Chapter 2

Basics of Signals

2.1 What are Signals?

As mentioned in Chapter XX, a system designed to perform a particular task often uses measurements obtained from the environment and/or inputs from a user. These in turn may be converted into other forms. The physical variables of interest are generally called *signals*. In an electrical system, the physical variables of interest might be a voltage, current, amount of charge, etc. In a mechanical system, the variables of interest might be the position, velocity, mass, volume, etc. of various objects. Financial examples might include the price of a stock, commodity, or option, an interest rate, or an exchange rate. In performing its tasks, the system may need to manipulate or combine various signals, extract information, or otherwise process the signals. These actions are called *signal processing* or *signal analysis*.

A convenient abstraction is to model the value of a physical variable of interest by a number. We are usually interested in the physical variable not at just a single time, but rather at a set of times. In this case, the signal is a *function* of time, say $f(t)$. For example, $f(t)$ might denote a voltage level, or the velocity of an object, or the price of a stock at time t.

In some cases, we might be interested in measuring the quantity as a function of some variable other than time. For example, suppose we are interested in measuring the water temperature in the ocean as a function of depth. In this case, the signal is a function of a spatial variable, with $f(x)$ denoting temperature at depth x.

A signal need not be a function of just a single variable. To continue the example above, suppose we are interested in the temperature at particular points in the ocean, not simply as a function of depth. In this case, we might let $f(x, y, z)$ denote the temperature at the point (x, y, z), so the signal is a function

† Lecture Notes for ELE201 Introduction to Electrical Signals and Systems.

‡ Thanks to Richard Radke for producing the figures.

of three variables. Now, if we are also interested in how the temperature evolves in time, the signal $f(x, y, z, t)$ would be a function of four variables.

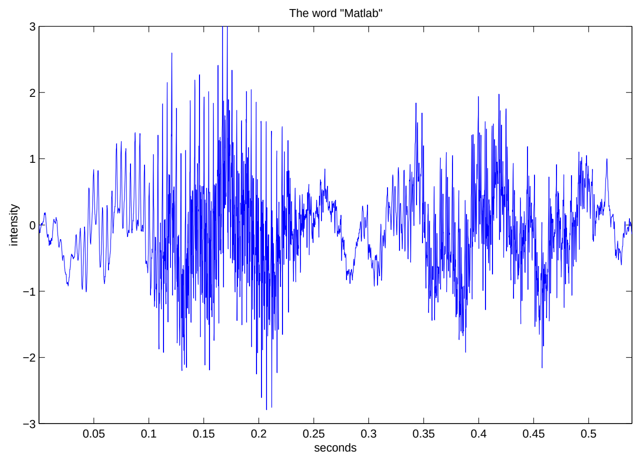

Figure 2.1: Someone saying the word "Matlab"

Examples of signals that we will encounter frequently are audio signals, images, and video. An audio signal is created by changes in air pressure, and therefore can be represented by a function of time $f(t)$ with f representing the air pressure due to the sound at time t. An example of an audio signal of someone saying "Matlab" is shown in Figure 2.1. A black and white image can be represented as a function $f(x, y)$ of two variables. Here (x, y) denotes a particular point on the image, and the value $f(x, y)$ denotes the brightness (or gray level) of the image at that point.

An example of a black and white image is shown in Figure 2.2. A video can be thought of as a sequence of images. Hence, a black and white video signal can be represented by a function $f(x, y, t)$ of three variables (two spatial variables and time). In this case, for a fixed t, $f(\cdot, \cdot, t)$ represents the still image/frame at time t, while for a fixed (x, y), $f(x, y, \cdot)$ denotes how the brightness at the point (x, y) changes as a function of time.

Three frames of a video of a commercial are shown in Figure 2.3. It turns out that color images (or video) can be represented by a combination of three intensity images (or video, respectively), as will be discussed later in Chapter XX.

2.2 Analog and Digital Signals

Often the domain and the range of a signal $f(x)$ are modeled as continuous. That is, the time (or spatial) coordinate x is allowed to take on arbitrary values (perhaps within some interval) and the value of the signal itself is allowed to take on arbitrary values (again within some interval). Such signals are called

Figure 2.2: A gray-scale image.

Figure 2.3: Video frames from a commercial.

analog signals. A continuous model is convenient for some situations, but in other situations it is more convenient to work with *digital* signals — i.e., signals that have a discrete (often finite) domain and range. Two other related words that are often used to describe signals are *continuous-time* and *discrete-time*, referring to signals where the independent variable denotes time and takes on either a continuous or discrete set of values, respectively.

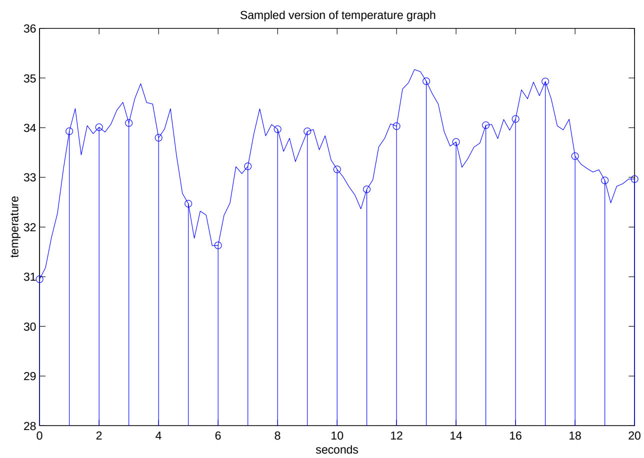

Figure 2.4: Sampling an analog signal.

Sometimes a signal that starts out as an analog signal needs to be *digitized* (i.e., converted to a digital signal). The process of digitizing the domain is called *sampling*. For example, if $f(t)$ denotes temperature as a function of time, and we are interested only in the temperature at 1 second intervals, we can sample f at the times of interest as shown in Figure 2.4.

