms. The ohmic overpotential is caused by the resistance of the electrolyte, and it may be nonlinear. The concentration overpotential is caused by a change in the ionic balance (the ratio of anions to cations) caused by the current. The activation overpotential results because current causes the energy of activation for oxidation to change. All of these characteristics are generally waveform-dependent and frequency-dependent.

14.3 Polarizability

Electrodes vary in the extent to which their surface charge changes over time. The extremes of this characteristic are polarizable electrodes and non-polarizable electrodes. While perfectly polarizable or perfectly non-polarizable electrodes do not exist, some practical electrodes come close to acquiring their characteristics.

14.3.1. Perfectly polarizable electrodes

For a perfectly polarizable electrodes, charge builds up on the electrode surface, and current is a displacement current, meaning that it displaces the charge buildup. Because the ions accumulate on the electrode surface, the current passing between the electrode and the electrolyte changes the concentration, primarily near the interface. Materials that approximate perfect polarizability are the noble metals like platinum and gold; because these materials are relatively inert, it is difficult for them to oxidize and dissolve. The electrical characteristics of such an electrode show a strong capacitive effect.

14.3.2. Perfectly non-polarizable electrodes

For a perfectly non-polarizable electrode, current passes freely across the electrode–electrolyte interface, requiring no energy to make the transition. Consequently, charge does not build up on the surface. Examples are the silver/silver chloride (Ag/AgCl) electrode and the mercury/mercurous chloride (Hg/Hg₂Cl₂) electrode.

14.4 Silver/silver chloride electrode

The Ag/AgCl electrode consists of silver that is coated with silver chloride. It is used in an electrolyte solution that contains a large amount of Cl[−] ion. The chemical reaction at the silver surface is an oxidation



The silver ion produced by this reaction then react with the chloride in solution according to



where the down-arrow (↓) indicates that the AgCl precipitates from the solution (does not remain dissolved) and accumulates on the electrode surface. It is of interest to know whether the

ionization of Ag^+ leads to a build-up of this ion that can alter the half-cell potential. The concentration of chloride in the body is approximately 0.09 moles/liter, so the chloride activity is on the order of 0.1 moles/liter. For a material such as AgCl that dissolves into a single monovalent anion (Cl^-) and a single monovalent cation (Ag^+), the product (anion activity) times (cation activity) is a constant, K_s at equilibrium, where $K_s \approx 1.6 \times 10^{-10}$. Activity, a , is the concentration of a substance that is available to react, and it is reasonable to approximate it as equal to the concentration (although it will be somewhat less than the concentration). This information allows the concentration of Ag^+ to be plotted as a function of the concentration of Cl^- . That is, if a solution is made from pure AgCl so that it has equal parts Ag^+ and Cl^- , but then Cl^- is added alone, the Ag^+ concentration must decrease as some of the Ag^+ reacts with Cl^- to form more AgCl . Specifically, the hyperbolic relationship

$$[\text{Ag}^+][\text{Cl}^-] = K_s,$$

must hold, where the square brackets indicate “concentration of.” The plot is shown in Figure 14-2.

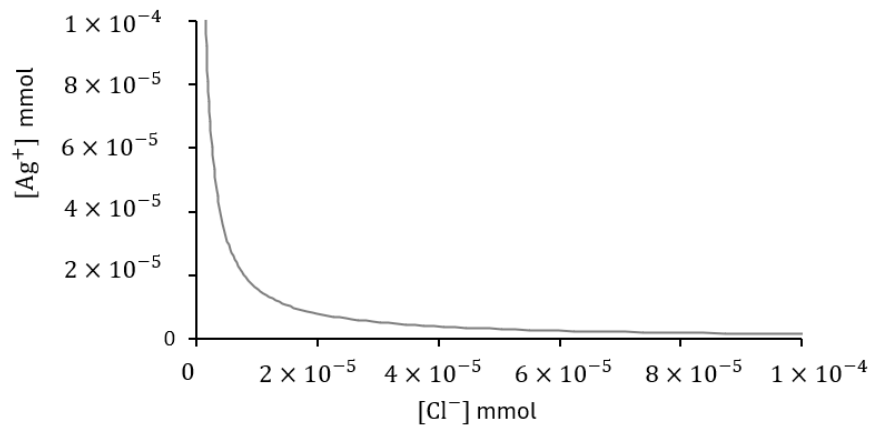


Figure 14-2: Hyperbolic relationship between $[\text{Ag}^+]$ and $[\text{Cl}^-]$.

As a result of this relationship, the stable ≈ 0.1 mole/liter concentration of Cl^- concentration in the body leads to a Ag^+ concentration of about 10^{-9} moles/liter. The electrode half-cell potential is

$$E = E_{\text{Ag}^+}^0 + \frac{RT}{nF} \ln[\text{Ag}^+] = E_{\text{Ag}^+}^0 + \frac{RT}{nF} \ln \frac{K_s}{[\text{Cl}^-]} \quad \left(\text{since } [\text{Ag}^+] = \frac{K_s}{[\text{Cl}^-]} \right)$$

From the addition property of logarithms

$$E = E_{\text{Ag}^+}^0 + \frac{RT}{nF} \ln K_s - \frac{RT}{nF} \ln[\text{Cl}^-]$$

Because E_{Ag}^0 and K_s are constant, changes in the half-cell potential are dominated by the stable $[\text{Cl}^-]$ in biological fluids and would not be influenced significantly by even an order of magnitude change in $[\text{Ag}^+]$.

14.5 Electrode fabrication

To fabricate a Ag-AgCl electrode, a silver electrode is placed into a Cl^- electrolyte solution. The positive side of a 1.5 volt battery is connected through a resistor to the electrode (anode), and the negative side of the battery is connected to another silver electrode with much larger surface area placed in the same solution. The resistor controls the amount of current delivered from the battery. The silver on the anode electrode becomes oxidized according to Equation 14-1 to Ag^+ as current enters into it, and Ag^+ reacts with Cl^- according to Equation 14-2. The resulting AgCl precipitates out of solution and deposits onto the silver. As more AgCl becomes deposited, the resistance to current increases. The deposition is complete when the current is near zero (about $10\ \mu\text{A}$).

14.6 Biopotential electrode equivalent circuit

The equivalent circuit for a biopotential electrode is shown in Figure 14-3.

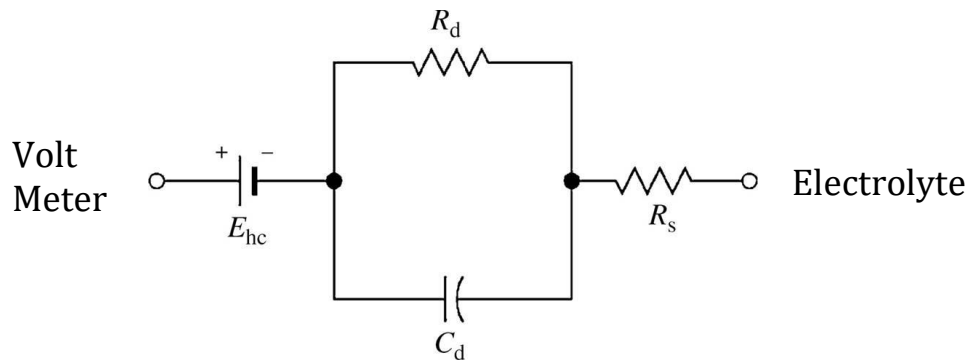


Figure 14-3: Equivalent circuit for a biopotential electrode.

The voltage source E_{hc} is the half-cell potential. C_d represents the capacitance across the double layer of charge at the electrode–electrolyte interface. R_d represents the leakage resistance across the double layer, and R_s represents the series resistance associated with interface effects and the resistance in the electrolyte. The parameter values in the circuit depend on the electrode material and geometry, and on the material of the electrolyte and electrolyte concentration.

The impedance as a function of frequency the electrode can be readily determined from the equivalent circuit.

$$Z_e = R_s + \frac{\frac{R_d}{j\omega C_d}}{R_d + \frac{1}{j\omega C_d}} = R_s + \frac{R_d}{j\omega R_d C_d + 1}$$

At high frequencies, where $j\omega R_d C_d \gg 1$, the impedance has a constant of R_s (impedance of $C_d \rightarrow 0$ and current bypasses R_d through the capacitor). At low frequencies, where $j\omega R_d C_d \ll 1$, the impedance is again constant but it has the larger value $R_s + R_d$ (impedance of $C_d \rightarrow \infty$). At frequencies between these extremes, the electrode impedance is frequency-dependent. The frequency dependance is strong for a pure Ag electrode, but the deposition of AgCl on top of the pure silver reduces the frequency dependance.

Chapter 15: Photoplethysmography & Pulse Oximetry

15.1 Introduction

Plethysmography (PPG) is the measurement of volume changes (from the Greek root *plethysmos* meaning to become full). One simple form is the measurement of changes in lung volume through the immersion of the body in a water bath. Photoplethysmography is an optical technique used to detect blood volume changes in the microvascular bed of tissue. The photoplethysmography device consists of a light source, such as an LED, and a photodetector, such as a photodiode or phototransistor. Blood absorbs light, so a larger volume of blood will absorb more light, causing variation in the signal at the photodetector. The principle can be used as a component of devices to monitor pulse rate, oxygen saturation, blood pressure, and blood vessel stiffness. The most common site at which the measurement is made is the fingertip.

15.2 Functional Principle

The light intensity returned to the photodetector from the blood and tissue has a pulsatile component, I_p , caused by changes in blood volume over a heartbeat, and a slowly varying baseline component, I_b , that corresponds to the scattering from tissue, venous blood, and the nonvarying (slowly varying) component of arterial blood. The pulsatile component is much smaller than the baseline component.

The peak of the PPG waveform occurs about 250 msec after the QRS complex of the ECG because time is required between ventricular contraction and the transmission of the blood pressure pulse to the fingertip.

15.3 Implementation

A block diagram of the PPG system is shown in Figure 15-1. The photodetector can be across the target tissue from the light source or on the same side. The band pass filter removes the I_0 component and noise in the frequency range greater than that of the I_p component.

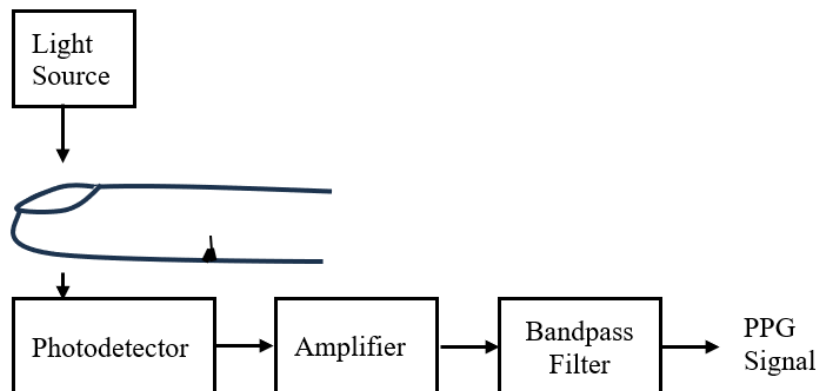


Figure 15-1: Block diagram of the PPG system.

15.4 Transmitted and Received Light Wavelengths

The amount of light on the detector depends on the degree to which the transmitted light is absorbed by the blood. The light absorbed by blood depends on the oxygenation state of hemoglobin. Oxygenated blood strongly absorbs blue light (~450 nm) and green-yellow light (500-600 nm). Red (~660 nm) and infrared (IR, ~940 nm) light pass through tissue. Green light is used for PPG because of the large propensity for absorption and the wide availability of high-power LEDs. IR is also used as will be explained in the next section.

15.5 Modes

In reflection mode the light source and photodetector are on the same side of the target tissue, so the signal is caused by reflected light. Green light has a higher absorption by oxyhemoglobin (O_2Hb) and deoxyhemoglobin (RHb) than IR, and reflection of green light is larger than reflected IR, resulting in better signal-to-noise ratio (SNR). A reflection mode PPG can be placed at any location on the body because light does not need to pass entirely through the target tissue. However, reflection mode PPG is more subject to motion artifact than transmission mode PPG. In transmission mode the light source and photodetector are on opposite sides of the target tissue, so the measurement location needs to be thin enough that sufficient light is transmitted. Common locations are the fingertip and the earlobe. For transmission mode, IR is more commonly used because it passes through tissue easily.

15.6 Factors that Affect the Signal Quality

Factors in addition to the light wavelength and the measurement site that affect the signal are contact pressure and ambient light. Too little contact pressure reduces the signal strength, but too much contact pressure can affect the blood flow. An optimal pressure is about 95 mm Hg. Ambient light can overwhelm the received signal and is particularly problematic when it is variable, so the photodetector must be shielded from light components that are not derived from the PPG light source.

15.7 Pulse Oximetry

While photoplethysmography detects changes in blood volume, pulse oximetry is a subcategory of photoplethysmography modified to detect oxygen saturation. Blood oxygen saturation, S , is defined as

$$S \text{ (in percent)} \equiv 100 \frac{[O_2Hb]}{[O_2Hb] + [RHb]},$$

where $[O_2Hb]$ is the concentration of oxygenated hemoglobin and $[RHb]$ is the concentration of deoxygenated (reduced) hemoglobin. It can be calculated from the amplitudes of the transmitted red and IR light signals. The total light, I , that reaches a distance d into a tissue is obtained from the Beer-Lambert law

$$I = I_0 e^{-\alpha D} \Rightarrow -\alpha D = \ln(I/I_0)$$

Where I_0 is the light intensity at location $D = 0$, i.e., at the surface of the tissue. Absorbance, A , is defined in terms of the log base 10 rather than the natural logarithm.

$$A \equiv \log(I_0/I) = -\ln(I/I_0)/2.303 = \alpha D/2.303.$$

The coefficient α depends on the extinction coefficient, E , of the absorbing component (in dL/g/cm) and the concentration, C , of that material such that*

$$A = ECD.$$

The structure of the finger consists of tissue, venous blood, and arterial blood. The change in the systolic and diastolic phases in arterial blood is caused by the change in blood volume, which is proportional to the distance D . The intensities corresponding to the systolic peak and diastolic minimum, I_s , and I_d , respectively, can be read from the PPG waveform.

$$A_s \equiv \log(I_0/I_s) = ECD_s \quad \text{Systolic Signal}$$

$$A_d \equiv \log(I_0/I_d) = ECD_d \quad \text{Diastolic Signal}$$

$$\Delta A = A_s - A_d = \log(I_0/I_s) - \log(I_0/I_d) \quad \text{Pulsatile Signal}$$

From the addition property of logarithms,

$$\Delta A = \log\left(\frac{I_0/I_s}{I_0/I_d}\right) = \log\left(\frac{I_s}{I_d}\right).$$

The intensity of each signal is now decomposed into a mean value $I_m = (I_s + I_d)/2$ and an amplitude $\Delta I/2 = (I_s - I_d)/2$. Hence, $I_s = I_m + \Delta I/2$, and $I_d = I_m - \Delta I/2$.

$$\Delta A = \log\left(\frac{I_m + \Delta I/2}{I_m - \Delta I/2}\right) = \log\left(\frac{1 + \frac{\Delta I}{2I_m}}{1 - \frac{\Delta I}{2I_m}}\right).$$

The expression for ΔA can be expanded as a Taylor series, and because the ratio $\Delta I/(2I_m)$ is much less than 1, only the first two terms need to be kept. With $x \equiv \Delta I/(2I_m)$, the function to be expanded is

$$h(x) = \log\left(\frac{1+x}{1-x}\right),$$

where

$$h(x) \approx h(0) + xh'(0).$$

But $h(0) = \log\left(\frac{1+0}{1-0}\right) = \log 1 = 0$, and with $h(x) = f(g(x))$ where $f(x) = \log x$ and $g(x) = (1+x)/(1-x)$,

$$\frac{d}{dx}\left(\log\left(\frac{1+x}{1-x}\right)\right) = \underbrace{f'(g(x))g'(x)}_{\text{chain rule}} = \underbrace{\frac{1-x}{\ln 10 (1+x)}}_{f'(g(x))} \underbrace{\left(\frac{(1-x) - (1+x)(-1)}{(1-x)^2}\right)}_{g'(x)}$$

* So clearly $\alpha \equiv 2.303EC$.

$$\left. \frac{d}{dx} \left(\log \left(\frac{1+x}{1-x} \right) \right) \right|_{x=0} = \frac{2}{\ln 10}.$$

Thus,

$$\log \left(\frac{1+x}{1-x} \right) \approx \frac{2}{\ln 10} x,$$

or more specifically,

$$\Delta A = \log \left(\frac{1 + \frac{\Delta I}{2I_m}}{1 - \frac{\Delta I}{2I_m}} \right) \approx \frac{2}{\ln 10} \left(\frac{\Delta I}{2I_m} \right) = \frac{\Delta I}{I_m \ln 10} \quad \text{Equation 15-1}$$

In pulse oximetry, two wavelengths of light are transmitted, one at 660 nm and the other at 940 nm. Deoxygenated hemoglobin (RHb) has a higher absorption at 660 nm and oxygenated hemoglobin (HbO₂) has a higher absorption at 940 nm. Thus, the signals received at the two wavelengths have different contributions from the RHb and HbO₂. Two signals are obtained

$$\begin{aligned} s_{660} &= a_1 \text{HbO}_2 + a_2 \text{RHb} \\ s_{940} &= b_1 \text{HbO}_2 + b_2 \text{RHb} \end{aligned}$$

Where a_1 , a_2 , b_1 , and b_2 are constants known from the absorbance spectra of Hb and HbO₂ (). We can easily solve these simultaneously for HbO₂ and RHb.

$$\begin{aligned} \text{HbO}_2 &= \frac{a_2 s_{940} - b_1 s_{660}}{a_2 b_1 - a_1 b_2} \\ \text{RHb} &= \frac{a_1 s_{940} - b_2 s_{660}}{a_2 b_1 - a_1 b_2} \end{aligned}$$

Therefore, the signal strengths s_{660} and s_{940} provide the relative amounts of HbO₂ and Hb.

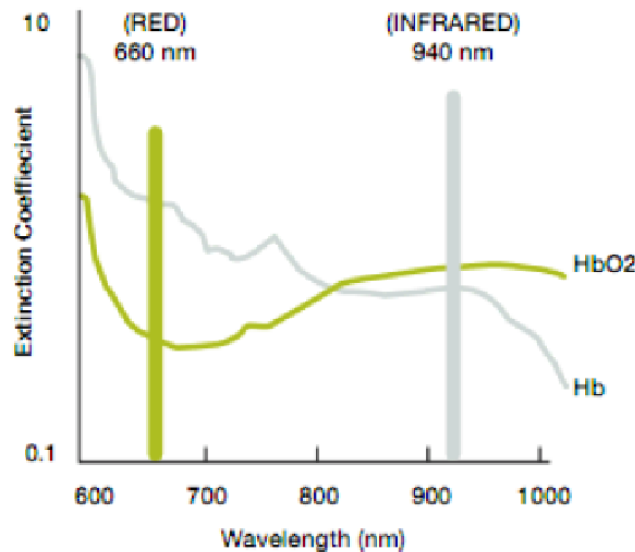


Figure 15-2: Absorbance spectra for Hb and HbO₂, from which coefficients a_1 , a_2 , b_1 , and b_2 can be deduced.

The 660 nm (red) to 940 nm (infrared) light ratio is expressed as

$$\phi = \frac{\Delta A_{660}}{\Delta A_{940}} = \frac{\Delta I_{660}/I_{m(660)}}{\Delta I_{940}/I_{m(940)}}$$

E_h , the extinction coefficient of tHb (total hemoglobin, O₂Hb plus RHb), can be calculated as the weighted average of E_{O_2Hb} and E_{RHb} , which corresponds to the concentration ratio.

$$E_h = E_{O_2Hb}S + E_{RHb}(1 - S),$$

where

$$S = \frac{[O_2Hb]}{[O_2Hb] + [RHb]}$$

and $O_2Hb + RHb = 1$. With subscripts (660) and (940) indicating the corresponding wavelengths,

$$\phi = \frac{\Delta I_{660}/I_{m(660)}}{\Delta I_{940}/I_{m(940)}} = \frac{E_{O_2Hb(660)}S + E_{RHb(660)}(1 - S)}{E_{O_2Hb(940)}S + E_{RHb(940)}(1 - S)}.$$

The E values are constants. Then, ϕ is used to calibrate a standard curve according to S . Thus, we can calculate S from ϕ , which can be measured from the PPG signal based on the changes in R and IR absorbance.

Chapter 16: Device Design

Bioinstrumentation often involves the design of instrumentation and the integration of instrumentation with other components, which may not be directly related to electrical engineering, such as mechanical or chemical components.* In designing devices, it is best to follow the engineering design process, which is a systematic approach. Different texts describe the engineering design process with varying perspectives and identify the steps differently, but they all follow a general pattern, even if they may attach different labels to the blocks in the diagram.

Figure 16-1 shows a block diagram of the process. While the diagram generally flows from needs analysis through design criteria, then to the design concept, and ultimately to the device, the student should realize that at any point in the process it is possible (sometimes necessary) to move backwards. One may find, for example, that the results from testing a given design concept demonstrate that the chosen concept cannot satisfy the design criteria that were specified, and a new design concept must be tried, modeled, and tested. Similarly, whereas a device may fulfill all of the design criteria, the designer may recognize that it is still not completely adequate or that it is possible to implement an additional design criterion.

In the figure, the Design Criteria block is circled in red, indicating a high degree of importance. This block is the heart of the process, as will become clear in the description of the other blocks.

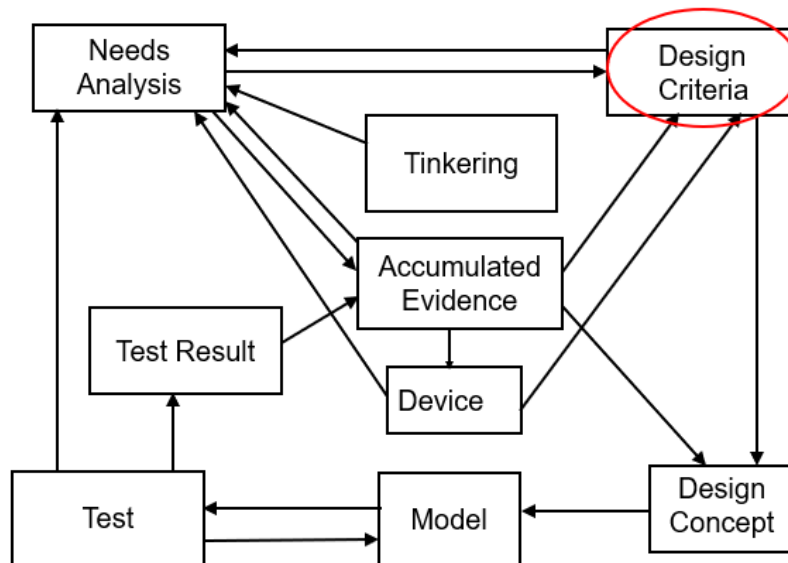


Figure 16-1: Conceptual diagram of the engineering design process.

