Sampling as Modulation

Douglas A. Kerr, P.E.

Issue 1
Date: November 21, 2011

ABSTRACT

In most methods of representing a continuous "function" (such as an audio or video waveform or a photographic image) in digital form the first step is to capture the value of the signal at regular intervals, a process called *sampling*. This process can be viewed as a form of (amplitude) modulation. That outlook assists us in understanding how the reconstruction of the original function from its samples works, and helps us understand the phenomenon of aliasing, which corrupts the reconstructed function in cases where the original function does not follow certain constraints. An appendix explains the concept of the *power spectral density* (PSD) plot, which will be encountered extensively in the article.

INTRODUCTION

Sampling

In most methods of representing a continuous function (such as an electrical audio or video waveform or a photographic image) in digital form the first step is to capture the value of the signal at regular intervals (usually of time or space, as applicable to the domain of the function), a process called *sampling*. A famous theorem formulated by Harry Nyquist (and later proven by Claude Shannon) states that if we sample more frequently than twice the highest frequency of any component of the signal, this suite of sample values (alone) **completely** describes the entire function. That means that from the train of sample values we can reconstruct the original function—all of it—not approximately but exactly.[1]

The "signal"

We will here for the most part consider a "continuous function" that is a voltage function of time, which, to facilitate recognition of its significance, we will speak of as a *signal*. The function of interest might equally well be the function of the air temperature of a storage cooler as a function of distance from the floor (along some specified vertical line). The concepts of course apply equally well to such functions.

---

[1] This does assume that the sample values are captured and maintained "absolutely precisely".

This signal function is *continuous* in two respects:

- In its *abscissa*: it has a value for however many instants of time as we may wish to imagine (over the span of our patience). Thus, strictly speaking, to describe the function over a certain span would require us to state an infinite number of values.

- In its *ordinate*: At any instant of time, its value may take on any value we can imagine (generally over some range). Thus, strictly speaking, to state its value at one instant of time would require an unlimited number of digits.

Of course, our practical need in representing this signal in digital form is to be able to do so with a finite amount of data. Overcoming the potential infinities in abscissa and ordinate are separate issues in the design of a digital representation system. Our work here, regarding sampling, only deals with disposing of the infinity in the abscissa: the apparent need to state an infinite number of values to describe the signal over some span of time. The Nyquist-Shannon sampling theorem of course offers us the solution to that challenge.

Aliasing

Note that the rule doesn't say that we must sample "more frequently than twice the highest frequency component in the signal **that we are interested in retaining**". If in fact the sampled signal contains components above a frequency of half the sampling rate (a limit called the *Nyquist frequency*, or *Nyquist rate*), these components aren't present in the reconstructed signal (about which we presumably don't really care), but are replaced by a component at a new frequency, below the Nyquist frequency. For example, if we sample 8000 times per second, and thus have a Nyquist frequency of 4000 Hz, an original signal component at 4800 Hz (4000 + 800) will appear in the reconstructed signal at 3200 Hz (4000 – 800).

Within the train of sample values, such an "out-of-band" component has the same representation as an "in-band" component (the one as which they reappear)—they are in effect traveling under another component's "identity". For this reason, this phenomenon is often called *aliasing*.

A MODEL

Introduction

Figure 1 is a block diagram of a totally-useless system, in which we can see the principle of representation of a function by samples, and its reconstruction from those samples, at work.

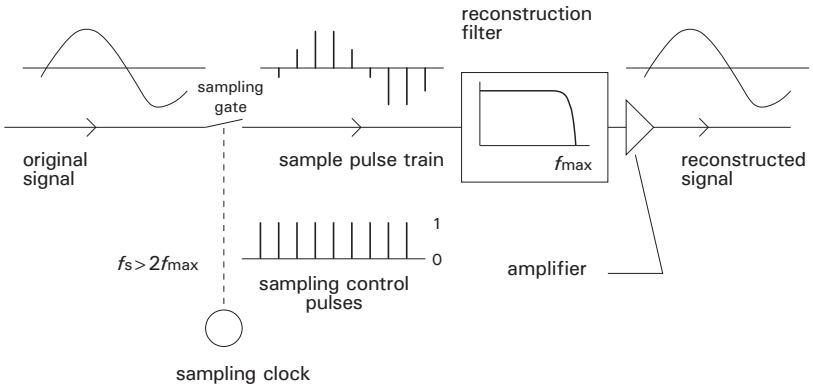

Figure 1. Useless sampling system

This is not a digital system at all, reminding us that the concept of sampling, although needed in most digital representation system, is not a creature only of digital representation, nor is it necessarily a digital process.

Design considerations

In the design of this "lecture room" model, we perhaps start with the aspiration to "capture" via sampling an electrical audio signal waveform which may contain components with frequencies up to 3500 Hz (a traditional "telephone bandwidth" audio signal). We learned earlier that Messrs Shannon and Nyquist have said that if we capture the instantaneous voltage of this waveform at periodic intervals, at a rate greater than twice the frequency of the highest frequency component in the signal, the suite of such sample values "completely describes" the signal waveform. The important corollary is that, from that suite of sample values, we can precisely reconstruct the original signal waveform.

