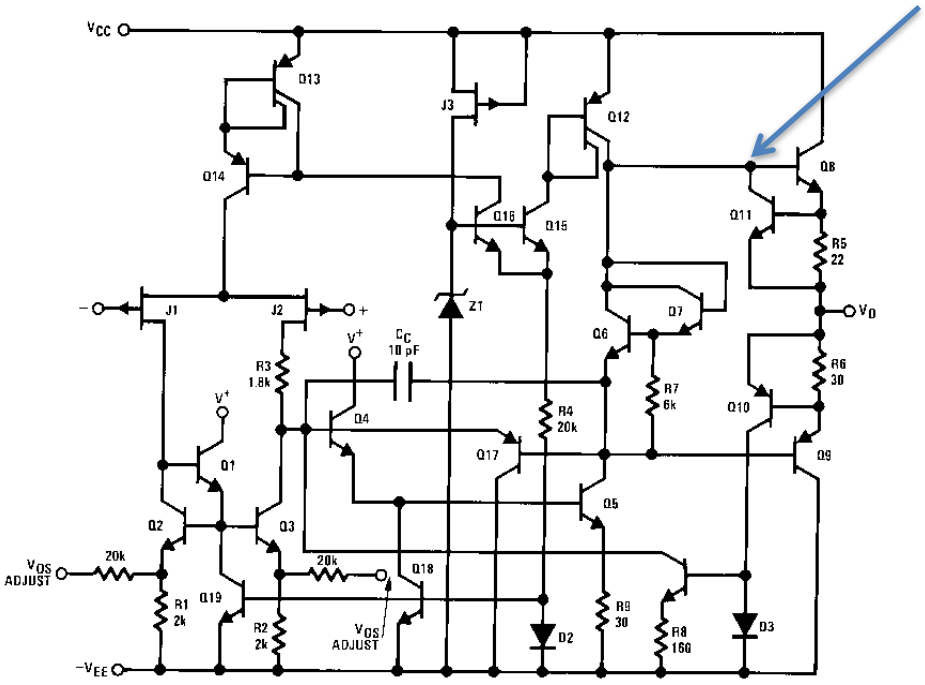
Lumped Circuit Analysis

Circuit schematics, like the one shown below, are the essential language of electronics. From the start it's important to understand the assumptions implicit in such a diagram.

1. The wires in a schematic are assumed to be perfect conductors, everywhere at the same electrical potential.

2. Wires are assumed to be electrically neutral. Any electrical charge is assumed to reside on the plates of the capacitors that are included explicitly in the schematic. All magnetic fields are assumed to be fully localized within the circuit elements such as inductors and transformers.



3. A black dot indicates that conductors are joined together at a *node*. If the conductors cross but there is no black dot, that usually means there is no connection but not all schematics adhere to this practice.

4. The spatial extent of the circuit plays no role. This is the case for *lumped* circuit circuit elements (capacitors, resistors, inductors, diodes, transistors, etc.) for which the size L of the circuit element obeys,

$$L \ll \lambda_{free} = \frac{c}{f}$$

where f is the frequency of operation. The behavior of a lumped circuit is entirely described by the current going through it and the voltage between its terminals. For a resistor, V = IR, for an inductor V = L dI/dt and for a capacitor, I = C dV/dt.

When the physical size of the circuit plays a role we're in the realm of *distributed circuits*. Transmission lines are a good example of this situation.

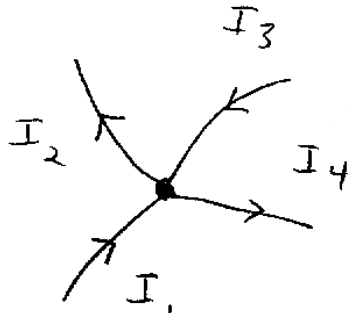
Kirchoff's Laws

The goal of circuit analysis is to find all the voltages and currents in a circuit. To do this we utilize Kirchoff's laws, which are derived from Maxwell's equations subject to the assumptions specific to lumped circuits. An excellent discussion can be found in the *Feynman Lectures on Physics*, Vol 2.

KCL

In circuit analysis, a *node* is defined to be a point where conductors join together, indicated by a black dot. If we assume there is no accumulation of charge at a node, then by charge conservation, the total electrical current entering a node must equal the total current leaving the node, *at each instant*. This is known as **Kirchoff's Current Law (KCL)**.

A node



The figure shows a node with 4 currents, some entering and some leaving. Currents entering the node have the opposite sign from those leaving. KCL applied to this node gives,

$$I_1 - I_2 + I_3 - I_4 = 0$$

In general, KCL states that for *each* node in the circuit,

$$\sum_i I_i(t) = 0 \qquad \boldsymbol{KCL}$$

Charge *can* accumulate at certain places in a circuit, but that will be accounted for by explicitly including capacitors.

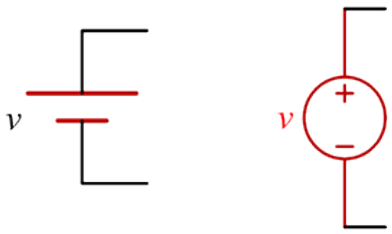
KVL

In lumped circuit analysis we assume that any time-dependent magnetic fields are confined to circuit elements such as inductors or transformers. Faraday's law then implies that *outside* these elements, $curl\ \vec{E} = 0$. From this we know the following things are true,

$$\oint \vec{E} \cdot d\vec{r} = 0 \qquad \vec{E} = -\nabla V$$

$$\rightarrow \sum_{Loop} \Delta V = 0 \qquad (KVL)$$

Voltage sources



DC voltage sources

AC voltage source

An ideal voltage source, otherwise known as a source of electromotive force (EMF) maintains the same potential difference across its terminals regardless of what is connected to the terminals. The two symbols shown on the left are typically used for DC (direct current) voltage sources. DC means that the voltage does not change with time. The long bar on the left-hand symbol indicates the (+) terminal. A battery is the most familiar DC source but it is far from ideal. Power supplies using transistors and feedback generally provide much more ideal sources of DC power. As stressed in the previous example, the magnitude and direction of current that flows through the voltage source depends on the rest of the circuit. Current can flow in either direction through the source!

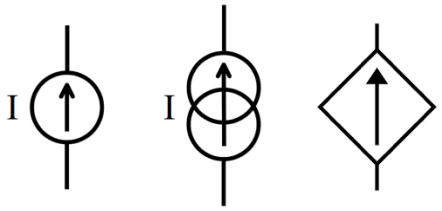
DC sources are essential but a world with nothing but DC sources would be a dull place. In general, voltage sources can depend on time. The schematic symbol for a time-dependent source shown to the left. We sometimes call this an AC (alternating current) source. The fact that it's an *ideal* voltage source means that it maintains this same time-dependent voltage across its terminals, regardless of what we connect to it. By far the most important source for analog electronics is harmonic:

$$V_1 - V_2 = V(t) = V_0\, cos(\omega t + \phi)$$

Harmonic sources are important because of Fourier's theorem. If the voltage source generates a more complicated waveform *V*(t), this can be written as a sum of harmonic waveforms, each with a different frequency. The response of a linear circuit to V(t) will then be the sum of the individual responses to each harmonic source.

There are obvious limits to "ideal" operation. If we we connect a wire across the terminals of either the AC or DC source then the potential across its terminals is zero, period. This is called short-circuiting the source. Real-world sources will generally provide constant voltage up until some limiting current is reached.

Current Sources

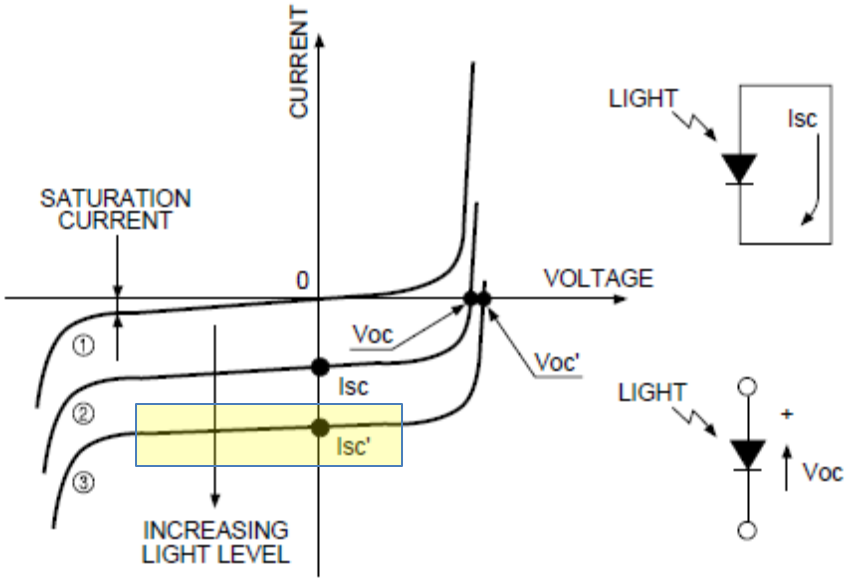


Ideal current sources

An ideal current source is something that maintains the same *current* regardless of the voltage across its terminals. Again, the current may or may not depend on time. Some common schematic symbols are shown. To my knowledge, there is no particular symbol distinguishing a DC from an (AC) current source.

A photodiode or solar cell is probably the best example of a current source, albeit an imperfect one. Its current - voltage characteristic is shown below. As the intensity of light increases the curve shifts downward, increasing the amount of photocurrent. For a given current (determined by the light intensity) the voltage across the photodiode is determined by the external circuitry. It can be negative, positive or zero. The yellow box shows a region over which the current is reasonably independent of the voltage for one particular value of light intensity. If it were an ideal current source the curve would be perfectly flat.

Current sources are less familiar than voltage sources but they are of utmost importance in modern electronics. That's because transistors, over a wide range of operating current and voltage, act like voltage-controllable current sources.



The last equation says that the *changes in electrical potential,* $\Delta V = (V_i - V_j)$ around any closed loop is zero. This result is called **Kirchoff's Voltage Law (KVL).** Labelling points around a circuit as 1, 2,… *N* we have,

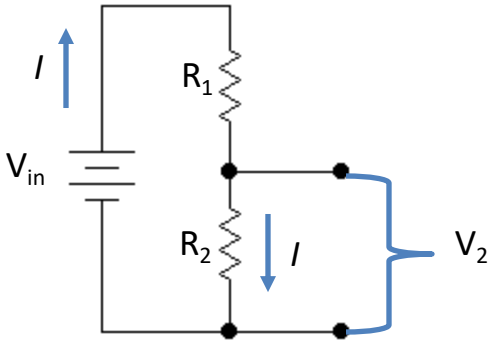
$$(V_1 - V_2) + (V_2 - V_3) + ... + (V_{N-1} - V_N) + (V_N - V_1) = 0 \qquad (KVL)$$

The *double subscript* notation is often used to denote the *difference* in voltage between points a and b, $V_a - V_b = V_{ab}$

Voltage divider

The voltage divider, shown below, is perhaps the most ubiquitous circuit in analog electronics. In the circuit shown, V_{in} is an ideal voltage source – a source of EMF that maintains the same voltage difference between its two terminals regardless of the current flowing through it. Think of it as an ideal battery. There is only one current *I* in this circuit. Defining a clockwise current as positive, using KVL and Ohms Law we have,

$$V_{in} - IR_1 - IR_2 = 0 \quad \rightarrow \quad I = \frac{V_{in}}{R_1 + R_2} \quad \rightarrow \quad V_2 = IR_2 = \frac{R_2}{R_1 + R_2} V_{in}$$

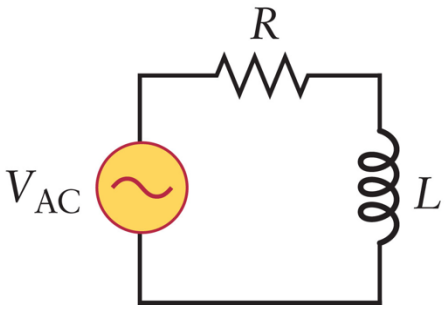


The voltage V_{in} divides between the voltage across R_1 and the voltage across R_2. This can be easily generalized to dividers in which the voltage source is time dependent. By far the most important time-dependent source for analog electronics is harmonic:

$$V_a - V_b = V(t) = V_0 \cos(\omega t + \phi)$$

The schematic symbol for a time-dependent source is shown on the right. The source can have an arbitrary time dependence but in linear circuits sinusoidal sources are most important because we can use Fourier analysis to build up an arbitrary waveform from sinusoids. The important thing is that $V_a - V_b$ is independent of what we hook across the terminals. Of course this has limits. We can't connect a wire across the terminals because then $V_a = V_b$.

Steady State AC circuit analysis



Practically all modern electronics involves time-dependent voltages and currents. To analyze these circuits we will make use of linearity and complex numbers. Consider the simple RL circuit driven by an ideal harmonic voltage source.

Case 1: let $V_{AC} = V_0 \cos(\omega t + \varphi)$ and call the current that flows in response to this voltage $I_{Re}(t)$. KVL gives,

$$V_0 \cos(\omega t + \varphi) = V_R + V_L = I_{Re} R + L \frac{dI_{Re}}{dt}$$

Case 2: let $V_{AC} = V_0 \sin(\omega t + \varphi)$ and call the current that flows $I_{Im}(t)$. KVL gives,

$$V_0 \sin(\omega t + \varphi) = V_R + V_L = I_{Im} R + L \frac{dI_{Im}}{dt}$$

Multiply the equation for case 2 by $i = \sqrt{-1}$ and add the result to the equation for case 1,

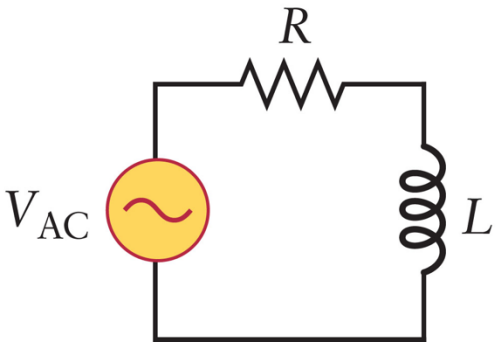
$$V_0(\cos(\omega t + \varphi) + i \sin(\omega t + \varphi)) = (I_{Re} + i\, I_{im}) R + L \frac{d\,(I_{Re} + i\, I_{im})}{dt}$$

$$V_0 e^{i\omega t} e^{i\varphi} = I_{Complex} R + L \frac{dI_{Complex}}{dt} \qquad I_{Complex} = I_{Re} + i\, I_{im}$$

We now substitute a trial solution for the complex current: $I_{complex}(t) = \hat{I}(\omega)e^{i\omega t}$

Plug this into our differential equation,

$$V_0 e^{i\omega t} e^{i\varphi} = \hat{I}(\omega)e^{i\omega t}R + L\frac{d}{dt}\left(\hat{I}(\omega)e^{i\omega t}\right)$$



The $e^{i\omega t}$ factors just cancel out leaving just algebra,

$$\hat{V}_0 = V_0 e^{i\varphi} = \hat{I}(\omega)R + i\omega L\,\hat{I}(\omega)$$

$$\hat{I} = \frac{\hat{V}_0}{R + i\omega L}$$

To recover the physical solution, suppose the excitation voltage is $V_0 \cos(\omega t + \varphi)$. Then corresponding current is,

$$I_{Re}(t) = Re\left[\hat{I}(\omega)e^{i\omega t}\right] = Re\left[\frac{\hat{V}_0}{R + i\omega L}\, e^{i\omega t}\right]$$

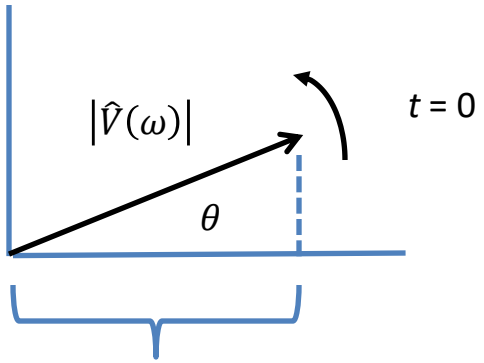
So the rule is to represent physical currents and voltages by the real part of complex quantities called *phasors*,

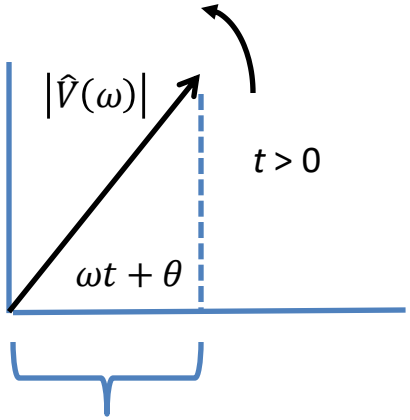
$$I(t) = Re\left[\hat{I}(\omega)e^{i\omega t}\right] \qquad V(t) = Re\left[\,\hat{V}(\omega)e^{i\omega t}\right]$$

With phasors, it's generally much more useful to write complex numbers in terms of amplitude and phase. In general,

$$V(t) = Re\left[\hat{V}(\omega)e^{i\omega t}\right] = Re\left[\left|\hat{V}(\omega)\right| e^{i\omega t}\, e^{i\theta}\right] = \left|\hat{V}(\omega)\right| Re\left[\, e^{i\omega t}\, e^{i\theta}\right] = \left|\hat{V}(\omega)\right| \cos(\omega t + \theta)$$

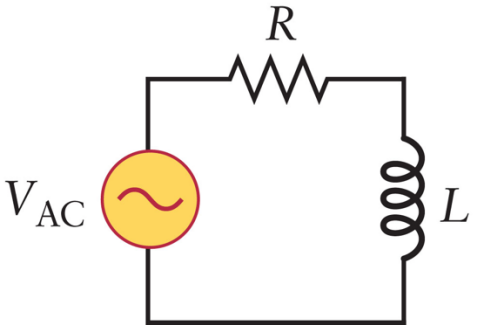
The phasor $\hat{V}(\omega)e^{i\omega t}$ rotates counterclockwise at angular velocity ω. The physical voltage or current is the projection of the phasor onto the real axis.