* Although, from a biomedical engineering perspective, everything is related to everything else.

16.1 Tinkering

Tinkering is not often included in descriptions of the engineering design process because it implies a lack of direction. However, this is the stage at which many design ideas begin. It is not a necessary step, but it can provide insight into how device components work together. It is also fun. Of course, the important thing to understand is when to stop tinkering and start the rest of the design process.

16.2 Needs Analysis

Design of a device is pointless if a need does not exist for that device. The designer needs to be aware of what (medical) problems need to be solved and what current solutions are available. Useful approaches for needs analysis include, (1) immersion, (2) the web, (3) product catalogs, (4) personal experience, and (5) recent journal articles. Of these approaches, immersion is key. For example, if one wishes to improve some aspect of bedside care for patients, it is best to immerse oneself into the hospital environment to determine how different related devices are used by the physicians and staff. Similarly, if the goal is to improve a home health device, a representative sample of the state of the art for that device should be obtained and used. In addition, conversations should be engaged with the physicians, nurses, staff, and patients relevant to the device.

Filling a need does not necessarily mean that the new device will add features that the old device lacks. In some cases, it can be as simple as making the device less expensive, more lightweight, easier to handle, more eco-friendly, or less complicated to operate. The needs analysis block is one where brainstorming (Section 16.8) can be particularly useful.

16.3 Design Criteria

While the needs analysis provides a description of the general criteria that a device must fulfill, those criteria must be defined quantitatively, and they must be supplemented with additional criteria that are important but not necessarily identified in the needs statement. For example, an artificial eye would be a great device, but would not be practical if it did not fit within the eye socket, if it weighed 10 pounds, if it required a massive battery to operate, or it lasted for only a week. The design criteria define the device and direct the experimental testing that will need to be done in the ensuing stages of the design process. Broadly, design criteria can be categorized as functional, environmental, safety-related, economic, and social.

16.3.1. Functional

Functional factors define the specific functions that the device must have. For a biomedical device, these factors include signal frequency, signal sensitivity, input impedance, transient response, accuracy, linearity, reliability, and output power.

16.3.2. Environmental

Environmental factors define how the device must be designed to fit the environment in which it is used. For example, a device for use in the ER has different requirements from one used by an EMT or for home use. Environmental factors do not address issues of environmental friendliness or ecological impact. Those issues are addressed under “social” criteria.

16.3.3. Medical safety

The device must be safe for the patient, for those treating the patient, and all others who will in some manner interact with the device, whether the interaction relates to the device's official use or arises by accident. For biomedical devices, one must clearly consider electrical safety, but other issues arise as well, such as packaging with edges that may cause danger or the need to maintain a sterile environment in the operating room.

16.3.4. Economic

Device affordability is important. A balance must be found between cost and value. For a company, a strong consideration is whether a device can be produced at a cost that is significantly less than standard reimbursement rates for that device's function. In some cases, the economics of medical devices overlaps with ethical issues in that one is considering the cost of a test or procedure one side and the value of one's quality of life, or of human life itself on the other. Other economic considerations that a hospital needs to consider in purchasing devices are product availability, warranty, and compatibility with existing equipment

16.3.5. Social

Social factors include environmental impact, compatibility with religious beliefs (objection to stem cells, vaccine aversion), acceptance by different ethnic groups, and embarrassment or stigma. A classical example of medical aversion is that of African American patients brought about by a history of physicians undertaking unethical experiments with this population. The Tuskegee experiments carried out by the public health service between 1932 to 1972 are probably the best-known example of such experiments. African American men who had contracted syphilis were recruited into the study and were misled as to the study's process. The purpose was to study the natural history of untreated syphilis, and 400 of the subjects were intentionally provided with no treatment, with 200 control subjects who did not have syphilis. It is clear that some ethnic aversion to medical treatment is not without reason.

16.4 Design Concept

After the design criteria are established, a concept must be devised that is likely to fulfill the criteria. Formation of the design concept is another appropriate application of brainstorming. Another useful tool is the generation of a decision matrix, in which the major design criteria are given a value rating its importance and different designs are given a number that rates the extent to which the design is likely to fulfill that criterion. The importance values and the likelihood values are multiplied for each criterion and then summed, and the design with the highest score is chosen.*

* While this approach appears to be objective, it must be used judiciously. It is possible for the design team to choose the weightings and likelihoods to skew the design in the direction that will lead to their desired approach. For best results, it is good to have the importance weightings determined by someone who is not familiar with the proposed designs and to have the likelihood ratings determined by someone who is not familiar with the importance weightings.

16.5 Simulation and Modeling

Once the design concept has been established, a plan must be made to test the fulfillment of all design criteria. These tests are performed through simulation and modeling. It is reasonable to construct multiple models, each of which tests different criteria. It may be possible to test some criteria with a relatively inexpensive model that can be constructed quickly, and it is unwise to construct an expensive model before these simpler models have shown that the design criteria associated with them can be fulfilled. Also, parts required for a complete prototype can be subject to paperwork delays and supply chain issues. Models can be as simple as a cardboard container that is used to optimize the size and shape of the device, a control panel to test the user interface, or a simple electrical component to verify input and output impedance.

16.6 Testing

Testing must be performed in a manner that is statistically valid. Prior to each test, the specific variables to be measured must be identified, and the statistical tests necessary to validate the associated design criterion need to be developed. Successful test results are generally collected to the “accumulated evidence” block. Unsuccessful tests generally require a step backward in the engineering design process diagram. The result may be a modification of the design, re-evaluation of the design criteria, or even a consideration of need. It is also possible that a test will reveal capabilities of the device that were not originally anticipated and that the design criteria and/or need are updated to incorporate those capabilities into the final product.

16.7 Prototype Device

The prototype is a version of the device that is constructed to fulfill all the design criteria. It is possible that the models pass all criteria, but that the final prototype reveals issues caused by combining all functions into one device. Therefore, after the prototype is constructed, the design criteria must be retested. Having already tested the individual models, however, greatly improves the likelihood that the prototype will operate correctly and provides insight into how to reconstruct the prototype in instances where problems arise.

16.8 Brainstorming

Brainstorming is not a block in the diagram, but it is a useful tool used throughout the process to establish the need, establish design criteria, formulate possible solutions, and design models and experiments to test specific design criteria. The idea behind brainstorming is to produce as many ideas as possible in a short amount of time. The main barrier to the production of ideas is over analysis. Many good ideas will be discarded if the participant is afraid of suggesting something that the other group members might consider stupid. The objective of a good brainstorming session is to relieve the participants of any fear of humiliation. Participants should knowingly raise stupid ideas, with the idea that they might stimulate new ideas from other participants. The brainstorming rules are.

1. Define the topic.
2. Form a small group. Resources suggest four to seven participants.
3. Assign one person to record all ideas that are produced and to indicate who produced each idea.

4. Announce an idea as soon as it enters your brain.
5. Do not discuss any ideas that are suggested, as doing so will slow the flow of ideas.
6. Produce as many ideas as possible, regardless of quality.
7. Do not criticize ideas that are presented by any of the group members.
8. Do not evaluate the ideas during the brainstorming session.
9. Let ideas that have been suggested stimulate your own new ideas.
10. Embrace stupid ideas and actively encourage their production.
11. Set a goal for the number of ideas to be produced by the group.
12. Encourage group members who seem reluctant to speak to participate.

If a brainstorming session is too short, i.e., less than 15 minutes, long, it may not reach a point where ideas flow smoothly. If it is too long, i.e., more than 45 minutes, the group may become exhausted and decide never to do a brainstorming session again.

Once the brainstorming session is complete, the ideas can be evaluated for relevance and feasibility. While it may seem that the “stupid” ideas are immediately removed from the list at this stage, sometimes a good idea can be formulated through the combination of two or more “stupid” ideas.

Chapter 17: Major Medical Instruments

17.1 Diagnostic

17.1.1. B-Mode Ultrasound

Working Principle

Applications

17.1.2. Doppler Ultrasound

Continuous-Mode

Pulsed Mode

Color Doppler

Power Doppler

Intravascular

17.1.3. Magnetic Resonance Imaging

17.1.4. Computer Aided Tomography

17.1.5. Positron Emission Tomography

17.1.6. SPECT

17.2 Treatment

17.2.1. Gamma Knife

Gamma knife is a form of radiation therapy used on brain tumors. Gamma radiation is potentially harmful to healthy brain tissue, but the radiation must pass through healthy tissue to reach the tumor. The strategy is to use a somewhat weaker level of gamma radiation that will not severely affect brain tissue and to pass that to the tumor from different angles so that the total dosage of radiation on the tumor, the sum from these different angles, is therapeutic.

The location and geometry of the tumor must be determined accurately prior to therapy, so the procedure occurs in three stages. First, an alignment frame is attached to the patient's head, and the patient is placed in an MRI machine, which acquires an image of the tumor. Next, the MRI image is processed and the optimal sequence of angles to be used by the gamma knife is computed. Third, the patient is placed in the gamma knife, where the radiation exposure happens. The procedure can take a few hours, and the head frame must be attached to the patient throughout so that the alignment on the MRI machine exactly matches that on the gamma knife. The attachment of the head frame is the most uncomfortable part of the procedure, and to alleviate some of the discomfort, a topical anesthetic is applied to the four locations where bolts hold the apparatus to the head and the patient is given a mild sedative.

17.2.2. Cochlear Implant

The cochlea is a component of the ear that converts sound waves to nerve impulses. The term cochlea comes from the Latin word for “snail shell” because it is shaped like a small snail. It is lined with sensitive hair cells whose vibrations are transduced by nerves and sent to the brain through the acoustic nerve. Hair cells near the opening of the cochlea (where the snail’s head would be) are more sensitive to low frequency sound, whereas deeper hair cells are more sensitive to high frequency sound. The cochlear implant is a device that can partially restore hearing in cases where the acoustic nerve is intact, but an aspect of sound transmission from the outer ear to the nerve connections on the cochlea does not function. Example problems include disruption of the ear bones or damage to the cochlea. Sound is converted to a signal by a microphone outside the ear and transmitted to a device that connects to locations along the cochlea. Multiple signals are obtained from the sound and sent to specific locations along the cochlea. The restored hearing is imperfect, but it is considered to be better than no hearing at all.

17.2.3. Lithotripsy

17.2.4. Pacemaker

17.2.5. Deep Brain Stimulation

Part II: Digital Bioinstrumentation

Chapter 18: Combined Analog and Digital Systems

18.1 Analog and Digital Signals

Analog signals are continuous in time. Digital signals are either generated numerically or derived through sampling of an analog signal. In the following MATLAB code, $s(t(k))$ is a digital signal that is defined only at the specific values of $t(k)$.

```
dt = 0.001; % seconds
tmax = 0.1; % seconds
t = 0:dt:tmax; % create a series of times
w = 2*pi*50; % Radians/s
s = cos(w*t); % create a cosine signal
```

If s is plotted as a function of t , it has the appearance of a continuous signal, but it is defined only at specific times. These discrete functions are sometimes plotted as stem plots, where the discrete points are not connected together, but are shown with vertical lines originating at the graph's horizontal axis, as in Figure 18-1. (Do not confuse the stem plot with a stem-and-leaf plot, which is an entirely different animal that will not be considered in this course.)

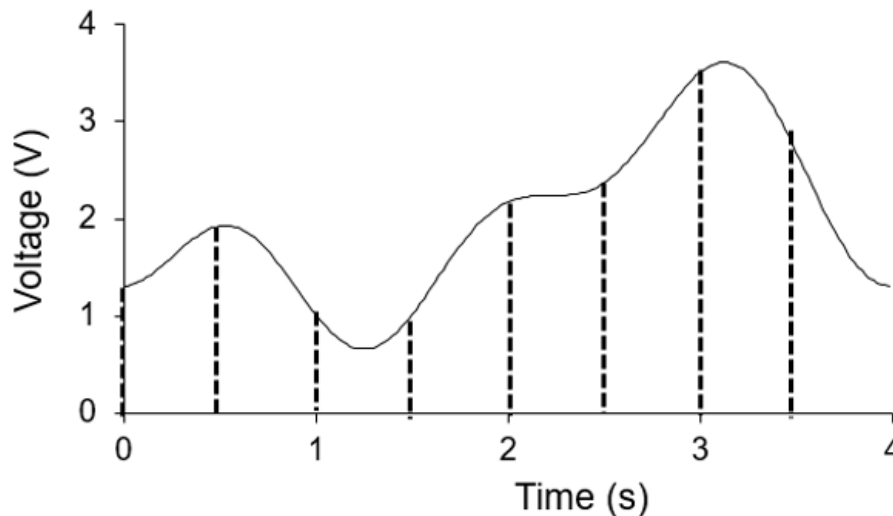


Figure 18-1: Stem plot for a digital signal. The solid line represents the original continuous signal and the dashed vertical lines are the “stems.” The sample rate for this example is 2 Hz (0.5 seconds per sample).

18.2 The Generalized System

Most biomedical instruments will have both analog and digital components. Because biological signals such as biopotentials, pressures, and temperatures, are continuous signals, an analog-to-

digital converter (digitizer) must be provided to convert these signals to a digital form if digital processing is to be used. Digital processing is preferred in many cases because it is easy to implement and modify, once the digital form of the signal has been obtained. For example, if one wishes to change the form of an analog filter used to extract important diagnostic information from a signal, the filter must be rewired, and a complete change in the board layout may be necessary. The same change in a digital system requires only modification of a few lines of code.

18.3 Appropriateness of Analog and Digital Approaches

Given the powerful benefit of a digital system, it may seem that in all cases the analog system should be used only to digitize the signal and that digital processing should be used after that, so your courses on analog electronics and basic circuits were a complete waste of time. However, analog systems cannot be completely replaced by digital systems. Fundamentally, it should be noted that digital hardware is, at heart, analog. It is constructed from nonlinear devices that are designed, in general, to have two stable states, True and False, or 0 and 1, or 0 volts and 5 volts, however one chooses to label them, but the states do not switch instantaneously from one state to the other, so they are controlled by analog principles. Also, if an application is simple, requiring perhaps at most a simple filter, implementation of an analog to digital converter may not be worthwhile or cost effective. Beyond these considerations, three areas in which digital systems cannot replace analog systems are signal pre-conditioning, downmixing, and feedback control.

18.3.1. Signal Pre-Conditioning

Signal pre-conditioning describes the processing that must be performed on a signal before it can be optimally digitized.

Adjustment for voltage range

Analog-to-digital converters are limited in the range of voltages that they can acquire and the frequency at which they can acquire the data. The digital representation of the signal is a binary number that has no specific meaning until it is converted to a voltage based on the reference voltage range that was used during the digitization. If the reference voltage range is 0 to 5 volts, then the digitizer will not be able to represent 6 volts or -0.8 volts, so the input signal must be kept between the 0 and 5 volt limits. The solution to this problem is to either amplify or attenuate the signal with an analog amplifier or attenuator. The gain of this device must be accounted for in the digital algorithm that recalculates the voltage of the original signal. It needs to be selected carefully because over-amplification leads to a signal that exceeds the voltage range, while over-attenuation leads to a signal that is too small for the bit resolution of the converter.

Anti-aliasing

Analog-to-digital converters sample the signal at a fixed frequency f_s .^{*} If the signal has components in it that are greater than $f_s/2$, the Nyquist frequency, then those components will be aliased to a different frequency. Once these components are aliased, no amount of digital

^{*} Signal sampling is performed at a non-constant sampling rate in some specialized applications, but the arguments here still apply in general.

filtering can remove them from the signal. Consider, for example, a sample rate of 2 kHz and a signal that contains a component of interest at 500 Hz and a noise component at 1500 Hz. The 1500 Hz signal will be aliased to 500 Hz and be indistinguishable from the signal of interest. It must be filtered from the signal prior to the digitization if it is to be eliminated. The cutoff frequency of the anti-aliasing filter should be less than the Nyquist frequency because analog filters always have a transition band between the pass band and the stop band, so signal components only a few percent larger than the filter cutoff frequency are only partially removed.

Pre-whitening

In some cases, signals have multiple components of different magnitude at different frequencies. Turbulent flows, for example, are composed of eddies of varying sizes. The velocities measured from rotation of the largest eddies tend to be large and low in frequency, while those measured from rotation of the smallest eddies tend to be small and high in frequency. However, velocities from both eddies are important to the understanding of turbulent flow dynamics. In particular, the smaller eddies relate to shearing stresses that may be important to blood damage in the turbulent jet from a prosthetic or stenotic valve. If the gain of the input signal is set so that the larger eddy voltages are within the range of the analog-to-digital converter, the smaller eddy voltages may be too small to be resolved by the converter. To avoid this problem, a pre-whitening filter is applied. The filter has a low gain at low frequencies and a higher gain at higher frequencies, e.g., $H(j\omega) = j\omega\alpha$ where α is the gain at $\omega = 1$ rad/s. Thus, before the signal is presented to the converter, the low and high frequency components have similar magnitudes and can be digitized with similar resolution. The name pre-whitening is derived from the concept of white noise, which has even spectral content at all frequencies. (Recall that white light, for example, is the combination of all colors in the light spectrum.) Pre-whitening is generally applied to signals like turbulence, where power decreases with increased frequency. Such signals would be categorized as brown if interpreted as a light spectrum (combination of red, orange, yellow, where red is the lowest frequency, orange is a higher frequency, and yellow is higher still).

18.3.2. Downmixing

Downmixing is a method used typically in the processing of Doppler signals, where Doppler ultrasound is the classic medical example. The sound used in Doppler ultrasound is typically in the 1 to 30 MHz range, and the implementation of an analogue-to-digital converter with a sample rate in this range is prohibitively expensive. However, in an ultrasound application, one does not care to extract the transmission frequency from the returned signal (one can read the user's manual to find that information) but rather the difference between the transmission frequency and the returned frequency. A trick is used to convert the returned signal to a signal with a frequency equal to only the Doppler frequency. In this process, the signal from the returned sound is multiplied by the transmitted frequency and the result is low pass filtered. Assume that the transmitted signal is $s_t = A_t \cos(\omega_0 t)$, and that the returned signal is $s_r = A_r \cos((\omega_0 + \omega_d)t + \varphi)$. A_t and A_r are the amplitudes of the transmitted and received signal, respectively, ω_0 is the transmitted frequency, ω_d is the Doppler shift for the returned signal, and φ is a phase shift that depends on the distance traveled by the sound wave and other acoustic parameters. The product of s_t and s_r is

$$s_t s_r = A_t A_r \cos(\omega_0 t) \cos((\omega_0 + \omega_d)t + \varphi) .$$

Now use the trigonometric identity $\cos x \cos y = \frac{1}{2}(\cos(x + y) + \cos(x - y))$, with $x = (\omega_0 + \omega_d)t + \varphi$ and $y = \omega_0 t$.

$$A_t A_r \cos(\omega_0 t) \cos((\omega_0 + \omega_d)t + \varphi) = \frac{1}{2} A_t A_r (\cos((2\omega_0 + \omega_d)t + \varphi) + \cos(\omega_d t + \varphi))$$

Because carrier frequency is orders of magnitude larger than the Doppler frequency, the frequency of the first cosine term, $2\omega_0 + \omega_d$, is much larger than that of the second cosine term, ω_d , so that the first cosine term can easily be filtered from the signal so that only the cosine term is present. The frequency of that term can be readily found through application of the Fourier transform.*

18.3.3. Feedback Control

In various medical applications, it is desirable to use the measured signal to control some aspect of the patient. The control can involve conversion of the digitized signal to an analog signal that is then used to drive an actuator, heater, or other device.

18.4 The 555 Timer as a Hybrid Analog and Digital Circuit

* While the derivation looks unwieldy when described in terms of cosines, it is in essence an application of the shifting theorem for Fourier transforms. $e^{-j\omega_0 t} \mathcal{F}(\omega - \omega_0) = \mathcal{F}(\omega)$, where $\mathcal{F}(\omega - \omega_0)$ is the spectrum of the returned signal and $\mathcal{F}(\omega)$ is that spectrum shifted by ω_0 .

Chapter 19: Digital Representation of Signals

Computers store numbers as binary values denoted by bits. If a bit is turned on, its value is one. If it is turned off, its value is zero.* The bits are combined into bytes, which are groups of 8 bits. With 8 bits, each of which can take on a value of 1 or 0, a byte can be composed of 2^8 combinations. More generally, with n bits, a number can take on n values.† A word is a combination of bytes, but words can have different lengths. Currently, home computers have a 64 bit architecture, meaning that a normal word length is $64/8 = 8$ bytes. A less familiar data unit is the nibble, which is 2 bits.

19.1 Binary Representation

19.1.1. As Numbers

Unsigned Integers

Consider an 8-bit word, as shown in Figure 19-1. The a_k s are the bit values, and the numbers in the top row are the bit numbers. The bit to the far right represents 0 or 1. The bit to the left of it represents 0 or 2. The next bit to the left represents 0 or 4, and the n^{th} bit to the left represents 0 or 2^n . Thus, the value of the number is

$$v = a_7 2^7 + a_6 2^6 + a_5 2^5 + a_4 2^4 + a_3 2^3 + a_2 2^2 + a_1 2^1 + a_0 2^0,$$

where a_0 through a_7 take on values of 0 or 1. For example, if only bits 3, 5 and 6 are turned on, then the value is $v = (1)2^6 + (1)2^5 + (1)2^3 = 64 + 32 + 8 = 104$.

7	6	5	4	3	2	1	0
a_7	a_6	a_5	a_4	a_3	a_2	a_1	a_0

Figure 19-1: Representation of an 8-bit number.

Signed Integers

The eight-bit binary representation in the previous subsection can represent any integer number between 0 and $2^8 - 1 = 255$. Similarly, a two-byte integer can represent any integer between 0 and $2^{16} - 1 = 65,535$. However, negative numbers must also be represented, and to do so, one of the bits, typically the highest order bit, is designated as the sign bit. If the sign bit is 0, then the rest of the bits represent a positive number, and if it is 1, then the rest of the bits represent a negative number. The eight bit number can then represent an integer value from -128 to $+127$, with -128 being represented with the sign bit on and all other bits zero. (This convention is opposed to having the sign bit along with all zeros represent -0 , which is the same as $+0$ and therefore a waste of a perfectly good number representation.) With a 32 bit operating system, a signed integer can represent a number from -2^{31} to $2^{31} - 1$ ($-2,147,483,648$ to $2,147,483,647$).

* For standard transistor-transistor logic, “on” refers to 5 volts and “off” refers to 0 volts. It is, of course, possible, though far less common to define “on” as 0 volts and “off” as 5 volts.

† For example, with two bits, $2^2 = 4$ values are possible. These values are 00, 01, 10, 11, equivalent to 0, 1, 2, and 3.