We then choose a sampling frequency, f_s, a bit greater than $2 \cdot f_{max}$. Perhaps we would choose 8000 Hz (as is in fact done in digital telephone transmission systems). (That may seem an excessive "margin", but the reason for that will emerge in due time.)

Sampling

The action begins at the *sampling gate*. We can look at this as a switch that closes f_s times per second, each time for a very brief period (more about that shortly), under control of a train of sampling control pulses, generated by a sampling clock. During each interval of closure, the voltage of the waveform is passed to the bus to the right of the switch; the rest of the time, the voltage on the bus is zero. We thus see a train of very short pulses of differing height. (We assume for the moment that the duration of the gate closure is so short that

the voltage on the bus is essentially constant over the time of closure; that is, the pulses are "flat topped".)

The height of any pulse is the "sampled value" of the original waveform at the corresponding instant of time. It is not in numerical form; it is just a voltage. If the voltage of the pulse is 0.753128... V, that means that the instantaneous voltage of the original waveform at the corresponding instant of time—the "sample value" for that instant—is 0.753128... V.

Thus, although the "waveform" on the bus contains only a tiny portion of the original waveform, Shannon and Nyquist tell us that it is a complete description of that waveform.

Reconstruction

We now take the train of sampling pulses and pass it through a low-pass filter, whose cut-off is the f_{max} we assumed in the design of the system. Amazingly, what comes out is our original signal waveform! (How that happens we will see later.)

But it is greatly reduced in amplitude. The reason is that, if the sampling gate closed for only 1% of the sampling interval, then over any span of time, only 1% of the energy of the original waveform would pass though onto the sample train bus. And that is of course only enough energy to (in a perfect situation) create a waveform of 1/10 the amplitude of the original waveform.

So we place an amplifier with a voltage gain of 10 in the chain. Now we are ready to deliver to the amazed audience a **perfect** duplicate of the original waveform.

This is in complete fulfillment of the holy writ of Messrs. Shannon and Nyquist. Note my emphasis on "perfect" (that of course assumes that the amplifier is perfectly linear and so forth).

A digital case

Now, if this were an actual digital system, the overall block diagram would be that of figure 2.

Here, each sample pulse has its voltage measured by an analog-to-digital converter (the *sample encoder*), and reported as a binary number (code word) with some number of bits (chosen as another consideration in system design). The code words are transmitted in some format over a *digital transmission channel* to the "receiving" part of the system (the lower portion of the figure). There, a digital-to-analog converter (the sample decoder) takes each code word and generates a voltage pulse whose voltage is as defined by the code word.

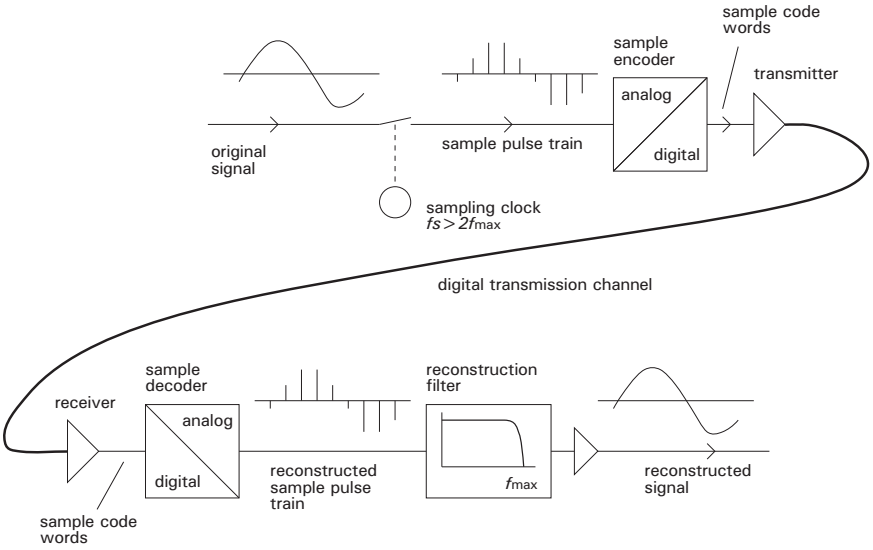

Figure 2. Digital transmission system

The train of pulses generated by the digital-to-analog convertor is a proxy for the original train of sample pulses. Thus, we can see that the overall operation is conceptually identical to the purely analog case seen in figure 1.

But there is one fly in the ointment. The digital words describe the voltages of the sample pulses only to a finite precision (exactly how much depending in the main on the number of bits in the code words). Thus, if a sample pulse had a voltage of 0.753128... V, the pulse from the digital-to-analog converter might have a voltage of 0.753 V. This is not because of any flaw in the system—it is a consequence of the finite precision of any numeric representation.

