Goal: Interpolation and Sampling Theorem

(a) Interpolation principles

(b) Pulse-type interpolations

(c) The sampling theorem

(d) Aliasing

1 Discrete-Time to Continuous-Time Conversion, aka Interpolation or Reconstruction

The task of interpolation is concerned with reconstructing a continuous-time signal from a discrete-time signal—a sequence of numbers.

Most practical interpolation is done in the following form

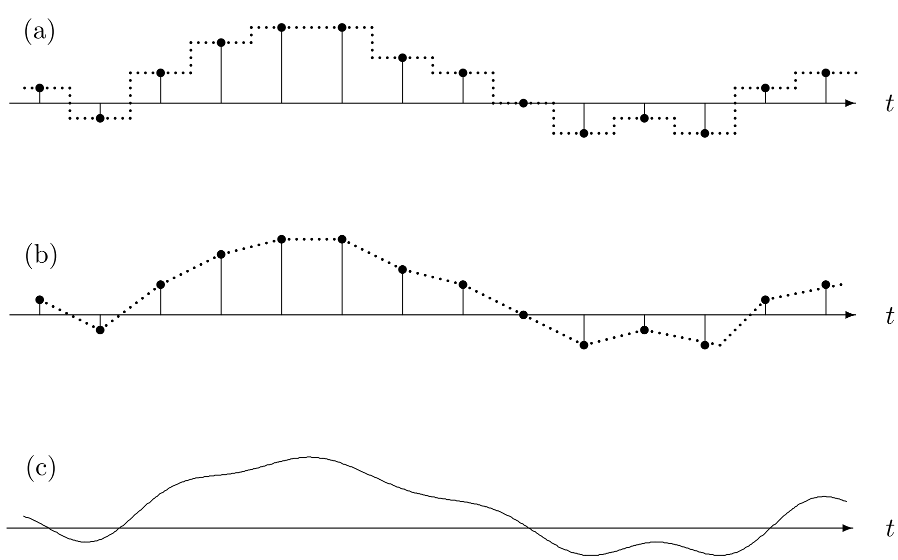
$$\tilde{x}(t) = \sum_{n=-\infty}^{\infty} x[n]p(t - nT_s).$$

where $p(t)$ is some basic interpolation (or reconstruction) pulse. The following three $p(t)$ are important.

(a) zero-order hold

(b) linear interpolation

(c) ideal interpolation



2 Issues in Interpolation

(1) Quality of the interpolated signal

Though we won't emphasize this much, one can use MSE to measure the quality of the interpolated signal, i.e.

$$\text{MSE} = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} \big(x(t) - \tilde{x}(t)\big)^2 \, dt.$$

(2) Good Interpolation Principles

Given a discrete-time signal $s[n]$, a good interpolation method should produce a continuous-time signal $s(t)$ such that

(a) (**Correspondence**) $s(t)$ has $s[n]$ as its samples, i.e.

$$s(nT_s) = s[n], \text{ for each } n.$$

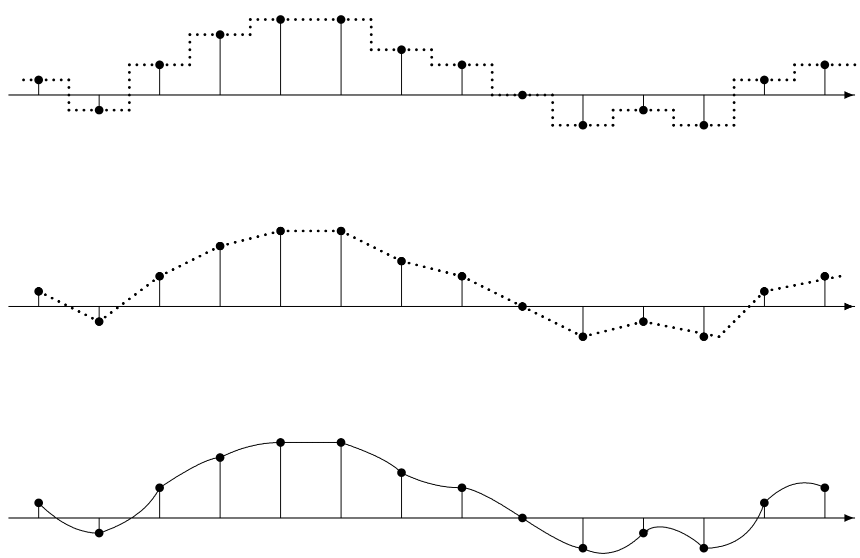
That is, the sampled reconstruction should yield the original samples.