Another example of sampling is shown in Figure 2.5. An original image $f(x, y)$ is shown together with sampled versions of the image. In the sampled versions of the image, the blocks of constant intensity are called *pixels*, and the gray level is constant within the pixel. The gray level value is associated with the intensity at the center of the pixel. But rather than simply showing a small dot in the center of the pixel, the whole pixel is colored with the same gray level for a more natural appearance of the image. The effect of more coarse sampling can be seen in the various images. Actually, the so-called "original" image in Figure 2.5a is also sampled, but the sampling is fine enough that we don't notice any graininess.

The process of digitizing the range is called *quantization*. In quantizing a signal, the value $f(x)$ is only allowed to take on some discrete set of values (as opposed to the variable x taking on discrete values as in sampling).

Figure 2.6 shows the original temperature signal $f(t)$ (shown previously in Figure 2.4) as well various quantized versions of f. Figure 2.7 shows the image from Figure 2.2 and various quantized versions. In the quantized versions of the images, the gray levels can take on only some discrete set of values. Actually,

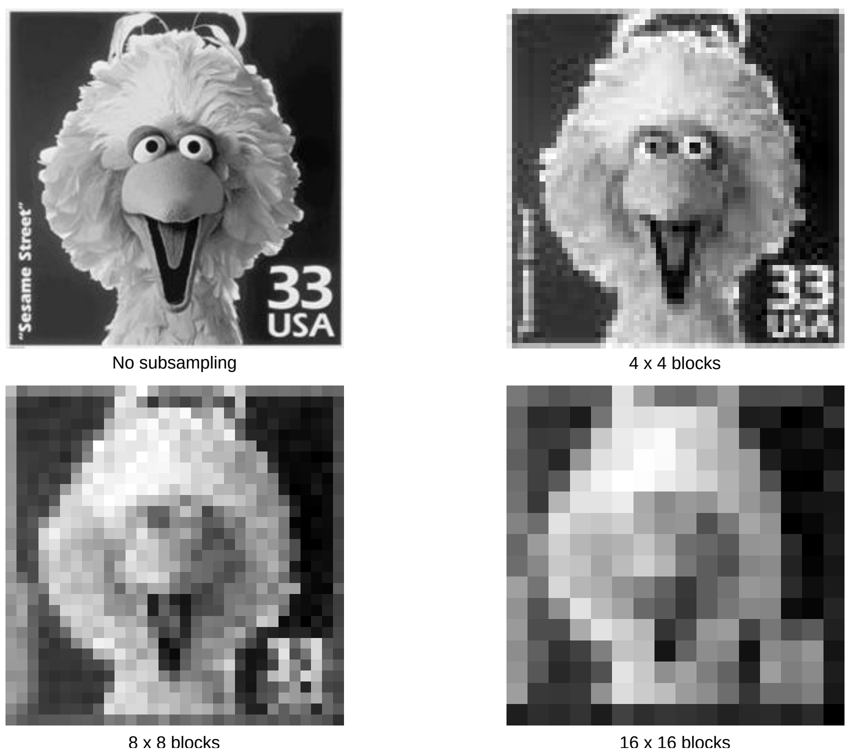

Figure 2.5: Sampling an image.

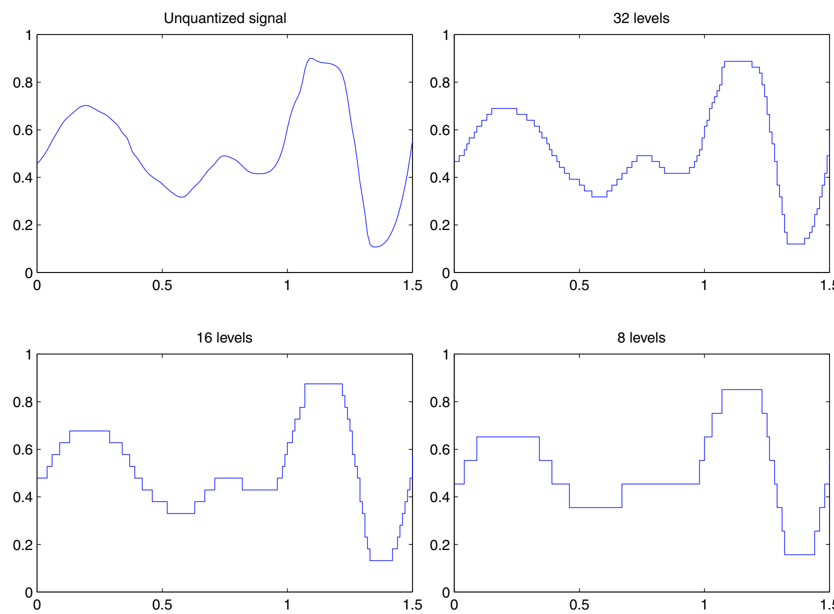

Figure 2.6: Quantized versions of an analog signal.

the so-called "original" image is also quantized, but because of the resolution of the printer and limitations of the human visual system, a technique known as halftoning (discussed in Chapter XX) can be used so that we don't notice any artifacts due to quantization. It is typical in images to let the gray level take on 256 integer values with 255 being the brightest gray level and 0 the darkest. In Figures 2.7d-f there are only 8, 4, and 2 gray levels respectively, and quantization artifacts become quite noticeable.