So of course, a waveform reconstructed from a series of samples that are not quite the same as the actual samples cannot be expected to be a perfect reproduction of the original waveform. This basic phenomenon of inaccuracy in the values is referred to as *quantizing error*, and the impact on the reconstructed waveform is often called *quantizing distortion*.

This is no failure of the Shannon-Nyquist sampling theorem, which did not say that we could precisely reconstruct the original function from **inaccurate** descriptions of its instantaneous values at regular intervals of time.

Dealing with this consideration is a gigantic topic in actual digital transmission systems, but we will not look into that here.

SAMPLING VIEWED AS AMPLITUDE MODULATION

Amplitude modulation

Amplitude modulation, as we most often encounter the term (in radio broadcasting, for example), refers to taking a sinusoidal *carrier wave* (of a frequency suitable to be propagated through space as an electromagnetic wave) and varying its amplitude to be proportional to some variable we wish to convey, such as the instantaneous value of an audio waveform.

But that can be described mathematically as taking the "carrier" signal and multiplying it by (in the familiar case, as in AM radio broadcasting) the instantaneous value of the audio waveform (assumed to have a maximum range of ± 1) plus 1.[2]

We see a simple block diagram of the process in figure 3:

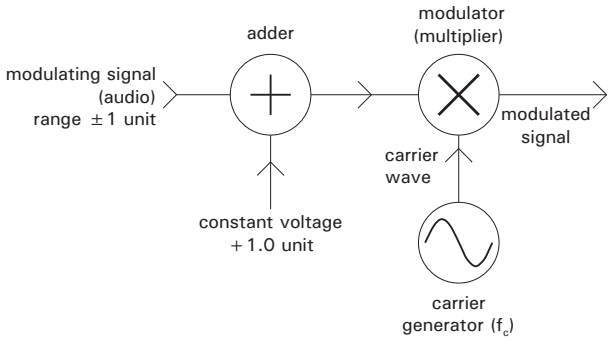

Figure 3. Amplitude modulator

But in other cases (called *double-sideband suppressed carrier* modulation), we do not add the constant "+1" term. Then the system becomes just as seen in figure 4.

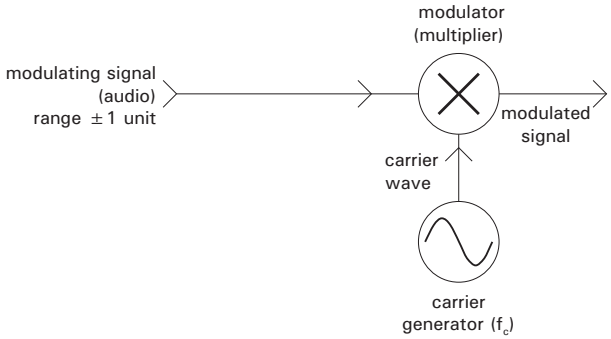

Figure 4. Amplitude modulator

---

[2] The plus one assures that the multiplier is never less than zero. It is perfectly workable to not have the constant "+1" term, and let the multiplier become negative for the negative parts of the audio waveform, but recovery of the audio signal at the receiver is more difficult, so this is not how AM broadcast radio works.

The spectrum of an amplitude modulation signal

In this article we will be very concerned with the matter of the different frequency components of various waveforms. In many cases, over some arbitrary interval of time, the situation is not of a collection of discrete components with explicit frequencies, but rather a distribution of the power in the signal, more-or-less continuously (but not necessarily uniformly), over some range of frequencies.

We can graphically present such as distribution on what is called a *power spectral density* plot, often just called a "spectrum" plot. Briefly, the height of the curve at any frequency tells us, in the signal represented by the waveform, the amount of power **per unit of frequency** at that frequency. Those not familiar with this concept may wish to see the explanation in Appendix A.

In figure 5, we will present such plots at various stages of an amplitude modulation process.

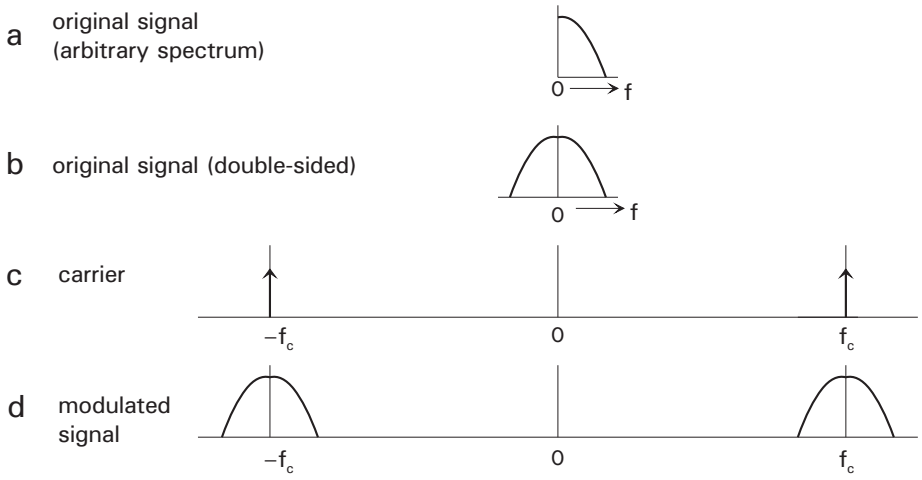

Figure 5. Spectrums in amplitude modulation

In panel **a**, we see the spectrum of our source signal. The spectrum shown is unrealistic and arbitrary, used just so we can recognize it when we see it later.