Big Endian/Little Endian

The location of a data value in memory is distinct from the value contained at that location, just as the contents of a post office box are separate from the post office box number. The bytes that make up a number are stored in adjacently numbered memory locations. If a four-byte integer is stored, for example, in address numbers 256, 257, 258, and 259, one might decide to store the lowest order byte at address 256, increasing to the highest order byte at address 259, or one might place the highest order byte at address 256, decreasing to the lowest order byte at address 259. Because computer architectures vary in the convention that they use a terminology is used to distinguish the convention. If the lowest order byte is stored at the lowest address location, the number is referred to as “Little Endian,” and if the highest order byte is stored at the lowest address location, the number is referred to as “Big Endian.” For the casual use of a computer, the byte order is transparent, but in cases where binary data is transferred between two devices with different architectures, it is critical that the byte order be correctly interpreted.

Floating Point vs. Integer

The number representation above allows representation of integers only. A different structure is used to represent a decimal number (referred to as “floating point,” or “real” architecture). The number is separated into a mantissa and an exponent, as shown for a 4-byte floating point number in Figure 19-2. The mantissa, in this case, is three bytes, and the exponent is one byte.

<u>0010110 01100011 00010010</u>	<u>00011111</u>
mantissa	exponent

Figure 19-2: Floating point number separated into the mantissa and exponent.

If the mantissa value is a and the exponent is b , then the number represented is $x = (1 + a) \times 2^b$. (Note that the base is 2, not 10.)

19.1.2. As Characters

Numbers can be interpreted by the software as characters. The ASCII encoding, for example, assigns the first 128 numbers to characters that are used in data storage. These characters include the keyboard characters, such as a through z, A through Z, 0 through 9, the special characters (e.g., !@#\$%^&() +-=|/?.<>) and others (such as “end of file,” “new line,” and “carriage return.”

Extended ASCII adds another 128 characters. However, extended ASCII is not unique and is language-dependent. Other encodings are also used, such as EBCDIC.

In ASCII, the letter ‘A’ is 65, ‘B’ is 66, and the alphabet follows that pattern. The letter ‘a’ is 97, ‘b’ is 98, and the lower case letters are similarly ordered alphabetically. The number ‘0’ is ASCII 48 (not ASCII 0). ASCII 0 is mapped to the “null” character.

19.1.3. Conversion to Readable Format

Analog to digital converters read data in binary format. The computer must work to change the binary format to human-readable (decimal) format. Also, the human readable format requires

more storage space. The decimal number 31,098, for example, requires at least five characters to store the digits, and if multiple numbers are to be stored, another character (e.g., a space or tab) is used to separate them. If the number is stored in its binary format, however, only two characters are needed (“yz”) and because all of these two-byte numbers are the same size, no separation character is needed. The file size is therefore smaller by more than a factor of three and the data can be transferred from one device to another three times as fast. For our laboratories, digitized data from the Arduino microprocessor will be transmitted in binary format to your laptop and then converted through the Python code that reads the data to a decimal format before it is written to a .CSV file. The Arduino analog-to-digital converter uses 10 bits, but it still transfers 2 complete bytes of data to the laptop. The extra six bits are always zero.

19.2 Quantization Error

Computers cannot store data to infinite accuracy, and every analog-to-digital converter uses a finite number of bits to represent each number. The voltage resolution of a digitizer is equal to 2^n divided by the voltage range of the converter, where n is the number of bits. For example, a 12-bit converter with a voltage range between -5 and $+5$ volts has a resolution of $\frac{5-(-5)}{2^{12}} \approx 0.00244$ volts. It cannot distinguish a voltage value of 0.032 from a voltage value of 0.033 . To obtain the best resolution for a digitized signal, the signal should be amplified such that its extreme values are near the limits of the digitizer, but not so close to those limits that the amplified voltages could exceed those limits. The consequences of having the voltage exceed the digitizer range vary with the digitizer design. In some cases, a voltage that exceeds the maximum will wrap to the minimum (e.g., if the digitizer range is -5 to 5 volts and the signal is at 5.3 volts, the digitized signal wraps to -4.7 volts, exceeding the lower limit by the same amount that the original signal exceeded the upper limit). In other cases, a voltage that exceeds the maximum limit will cause a digitized value equal to the maximum limit.

For the Arduino microprocessor, the minimum value is zero, and the maximum value is set by the selected reference voltage, either 3.5 or 5 volts. Thus, the device cannot digitize negative voltages. Any waveform that has negative values must have an offset added to it before the signal is digitized. The offset is then subtracted from the digitized signal.

Quantization error manifests as a white noise that is added to the digitized signal. The power inherent in the white noise signal can be obtained with the aid of the relationships derived in Section 6.2 on basic statistics. Because an exact number cannot be represented by a finite number of bits, the number represented by the digitizer will be the closest number to the analog voltage. For example, if the digitizer has 10 bits and can therefore represent integers from 0 to 1023 , a voltage whose digitized value should be 384.393 would be represented as 384 , and one whose digitized value should be 384.792 would be represented as 384 . The difference between the stored number and the exact number is evenly distributed across values from -0.5 to $+0.5$, with an average (expected value) of zero. The error $e(t_k)$ is a time sequence that is essentially added to the signal of interest. That is, if $s_d(t_k)$ is the digitized signal and $s_0(t_k)$ is the original signal, $s_d(t_k) = s_0(t_k) + e(t_k)$. The sequence $e(t_k)$ can be characterized by the amount of noise power it contributes, where power (P) is

$$P = \frac{1}{T} \int_0^T e^2 dt.$$

For a random signal, such as quantization noise, this integral is equivalent to the expected value of the variance of the signal (Equation 6-5).

$$\text{Power} = E\{(e_k - \bar{e})^2\} = \int_{-\infty}^{\infty} (e - \bar{e})^2 f(e) de$$

where $f(e)$ is the probability density function for e and \bar{e} is the average value of e , which is zero in this case. Because e is uniformly distributed between $-\frac{1}{2}$ the quantization level and $+\frac{1}{2}$ the quantization level, and because the integral of all probability density functions must be 1 (i.e., it is certain that an outcome will happen, even if it is not certain what it will be), the probability density function is

$$f(e) = \begin{cases} 0 & \text{if } e > \frac{q}{2} \text{ or } e < -\frac{q}{2} \\ \frac{1}{q} & \text{if } -\frac{q}{2} < e < \frac{q}{2} \end{cases}$$

With this probability density,

$$\int_{-\infty}^{\infty} (e - \bar{e})^2 f(e) de = \int_{-\frac{q}{2}}^{+\frac{q}{2}} \frac{e^2}{q} de = \frac{1}{3q} \left(\left(\frac{q}{2} \right)^3 - \left(-\frac{q}{2} \right)^3 \right) = \frac{q^2}{12}$$

19.3 Considerations for Number of Bits

Digitization error decreases with the number of bits. However, analog-to-digital hardware cost increases with the number of bits, 12 and 16 bit converters are typical.

A 12 bit converter allows 4096 levels of quantization, corresponding to about 0.024% error, which is generally sufficient, if the signal has been properly pre-conditioned. (Consider that 1% error is typical of most engineering measurements).

Data from a 12 bit converter will take 1 ½ bytes (8 bits in one byte and 4 bits in another), so that four bits are unused.

19.4 Sample Time vs. Sample Number

In most cases, signals are sampled at a constant sample rate, f_s , so that the time between samples is $dt = 1/f_s$. Even sampling lends itself readily to the use of analysis methods such as the fast Fourier transform. In some cases, the time between samples is variable because the events being investigated arrive discontinuously. An example is heart rate variability, where the measurement is heart rate and event is the occurrence of the QRS complex of the electrocardiogram. Another example is laser Doppler velocimetry, where the measurement is fluid velocity but a data point does is recorded at the random times when a light-scattering particle enters the measurement location. One way to use the Arduino digitizer is to have the device read a voltage sample, then read its clock time, and send both numbers through the serial port. This method also leads, unnecessarily, to uneven sample times. A better approach is to use a clock, which is a square

wave that tells the digitizer when to read each data value. For example, a sample might be acquired when the clock transitions from 0 to 5 volts. A constant clock frequency then leads to an evenly sampled signal.

With evenly sampled signals, the signal data are a series of numbers numbered in the sequence of their acquisition. The k^{th} sample is designated as $s[k]$, where k is an integer. The signal as a function of time is designated as $s(t_k)$, where $t_k = k(dt)$.

19.5 Sample Frequency and Aliasing

To prevent aliasing, a signal must not have a component at a frequency larger than $f_s/2$, which is the Nyquist frequency. Recall that, from the point of view of Fourier transforms, a frequency can be positive or negative. If a real signal has a component at a positive frequency, it must also have a component at the matching negative frequency, but a signal with an imaginary component can have different positive and negative frequency components. In illustration, Euler's rule leads to the following representation for a cosine wave.

$$\cos(\omega t) = \frac{e^{j\omega t} + e^{-j\omega t}}{2}$$

The portion of this signal in the positive frequency spectrum is $\frac{1}{2}e^{j\omega t}$, and the portion in the negative frequency spectrum is $\frac{1}{2}e^{-j\omega t}$. The signal $e^{j\omega t}$ has a component at $+\omega$ that is not matched by a component at $-\omega$. This asymmetry is possible because the signal is complex, rather than real. To take the Fourier transform of $e^{-j\omega t}$, fill the real part of the FFT input array with $\cos(\omega t)$ and fill the imaginary part of that array with $\sin(\omega t)$.

If a frequency component, f , is greater than $f_s/2$, it is aliased to a frequency of $f_a = f - nf_s$, where n is $\text{mod}((f + f_s/2)/f_s)^{*†}$. Assume that $f_s = 1$ kHz and the signal has a component at 600 Hz. Then n is $\text{mod}\left(\frac{600+500}{1000}\right) = 1$ and $f_a = 600 - (1)(1000) = -400$ Hz. If the signal is real, it will also have a component at -600 Hz that will be aliased to $+400$ Hz.

The method for identification of the aliased frequency is most easily understood by a few examples. In essence, it amounts to shifting the original component by multiples of f_s until it is within the range $\pm f_s/2$, as in Figure 19-3. Panel A) shows a case where the original frequency is modestly larger than the Nyquist frequency. After the signal is shifted by $-f_s$ to place it within the Nyquist range, it appears at a negative frequency (-300 Hz). In Panel B), the original frequency is modestly less than $-f_N$. Here, f_s must be added to move the component into the Nyquist range, and the aliased component appears at a positive frequency. In Panel C), the original frequency is substantially larger than the Nyquist frequency, and the shift of $-f_s$ places the aliased signal in the positive frequency range. In Panel D), the original component is larger than the Nyquist frequency by more than f_s . A shift of $-f_s$ is not sufficient to place it in the Nyquist range, so the shift must be $-2f_s$. This shift places the aliased component in the negative frequency range of the spectrum.

* The mod function (modulo) means to take the integer part of the number.

† If f is negative, then $n = -\text{mod}((|f| + f_s/2)/f_s)$

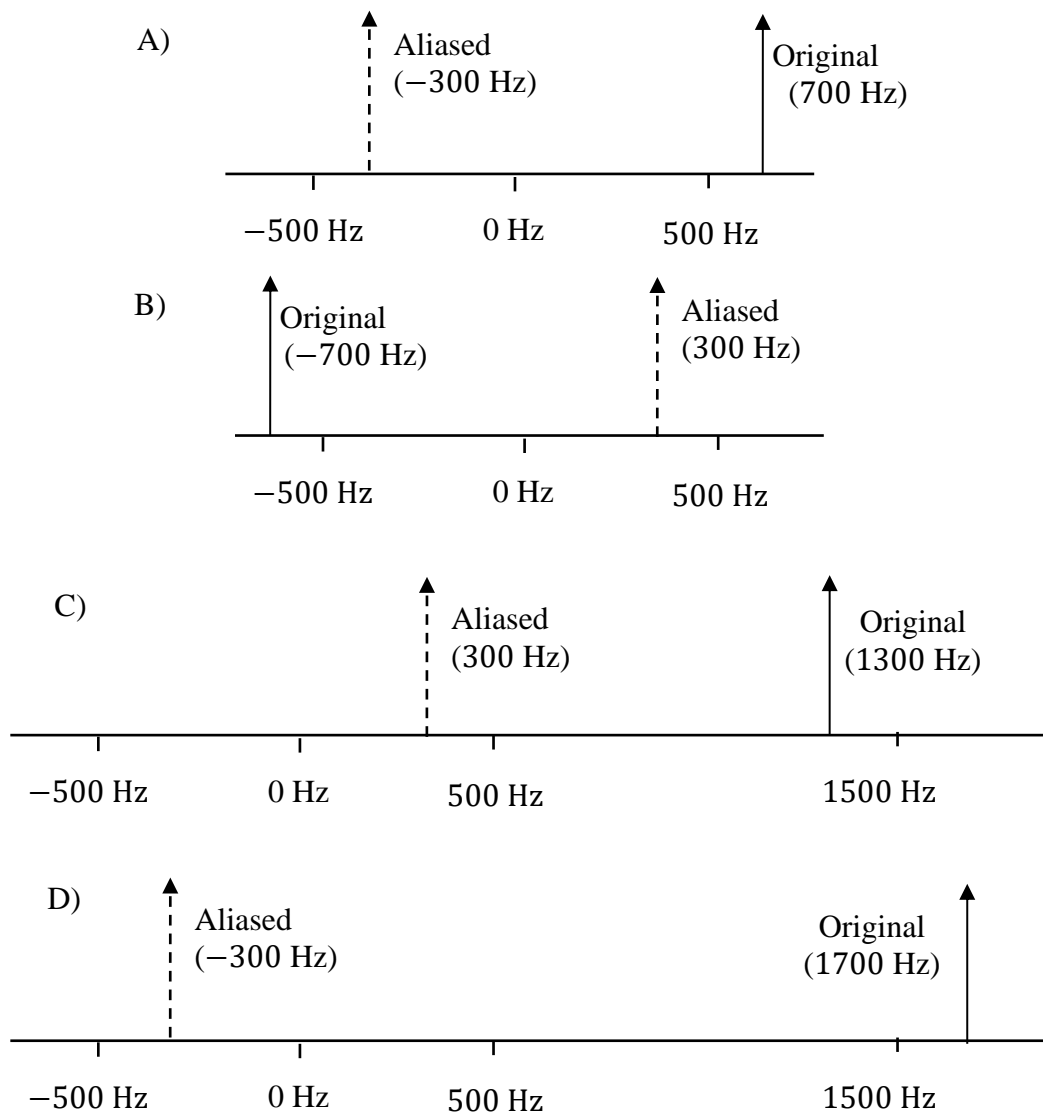


Figure 19-3: Process used to predict where frequency components will appear upon aliasing. In all examples, the sample frequency is $f_s = 1,000$ Hz, so the Nyquist range is -500 Hz to $+500$ Hz. The signal corresponds to $e^{2\pi f t}$. Solid arrows are the original signal, and dashed arrows are the aliased signal. A) The original frequency component is at 700 Hz, which is 200 Hz larger than the Nyquist frequency. The original component becomes shifted by $-f_s = 1000$ Hz to -300 Hz. B) The original component is -700 Hz, which is 200 Hz less than the Nyquist frequency of -500 Hz. The component is shifted upward by $f_s = 1000$ Hz so that it is within the Nyquist range, and it appears at $-700 + 1000$ Hz = 300 Hz. C) The original component is 1300 Hz, which is 800 Hz more than the Nyquist frequency of $+500$ Hz. The component is shifted downward by $f_s = -1000$ Hz so that it is within the Nyquist range and it appears at $1300 - 1000$ Hz = 300 Hz. D) The signal is at 1700 Hz, which is 1200 Hz larger than the Nyquist frequency. A shift of -1000 Hz is not sufficient to place it within the Nyquist range, so the shift must be $(2)(-1000) = 2000$ Hz. The aliased signal then appears at -300 Hz.

Chapter 20: Methods for Analog-to-Digital Conversion

The most straightforward method for control of an analog-to-digital converter is to present the input with a voltage signal and to use a clock signal to trigger the acquisition. This method provides evenly spaced samples in time with a sample frequency equal to the clock frequency. The technique is modified in various ways that depend on the needs of the data to be acquired.

20.1 Multiplexing

Frequently, more than one signal must be acquired in the same time frame. One might, for example, wish to monitor and record both the EKG, and oxygen saturation on a patient. One option is to use separate analog-to-digital converters, but that option is rarely pursued. Instead, a single converter is used to alternately sample the multiple channels of data. The signals are introduced to the input of a multiplexer, which sends each one successively to an output port that is connected to the converter input. The location of each data point in the digitized data array is used to identify the signal. With this configuration, a single clock can control the multiplexer sequence and the digitization. I.e., when the clock transitions from low to high, the multiplexer switches to the next signal and the digitizer acquires a sample.

20.2 Sample and Hold

If the multiplexer is used alone, the samples are acquired at different times. In some cases, it may be acceptable for the signals to be unsynchronized, but in other cases it is not. Since the digitizer cannot acquire multiple signals simultaneously, an alternative is used where, at the desired acquisition time, all signals are frozen at their current value until they have all been converted. They are then unfrozen, and the process starts anew for the next clock cycle. The heart of each channel of the sample and hold device is simply a voltage-controlled switch, where the clock provides the control signal. During the acquisition, the switch disconnects the inputs of the sample and hold from their respective voltage signals and stores the voltage level on a capacitor until it can be read.

Use of the combined multiplexer, sample and hold, and digitizer requires two clocks. The clock that controls the sample and hold circuit is referred to as the frame clock, and the clock that controls the digitizer and the multiplexer is referred to as the burst clock.

20.3 Arduino Characteristics and Limitations

The Arduino is an inexpensive microprocessor. Because it is open-source, software for many applications have been developed for it. We will use it mainly for data acquisition. It can reliably acquire a signal at a data rate of about 6 kHz, which is adequate for a wide range of biomedical applications. The voltage range is between 0 and V_{ref} , where V_{ref} is the selected reference voltage, selectable as 3.5 volts or 5 volts. Ten bits of digitization are used, which is sufficient for most engineering applications if the signal is properly pre-conditioned (e.g., amplified to match the voltage range and pre-whitened, if necessary). The Arduino does not have a dedicated DAC, but an analog signal can be simulated from the digital pins. These pins

can generate pulses at varying frequencies, and if a digital signal controls the frequencies, the pulses can be low pass filtered so that frequency becomes converted to voltage in a manner similar to the operation of action potential trains in nerves. This method is referred to as pulse wave modulation.

Control of the Arduino is based on three basic concepts, setup, loop, and interrupts. The setup initializes variables, pins, and other hardware aspects. After the setup, the device repeats the loop code indefinitely. For example, code that would cause one of the Arduino LEDs to blink would first set the LED as an output in the setup section and then repeatedly send HI and LOW signals to it in the loop section. Interrupts allow the device to respond to external events. In the case of data acquisition used here, the external event is a clock signal that tells the Arduino to jump to the code that reads the digital number from the analog-to-digital converter.

The Arduino integrated development environment can be downloaded from www.arduino.cc/en/software. The code we will use to run the Arduino as an analog-to-digital converter is included in the appendix to the laboratory manual for this course.

20.4 Python

Python is an open-source, general-use programming language, often compared to MATLAB. Python will be used in this course to read the data sent through the serial port to from the Arduino. The code required to do so is provided in the appendix to the laboratory manual. Multiple development environments exist for Python, and the code provided was developed under Thonny, which can be downloaded from thonny.org.

Chapter 21: Digital-to-Analog Conversion

Digitization of an analog signal has many advantages, particularly in terms of data analysis and storage. However, the conversion of the digitized signal to an analog signal is also of interest, so a digital-to-analog converter is required. Digital-to-analog converters (DACs) are also important because they are a major component in the design of many analog-to-digital converters; a digital value is presented to the DAC, the DAC generates an analog signal that is compared to the voltage to be digitized, and then the digital value is updated depending on whether the DAC output is greater than or less than the voltage of interest. The process continues iteratively until the two analog signals match.

21.1 Applications

21.1.1. Generation of a Perceivable Signal (e.g. Music)

Speakers are driven by analog signals. Digital methods can be used to remove noise and create other effects, but ultimately the signal cannot be considered sound until it drives a speaker. A compact disc will send the sound signal to the speaker with a sample frequency of 44,100 Hz on each stereo channel. This rate leads to a Nyquist frequency about 10% above the accepted maximum range of 20 kHz for human hearing. Upon playback, the data values are converted successively to voltage values and filtered before they are amplified and presented to the speaker.

21.1.2. Feedback Control

Digital signals are also converted to analog signals so that they can control motors, valves, and pumps. One biomedical example is the BIPAP machine, which senses pressure at the mouth of a patient and drives a pump to assist breathing during inhalation. Processing of the measured pressure signal is implemented digitally, but the respiratory assist pump must be controlled with an analog signal.

21.2 Hardware

Several concepts can be implemented to design a DAC.

21.3 Post-Processing

Post-processing of the signal after it is converted to a series of sequential voltages will be described here in terms of an audio playback.

Chapter 22: Digital Versions of Analog Signals

The familiar analog signals, e.g. the delta function, step function, and sine wave, have digital equivalents. The behaviors of analog and digital signals are similar with a few subtle differences. In the analysis of these signals, the integral is replaced by a summation. Typically, as the time between samples becomes smaller, the signals and data analysis methods behave more closely to their continuous analog counterparts.

22.1 Special Functions

22.1.1. The Impulse (Delta) Function

In continuous time, the delta function is defined as^{*}

$$\delta(t) = \begin{cases} 1, & t = 0 \\ 0, & t \neq 0 \end{cases}; \quad \int_{-\infty}^{\infty} \delta(t) dt = 1. \quad \text{Equation 22-1}$$

The discrete time delta function is

$$\delta[n] = \begin{cases} 1, & n = 0 \\ 0, & n \neq 0 \end{cases} \quad \text{Equation 22-2}$$

The integral part of the definition reduces to a sum and need not be part of the definition itself because follows from the first part (unlike the continuous time case).

$$\sum_{n=-\infty}^{\infty} \delta[n] = 1 \quad \text{Equation 22-3}$$

If the argument is changed from n to $n - k$, the location at which the function is 1 is shifted to the right, as shown in Figure 22-1.

