SOPHOMORE COLLEGE
MATHEMATICS OF THE INFORMATION AGE
SECRETS OF THE UNIVERSE, PART 2
FOURIER TRANSFORMS

Let's remember the banner headlines associated with the formula

$$f(t) = \sum_{n=-\infty}^{\infty} \hat{f}(n) e^{2\pi n i t}.$$

Point by point, you should understand:

- We're dealing with a periodic function of period 1 that is square integrable. We pass from the time domain to the frequency domain via the formula for the Fourier coefficients:
$$\hat{f}(n) = \int_0^1 e^{-2\pi i n t} f(t)\, dt.$$
- Thus to the time domain form of the function (*i.e.* to $f(t)$) we have associated spectral information – the harmonics, $e^{2\pi i n t}$, and the strengths, $\hat{f}(n)$, by which they contribute to $f(t)$.
  - The spectrum of a given function is the set of frequencies that are present, *i.e.* those values of n for which $\hat{f}(n) \neq 0$. Not all (possible) harmonics need be present in the spectrum of a given function, since $\hat{f}(n)$ can be zero for some values of n. The square wave we worked with, for example, has only odd harmonics.
  - The spectrum is a *discrete* set, though possibly infinite.
- From the frequency domain description we can reconstruct the time domain description via
$$f(t) = \sum_{n=-\infty}^{\infty} \hat{f}(n) e^{2\pi n i t}.$$
  - This equality is understood in terms of mean square convergence, or convergence in $L^2([0,1])$.[1]

What if it isn't periodic? Most functions *aren't* periodic. A *sustained* note repeats in time, but how helpful is it to regard rapidly changing music, or speech, as a periodic phenomenon? Not very helpful. For the purposes of applying what we've learned, however, let's regard a function which is not periodic as having 'infinite period' and see if we can do something. More precisely, we look at the limiting case of a function which is periodic of period p as $p \to \infty$.

What do the formulas for Fourier coefficients and the Fourier series look like when the period is p, for any p? We can figure this out from what we know about period 1, by scaling. Suppose $f(t)$ has period p and let

$$g(t) = f(pt)\,.$$

[1] If the function is continuously differentiable, for example, then one can prove other, more traditional results on convergence.

Then

$$\begin{aligned} g(t+1) &= f(p(t+1)) \\ &= f(pt+p) \\ &= f(pt) \quad \text{(since } f(t) \text{ is periodic of period } p) \\ &= g(t) \end{aligned}$$

Thus $g(t)$ is periodic of period 1. Now expand $g(t)$ is a Fourier series

$$f(pt) = g(t) = \sum_n \hat{g}(n) e^{2\pi int}.$$

Write $u = pt$ or $t = u/p$, to get

$$f(u) = \sum_n \hat{g}(n) e^{2\pi inu/p}.$$

The harmonics are now the complex exponentials $e^{2\pi int/p}$ and the frequencies are the points n/p. Notice that the frequencies are spaced $1/p$ apart, *i.e.*, if we looked at a spectrum analyzer plot of signal we'd see the bars drawn at the points $1/p$, $2/p$, $3/p$, $\ldots$, rather than at 1, 2, 3,$\ldots$. The larger p is the closer the spacing of the points in the spectrum.

What about the Fourier coefficients? We have

$$\begin{aligned} \hat{g}(n) &= \int_0^1 e^{-2\pi int} g(t)\, dt \\ &= \int_0^1 e^{-2\pi int} f(pt)\, dt \\ &\qquad \text{(now substitute } u = pt \text{ to write the integral as)} \\ &= \int_0^p e^{-2\pi iinu/p} f(u)\, \frac{1}{p} du \end{aligned}$$

Changing the name of the variable back to t (purely for psychological comfort), the Fourier coefficients of $f(t)$ is

$$\hat{f}(n) = \frac{1}{p} \int_0^p e^{-2\pi int/p} f(t)\, dt.$$

For what's to come, let's write this expression as an integral from $-p/2$ to $p/2$:

$$\hat{f}(n) = \frac{1}{p} \int_{-p/2}^{p/2} e^{-2\pi int/p} f(t)\, dt.$$

What happens as $p \to \infty$? With a certain amount of bravado and disregard for mathematical politeness we can answer this. Write

$$\frac{1}{p} = \Delta\xi$$

and the points n/p as ξ_n, that is,

$$\frac{n}{p} = n\Delta\xi = \xi_n.$$

Then the equation for $f(t)$ becomes

$$f(t) = \sum_{n=-\infty}^{\infty} \frac{1}{p} \left\{ \int_{-p/2}^{p/2} e^{-2\pi i n t/p} f(s)\, ds \right\} e^{2\pi i n t/p} = \sum_{n=-\infty}^{\infty} \left\{ \int_{-p/2}^{p/2} e^{-2\pi i \xi_n s} f(s)\, ds \right\} e^{2\pi i \xi_n t} \Delta\xi.$$

Looks like a Riemann sum. Looks like as $p \to \infty$ that expression becomes

$$f(t) = \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} e^{-2\pi i s \xi} f(s)\, ds \right\} e^{2\pi i t \xi}\, d\xi.$$

The Fourier transform

Let's turn this chicanery into a *definition.* For a function $f(t)$ (almost any function) we define its *Fourier transform* to be

$$\hat{f}(s) = \int_{-\infty}^{\infty} e^{-2\pi i s t} f(t)\, dt.$$

The integration is with respect to t but the complex exponential involves both the variables t and s. After integation what remains is a function of s. The formula looks similar that for the Fourier coefficient, but unlike the discrete Fourier coefficients $\hat{f}(n)$ for a periodic function, the Fourier transform of a general (nonperiodic) function is now a (transformed) function of a *continuous* real variable s. We still think of s as a 'frequency' variable, and the Fourier transform $\hat{f}(s)$ as the 'frequency domain' representation of the function, but instead of discrete frequencies as in the periodic case we have a continuum of frequencies in the nonperiodic case. The *spectrum* of $f(t)$ is the set of real numbers s where $\hat{f}(s) \neq 0$.

There are various notations for the Fourier transform in day-to-day use. One sees the transform written as

$$\mathcal{F}f(s) = \int_{-\infty}^{\infty} e^{-2\pi i s t} f(t)\, dt.$$

It's also common practice for people to use a lower case letter for the signal, like $f(t)$, and the corresponding upper case letter for the transform, like $F(s)$.

As our derivation, above, indicates, we recover the time domain version of the function by summing – or, in this case by integrating – the spectral information; the 'strength' $\hat{f}(s)$ (which is generally complex valued) times the corresponding 'harmonic', $e^{2\pi i s t}$;

$$f(t) = \int_{-\infty}^{\infty} e^{2\pi i s t} \hat{f}(s)\, ds.$$

This is known as the *Fourier Inversion Theorem.*

Recall again our rallying cry:

Every signal has a spectrum. A signal is determined by its spectrum.

Give an engineer a signal and he – or she – will take the Fourier transform. 'Tell me about the spectrum' she'll say. Guaranteed.

Mathematically, the statements translate to: 'Every function has a Fourier transform,' and 'Fourier inversion holds.' How literally are we to take these statements? Pretty literally.