Actually, the mathematician sees this in a little different light. To the mathematician, the spectrum of the original signal has two parts, the one we usually think of (spanning a range of positive frequencies) and another, with mirror-image shape, extending into the negative frequencies (panel **b**). Now negative frequencies have no real physical significance—they are a mathematical fiction. (A sinusoidal signal with a negative frequency is exactly like one with the same positive frequency.) But this view makes many analyses work very tidily.

In panel **c**, we see the spectrum of the carrier signal, a sinusoidal waveform at frequency f_c, also shown in "double-sided" presentation. Its spectrum shows just a "spike" at frequency f_c (the carrier is at a discrete frequency, not a range of frequencies), plus another for its "alter ego", at $-f_c$.

There is a paradox about such "spikes". If there is power in the carrier wave (and if not, there is no carrier wave), then the area under the two spikes must equal that power. Since the width of the spike is theoretically zero, its height must theoretically be infinite. But we of course don't draw it that way. Instead, we sometimes draw a convenient-length line with an arrow head, a convention that tells us that the real height of the "spike" is (theoretically) infinite. (I don't do that in the figure here.)

When we amplitude modulate a sinusoidal carrier, we often describe the result by saying that the frequency spectrum of the resulting modulated signal consists of two "sidebands", one on either side of the carrier frequency, the upper one having the same shape as the single-sided spectrum of the original signal and the lower one being its mirror image. And we see that here in the vicinity of f_c. And of course since we look at the spectrum of the carrier itself as double sided, we see the same again in the vicinity of $-f_c$.

But, having accepted the fact that the spectrum of our original signal is double-sided, we can also say that the spectrum of the modulated signal is just the same as the spectrum of the original signal, shifted in the frequency domain by the amount of the carrier frequency (again, once for each of the two "alter egos" of the carrier).

Sampling as amplitude modulation

For convenient reference, I show here again figure 1.

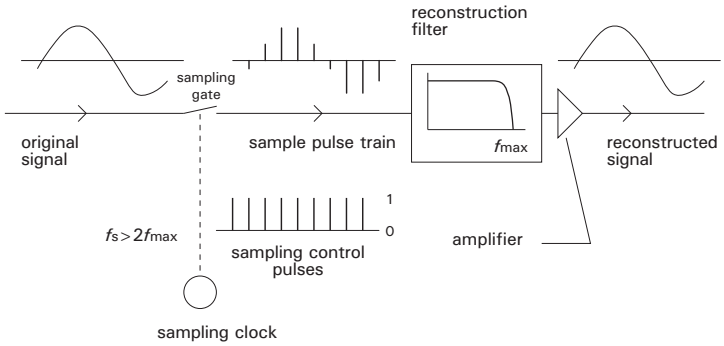

Figure 1 again. Useless sampling system

This is the conventional way of showing the capture of a signal by sampling and its later reconstruction from the train of samples.

But we can just as well imagine that the process of sampling is conducted as seen in figure 6.

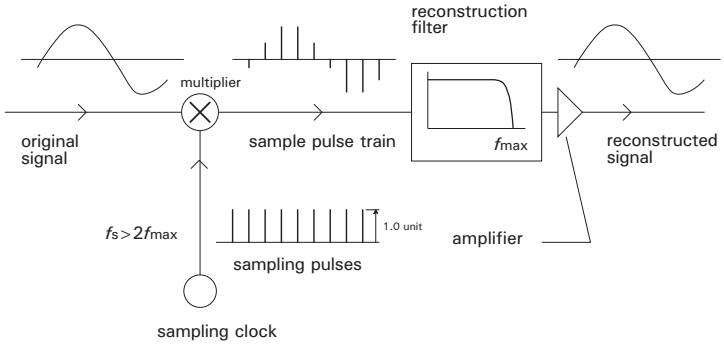

Figure 6. Useless sampling system—alternate implementation

Here we have replaced the sampling gate, closed briefly by a sampling clock at a rate of f_s, with a multiplier, whose two inputs are:

- The original waveform
- A train of very brief pulses, of height 1.0 unit, at a rate of f_s (generated by a slightly different type of sampling clock).

We should be able to see that the result is identical to the result of the arrangement of figure 6.