Sampling and quantization to digitize a signal seem to throw away much information about a signal, and one might wonder why this is ever done. The main reason is that digital signals are easy to store and process with digital computers. Digital signals also have certain nice properties in terms of robustness to noise, as we'll discuss in Section XX. However, there are also situations in which analog signals are more appropriate. As a result there is often a need for *analog-to-digital conversion* and *digital-to-analog conversion* (also written A/D and D/A conversion). In digitizing signals, one would also like to know how much information is lost by sampling and quantization, and how best to do these operations. The theory for sampling is clean and elegant, while the theory for quantization is more difficult. It turns out that choices for sampling rates and number of quantization levels also depend to a large extent on system and user requirements. For example, in black-and-white images, 256 gray levels is adequate for human viewing – much more than 256 would be overkill, while much less would lead to objectionable artifacts. We defer a more detailed consideration of sampling and quantization until Chapter XX after we have covered

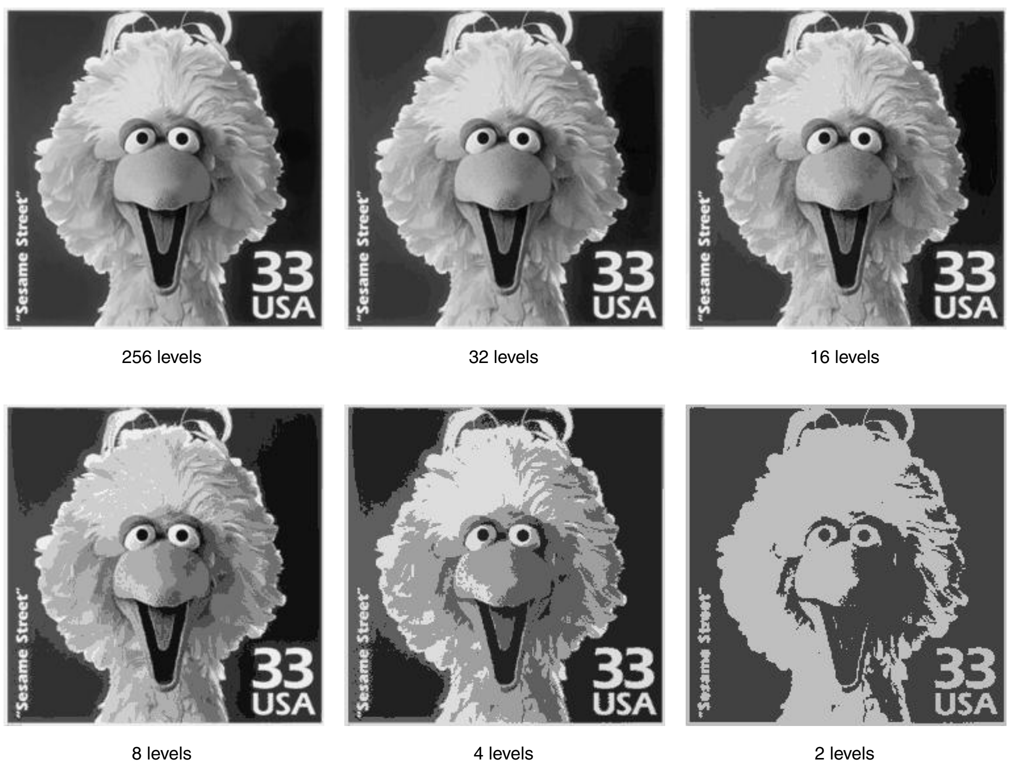

Figure 2.7: Quantized versions of a gray-scale image.

some additional background material.

Before discussing some basic operations on signals, we describe a fairly common notational convention which we will also follow. Continuous-time signals will be denoted using parentheses, such as $x(t)$, while discrete-time signals will use brackets such as $x[n]$. This convention also applies even if the independent variable represents something other than time. That is, $y(u)$ denotes a signal where the domain is continuous, while $y[k]$ indicates a discrete domain, whether or not the independent variables u and k refer to time. Often the letters i, j, k, l, m, n are used to denote a discrete independent variable.

2.3 Some Basic Signal Operations

In addition to the obvious operations of adding or multiplying two signals, and differentiating or integrating a signal, certain other simple operations are quite common in signal processing. We give a brief description of some of these here. The original signal is denoted by $x(t)$.

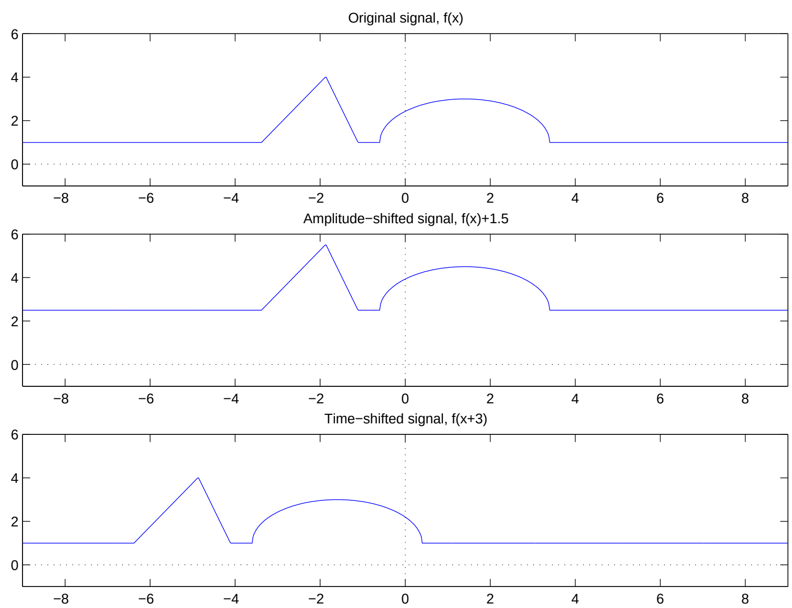

Figure 2.8: Amplitude- and time-shifted versions of a signal.

The signal $a + x(t)$ where a is some number is just adding a constant signal to $x(t)$ and simply shifts the range (or amplitude) of the signal by the amount a. A somewhat different operation is obtained when one shifts the domain of the signal. Namely, the signal $x(t - t_0)$ is a *time-shift* of the original signal $x(t)$ by the amount t_0. It's like a delayed version of the original signal. Figure 2.8 shows amplitude and time-shifted versions of a signal.