(b) (**Smoothness**) $s(t)$ is as smooth as possible.

The motivation for (b) is a kind of *Occam's razor principle*, i.e. that the simplest explanation for some phenomenon is the best explanation.

Here we assert that *the smoothest and least fluctuating interpolation is the best interpolation*, because it is in some sense the simplest. **More generally, we look for interpolations whose spectrum is concentrated at the lowest possible frequencies**, because interpolations with larger high frequency components will fluctuate more and be less smooth.

Example In the previous figure, one can easily identify the smoothest and least fluctuating interpolation of the three shown. With smooth interpolations in mind, parabolic interpolation is better than linear, which in turn is better than zero-order hold.

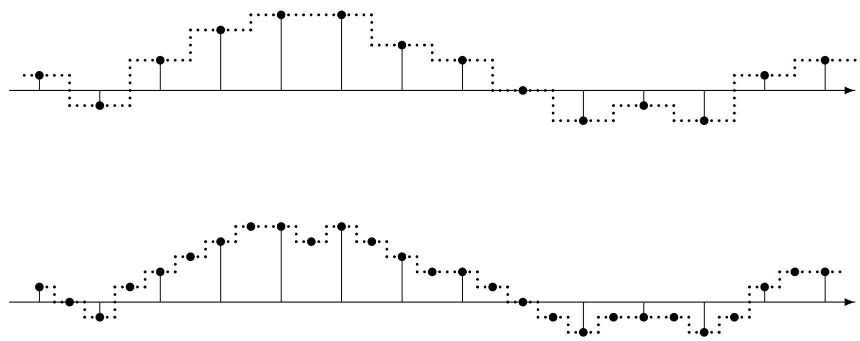


(3) The effect of increasing the sampling rate f_s

- With zero-order hold, linear interpolation, parabolic interpolation and most pulse-type interpolations, it should be intuitive that $\tilde{x}(t)$ becomes a better approximation to $x(t)$ as f_s increases.

  For example,

$$\text{MSE} \to 0 \quad \text{as } f_s \to \infty.$$



  On the other hand, we'd prefer to be able to use as small a sampling rate as possible, because a smaller sampling rate generates fewer samples for us to have to save and/or process. As a result, at some point MSE is sufficiently small and further increases in the sampling rate are not worthwhile, i.e. there is a point of diminishing returns.

- Surprisingly, however, there is one particular choice of $p(t)$ that creates *perfect* interpolations. And, surprisingly, f_s need not grow without bound. Instead it is only required that f_s be larger than twice the frequency of all spectral components of the signal. This remarkable result stems from the sampling theorem.

3 The Sampling Theorem

Textbook: Sections 4.1.2 and 4.5

If the sampling frequency f_s is greater than the twice the frequency of all spectral components of the signal $x(t)$, then

$$x(t) = \tilde{x}(t) \triangleq \sum_{n=-\infty}^{\infty} x[n] p^*(t - nT_s),$$

where $p^*(t)$ is the *sinc* pulse, i.e.,

$$p^*(t) = \frac{\sin(\frac{\pi}{T_s} t)}{\frac{\pi}{T_s} t}.$$

Notes

1. (**Perfect Reconstruction**) This theorem shows that, under appropriate conditions, the signal $x(t)$ equals, without approximation, the interpolation $\tilde{x}(t)$ produced from its samples using the pulse $p^*(t)$.

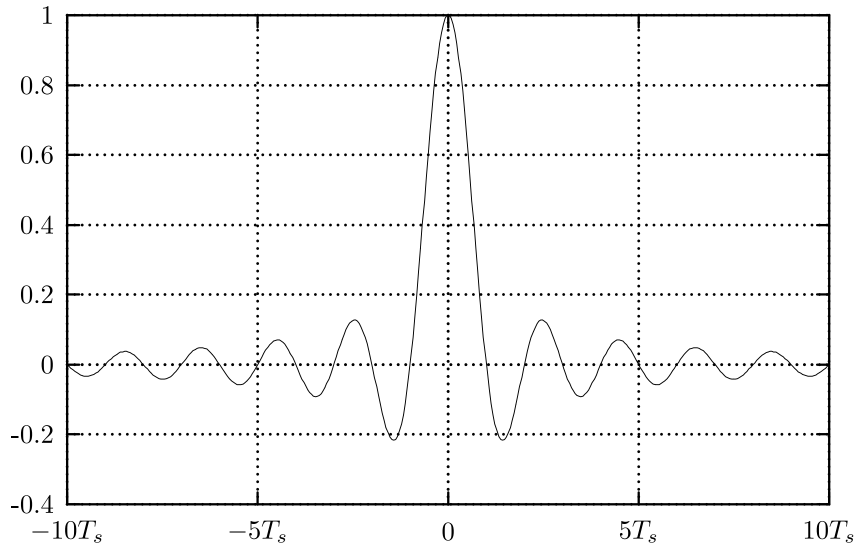
2. (Example) Suppose $x(t) = 2\cos(2\pi 3t + .1) + 2\cos(2\pi 5t + .2)$. Plot the spectrum. The theorem shows that interpolation $\tilde{x}(t)$ from the samples of $x(t)$ equals $x(t)$ if we choose sampling rate $f_s > 2f_{\max} = 2 \times 5 = 10$ samples/sec.