However, the mathematical issues surrounding their validity are not to be taken lightly. Think about what's mixed together. The formula for the Fourier transform,

$$\hat{f}(s) = \int_{-\infty}^{\infty} e^{-2\pi i s t} f(t)\, dt,$$

involves an improper integral from $-\infty$ to ∞, which already calls for special consideration, and multiplying $f(t)$ by the complex exponential $e^{-2\pi i s t}$ makes the integrand oscillate, which makes questions of convergence of the integral that much more delicate. Furthermore, applying the Fourier inversion formula

$$f(t) = \int_{-\infty}^{\infty} e^{2\pi i s t} \hat{f}(s)\, ds$$

involves knowing first that $e^{2\pi i s t}\hat{f}(s)$ *can* be integrated *and* that the integration leads back to $f(t)$.

As a cautionary example, the trig functions sine and cosine – the building blocks of periodic phenomena! – do *not* have simple Fourier transforms. More general theories of integration notwithstanding, there is no way to make sense of

$$\int_{-\infty}^{\infty} e^{-2\pi i s t} \cos 2\pi t\, dt \quad \text{or} \quad \int_{-\infty}^{\infty} e^{-2\pi i s t} \sin 2\pi t\, dt$$

as any kind of classical functions.

Plainly there's a lot going on. Why then the certainty in employing these Secrets of the Universe – and believe me, they're fully employed. Several reasons. First, by now there is a quite sophisticated mathematical understanding of the Fourier transform and Fourier inversion, including definitions that go beyond the formulas via integrals. This culminates with the theory of 'distributions,' which is the rigorous treatment of 'delta functions' and the like.[2] Second, nature seems to take care of the existence of the integrals quite nicely, thank you, not to mention Fourier inversion. Recall that your inner ear finds the frequency content of a given signal (signals which generally are not periodic) and that your brain must then take the inverse transform to get the signal back. In the field of optics one discovers that lenses and prisms have the effect of taking a Fourier transform. Finally, for modern, computational uses of Fourier analysis, which are ubiquitous, one works with discrete data not continuous signals. It's a discrete form of the Fourier transform and the inverse transform that are used, with finite sums replacing integrals. There are no questions of convergence or existence for such finite sums.[3] There are other questions, but not the same questions.

[2] Though delta functions are often associated with P. Dirac and his treatment of quantum mechanics, their operational use was realized much earlier by O. Heaviside in his elucidation of Maxwell's theory of electromagnetism.

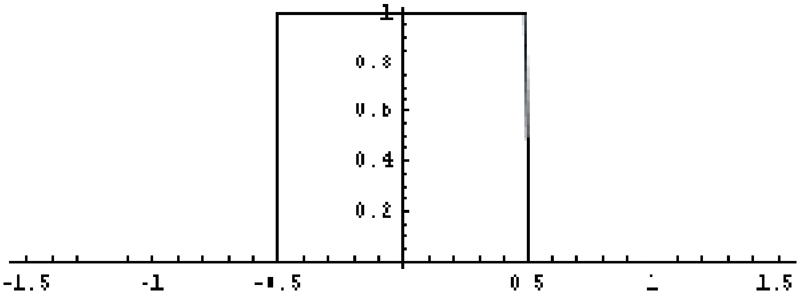
[3] The discrete form of the Fourier transform is essentially a Riemann sum approximation to the integral, or at least it can be developed from that point of view. This, the DFT (discrete Fourier transform), is the fundamental operator in applications. It is *not* to be confused with the FFT (Fast Fourier Transform). The FFT is an *algorithm* for the efficient calculation of the DFT. Keep this straight. It's easy to detect when somebody doesn't know what they're talking about when they casually refer to the FFT when they mean, or should mean the DFT.

An example. Before doing anything else grand and sweeping, let's do something modest and confined. Let's compute an example.

Take

$$\Pi(t) = \begin{cases} 1, & -\frac{1}{2} \leq t \leq \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$$

This is called, variously, the 'top hat' function, the 'rect' function (for rectangle), the 'indicator' function for the interval $[-1/2, 1/2]$, or the 'characteristic' function for the interval $[-1/2, 1/2]$. One can debate whether to define the function at the endpoints as I have. I refuse to be drawn into the debate – for our purposes it will never matter. The graph is:



The function is zero outside of $[-1/2, 1/2]$ so the limits on the integral defining $\widehat{\Pi}$ only go from $-1/2$ to $1/2$. The integration is straightforward:

$$\begin{aligned} \widehat{\Pi}(s) &= \int_{-\infty}^{\infty} \Pi(x) e^{-2\pi i s x}\, dx = \int_{1/2}^{1/2} e^{-2\pi i s x}\, dx \\ &= \frac{1}{-2\pi i s} e^{-2\pi i s x} \Big]_{-1/2}^{1/2} \\ &= \frac{1}{-2\pi i s}(e^{-\pi i s} - e^{\pi i s}) \\ &= \frac{1}{2\pi i s} 2i \sin \pi s \\ &= \frac{\sin \pi s}{\pi s} \end{aligned}$$

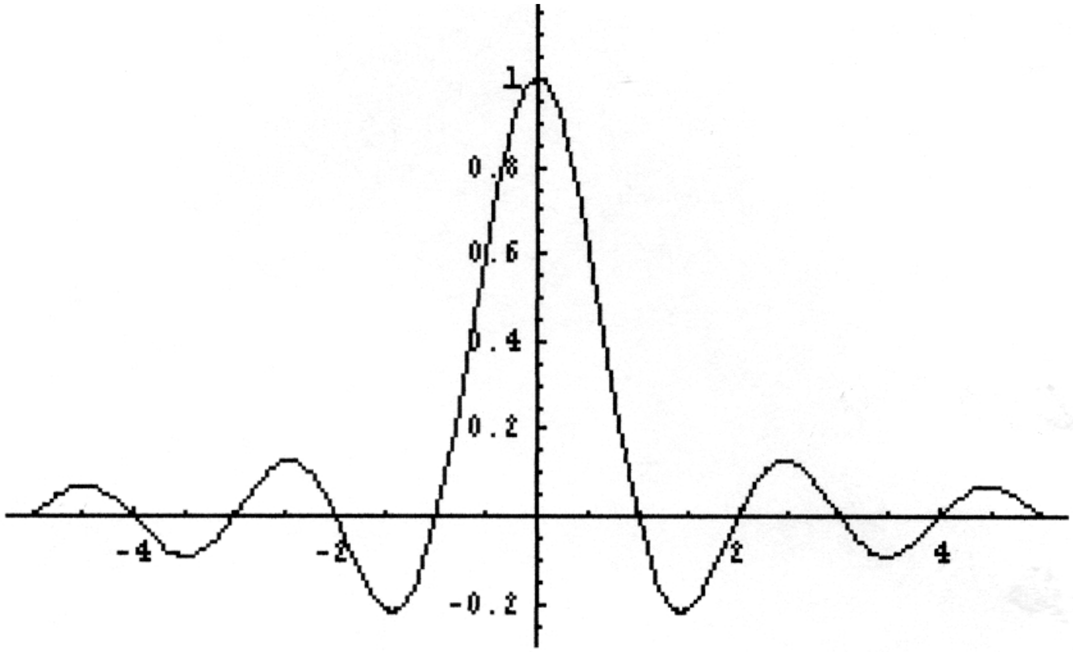
Note the special values

$$\operatorname{sinc} s = \begin{cases} 1, & s = 0, \\ 0, & s \text{ an integer other than } 0. \end{cases}$$