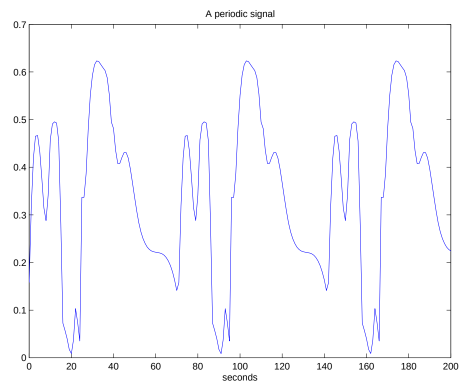

Figure 2.9: A periodic signal.

For some signals, appropriate time shifts can leave the signal unchanged. Formally, a signal is said to be *periodic* with period P if $x(t-P) = x(t)$ for all t. That is, the signal simply repeats itself every P seconds. Figure 2.9 shows an example of a periodic signal.

Amplitude scaling a signal to get $ax(t)$ is simply multiplying $x(t)$ with a constant signal a. However, a rather different operation is obtained when one scales the time domain. Namely, the signal $x(at)$ is like the original signal, but with the time axis compressed or stretched (depending on whether $a > 1$ or $a < 1$). Of course, if $a = 1$ the signal is unchanged. Figure 2.10 shows the effects of amplitude and time scaling. For negative values of a, the signal is "flipped" (or "reflected") about the range axis, in addition to any compression or stretching. In particular, if $a = -1$, the signal is reflected about the range axis, but there is no stretching or compression. For some functions, the reflection about the range axis leaves the function unchanged, that is, the signal is symmetric about the range axis. Formally, the property required for this is $x(-t) = x(t)$ for all t. Such functions are called *even.* A related notion is that of an *odd* function, for which $x(-t) = -x(t)$. These functions are said to be symmetric about the origin, meaning that they remain unchanged if they are first reflected about the range axis and then reflected about the domain axis. Figure 2.11 shows examples of an even function and an odd function.

The signal $x(y(t))$ is called the *composition* of the two functions $x(\cdot)$ and $y(\cdot)$. For each t, it denotes the operation of taking the value $y(t)$ and evaluating $x(\cdot)$ at the time $y(t)$. Of course, we can get a very different result if we reverse the order and consider $y(x(t))$.

One other operation that is extremely useful is known as *convolution.* We will defer a description of this operation until Section XX.

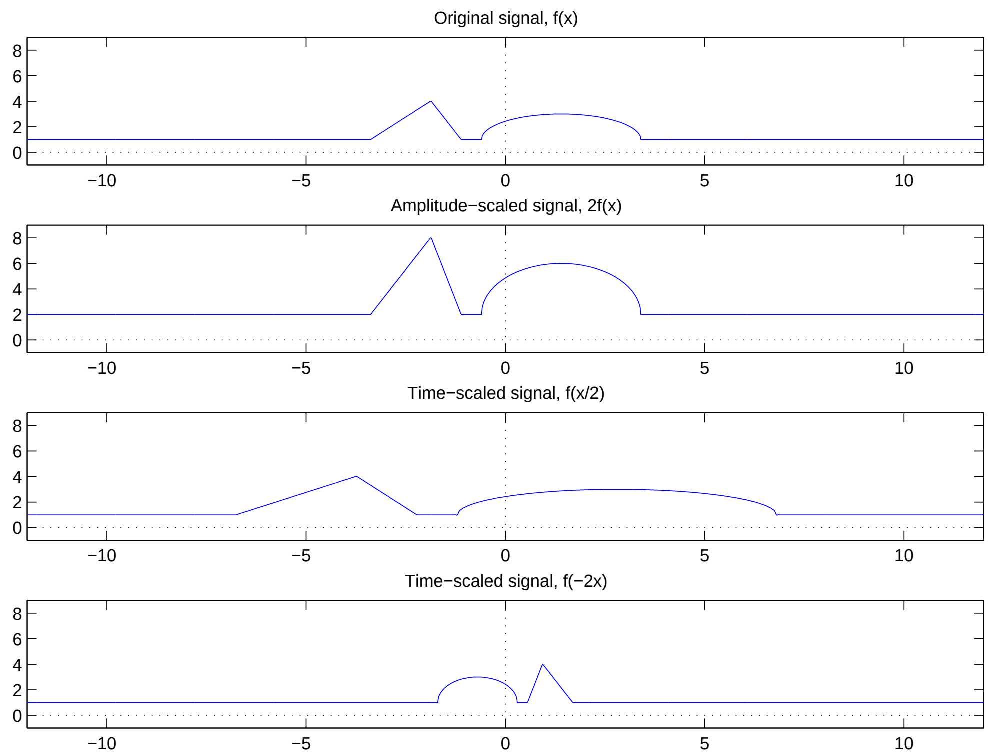

Figure 2.10: Amplitude- and time-scaled versions of a signal.

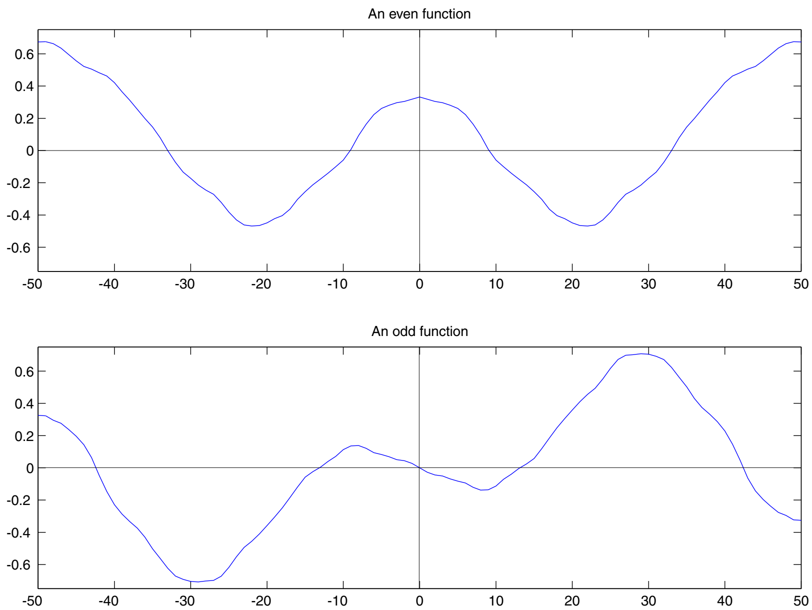

Figure 2.11: Examples of even and odd functions.