The point of this analysis is that we can apply KVL and KCL to the complex amplitudes alone. The time dependence has been factored out. Returning to the original circuit, we ended up with,

$$\hat{V}_0 = \hat{I}\, R + i\omega L\, \hat{I}$$



This looks like KVL applied to a DC circuit except that the coefficients of current have been replaced by complex numbers,

$$\hat{V}_0 = \hat{I}\, Z_R + \hat{I}\, Z_L$$

$$Z_R = R \quad Z_L = i\omega L$$

For each linear circuit element, the complex amplitude of the voltage is proportional to the complex amplitude of the current through it. The constant of proportionality is called the *impedance Z.*

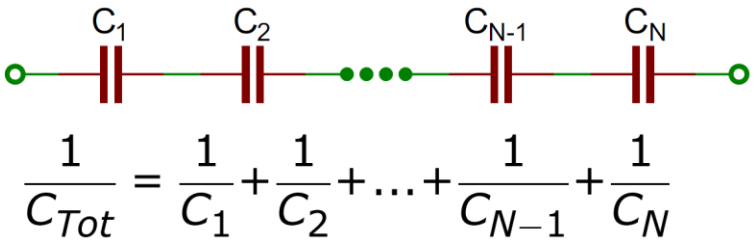
$$\hat{V}\,(\omega) = Z(\omega)\,\hat{I}\,(\omega)$$

The impedances for R, L and C are given by, $Z_R = R \quad Z_L = i\omega L \quad Z_C = \frac{1}{i\omega C}$

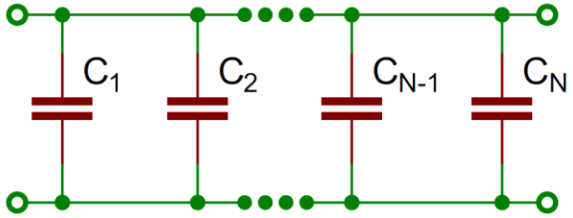
Using KCL and KVL, as we did for combinations of resistors, it's easy to show that series and parallel combinations of impedances obey,

$$Z_{series} = \sum_{i=1}^{N} Z_i \qquad \frac{1}{Z_{parallel}} = \sum_{i=1}^{N} \frac{1}{Z_i}$$

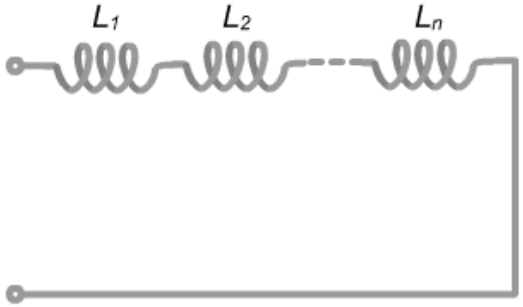
From here, the usual expressions for parallel and series capacitors and inductors easily follow:



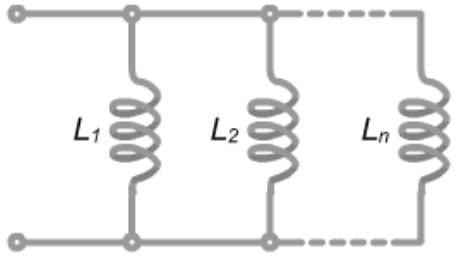
$$\frac{1}{C_{Tot}} = \frac{1}{C_1} + \frac{1}{C_2} + \ldots + \frac{1}{C_{N-1}} + \frac{1}{C_N}$$



$$C_{Tot} = C_1 + C_2 + \ldots + C_{N-1} + C_N$$

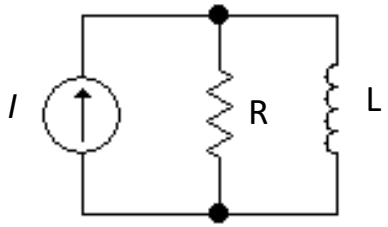


$$L_T = L_1 + L_2 + L_3 + \ldots L_n$$



$$\frac{1}{L_T} = \frac{1}{L_1} + \frac{1}{L_2} + \ldots \frac{1}{L_n}$$

Example



In the circuit shown $I_{gen}(t) = I_0\, sin(\omega t)$. Find the current through the inductor.

First, the resistor and inductor form a parallel impedance given by,

$$Z = Z_R \parallel Z_L = \frac{Z_R\, Z_L}{Z_R + Z_L} = \frac{R\, i\omega L}{R + i\omega L}$$

The voltage across this impedance is $\hat{V}_Z = \hat{V}_R = \hat{V}_L = \hat{I}\, Z$. Next, find the phasor representing the current source,

$$I_{gen}(t) = I_0\, sin(\omega t) = Re\left(\hat{I}\, e^{i\omega t}\right) \rightarrow \hat{I} = -i\, I_0$$

The phasor amplitude representing the inductor current is given by

$$\hat{I}_L = \frac{\hat{V}_L}{Z_L} = \frac{R\,\hat{I}}{R + i\omega L} = \frac{-\,i\, I_0\, R}{R + i\omega L} = \frac{I_0\, R}{\sqrt{R^2 + (\omega L)^2}}\, e^{i\,\phi} \qquad \phi = \pi + arctan\left(\frac{R}{\omega L}\right)$$

The physical current is now given by, $$I_L(t) = Re\left(\hat{I}_L\, e^{i\omega t}\right) = \frac{I_0\, R}{\sqrt{R^2 + (\omega L)^2}}\, cos(\omega t + \phi)$$

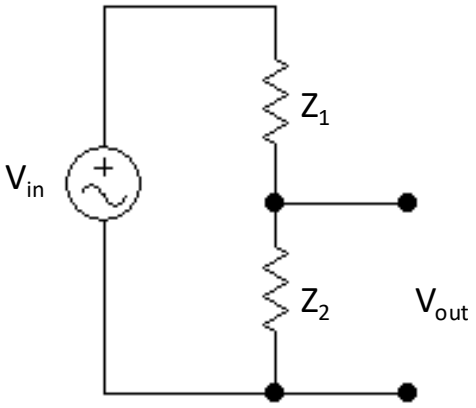
Please don't do this:

$$I_L(t) = Re\left(\frac{R}{R + i\omega L}\, I_0\, sin\omega t\right)$$

This expression has no phase shift for $sin\omega t$. You need to *first* convert the generator current into a phasor ($\hat{I} = -i\, I_0$) and *then* take $Re\left(\hat{I}_L\, e^{i\omega t}\right)$. The point of the phasor method is to factor out the time dependence from the circuit equations leaving just complex algebra. Then put the time dependence back in at the end.

RC low pass filter

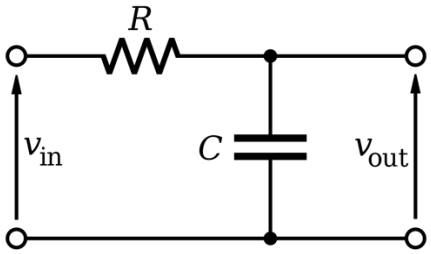
We can now solve the differential equations for harmonically driven linear circuits with just algebra, albeit with complex numbers. Let's do an important example – the complex voltage divider. V_{in} is the sinusoidal source voltage driving the series combination of Z_1 and Z_2. Representing V_{in} and V_{out} by their complex amplitudes we have,



$$\hat{V}_{out} = \frac{Z_2}{Z_1 + Z_2}\hat{V}_{in}$$

Z_1 and Z_2 can be any combination of resistors, capacitors or inductors. Specializing to the case where $Z_1 = R$ and $Z_2 = 1/i\omega C$ we have the output phasor amplitude,

$$\frac{\hat{V}_{out}}{\hat{V}_{in}} = \frac{1}{1 + i\,\omega RC}$$



The ratio of the complex output to the complex input is known as the *transfer function*:

$$\frac{\hat{V}_{out}}{\hat{V}_{in}} = \frac{1}{1 + i\,\omega RC} = \left|\frac{\hat{V}_{out}}{\hat{V}_{in}}\right| e^{i\varphi(\omega)}$$

The magnitude and phase of the transfer function are given by,

$$\left|\frac{\hat{V}_{out}}{\hat{V}_{in}}\right| = \frac{1}{\sqrt{1+(wRC)^2}} = magnitude$$

$$\varphi(\omega) = \arctan(-\omega RC) = -\arctan(\omega RC)$$

The magnitude decreases as the frequency goes up. This circuit is therefore a *low-pass filter*. It *passes* signals with angular frequencies well below $\omega_0 = 1/RC$ and it *attenuates* signals with frequencies well above this value. Right at $\omega_0 = 1/RC$ the amplitude of the output has fallen by $1/\sqrt{2}$ relative to the input.

Decibels and Bode plots

In electronics we usually need to know the operation of a circuit over many decades in frequency. As such, it is useful to plot things on a log scale. Define the decibel (dB) unit for any transfer function as,

$$\left|\frac{\hat{V}_{out}}{\hat{V}_{in}}\right|_{dB} = 20\log_{10}\left|\frac{\hat{V}_{out}}{\hat{V}_{in}}\right|$$

Specializing to the RC low pass filter we have,

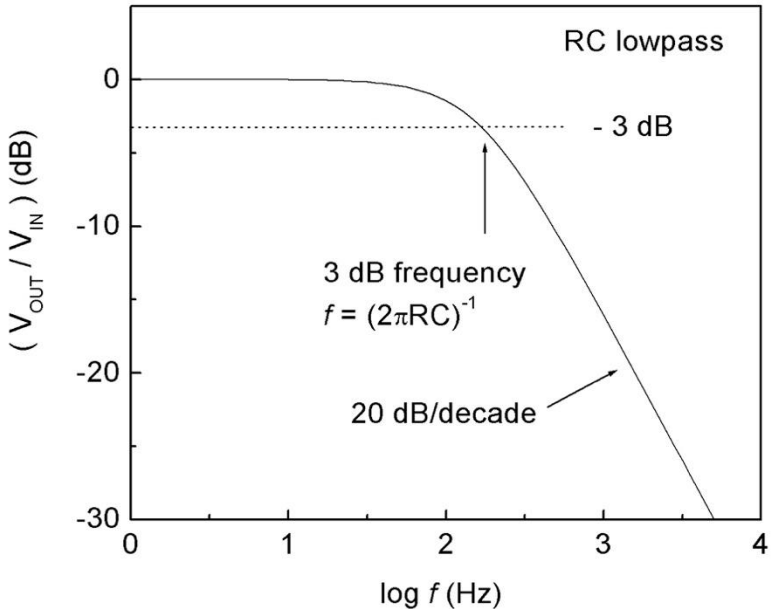
$$\left|\frac{\hat{V}_{out}}{\hat{V}_{in}}\right|_{dB} = -20\ log_{10}\sqrt{1+(\omega RC)^2}$$

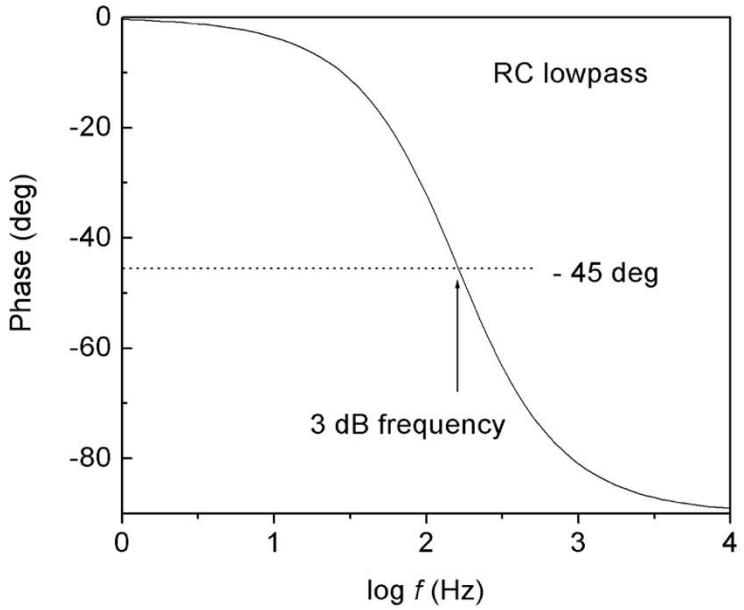
When ω >> 1/RC this expression looks like,

$$\left|\frac{\hat{V}_{out}}{\hat{V}_{in}}\right|_{dB} = 20\log_{10}\frac{1}{\sqrt{1+\left(wRC\right)^2}} \xrightarrow{w>>1/RC} -\ 20\log_{10}\left(wRC\right) = -\ 20\log_{10} w + const$$

So when ω >> 1/RC, plotting the amplitude of the transfer function (in dB) versus $\log\omega$ yields a straight line with slope -20 dB/decade. Alternatively, we could say that when ω >> 1/RC the output falls at approximately 6 dB/octave in frequency. Right at ω = 1/RC the amplitude has fallen by exactly $1/\sqrt{2}$ or approximately 3 dB since -3 = 20 $\log_{10}$ ($1/2^{1/2}$)) . ω_c =1/RC is therefore called the *3 dB frequency* of the filter.

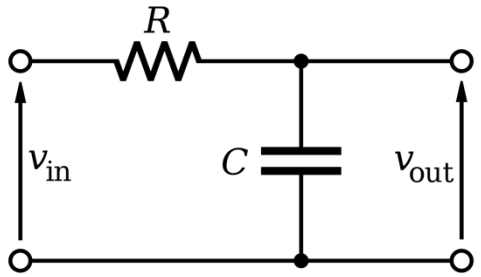
The figures below show the amplitude and phase of the low pass filter response plotted versus log*f* where the frequency *f* = ω/2π . This kind of figure is known as a *Bode plot*. The phase of the low pass filter begins at 0° , reaches - 45° at the 3 dB frequency and saturates at
– 90° when *f* = ∞.



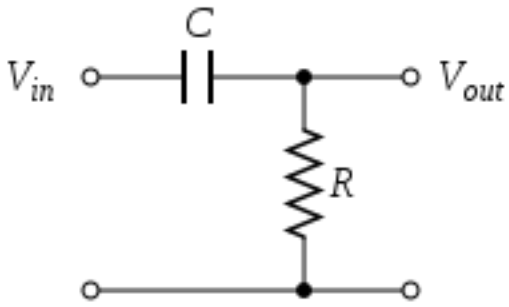


Bode Plots

Bode plots of amplitude and phase of the transfer function for an RC low pass filter with R = 1 kΩ and C = 1 µF. 3 dB frequency = 159.15 Hz



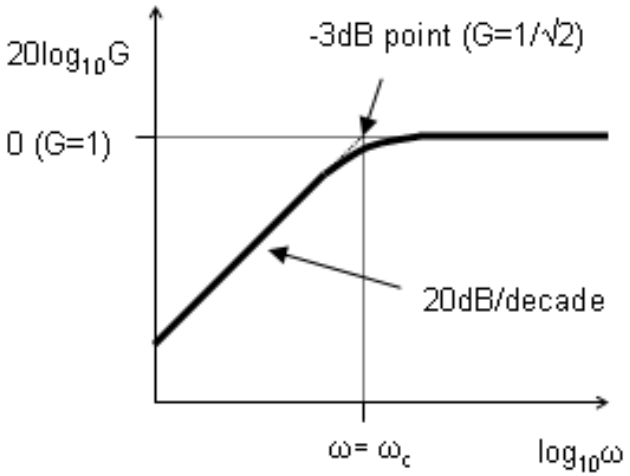
High pass filter



Switching the roles of resistor and capacitor leads to an RC high pass filter, shown on the left. The transfer function is,

$$\left.\frac{\hat{V}_{out}}{\hat{V}_{in}}\right]_{HP} = \frac{i\,\omega RC}{1 + i\,\omega RC}$$

$$\left|\frac{\hat{V}_{out}}{\hat{V}_{in}}\right| = \sqrt{\frac{\omega RC}{1 + (\omega RC)^2}} \qquad \varphi(\omega) = \arctan\left(\frac{1}{\omega RC}\right)$$

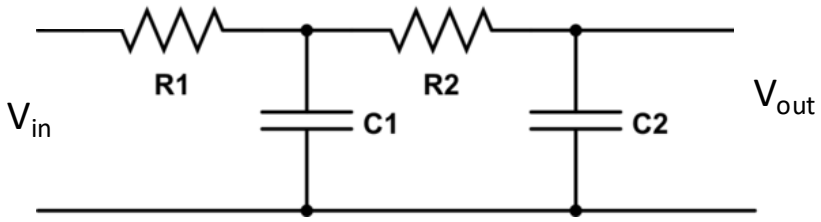


The Bode plot for the magnitude is shown on the left. When f << 1/2πRC the magnitude is approximately ωRC so it *increases* by 20 dB/decade. For f >> 1/2πRC the gain approaches 1, as expected for a high pass filter.

For f = 0 the phase is 90° , reaching 45° when f = 1/2πRC and approaching 0° for f >> 1/RC .