But we also recognize that the arrangement here is in fact just an amplitude modulation process, with one difference from the familiar situation: the carrier, rather than being a sinusoidal signal at frequency f_c is the train of sampling pulses at rate f_s.

So we see that we can look at the process of sampling as a case of amplitude modulation. And doing so will help us understand a number of things about the process.

Our next step is to look into the various spectrums involved in the system. We can follow the action on figure 7.

In panel **a** is the spectrum of our familiar arbitrary audio signal, and in panel **b** we see it "double sided".

Now we need to consider the spectrum of our rather peculiar carrier wave (*i.e.*, the train of sampling pulses). We see that in panel **c**.

Theoretically, this line of frequency "spikes" goes to both plus and minus infinite frequency.[3]

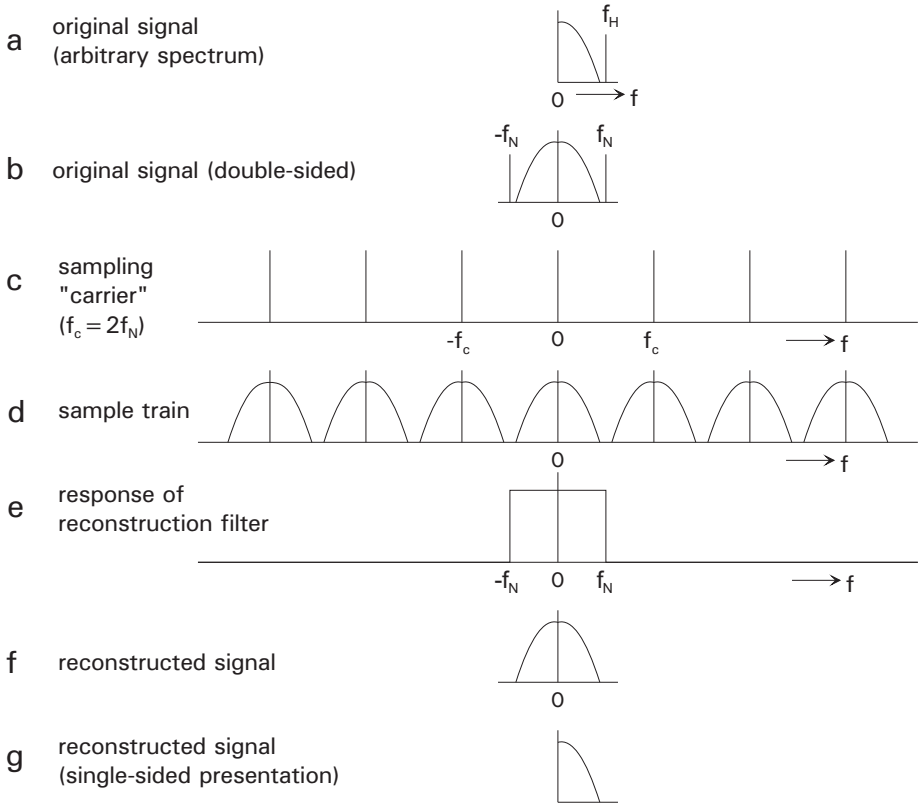

Figure 7. Waveforms in the sampling process

Because the time width of each of the sampling pulses is very small, the area under each of them is very small, and thus the total power in the sequence of sampling pulses would be very small, and thus the height of the spectrum would be very small. We have "blown it up" here to make the story easy to follow.

In panel **d**, we see the spectrum of the train of sample pulses (again blown up for convenience). We should expect this from the previous discussions regarding amplitude modulation. As a consequence of what we see in panel **c**, this too extends to infinite frequency in both directions.

Now, in our "useless" system, we immediately proceed to reconstruct the original signal, which we do by feeding the train of sample pulses to a low pass filter (the *reconstruction filter*). It has a cutoff frequency

---

[3] The electrical engineer will recognize that the presence of the spike at zero frequency, which we can think of as representing the "DC component" of the sampling pulse train, comes from the fact that the average value of the sampling pulse train is non zero.

of f_N. Of course, recognizing the mirror-image world of the mathematician, we show the response of this filter for both positive and negative frequencies.

In panel **e**, we see the spectrum of the output of the reconstruction filter. It of course only allows to pass the range of frequencies that happens to span the "haystack" centered around zero frequency. The spectrum of its output is seen in panel **f**.

But of course the spectrum we see in panel **f** is identical to that we see in panel **b**, which is the double-sided spectrum of the original signal. We see it in single-sided form in panel **g** (identical to the original signal seen in single-sided form in panel **a**).

Thus, the original signal has been recovered from the train of sample pulses, just as Messrs. Shannon and Nyquist assured us can be done.

The emergence of aliasing

Now we have supposed that we have chosen the sampling frequency high enough to accommodate all the frequency components we expect to have in our original signals. But for some reason, on a particular occasion the original signal contains frequency components above the limit (components in whose transport we are presumably not interested). Figure 8 illustrates this situation through our whole sampling system.