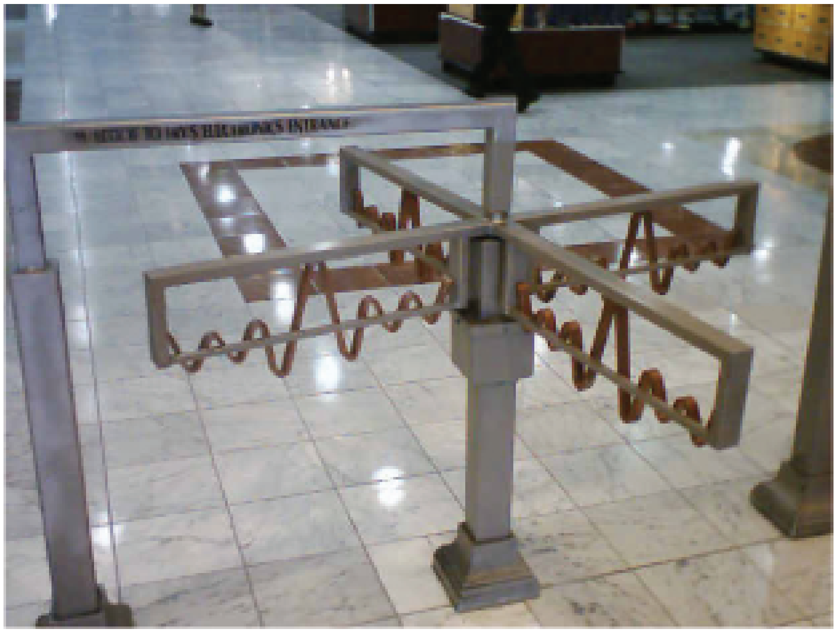
You may have thought that $\sin x/x$ only comes up as an example of $0/0$ in calculus classes and could have no possible interest beyond that. No, no! It's in your CD player, for example, for reasons we'll explain. In fact, the function comes up so often it's given a name, the *sinc function*, and denoted

$$\operatorname{sinc} s = \frac{\sin \pi s}{\pi s}.$$

Here's the graph.



Engineers see this curve when they sleep, and when they shop:



A warning. The computation we did of $\widehat{\Pi}(s)$ is just about the simplest one that can be done of a Fourier transform, but look at the trouble it causes: Fourier inversion would have us calculate

$$\int_{-\infty}^{\infty} e^{2\pi i s t} \frac{\sin \pi s}{\pi s}\, ds.$$

This is not an elementary integral by any means, and it's not at all clear that it even exists. By special techniques one can show that

$$\int_{-\infty}^{\infty} e^{2\pi i s t} \frac{\sin \pi s}{\pi s}\, ds = \begin{cases} 1, & |s| < \frac{1}{2}, \\ \frac{1}{2}, & |s| = \frac{1}{2}, \\ 0, & |s| > \frac{1}{2}. \end{cases}$$

(This result is an argument for defining Π to have value $1/2$ at $\pm 1/2$, but, as I said, this won't be an issue.)

A Theorem. Let me also provide you with two simple results, of a general nature, each showing a way in which the time and frequency domain representations of a signal interact. Simple, but very important in applications.

Suppose we multiply the signal by a complex exponential. What happens to the Fourier transform? More precisely, we ask: What is the Fourier transform of $e^{2\pi i s_0 t} f(t)$, where s_0 is fixed. I'll explain in a moment why we ask the question this way, and why we ask the question at all, but first let's do the calculation. The Fourier transform of $e^{2\pi i s_0 t} f(t)$ is

$$\int_{-\infty}^{\infty} e^{-2\pi i s t} e^{2\pi i s_0 t} f(t)\, dt = \int_{-\infty}^{\infty} e^{-2\pi i (s-s_0) t} f(t)\, dt = \hat{f}(s - s_0).$$

On the right hand side we see the Fourier transform of f shifted to s_0. In words, multiplying by a complex exponential of frequency s_0 in the time domain corresponds to a shift of the spectrum by s_0 in the frequency domain. This has *everything* to do with AM radio.

Let's also note a corresponding result. This time we ask, if we shift the time what happens to the spectrum? More precisely, what is the Fourier transform of $f(t - t_0)$? We get the answer by a direct calculation again, just slightly more involved than the one we just did. The Fourier transform of $f(t - t_0)$ is, by definition,

$$\int_{-\infty}^{\infty} e^{-2\pi i s t} f(t - t_0)\, dt$$

We make a change of variable:

$$\begin{aligned}
\int_{-\infty}^{\infty} e^{-2\pi i s t} f(t - t_0)\, dt &= \int_{-\infty}^{\infty} e^{-2\pi i s (u + t_0)} f(u)\, du \quad (\text{using } u = t - t_0) \\
&= \int_{-\infty}^{\infty} e^{-2\pi i s u} e^{-2\pi i s t_0} f(u)\, du \\
&= e^{-2\pi i s t_0} \int_{-\infty}^{\infty} e^{-2\pi i s u} f(u)\, du \\
&= e^{-2\pi i s t_0} \hat{f}(s)
\end{aligned}$$

This says that a shift in the time domain corresponds to multiplying the Fourier transform by a complex exponential in the frequency domain.

Everything in modulation. An operation on a signal or on its spectrum is often referred to generally as *modulation* of some sort. The tools one uses to modulate, typically electrical circuits, may either operate on the signal directly in the time domain or operate by changing the frequencies in the spectrum of the signal.

We'll use the problem of transmitting audible sound as an example of the use of modulation, Let's recall the following important facts, which we mentioned earlier in connection with hearing:

> The sound signals you can hear have a bandwidth of about $20,000$ Hz, from a low frequency of about 20 Hz to a high frequency of about $20,000$ Hz. That's the complete range, including music. Spoken speech has a bandwidth of about $4,000$ Hz.

Suppose we wanted to broadcast a musical note, like an A at 440 Hz. Why not convert a 440 Hz sound wave into a 440 Hz electromagnetic wave and transmit that (assuming that we have transducing devices that can convert back and forth between sound waves and electromagnetic waves)? There are three reasons why this would not be effective.

(1) There are many sources of low frequency (audio frequency) around in the environment ('random' electromagnetic signals, the 60 Hz 'hum' of electrical devices, *etc.*). These could interact with the broadcast A add a lot of noise to the actual signal.

(2) Only one radio station could transmit a signal at a time! If two radio stations were both transmitting signals at the same frequency or around the same frequencies (one transmitting an A and one transmitting a B, say) a radio would pick up a combination of both and be unable to sort the signals out.