2.4 Noise

In many applications desired signals are subject to various types of degradations. These degradations can arise from a variety of sources such as limitations of the sensing device, random and/or unmodeled fluctuations of underlying physical processes, or environmental conditions during sensing, transmission, reception, or storage of the data. The term *noise* is typically used to describe a wide range of degradations.

It is often useful to try and model certain properties of the noise. One widely used model is to assume that the original (desired) signal is corrupted by *additive noise*, that is, by adding another unwanted signal. Of course, if we knew the noise signal that was added, we could simply subtract it off to get back the original signal, and the noise would no longer be an issue. Unfortunately, we usually do not have such detailed knowledge of the noise signal. More realistically, we might know (or assume) that the noise satisfies certain properties without knowing the exact values of the noise signal itself. It is very common to model the noise as random, and assume that we know something about the distribution of the noise. For example, we might assume that the noise is randomly (uniformly) distributed over some interval, or that it has a Gaussian (normal) distribution with a known mean and variance. Even this minimal type of knowledge can be extremely useful as we will see later.

Robustness to effects of noise can be a major design consideration for certain systems. This can be one reason why for many applications a digital system

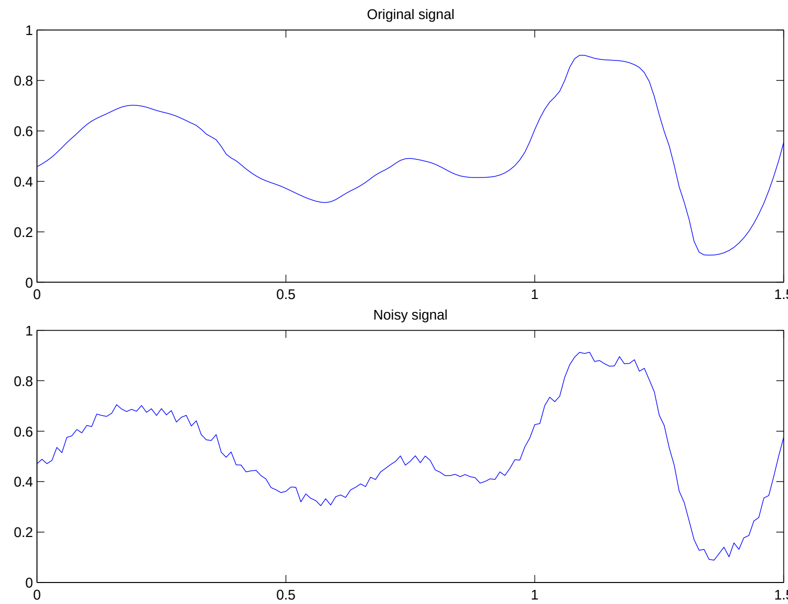

Figure 2.12: Adding noise to an analog signal.

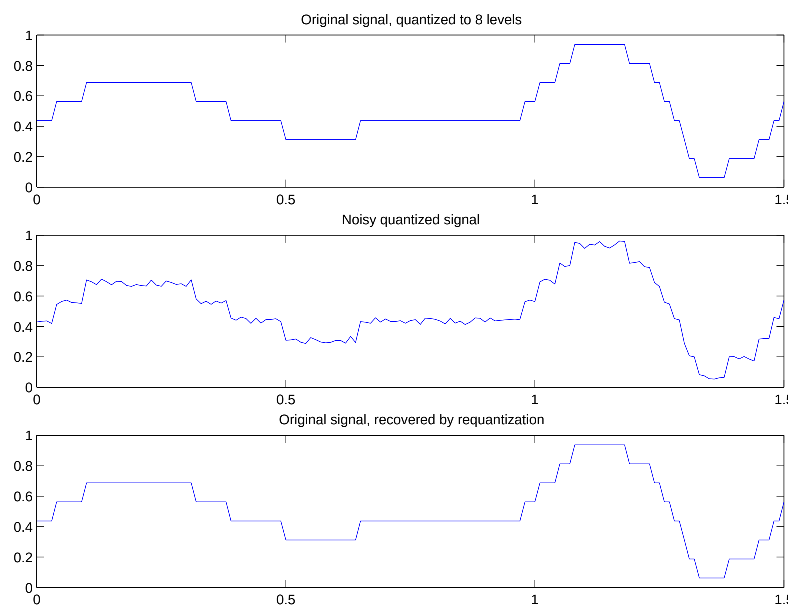

Figure 2.13: Adding noise to a quantized signal.

might be preferred over an analog one. Of course, the power of digital computing is also a key reason for the prevalence of digital systems, and robustness to noise is one factor that makes digital computing so reliable. Figures 2.12 and 2.13 illustrate the effect of adding noise on an analog signal and a quantized (although still continuous-time) signal. Without further knowledge of signal and noise characteristics, the noise cannot be removed from an analog signal since any possible value could be a valid value for the signal. On the other hand, if we know the original signal is quantized (so it takes on only a discrete set of values), then depending on the noise level, it may be possible to remove much of the noise by simply re-quantizing the noisy signal. This process simply maps the observed signal values to one of the possible original levels (for example, by selecting the closest level).

2.5 Some Common Signals

Here we briefly define some signals that we will commonly encounter. Perhaps the most basic and frequently used signal is a sinusoid defined by

$$x(t) = A\sin(\omega t)$$

and shown in Figure 2.14.