^{*} It follows from this definition that

$$\delta(t - t_0) = \begin{cases} 1, & t = t_0 \\ 0, & t \neq t_0 \end{cases}; \quad \int_{-\infty}^{\infty} \delta(t - t_0) dt = 1.$$

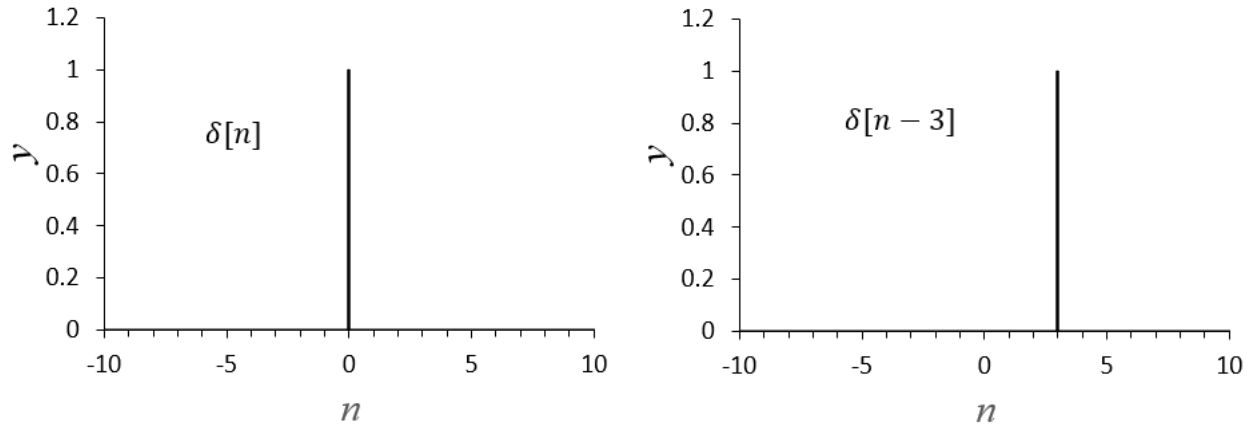


Figure 22-1: Stem plot of the discrete delta function, unshifted and shifted.

In continuous time, the delta function has the property that

$$\int_{-\infty}^{\infty} \delta(t - t_0) f(t) dt = f(t_0). \quad \text{Equation 22-4}$$

Thus, if $\delta(t)$ is shifted by t_0 and then integrated against a function of time, it selects the value of the function at the shifted time point.

Proof: $\delta(t - t_0)$ is zero everywhere except at $t = t_0$. Therefore, Equation 22-4 can be rewritten with the limits from $t - \epsilon$ to $t + \epsilon$, where ϵ is infinitesimally small. If $f(t)$ is a continuous function, then it must have the value $f(t_0)$ within the range of the integral. It then becomes a constant value that can be taken out of the integral, and Equation 22-4 becomes

$$f(t_0) \int_{t_0-\epsilon}^{t_0+\epsilon} \delta(t - t_0) dt = f(t_0)(1) = f(t_0).$$

For the discrete delta function,

$$\sum_{n=-\infty}^{\infty} \delta[n - k] f[n] = f[k] \quad \text{Equation 22-5}$$

Because $\delta[n - k]$ is zero unless $n = k$, so the only surviving term in the sum is where $n = k$, i.e., $f[k]$.

22.1.2. The Doublet Function

Definition

The continuous time doublet function, $\delta'(t)$, is defined as the first derivative of $\delta(t)$. This function is even more counterintuitive than the delta function. It is zero everywhere, except at zero, where it is both $+\infty$ and $-\infty$; the derivative of the delta function must be $+\infty$ a small time ϵ to its left so that it can reach a value of ∞ in zero time and must be $-\infty$ a small time ϵ to its right so that it can return from $+\infty$ back to zero.

For discrete signals,

$$\delta'[n] = \begin{cases} 1 & \text{if } n = 0 \\ -1 & \text{if } n = 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{Equation 22-6}$$

Relationship to Derivative Sampling

In analogy to Equation 22-4, if the doublet function is integrated against a function, it selects the derivative of that function.

$$\int_{-\infty}^{\infty} \delta'(t_0 - t) f(t) dt = f'(t_0) \quad \text{Equation 22-7}$$

The result can be found through integration by parts, with $u = f(t)$, $du = f'(t) dt$, $dv = \delta'(t_0 - t) dt$, $v = -\delta(t_0 - t)$. The integral becomes

$$-f(t)\delta(t_0 - t)|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \delta(t_0 - t) f'(t) dt.$$

The first term is zero because $\delta(t - t_0)$ is zero at $\pm\infty$. The second term is $f'(t_0)$, according to Equation 22-4.

The analog to differentiation for discrete signals is the difference between adjacent elements. The discrete version of Equation 22-7 is

$$\sum_{n=-\infty}^{\infty} f[n]\delta'[k - n] = f[k] - f[k - 1] \quad \text{Equation 22-8}$$

22.1.3. The Unit Step Function

The step function, $u_s(t)$, is discontinuous at 0, with definition

$$u(t) = \begin{cases} 1, & t \geq 0 \\ 0, & t < 0 \end{cases}.$$

In discrete time, the step function $u[n]$ is similarly defined.

$$u[n] = \begin{cases} 1, & n \geq 0 \\ 0, & n < 0 \end{cases}$$

Just as $u_s(t)$ is the integral of $\delta(t)$,

$$u[n] = \sum_{k=-\infty}^n \delta[k].$$

22.2 Sinusoids

A continuous-time signal $x(t)$ with angular frequency ω_0 , is given by

$$x(t) = A \cos(\omega_0 t + \theta)$$

where A is the amplitude, $\omega_0 = 2\pi f_0$, f_0 being frequency in Hz, $f_0 = \frac{1}{T_0}$, and θ is the phase shift. For a discrete time sinusoid sampled every T_s seconds, replace t with nT_s .

$$x[n] = x(nT_s) = A \cos(\omega_0 nT_s + \theta).$$

The discrete signal may not have the same fundamental period as the continuous signal from which it is derived, as illustrated by Figure 22-2. Although the original signal has a period of 0.01 seconds, the sampled signal repeats itself over periods of 0.04 seconds.

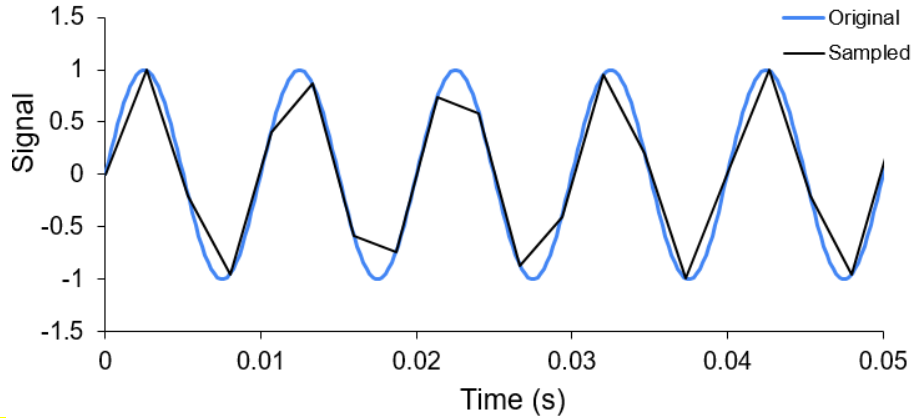


Figure 22-2: A 100 Hz signal sampled at 375 Hz. Four cycles of the original signal are required before the sampled signal repeats.

22.2.1. Fundamental Period

The sinusoid is discretized as

$$\cos(\omega_0 t) = \cos(\omega_0 nT_s)$$

From the definition of periodicity, the sampled signal is periodic with fundamental period N_0 if, for all integer times, n , $x[n] = x[n + N_0]$. I.e. N_0 is the number of samples required before the signal repeats itself. Then from the definition of periodicity

$$\cos(\omega_0 nT_s) = \cos(\omega_0 (n + N_0)T_s) = \cos(\omega_0 nT_s + \omega_0 N_0 T_s).$$

The phase difference $\omega_0 N_0 T_s$ is the number of radians required for the discrete sinusoid to repeat. It must be the smallest possible integer multiple of 2π , thus

$$\omega_0 N_0 T_s = 2\pi k \Rightarrow N_0 = \frac{k}{T_s f_0}.$$

Thus, we need to find N_0 such that $k = N_0 T_s f_0$ is an integer. N_0 will be finite only if $f_0 T_s$ can be expressed as a ratio of integer values. E.g., assume that $f_0 T_s = \frac{5}{12}$. Then $k/N_0 = 5/12$, or $N_0 = 12k/5$. The smallest value of k that makes N_0 an integer is 5, which makes $N_0 = 12$. I.e., after

12 samples, the discrete signal will repeat, and these 12 samples will represent 5 cycles of the analog signal.

The most important lesson from this analysis is that the fundamental period of a discrete signal is different from the frequency of the sine wave. Also, it is easy to test you on the fundamental period.

However, being able to calculate the fundamental period is probably far less useful than it sounds. The fundamental period exists only if the sample rate divided by the sinusoidal frequency is a rational number. The odds of it being a rational number in a physical measurement are infinitely small because one of the basic results of real analysis is that

$$\frac{\text{The number of irrational numbers}}{\text{The number of rational numbers}} = \infty$$

22.2.2. Dimensional Frequency

For a discrete sinusoid with sample time T_s , $x[n] = x(nT_s) = \cos(\omega_0 nT_s)$. If ω_0 and T_s are combined as $\omega_0 T_s$, they are a non-dimensional number, so $x[n] = \cos((\omega_0 T_s)n)$. Thus, for example, $\cos(0.2n)$ represents an unlimited number of sample times and signal frequencies as long as $\omega_0 T_s = 0.2$. Once either ω_0 or T_s is known, we can calculate the other. For example, if

$$x[n] = \cos(0.03n)$$

and if we know that the sample time is 0.001 seconds, then the angular frequency is $\omega_0 = 0.03/0.001 = 30$ radians/s.

22.2.3. Aliasing

Aliasing occurs when the sample frequency is slow low to capture the high frequency changes in a signal. The Nyquist theorem^{*} states that a signal can be reconstructed from its sampled version if the signal has no components with frequencies greater than $f_s/2$, where f_s is the sample frequency. If higher frequencies are present in the original signal, they will appear at aliased frequencies in the reconstructed signal. Computation of the frequencies at which signals will be aliased was covered in Section 19.5.

22.2.4. Sampling Considerations for Biomedical Measurements

Table lists several of the basic measurements that are performed clinically, along with the range of the variable, the frequencies of interest, and the methods through which the measurements are made. Requirements for sampling frequency are relatively modest, except for electromyography.

Measurement	Range	Frequency (Hz)	Method
-------------	-------	----------------	--------

^{*} This theorem is also referred to as the Nyquist-Shannon theorem or the Whittaker–Nyquist–Shannon theorem. Here, it will be referred to simply as the Nyquist theorem.

Blood flow	1 to 300 mL/s	0 to 20	Electromagnetic or ultrasonic
Blood pressure	0 to 400 mmHg	0 to 50	Cuff or strain gage
Cardiac output	4 to 25 L/min	0 to 20	Fick, dye dilution
Electrocardiography	0.5 to 4 mV	0.05 to 150	Skin electrodes
Electroencephalography	5 to 300 μ V	0.5 to 150	Scalp electrodes
Electromyography	0.1 to 5 mV	0 to 10000	Needle electrodes
pH	3 to 13 pH units	0 to 1	pH electrode
pCO ₂	40 to 100 mmHg	0 to 2	pCO ₂ electrode
pO ₂	30 to 100 mmHg	0 to 2	pO ₂ electrode
Respiratory rate	2 to 50 breaths/min	0.1 to 10	Impedance
Temperature	32 to 40 °C	0 to 0.1	Thermistor

22.3 Signal Transformations

22.3.1. Time Shifting

The signal $f[n - k]$ has the same shape and amplitude as the signal $f[n]$, but is shifted to the right by k places (Figure 22-3).

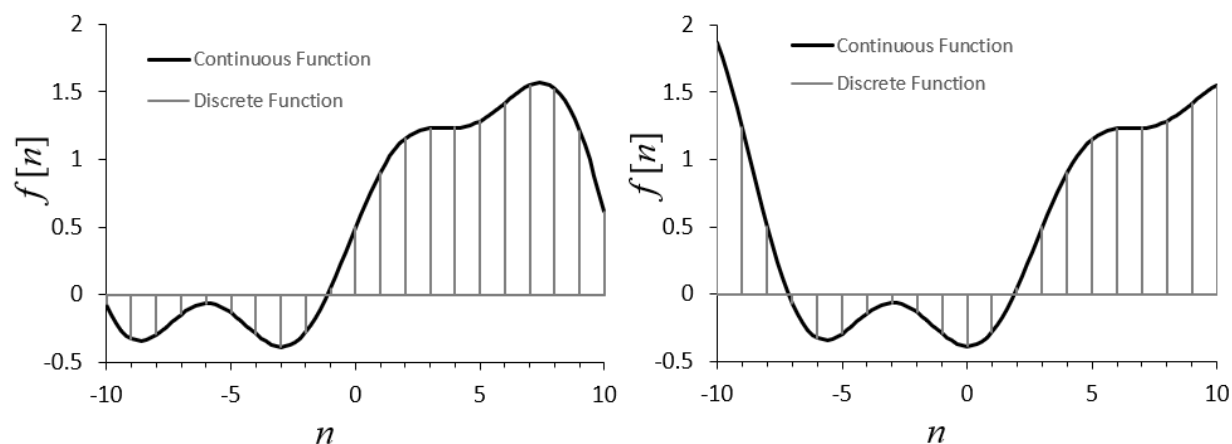


Figure 22-3: An arbitrary signal $f[n]$ unshifted (left) and its counterpart $f[n - k]$ shifted to the right by 3 time steps (right).

22.3.2. Time Reversal

The function $f[-n]$ is the mirror image of $f[n]$, reflected across the $n = 0$ axis (Figure 22-4).

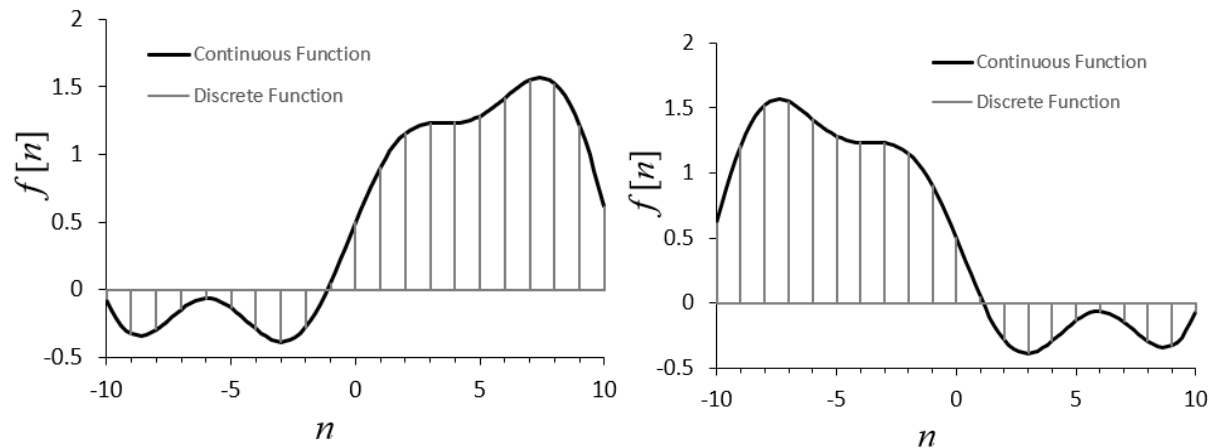


Figure 22-4: The function $f[n]$ (left) and its mirror image $f[-n]$ (right).

22.3.3. Time Expansion/Contraction

The function $f[\alpha n]$ has the same shape as $f[n]$ but is compressed if $\alpha > 1$ and expanded if $\alpha < 1$. The directionality (compressed or expanded) is easy to remember. If the signal is a sine wave, for example, and the argument changes from $\omega_0 t$ to $2\omega_0 t$, the new sine wave has double the frequency, which means it has more cycles in a given amount of time and therefore looks compressed. Similarly, an α less than 1 leads to a lower frequency and a sine wave that appears expanded.

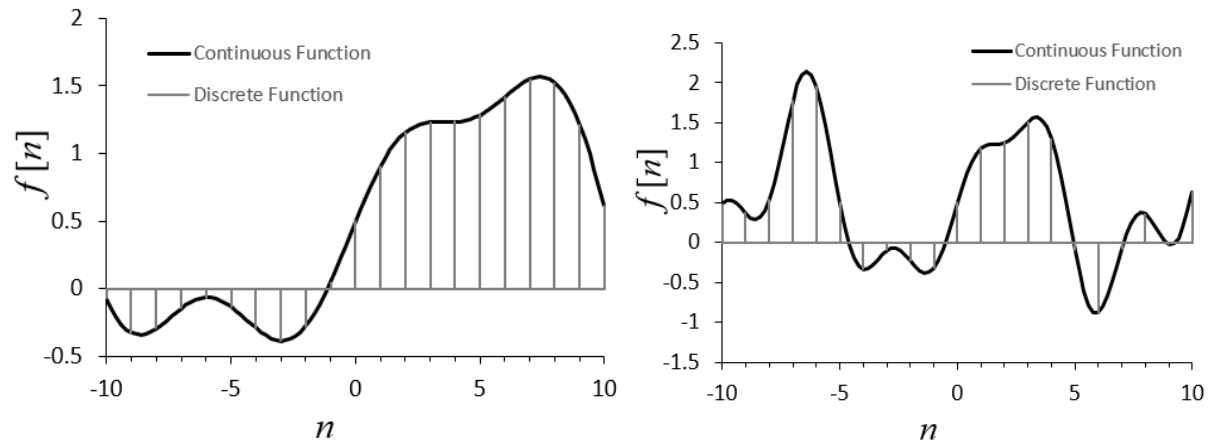


Figure 22-5: The function $f[n]$ (left) and its compressed version $f[2.2n]$ (right).

22.3.4. Amplification

22.3.5. Offset Shift

22.3.6. Addition

22.3.7. Windowing

In Fourier Analysis

In Sampling

22.3.8. Convolution

Chapter 23: Digital Fourier Analysis

23.1 Flavors of Fourier Analysis

The student has been introduced to several types of Fourier analysis, such as the Fourier transform, Fourier series, discrete Fourier transform, and fast Fourier transform. It is important to understand the differences between these techniques and the conditions under which they can be used.

23.1.1. Fourier series

(Show the reconstruction of a carotid artery waveform.)

(Show the reconstruction of a square wave. Why does it not work, technically?)

23.1.2. Fourier transform (continuous)

23.1.3. Discrete Fourier transform

23.1.4. Fast Fourier transform

23.2 Review of Fourier Transform Properties

23.2.1. Positive and Negative Frequencies

23.2.2. Symmetry for a Real Signal

23.2.3. Examples of Square and Triangle Wave

23.2.4. Non-Symmetry for a Complex Signal

23.2.5. Example of a Complex Signal (Quadrature)

23.3 Discrete Fourier Transform

The most complicated process in the use of Fourier analysis on digital signals is the fast Fourier transform (FFT) algorithm. Because nearly any software package, including MATLAB, has a built-in FFT function, the details of this process will not be described in depth.* However, appropriate use and interpretation of the transform requires that both the spectral frequencies and the magnitude be correctly scaled. Built in software will not typically return a frequency array, and the transform itself will not be scaled. Furthermore, the scaling of the transform depends on the type of spectrum to be generated, i.e., the amplitude spectrum or the power spectrum. This chapter is mainly concerned with tests and methods that you can use to ensure that the scaling is correct for a given application.

* You should know, however that (1) it is based on the Cooley-Tuckey algorithm and (2) it employs some clever mathematical tricks.

23.3.1. Scaling the Frequency Axis

To generate a correctly scaled frequency axis for the transform, one must know the number of data points (record length) processed by the FFT and the sample frequency. Four basic rules will be needed.

1. The maximum frequency of the FFT spectrum is half the sampling frequency of the signal ($\frac{1}{2}f_s$).
2. The minimum frequency of the FFT spectrum is $-\frac{1}{2}f_s$.
3. The total number of output data points in the FFT is equal to the total number of input data points fed to the routine.
4. If N data points are fed into the FFT, the first $N/2$ data points of the output correspond to positive frequencies, beginning at zero, and the second $N/2$ data points correspond to negative frequencies, beginning at the negative of the Nyquist frequency.

Assume that N data points are fed into the FFT, and that these data points are sampled at a data rate of f_s . The frequencies along the horizontal axis will be separated by $\Delta f = \left(\frac{1}{2}f_s\right) / \left(\frac{N}{2}\right) = \frac{f_s}{N}$. It may be easier simply to remember that the frequencies span from the negative of the Nyquist frequency to the Nyquist frequency, so that the entire spectrum must range between $-f_s/2$ and $f_s/2$ and that the spectrum has N data points, so that one can easily remember that $\Delta f = f_s/N$ without having to worry about the factors of 2. Because the positive frequencies are considered by convention to start at 0, and because we choose to have $N/2$ data points in each of the positive and negative parts of the spectrum, the positive part of the spectrum is considered to go up to, but not to include, $f_s/2$, so the last positive frequency is $f_s/2 - \Delta f$.* With these concepts in mind, we can generate an array for the frequency axis that starts at $-f_s/2$, goes to $+f_s/2 - f_s/N$, and has data points separated by f_s/N . In MATLAB, the command to create this array, given f_s and N is as follows.

```
deltaf = fs/N;
fmax = fs/2; % maximum frequency for the spectrum
farray = -fmax:deltaf:fmax-deltaf;
```

The concepts are easily visualized by the axis in Figure 23-1.

.

* We could just as easily make the extreme negative data point $-f_s/2 + \Delta f$ because the FFT is periodic, so mathematically the data point at $+f_s/2$ is equal to the data point at $-f_s/2$.

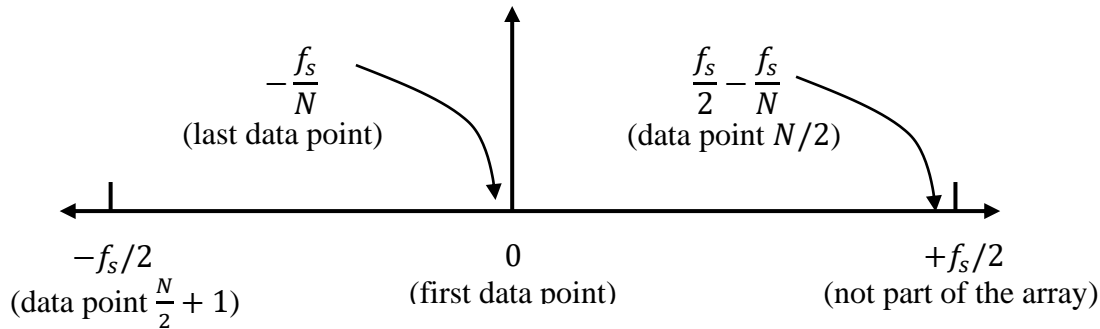


Figure 23-1: Frequency axis for the scaled FFT-based spectrum. The first data point corresponds to zero frequency. Data point $N/2$ corresponds to $f_s/2 - \Delta f$, where $\Delta f = f_s/N$. Data point $N/2 + 1$ corresponds to $-f_s/2$, and data point N corresponds to $-\Delta f$.

Next, the output of the FFT must be rearranged so that the amplitude array lines up with the frequency array. The amplitude array from the FFT algorithm will align with frequencies as in Figure 23-2.