3. (**Sinc** Interpolation Pulse) An interpolator that interpolates using $p^*(t)$ is called an *ideal interpolator*. This particular pulse is often called a "sinc function" or "sinc pulse". Notice that

(a) its value at zero is 1

(b) it has infinite support

(c) it equals zero at times $\pm T_s, \pm 2T_s, \ldots$.

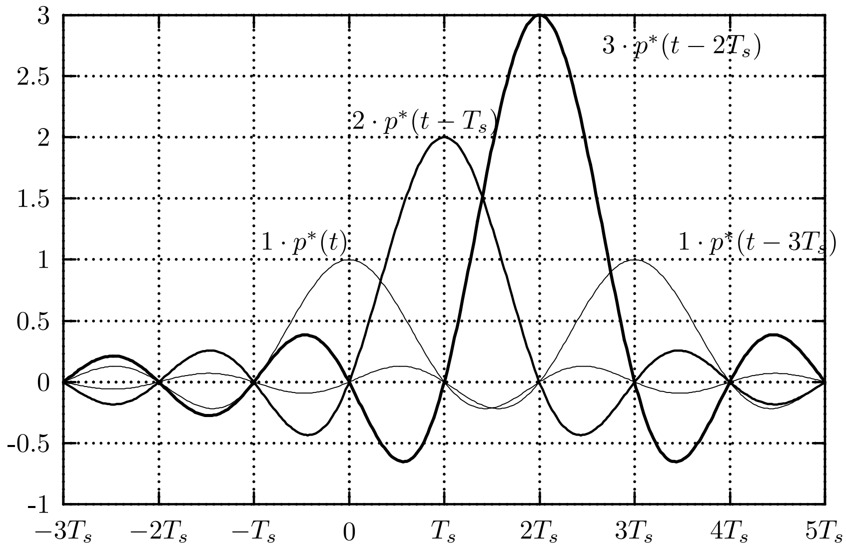


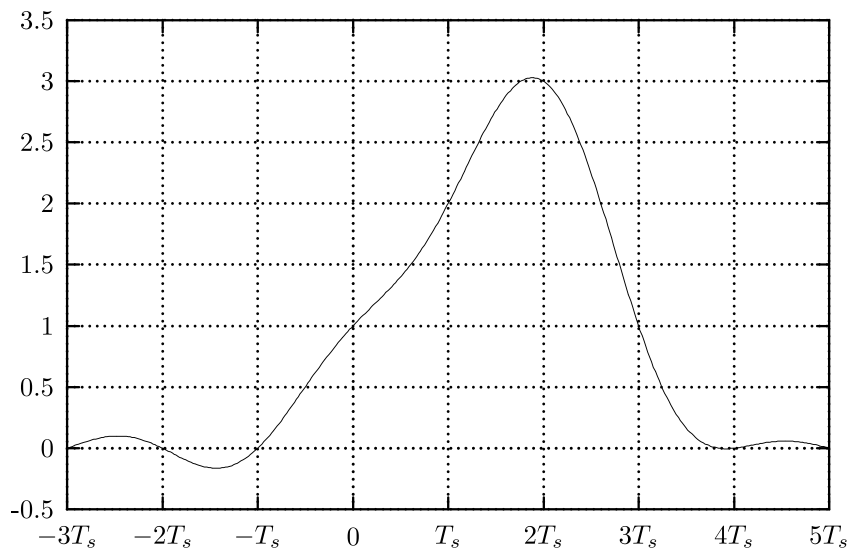
4. (Example) Illustration of the interpolation of a set of samples using the sinc pulse:

$$\tilde{x}(t) = \sum_{n=-\infty}^{\infty} x[n] p^*(t - nT_s),$$

where

$$x[n] = \begin{cases} 1, & n = 0, \\ 2, & n = 1, \\ 3, & n = 2, \\ 1, & n = 3, \\ 0, & \text{else.} \end{cases}$$





5. This is a remarkable and surprising theorem. A complete derivation is beyond the scope of EECS 206, but is included in EECS 306. It requires the frequency domain analysis of aperiodic continuous-time signals via the "continuous-time Fourier transform". Later we'll have just a brief discussion about its derivation. The theorem is often called the *Shannon Sampling Theorem*, after UM alumnus Claude Shannon who used the theorem in his pioneering 1948 paper, which among other things made it widely known to engineers. The earliest versions of the theorem go back 1847.