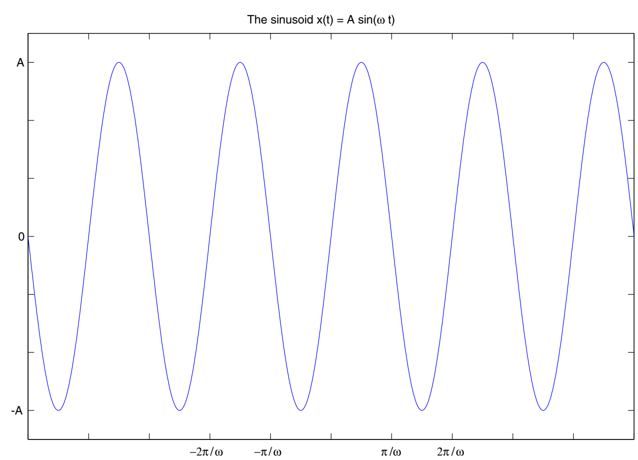

Figure 2.14: The sinusoid $x(t) = A\sin(\omega t)$.

Here A is the amplitude, and ω is the radian frequency. The units of ω are radians/sec so that when multiplied by time t (in sec) we get radians. An equivalent form for the sinusoid that is often used is

$$x(t) = A\sin(2\pi f t).$$

The frequency f is in units of *Hertz* (abbreviated *Hz*) which is $\sec^{-1}$, or often called cycles per second. Of course, f and ω are related by $\omega = 2\pi f$. Also, it's clear that since $\sin(\theta + \pi/2) = \cos\theta$, we could have equivalently written the sinusoid as

$$x(t) = A\cos(2\pi f t - \pi/2).$$

Up to this point, we have only considered real-valued signals. Although physical quantities can generally be represented in terms of real-valued signals, it turns out to be extremely useful to consider signals taking on complex values. The most basic complex-valued signal we will use is the complex exponential $e^{j\omega t}$. (Note that here we have used the symbol j instead of i to denote the imaginary number $\sqrt{-1}$. This is common in electrical engineering since the symbol i has traditionally been used to represent an electrical current.) The well-known Euler identity can be used to write the complex exponential in terms of standard sinusoids. Namely,

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

As with sinusoids, the complex exponential can also be written in terms of frequency in Hertz rather than radian frequency.

Some other signals that we will use on occasion and therefore give special symbols to are the step function, ramp, square wave, triangle wave, and the sinc function (pronounced like "sink"). These signals are defined by

$$\text{step}(t) = \begin{cases} 0 & \text{if } t < 0 \\ 1 & \text{if } t \geq 0 \end{cases}$$

$$\text{ramp}(t) = \begin{cases} 0 & \text{if } t < 0 \\ t & \text{if } t \geq 0 \end{cases}$$

$$\text{rect}(t) = \begin{cases} 1 & \text{if } -1/2 \leq t < 1/2 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{tri}(t) = \begin{cases} 1 - |t| & \text{if } -1 \leq t \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

and

$$\text{sinc}(t) = \frac{\sin(\pi t)}{\pi t}$$

and are shown in Figure 2.15.

2.6 Delta Functions

The notion of a delta function is extremely useful in the analysis of signals and systems, although it may feel unnatural on first exposure. Although the concept of the delta function can be made completely rigorous, rather than get side-tracked with too much mathematical detail and sophistication, our aim here is to provide some intuition and ability to work with the delta function. On the other hand, it is important to have enough rigor so that this important tool is used properly.

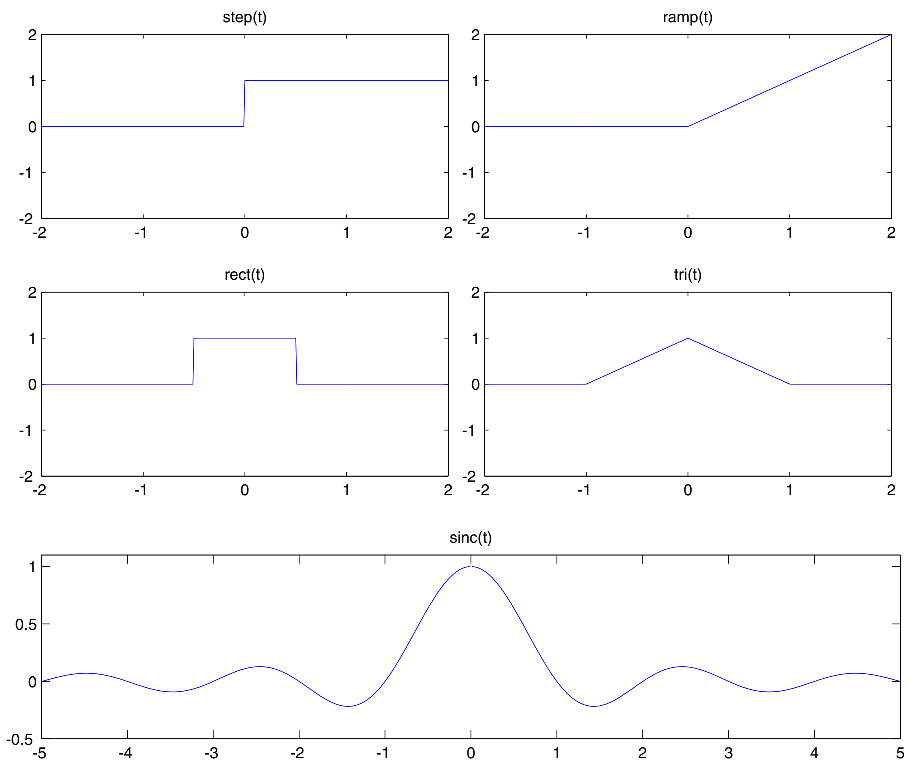

Figure 2.15: (a) Step function. (b) Ramp function. (c) Rectangle function. (d) Triangle function. (e) Sinc function.

The delta function in continuous-time is also called the Dirac delta function, unit impulse function, or sometimes just the impulse function. It is defined implicitly through its behavior under integration as follows:

Definition: $\delta(t)$ is the Dirac delta function if it satisfies

$$\int_{-\infty}^{\infty} f(t)\delta(t)\,dt = f(0) \tag{2.1}$$

for every function $f(t)$ that is continuous at $t = 0$.