Multiple pole filters

The RC low pass is an example of a what is called a *single pole* filter. Suppose we make a more complicated filter with 2 R's and 2 C's. This is a 2-pole filter where a pole is a value of iω that sends the denominator to zero. The transfer function now has the form,



$$\frac{\hat{V}_{out}}{\hat{V}_{in}} = \frac{1}{1 + i\,\omega/\omega_1}\frac{1}{1 + i\,\omega/\omega_2}$$

With some algebra we can find ω_1 and ω_2 in terms of R_1, R_2, C_1, C_2 but the result is not particularly illuminating. Using $f = \omega/2\pi$ the magnitude is given by,

$$\left|\frac{\hat{V}_{out}}{\hat{V}_{in}}\right| = \frac{1}{\sqrt{1 + (f/f_1)^2}}\frac{1}{\sqrt{1 + (f/f_2)^2}}$$

To make a Bode plot, express the magnitude in decibels:

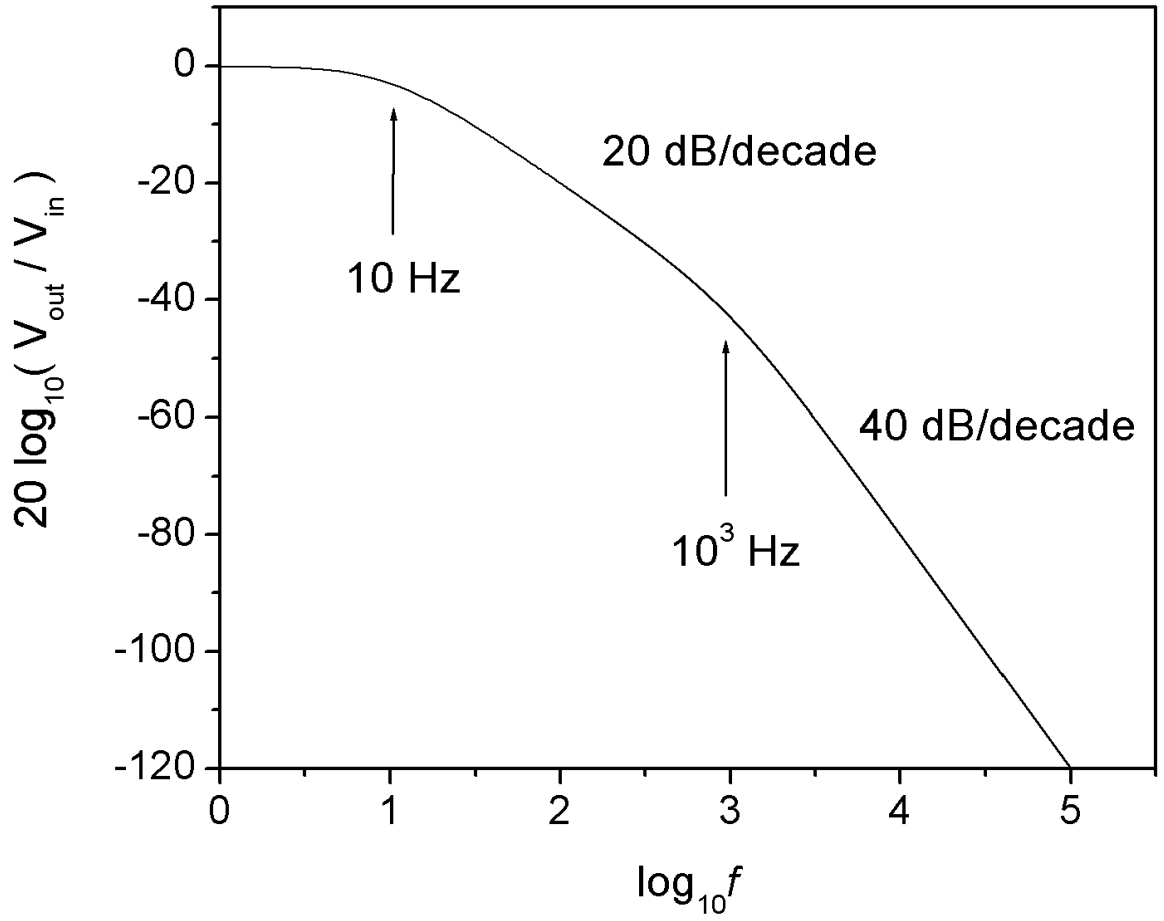
$$\left|\frac{\hat{V}_{out}}{\hat{V}_{in}}\right|_{dB} = 20\,log_{10}\left[\frac{1}{\sqrt{1 + (f/f_1)^2}}\frac{1}{\sqrt{1 + (f/f_2)^2}}\right]$$

$$= 0 \quad (f \ll f_1)$$

$$= -20\,log_{10}\,f + constant \quad (f_1 \ll f \ll f_2)$$

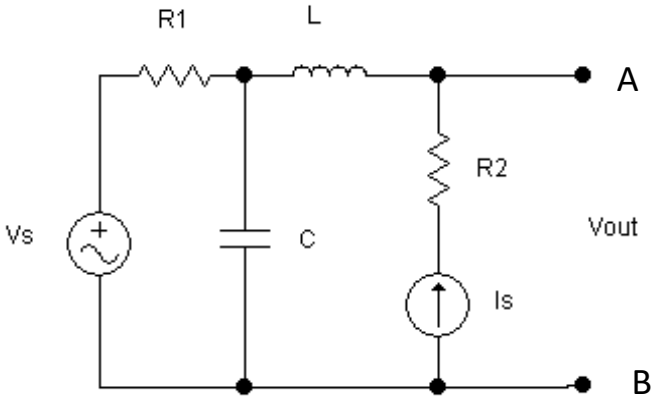
$$= -40\,log_{10}\,f + constant \quad (f_2 \ll f)$$

The Bode plot now shows 2 corner frequencies, at f_0 =10 Hz and f_1 = 1000 Hz. Between 10 Hz and 1000 Hz the magnitude of V_{out}/V_{in} drops by 20 dB/decade. Above 1000 Hz it drops by 40 dB/decade. The phase goes from 0° at f = 0 to -45° ($f \approx f_1$) to -135° ($f \approx f_2$) to $-180°$ as $f \rightarrow \infty$. The more poles, the more the filter attenuates as the frequency goes up. For an n-pole filter, the transfer function falls by 20n dB/decade once the frequency is well above the highest pole frequency.



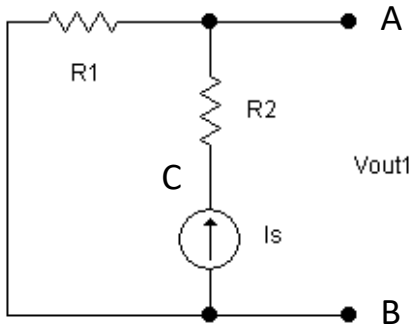
Superposition

Kirchoff's laws, being derived from Maxwell's equations, are linear. If the circuit elements are also linear then we can use the principle of *superposition* to simplify the analysis. Superposition implies that the response of the circuit to all the voltage and current sources is the sum of the individual responses. To apply superposition, turn off all but one voltage (or current) source and solve for the voltages and currents in the circuit. Then turn all but the next source off and find the voltages and currents, and so on. Then add all the solutions together to obtain the response when all the sources are on simultaneously. Consider the circuit shown below. Consistent with the vast majority of modern circuits, there are both AC and DC sources. In this example, assume the current source is DC and the voltage source is harmonic. Now use superposition to find $V_A - V_B = V_{out}$.



Step 1: Turn off the V_S and find the output voltage due just to the DC current source I_S. Turning V_S off amounts to setting $V_S = 0$ which means replacing it with a wire. The effective circuit now becomes the series combination of R_1 and R_2. Recall that for DC, a capacitor is an open circuit and an inductor is a short circuit. The effective circuit with $V_S = 0$ is shown next.

V_{out1} is the terminal voltage due just to I_S. It is given by $V_{out1} = I_s R_1$



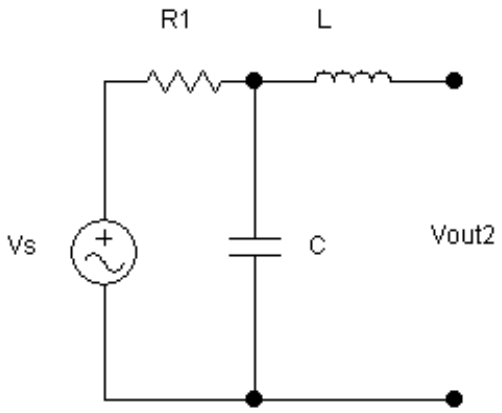
As an aside, what's the voltage $V_C - V_B$ across the current source? Just use KVL.

$$(V_B - V_C) + (V_C - V_A) + (V_A - V_B) = (V_B - V_C) + I_S R_2 + I_s R_1 = 0$$

$$(V_C - V_B) = I_S (R_1 + R_2)$$

If this confuses you, remember that the current source generates *constant current*. The voltage across it is determined by I_S and the rest of the circuit.

Step 2. Turn V_s back on and set $I_S = 0$. Turning off a current source implies zero current, i.e., an *open circuit*. Therefore the effective circuit with $I_S = 0$ is shown. If we measure the output voltage with a device that draws no current (an ideal voltmeter) then no current passes through the inductor so it has not voltage across it. Therefore V_{out2} becomes the response of an RC low pass filter,



$$\hat{V}_{out2}(\omega) = \frac{\hat{V}_S}{1 + i\omega R_1 C}$$

By superposition, the total voltage on the output terminals is now given by the sum,

$$V_{out}(t) = V_{out1} + V_{out2} = I_S R_1 + Re\left[\frac{\hat{V}_S}{1 + i\omega R_1 C}\, e^{i\omega t}\right]$$

Suppose the AC voltage is given by $V_s(t) = V_0 \sin \omega t$. Then we need to represent it by a phasor:

$$V_S(t) = V_0 \sin(\omega t) = Re\left[\hat{V}_S\, e^{i\omega t}\right] \quad \rightarrow \hat{V}_S = -i\, V_0$$

The physical output voltage as a function of time is now given by,

$$V_{out}(t) = I_S R_1 + Re\left[\frac{\hat{V}_S}{1 + i\omega R_1 C}\, e^{i\omega t}\right] = I_S R_1 + Re\left[\frac{-i\hat{V}_0}{1 + i\omega R_1 C}\, e^{i\omega t}\right]$$

$$V_{out}(t) = I_S R_1 - \frac{V_0}{\sqrt{1 + (\omega R_1 C)^2}} \cos\big(\omega t + \varphi(\omega)\big) \qquad \varphi(\omega) = arctan\left(\frac{1}{\omega R_1 C}\right)$$

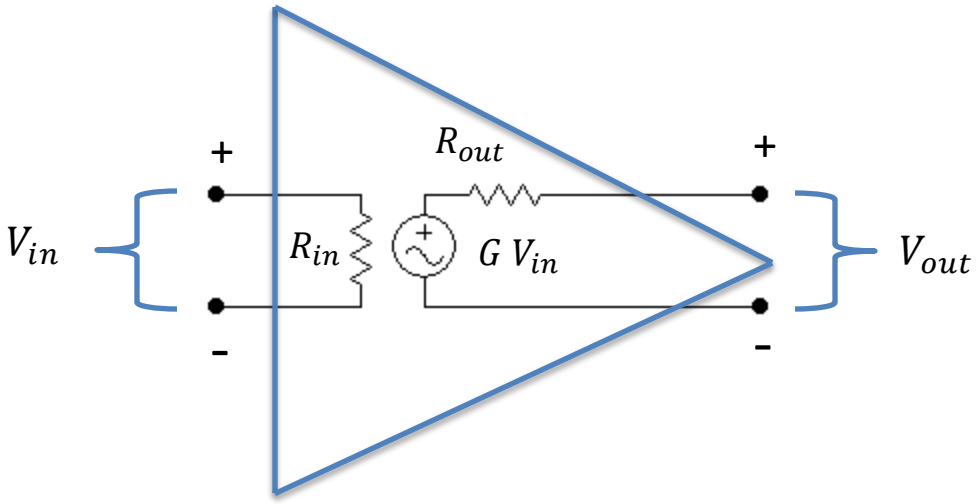
Transforming the complex amplitudes back into the real time signal is often not needed. The complex amplitude of the response, by itself, is often all we need to know.

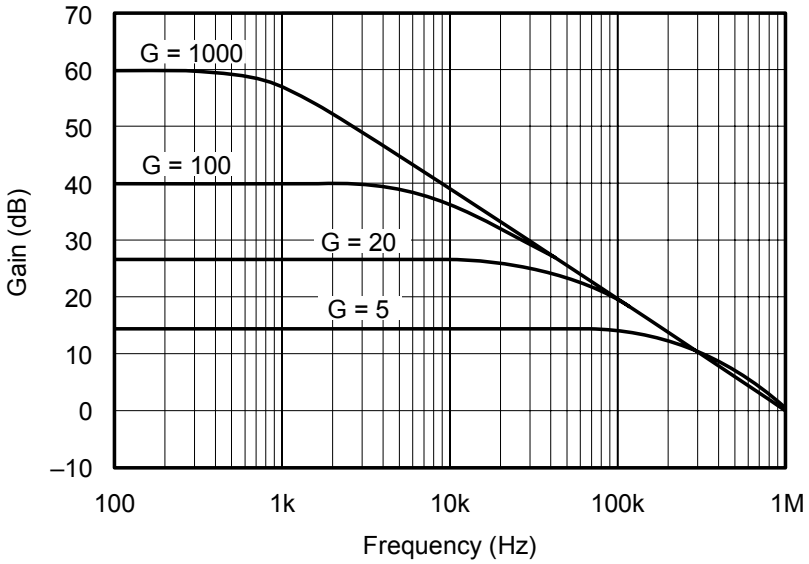
Amplifiers

Amplifiers are indispensable to electronics. The schematic shows the effective circuit of a voltage amplifier. A voltage V_{in} is applied between the (+) and (-) input terminals and a voltage $V_{out} = G\,V_{in}$ is generated across the output terminals. *G* is called the *gain*. It is often shown in decibels where,

$$G(dB) = 20\,log_{10}G$$

In all amplifiers the gain depends on the frequency of the signal its amplifying. I've shown the gain versus frequency for a typical amplifier in which you can set G externally to values anywhere from 5 to 10,000. Above a characteristic frequency, often called the *bandwidth (BW)* , the gain begins to fall off. To be more precise, the bandwidth is the frequency as which the gain falls to $1/\sqrt{2}$ of its value at very low frequency. That's also called its 3 dB frequency because the gain has fallen by about 3 decibels from its low frequency value. For this amp, if you set the gain to G = 100, the bandwidth is about 5 kHz. If you look closely you'll see that $G * BW = constant$.

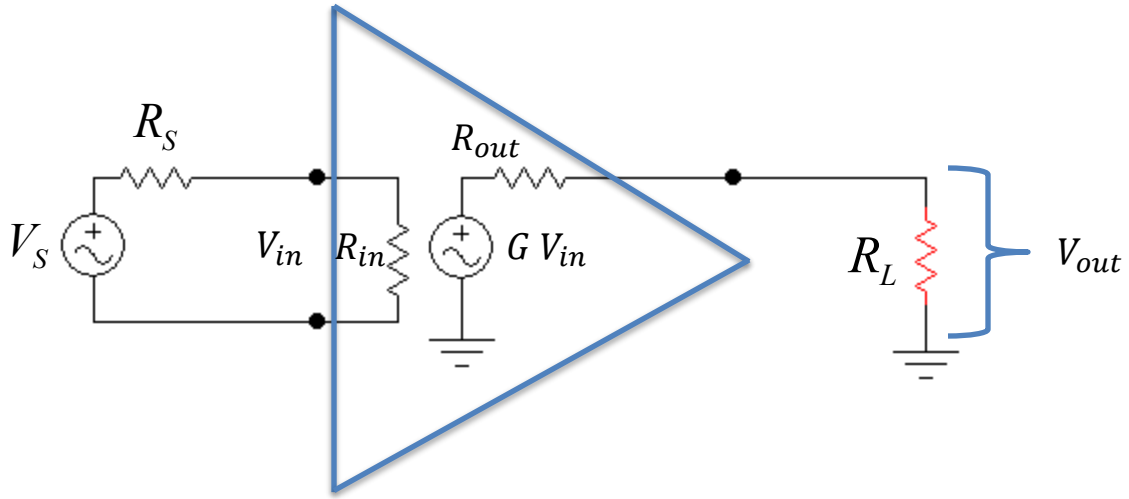




To an incoming signal (V_{in}) the amplifier appears like a resistor R_{in} called the *input resistance*. This could be anywhere from a few Ohms to maybe 10^8 Ohms, depending on the amp. The amplifier senses the voltage between its (+) and (-) input terminals, multiplies it by G and applies this voltage to the output circuitry. There is an unavoidable resistance R_{out} in series with the output voltage generator. R_{out} is called the *output resistance* or more generally the output impedance. For voltage amplifiers R_{out} might range from a few Ohms to a few thousand Ohms. For high frequency amps it is generally made to be 50 Ohms.

Effect of input and output impedance

Both the input and output impedance of an amplifier affect the final signal size. Imagine we drive the amp with a voltage source V_S whose source resistance is R_S. The voltage V_S is generated by sound waves moving a membrane in the microphone.



Before proceeding, notice that I've attached one side of the output voltage to ground (i.e., 0 Volts). That's usually the case but not always. You should assume that all points connected to the ground symbol are at the same electrical potential and that current can move freely between then.

1. Look first at the input circuit. By voltage division we have,

$$V_{in} = \frac{R_{in}}{R_S + R_{in}} V_S$$

For a typical electret microphone R_S might be $2\ k\Omega$. Suppose the amplifier has $R_{in} = 200\ \Omega$. Then $V_{in} = 0.1\ V_S$. V_S is what we care about but V_{in} is what the amplifier gets to amplify, so we've already cut down our intended signal by 10. However, if we choose an amp with $R_{in} \gg R_S$ then $V_{in} \approx V_S$. You can see why it is usually desirable for a voltage amplifier to have R_{in} as large as possible.