6. (**Bandlimited Signals**) Let $f_{\max}$ denote the highest frequency of any spectral component of the signal $x(t)$. If $f_{\max} < \infty$, then $x(t)$ is said to be **bandlimited** because its bandwidth is finite. Moreover, we say $x(t)$ **is bandlimited to frequency** $f_{\max}$.

7. (Applicability of the Sampling Theorem)

The Sampling Theorem applies to bandlimited signals, for example a finite sum of sinusoids. It shows that such signals can be perfectly recovered from their samples. Moreover, it indicates that the sampling frequency f_s need not grow without bound to obtain very good interpolations. We need only have

$$f_s > 2f_{\max}$$

or equivalently

$$f_{\max} < \frac{f_s}{2}.$$

The sampling freqeuncy $f_s = 2f_{\max}$ is often called the **Nyquist frequency**.

8. (**Nonbandlimited Signals**) If a signal $x(t)$ has $f_{\max} = \infty$, then it is not bandlimited and the sampling theorem does not apply. For example, a periodic square wave is not bandlimited. We'll briefly discuss sampling nonbandlimited signals later.

9. (Sinc pulses for theory purpose) The fact that the pulse $p^*(t)$ has infinite support can make it difficult to build a system that implements ideal interpolation. For example, whereas zero-order hold and linear interpolation use just one and two samples, respectively, when producing the value of $\tilde{x}(t)$ at any particular time t, the ideal interpolator uses an infinite number of samples.

   In practice, few systems attempt to use ideal interpolation. Instead most use zero-order hold, linear interpolation, or some other simple scheme. Because of this, they generally need to use a sampling rate that is the larger than the Nyquist rate $2f_{\max}$. The ratio $f_s/(2f_{\max})$ is sometimes called "the oversampling ratio".

   If ideal interpolation is not commonly used, what then is the value of the sampling theorem? Its main value is in the understanding that it provides. For example, it tells us that good interpolation is possible only when $f_s/(2f_{\max})$.

10. (Sinc interpolation is the smoothest) It can be shown (e.g. in EECS 306) that the signal produced by the ideal interpolator,

$$\tilde{x}(t) = \sum_{n=-\infty}^{\infty} x[n]p^*(t - nT_s),$$

   is itself bandlimited to frequency $f_s/2$ and that it is the only signal that is bandlimited to frequency $f_s/2$ that passes through the samples. That is, any other interpolation of the samples has components at frequencies greater than $f_s/2$. Thus, s(t) is the "smoothest" possible interpolation of the samples. This is an important property.

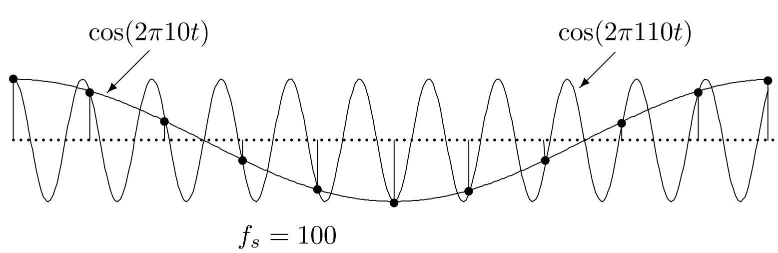
11. (**Importance of $f_s > 2f_{\max}$**) What's so special about $2f_{\max}$? What goes wrong when $f_s < 2f_{\max}$?