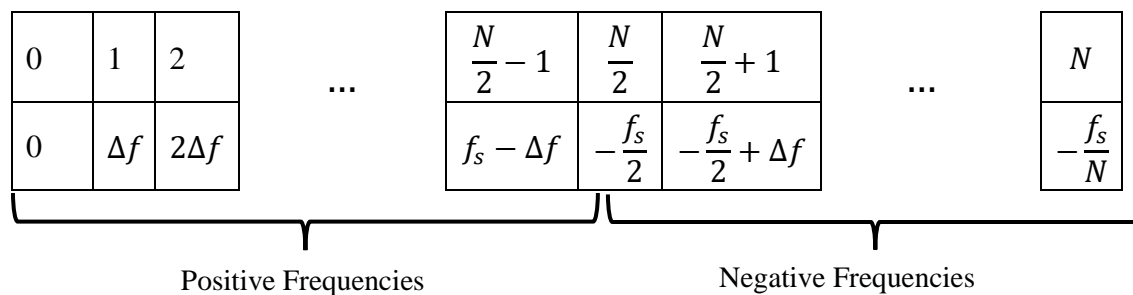


Figure 23-2: Alignment of the amplitude array from the FFT routine with frequencies. Element 0 aligns with 0 Hz. Element $N/2$ aligns with the negative Nyquist frequency.

For plotting purposes, we would like the first amplitude in the array to correspond to the lowest frequency instead of zero, so we need to move the second half of the array to the first half and the first half to the second half. The rearrangement is not difficult to code, but one has to make sure that one does not write over the parts of the array before they have been moved. For example, if we want to swap the positions of the $-f_s/2$ amplitude with that of the 0 amplitude, we cannot simply write

```
array(N/2) = array(1);
array(1) = array(N/2);
```

because the value of `array(N/2)` will have been overwritten by `array(1)` on the second line. We would need to write

```
temp = array(N/2);
array(N/2) = array(1);
array(1) = temp;
```

MATLAB, however, saves us some work because it has the built-in command `circshift` that shifts the array as needed. The shift is thus performed with the following command.

```
array = circshift(array, N/2);*
```

23.3.2. Amplitude Spectrum (Fourier Transform)

The next task is to ensure that the physical dimensions are correct for the transform. We must ensure that the forward transform is correct and that the reverse transform yields the original signal, with the original scaling and the original units. We also need to ensure that the magnitude is independent of the number of data points used in the transform and the frequency of the test signal. A simple test signal is the pulse function of width W . The signal (Figure 23-3) is defined as

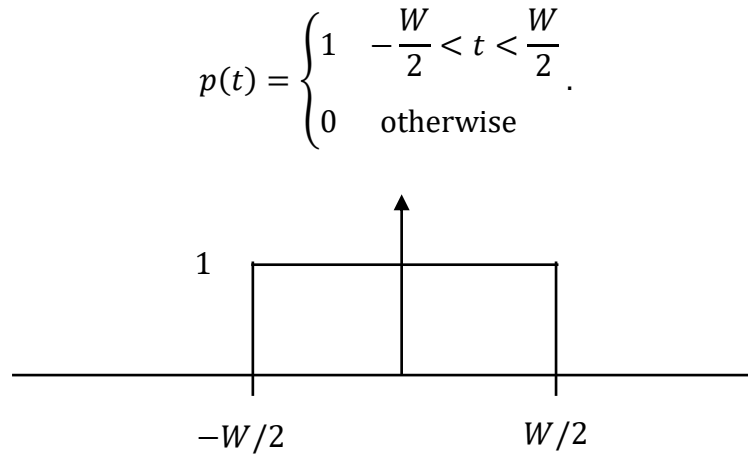


Figure 23-3: A convenient test signal to ensure that the spectrum is correctly scaled.

The transform is

$$\begin{aligned} \mathcal{F}\{p(t)\} &= \int_{-\frac{W}{2}}^{\frac{W}{2}} e^{j\omega t} dt \\ \mathcal{F}\{p(t)\} &= \frac{1}{j\omega} e^{j\omega t} \Big|_{-\frac{W}{2}}^{\frac{W}{2}} \\ \mathcal{F}\{p(t)\} &= \frac{1}{j\omega} (e^{j\omega W/2} - e^{-j\omega W/2}) \end{aligned}$$

But

$$\sin(\omega W/2) = \frac{e^{j\omega W/2} - e^{-j\omega W/2}}{2j}$$

so

* This code snippet and the code snippet above Figure 23-2 are the main practical points that need to be derived from this section.

$$\mathcal{F}\{p(t)\} = \frac{2j \sin(\omega W/2)}{j\omega}.$$

This function is typically manipulated slightly to the form

$$\mathcal{F}\{p(t)\} = \frac{W \sin(\omega W/2)}{\omega W/2},$$

and a new function, the sinc function, is defined as

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

so that we can write

$$\mathcal{F}\{p(t)\} = W \text{sinc}\left(\frac{\omega W}{2}\right).$$

The maximum value of the function occurs at $\omega = 0$, and it is

$$\mathcal{F}\{p(t)\}|_{\max} = \lim_{\omega \rightarrow 0} \left(\frac{W \sin(\omega W/2)}{\omega W/2} \right) \rightarrow W \lim_{\omega \rightarrow 0} \left(\frac{\omega W/2}{\omega W/2} \right) \rightarrow W$$

Thus, for a correctly scaled FFT, the output for a pulse of width W must have a DC value of W^* . However, because the FFT takes the amplitude data and the number of points as input parameters, it receives no indication of the time scale, and hence the value of W . We will begin with the following code.

```
% Scaled spectrum
W = 0.2; % (seconds) Start with an arbitrary value of W, here 0.3
seconds
Npts = 2048; % We will use 1024 data points in the time series
p = zeros(1, Npts);
fs = 200; % set the sample rate to 200 Hz, so we will have about 10
seconds of data total.
dt = 1/fs; % time between samples
Ttotal = dt*Npts;
fm = fs/2; % maximum frequency for the spectrum
deltaf = fs/Npts;
farray = -fm:deltaf:fm-deltaf; % Generate a scaled frequency array.
Nlast = W/(2*dt); % half width of the pulse in terms of data points
p(1:Nlast) = ones(1, Nlast); % fill the positive-time part of the pulse
p(Npts-Nlast+1:Npts) = ones(1, Nlast); % fill the negative-time part
of the pulse
time = 0:dt:Ttotal-dt;
```

* Because the input to the FFT starts at $t = 0$, we need to be clever in the way that we input the signal from $-W$ to 0. Because the FFT is discrete, it implicitly assumes that the input is periodic, so if the input ranges from 0 to T , the input from 0 to $\frac{W}{2}$ is the part of the signal for positive time, and the input from $T - \frac{W}{2}$ to T is the input for negative time. Therefore, the code that creates the signal array is broken into two pieces, one at the beginning of the array ($p(1:Nlast)$), and one at the end of the array ($p(Npts-Nlast+1:Npts)$).

```

figure(1);
plot(time,p);
ylim([0 1.5]);
spectrum = fft(p, Npts);
spectrum = circshift(spectrum,Npts/2);
figure(2);
plot(farray,real(spectrum));
xlabel('frequency (Hz)');
ylabel('Amplitude');

```

The last MATLAB plot generated by this code is shown in Figure 23-4. The peak at zero frequency is 30, which is much greater than the expected value of 0.2. Therefore, we need some scaling factor. Since we notice that the ratio of the obtained maximum to the expected maximum is $40/0.2 = 200$, we can guess that the output must be scaled by the sampling frequency. We therefore add the following line after the circshift command.

```
spectrum = spectrum/fs;
```

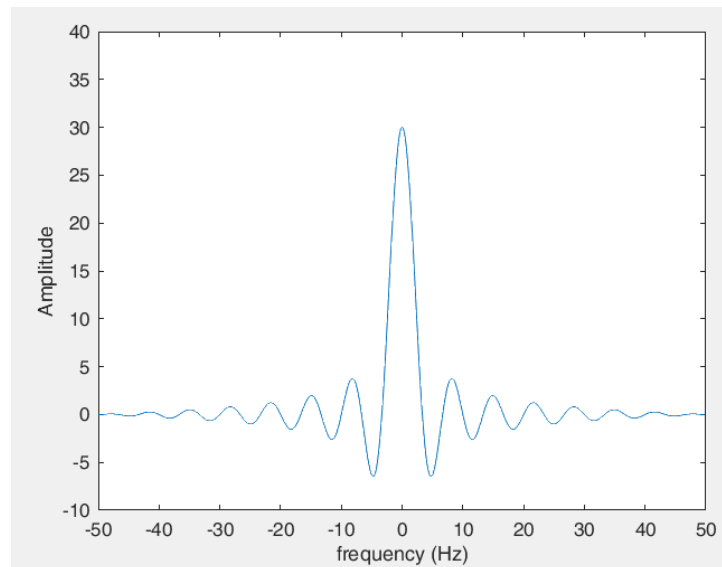


Figure 23-4: Sinc function generated by the MATLAB code, prior to scaling.

To determine whether this scaling is sufficient, the code is run for different values of N_{pts} , W , f_s . The other parameters do not need to be considered because they can all be calculated from these three. The tests show that the spectrum always peaks at a value of W , indicating that the scaling is correct. *Even if you are given a subroutine that was designed to correctly scale the spectrum, you should use the sinc function as a test signal to ensure that the code works as indicated and that you are using the code correctly.*

23.3.3. Amplitude Spectrum (Fourier Series)

Scaling for the Fourier series differs from scaling for the Fourier transform. The Fourier series for the signal $A \cos(2\pi f_0 t)$ has the component $\frac{A}{2}$ at $\pm f_0$. However, if a cosine wave of amplitude A and frequency f_0 is scaled according to the previous section, the value of the spectrum at frequencies $\pm f_0$ will not be $\frac{A}{2}$ because the Fourier *transform* for this signal is $\frac{A}{2}(\delta(\omega - \omega_0) + \delta(\omega + \omega_0))$. The transform amplitude is infinite at the frequency locations of the two delta functions. For a discrete signal, the delta functions must be interpreted in terms of their integrals, rather than their peak value. Recall that part of the definition of the delta function is that $\int_{-\infty}^{\infty} \delta(\omega) d\omega = 1$, so that a cosine wave $\int_{-\infty}^{\infty} \mathcal{F}(\omega) d\omega = \int_{-\infty}^{\infty} \frac{A}{2}(\delta(\omega - \omega_0) + \delta(\omega + \omega_0)) d\omega$, which is equal to A . As a result of the $d\omega$ in this integral, the amplitudes of the two peaks of \mathcal{F} must get larger as the spacing between frequencies on the horizontal axis becomes smaller.

It may in some cases be desirable to interpret the FFT as a Fourier series instead of a Fourier transform, so that the magnitudes of the sinewaves are recovered. To obtain that spectrum one must divide the spectrum by the number of input points, rather than the sample frequency. Note, however, that if the signal's frequency is between two of the frequencies calculated on the frequency axis, the component will be partly represented by the frequency above it and partly by the frequency below it, so the amplitude will not be exactly equal to the cosine amplitude.

23.3.4. Energy Spectrum and Power Spectrum

The energy spectrum of a signal is defined as

$$S_{xx}(\omega) = |\mathcal{F}(\omega)\mathcal{F}^*(\omega)|. \quad \text{Equation 23-1}$$

where \mathcal{F}^* denotes the complex conjugate of \mathcal{F} . This definition is appropriate for a function that is defined over a finite period of time, as in a pulse function. However, if one is interested in a stochastic process, such as steady flow turbulence that is interpreted as a continuous signal from $-\infty < t < \infty$, the Fourier transforms in Equation 23-2 do not converge, and a different approach is needed. First, a Fourier transform is defined over a finite time.

$$\hat{\mathcal{F}}(\omega) \equiv \int_0^T x(t)e^{-i\omega t} dt \quad \text{Equation 23-2}$$

Here, the limits of the integral are from 0 to T instead of $-\infty$ to ∞ . The power spectrum of the signal is then defined as

$$P(\omega) = \lim_{T \rightarrow \infty} E \left\{ \frac{1}{T} \hat{\mathcal{F}}(\omega) \hat{\mathcal{F}}^*(\omega) \right\} \quad \text{Equation 23-3}$$

Equation 23-3 recognizes that the energy of the signal is infinite, so one must normalize it by time and use power instead. The operator E is expected value (Section 6.2.1). Thus, over a given time window, the spectral content at frequency ω may not be $P(\omega)$, but the spectra

averaged over the entire signal will converge to $P(\omega)$. From a practical perspective, if the signal is stationary,* the power spectrum can be approximated by

$$P(\omega) = \frac{1}{NT} \sum_{k=1}^N \hat{F}(\omega) \hat{F}^*(\omega) . \quad \text{Equation 23-4}$$

This equation is the basis of the periodogram, which will be discussed in Section 23.3.6. A test of the correct scaling for the power spectrum depends on the application of Parseval's theorem.

23.3.5. Parseval's Theorem

Parseval's theorem relates the energy in a signal to the energy spectrum.

$$\int_{-\infty}^{\infty} |s(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} E(\omega) d\omega = \int_{-\infty}^{\infty} E(f) df \quad \text{Equation 23-5}$$

where $E(\omega)$ is the energy spectrum at frequency ω (radians/sec). The expression on the right-hand side follows from the expression in the middle because $df = \frac{d\omega}{2\pi}$. This theorem demonstrates that the energy in the signal (on the left-hand side) must be the integral of the energy spectrum over all frequencies. To ensure that one has correctly scaled the output of a function or subroutine that calculates an energy spectrum, input a function with a known energy into the energy spectrum function, then numerically integrate the output and ensure that the integral of the signal squared over time is equal to the integral of the spectrum over frequency.

If the signal has a finite duration, T , we can divide Parseval's theorem by T to rewrite it in terms of signal power instead of signal energy, where $P(f) = E(f)/T$.

$$\frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} |s(t)|^2 dt = \int_{-\infty}^{\infty} P(f) df \quad \text{Equation 23-6}$$

This expression states that the total power in the signal can be obtained from the integral of the power spectrum, and it can be used to ensure that the power spectrum is correctly scaled, as will be shown below.

23.3.6. Periodogram

Often, one wishes to calculate the power spectrum directly from a time series. The technique for this calculation involves the periodogram, defined as

$$P_k(\omega) = \frac{1}{T} \left| \int_{kT}^{(k+1)T} s(t) e^{i\omega t} dt \right|^2 \quad \text{Equation 23-7}$$

* Stationary means that the underlying statistics of the signal do not change over time. The pressure fluctuations in a turbulent flow, for example, would be stationary if the flow rate does not change, but if the flow does change, it changes the underlying statistics and leads to a nonstationary signal.

This calculation must be performed carefully for stochastic signals. Typically, one will divide the signal into N equal time windows of length T , and then average together the N periodograms.

$$P(\omega) = \frac{1}{N} \sum_{k=0}^{N-1} P_k(\omega) \quad \text{Equation 23-8}$$

For example, if the signal duration is 100 seconds, sampled at 1024 samples/second, and if the frequency resolution that is desired in the spectrum is $\Delta f \approx 1$ kHz, the signal can be broken into 100 windows of 1 second each. This process ensures that the power spectrum consists of 1024 frequency components, each of which is an average of multiple independent estimates, where each window provides one estimate. If, instead, one elects to calculate the Fourier transform of all 102400 points at once, the power spectrum will contain 102400 frequencies, each of which is only a single random estimate. This alternative approach does not lead to a converged spectrum, but rather to a larger number of spectral components, each of which is strongly random.

The power spectrum defined in Equation 23-8 differs from the single spectrum in several ways. First, it is related to the magnitude squared of the power spectrum, and as such, it is a real valued function. Second, it is averaged over several relatively short signal records. It contains no phase information. It is used when one wishes to find the frequency content of a (more or less) continuous signal, typically a stochastic (random) signal, such as white noise, an electroencephalogram, an electromyogram, or the velocity in a turbulent flow (which in turn is related to the relative sizes of the eddies in the flow). Assume that we have fluid velocity data $v(t)$ (perhaps taken with a hot film anemometer) collected for a relatively long time T_{total} , sampled at a frequency f_s . We will break this data series into N shorter records, each of which is of length $T = T_{\text{total}}/N$ and calculate the Fourier transform of each record. Each of the shorter records will have a length of $N_{\text{rec}} = (T)(f_s)$ points. We will then calculate the magnitude-squared of the transformed data for each record and average the resulting spectra, frequency by frequency. From Equation 23-2, the dimension of each individual transform is velocity-time. From Equation 23-3, the dimensions of the power spectrum will be $\frac{(\text{velocity-time})^2}{\text{time}} = \text{velocity}^2\text{-time}$, which, because the spectrum is in the frequency domain rather than the time domain, is more meaningfully written as $\frac{\text{velocity}^2}{\sqrt{\text{Hz}}}$. We need to make sure that the power is scaled correctly and that it does not change with T , N , N_{rec} , and that it correctly takes into account f_s . To ensure this proper scaling, we will use a test signal $A \sin(\omega t)$ along with Parseval's theorem in the form of **Equation 23-6**. The integral on the left of **Equation 23-6** is relatively easy to evaluate for a sine wave if we assume that T is a integral number of cycles of the sine wave.

$$\frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} |s(t)|^2 dt = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} A^2 \sin^2(\omega t) dt$$

Use the identity $\sin^2(x) = \frac{1}{2}(1 - \cos(2x))$.

$$\frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} |s(t)|^2 dt = \frac{A^2}{2T} \int_{-\frac{T}{2}}^{\frac{T}{2}} (1 - \cos(2\omega t)) dt$$

If T is n cycles of $\sin(\omega t)$, then it is $2n$ cycles of $\cos(2\omega t)$, so the integral over $\cos(2\omega t)$ is zero, and the integral becomes $A^2/2$.

To find the correct scaling, we will begin with the following section of code.

```
% Calculate the power spectrum of "signal"
dt = 0.001;
npts = 256;
Nrecs = 10;
t = 0:dt:(Nrecs*npts-1)*dt;
w = 2*pi*200; % frequency of the signal in radians/s
A = 5.; % signal amplitude
signal = A*sin(w*t);
ntotal = length(signal);
T = ntotal*dt;
% create the frequency array for the power spectrum
fmax = 1/(2*dt);
deltaf = 2*fmax/npts;
farray = -fmax:deltaf:fmax-deltaf;

% Determine how many records to average
nrecs = ntotal/npts;

% initialize the output array
pspec = zeros(1,npts,'double');

for k = 1:nrecs % loop for each record to be averaged

    b = fft(signal((k-1)*npts+1:k*npts)); % take the FFT
    pspec = pspec + abs(b).*abs(b); % find and sum the power
    spectrum

end

pspec = circshift(pspec,[0 npts/2]); % Rearrange the spectrum

% Approximate the integrals in parseval's theorems with sums
psignal = sum(signal.*signal)*dt/T; % Calculate power from the
    signal.
pspectrum = sum(pspec)*deltaf; % Calculate power from the
    spectrum.

fprintf('Power in the signal is %f\n',psignal);
fprintf('Power in the spectrum is %f\n',pspectrum);

% Plot the power spectrum
```

```
plot(farray, pspec);
```

When the code is run with an amplitude of 5, the power in the signal as calculated directly from the signal is 12.5, which agrees with our theoretical value ($12.5 = 5^2/2$). However, the value calculated from the spectrum is 32,000,000, indicating that some normalization is needed. We will first check to see if the output depends on the number of points per spectrum. When `npts` is changed from 256 to 512, The power calculated from the spectrum doubles, so we will normalize by `npts`. Now the result is 12,500, which is exactly 12.5 multiplied by the sample frequency multiplied by the number of records. We therefore will divide the power spectrum by the sample frequency (which is the same as multiplying it by the sample time) and divide by the number of records. Once these adjustments are made, the power from the spectrum is independent of the number of records, the sampling frequency, and the number of points per record. It also changes as expected when the amplitude changes. In all, to get the correctly scaled spectrum we needed only to add the line

```
pspec = pspec*dt/npts/Nrecs; % Scale the power spectrum
```

The spectrum is as shown in Figure 23-5: Power spectrum of the 200 Hz sine wave..

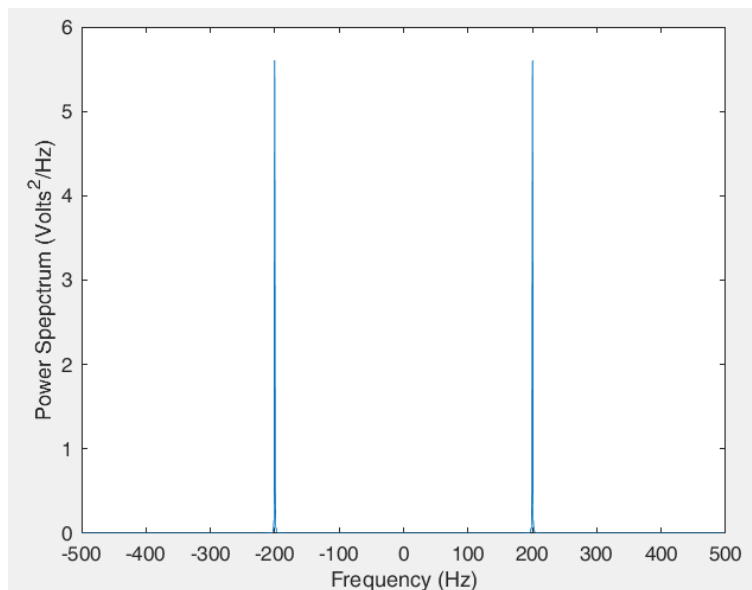


Figure 23-5: Power spectrum of the 200 Hz sine wave.

The height of the two spikes is not 5 or 12.5 because the power is the integral under the spectrum, not the amplitude of the spectrum. Thus if h is the height of each spike, the power under each spike is $h\Delta f$, so here $2h\Delta f = 12.5$. Because the Δf on the frequency axis depends on the number of points in the spectrum and on f_{\max} , the height of the spikes will change if we change these two parameters. Because f_{\max} depends on the sampling frequency, the height of the spikes will also depend on dt .

23.4 Data Simulation

Data simulation is useful for the validation of data processing algorithms and as a component of physiological models. Data sequences are generally categorized as periodic (deterministic) or random (stochastic).

23.4.1. Periodic Signals

Sinusoids

Square Waves

Triangle and Sawtooth Waves

Periodic Signals Based on Fourier Series

23.4.2. Random Signals

Gaussian White Noise

Probability Density

Spectral Shape

23.5 Graphical User Interface (GUI)

23.6 Simulink

Chapter 24: Digital Filtering

24.1 Filtering

24.1.1. Applications of Filtering

24.1.2. Advantages of Digital vs. Analog Filtering

24.2 Some Practical MATLAB Filtering Commands

Most of the Matlab filtering commands are performed in two steps. The first step generates the numerator and denominator of the transfer function (the ARMA coefficients), and the second step applies the filter to the signal.