(3) It turns out that for an antenna to transmit efficiently (measured by power) it should have length about a quarter of the wavelength of the signal being transmitted. This is a fact of the physics of antennas that we won't derive, but from it we can obtain an important conclusion. The antenna required to transmit electromagnetic signals in the audio range would be *huge*. Let's do a 'back of the envelope' calculation' – a simple estimate that will make the point. We use the relation

$$\text{speed} = c = \lambda\nu = \text{wavelength} \times \text{frequency}$$

where c, the speed of light, is about 3.0×10^8 m/sec. For a frequency of 1000 Hz, say, the wavelength is

$$\begin{aligned} \text{wavelength} &= \frac{\text{speed}}{\text{frequency}} \\ &= \frac{3.0 \times 10^8\,\text{meters/sec}}{10^3\,\text{cycles/sec}} \\ &= 3.0 \times 10^5\,\text{meters/cycle} = 300,000\,\text{meters/cycle} \end{aligned}$$

That's about 180 miles, so the corresponding antenna would have to be around 45 miles long!

The solution is to modulate the electromagnetic waves modeling sound to a *higher band of frequencies*. The higher the frequency the shorter the wavelength and hence the shorter the transmitting (and receiving!) antennas can be. Furthermore, different radio stations can transmit simultaneously using different band of frequencies which do not overlap, and which can then be separated by a receiver.

Of course, *producing* signals of very high frequency is a technological challenge. These devices are called *oscillators*. (Receiving and selecting bands of frequencies is another problem.) We won't go into the design of oscillators, but the fact that they can be built, and improved and improved to yield higher and higher frequencies means that more and more kinds of communications systems are possible. We'll look at AM (Amplitude Modulation) radio as an example. AM is simpler than FM (Frequency Modulation), at least as far as the description of the spectrum goes, and it's using the spectrum, as an application of Fourier transforms, that we want to emphasize.

You may already be a winner in the Electromagnetic Spectrum Sweepstakes! The possibility of broadcast systems raises important questions of public policy: Who decides who gets to transmit over which parts of the spectrum? The 'airwaves' belong to the public, and transmission crosses state borders. Thus regulation of the use of the electromagnetic spectrum is vested in the national government, specifically in the Federal Communications Commission in the Executive branch. Here's a quick description of the agency:

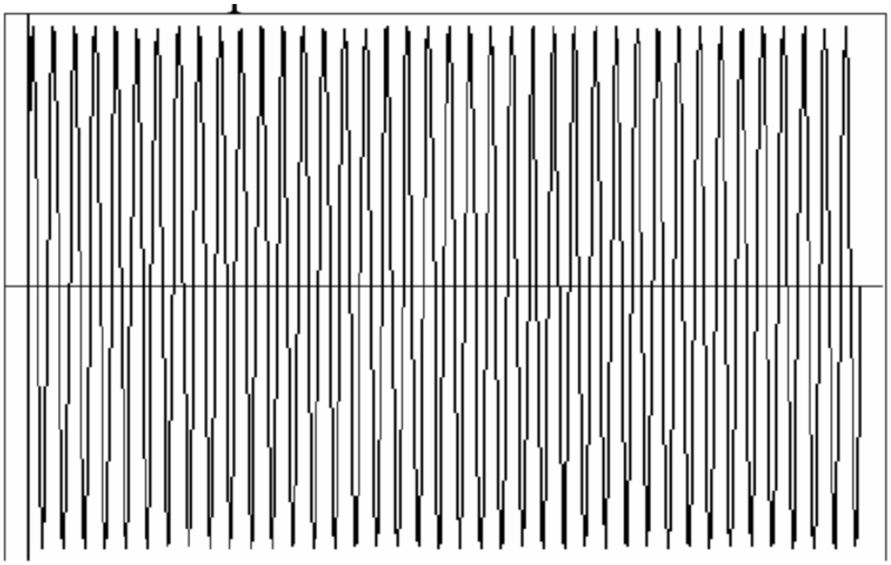
> Federal Communications Commission (FCC), independent executive agency of the U.S. government established in 1934 to regulate interstate and foreign communications in the public interest. The FCC is composed of five members, not more than four of whom may be members of the same political party, appointed by the president with the consent of the U.S. Senate. The commissioners are authorized to classify television and radio stations, to assign broadcasting frequencies, and to prescribe the nature of their service. The FCC has jurisdiction over standard, high-frequency, relay, international, television, and facsimile broadcasting stations and also has authority over experimental, amateur, coastal, aviation, strip, and emergency radio services; telegraph and interstate telephone companies; cellular telephone and paging systems; satellite facilities; and cable companies. The commission is empowered to grant, revoke, renew, and modify broadcasting licenses. It superintended the relations between AT&T and its successor phone companies and later promoted competition between long-distance phone companies. In the 1990s the FCC was involved in battles over the regulation of both pricing and content in the cable television industry. With the rapid development of telecommunications technologies, particularly mobile communications systems, and the blurring of distinctions between cable television and local and long-distance telephone companies, the job of the FCC continues to become more complex.

You may not have thought of the electromagnetic spectrum as a valuable commodity, but one of the most watched for events at the FCC is an 'auction' of a portion of the EM spectrum for commercial use. By analogy think of a supermarket offering to stock its shelves with products. Companies compete for – and pay for – 'shelf space'. For a chart showing the regulation of the EM spectrum, see

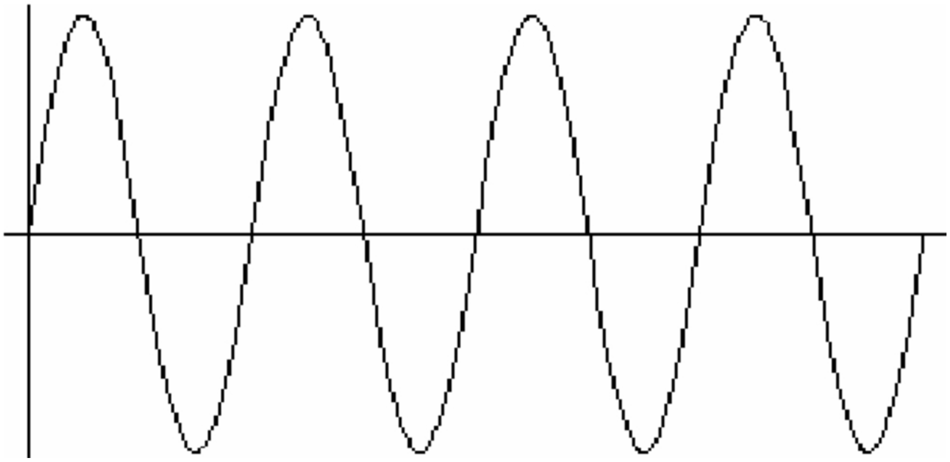
`http://www.ntia.doc.gov/osmhome/allochrt.pdf`.