2. Now look at the output signal. Usually we want the output to drive something, maybe a loudspeaker. This has some resistance we'll call R_L. (For a loudspeaker it might be $8\ \Omega$.) By voltage division, the voltage that appears across the speaker resistance is given by,

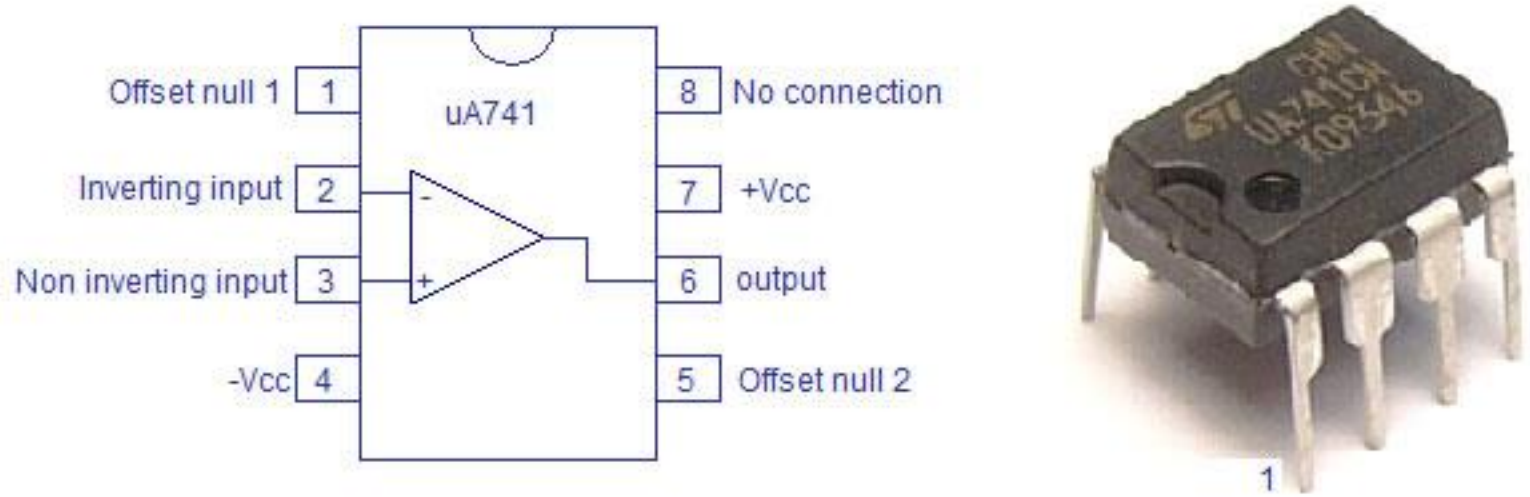
$$V_{out} = \frac{R_L}{R_{out} + R_L} G\ V_{in}$$

If $R_{out} \gg R_L$ then $V_{out} \ll G\ V_S$ and again, we'll lose a lot of the amplified signal. Therefore, for the output stage, it is desirable to have $R_L \gg R_{out}$ in which case $V_L \approx G\ V_{in}$. Therefore voltage amplifiers are often designed to have very *low* output resistance.

Operational Amplifiers

The op amp is the building block of analog electronics. It is essentially an extremely high gain ($10^5 – 10^6$) differential amplifier with high input resistance (> 1 MΩ) moderately low output resistance (100 Ω or less) and a bandwidth of about 10 Hz. You can now buy them with much larger bandwidths but there is a sacrifice in performance. And even with 10 Hz of bandwidth, we'll see that the very large gain makes the op amp useful over a much wider range of frequency. The figure below shows the pinout for the venerable 741 op amp. This chip style is called DIP for *dual inline package*. The circular indentation marks pin 1. The 741 is long out of date but the principles are the same for most op amps.

Op amps can operate over wide ranges of supply voltage. The 741 will operate V_{CC} =10 - 15 V and $–V_{CC}$ = -10 to - 10 V. The LF 411 will operate from ± 18 V down to ± 4.5 V and the ± supplies don't even have to be symmetric. Newer op amps, designed for higher speeds and lower power consumption may operate with as little as ± 2.5 V or just a single 5 V supply. How all of this is achieved is the subject of transistor amplifier design. We'll just talk about how to *use* op amps.

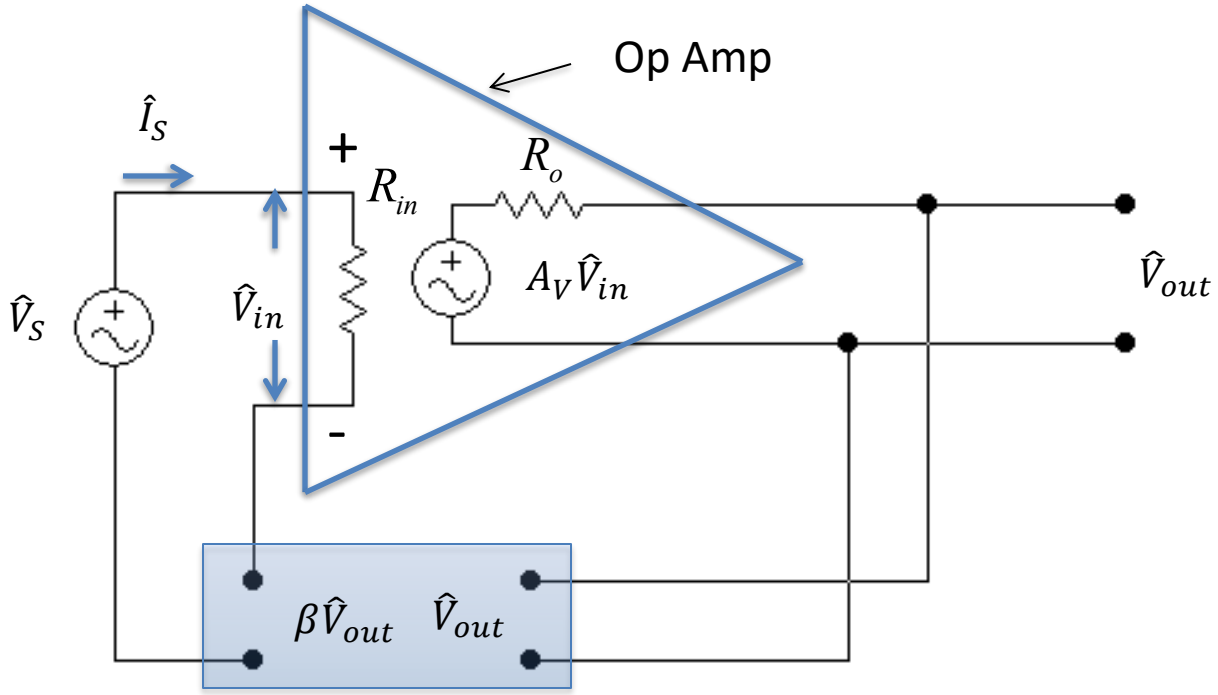


uA741 opamp Pinout and External appearance

Feedback

To be consistent with specification sheets, op amp gain is generally denoted A_V, not G. A_V is called the *open loop gain*. It should be considered a complex quantity which is real at low frequencies but picks up a phase shift at higher frequenices, essentially like a low pass filter. Therefore I will use phasors to denote the currents and voltages. Op amps are rarely used "open loop", i.e., to directly amplify signals. With a gain of 10^5 or more, even a tiny signal across the input terminals will drive the output to a voltage beyond the power supply levels and render the circuit useless. To make sure the amplifier operates as a linear device it is generally used with *feedback* as shown below. With feedback applied, signals (V_S) of even a few volts can be amplified while V_{in} remains extremely small. Feedback will magically improve many other things too.

Feedback is achieved by connecting the output of the op amp to the shaded box. Often this feedback box is just a voltage divider, but in principle it could be anything. This particular way of connecting things is known as *voltage-series feedback.* We sample the output voltage and feed back a portion βV_{out} in series with the input signal. Think of the β box as sampling V_{out} without drawing any significant current from the output. In general, β can be complex. We will now calculate the gain *with* feedback, $\hat{V}_{out}/\hat{V}_S$ otherwise known as the *closed loop gain*.



Closed loop gain

Apply KVL around the input loop and use the definition of the open loop gain A_V,

$$\hat{V}_S = \hat{V}_{in} + \beta\hat{V}_{out} = \frac{\hat{V}_{out}}{A_V} + \beta\hat{V}_{out}$$

$$\frac{\hat{V}_{out}}{\hat{V}_S} = \frac{A_V}{1+\beta A_V} = A_V^{closed}$$

Now use the fact that the magnitude of the open loop gain $A_V >> 1$. In that case, the closed loop gain reduces to,

$$A_V^{closed} = \frac{A_V}{1+\beta A_V} \approx \frac{1}{\beta} \qquad |A_V| \gg 1$$

By making the open loop gain very large, we've made the closed loop gain *independent* of the amplifier details. It depends only on β of the feedback network, which might be as simple as a voltage divider. And if so, it can be made with stable and predictable components. By contrast, the magnitude of the open loop gain might vary from 82,000 for one op amp to 88,513 for another one. But if it's large enough, then it drops out of the formula for the closed loop gain.

Going back to the first equation we can also solve for V_{in},

$$\hat{V}_{in} = \frac{\hat{V}_{out}}{A_V} = \frac{\hat{V}_S}{1+\beta A_V} \ll \hat{V}_S \quad A_V \gg 1$$

In the limit of very large A_V, V_{in} becomes so small compared to V_S that we can approximate it by zero. This leads to Op Amp Rule #1:

Rule #1: In an ideal op amp with negative feedback, V_{in}, the voltage between the input terminals, is reduced to zero.

Next, let's look at how much current flows into the op amp input terminals. Using the definition of the input resistance,

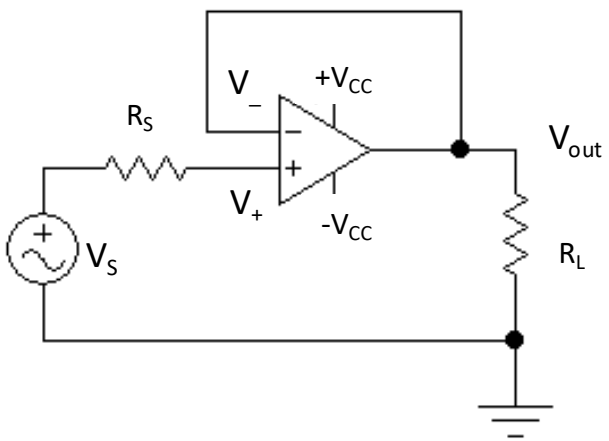
$$\hat{I}_s = \frac{\hat{V}_{in}}{R_{in}}$$

But we just showed that V_{in} is extremely small. (Typically less than 10^{-5} V.) And for op amps, R_{in} can be anywhere from $10^6 - 10^9 \Omega$ so the current is of order 10^{-11} A or less. That's *very* small compared to the other currents typically running around in circuits. Therefore we'll lay down ideal op amp rule #2:

Rule #2. When an ideal op amp is wired with negative feedback, no current flows in or out of its input terminals.

These rules make the analysis of op amp circuits simple. They are both approximations, of course. In rule #1, V_{in} is very small but not exactly zero, otherwise there would be no output. Strictly speaking the rules hold in the limit that $\beta A_V \gg 1$. However, as the frequency increases A_V decreases and the amp becomes less and less ideal.

The Follower



This is the simplest op amp circuit. *All* of the output is fed back to the input. Op amp rule #1 says that $V_+ = V_-$. Since the output is connected directly to the (-) terminal then $V_{out} = V_- = V_+$. We're applying a signal V_S but by rule #2 no current flows in or out of the (-) terminal so no current flows through R_S so $V_S = V_- = V_+ = V_{out}$. The ratio of the output to the input we apply is called the *closed loop gain*:

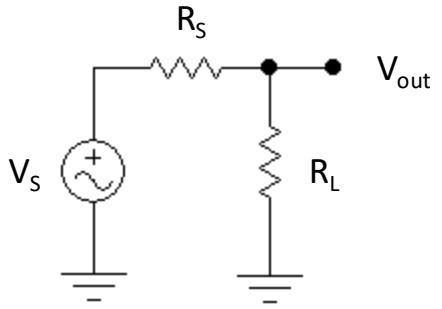
$$Closed\ Loop\ Gain = \frac{\hat{V}_{out}}{\hat{V}_S} = 1$$

That's to be distinguished from the open loop gain $A_V = V_{out}/(V_+ - V_-)$, which might be 100,00. Since $V_{out} = V_S$ then this configuration is called a *follower*.

I've also explicitly shown wires labelled $+V_{CC}$ and $-V_{CC}$. These are DC voltages required to power the op amp. Typical values are $\pm 5\ V$. After this point I won't show these power connections but assume they are there. Also, I've explicitly shown the connection to ground. For most op amps the output voltage is defined relative to ground potential.

Why build an amp with a gain of 1? Think about making a direct connection between the transducer and the load resistor R_L. That's just a voltage divider so,

$$\hat{V}_{out} = \frac{R_L}{R_L + R_S}\hat{V}_S \qquad (divider)$$

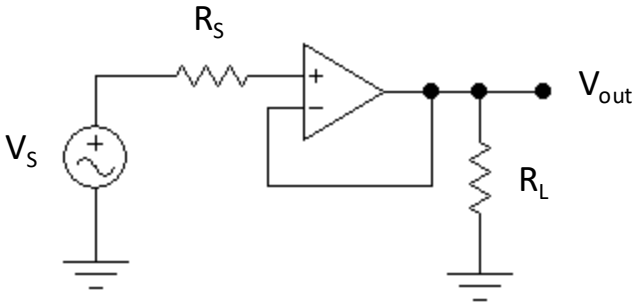


If $R_L << R_S$ (often the case), we'd lose most of our signal voltage. Now insert the op amp follower. This time we get all the signal voltage across R_L because the op amp's input terminals draw negligible current.

$$\hat{V}_{out} = \hat{V}_S \qquad (with\ follower)$$

For this reason, the follower is often called a unity gain *buffer*. Apropos of the earlier discuss of amplifier input and output resistance, what happened to the op amp's output impedance? Doesn't that play a role? It takes a little work to show but feedback reduces the op amps original output resistance to a very small value,

$$R_{out} \longrightarrow R_{out}/(1 + \beta A_V) \ll R_{out}$$

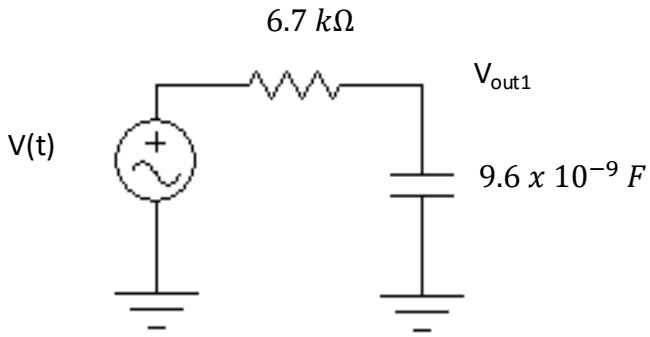


So again, the voltage division problem we might have in an amplifier with large output resistance is eliminated.

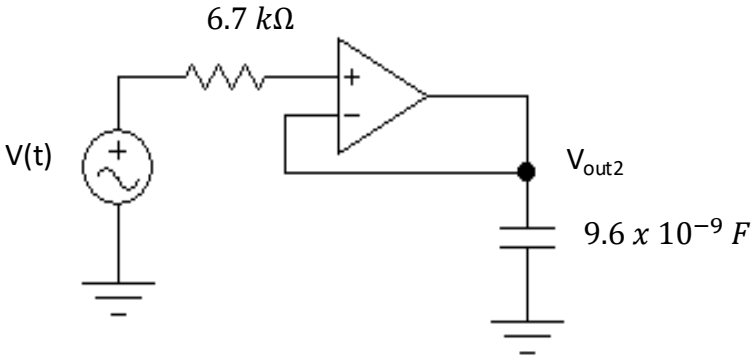
Note: I've simplified the schematics a bit. If you see ground symbols at different places in a schematic you should assume they are always connected by conductors of zero resistance. There is no need to explicitly draw a wire connecting them. Also, the power supply connections to $+V_{CC}$ and $-V_{CC}$ are assumed to be there.

Current limits

Op amps greatly simplify analog design but they are generally small signal devices and can typically output no more than about 20 mA before all the wonderful properties I've described no longer hold true. To drive a big motor you would need op amps working in conjunction with big transistors that can carry many amps.

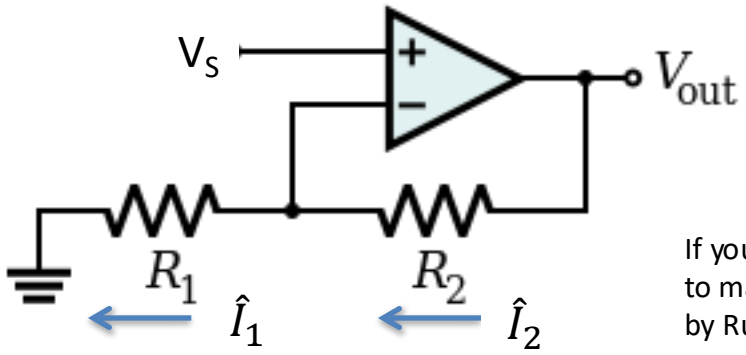


$$f_{3\,dB} = \frac{1}{2\pi RC} = 2.5\ kHz$$



Noninverting Amplifier

Followers are useful but suppose you want some gain. Then wire up the circuit shown below. V_S is the applied signal voltage and V_{out} is the output. Both are defined relative to ground potential. Using rule #1, $V_S = V_+ = V_-$. Therefore $I_1 = V_S/R_1$ by Ohm's law. But by rule #2, no current flows in or out of either input terminal so $I_1 = I_2$. Therefore,



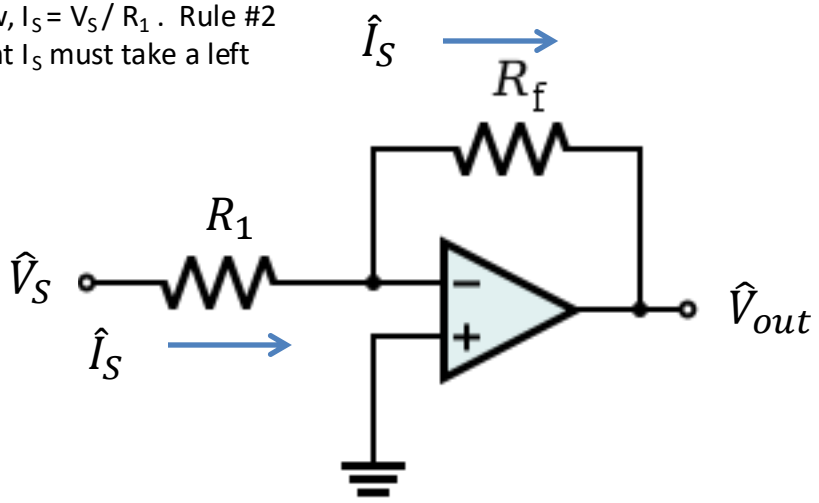
$$V_{out} = V_S + I_2 R_2 = V_S + I_1 R_2 = V_S + \frac{V_S}{R_1} R_2$$

$$A_V^{closed} = \frac{V_{out}}{V_S} = 1 + \frac{R_2}{R_1} = \frac{1}{\beta}$$

If you want a closed loop gain of exactly 153 then just choose two precision resistors to make it so. These can be made precise and stable with temperature. And again, by Rule #2, even if there were a source resistance R_S in series with V_S it wouldn't matter since no current would flow through it. The op amp multiplies the entire signal V_S by the closed loop gain.