24.2.1. Butter

The `butter()` command applies a Butterworth filter of type (type is 'low', 'high', 'bandpass', or 'stop') the specified order (order) and the specified cutoff frequency (fcutoff) to a signal. The command syntax is

```
[b, a] = butter(order,fcutoff,type);
```

Where the filter coefficients b and a are applied to the signal with

```
s_filtered = filter(b,a,s_unfiltered);
```

Variables "s_unfiltered" and "s_filtered" are the unfiltered and filtered versions of the signal, respectively. The b array contains the numerator of the transfer function and the a array contains the denominator of the transfer function. The cutoff frequency, fcutoff, is specified as a fraction of the Nyquist frequency. If the Nyquist frequency is, for example, 2 kHz and the filter cutoff is to be 300 Hz, then fcutoff must be equal to $300/1000 = 0.3$. For a bandpass or stopband filter, fcutoff is specified with a two-element array; to bandpass filter a signal between 100 Hz and 300 Hz, with a sample frequency of 2000 Hz, use the command

```
[b, a] = butter(order,[0.1 0.3],'bandpass');
```

If 'bandpass' or 'stop' is selected, the order of the filter is $2n$ (because the filter is equivalent to cascading two Butterworth filters of order n).

24.2.2. Lowpass, Highpass, and Bandpass

These filters are somewhat easier to implement, requiring only two steps, but are not recommended by this author because they allow less control over the filter structure. Based on previous experience, manipulation of these filters by students leads to unpredictable results. Nonetheless, the syntax (for the lowpass version) is

```
s_filtered = lowpass(s_unfiltered,fcutoff);
```

24.3 Heart Rate Variability

24.3.1. Thresholding

24.3.2. Resampling

Chapter 25: Analog and Digital Domains

25.1 Analog Signal Pre-Processing

25.1.1. Anti-Aliasing Filter

25.1.2. Amplification, Attenuation, and Offset

25.1.3. Pre-Whitening

25.2 Digital Operations

25.2.1. Offset Subtraction

25.2.2. Windowing

25.2.3. Fourier Analysis

25.2.4. Cross Correlation

25.2.5. Neural Networks

25.2.6. Lyapunov Exponent

25.2.7. Autoregressive and ARMA Analysis

25.2.8. Feature Extraction

Filtering

Rectification

Thresholding

25.2.9. Graphical Presentation

Chapter 26: Discrete Functions

26.1 Applications of Digital Systems

26.1.1. Low Pass

26.1.2. High Pass

26.1.3. Band Pass

26.1.4. Notch

26.1.5. Pre-Whitening

26.2 AR, MA, and ARMA Systems

A discrete-time system is a device or process that accepts a discrete-time signal $x[n]$ as its input and produces another discrete-time signal $y[n]$ as its output. If the system is time invariant, the output depends on past and present values of $x[n]$ and on past values of $y[n]$.^{*} Time invariant systems are categorized as moving average (MA), autoregressive, or autoregressive moving average (ARMA). MA and autoregressive filters are special cases of the more general ARMA filter.

26.2.1. Difference Equations

Discrete approximation of the derivative

The difference equation is the DT counterpart to a continuous-time differential equation. For the first order differential equation

$$\frac{dy}{dt} + \alpha y(t) = \beta x(t), \quad \text{Equation 26-1}$$

recall that the derivative is defined by

$$\frac{dy}{dt} = \lim_{\Delta t \rightarrow 0} \left\{ \frac{(y(t) - y(t - \Delta t))}{\Delta t} \right\}.$$

Sample the signals y and x with time steps of Δt to convert the equation to discrete time.

$$\begin{aligned} x(t) &\rightarrow x(n\Delta t) = x[n] \\ y(t) &\rightarrow y(n\Delta t) = y[n] \end{aligned}$$

$$\frac{dy}{dt} \approx \frac{y[n] - y[n-1]}{\Delta t}, \text{ for } \Delta t \text{ small but } \Delta t \neq 0 \quad \text{Equation 26-2}$$

Equation 26-1 is then approximated by

^{*} A non-causal system would depend also on future values of $x[n]$ and $y[n]$.

$$\frac{y[n] - y[n-1]}{\Delta t} + \alpha y[n] = \beta x[n] \quad \text{Equation 26-3}$$

Recursive Format

Recursive format is an equation form where the current value of y (i.e. $y[n]$) is written in terms of past values of y and past and present values of x . It requires only that one solve the difference equation (in this case, Equation 26-3) for $y[n]$.

$$y[n] - y[n-1] + \alpha \Delta t y[n] = \beta \Delta t x[n]$$

$$y[n] + \alpha \Delta t y[n] = \beta \Delta t x[n] + y[n-1]$$

$$y[n] = \frac{1}{1 + \alpha \Delta t} (\beta \Delta t x[n] + y[n-1])$$

This equation has the form

$$y[n] = ax[n] + by[n-1],$$

where the constant coefficients a and b are

$$a \equiv \frac{\beta T_s}{1 + \alpha T_s}; \quad b \equiv \frac{1}{1 + \alpha T_s}$$

I.e., to obtain $y[n]$, update a previous value $y[n-1]$ through a scaling factor a added to a scaled value b of the current input $x[n]$.

Example: What is the difference equation in recursive format of the following differential equation with a sampling frequency $f_s = 100$ Hz?

$$\frac{dy}{dt} + 7y(t) = 11x(t)$$

Answer: The equation is discretized as

$$\frac{y[n] - y[n-1]}{T_s} + 7y[n] = 11x[n] \Rightarrow y[n] + 7\Delta t y[n] = 11T\Delta t_s x[n] + y[n-1]$$

$$\Rightarrow y[n] = \frac{11\Delta t}{1 + 7\Delta t} x[n] + \frac{1}{1 + 7\Delta t} y[n-1].$$

Δt is the reciprocal of the sample frequency.

$$\Delta t = \frac{1}{f_s} = \frac{1}{100} = 0.01$$

$$y[n] = 0.1028 x[n] + 0.9346 y[n-1]$$

Although the concern for this course is to use the discrete version of the signals and equation for digital filtering, this decomposition can be used as a simple method to solve a differential

equation* numerically. For example, assume that one wishes to solve Equation 26-1 with the initial condition $y(0) = 3$ and input function $x = u_s(t)$. For the first step ($n = 1$),

$$y[1] = ax[1] + by[0], = 1a + 3b$$

For the second step ($n = 2$)

$$y[2] = ax[2] + by[1] = 1a + b(1a + 3b).$$

The process is continued for each step n . Generally, if the time step Δt is small enough so that the limit in the derivative definition is approximated closely, the result will be close to the analytic solution to the problem.

Example: The solution to the equation

$$\frac{dy}{dt} + 2y = u_s(t); \quad y(0) = 0$$

Is $y = \frac{1}{2}(1 - e^{-2t})u_s(t)$. Use Excel to plot this solution along with an approximate solution obtained through the above method.

Answer:

The equation is discretized as

$$\begin{aligned} \frac{y[n] - y[n-1]}{\Delta t} + 2y[n] &= u_s[n] \\ y[n] + 2\Delta t y[n] &= u_s[n]\Delta t + y[n-1] \\ y[n] &= \frac{1}{1 + 2\Delta t} (u_s[n]\Delta t + y[n-1]) \end{aligned}$$

This equation is coded in Excel as follows, where Column A is the time base, Column B is the analytic solution, and Column C is the numerical solution. The variable dt is defined as Δt .

	A	B	C
1	0	=(1-EXP(-2*A1))/2	0
2	=A1+dt	=(1-EXP(-2*A2))/2	=1/(1+2*dt)*(dt+C1)
3	=A2+dt	=(1-EXP(-2*A3))/2	=(1-2*dt)*(dt+C2)
4	=A3+dt	=(1-EXP(-2*A4))/2	=(1-2*dt)*(dt+C3)
5	=A4+dt	=(1-EXP(-2*A5))/2	=(1-2*dt)*(dt+C4)
6	=A5+dt	=(1-EXP(-2*A6))/2	=(1-2*dt)*(dt+C5)
7	=A6+dt	=(1-EXP(-2*A7))/2	=(1-2*dt)*(dt+C6)
8	=A7+dt	=(1-EXP(-2*A8))/2	=(1-2*dt)*(dt+C7)
9	=A8+dt	=(1-EXP(-2*A9))/2	=(1-2*dt)*(dt+C8)
10	=A9+dt	=(1-EXP(-2*A10))/2	=(1-2*dt)*(dt+C9)

The results are shown in Equation 26-1. Even for a relatively crude step of 0.2 seconds, the numerical solution follows the character of the analytic solution, but the numerical solution approaches the analytic solution as $\Delta t \rightarrow 0$.

* The distinction between filtering and solving a differential equation should be recognized as artificial, as the system described by the differential equation is a low pass filter.

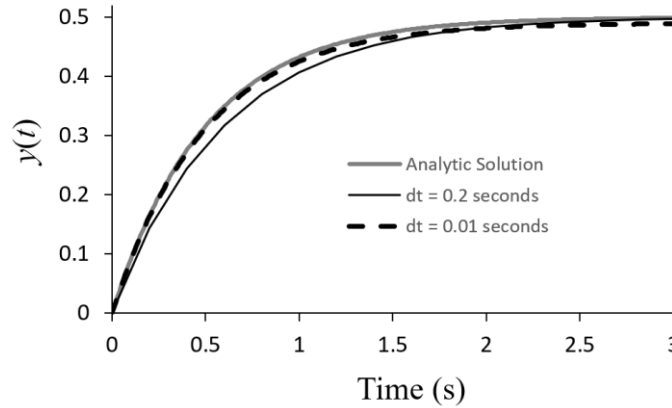


Figure 26-1: Theoretical solution to the first order differential equation and numerical solutions with time steps of 0.2 and 0.01 seconds. As the time step decreases, the solution becomes closer to the analytical solution.

Higher order derivatives

Higher order derivatives can be calculated in a similar manner. For example, the second order derivative in continuous time is

$$\frac{d^2y}{dt^2} = \lim_{\Delta t \rightarrow 0} \frac{\left(\frac{dy(t)}{dt} - \frac{dy(t - \Delta t)}{dt} \right)}{\Delta t}$$

The first order derivatives can be replaced with Equation 26-2 and Δt is kept as the sample time (i.e., the limit is not taken).

$$\begin{aligned} \lim_{\Delta t \rightarrow 0} \frac{\left(\frac{dy(t)}{dt} - \frac{dy(t - \Delta t)}{dt} \right)}{\Delta t} &\approx \frac{1}{\Delta t} \left(\underbrace{\frac{y[n] - y[n-1]}{\Delta t}}_{\sim dy(t)/dt} - \underbrace{\frac{y[n-1] - y[n-2]}{\Delta t}}_{\sim dy(t-dt)/dt} \right) \\ &= \frac{y[n] - 2y[n-1] + y[n-2]}{(\Delta t)^2} \text{ discrete} \end{aligned}$$

Similarly,

$$\begin{aligned} \frac{d^3y}{dt^3} &\approx \frac{1}{\Delta t} \left(\underbrace{\frac{y[n] - 2y[n-1] + y[n-2]}{(\Delta t)^2}}_{\text{Second derivative at step } n} - \underbrace{\frac{y[n-1] - 2y[n-2] + y[n-3]}{(\Delta t)^2}}_{\text{Second derivative at step } n-1} \right) \\ &= \frac{y[n] - 3y[n-1] + 3y[n-2] - y[n-3]}{(\Delta t)^3}. \end{aligned}$$

Example: Find a discrete representation of the ordinary differential equation

$$y \frac{dy}{dt} = x.$$

(While the equation is nonlinear and does not therefore lead to a linear discrete version, the methods from the above derivatives still apply.)

Answer: Replace y with $y[n]$, x with $x[n]$, and dy/dt with $(y[n] - y[n-10])/\Delta t$.

$$y[n] \frac{y[n] - y[n-1]}{\Delta t} = x[n]$$

$$y[n]^2 - y[n]y[n-1] = x[n]\Delta t$$

The equation is quadratic in $y[n]$. The solution as an explicit function of $y[n]$ is

$$y[n] = \frac{1}{2}y[n-1] \pm \frac{1}{2}\sqrt{y[n-1]^2 - 4x[n]\Delta t}$$

If Δt is sufficiently small, then $y[n]$ can be predicted at each time step from known previous values of y and the current value of x .

26.2.2. Moving Average

A simple example of a discrete time system is the weighted moving-average system

$$y[n] = \frac{6}{11}x[n] + \frac{3}{11}x[n-1] + \frac{2}{11}x[n-2].$$

Here, $y[n]$ is the weighted average of the three most recent values of x . The weights are frequently chosen such that the most recent value carries the most weight (i.e. $\frac{6}{11}x[n]$), and others get successively smaller ($\frac{3}{11}x[n-1]$, $\frac{2}{11}x[n-2]$). With positive weights, this system is a discrete low-pass filter, analogous to the continuous time filter

$$y(t) = \frac{1}{T} \int_t^{t+t_0} \mathcal{C}(\tau)x(t-\tau) d\tau.$$

Despite the simplicity of this filter and the ease with which it can be implemented digitally, it is not easily implemented with our operational amplifier methods. Although the equation is similar to that given for linear systems in **Equation 2-2**, the upper integral limit is not ∞ .

The weighted moving average system can be expressed as a system diagram (Figure 23-5).

The delay blocks, symbolized by z^{-1} ,* delay their inputs by 1 time step, i.e., $x[n]$ is shifted by 1 time step to $x[n-1]$, and $x[n-1]$ is shifted by one time step to $x[n-2]$. The number labels on the arrows are the multiplication factors for $x[n]$ and its lags. I.e., $x[n]$, $x[n-1]$, and $x[n-2]$ are multiplied by 2, 0.7 and 0.3, respectively. The sum block, symbolized by Σ , sums all of its inputs to generate its output.

* The reason for this notation will be clear after the section on z transforms. Note for now that z^{-2} is a time delay of 2 time steps and in general z^{-k} is a time delay of k time steps.

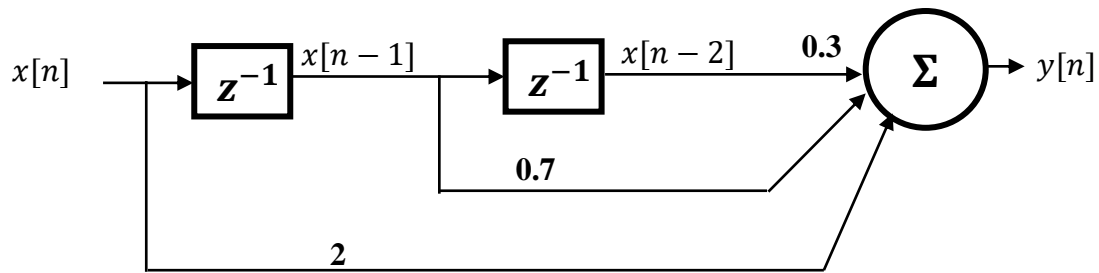


Figure 26-2: System diagram for a moving average filter $y[n] = 2x[n] + 0.7x[n - 1] + 0.3x[n - 2]$.

The general formula for the moving average filter is

$$y[n] = \sum_{k=0}^{N-1} C[k]x[n - k], \quad \text{Equation 26-4}$$

Where N is the length of the filter and $C[k]$ is the set of weighting coefficients. A general diagram for the filter is shown in Figure 26-3, where the dashed lines represent the additional delays and coefficients required to reach $N - 1$ elements.

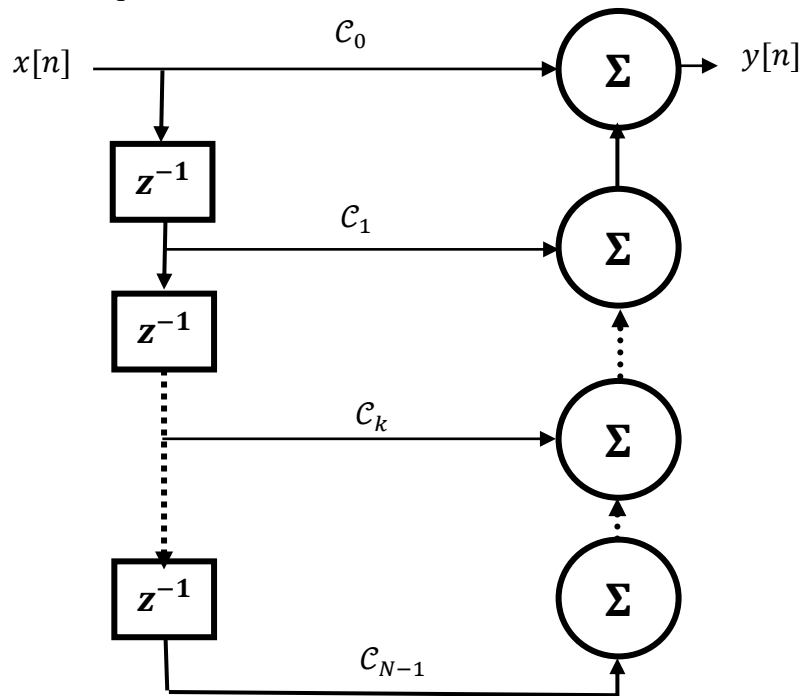


Figure 26-3: General diagram for the MA filter.

Moving average filters are finite impulse response (FIR) filters, meaning that when the input $x[n]$ is an impulse ($\delta[n]$) the output $y[n]$ will become and remain zero after $n > N$.

26.2.3. Autoregressive

Autoregressive filters are IIR filters of the form

$$\sum_{k=0}^M c_k y[n-k] = x[n]. \quad \text{Equation 26-5}$$

In this form, the output is lumped together with lags of the output on the left-hand side. M is the number of lags of y that are used. The equation can be rewritten to isolate $y[n]$ on the left-hand side.

$$y[n] = \frac{1}{c_0} x[n] - \frac{1}{c_0} \sum_{k=1}^M c_k y[n-k].$$

This form shows that the output at step n is a linear combination of previous outputs and the current input ($x[n]$), but does not depend directly on previous values of the input. The general diagram for an autoregressive filter is shown in Figure 26-4, where $a_0 \equiv 1/c_0$ and $a_k \equiv c_k/a_0$.

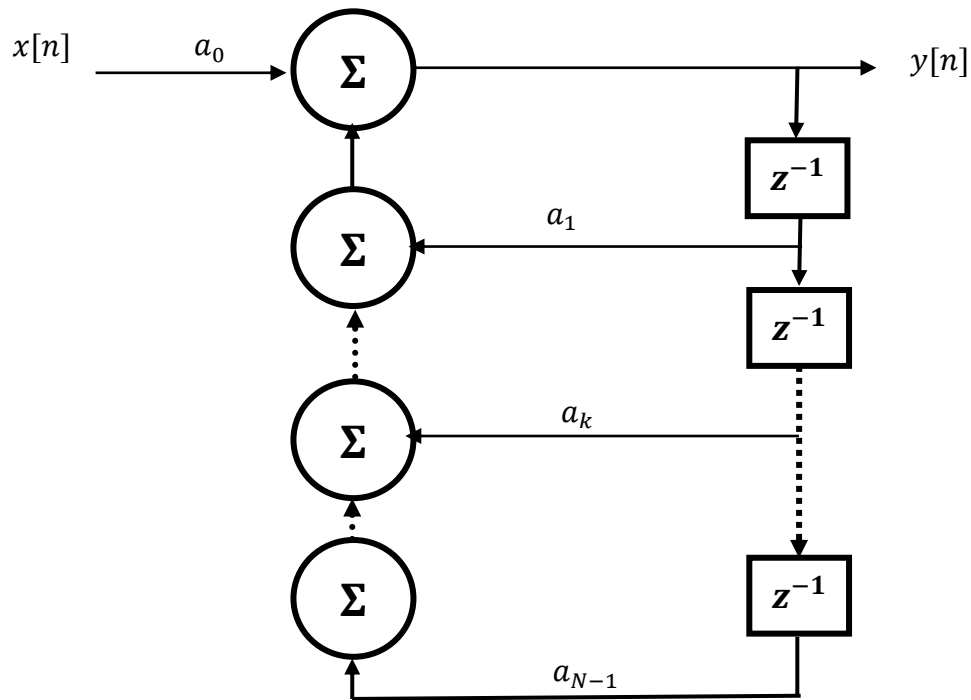


Figure 26-4: General diagram for the autoregressive filter.

Autoregressive filters are infinite impulse response filters. If an impulse is presented as the input, $x[n] = \delta[n]$, the filter will continue to have a non-zero output. The output can temporarily be zero, but it will not remain zero. The output can also decrease toward zero.

Example: Show that the autoregressive filter $y[n] = x[n] - y[n-1] - y[n-2]$ with an impulse input reaches zero but only temporarily and is therefore an infinite impulse response filter.

Answer: With $x[n] = \delta(n)$

$$\begin{array}{rclcl}
 y[n] & = & x[n] & -y[n-1] & -y[n-2] \\
 y[0] & = & 1 & -0 & -0 & = 1 \\
 y[1] & = & 0 & -1 & -0 & = -1 \\
 y[2] & = & 0 & -(-1) & -1 & = 0 \\
 y[3] & = & 0 & -0 & -(-1) & = 1 \\
 y[4] & = & 0 & -1 & -1 & = -1
 \end{array}$$

One can extrapolate these results to see that the output will continue the pattern 0, 1, -1 forever.

26.2.4. ARMA

The ARMA filter combines the autoregressive and MA filters.

$$\sum_{i=0}^N a_i y[n-i] = \sum_{i=0}^M b_i x[n-i]. \quad \text{Equation 26-6}$$

In standard ARMA form, the output at step n is a linear combination of the current and previous inputs ($x[n-i]$) and previous outputs ($y[n-i]$). The left-hand side is the dependency on the outputs, and the right-hand side is the dependency on the input. Because both sides of the equation are equal, they can be set to an intermediate variable $w[n]$.

$$\sum_{i=0}^N a_i y[n-i] = w[n]; \quad w[n] = \sum_{i=0}^M b_i x[n-i]. \quad \text{Equation 26-7}$$

Comparison of the left equation of Equation 26-7 to the autoregression equation shows that $w[n]$ is the input of an autoregressive filter with output $y[n]$. Comparison of the right equation of Equation 26-7 shows that $w[n]$ is the output of a MA filter with input $x[n]$. Thus, the overall filter can be considered to be an autoregressive filter cascaded with a MA filter, as shown in Figure 26-5.

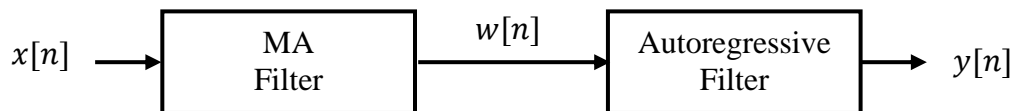


Figure 26-5: ARMA filter as an autoregressive filter cascaded with a MA filter.

Because these filters are linear, their order is irrelevant, so the MA filter position can be switched with the autoregressive filter position, as in Figure 26-6. The intermediate variable is no longer $w[n]$, but that value has no specific relevance.

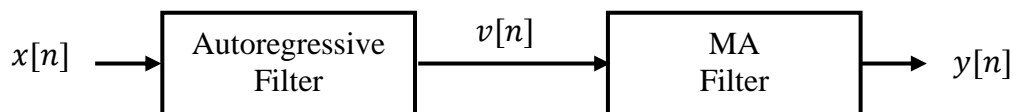


Figure 26-6: ARMA filter with the order of the autoregressive and MA filters switched.