AM Radio: News, Talk, Weather, Sports, Traffic. Let's go back to the problem of a radio station transmitting an 'A'. That's a signal of 440 Hz, and we assume that the station can convert this to an electromagnetic wave of that frequency. In radio lingo, the range of signals we want to transmit are often referred to as the *baseband frequencies*,, and the actual sound wave to be transmitted is called the *baseband signal.* As we remarked earlier, the station can't transmit at 440 Hz. Instead, the station starts with a *carrier wave* oscillating at a much higher frequency, say 1000 kHz.[4] The carrier wave starts out with a constant amplitude. At the radio station, the amplitude of the carrier wave is then made to vary with a periodicity of 440 Hz, so the sound signal is encoded in the variation in the amplitude of the carrier wave – thus the name Amplitude Modulation. This is actually not hard to write down in formulas – it amounts to multiplying two sinusoids. In pictures, the carrier wave looks like:

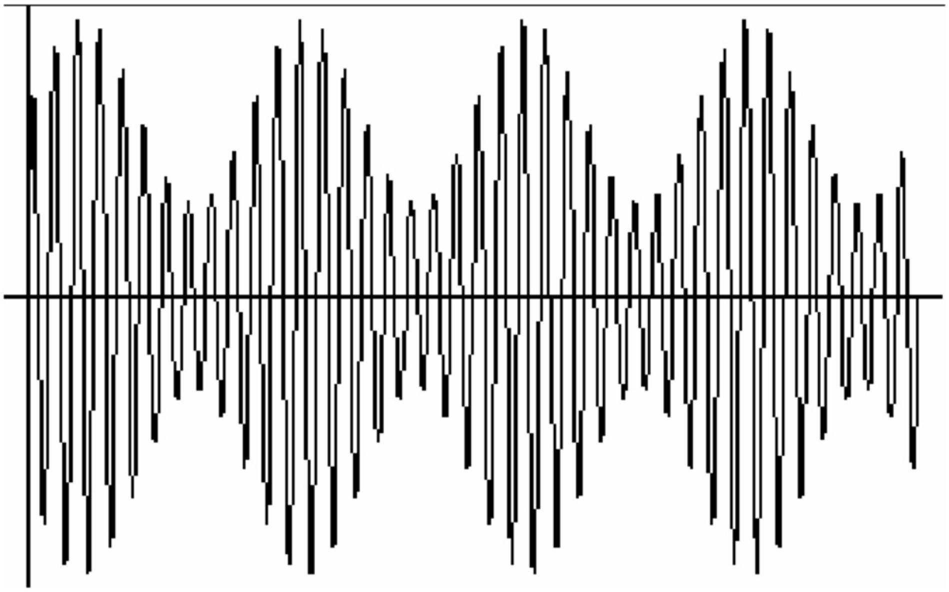
[4] Note: *kilo*Hz, is a unit of thousands of cycles per second. So $1,000$ kHz is 1 MHz (mega Hertz), 1,000,000 cycles per second, or about 2000 times faster than the frequency of the A.



The baseband looks like:

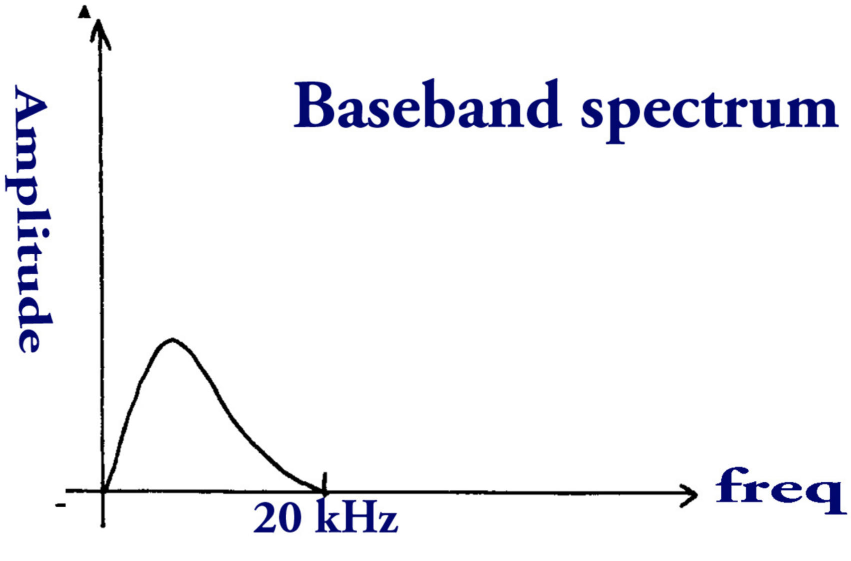


And multiplying the two together produces the baseband modulated by the carrier:

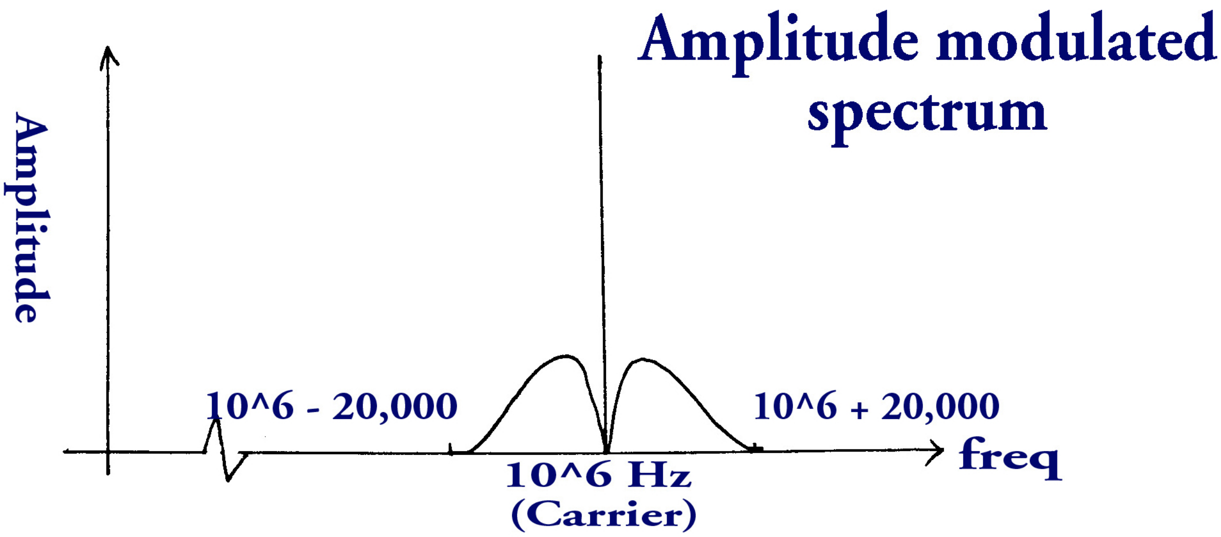


We see that the signal we're interested in hearing together with its reflection is an 'envelope' of the carrier wave; the carrier wave oscillates really fast in between the signal and its reflection, and it's amplitude varies according to that signal. At the receiving end, the circuitry in the radio strips away the carrier wave leaving the signal that we want. (Amplification is also necessary, since the signal usually reaches the radio at too low a power to drive the speaker.)