Inverting amplifier

Since the (+) terminal of the op amp is grounded then $V_+ = 0$ and by rule #1 the other terminal must also be at zero volts. Therefore $V_- = 0$ and by Ohm's law, $I_S = V_S / R_1$. Rule #2 says that no current goes into the op amp input terminals so the current I_S must take a left turn and flow through the feedback resistor R_f. Therefore,

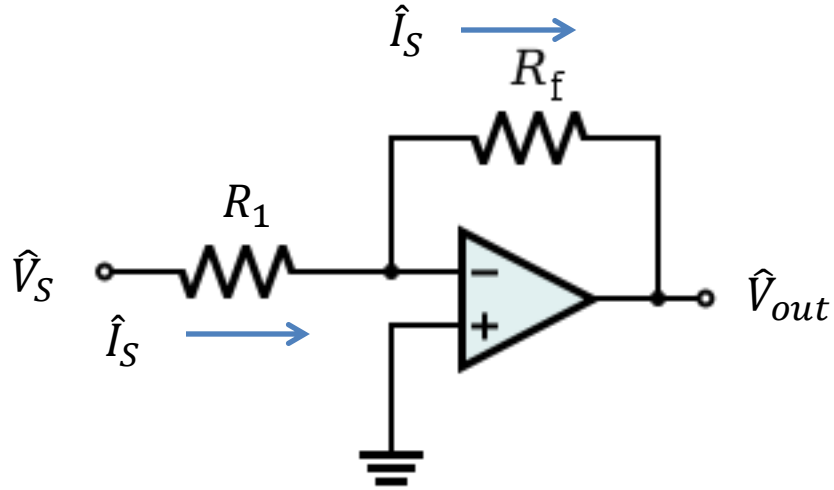


$$\hat{V}_{out} = \hat{V}_- - \hat{I}_S R_f = 0 - \frac{\hat{I}_S}{R_1} R_f = -\frac{\hat{I}_S}{R_1} R_f$$

The closed loop gain is therefore,

$$A_V^{closed} = \frac{\hat{V}_{out}}{\hat{V}_S} = -\frac{R_f}{R_1}$$

Virtual Ground



In this configuration the input resistance seen by the signal source V_S is *not* infinite but is,

$$\frac{\hat{V}_S}{\hat{I}_S} = R_1$$

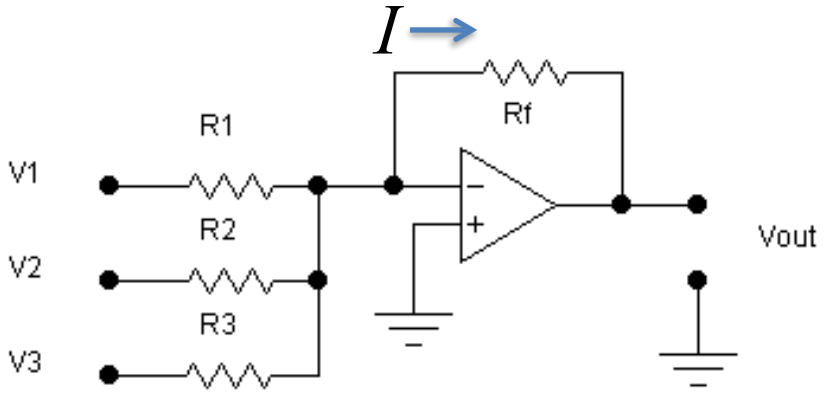
This configuration brings up another important point. Rule #1 says that feedback has forced the voltage between the (+) and (-) op amp terminals to be zero. Therefore if $V_+ = 0$ then $V_- = 0$. Both are at the same electrical potential but the (+) terminal is connected to the *physical* ground of the circuit.

Currents flow through the real conductors that are at ground potential. By contrast, the (-) terminal is at zero potential but it is not connected through a wire to the real ground. It's forced to zero voltage by feedback. Therefore it is said to be a *virtual ground*.

Adding voltages

Sometimes we want to add voltage from different sources. We can do that by adding more inputs to the inverting amplifer, as shown. No current flows to the inverting terminal so the total current through R_f is (by Kirchoff's current law) $I = I_1 + I_2 + I_3 = V_1/R_1 + V_2/R_2 + V_3/R_3$. By op amp rule #1 and Ohm's law, the output is given by,

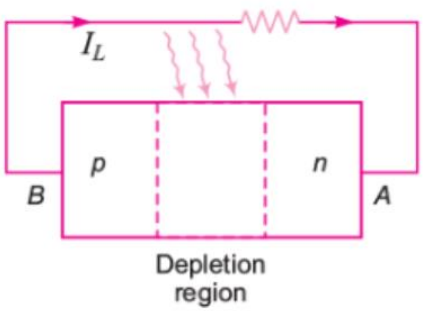
$$V_{out} = 0 - IR_f = -R_f\left(\frac{V_1}{R_1} + \frac{V_2}{R_2} + \frac{V_3}{R_3}\right)$$

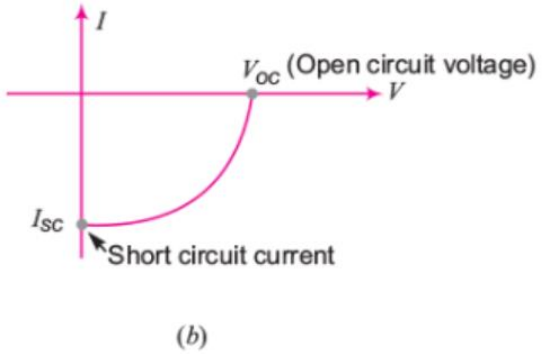


When used in this configuration the virtual ground point is also called the *summing point*.

Current to voltage converter

Usually the signal is treated as a voltage source but some transducers are better characterized as constant current sources. A photodiode is the best example. Incoming photons generate a photocurrent which flows in a direction to the normal forward-biased diode current. When the photodiode is short-circuited, all of the current flowing is photocurrent, labelled I_{SC} in the IV plot below.

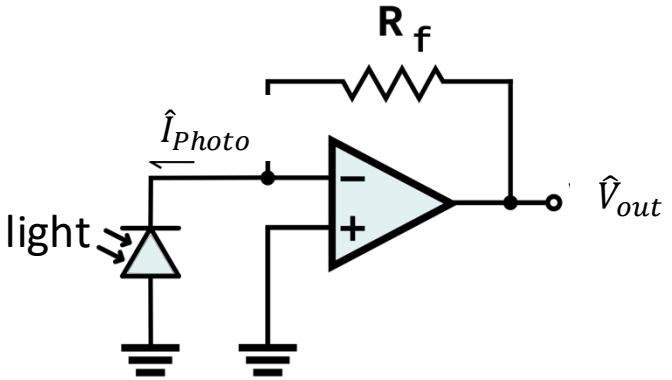




An op amp in the inverting configuration is often used to detect this photocurrent. By rule #1 the (-) terminal is a virtual ground so there is no voltage across the diode. By rule #2, all of the photocurrent flows through R_f, generating a voltage proportional to the photocurrent,

$$\hat{V}_{out} = R_f \hat{I}_{SC} = R_f \hat{I}_{Photo}$$

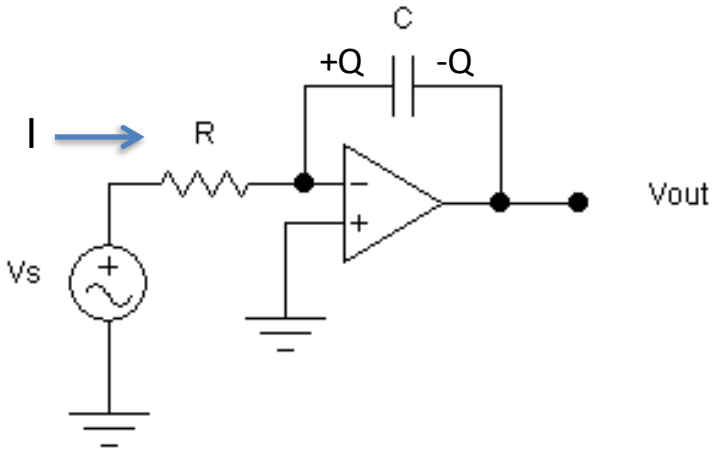
Amps which take a current and turn it into a voltage are often called *transimpedance amplifiers* or *current to voltage converters*.



Wikipedia, Transimpedance Amplifier.

Integrator

In this circuit the feedback element is a capacitor for which Q = CV and I = dQ/dt. It's better to look at the currents and voltages in the time domain. Using Ohm's law and the op amp rules we have,

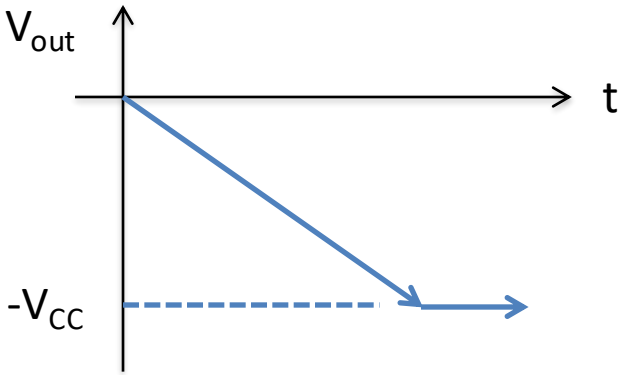


$$V_{out} = 0 - Q/C \qquad \frac{d\,V_{out}}{dt} = -\frac{I}{C} = -\frac{V_S}{RC}$$

$$V_{out}(t) = -\frac{1}{RC}\int_0^t V_S(t')\,dt' - \frac{Q(t=0)}{C}$$

If the capacitor is short-circuited at t = 0, then we can drop the second term. For example, if the input is a constant positive voltage then the output is a ramp with slope – V_S/RC.

$$V_{out}(t) = -\frac{1}{RC}\int_0^t V_S\;dt' = -\frac{V_S}{RC}\;t \qquad V_S = constant$$



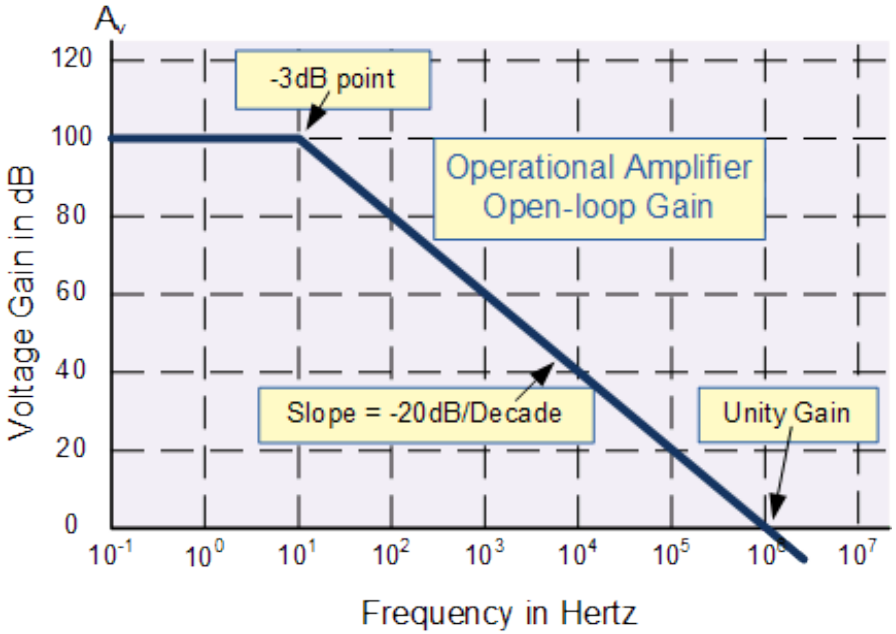
Starting at t = 0 the output would keep ramping until it hits the power supply voltage and can go no further. Integrators are widely used in filter circuits and in analog to digital conversion circuits.

Open loop gain and bandwidth.

All of this is wonderful but of course it can't go on forever. Op amps have many important limitations, the first of which is bandwidth. In order to keep circuits with op a, the amplifier is *designed* to have an open loop gain that has the frequency dependence of a low pass filter, but multiplied by the amplifier gain at DC:

$$A_V = \frac{V_{out}}{V_+ - V_-} = \frac{A_0}{1 + i\,\omega/\omega_0} = \frac{A_0}{1 + i\,f/f_0}$$

For a 411 op amp, $A_0 \approx 10^5$ and $f_0 \approx 10$ Hz. For some op amps, $A_0 > 10^6$ and f_0 might be < 2 Hz. The Bode plot shows the magnitude of A_V for a typical op amp. The low pass filter characteristic reduces the gain to 1 at about 1 MHz. This is is called the unity gain frequency or unity gain bandwidth f_T. You can now find op amps with f_T > 100 MHz. In exchange for higher speed they will suffer in some other area of performance.



It would seem that the very low frequency 3 dB point would severely limit op amp performance but the point is that when you begin (at DC) with lots of gain you can afford to throw some of it away at higher frequencies and still achieve near-ideal behavior.

Figure from electronics-tutorials.ws/opamp/opamp_1.html

Analog/Digital Conversion

Signals coming from microphones, temperature sensors, photodiodes are all voltages or currents that we wish to convert into numbers. That's the job of an analog to digital converter (ADC). Conversely, the computer algorithm to control an airplane or a robot must convert digital information into currents and voltages, which is the job of a digital to analog converter (DAC). We'll cover a few examples of ADC methods.

ADC and DAC symbols and parameters

The schematic symbols for ADCs and DACs are shown. Two important parameters are required to characterize them: speed and resolution.

Resolution

The resolution of an ADC refers to how precisely it can convert an analog voltage into a string of 1's and 0's. That's determined by the number of *bits* and the voltage *span*. For example, the Arduino Uno has an ADC with 10 bit resolution and a voltage span from 0-5 V. That means it will convert input voltages in the range from 0-5 V into 2^{10} possible values ranging from 0 to 2^{10}-1. The voltage *resolution* is therefore about 5 mV. The most precise ADCs today employ what are called *delta-sigma* methods. They have up to 24 bit resolution corresponding to 16,777,216 possible binary values. For a span of 5 V that's a resolution of 0.3 microvolts, which is rather amazing.

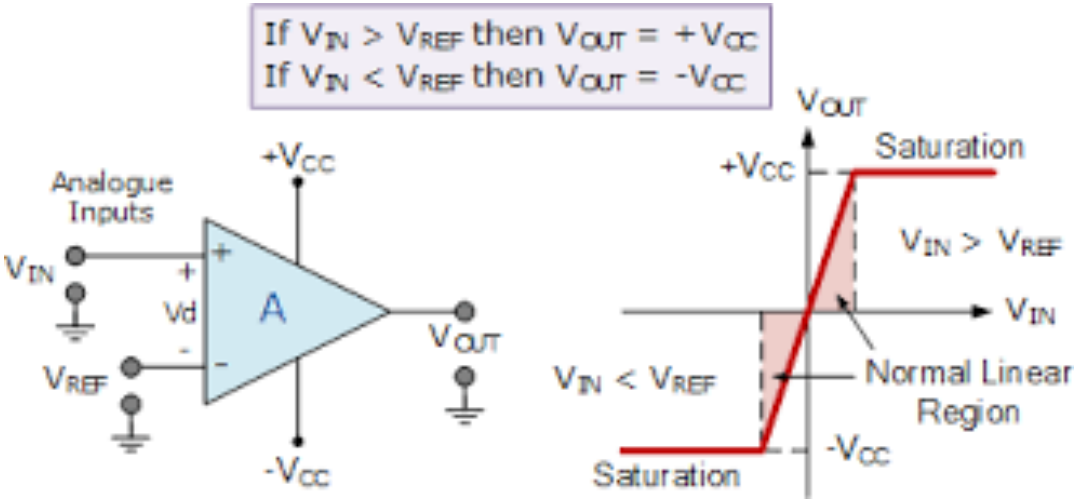
Speed

Speed refers to the time it takes the ADC to convert a voltage into a binary number. Generally the better the resolution the slower the ADC. The bench oscilloscopes in a typical lab might have 8 bit resolution but they can convert a voltage into a number in maybe 1 nsec. Analog devices now has a 12-bit ADC that can convert in a fraction of a nanosecond. A 24 bit ADC might have a conversion of time of 100 msec.