Example 1 Consider sampling the signal

$$x_0(t) = \cos(2\pi f_0 t + \phi)$$

with sampling rate f_s such that $f_0 = 1.1 f_s$. Notice that

$$f_s < f_0 = f_{\max} < 2f_{\max}.$$



Notice that the interpolation looks like a sinusoid with a much lower frequency. Let us see what is happening. The sampled signal is

$$\begin{aligned}
x_0[n] &= \cos(2\pi f_0 n T_s + \phi) \\
&= \cos(2\pi(1.1) f_s n T_s + \phi) \\
&= \cos(2\pi(1.1)n + \phi) \quad \because f_s = 1/T_s \\
&= \cos(2\pi(.1)n + \phi) \quad \text{equivalent frequencies } 2\pi(.1) \text{ and } 2\pi(1.1)
\end{aligned}$$

Now, observe that the samples of $x_0(t)$ are *exactly* the same as the samples from another signal $x_1(t) = \cos(2\pi f_1 t + \phi)$ with the much lower frequency $f_1 = 0.1 f_s$:

$$\begin{aligned}
x_1[n] &= \cos(2\pi f_1 n Ts + \phi) \\
&= \cos(2\pi(.1) f_s n T_s + \phi) \\
&= \cos(2\pi(.1)n + \phi) \quad \because f_s = 1/T_s
\end{aligned}$$

Thus we see that sampling $x_0(t)$ and $x_1(t)$ at the given sampling frequency produces $x_0[n]$ and $x_1[n]$ that are identical because they are sinusoids with equivalent frequencies.

Recalling the basic principles of interpolation, we recognize that *any reasonable interpolator will attempt to produce the sinusoid $x_1(t)$ because it fluctuates less.* (Indeed, the sampling theorem indicates that the ideal interpolator would produce $x_1(t)$ exactly, because f_s is more than twice as large as the frequency of all of its components.) Thus, we have the unpleasant situation that one signal $x_0(t)$ is the input to the sampler, but a rather different signal $x_1(t)$ comes out of the interpolator.

Example 2 Consider sampling the signal

$$x_0(t) = \cos(2\pi f_0 t + \phi)$$

with sampling rate f_s such that $f_0 = 0.6 f_s$. Notice that

$$f_s < 2f_0 = 2f_{\max}.$$

Draw the signal, its samples, and the linear interpolation of the samples.

Notice that the interpolation looks like a sinusoid with a lower frequency. Let us see what is happening. The sampled signal is

$$\begin{aligned}
x_0[n] &= \cos(2\pi f_0 n T_s + \phi) \\
&= \cos(2\pi(.6) f_s n T_s + \phi) \\
&= \cos(2\pi(.6)n + \phi) \quad \because f_s = 1/T_s
\end{aligned}$$

Now, observe that the samples of $x_0(t)$ are exactly the same as the samples from another sinusoid $x_1(t) = \cos(2\pi f_1 t - \phi)$ with the lower frequency $f_1 = 0.4 f_s$:

$$\begin{aligned}
x_1[n] &= \cos(2\pi(.4) f_s n T_s - \phi) \\
&= \cos(2\pi(.4)n - \phi) \quad \because f_s = 1/T_s \\
&= \cos(-2\pi(.4)n + \phi) \quad \cos(-\theta) = \cos(\theta) \\
&= \cos(2\pi(.6)n + \phi) \quad \text{equivalent frequencies } -2\pi(.4) \text{ and } 2\pi(0.6)
\end{aligned}$$

We see that sampling $x_0(t)$ and $x_1(t)$ produces $x_0[n]$ and $x_1[n]$ that are identical because they are sinusoids with equivalent frequencies. And as in the previous example, any reasonable interpolator will produce, at least approximately, the sinusoid $x_1(t)$ because it has the lower frequency. (And ideal sampling would produce $x_1(t)$ exactly because f_s is more than twice as large as its frequency.)

12. (**Alias**) Two continuous-time signals that have the same samples, such as in the previous two examples, are said to be **aliases** of each other. **Aliasing** is said to occur when, as in the previous example, a continuous-time signal $x_0(t)$ is sampled, but a very different continuous-time signal $x_1(t)$ is produced, at least approximately, by the interpolator. By "very" different we mean that the difference is not simply due to a crude interpolation, like zero-order hold, but is due to the fact that the interpolator has produced a signal that is bandlimited to a lower frequency than the original signal. It is also common to say that $x_0(t)$ has **aliased** to $x_1(t)$.

13. (Examples of aliases)

    Sinusoids and complex exponentials that are aliases of each other.

    (a) Sinusoids

    $$x_0(t) = \cos(2\pi f_0 t + \phi)$$
    $$x_1(t) = \cos\big(2\pi(f_0 + m f_s)t + \phi\big),$$

    where m is any positive or negative integer.

    (b) Sinusoids

    $$x_0(t) = \cos(2\pi f_0 t + \phi)$$
    $$x_1(t) = \cos\big(2\pi(m f_s - f_0)t - \phi\big),$$

    where m is any positive or negative integer. In this case, it is sometimes said that the frequencies of $x_0(t)$ and $x_1(t)$ are related by **folding**.