What about the spectrum? Amplitude modulation is exactly the process of multiplying a baseband signal $f(t)$ by a carrier signal $e^{2\pi i s_0 t}$. (Don't trouble yourself over using a complex exponential here. Think sine and cosine if you wish; just don't trouble yourself.) We have a theorem that tells us what happens to the spectrum – it's shifted. Here's a picture. If the baseband signal has a spectrum that look like this (I've only drawn the positive frequencies):



Then the spectrum of the signal modulated by a carrier wave of frequency 1000 kHz, for example, looks like this:

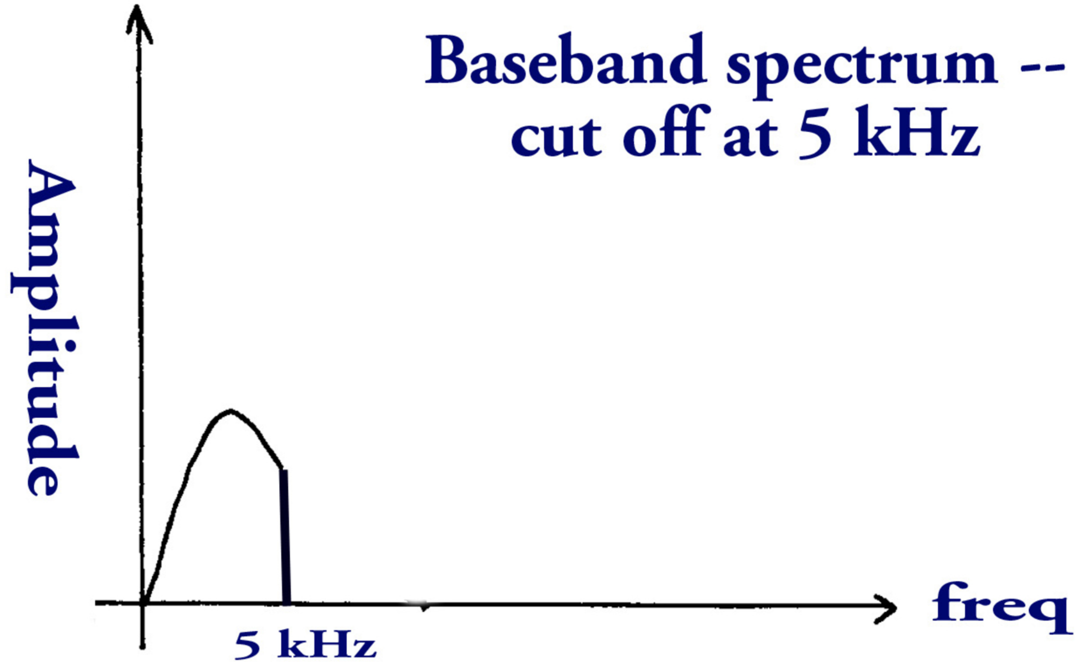


For broadcast, the spectrum has been shifted to a higher band of frequencies. At the receiving end – your radio – the carrier wave is stripped away (leaving the shape of the baseband signal) and the spectrum is demodulated down to the audible range. Note that the bandwidth for an AM station (the total width of the spectrum of the signals it would like to transmit) is *twice* the bandwidth of the original baseband spectrum.

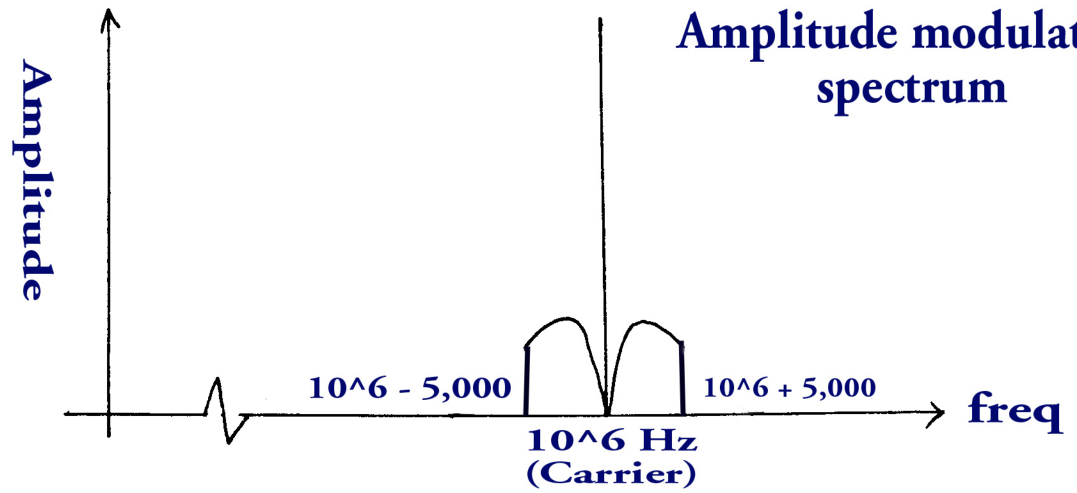
The AM radio band, or, Why does music sound so bad on AM?. You probably know that AM radio is *not* so good for music. It's OK for speech (talk radio), but the quality of sound for music just isn't there. It's not the speakers, and it's not really the method of modulation *per se.* It's the law!

AM radios receive signals (amplitude modulated electromagnetic waves) in the range 535 kHz to 1605 kHz, as you can check by reading off the AM dial. Furthermore, if you look at a paper that lists the AM stations, they're typically spaced 20 or 30 kHz apart. In New York City you'd find that they're only 10 kHz apart. On a 'digitally tuned' AM receiver you can push a button that will step you up or down the dial. It does so (or at least mine does) in steps of precisely 10 kHz. Where does this spacing come from? What are the reasons and consequences? The reasons are partly engineering and partly political. The use of the AM part of the electromagnetic spectrum is regulated by the FCC. Why it starts at 535 and ends at 1605 I don't know, but the rules on the spacing come from looking at the spectrum of an AM signal and parceling out a precious resource.

For radio communication we would like to be able to transmit all frequencies in the audible range, from about 20 Hz to 20,000 Hz (20 kHz); rather, we would like to be sure that all frequencies in this range can be used to modulate a carrier wave, and hence can be recovered by the radio at the receiving end. To have their signals not overlap when broadcasting in the same geographic region, two radio stations want their carrier signals to be spaced 20 kHz apart – 10 kHz on either side of the carrier frequency *for each station.* Way back at the beginning of commercial radio the FCC thought this was too luxurious. They had a limited amount of EM spectrum space to allocate, and they wanted to allow for as many stations as possible, so they insisted that the carrier frequencies could be at most 10 kHz apart, as is the case in New York City. So, now, if two radio stations are just 10 kHz apart on the radio dial, what is the range of baseband frequencies they can transmit without their signals overlapping in frequency? Only 5 kHz! AM radio stations are only allowed to broadcast a baseband spectrum 5 kHz wide – they have to squeeze everything into that, or cut everything else out. So AM stations filter out the high frequencies in the baseband signal *before* using it to modulate. Their baseband spectrum looks like this:



Now, that's fine for speech – remember the bandwidth of speech is about 4 kHz., but our hearing is sensitive up to about 20 kHz, and for *music* we need that bandwidth! We cannot – by law, and *only* by law – get it on AM radio and that's why AM radio sounds so bad in broadcasting music. The spectrum of the broadcast signal looks like this:

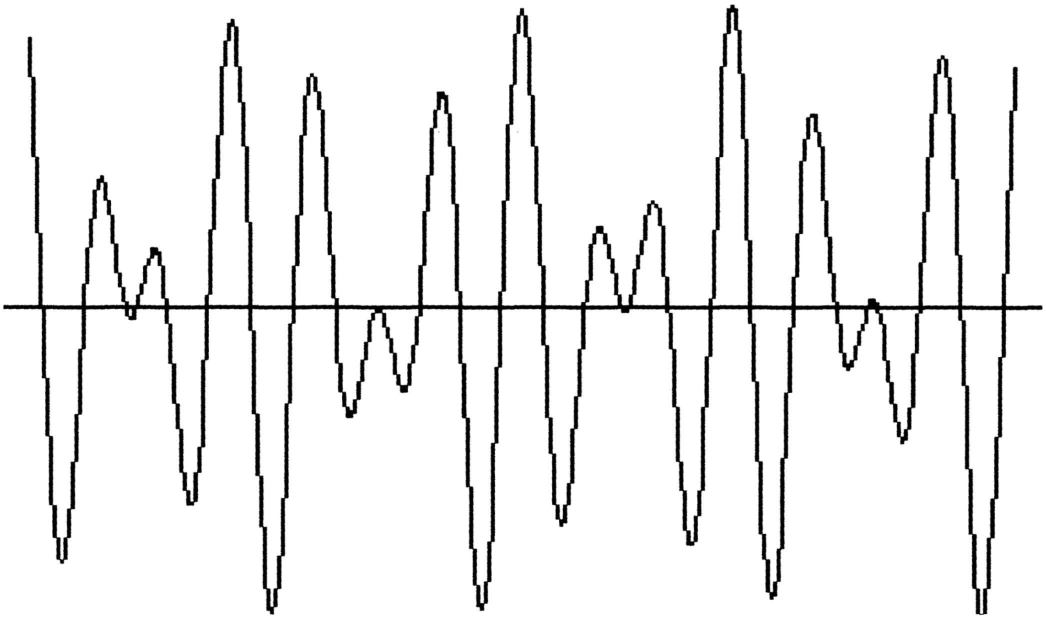


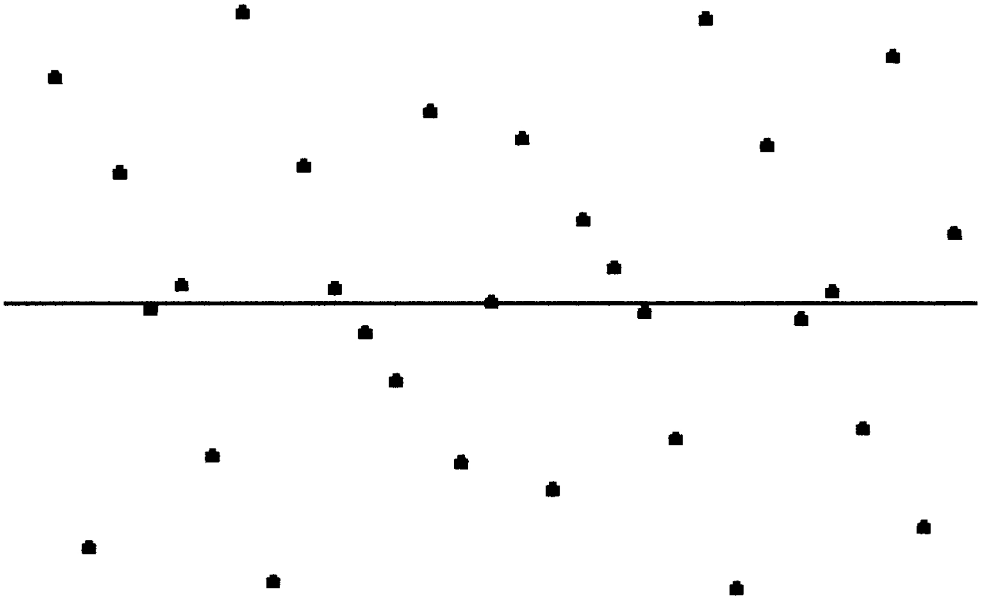
From continuous to discrete: The Sampling Theorem

Understanding the Fourier transform and the spectrum is the foundation for turning analog information into digital information. The question is:

> How accurately can a discrete set of sampled values of a continuous function represent the function at other values? That is, if we know the values of a function at a discrete set of points, how well can we interpolate the values in between those points?

This hardly seems reasonable – it even seems kind of crazy. A function, a signal, can jump all over the place, so how on earth would you expect to be able to say anything about *all* of its values by only knowing *some* of its values? Consider the following two pictures. The first is a sum of sine curves. The second is a bunch of points selected from the first.





How do you propose to reconstruct the curve in the first picture from the bunch of dots in the second? The craziest thing is that this isn't crazy. It's the representation of the signal in the frequency domain that leads to an answer.

You don't get something for nothing. Let's start with *assuming* something about the signal. The simplest case: Suppose we already *know* that the signal is a single, though general, sinusoid. We can write it either as,

$$A\sin(2\pi\nu t + \phi),$$

or as

$$a\cos(2\pi\nu t) + b\sin(2\pi\nu t).$$

To pin down *which* particular sinusoid we have to find the unknowns; either an amplitude, frequency and phase, in the first case, or the coefficients a and b and the frequency, in the second case. Same thing.

Remember the thing about periodic function is that if we know it for one period we know all of it. So how many points on the curve in one period, how many values, or samples, of the signal, would I need to know in order to pin it down – in order to find the unknowns. I need at least three; three unknowns, three equations, the values of the function at least three points *strictly within* a cycle, any cycle. Two points would not be enough, and neither would three if two of them are at the boundary of one cycle.

What if the signal is a sum of sinusoids,

$$\sum_{n=1}^{N}(a_n\cos(2\pi nt) + b_n\sin(2\pi nt)).$$

We'd want to sample all the terms, so to speak, and have enough samples to determine all of them. How many sample points might we need? Sample points for the sum, *i.e.* values of the sum at a set of points, are 'morally' sample points for the individual harmonics, though not explicitly. We need to take enough samples to get sufficient information to determine all of the harmonics. Now, in the time it takes for the combined signal to go through one cycle, the harmonics will have gone through several cycles. We have to take enough samples of the combined signal so that as the individual harmonics go rolling along, we'll be sure to have at least three samples in *some* period of *every* harmonic. To do this, 'number of samples' is not as useful an idea as 'rate of sampling', *i.e.* samples/sec. This, in turn, depends on taking evenly spaced samples.

Here's what I mean. The sampling rate is the number of samples we take per second. (So the sampling rate has the units of 1/sec, or Hz.) For a given harmonic of frequency ν we want

$$\text{Samples/cycle} > 2.$$

and since

$$\frac{\text{Samples/sec}}{\text{cycles/sec}} = \text{Samples/cycle}$$

we want

$$\text{Sampling rate} = \text{Samples/sec} > 2 \cdot \text{cycles/sec} = 2\nu.$$

This is the rate at which we should sample a given sinusoid to guarantee that some single cycle will contain at least three sample points. Furthermore, if we sample at that rate for a given frequency, we'll certainly be taking more than enough samples for a harmonic of *lower* frequency.

For our combined signal, the thing that's driving the sampling rate up are the higher frequencies – specifically, the *highest* frequency. If we sample at a rate greater than twice the highest frequency, our sense is that we'll be sampling often enough for all the lower harmonics as well and we can determine everything!

At this point we're thinking that we really can determine a signal from its samples *provided* that there is a highest frequency in its frequency spectrum. Though we've been working here with periodic functions, we want to apply this reasoning more generally, and we adopt the following important definition.