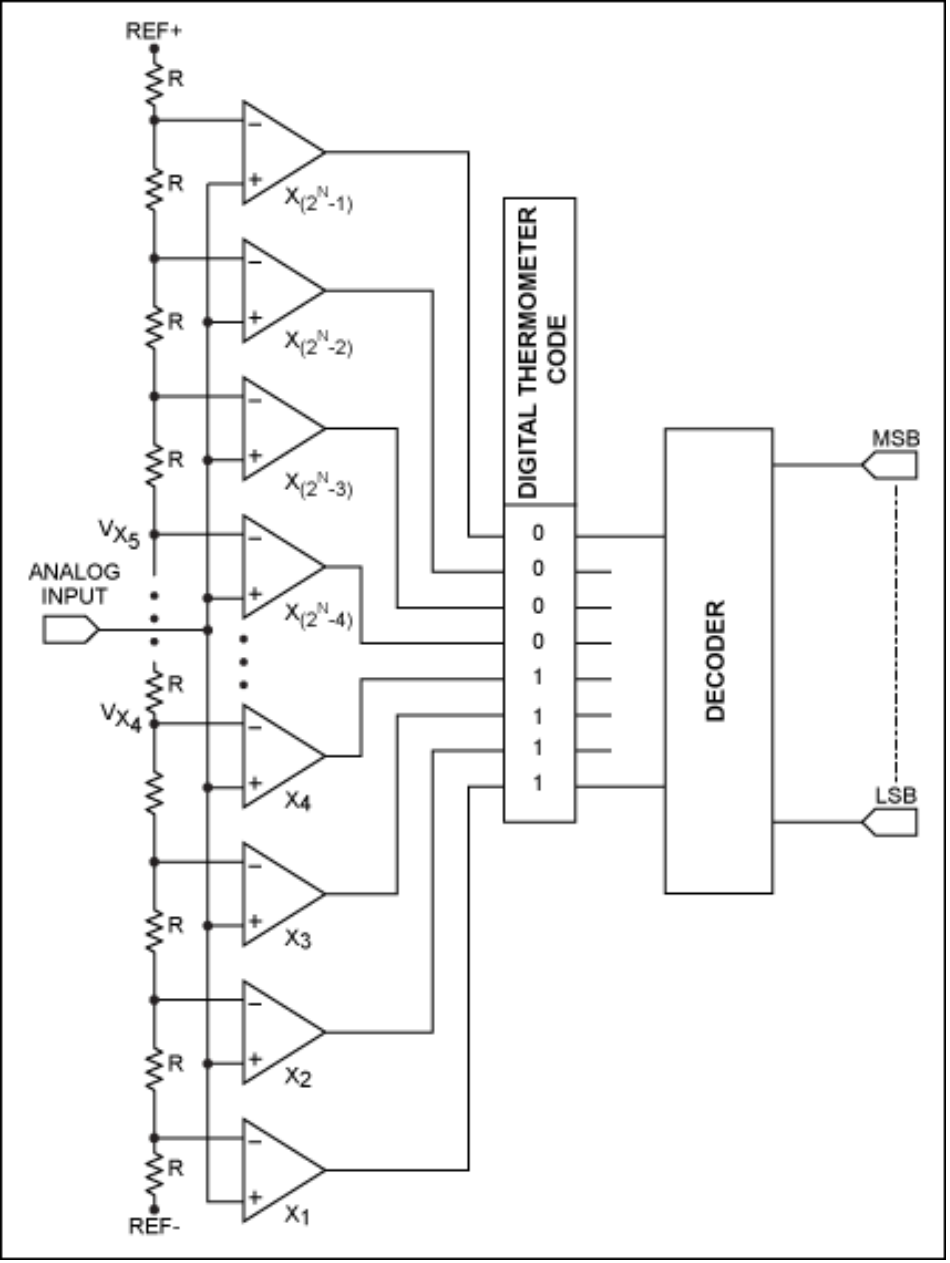
The same concepts apply to DACs. A 16-bit DAC (generally required for good audio quality) will produce 2^{16} possible outputs which might be voltages or currents. Speed refers to the time between successive outputs. The fastest DACs now output analog signals at GHz speeds with at least 8-bit resolution.

Comparator – 1 bit ADC

The simplest ADC is a comparator. It compares an input voltage Vin to a reference voltage V_{REF}. If $V_{IN} > V_{REF}$, the comparator output goes to logical HI and if $V_{IN} < V_{REF}$ the output goes to logical LO. You can do this with an op amp, as shown below. There is no feedback in this circuit. Since the open loop gain of the op amp is of order 100,000, if $I\, V_+ - V_- \,I$ exceeds 10^{-4} V the output will go to +/- V_{CC}. Defining logical HI as $+V_{CC}$ and logical LO as $-V_{CC}$ we have a 1 bit ADC meaning that the output can be either logical 0 or logical 1 depending on $V_{IN} - V_{REF}$.



https://www.electronics-tutorials.ws/opamp/op-amp-comparator.html



Flash ADC

Suppose you want better than 1 bit resolution. One way to do it is to build a voltage divider. The one on the left has N resistors in series. It divides the voltage between REF+ and REF- (which might be +1 V and -1 V) into 2^N-1 different equally spaced levels labelled Vx1, Vx2 , Vx3 etc. Now connect each of these points to the (-) terminal of a comparator. Connect the incoming analog voltage to the (+) terminal of all the comparators.

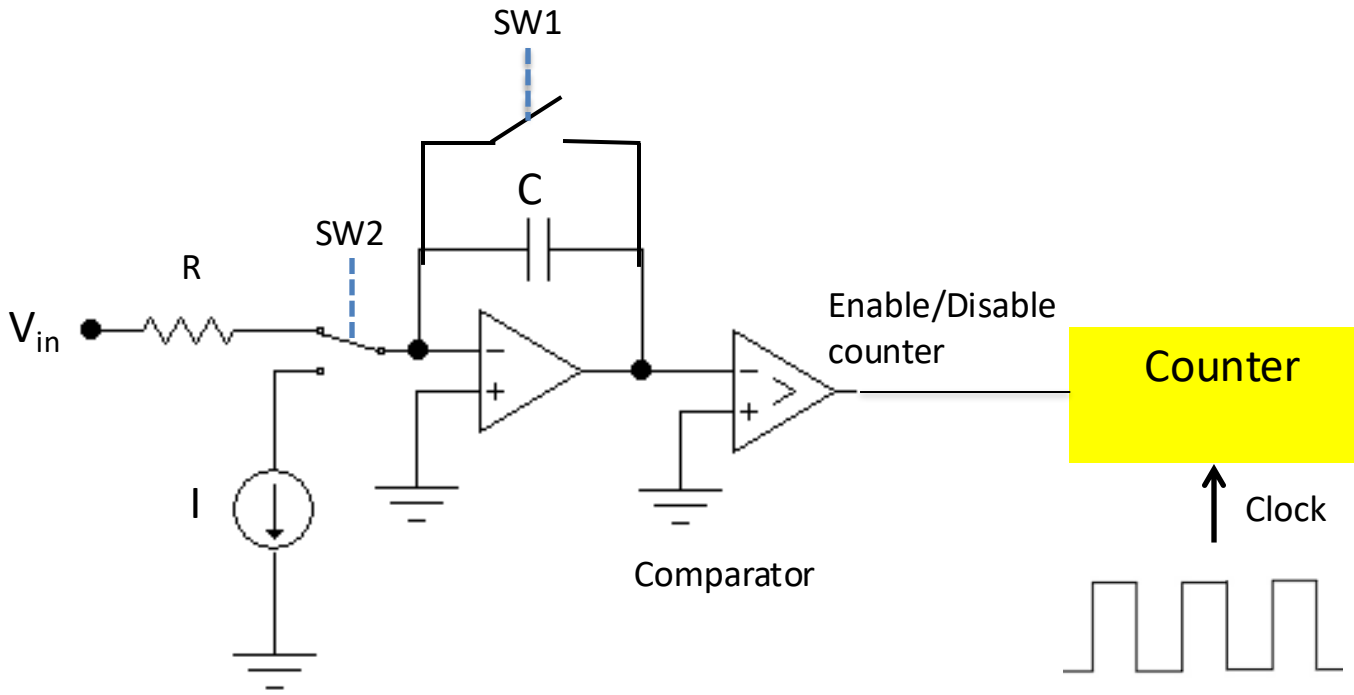
Suppose the analog input voltage is between Vx_4 and Vx_5. Then the input voltage exceeds the threshold for comparators X1, X2, X3 and X4 and they will all go to logic 1. All the other comparators will remain at logic 0.. It's called a "thermometer code" because the 1's fill up the memory register the same way an old mercury thermometer rises as the temperature increases. A *decoder* then converts the thermometer code into a binary number that represents the input voltage level.

Flash ADC's typically have the highest conversion rate but they don't have great resolution. They are often used in digital scopes. The resolution might be 8 bits (256 voltage levels) because it's difficult to make the voltage levels in the divider extremely stable and precise and the comparators are subject to errors and drifts.

https://www.maximintegrated.com/en/design/technical-documents/tutorials/8/810.html

Dual Slope Integration ADC

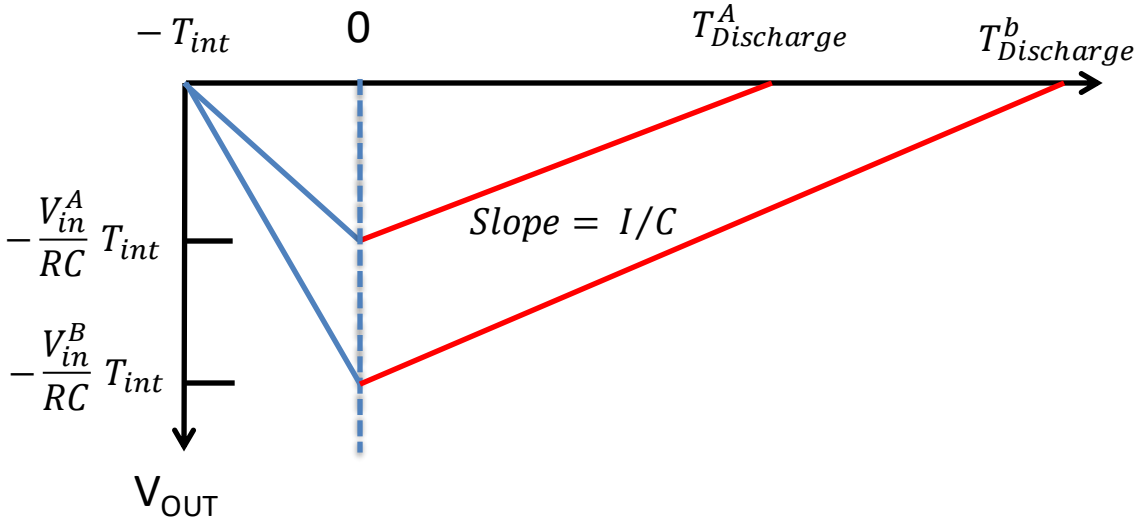
Dual slope integration is a much slower but more accurate kind of conversion technique. Variations of this are used in many existing bench digital voltmeters. SW1 and SW2 are MOSFET electronic switches. V_{IN} is the voltage to be converted. The basic idea is to integrate the input voltage (i.e. charge the capacitor) for a fixed time and then discharge the capacitor at a fixed rate. The time it takes to discharge, in clock cycles, is then a digital measure of V_{IN}. The conversion process is described below.



Initially close SW1 to short circuit the capacitor, setting $V_{OUT} = 0$. Now open SW1 and begin integrating for a time T_{int} as measured in clock cycles. The output of the integrator will be,

$$V_{out}(t) = -\frac{1}{RC}\int_{-T_{int}}^{0} V_{in}\, dt = -\frac{T_{int}}{RC}\, V_{in}$$

Now move SW2 to the lower position. Begin counting as the current sink *discharges* the capacitor. When V_{OUT} reaches 0, the comparator sends a logic signal to stop counting clock pulses. The figure shows what happens for two different input voltages, V_{in}^{A} and V_{in}^{B}. They charge up a *different* rates but discharge at the *same* rate, I/C, so they cross zero at different discharge times. The discharge time is measured by counting clock cycles. It's easy to see that,



$$\frac{T_{Discharge}}{T_{integ}} = \frac{V_{in}}{RI}$$

This scheme has several nice features.

1. It does not depend on the capacitor value. C doesn't need to be accurately known and it can even drift over long time periods without affecting the measurement.

2. V_{in} is proportional to the *ratio* of the discharge time to the integration time. Those times are measured in clock cycles so even if the length of the clock cycle (i.e. the period) drifts, that cancels out of the ratio.

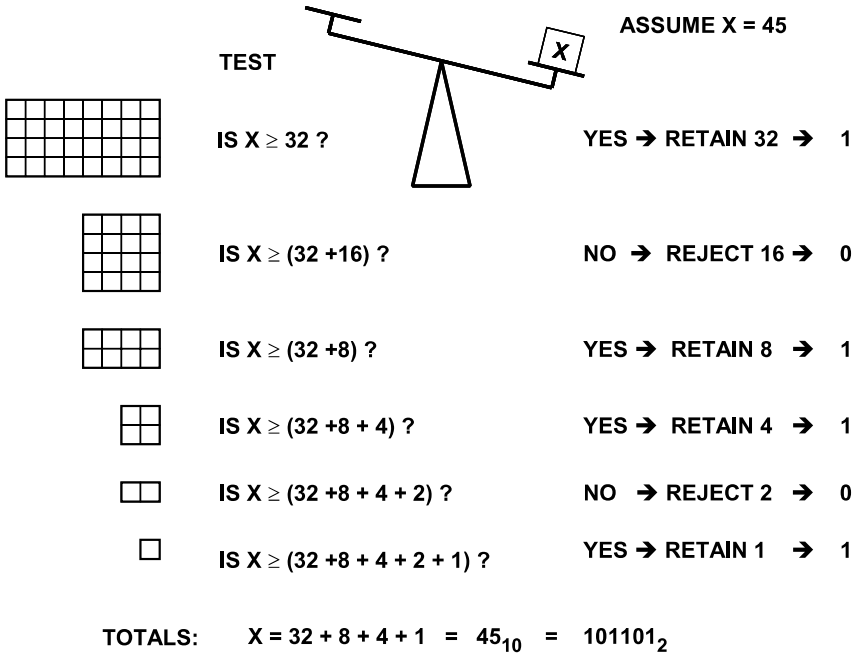
3.If there are interfering signals whose frequency is some multiple of $f_{Int} = 1/T_{Integ}$ they are *eliminated*. That's because,

$$V_{out}(t) = -\frac{1}{RC}\int_{-T_{int}}^{0} \cos\left(\frac{2\pi n}{T_{integ}}t\right)dt = 0 \qquad n = 1, 2, 3 \ \ldots$$

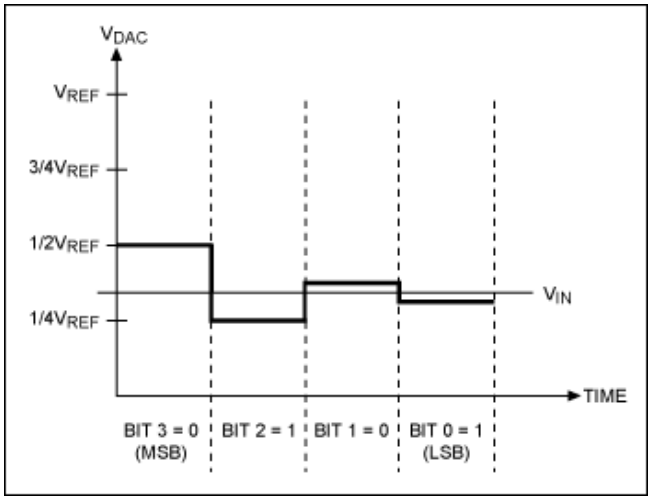
The most common interference comes from power lines with a frequency of 60 Hz. If $T_{integ} = 1/60$ sec, interfering signals at 60 Hz, 120 Hz, 180 Hz, etc., are eliminated. Many digital multimeters allow you to set the integration time.

Successive Approximation Register ADC

The successive approximation register (SAR) approach is very old. The basic idea first written down (to my knowledge) back in 1563. Suppose you're a farmer bringing potatoes to market and you need them weighed. Assume you have 45 lbs (or whatever they used for units in 1563) of potatoes. Put the basket on a balance and first try a counterweight of 32 lbs. If it tilts as shown then retain 32 and mark the most significant digit as 1. Next, add 16 lbs. This time the balance tilts the other way so the next significant digit is 0. You get the idea. The diagram shows a 6 bit conversion. This successive approximation process can also be done electrically as shown next.



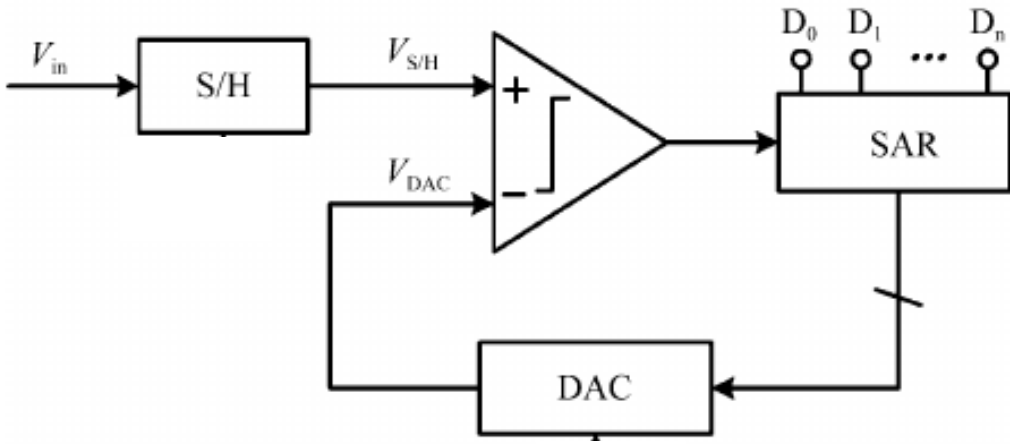
Successive approximation process, ca 1563



https://wiki.analog.com/university/courses/alm1k/alm-signals-labs/alm-sar-adc-1

The modern, electronic version is shown here. The incoming analog voltage is V_{in}. It's first compared to a voltage of $V_{ref}/2$. If $V_{in} < V_{ref}/2$ then the MSB is set to 0. Next, compare V_{in} to $V_{ref}/4$. This time $V_{in} > V_{ref}/4$ so we set the next significant bit to 1, etc.