From this definition we can infer the following two properties of the delta function. First, by considering the function $f(t) = 1$ in Equation (2.1), we get

$$\int_{-\infty}^{\infty} \delta(t)\,dt = 1 \tag{2.2}$$

This result implies that the area under the delta function is equal to 1.

The second property gives the value of $\delta(t)$ for $t \neq 0$. Suppose $\delta(t)$ took on positive values in even a very small interval away from $t = 0$. Then we could choose a function $f(t)$ that also took positive values inside a portion of this same interval, but with $f(t) = 0$ elsewhere (including $t = 0$) and with $f(t)$ continuous at $t = 0$. However, in this case the left hand side of Equation (2.1) must be positive, but the right hand side is 0. Therefore, $\delta(t)$ cannot take on positive values in any interval. A similar argument leads us to the conclusion that $\delta(t)$ cannot take on negative values in any interval. Thus, $\delta(t) = 0$ for $t \neq 0$.

These two results (namely, that the area under $\delta(t)$ is 1 and that $\delta(t) = 0$ for all $t \neq 0$) are inconsistent with our usual notions of functions and integration. If $\delta(0)$ was any finite value, then the area under $\delta(t)$ would be zero. Strictly speaking, $\delta(0)$ is undefined although it is convenient to think of $\delta(0) = \infty$. Thus, although we call $\delta(t)$ the delta "function," it is technically not a function in the usual sense. It is what is known as a distribution. However, it turns out that for many manipulations we can treat $\delta(t)$ like a function.

It is also convenient to have a graphical representation as shown in Figure 2.16. The arrow indicates that the value at $t = 0$ is infinite (or undefined), with the height of the arrow indicating the area under $\delta(t)$. To depict $A\delta(t)$ where A is some constant, we would draw the height of the arrow to be A.

It is sometimes also helpful to think of $\delta(t)$ as a limit of a sequence of approximating functions. Consider the function $a\,\mathrm{rect}(at)$. This has area 1, but if $a > 1$ it is more concentrated around $t = 0$. As we let $a \to \infty$ we get a sequence of approximations as shown in Figure 2.17, which intuitively get closer and closer to $\delta(t)$. In fact, it is not hard to verify that for $f(t)$ continuous at $t = 0$ we have

$$\int_{-\infty}^{\infty} f(t)\,a\,\mathrm{rect}(at)\,dt \to f(0) \quad \text{as } a \to \infty$$

so that in the limit $a \to \infty$ the defining property of $\delta(t)$ is indeed satisfied. It turns out that many other choices for the approximating functions will also work if the area is 1 and scaling is done to get concentration at $t = 0$.

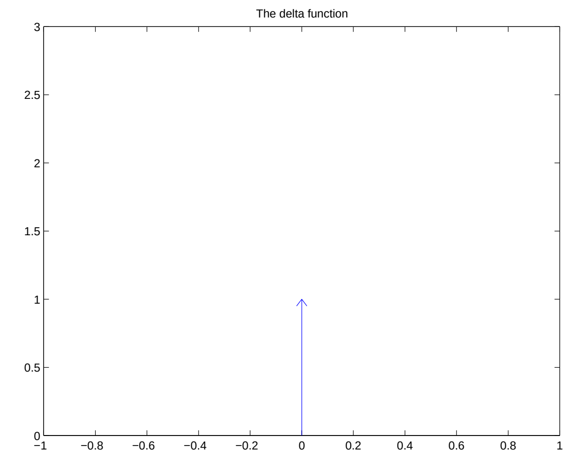

Figure 2.16: Representation of a delta function.

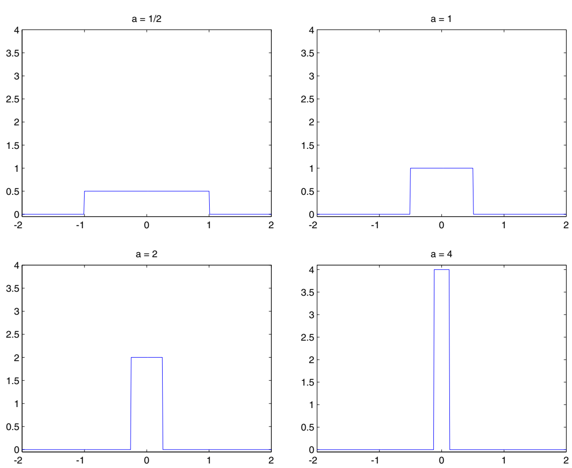

Figure 2.17: The delta function is a limit of rectangle functions with area 1.

The primary motivation for introducing the delta function is that it is a useful tool in analyzing systems and signals as we'll see in Chapter XX. There is also a physical motivation involving modeling of physical phenomena that is actually closely related to the analytical motivation. In some situations we would like to model physical phenomena that occur in a time interval short compared to the resolution of a measuring device. For example, we may be interested in the energy entering a camera due to a flash of light as the camera shutter opens and closes very quickly. For a fixed shutter speed, many models for the light source will be good enough to represent a "flash" or "impulse" of light. However, if we fix a particular function to model the flash of light, it may not represent a true "flash" for faster shutter speeds. Modeling the flash of light as a delta function is an idealization that works for *any* (non-infinite) shutter speed.

A number of properties of $\delta(t)$ can be obtained directly from the definition by utilizing the usual rules of integration. Intuition can also sometimes be gained by considering approximations of $\delta(t)$, although a formal justification generally requires verifying the defining property.