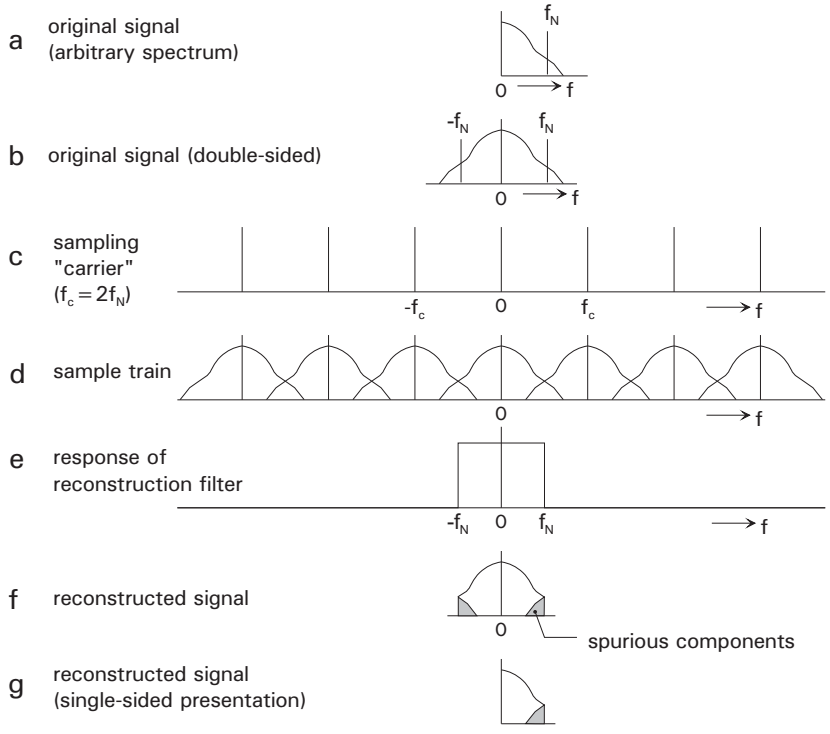

Figure 8. Sampling with foldover distortion (aliasing)

Note in panels **a** and **b** the new components in the spectrum of our original source signal above the limit dictated by the sampling frequency we chose.

We follow this signal through its journey as before. We see on panel **d** that the multiple copies of the original signal spectrum (generated thanks to the different harmonic components of the carrier) now overlap in the frequency domain.

When we apply the reconstruction filter, it lifts out most of, but not all of, one copy of the original signal (the copy centered about zero frequency) plus the "tails" of two other copies—those centered about the first component of the "carrier" on each side of zero frequency. Note that these ingredients are opposite in their frequency orientation from what they were in the original signal. These are anomalous components in the reconstructed signal. They corrupt the reconstructed signal. This aberration is called *foldover distortion*, or *aliasing*.

The lesson here is that we must put in our system, ahead of the point where the sampling is done, a baseband filter which eliminates any "unexpected" components in the original signal at or above the "limit" frequency, f_N, dictated by the sampling frequency. They are presumably irrelevant to our purpose, or we wouldn't have ignored them in choosing the sampling rate.

We see that filter added in the actual digital model system in figure 9.

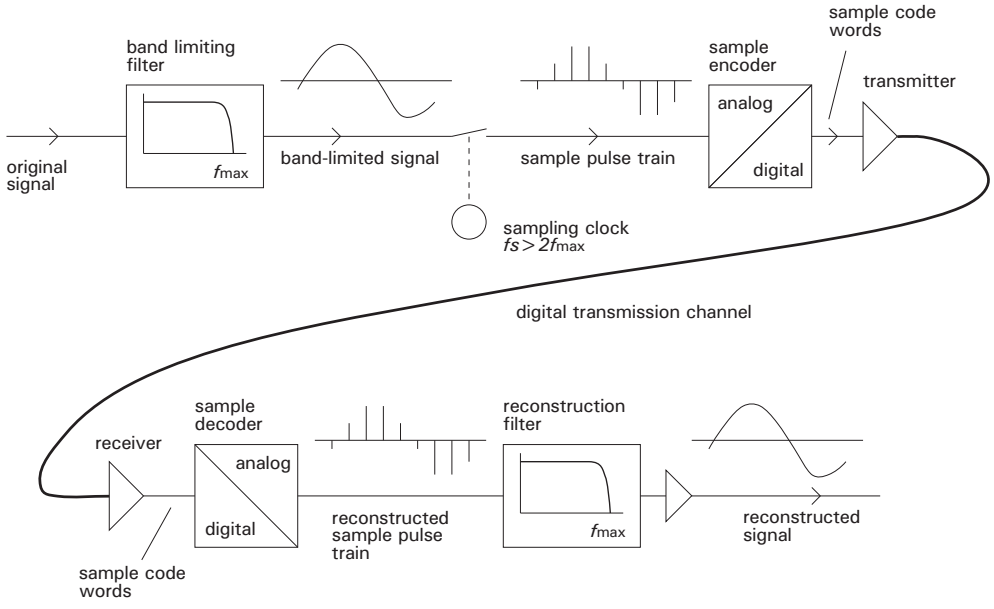

Figure 9. Addition of the band limiting filter

This band limiting filter is often called, for reasons that should by now be evident, an *anti-aliasing* filter.