> A signal $f(t)$ is *bandlimited* if its frequency spectrum is contained in a closed bounded interval on the real axis. That is $\hat{f}(s) = 0$ for $|s| \geq B$

Not all signals are bandlimited, because, for instance, some require an infinite Fourier series for their frequency domain representation, *e.g.* a square wave. For the square wave there is no highest frequency.

For bandlimited signals we have the following remarkable result of Harry Nyquist.

Nyquist Sampling Theorem Suppose a signal is bandlimited. Let ν_{max} be the maximum frequency in its frequency spectrum. If the signal is sampled a rate $> 2\nu_{\text{max}}$ then the signal can be reconstructed *exactly* from its samples.

The number $2\nu_{\text{max}}$ is often referred to as the *Nyquist rate* and ν_{max} itself is often referred to as the *Nyquist frequency.*

I'm going to sketch a proof. It makes essential use of periodicity and Fourier series, applied in the *frequency* domain. We're supposing that $\hat{f}(s)$ is identically zero for $|s| \geq B$ so consider $\hat{f}(s)$ on the interval from $-B$ to B and extend it to be periodic of period $2B$. Expand $\hat{f}(s)$ as a Fourier series:

$$\hat{f}(s) = \sum_{n=-\infty}^{\infty} c_n e^{2\pi i n s/2B}.$$

Next, according to the Fourier inversion theorem,

$$\begin{aligned} f(t) &= \int_{-\infty}^{\infty} e^{2\pi i s t} \hat{f}(s)\, ds \\ &= \int_{-B}^{B} e^{2\pi i s t} \hat{f}(s)\, ds, \quad \text{since } \hat{f}(s) = 0 \text{ for } |s| \geq B. \end{aligned}$$

Now substitute the expression for $\hat{f}(s)$ via Fourier series:

$$\begin{aligned} f(t) &= \int_{-B}^{B} e^{2\pi i s t} \sum_{n=-\infty}^{\infty} c_n e^{2\pi i n s/2B}\, ds \\ &= \sum_{n=-\infty}^{\infty} c_n \int_{-B}^{B} e^{2\pi i s t}\, e^{2\pi i n s/2B}\, ds \\ &= \sum_{n=-\infty}^{\infty} c_n \int_{-B}^{B} e^{2\pi i (t + n/2B) s}\, ds \end{aligned}$$

The integral is straightforward, of the type we've done before. To ease the notation, write

$$\alpha = 2(t + \frac{n}{2B}).$$

Then

$$\begin{aligned}\int_{-B}^{B} e^{\pi i\alpha s}\,ds &= \frac{1}{i\alpha}(e^{i\alpha B} - e^{-i\alpha B})\\ &= 2B\operatorname{sinc}\alpha B,\end{aligned}$$

recalling that

$$\operatorname{sinc} x = \frac{\sin \pi x}{\pi x},$$

We thus find

$$f(t) = \sum_{n=-\infty}^{\infty} c_n 2B \operatorname{sinc}(2Bt + n).$$

Let's now solve for the coefficients. To get c_k let

$$t = \frac{-k}{2B}.$$

Then

$$\operatorname{sinc}(2Bt + n) = \operatorname{sinc}(-k + n) = \begin{cases} 1, & k = n \\ 0, & k \neq n. \end{cases}$$

That is

$$f(-\frac{k}{2B}) = \sum_{n=-\infty}^{\infty} c_n 2B \operatorname{sinc}(-k + n) = 2Bc_k, \quad \text{hence} \quad c_k = \frac{1}{2B} f(-\frac{k}{2B}).$$

Thus

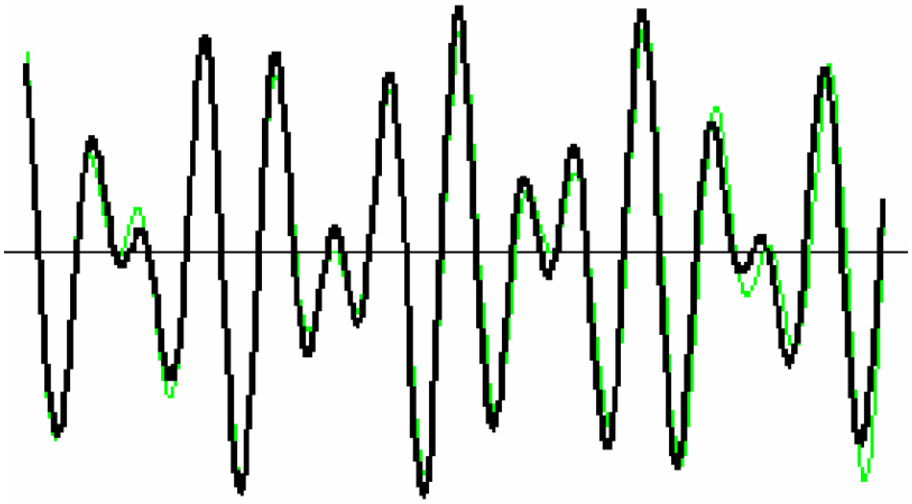
$$f(t) = \sum_{n=-\infty}^{\infty} f(-\frac{n}{2B}) \operatorname{sinc}(2Bt + n).$$

It's more common to replace $-n$ by n (we're summing over all positive and negative n) and to write the formula as

$$f(t) = \sum_{n=-\infty}^{\infty} f(\frac{n}{2B}) \operatorname{sinc}(2Bt - n).$$

Look carefully at what we've done here. On the right hand side we have discrete values of the function, the values $f(n/2B)$ at the *sample points* $n/2B$, and the formula says that via a sum of shifted sinc functions we can interpolate *any* value of the function $f(t)$ if we know its values only at the sample points. You can get it all back. That's what seemed so crazy when we first raised this as a possibility, but there it is. Remember, however, that there is the assumption that the signal is band-limited. That's not a trivial assumption, but it's also not a terribly restrictive one, certainly not in practice. This form of the sampling theorem, by the way, is usually credited to Claude Shannon (who else).

Now, in theory we have – and we need – infinitely many sample points in the interpolating sum. In practice the function $f(t)$ itself is only of finite duration and we take only a finite number of sample points. In that case we get an approximation, not an exact fit. But look how good it is! Here's a graph of that sampled sine function I showed earlier and the sinc interpolation based at the sample points. The green curve is the approximation.



CDs are as good as records. Sampling is the first step in digitizing a signal. We don't have to know all the steps, however, to have some appreciation for how the sampling theorem is used and why it is so important. Consider music. That is, consider reproducing music electronically. CDs are digital technology, and the first step in digitizing music to put it on a CD is to sample it. What should the sampling rate be? Once again, people hear in the range from about 20 Hz to 20,000 Hz. Some musical instruments may produce frequencies higher than that, but we don't hear them. Thus a piece of music is a bandlimited signal, with maximum frequency 20 kHz. According to the Nyquist theorem, we'll be able to reconstruct the signal exactly from its samples if we sample at a rate greater than 40 kHz. In fact, CDs are sampled at a rate of 44.1 kHz. (Where this exact number comes from I'm really not sure – it's partly due to the hardware design of some of the original tape recording machines used in the process.)