The diagrams for the MA and autoregressive filters can therefore be combined as in Figure 26-7.

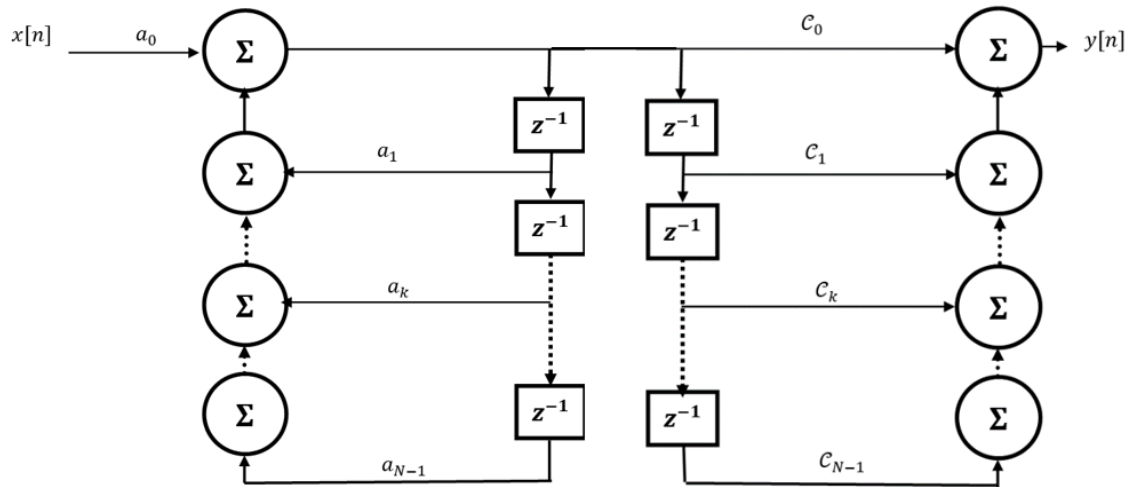


Figure 26-7: Combined MA and autoregressive filter diagrams.

The diagram can be simplified because the lags of $v[n]$ derived from the autoregressive filter are the same as those derived from the MA filter. Specifically, the input to c_1 is identical to the input to a_1 , the input to c_2 is identical to the input to a_2 , and, in general, the input to c_k is identical to the input to a_k . The parallel z^{-1} blocks can therefore be combined as in

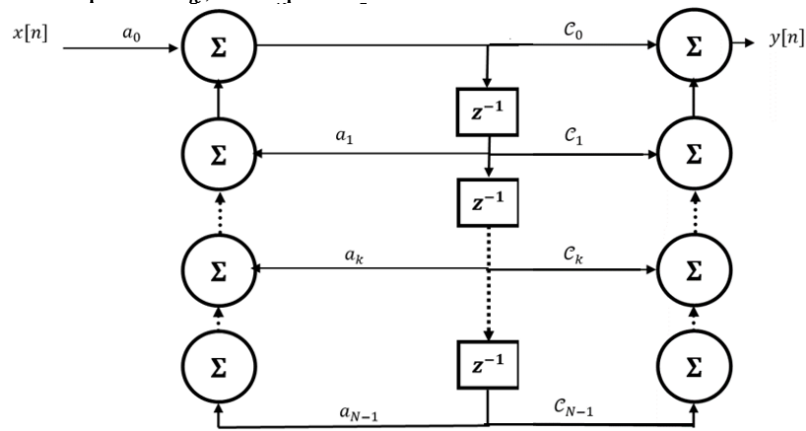


Figure 26-8.

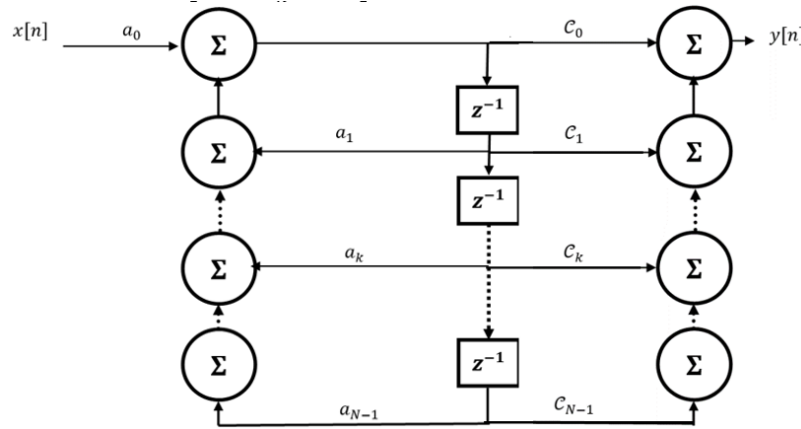


Figure 26-8: Simplified ARMA filter diagram.

26.3 Properties of Linear Time-Invariant Systems

26.3.1. Linearity

Linearity has the same meaning for discrete systems as for analog systems. Specifically, if a system is linear and two inputs x_1 and x_2 lead to outputs y_1 and y_2 , respectively, and if a_1 and a_2 are constants, then the output to $a_1x_1 + a_2x_2$ is $a_1y_1 + a_2y_2$, as illustrated in .

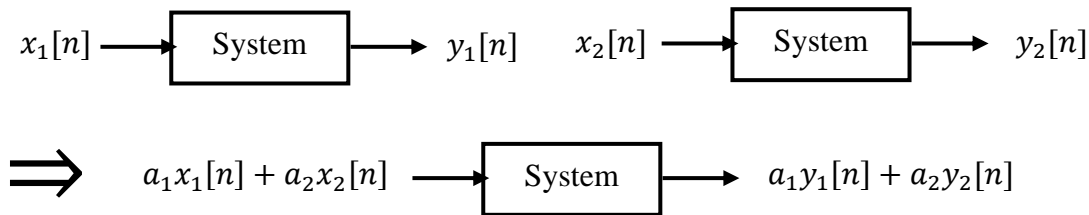


Figure 26-9: Linear system. If x_1 leads to y_1 and x_2 leads to y_2 then $a_1x_1 + a_2x_2$ leads to $a_1y_1 + a_2y_2$ for constants a_1 and a_2 .

26.3.2. Memory

A discrete time system has memory if its output depends on a value of the current input $x[n]$ and any previous values of its input or output (ie. $x[n-1]$, $y[n-1]$). Systems with memory are referred to as dynamic systems. If a system does not have memory, it is referred to as a static system.

For example, a system that squares its input is a static system. An autoregressive system is dynamic. A moving average, other than $y[n] = a_0x[n]$ is dynamic.

Example: The continuous time equation that for an RC circuit

$$RC \frac{dy}{dt} + y = x$$

describes a system with memory. Does the discrete version describe a dynamic system?

Answer: The equivalent discrete equation is

$$RC \frac{y[n] - y[n - 1]}{\Delta t} + \alpha y[n] = x[n].$$

Because the system depends on the $n - 1$ lag of y , it has memory and is therefore dynamic. In the case of the RC circuit, the physical source of the memory is the charge on the capacitor.

26.4 Examples of Nonlinear and Time-Variant Systems

Chapter 27: Filters

27.1 Applications

27.2 Ideal vs. Real Filters

27.3 Implementation

27.3.1. By Transfer Function

27.3.2. By Convolution

27.4 Advantages of Digital vs. Analog Filters

27.5 FIR Filters

27.6 IIR Filters

27.7 Impulse Response

27.8 Transient vs. Steady State Response

27.9 Relationship between Impulse Response and Transfer Function

27.9.1. The Convolution Theorem

27.9.2. Time Response for a Brick Wall Filter

27.9.3. Effect of an Offset on High Pass Filter Output

27.10 Filter Characteristics

27.10.1. Pass Band

27.10.2. Stop Band

27.10.3. Transition Band (Rolloff)

27.10.4. Ripple

27.11 Decibel Scale

27.11.1. For Voltage

27.11.2. For Power

27.11.3. Why Voltage and Power are Different

27.12 Windowing

27.12.1. Why Windowing is Used

27.12.2. Common Windows and Their Advantages/Disadvantages

27.13 Filters Derived from Other Filters

27.13.1. Low Pass to High Pass

Spectral Inversion

Spectral Reversal

27.13.2. Band Pass from Cascaded Low Pass and High Pass Filters

27.13.3. Notch from Parallel Low Pass and High Pass Filters

Chapter 28: Z Transforms

28.1 Differential Equations and Transfer Functions

28.1.1. Relation to Continuous Systems

The z-transform is useful in determining the transfer function of a discrete signal or system. It is derived from the Fourier transform for continuous systems. First, recall the transform process for a continuous system. Let the system be described by

$$\frac{dy(t)}{dt} + \alpha y(t) = x(t).$$

The Fourier transform of both sides was taken.

$$\mathcal{F}\left\{\frac{dy(t)}{dt} + \alpha y(t)\right\} = \mathcal{F}\{x(t)\}$$

$$j\omega Y(j\omega) + \alpha Y(j\omega) = X(j\omega)$$

$$\frac{Y(j\omega)}{X(j\omega)} = \frac{1}{j\omega + \alpha}; \quad \left|\frac{Y(j\omega)}{X(j\omega)}\right| = \frac{1}{\sqrt{\omega^2 + \alpha^2}}, \quad \angle \frac{Y(j\omega)}{X(j\omega)} = -\text{atan}\left(\frac{\omega}{\alpha}\right)$$

The system was then reduced to a block diagram, where the Fourier transform of the output was equal to the Fourier transform of the input, multiplied by the transfer function $H(j\omega) \equiv |Y(j\omega)/X(j\omega)|$.

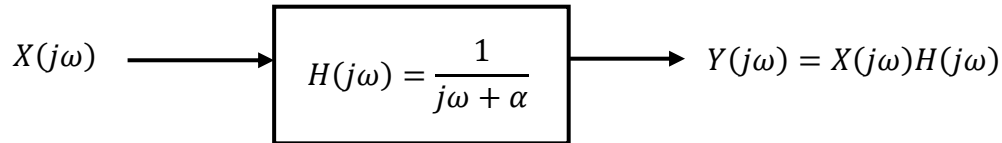


Figure 28-1: Continuous system reduced to a block diagram.

28.1.2. Discrete Systems

For a discretized system, the derivatives become differences, and the continuous differential equation equivalent is

$$y[t] + a_1 y[t - \Delta t] + \dots + a_k y[t - k\Delta t] = b_0 x[t] + b_1 x[t - \Delta t] + \dots + b_m x[t - m\Delta t].$$

We can again take the Fourier transform. From the shifting theorem, $\mathcal{F}\{f(t - t_0)\} = e^{-i\omega t_0} \mathcal{F}\{f(t)\}$. With $Y \equiv \mathcal{F}\{y[t]\}$ and $X \equiv \mathcal{F}\{x[t]\}$,

$$Y + a_1 e^{-i\omega \Delta t} Y + \dots + a_k e^{-ik\omega \Delta t} Y = b_0 X + b_1 e^{-i\omega \Delta t} X + \dots + b_m e^{-im\omega \Delta t} X$$

Factor out Y on the left and X on the right.

$$(1 + a_1 e^{-i\omega\Delta t} + \dots + a_k e^{-ik\omega\Delta t})Y = (b_0 + b_1 e^{-i\omega\Delta t} + \dots + b_m e^{-im\omega\Delta t})X$$

Thus,

$$H(j\omega) \equiv \frac{Y(j\omega)}{X(j\omega)} = \frac{b_0 + b_1 e^{-i\omega\Delta t} + \dots + b_m e^{-im\omega\Delta t}}{1 + a_1 e^{-i\omega\Delta t} + \dots + a_k e^{-ik\omega\Delta t}}.$$

As with continuous systems, $H(j\omega)$ as a magnitude and phase and can be plotted as a Bode plot.

As a simple example, the difference filter is the equivalent of the continuous derivative, as deduced from the process described in Section 26.2.1. If all a and b coefficients are zero except for $b_0 = \frac{1}{2}$ and $b_1 = -\frac{1}{2}$

$$y[t] = \frac{1}{2}x[t] - \frac{1}{2}x[t - \Delta t]$$

The transfer function is

$$H(j\omega) = \frac{Y(j\omega)}{X(j\omega)} = \frac{1}{2}(1 - e^{-j\omega\Delta t}).$$

Given, from Euler's rule, that $e^{-j\omega t} = \cos(\omega t) - j \sin(\omega t)$, the real part is $\frac{1}{2}(1 - \cos(\omega t))$, and the imaginary part is $\frac{1}{2}\sin(\omega t)$. The magnitude is then

$$\left| \frac{Y(j\omega)}{X(j\omega)} \right| = \frac{1}{2} \sqrt{(1 - \cos(\omega\Delta t))^2 + \sin^2(\omega\Delta t)},$$

Which expands to

$$= \frac{1}{2} \sqrt{1 - 2\cos(\omega\Delta t) + \underbrace{\cos^2(\omega\Delta t) + \sin^2(\omega\Delta t)}_{=1}} = \frac{1}{\sqrt{2}} \sqrt{1 - \cos(\omega\Delta t)},$$

Where the right-hand expression is a result of the well-know trigonometric expression $\cos^2(x) + \sin^2 x = 1$. The phase is

$$\angle \left(\frac{Y(j\omega)}{X(j\omega)} \right) = \text{atan} \left(\frac{\sin(\omega\Delta t)}{1 - \cos(\omega\Delta t)} \right).$$

Recall that the derivative filter, $y = dx/dt$ has the transfer function $H(j\omega) = \omega$. The magnitude and phase of the derivative and difference filters are compared in Figure 28-2. The magnitudes are close to a scaled frequency of 1 radian/s. As the signal frequency approaches the sample rate of the signal, the difference filter transfer function magnitude deviates from the difference filter. The phases for the two signals are equal at $\omega = 0$, but the derivative filter's phase remains at 90° while the difference filter's phase decreases to 0 at the Nyquist frequency. (Here we take $\Delta t = 1$, so the frequency scale ranges from 0 to a scaled Nyquist frequency of $\omega = \pi$).

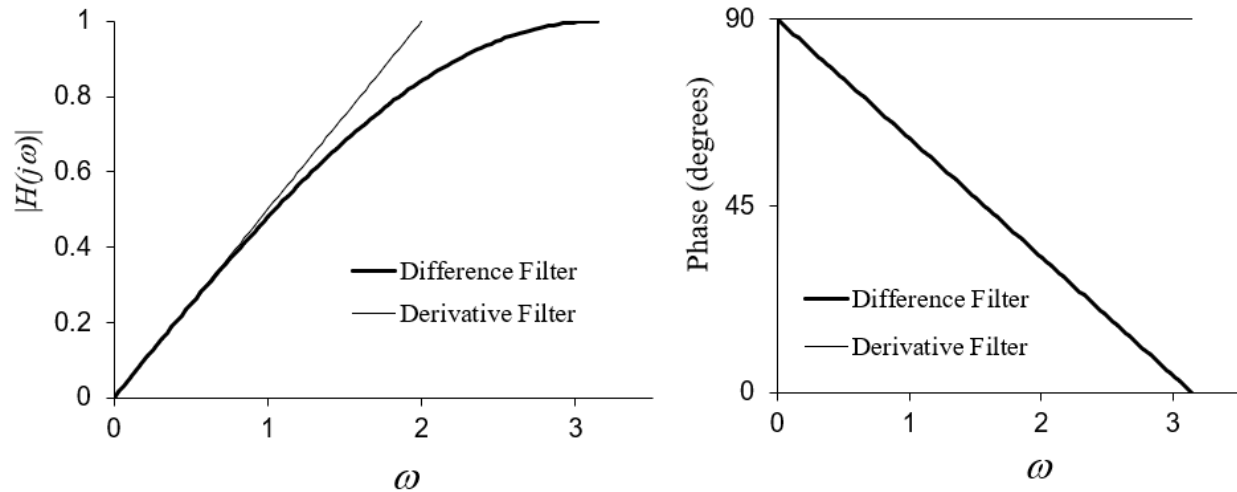


Figure 28-2: Magnitude and phase of the difference and derivative filters.

28.1.3. Comparison of the Transfer Function to Fourier Series

In the following sense, the process of transforming a discrete time series is the opposite of Fourier series. The Fourier series takes a continuous periodic signal of period T , and transforms it to a set of discrete frequencies, separated by $\Delta f = 1/T$. The above process takes a time signal evaluated at discrete times separated by Δt and transforms it into a continuous, periodic spectrum with period $f = 1/\Delta t$.

28.2 Conversion to a Z Transform

The general transfer function

$$\frac{Y(j\omega)}{X(j\omega)} = \frac{(b_0 + b_1 e^{-i\omega\Delta t} + \dots + b_m e^{-im\omega\Delta t})}{(1 + a_1 e^{-i\omega\Delta t} + \dots + a_k e^{-ik\omega\Delta t})}$$

can be written more compactly if we use the shorthand $z \equiv e^{i\omega\Delta t}$. Then

$$\frac{Y(j\omega)}{X(j\omega)} = \frac{(b_0 + b_1 z^{-1} + \dots + b_m z^{-m})}{(1 + a_1 z^{-1} + \dots + a_k z^{-k})}$$

With this notation change, the transform is referred to as the z transform.

28.3 Definition of the Z Transform

The z-transform in discrete time is analogous to the Laplace (or Fourier) transform in continuous time. The Laplace transform (continuous time) is

$$\mathcal{L}[x(t)] = X(s) = \int_0^{\infty} x(t) e^{-st} dt$$

The Fourier transform (continuous time) is

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

The z-transform (discrete time) is

$$\mathcal{Z}[x[n]] = X(z) = \sum_{n=0}^{\infty} x[n] e^{-i\omega n \Delta t} = \sum_{n=0}^{\infty} x[n] z^{-n}$$

For discrete time, the integral becomes a sum because time is defined at only discrete points. The connection of the z transform and the Fourier transform is derived as follows.

For a discrete system, $t = n\Delta t$, where $\Delta t = 1/f_s$, f_s being the sample frequency. Furthermore, $\omega = 2\pi m \Delta f$, where $\Delta f = f_s/N$, N being the number of data points.

$$\mathcal{F}\{x(t)\}_{\text{sampled}} = \int_{-\infty}^{\infty} x[n/f_s] e^{-j(\frac{2\pi m f_s}{N})(\frac{n}{f_s})} dt$$

The integral is non-zero only for integral values of n , so it can be replaced by a sum over n (from 0 to N). Also, f_s cancels in the exponent.

$$\mathcal{F}[m]_{\text{discrete}} = \sum_{n=0}^N x[n] e^{-j(\frac{2\pi mn}{N})}$$

28.4 Derivation of Z Transforms

28.4.1. Impulse

28.4.2. Step

28.4.3. Finite Duration Signal

28.4.4. Rational Transfer Function Form

28.4.5. Causal Exponential

28.4.6. Other Common Z Transforms

28.5 Properties of Z Transforms28.5.1. Linearity28.5.2. Delay (Lag)28.5.3. Multiplication by a^n 28.5.4. Multiplication by n 28.5.5. Time Scaling28.5.6. Time Reversal28.5.7. Summation28.5.8. Convolution28.5.9. Initial Value28.5.10. Final Value**28.6 Examples of Z Transform Properties****28.7 BIBO Stability**28.7.1. Partial Fraction Expansion28.7.2. Requirements on Poles*For Laplace Transforms**For Z Transforms***28.8 System Behaviors Based on Pole Placement**28.8.1. $|p| < 1$ with $\Re\{p\} > 0$ Exponential Decay28.8.2. Complex poles with $|p| < 1$ Underdamped28.8.3. $|p| < 1$ with $\Re\{p\} < 0$ Oscillating Exponential Decay*Relationship to the Underdamped Case**Illustration of the Origin of the Oscillation*28.8.4. $|p| > 1$, Positive Pole ($\Re\{p\} > 0$) Unstable28.8.5. $|p| > 1$, Negative Pole ($\Re\{p\} < 0$) Unstable, Oscillatory28.8.6. $|p| < 1$, $\Re\{p\} < 0$ Stable, Underdamped, "Oscillatory"28.8.7. $|p| > 1$, $\Re\{p\} < 0$ Unstable, Underdamped, "Oscillatory"**28.9 Example of Matlab Code to Plot the Transfer Function**

Chapter 29: Digital Filter Design

29.1 Relationship between Ω & ω and between T_s & n

In the translation from a continuous to a discrete signal, some of the dimensionality becomes disassociated from the basic parameters. For example, with a discrete signal, the frequency ω_0 has the units of radians/s, whereas the discrete signal uses the parameter Ω , with units of radians. It is important to be able to translate among these parameters.

A continuous time sinusoid of angular frequency ω_0 , is given by

$$x(t) = A \cos(\omega_0 t + \varphi),$$

with period $T_0 = 2\pi/\omega_0$ (seconds) and frequency $f_0 = 1/T_0$ (Hz).

The DT sinusoid sampled every T_s seconds is

$$x[n] = \underbrace{x(nT_s)}_{\substack{\text{Substitute} \\ t=nT_s}} = A \cos(\omega_0 nT_s + \varphi) = \underbrace{A \cos(\Omega_0 n + \theta)}_{\substack{\text{Substitute} \\ \omega_0 T_s = \Omega_0}}$$

The parameter $\Omega_0 \equiv \omega_0 T_s$ is the discrete time angular frequency, and it has the dimensions of radians, as it is the product of radians/second and seconds. Recall that $-\omega_N \leq \omega_0 < \omega_N$, so that $-\omega_N T_s \leq \omega_0 T_s < \omega_N T_s$, where ω_N is the Nyquist frequency.

$$\omega_N = 2\pi \left(\frac{f_s}{2} \right) = \frac{2\pi}{\underbrace{2T_s}_{f_s=1/T_s}} = \frac{\pi}{T_s}.$$

Thus, $\omega_N T_s = \pi$, so $-\pi \leq \omega_0 T_s \leq \pi$, or $-\pi < \Omega_0 < \pi$. To convert between Ω_0 and ω_0 , one must simply multiply by T_s .

29.2 Why we write $H(e^{j\Omega n})$ instead of $H(\Omega n)$

For a linear system in continuous time, the input $A \cos(\omega_0 t)$ leads to the output $A |H(j\omega_0)| \cos(\omega_0 t + \varphi)$, where $\varphi = \angle H(j\omega_0)$. Similarly, for a discrete system with $\Omega_0 = \omega_0 T_s$, the input $A \cos(\Omega_0 n)$ leads to the output $A |H(e^{j\Omega_0 n})| \cos(\Omega_0 n + \varphi)$. Generally, we write $H(e^{j\Omega n})$ instead of $H(\Omega n)$ simply because it conforms to our idea of a z transform, where $z = e^{j\Omega}$.

29.3 The z-Plane

The variable z is complex so it has a real and imaginary part. If we consider all possible values of z , not merely those for which $|z| = 1$, we can conceptualize a z-plane, in analogy to the s-plane for Laplace transforms. Values of $z = e^{j\Omega}$ lie on the unit circle within this plane, as illustrated in Figure 29-1.