SUMMARY

We have seen that the capture of an electrical waveform in the form of regularly-spaced “samples” of its instantaneously value is exactly equivalent to the amplitude modulation of a special “carrier wave” (a train of narrow pulses) by the waveform. The reconstruction of the original waveform, from the train of samples, by a low-pass filter, and the working of the phenomenon of aliasing, can be intuitively seen through such an outlook.

#

APPENDIX A

The power spectral density function

Introduction

In this appendix, we discuss the concept of the *power spectral density* function (PSD). In the process, we will illuminate the general concept of a density function.

Context

To make the discussion concrete, I will work in the context of an electrical waveform, which is a variation of voltage with time. The concepts are equally applicable to other continuous functions, such as a variation of temperature in a storage room with distance from the floor, or the variation in the illuminance on a camera image as a function of distance along a "track" across the image.

Power

Given that the waveform is electrical, it can be thought of as representing a voltage "signal", and if that voltage signal is applied to a "resistive load", power will flow. We might be interested in the instantaneous power at an instant of time, or the average power over some interval of time.

When we are dealing in the abstract with a waveform (as we do here), we often assume an arbitrary value of the resistance of the load (typically one ohm), so the conversion from voltage to power will be trivial ($p = v^2$).

Composition of a waveform

The "simplest" waveform (from the standpoint of interest to us here) is the *sine wave*. It is called that since its shape is identical to a plot of the trigonometric sine of an angle vs. the angle (when we allow the angle to advance over multiple revolutions.[4]

If we have a waveform that repeats over time but is not a sine wave, it can be shown that it can be considered to be the sum of multiple sine waves whose frequencies are the *fundamental frequency* of the waveform (the reciprocal of its period, the time it takes to repeat) and

---

[4] Curiously enough, in mathematical work, the most common "sine wave" function is actually the **cosine** function! Of course this has the same shape, but has a different origin in the time domain. But we still speak of it as a "sine wave".

integer multiples of that frequency, each with an appropriate amplitude (peak voltage) and time phase.

This representation of a recurrent waveform is called the *Fourier series* representation. We can think of it as an understanding of the waveform as being composed of multiple components, each of which is a sine wave with a certain frequency.

We could completely describe the waveform numerically in this way with a list of its components, stating for each its frequency, amplitude, and phase.

Now for some waveforms, there are an infinite number of these components, with frequencies extending to infinite frequency. Of course for those, a description in a "list" is not practical!

If we have a waveform that continues over time but is not (so far as we can tell) repetitive, then (from the perspective of any point in the time of its life) it can also be understood as comprising (sine wave) components at various frequencies, with various amplitudes and time phases, but in this case an infinity of those components (although perhaps over a finite range of frequencies). These are spaced in frequency by infinitesimal amounts. In other words, the distribution of the content of the waveform is not *discrete* with frequency (as in the case of a repetitive waveform) but rather is *continuous* with frequency.

We portray such a distribution with a graphical plot called the *power spectral density* (PSD) plot. We see an example in figure 10.

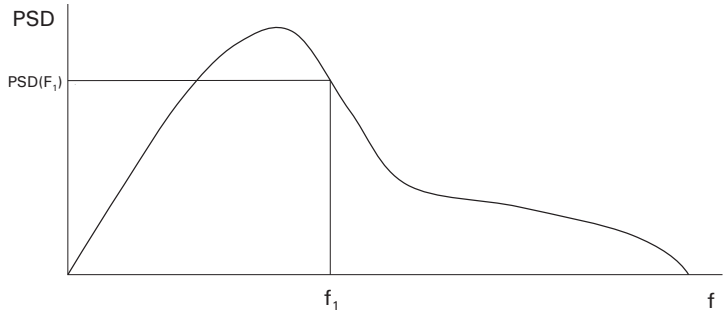

Figure 10. Power spectral density (PSD) plot

The x axis here represents frequency, the vertical axis the *power spectral density* (PSD) value. This plot, by the way, is often called the *spectrum* of the waveform.

In this case the nature of the waveform is that it contains no components (and thus no power) at frequencies higher than the frequency f_{max}; thus, the plot is not infinitely-wide (handy for all of us!).

It is tempting to think that for some frequency (for example, f_1) the height of the plot tells us the amount of power in the waveform. But in fact, for a continuous distribution of the power, there is zero power at any particular frequency. This is often startling when we first encounter it. But a little thought will show that this must be so. There are an infinite number of distinct frequencies over the range of the plot. If there was **any** power at each of them, the total power in the signal would have to be infinite.