    (c) Complex exponentials

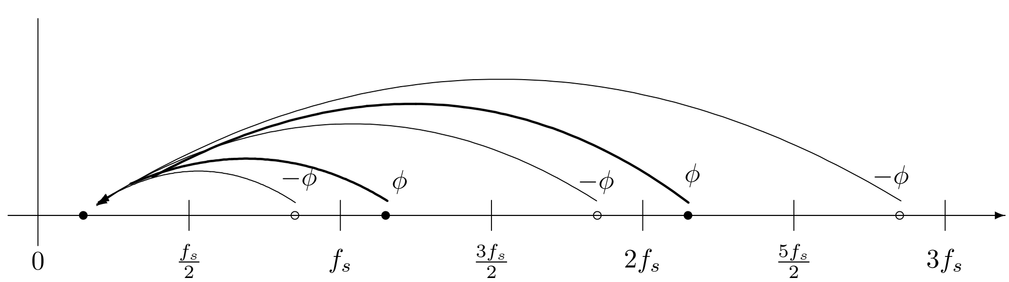
    $$x_0(t) = e^{j(2\pi f_0 t + \phi)}$$
    $$x_1(t) = e^{j\big(2\pi(f_0 + m f_s)t + \phi\big)},$$

    where m is any positive or negative integer.

14. From the above property we may deduce:

    > Any sinusoid $x_0(t)$ with frequency greater than $f_s/2$ has an alias $x_1(t)$ with frequency less than $f_s/2$.

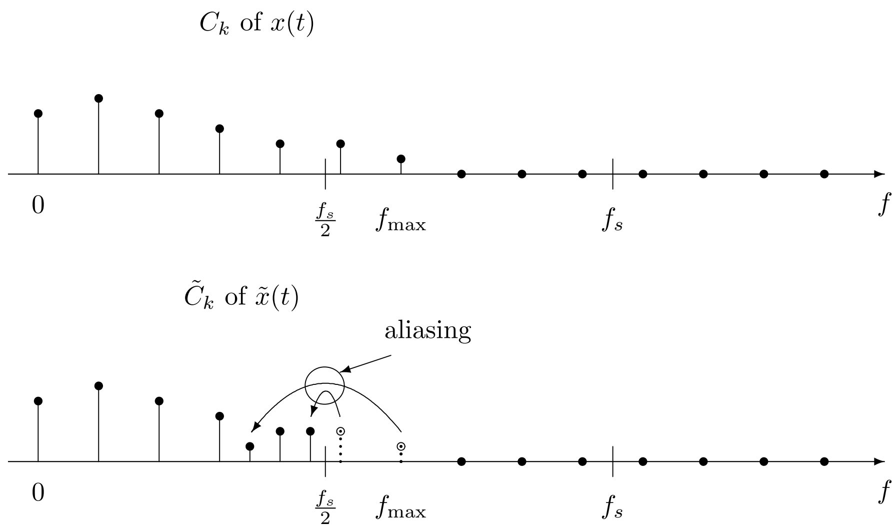
    The following diagram illustrates the higher frequencies that alias to a given frequency less than $f_s/2$. It also indicates whether the aliasing results in the same phase or the negative of the phase (folding).



15. Let us return to the question of what goes wrong when $f_s < 2f_{\max}$ or, equivalently, when $f_{\max} > f_s/2$.

(i) (Sinusoids) We begin by considering sinusoids. From the previous property, any sinusoid $x_0(t)$ with frequency $f_0 = f_{\max} > f_s/2$ has an alias with frequency $f_1 = f_{\max} < f_s/2$, which any reasonable interpolator will produce, at least approximately. This shows clearly what goes wrong.

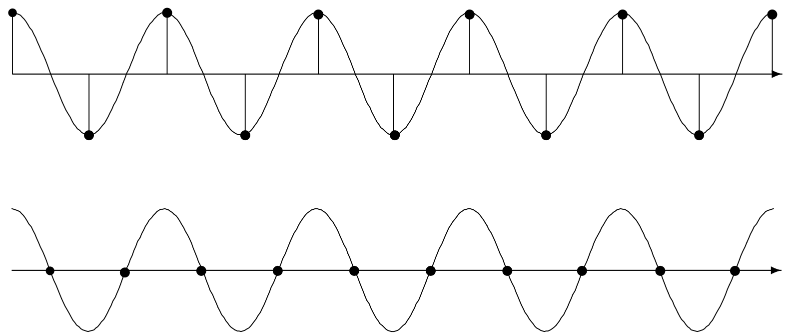
(ii) (Periodic signals) Next, consider an arbitrary periodic signal $x(t)$. By the Fourier series theorem, $x(t)$ is a sum of sinusoidal components. We also note that sampling is a linear operation. Thus, the samples of the periodic signal are simply the sum of the samples of its sinusoidal components. If the signal has $f_{\max} > f_s/2$, then at least one sinusoidal component will suffer aliasing, and consequently the original signal $x(t)$ suffers aliasing. In particular, the interpolator will produce, at least approximately, the sum of the aliased sinusoids, rather than the sum of the original sinusoids.