The actual electronics is shown below. First, V_{in} is stored in a *sample and hold* (S/H) circuit. That's basically a capacitor with some op amps to hold the charge while the successive approximation process takes place. A digital to analog converter (DAC) now applies $V_{ref}/2$ to the (-) terminal of a comparator. In this example $V_{in} < V_{ref}/2$ so the comparator output goes to 0 and bit D0, the most significant bit, is set to 0 in the successive approximation register (SAR). Some logic circuitry next tells the DAC to output $V_{ref}/4$. When the comparator sees that $V_{in} > V_{ref}/4$ its output goes to 1 and this sets D1 to 1 in the SAR. After *N* such comparisons the number $D_0D_1...D_N$ gives the digital representation of V_{in}.

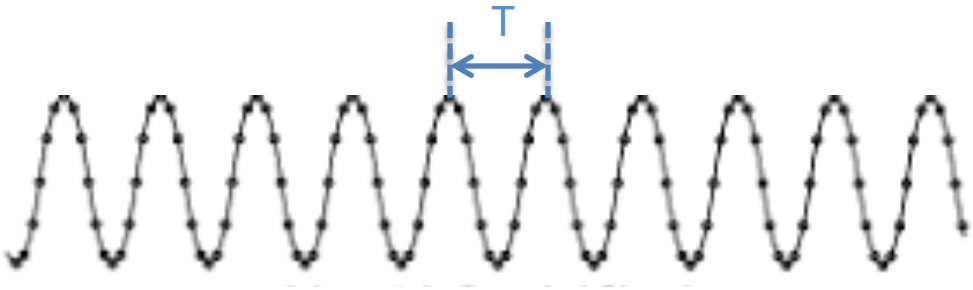


SAR converters are widely used for resolution up to 18 bits and moderate speeds. The ADC in an Arduino is an SAR type with 10 bit resolution that will perform a conversion in about 100 μsec. Analog Devices makes an 18 bit ADC (the AD7641) that does the conversion in more like $0.5\ \mu sec$.

The conversion process takes time so if V_{in} is, for example, a sinewave, the conversions must take place on a time scale shorter than the period of the sinewave. More on that next.

Sampling and aliasing

If you've ever been to a party with a strobe light going or seen an old western movie then you're familiar with *sampling* and *aliasing*. Sampling is what an ADC does. It generates a number based on samples of an analog voltage. This process takes some time which we'll call the *sampling time* t_S. During this time the circuit is converting the input into a binary number. The conversion process might take anywhere from milliseconds to less than a nanosecond. The idea is shown below, in which the sampling time t_S is the amount of time between the dots. Think of a successive approximation ADC in which the trial and error conversion process takes t_S seconds before we get a digital representation of the voltage.



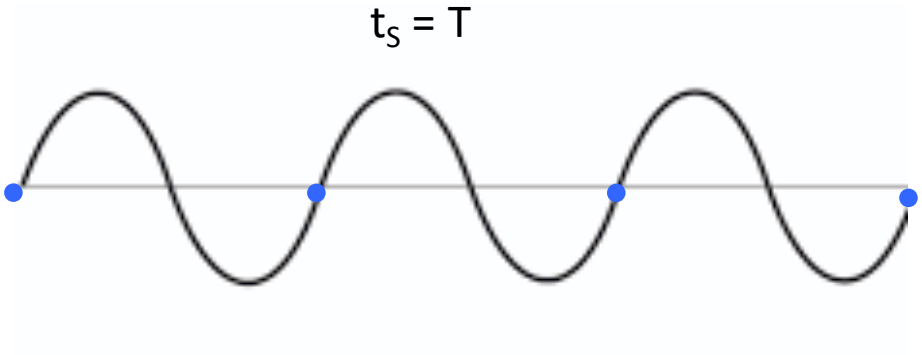
$t_S << T$

$t_S > T$

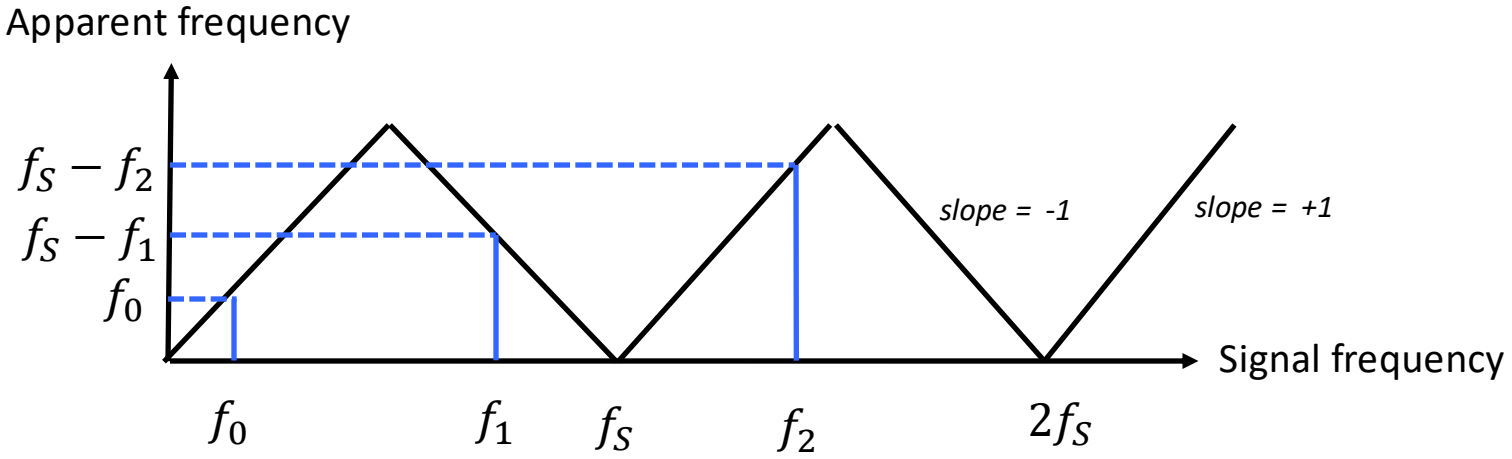
Consider digitizing a sine wave for different sampling times t_S compared to the period T of the waveform. In the top figure $t_S << T$. The ADC does a conversion rapidly compared to the period of the sine wave. If you were to now interpolate between the dots you'd have a good approximation to the original analog waveform.

In the bottom figure we've used a much slower ADC. Now $t_S > T$. Interpolating between the dots gives a sine wave with a period of 5T or a frequency 1/5 as large as the actual signal. This process is called *aliasing*. The slower, "apparent" waveform is an *alias* for the original waveform.

https://zone.ni.com/reference/en-XX/help/371361R-01/lvanlsconcepts/aliasing/



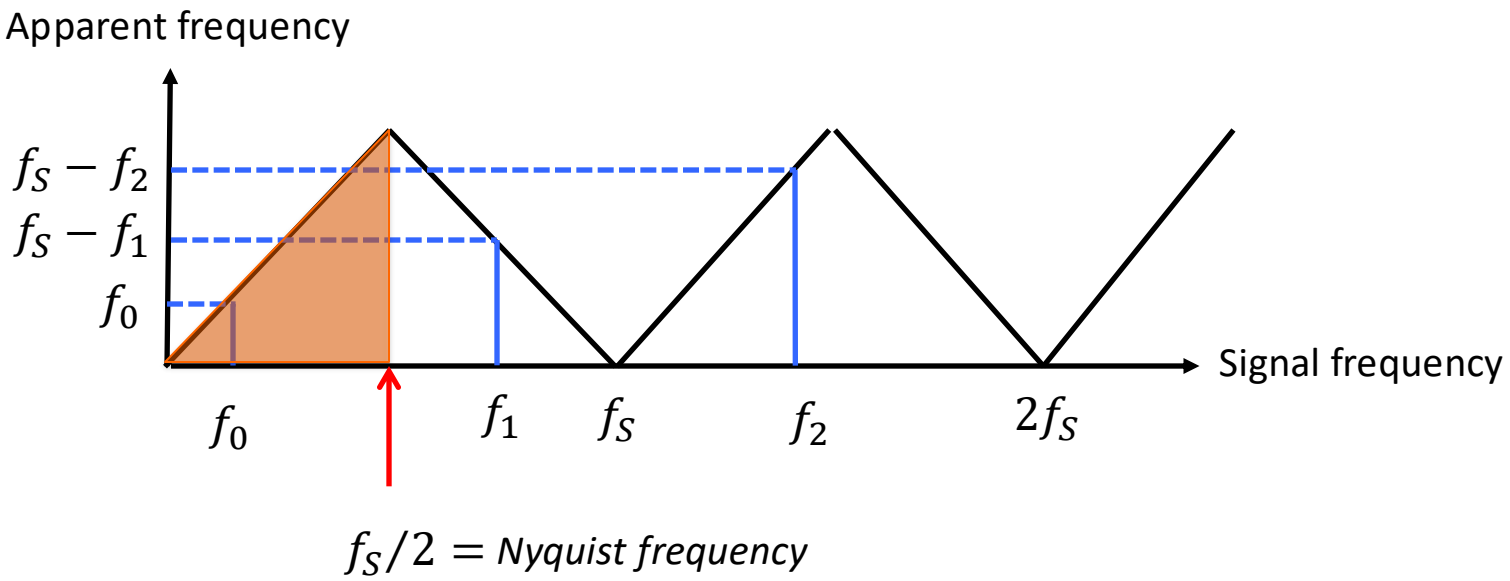
For an extreme example of aliasing, imagine that the sampling time is *equal* to the period of the sine wave, as shown on the left. Interpolating between the sampled data points (blue dots) gives a *constant* signal, i.e., the *apparent frequency* is zero! All of this this can be summed up by the plot shown below.



The triangle wave has slope +1 or -1 and $f_S = 1/t_S$ is the *sampling frequency*. Now look at 3 different incoming signals with frequencies f_0 , f_1 , f_2 respectively. Each signal is digitized by an ADC with with a sampling frequency f_S .

The signal at f_0 has an apparent frequency f_0 so is *not* aliased.
The signal at f_1 has an apparent frequency of $f_S - f_1$. It is aliased down to frequency $f_S - f_1$.
The signal at f_2 is aliased down to $f_2 - f_S$.

You can also see that a signal whose actual frequency is an integral multiple of the sampling frequency will be aliased down to an apparent frequency of 0.



The figure shows that if the actual signal frequency is *f* then its alias is at frequency ,

$$f_{apparent} = f_{Alias} = Minimum\,|nf_S - f| \quad n = 1, 2, 3 \ \ldots$$

In other words, find that multiple *n* of the sampling frequency that comes closest to the signal frequency. The expression above then gives the aliased frequency. Signals in the red shaded region have a frequency < $f_S/2$ so they are *not* aliased. All signals outside that zone *are* aliased. It is also said that they are *undersampled*. The dividing line at $f_S/2$ is called the *Nyquist frequency*, f_N, named after famed Bell Telephone Labs engineer Harry Nyquist. An arbitrary signal has many Fourier components so to avoid aliasing any of them, all signal frequencies must lie within the red shaded region. The general rule is,

To avoid aliasing, all signal frequencies must be less than the Nyquist frequency. Alternatively, the sampling frequency must be at least twice the highest frequency present in the incoming signal.

Wheel and Strobe Light

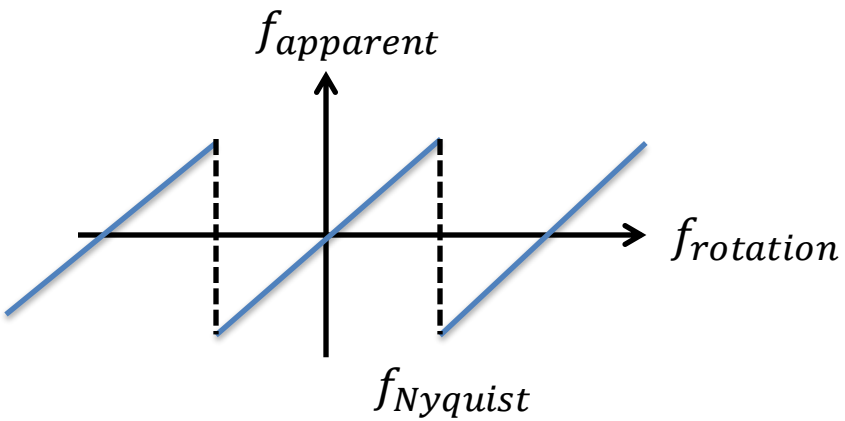
Just to remind you that you are familiar with aliasing, think about a wheel with a black mark on it that rotates clockwise by one revolution every T seconds. So the actual frequency is $f_{rotation}$ = 1/T. Suppose the mark lights up under infrared light.

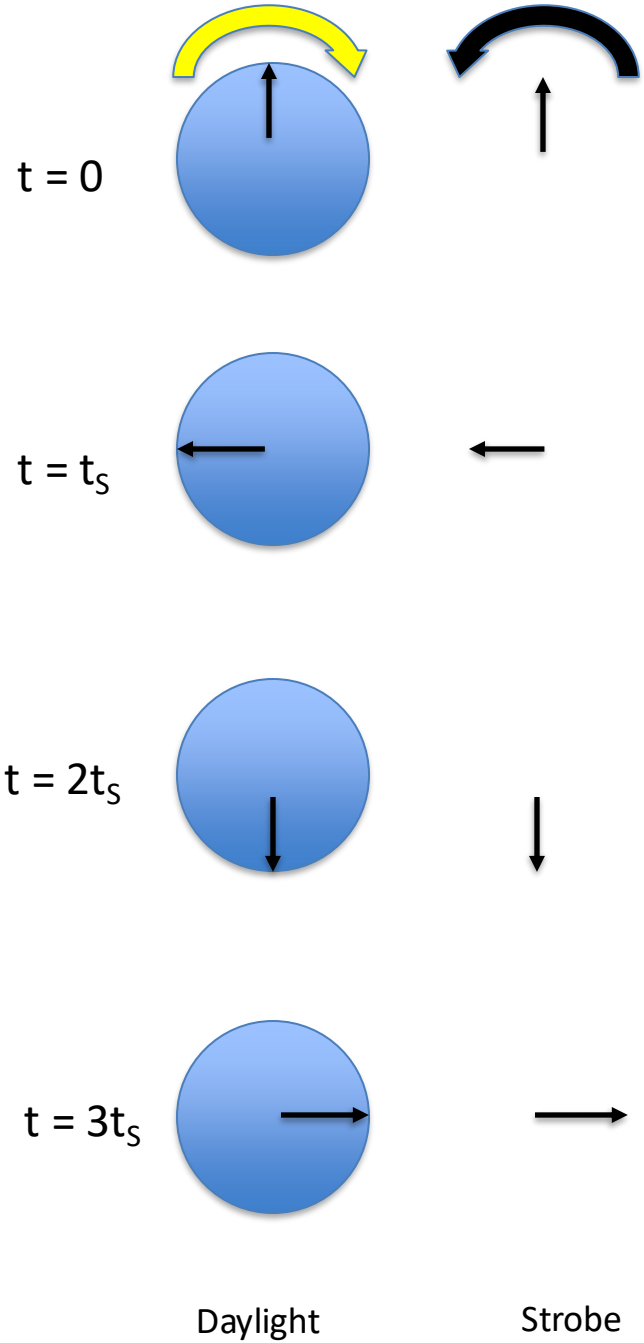
Now imagine that everything is dark and you turn on an infrared strobe light that flashes every (3/4)T seconds. So it *samples* the configuration of the wheel with t_S = (3/4)T.