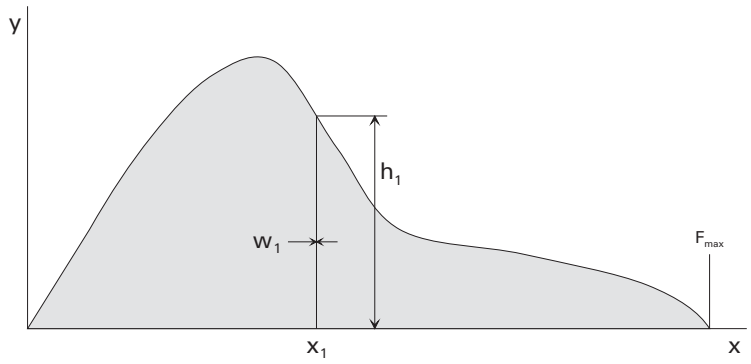

Figure 11. Aluminum plate analogy

We can come to grips with this perhaps-troublesome fact through an analogy. Figure 11 shows the profile of an aluminum plate (in light gray).

We use a system of x and y coordinates to refer to locations on it.

Now consider the "place" on the plate where $x\text{-}x_1$. It is shown by the vertical line on the figure.

Now we ask, "how much aluminum (in area) is there at this 'place' on the plate?" The answer is "none". That "place" on the plate, defined by a line ($x = x_1$), has zero width. The area of the aluminum there is the product of the "height" of the place (h_1) and the "width" of the place (w_1) which is zero. So the area is zero.

We will continue our work on figure 12.

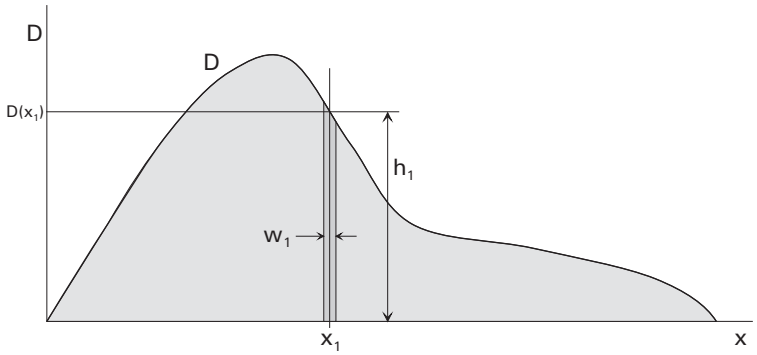

Figure 12. Aluminum plate analogy (2)

Here we will consider a very narrow stripe of the plate centered on $x = x_1$, with width w_1 (shown in darker gray).

The amount of aluminum (in area) in that little stripe is $h_1 \bullet w_1$. But if we take the ratio of that area to h_1, this is of course h_1.

Thus we can look at h_1 as not just telling us the height of the plate at $x = x_1$, but in fact also as telling us the "density" of the plate in terms of area "gathered" per unit distance in x.

This is the concept of a *density function*. Consider the curve, D, that follows the upper edge of the plate. The value of curve D does not tell us the amount of aluminum (in area) at some value of x, but rather the amount of area **per unit of width** (for very small widths around that value of x)[5]. For location x1, that value of D is what I have labeled $D(x_1)$. The curve D is called the *area density function* for the plate.

Thus it is for the PSD curve for a waveform. It tells us not the amount of power at a specific frequency (which amount is zero), but rather the amount of power **per unit of frequency** for a very small range of frequency (approaching zero, to be mathematically rigorous) centered about that specific frequency.

Finite bands of frequency

If in fact we consider two frequencies and draw vertical lines at those values, then the (graphical) area under the PSD curve between those two lines corresponds to the amount of power in the waveform for components (an infinite number of them!) whose frequencies lie between those two frequencies.

Assuming that the PSD curve is of finite overall width, then the total area under it must equal the total power in the waveform. The scale of the PSD function is in fact chosen to make this so.

The Fourier transform

If we take the voltage of a waveform (seen as a mathematical function of the variable *time*), and take its Fourier transform, we get what is sometimes called the *amplitude spectral density* (ASD) function, which we can think of as a curve against frequency. If we take the square of the ordinate of that curve (for every point of frequency), we get the *power spectral density* curve we have been discussing here.

---

[5] Formally, it is the ratio of the amount of power to the "width" in frequency, in the limit as the width in frequency approaches zero.

The ASD function is in fact a density function, but it is hard to give direct “graphical” significance to it. For example, we might suspect that for a certain range of frequencies, the area under the curve between those frequencies would be the total amount of something in the waveform between those two frequencies. But there is in fact no corresponding physical quantity for that “something”.

In fact, from an “electrical” standpoint, the amplitude spectral density curve must be thought of as just the square root of the PSD curve (at every frequency).

#