For example, since $\delta(t) = 0$ for all $t \neq 0$, it seems obvious that $\delta(t)$ is even. However, since $\delta(t)$ is technically not a function, we should really verify directly that $\delta(-t)$ behaves just as $\delta(t)$ in the defining property of Equation (2.1). This can be done by a simple change of variable as follows. For any function $f(t)$ continuous at $t = 0$, we have

$$\int_{-\infty}^{\infty} f(t)\delta(-t)\,dt = \int_{-\infty}^{\infty} f(-u)\delta(u)\,du = f(-u)\ \big|_{u=0} = f(0)$$

where the first equality is obtained by the change of variable $u = -t$, and the second equality follows from the definition of $\delta(u)$. The conclusion is that $\delta(-t)$ satisfies the required property of $\delta(t)$, and so $\delta(-t) = \delta(t)$.

By the change of variable $u = at$ and considering the cases $a > 0$ and $a < 0$ separately, it is easy to show that

$$\delta(at) = \frac{1}{|a|}\delta(t).$$

By the change of variable $u = t - t_0$, it follows that

$$\int_{-\infty}^{\infty} x(t)\delta(t - t_0)\,dt = x(t_0).$$

Therefore, the time-shifted delta function $\delta(t-t_0)$ behaves like we would expect. This property is sometimes called the sifting property of the delta function. The natural graphical depiction of $\delta(t - t_0)$ is shown in Figure 2.18.

We now turn to the discrete-time delta function, also called the Kronecker delta function. The Kronecker delta function is denoted by $\delta[n]$ and defined as

$$\delta[n] = \begin{cases} 1 & \text{if } n = 0 \\ 0 & \text{otherwise} \end{cases}$$

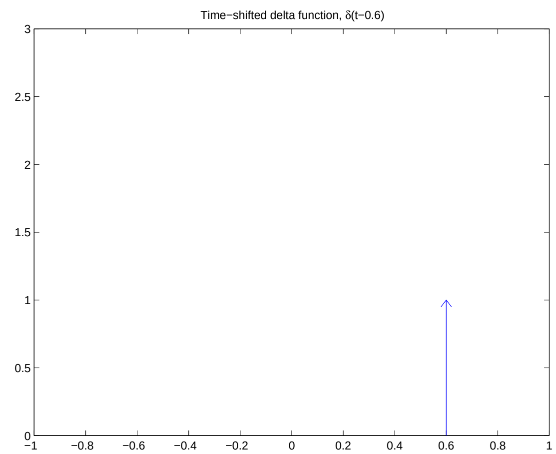

Figure 2.18: Time-shifted delta function.

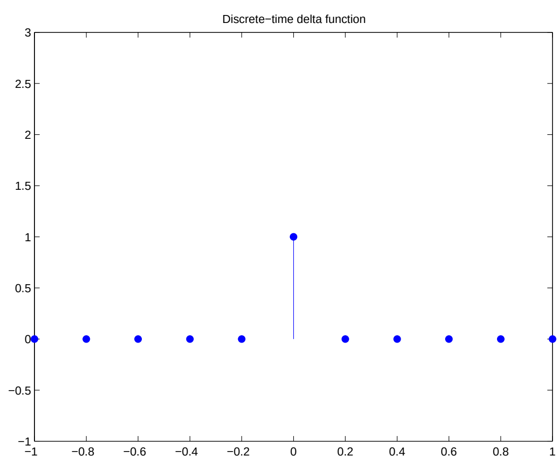

Figure 2.19: Discrete-time delta function.

Figure 2.19 shows the graph of $\delta[n]$. Hence, in discrete-time, the delta function is in fact a function in the proper sense. There are none of the mathematical subtleties/difficlties associated with the continuous-time delta function. In fact, $\delta[n]$ is rather simple to work with.

Many properties of $\delta(t)$ have analogous counterparts in discrete-time, and the discrete-time properties are generally easier to verify. For example, the result

$$\sum_{n=-\infty}^{\infty} f[n]\delta[n] = f[0]$$

follows trivially from the definition. Recall that in continuous time, the analogous property was actually the definition. Also trivial is the fact that $\delta[n]$ is an even function of $n = \ldots, -1, 0, 1, \ldots$. It is easy to see that the time-shifted delta function $\delta[n - n_0]$ satisfies the discrete-time sifting property

$$\sum_{n=-\infty}^{\infty} f[n]\delta[n - n_0] = f[n_0].$$

It turns out that for some properties the discrete-time counterpart is *not* analogous. For example, in discrete-time if a is an integer we have $\delta[an] = \delta[n]$.

2.7 2-D Signals

One useful notion that arises in two (and higher) dimensions is separability. A function $f(x, y)$ is called *separable* if it can be written as $f(x, y) = f_1(x)f_2(y)$. Many of the commonly encountered 2-D functions are simply separable extensions of the corresponding 1-D functions. For example, the 2-D version of the complex exponential is

$$e^{j(\omega_1 x+\omega_2 y)} = e^{j\omega_1 x}e^{j\omega_2 y}$$

where ω_1 and ω_2 are the radian frequencies in the x and y directions, respectively. That is, the 2-D complex exponential is simply the product of a 1-D complex exponential in each direction.

Likewise, the 2-D Dirac delta function $\delta(x, y)$ is given by

$$\delta(x, y) = \delta(x)\delta(y)$$

Formally, $\delta(x, y)$ would actually be defined by the property

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\delta(x, y)\, dx\, dy = f(0, 0)$$

for any function $f(x, y)$ continuous at $(0, 0)$, but showing equivalence with the separable expression is straightforward.

Similarly, in the discrete case, the Kronecker delta $\delta[m, n]$ is defined by

$$\delta[m, n] = \begin{cases} 1 & \text{if } m = n = 0 \\ 0 & \text{otherwise} \end{cases}$$

Clearly the Kronecker delta is also separable

$$\delta[m,n] = \delta[m]\delta[n].$$

Many other 2-D functions like $\mathrm{rect}(x, y)$ and $\mathrm{sinc}(x, y)$ have natural separable extensions from the 1-D versions as well.

Separability can offer some nice analytical and computational advantages. For example, terms involving the two independent variables can sometimes be separated, reducing a two-dimensional analysis to two separate one-dimensional analyses. This can result in computational savings by allowing processing to be done along the two dimensions separately.