If all you see is the mark light up when the strobe flashes, you would conclude that the wheel is rotating *counterclockwise* by ¼ turn every t_S seconds. The apparent frequency would be,

$$f_{apparent} = -\frac{f_{rotation}}{3}$$

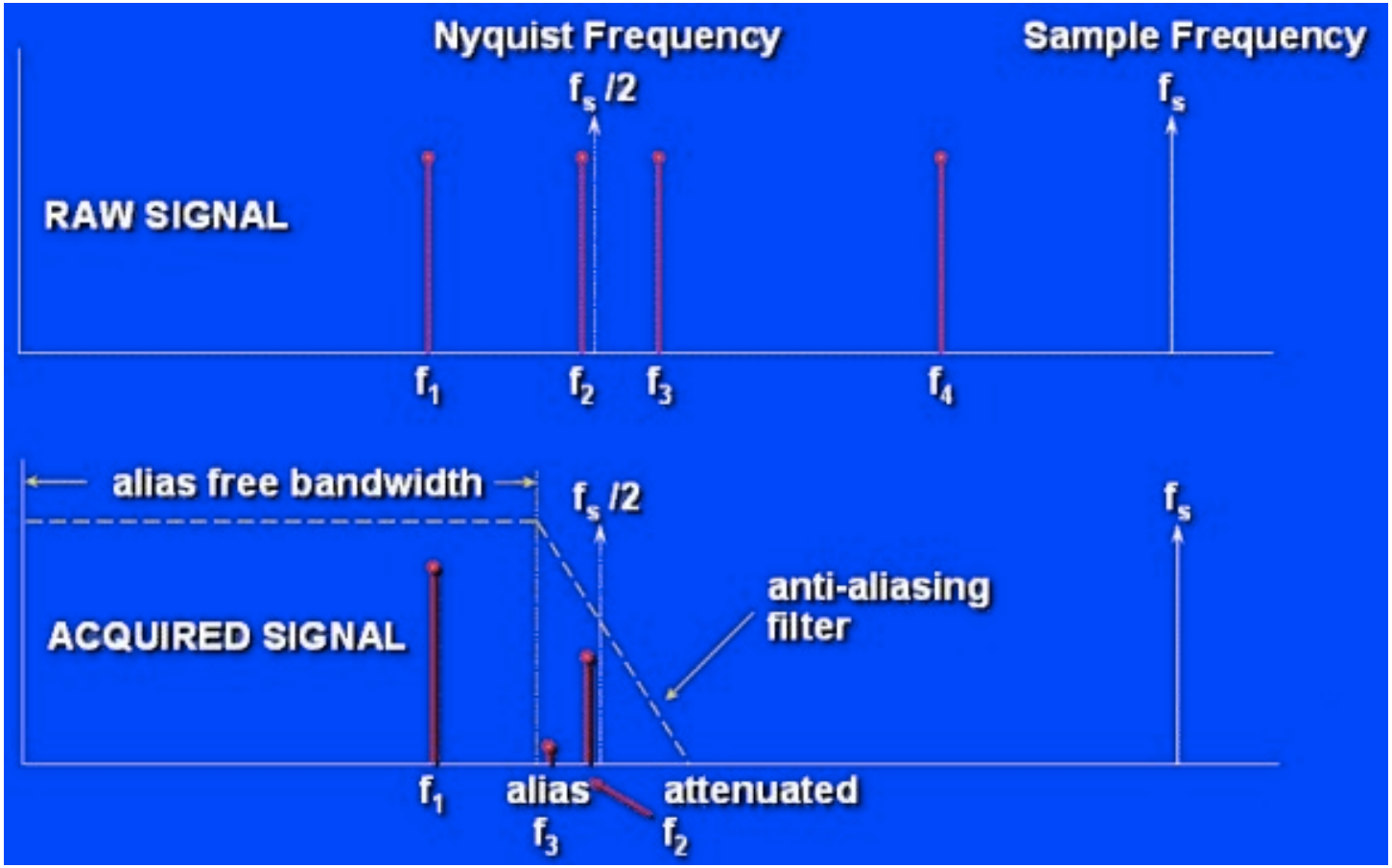
The actual frequency (1/T) is larger than the Nyquist frequency = 2/(3T) so the image under the strobe is an alias. For rotations we can distinguish between + and – frequencies so the diagram looks a little different:





Anti-aliasing filter

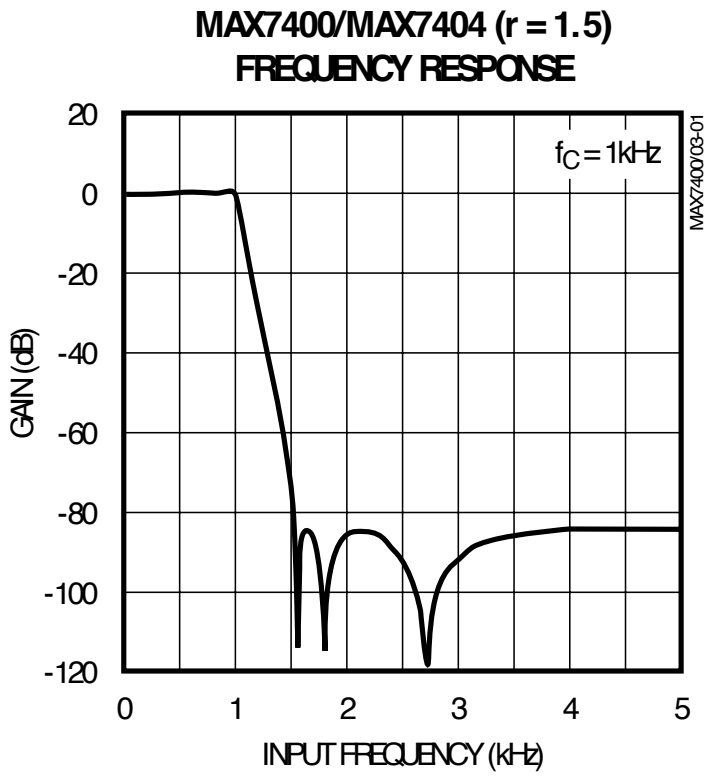
To avoid aliasing, we need to make sure that *no* signal with a frequency above the Nyquist frequency makes it through to the ADC. The figure illustrates the situation. The raw signals at f_3 and f_4 will both be aliased. To eliminate them, an *anti-aliasing filter* is placed ahead of the ADC. The *acquired signal* is what we get after the ADC. The filter *does* eliminate the signal at f_4 but it doesn't cut off sharply enough to eliminate f_3 entirely. Some of that signal gets through the filter, is folded about the Nyquist frequency and ends up in the acquired signal.



https://www.ni.com/en-us/innovations/white-papers/18/anti-aliasing-filters-and-their-usage-explained.html

A perfect anti-aliasing filter would completely eliminate the raw signals at f_3 and f_4. This is a good example of when it would be nice to have a brick wall filter with a cutoff at the Nyquist frequency. (The alternative, of course, is to get a faster ADC.) As we've seen, designing a brick wall filter requires many poles. You can make one using op amps but for many poles that gets clumsy and difficult to tune.

Fortunately there are chips for this purpose made by companies like Maxim Electronics and Analog Devices. For example, the Maxim model Max7403 is an 8-pole elliptic filter using switched capacitor technology. You can change the cut-off frequency from 1 Hz to 10 kHz by changing the clock frequency of the switched capacitors. Even the clock is built-in so the user just needs to change a capacitor to tune it. It runs off a 5 V supply.



https://datasheets.maximintegrated.com/en/ds/MAX7400-MAX7407.pdf

The 7403 response is shown for a 1 Hz cutoff frequency. The attenuation in the stop band is greater than 80 dB (i.e., an attenuation of 1/10,000). The cutoff frequency is set by the clock frequency:

$$f_{cutoff} = \frac{f_{Clock}}{100}$$

Notice that the attenuation in the stop band does not monotonically decrease. That's characteristic of an elliptic filter. It has the sharpest cutoff, per pole, of any design (Butterworth, Chebyshev, etc.) but the attenuation in the stop band is more flat as the frequency goes up. That's okay because the ADC itself has noise so if the anti-aliasing filter reduces the possible aliased signals to below the inherent noise of the ADC, there is no need for a better filer.

19-4764; Rev 2; 6/99

MAXIM

8th-Order, Lowpass, Elliptic, Switched-Capacitor Filters

General Description

The MAX7400/MAX7403/MAX7404/MAX7407 8th-order, lowpass, elliptic, switched-capacitor filters (SCFs) operate from a single +5V (MAX7400/MAX7403) or +3V (MAX7404/MAX7407) supply. These devices draw 2mA of supply current and allow corner frequencies from 1Hz to 10kHz, making them ideal for low-power anti-aliasing and post-DAC filtering applications. They feature a shutdown mode that reduces the supply current to 0.2µA.

Two clocking options are available: self-clocking (through the use of an external capacitor) or external clocking for tighter cutoff-frequency control. In addition, an offset adjustment pin (OS) allows for the adjustment of the DC output level.

The MAX7400/MAX7404 provide 82dB of stopband rejection and a sharp rolloff with a transition ratio of 1.5. The MAX7403/MAX7407 provide a sharper rolloff with a transition ratio of 1.2, while still delivering 60dB of stopband rejection. The fixed response of these devices simplifies the design task to corner-frequency selection by setting a clock frequency. The MAX7400/MAX7403/MAX7404/MAX7407 are available in 8-pin SO and DIP packages.

Applications

ADC Anti-Aliasing

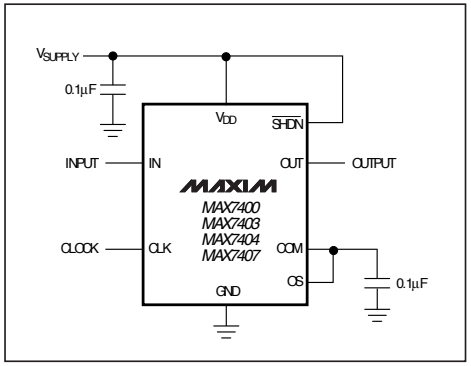
Post-DAC Filtering

CT2 Base Stations

Speech Processing

Air-Bag Electronics

Typical Operating Circuit



Features

- **8th-Order Lowpass Elliptic Filter**
- **Low Noise and Distortion**
  -82dB THD + Noise (MAX7400)
- **Clock-Tunable Corner Frequency (1Hz to 10kHz)**
- **100:1 Clock-to-Corner Ratio**
- **Single-Supply Operation**
  +5V (MAX7400/MAX7403)
  +3V (MAX7404/MAX7407)
- **Low Power**
  2mA (Operating Mode)
  0.2µA (Shutdown Mode)
- **Available in 8-Pin SO and DIP Packages**
- **Low Output Offset: ±5mV**

Ordering Information

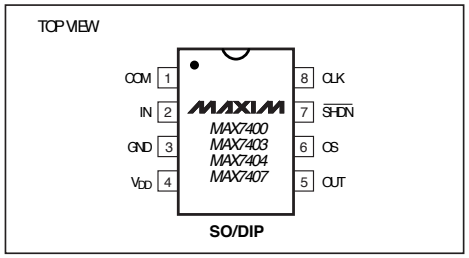
PART	TEMP. RANGE	PIN-PACKAGE
MAX7400CSA	0°C to +70°C	8 SO
MAX7400CPA	0°C to +70°C	8 Plastic DIP
MAX7400ESA	-40°C to +85°C	8 SO
MAX7400EPA	-40°C to +85°C	8 Plastic DIP

Ordering Information continued at end of data sheet.

Selector Guide

PART	FILTER RESPONSE	OPERATING VOLTAGE (V)
MAX7400	Elliptic (r = 1.5)	+5
MAX7403	Elliptic (r = 1.2)	+5
MAX7404	Elliptic (r = 1.5)	+3
MAX7407	Elliptic (r = 1.2)	+3

Pin Configuration



SO/DIP

Aliasing can be useful

It's possible to put aliasing to use and in fact it's done all the time. In the figure below, the aliased signal certainly does *not* reproduce the original sine wave. But maybe all you want is the amplitude of the original signal. In that case the aliased signal waveform will do just fine. You can then use much slower electronics to find its amplitude (and phase). This is a form of *frequency down conversion*. The aliasing process just replaces the original fast signal (typically radio frequency) by a lower frequency version that's easier to work with. We'll see the same thing later under the guise of *mixing*.

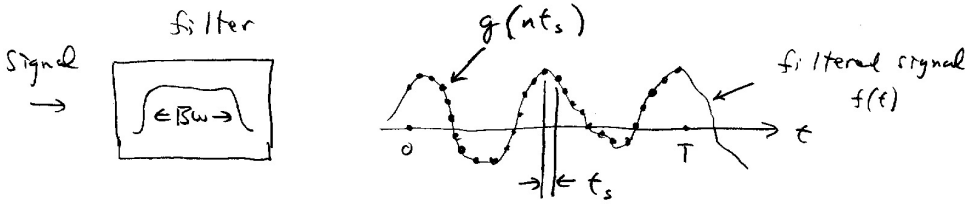
$t_S > T$

Sampling Theorem

From earlier remarks you might have the impression that after sampling a signal we need to somehow "connect the dots" to get some approximation of the original signal. That seems arbitrary but in fact, it was shown some time ago that if the function being sampled is *band-limited*, meaning it has no Fourier components beyond a certain limit, then there is a unique way to connect the dots. This is called the Sampling Theorem:

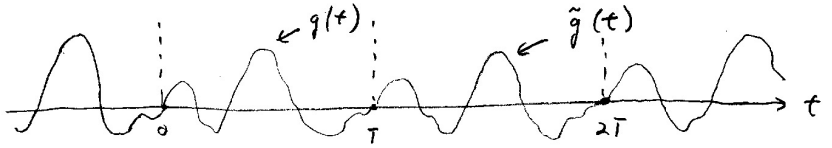
Take N samples of a continuous function F(t) . If F(t) has no Fourier components at frequencies above the Nyquist frequency, then F(t) can be completely reproduced using just the N samples. A handwaving explanation is given next.

The sampling theorem



So we sample over a time period $0 \le t \le T$ and take $N+1$ samples: $\{ g(0), g(t_s), \dots, g(Nt_s) \}$.

Next, imagine that we construct a function $\tilde{g}(t)$ such that $\tilde{g}(t) = g(t)$ for $0 \le t \le T$. Outside of this interval, $\tilde{f}(t)$ is just a periodic replica of $g(0 \le t \le T)$:

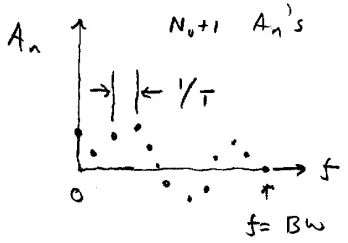


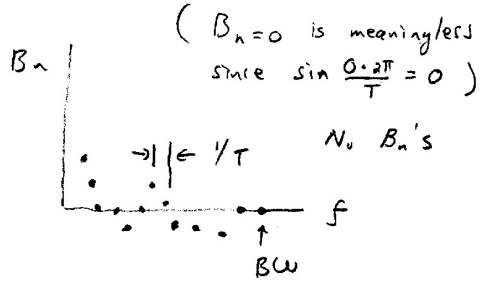
Since $\tilde{g}$ is <u>periodic</u> with period T, it can therefore be represented by a Fourier series with fundamental frequency $1/T$:

$$\tilde{g}(t) = \sum_{n=0}^{\infty} \left(A_n \cos \frac{2\pi}{T} nt + B_n \sin \frac{2\pi}{T} nt \right)$$

But remember that $g(t)$, and therefore $\tilde{g}(t)$, is <u>band-limited</u>. So the highest frequency the Fourier series may contain is $f_{max} = BW$. So we have

$1 + N_0 = 1 + \frac{BW}{1/T} = 1 + T \cdot BW$, A_n coefficients (corresponding to frequencies $\{0, \frac{1}{T}, \frac{2}{T}, \ldots, BW\}$ and we have $N_0 = \frac{BW}{1/T} = T \cdot BW$, B_n coefficients $\{\frac{1}{T}, \frac{2}{T}, \ldots BW\}$.





These Fourier coefficients completely determine $\hat{g}(t)$ for all times and therefore $g(t)$ for $0 \le t \le T$. How many coefficients? We have $2N_0 + 1 = 1 + 2T \cdot BW$. To find these $\{A_n, B_n\}$ we use the $N+1$ samples of $g(t)$: $\{g(0), g(t_s), \cdots, g(Nt_s)\}$:

$$g(0) = \sum_{n=0}^{N_0} A_n$$

$$g(t_s) = \sum_{n=0}^{N_0} \left(A_n \cos \frac{2\pi}{T} n \cdot t_s + B_n \sin \frac{2\pi}{T} n \cdot t_s \right)$$

$$g(2t_s) = \sum_{n=0}^{N_0} \left(A_n \cos \frac{2\pi}{T} n \cdot 2t_s + B_n \sin \frac{2\pi}{T} n \cdot 2t_s \right)$$

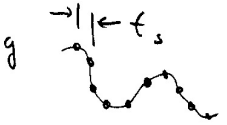
$$\vdots \qquad\qquad \vdots$$

$$g(Nt_s) = \sum_{n=0}^{N_0} \left(A_n \cos \frac{2\pi}{T} n \cdot Nt_s + B_n \sin \frac{2\pi}{T} n \cdot Nt_s \right)$$

For this to work we need as many sampled data points $\{g(0), g(t_s), \ldots, g(Nt_s)\}$ ($N+1$ of them) as we have Fourier coefficients, $2N_0+1 = 2 \cdot T \cdot BW + 1$

So we need $\boxed{N+1 = 2T \cdot BW + 1}$ sampled data points.

Now recall that we sample every t_s seconds over an interval $0 \le t \le T$ so $N = T/t_s$. So



$$N+1 = \frac{T}{t_s} + 1 = T \cdot (2BW) + 1 \quad \text{or}$$

$$\boxed{\frac{1}{t_s} = 2 \cdot BW = f_s = 2 f_{Nyquist}}$$

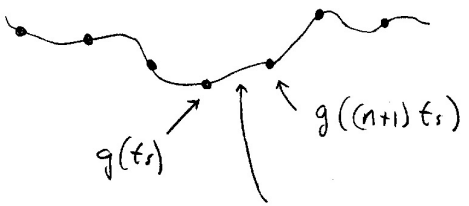
So, as expected, the Nyquist frequency (set by the ADC) must be equal to (or greater than) the highest frequency in the filtered signal. That is, $\boxed{BW \le f_{Nyquist}}$

If this is satisfied then we can find $\{A_n, B_n\}$, plug back into the Fourier and rewrite it to get,

$$g(t) \underset{0 \le t \le T}{=} \sum_{n=0}^{N_0 = BW \cdot T} g(nt_s) \frac{\sin\left[\frac{\pi(t - nt_s)}{t_s}\right]}{\left[\frac{\pi(t - nt_s)}{t_s}\right]}$$

"Sampling Theorem"

Sampling theorem tells us that if $BW < f_{Nyquist}$, we don't need to sample any faster than $2 \cdot f_{Nyquist}$ in order to reconstruct the signal at <u>all</u> times, <u>not</u> just at $t = nt_s$. It's an <u>interpolation formula</u> for the function between the sampled times.



Sampling theorem gives us these intermediate values of $g(t)$.

If we <u>undersample</u> ($BW > f_{Nyquist}$) then the reconstructed function will have <u>aliases</u>.