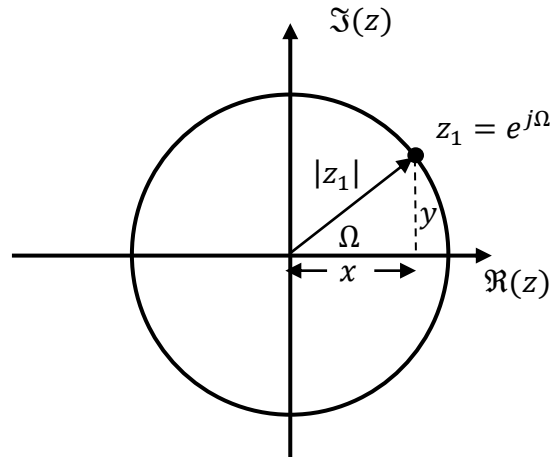


Figure 29-1: The z-plane

The real part of z_1 is the length marked x on the figure, and the imaginary part of z_1 is y , which is consistent with $z_1 = e^{j\Omega}$ because $e^{j\Omega} = \cos(\Omega) + j \sin(\Omega)$; $x = \cos(\Omega)$ is the adjacent side of the right triangle shown, and $y = \sin(\Omega)$ is the opposite side. The variable Ω is the angle for this triangle, and Ω must range between 0 and π radians. Since $\Omega = \omega T_s$, it increases to π as the signal frequency increases to the Nyquist frequency.

29.4 Digital Filter Characteristics

29.4.1. Frequency Response

As with continuous time systems, the output ($y[n]$) of a discrete time system with sinusoidal input ($x[n]$) depends on the frequency response of the system (filter).

29.4.2. Stability

A linear time invariant system can be created that is not stable. Here we consider bounded-input bounded-output (BIBO) stability, which means that any bounded input will lead to a bounded output. A simple example is the filter

$$y[n] = x[n] + 1.1y[n - 1].$$

Use as the bounded input the delta function $x[n] = \delta[n]$. For $n = 0$, $y[0] = 1$. For $n = 1$, $y[1] = 1.1(1)$. For $n = 2$, $y[2] = 1.1(1.1)$. For $n = 3$, $y[3] = 1.1(1.21)$. Thus, with each successive increment in n , the output grows by a factor of 1.1, which is an exponential increase, as shown in Figure 29-2.

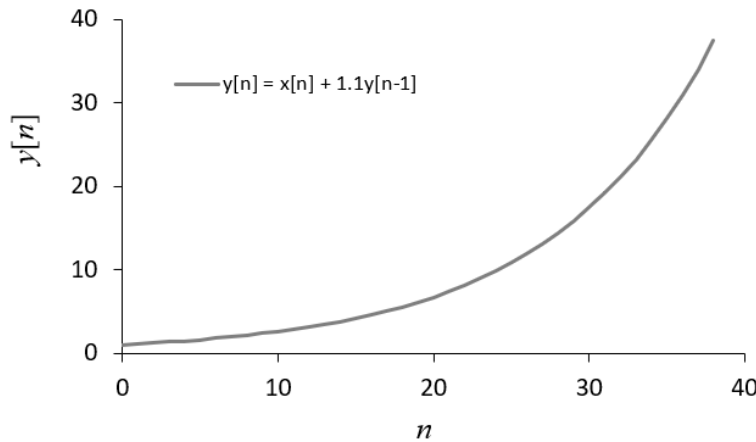


Figure 29-2: Exponential growth of an unstable system.

Interpret this behavior as a physical system. It would be equivalent to giving a patient an injection of a drug and having the concentration of the drug in the patient's blood increase without bound. Regardless of what the drug does, the patient would ultimately die. In contrast, for a stable system, the amount of the drug would decrease exponentially to zero, for example, by excretion in the kidney.

The transfer function of this filter is obtained as

$$\begin{aligned}
 Y &= X + 1.1z^{-1}Y \\
 (1 - 1.1z^{-1})Y &= X \\
 \frac{Y}{X} = H(z) &= \frac{1}{1 - 1.1z^{-1}} = \frac{z}{z - 1.1}.
 \end{aligned}$$

The value $p \equiv 1.1$ is the pole of this system (because $z - 1.1$ causes $H(z)$ to become infinite). For a system that has poles,* stability of the filter depends on the location of poles; they must be within the unit circle for the system to be BIBO stable, meaning that $|p| < 1$. This condition is analogous to stability for a continuous time system, where the criterion is that $\Re\{p\} < 0$. These two conditions are shown in Figure 29-3.

* An ideal differentiator is an example of a system that has no poles and yet is BIBO unstable; the output to a step function input goes to infinity because the derivative of the step function is infinite at time zero.

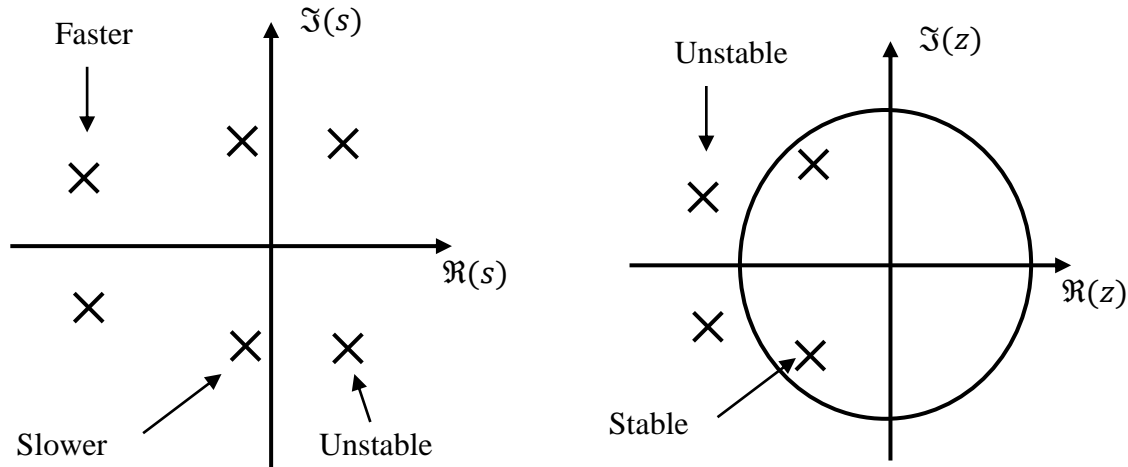


Figure 29-3: (Left) For a continuous system, stability requires $\Re\{p\} < 0$, so that all poles p must be to the left of the imaginary axis in the s plane. (Right) For a discrete time system, stability requires that $|p| < 1$, so that all poles must be inside the unit circle $|z| = 1$.

29.4.3. Discrete Frequency to Physical Frequency (Ω to ω)

To convert from Ω to ω , recognize that $\omega = 2\pi f_N$ corresponds to $\Omega = \pi$. Because $f_N = f_s/2$,

$$\frac{\omega}{2\pi f_N} = \frac{\omega}{2\pi(f_s/2)} = \frac{\omega}{\pi f_s} = \frac{\Omega}{\pi} \Rightarrow \omega = \Omega f_s.$$

Therefore, one need only multiply Ω by the sample frequency.

Example: A signal is discretized at 2 kHz. What physical frequency corresponds to $\Omega = \pi/4$?

Answer: $\omega = \frac{\pi}{4}(2000) = 500\pi$ radians/s, which corresponds to 250 Hz.

29.5 Geometric Relationship between Pole Location and Transfer Function

Consider a transfer function $H(z)$ for a system with a single real pole. The transfer function is

$$H(z) = \frac{1}{z - p}, \quad \text{Equation 29-1}$$

and the diagram in the z -plane is shown in Figure 29-4. One can show that as z moves closer to the pole, the magnitude of the transfer function must increase. Specifically in this case, as Ω approaches π , the magnitude of the transfer function must increase.

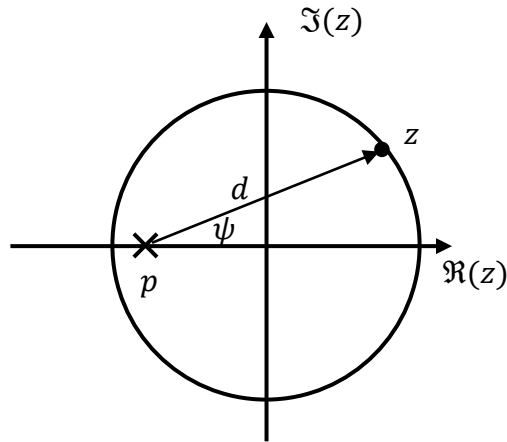


Figure 29-4: Diagram of a first order (single pole) system in the z-plane.

The denominator of Equation 29-1 is the difference between z and p . That difference is a complex number that can be written as $|z - p|e^{i\psi}$. However, $|z - p|$ is d as marked in the figure. The transfer function is therefore

$$H(z) = \frac{1}{de^{i\psi}},$$

so if d becomes shorter, the transfer function magnitude becomes larger. In this case, as Ω approaches π , d becomes as short as possible, so the transfer function will demonstrate a peak at the Nyquist frequency.

The opposite effect occurs for a zero. As z approaches a zero, because the zero is in the denominator, the transfer function magnitude becomes smaller. If the zero is placed on the unit circle, the transfer function magnitude will become zero at the corresponding frequency. Overall, for a transfer function with multiple poles and multiple zeros, the result will be a balance between the changes in distances to the zeros and the changes in distances to the poles.

29.6 Filter Designs Based on Pole Placement

29.6.1. Low Pass

29.6.2. High Pass

29.6.3. Band Pass

29.6.4. Notch

29.6.5. Band Reject

29.7 MATLAB Filter Design and Analysis Tool

29.7.1. Pole-Zero Placement Examples

29.7.2. Filter Coefficients

29.7.3. Other Icons

29.7.4. Pre-Whitening Filter

First Order

Fourth Order

29.8 Direct Transfer Function Calculation from Poles and Zeros

If a transfer function is known in the z-domain, frequency response can be easily calculated in MATLAB where MATLAB's ability to work with complex numbers is employed. Here, the frequency response of three 60 Hz notch filters will be compared. The first has zeros on the unit circle with no poles, the second has zeros on the unit circle and poles with magnitude 0.8, and the third has the same zeros on the unit circle and poles with magnitude 0.95. The sample frequency will be assumed to be 720 Hz. The value of Ω must be such that $\Omega/\pi = 60/720$, so $\Omega = \pi/12$.

The Matlab code is

```
% Compare the frequency response of 60 Hz three notch filters. One with zeros
% on the unit circle and no poles, one with zeros on the
% unit circle and poles inside the unit circle (at the notch frequency),
% but with magnitude 0.8, and a third with the same two zeros and poles on
% the unit circle with magnitude 0.95. Let the sample frequency be fs = 720 Hz.

Omega = 0:.01:pi; % Frequency axis
fs = 720;          % Hz (sample frequency)
w = Omega*fs/(2*pi); % Physical frequency in radians/s
fN = 60;           % Hz (notch frequency)
wN = 2*pi*fN;      % Notch frequency in radians/s
OmegaNotch = wN/(fs); % Notch frequency as % of the Nyquist frequency.
zero1 = exp(i*OmegaNotch); zero2 = exp(-i*OmegaNotch); % Define the zeros
pole1 = 0.8*exp(i*OmegaNotch); pole2 = 0.8*exp(-i*OmegaNotch); % Define the first poles
pole3 = 0.95*exp(i*OmegaNotch); pole4 = 0.95*exp(-i*OmegaNotch); % Define the second poles
z = exp(i*Omega); % Convert omega to z
Tf1 = (z-zero1).*(z-zero2); % Calculate the first transfer function
plot(w,abs(Tf1));
Tf2 = Tf1./((z-pole1).*(z-pole2)); % Calculate the second transfer function
hold on;
plot(w,abs(Tf2));
Tf3 = Tf1./((z-pole3).*(z-pole4)); % Calculate the third transfer function
hold on;
plot(w,abs(Tf3)); xlabel('Frequency (Hz)'); ylabel('Magnitude');
legend('No Poles','Poles Magnitude 0.8','Poles Magnitude 0.95','location','northwest');
legend('boxoff');
```

The code initially defines the Ω space (0 to π), then defines the sample and notch frequencies and determines the Ω value for the notch based on these two parameters. It also defines an ω array that will be used later to plot the results as a function of ω instead of Ω . The zeros and poles are calculated from the Ω_{Notch} value. An array of z values is then generated, and these are used directly in the transfer function equation to calculate $H(z)$. Finally, the magnitudes of the transfer functions are plotted. The results are shown in Figure 29-5. With no poles, the notch is broad, and the higher frequency components are overly amplified. With poles of magnitude 0.8, the transfer function is relatively symmetric around the notch frequency, but still somewhat larger at higher frequencies. With poles of magnitude 0.95, the transfer function is again symmetric around the notch frequency, but the notch is narrower.

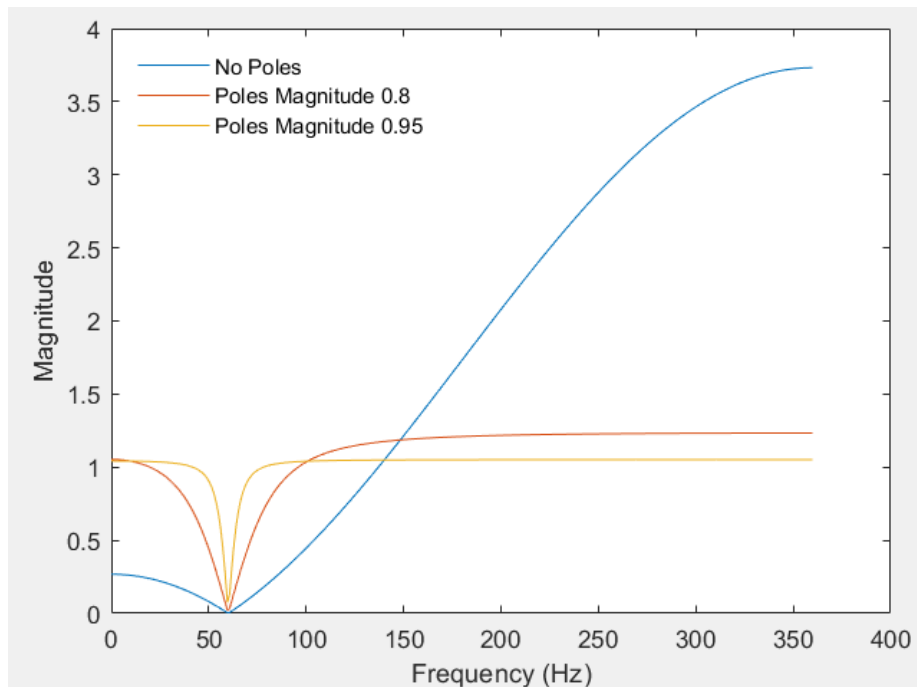


Figure 29-5: Calculated transfer functions for the three notch filters.

Chapter 30: Analysis of Biopotential Signals

30.1 Electromyogram

30.1.1. Origin of the Signal

Properties of Skeletal Muscle

Summation of Action Potentials

30.1.2. Relation of the EMG to Pathology

30.1.3. Electrode Placement

30.1.4. Time Domain

Amplitude

RMS vs. Time

Relationship between Force and Time

Intermittency

Noise Sources

30.1.5. Spectral Analysis

Typical EMG Spectrum

Frequency Content, Needed Sample Frequency

Total Power, Peak Power, Median, Mean Frequencies

Effects of Noise Sources

Energy or Power Spectrum?

Cumulative Power

30.1.6. Maximum Voluntary Contraction

30.1.7. Changes Caused by Fatigue

30.1.8. Useful Signal Measures

Full-wave rectified (FWR)

Mean Absolute Value (MAV)

Root-Mean-Squared (RMS)

Linear Envelope (LE)

Integrated EMG (iEMG)

Frequency/Power Spectrum (FFT)

Fatigue Analysis (Sequential FFT)

Power in the Band

30.2 Electroencephalogram

30.3 Electrooculogram

30.3.1. Origin of the Signal

30.3.2. Applications

30.3.3. Behavior of the Signal

Chapter 31: Relating Power Spectra to Physics (Music)

31.1 Introduction

31.1.1. How Music Relates to Instrumentation

31.1.2. Some Initial Questions

Why do we use 12 notes?

Why are 7 of these “in key?”

Why the “Circle of Fifths?” (Why not the “Circle of Thirds?”)

What’s all this about Major and Minor?

What do Augmented and Diminished Mean?

Why can a given chord have multiple names?

What is “color?”

31.2 Musical Frequencies

31.2.1. Musicians Terminology

31.2.2. In Hz

12 Tones

Each Tone is a Fraction Times the Previous One

$C^\# = xC$; $D = xC^\#$

If we go n steps, Note $n = x^n$ (Note 0)

Therefore $D = x^2C$ and $D_{\text{Octave 2}} = x^{12}D_{\text{Octave 1}} = 2D_{\text{Octave 1}}$

$\therefore x^{12} = 2 \Rightarrow x = 2^{\frac{1}{12}}$

In Particular, if $n = 7$, $2^{\frac{7}{12}} = 1.4983 \approx \frac{3}{2}$

This Simple Fraction Sounds Pleasing to the Ear

Frequencies Mapped to Notes

31.2.3. Circle of Fifths

Can Go Backwards

Follows the Key Signature Notes (F#, C#, G#, D#, A#, E#, B#=C, Fx=G, Db, Ab, Eb, Bb, Fb).

31.2.4. Piano Keyboard (Black vs. White Keys)

31.2.5. Energy Spectra

Major Chord

Minor Chord

All Notes in an Octave Played at Once

31.3 The Components of a Guitar

31.4 The Wave Equation

31.4.1. General Solution

31.4.2. Forward and Backward Waves

31.5 Boundary Conditions

31.5.1. Plucked String

31.5.2. Struck String

31.6 Solution

31.6.1. Functional Form

31.6.2. Allowed Frequencies

31.6.3. Illustration of String Shapes

31.6.4. Superposition (Linearity of the Fourier Transform)

31.7 Harmonics

31.8 Color

31.9 Beating of Adjacent Notes

31.10 Alternative Music Theories

31.10.1. Pythagorean

31.10.2. Just Intonation

31.10.3. Well-Tempered Scale

31.11 Insights from Fourier Transform Theorems

31.11.1. Finite Duration and Spectral Broadening

31.11.2. Effect of Envelope Shape

31.11.3. Dissonance

31.11.4. Rhythm (Need to think about this)

Fourier Spectrum of Rhythmic Clapping

31.12 Capabilities of Different Instruments

31.13 Interpretation

31.13.1. Speeding up & slowing down (rubato, firmata)

31.13.2. Increasing/decreasing volume (dynamics, piano, forte)

31.13.3. Modulating notes (vibrato)

31.13.4. Ornamentation (trills, slurs)

31.13.5. Changes in Color

Chapter 32: Project Formulation

32.1 Hypothesis

32.2 Human Subject Approval

32.3 Test Setup

32.4 Protocol

32.5 Data Analysis

Chapter 33: Final Laboratory on Hypothesis Testing

33.1 Introduction

You will propose a hypothesis about a signal of your choice (EKG, EMG, EEG, EOG, or other, subject to approval) perform experiments to test the hypothesis.

33.2 Process

This process will be carried out over the last two weeks of the course. Throughout those three weeks, you will be expected to meet regularly with your group members (nominally two times per week) and keep track of your meeting agenda and meeting minutes. These two items will be recorded on forms that are provided to you (in an Excel file).

1. Get together with your lab group.
2. Choose a physiological signal and propose a hypothesis about that signals behavior under some experimental condition (intervention). (The most likely signals are biopotentials, but you could consider other signals, such as respiration, force, temperature, as long as the signal is time-dependent).
3. Provide a proposal for an experiment to be conducted.
4. Submit (to Dr. Jones) a Human Use form for the proposal.
5. Build an Arduino-based signal acquisition system to record the signal, being sure to pre-process the signal in an optimal manner for the data acquisition.
6. Appropriately filter and condition the signals to extract the information of interest.
7. Apply hypothesis testing (statistics) on your acquired data to determine whether the effect of your intervention on the signal characteristic of interest is statistically significant.
8. Write up the results in a professional report.

33.3 Meetings

You will need to meet with your group at least three times to monitor progress. Each meeting must have an agenda, and the results of each meeting must be documented in minutes. A template has been provided for both the agenda and the minutes. A key part of the minutes is the designation of action items. Each action item must be assigned to a team member with a due date. Indicate, after the due date, whether the assigned task was completed on time. If it was not, provide a reason for the delay.

33.4 Report

You will submit a formal group report (one submission per group). The general instructions for the formal laboratory report are given in the laboratory manual.

- 1) Title Page: project name, course name and number, date of submission, group members' names.
- 2) A statement of the roles played by each of the group members. The statement should be more detailed than stating which part of the final report each member wrote.
- 3) Abstract

- 4) Introduction
- 5) Methods
- 6) Results
- 7) Discussion
- 8) Conclusion
- 9) References Cited (Nominally, 3 references are appropriate for this project. To count as a reference, the work must be a journal article and must be specifically cited within your narrative to back up a point).
- 10) Supplementary Material (Excel data sheets, Matlab codes, Arduino codes, Python codes).

References

- [1] J. S. Hausmann *et al.*, “Using Smartphone Crowdsourcing to Redefine Normal and Febrile Temperatures in Adults: Results from the Feverprints Study,” *J Gen Intern Med*, vol. 33, no. 12, pp. 2046–2047, Dec. 2018, doi: 10.1007/s11606-018-4610-8.
- [2] B. J. Hefflin, T. P. Gross, and T. J. Schroeder, “Estimates of medical device–associated adverse events from emergency departments,” *American Journal of Preventive Medicine*, vol. 27, no. 3, pp. 246–253, Oct. 2004, doi: 10.1016/j.amepre.2004.04.005.
- [3] N. A. Maffiuletti, A. J. Herrero, M. Jubeau, F. M. Impellizzeri, and M. Bizzini, “Differences in electrical stimulation thresholds between men and women,” *Ann Neurol*, vol. 63, no. 4, pp. 507–512, Apr. 2008, doi: 10.1002/ana.21346.
- [4] C. F. Dalziel, “Effects of Electric Shock on Man,” *IRE Trans. Med. Electron.*, vol. PGME-5, no. 0, pp. 44–62, Jul. 1956, doi: 10.1109/IRET-ME.1956.5008573.
- [5] J. G. Webster, *Bioinstrumentation*. John Wiley & Sons, 2004.
- [6] P. A. Fletcher, I. Marinelli, R. Bertram, L. S. Satin, and A. S. Sherman, “Pulsatile Basal Insulin Secretion Is Driven by Glycolytic Oscillations,” *Physiology*, vol. 37, no. 4, pp. 216–223, Jul. 2022, doi: 10.1152/physiol.00044.2021.