In summary, if a signal has $f_{\max} > f_s/2$, then the interpolator produces an alias with $f_{\max} < f_s/2$.

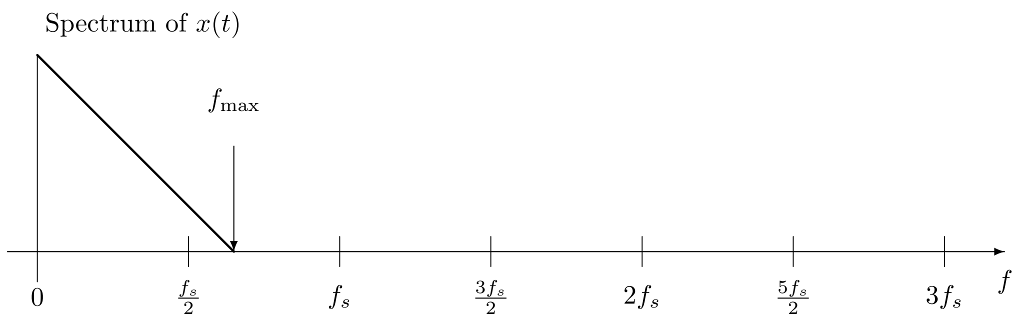
16. Derivation of the sampling theorem: Using the Fourier series theorem as in the previous note, one can argue that if a periodic signal $x(t)$ has $f_{\max} > f_s/2$ (i.e. all spectral components have frequencies less than $f_s/2$), then no other periodic signal with $f_{\max} < f_s/2$ has the same samples. This indicates that when $f_{\max} < f_s/2$, it should be possible to reconstruct the signal from its samples. However, to show that this can be done with the sinc pulse based interpolator requires methods beyond our scope.

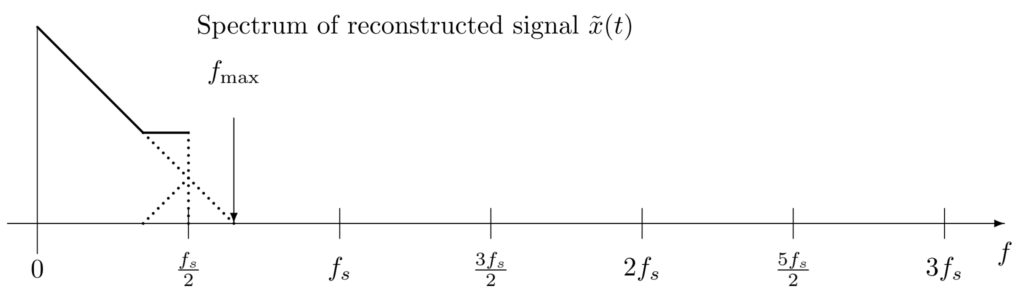
17. (Sampling with the Nyquist rate) What happens if we sample exactly at the Nyquist rate, i.e. with $f_{\max} = f_s/2$. Aliasing might or might not occur. For example consider taking two samples per period from a sinusoid, which means f_s is exactly twice the frequency of the sinusoid. These samples can be taken at the zero crossings, in which case the sinusoid aliases to the all zero signal. Or they can be taken at the peaks, in which case aliasing does not occur. Or they can be taken at other times, in which case the sinusoid has an alias at the same frequency but a different phase.



18. What happens if a signal is bandlimited and we sample at too low a frequency?

    In the case of a periodic signal and ideal interpolation, sampling and interpolation results properly reconstructs all spectral components at frequencies less than $f_s/2$, but it aliases all spectral components at frequencies greater than $f_s/2$ to frequencies less than $f_s/2$.



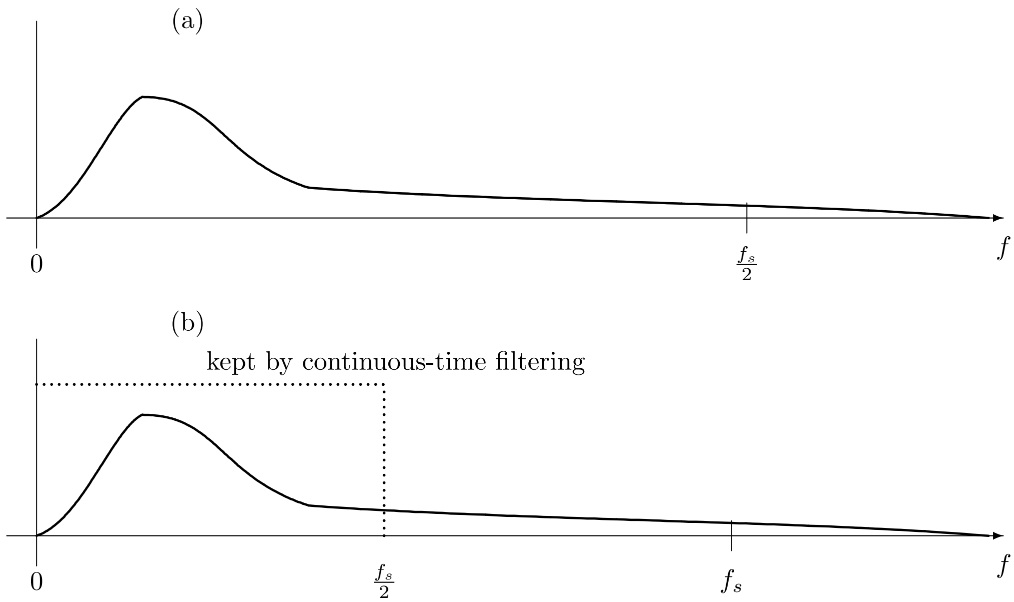


19. How to sample a signal that is not bandlimited?

    There's no perfect way. But one must pick a sampling rate f_s. All spectral components at frequencies above $f_s/2$ will alias to frequencies below $f_s/2$.

    (a) If possible, one chooses f_s so large that the frequency components above $f_s/2$ are very small.

    (b) If possible, one precedes the sampler with a "continuous-time filter" that eliminates all frequency components above $f_s/2$. This reduces the interpolation MSE by approximately a factor of two. Continuous-time filters are discussed in EECS 306.



20. Sampling and interpolation for the signal recovery task.

In the signal recovery task, the signal $x(t) = s(t) + n(t)$ is sampled with the goal of eventually producing an approximation $\tilde{s}(t)$ to the desired part of the signal, namely, $s(t)$. Though we are not trying to reconstruct $x(t)$, it makes sense to sample it at a rate greater than $2f_{\max}$ for $x(t)$, because then the samples contain all the information in $x(t)$. (From the samples one could reconstruct $x(t)$.) It is not essential that one sample at a frequency significantly greater than $2f_{\max}$, but in some cases, this may simplify the processing that must be performed. The digital-to-analog converter, which is the last step of the signal recovery system, is not actually reconstructing a signal from the samples of the signal, rather it is constructing a signal $\tilde{s}(t)$ from samples $\tilde{s}[n]$ created by a digital processor. The interpolation used by the digital-to-analog converter could be zero-order hold, linear interpolation, ideal interpolation, or some other form of interpolation. If ideal interpolation is chosen, then $\tilde{s}(t)$ is determined by

$$\tilde{s}(t) = \sum_{n=-\infty}^{\infty} \tilde{s}[n] p^*(t - nT_s).$$

There is also, sometimes, a shortcut to finding $\tilde{s}(t)$. If $\tilde{s}[n]$ happens to be a sinusoid, e.g.

$$\tilde{s}[n] = A\cos(\hat{\omega}n + \phi), \quad 0 \le \hat{\omega} \le \pi,$$

then we know that the ideal interpolator will produce the unique continuous-time signal that is bandlimited to frequency $f_s/2$ and has $\tilde{s}[n]$ as its samples. What is this signal? It is easy to see by inspection that following signal has these two properties:

$$A\cos(\hat{\omega} f_s t + \phi).$$

Thus it must be the output $\tilde{s}(t)$ of the ideal interpolator.

This method can also be applied to a signal that is a sum of sinusoids, by applying it separately to each sinusoidal component. It also applies to arbitrary periodic signals, because by the DFT Theorem, any periodic signal is the sum of sinusoids.

21. Sampling for signal detection

In the signal detection problem, we sample the signal $x(t) = s(t) + n(t)$ with the goal of a making a decision about $s(t)$ based on the samples. The system does not output a continuous-time signal. As in the signal recovery problem, it makes sense to sample $x(t)$ at a rate greater than $2f_{\max}$, because in this case, the samples contain all the information that was originally in $x(t)$. Sometimes sampling at a significantly higher rate simplifies the processing that must be done.