What is “Bandwidth”

James D (jj) Johnston

Immersion.net

Yeah? Who cares?

- Well, everybody does, but lots of people don’t know that (in the audio industry).

- Digital equipment operates with a given maximum bandwidth, given in PCM systems by half of the operating sample rate. (maximum bandwidth < half of the operating sample rate)

- Analog systems have a limited bandwidth, as well, but the form of limiting can be very different.

Some questions I will answer today:

- But what happens if the sampling time completely misses an input pulse?

- What is the actual frequency content of a tone burst?
  - What **IS*** the actual bandwidth of that 20kHz tone burst of one cycle?
  - How many cycles does it take to make a “frequency” into a “pure frequency”?
  - What does “pure frequency” even mean in the real world?

Let’s talk about the math, as well

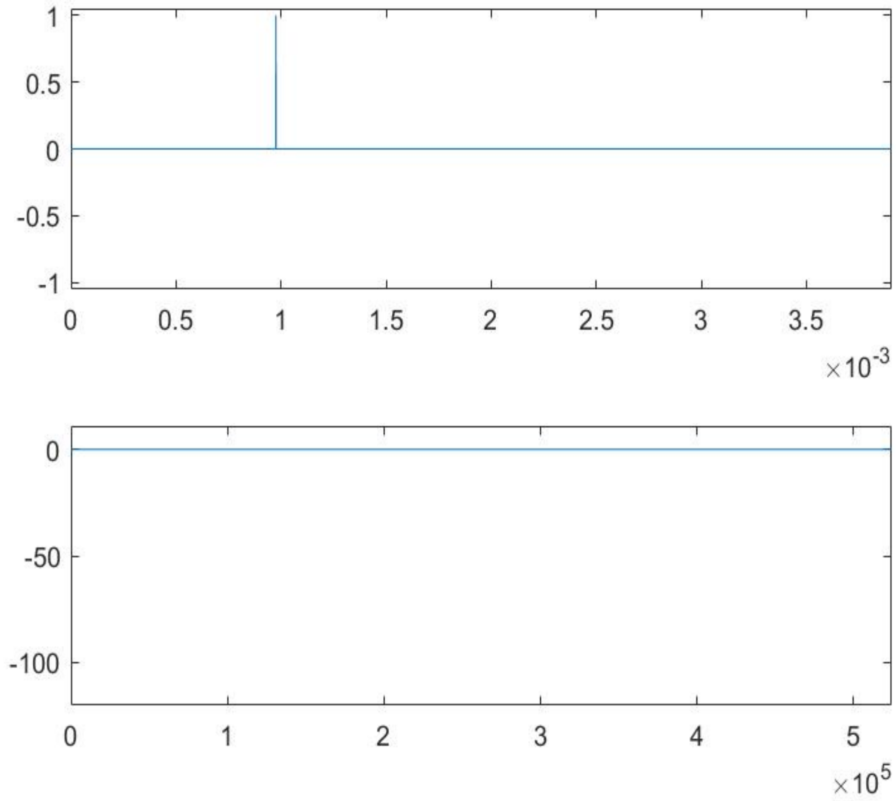
- Strictly speaking all signals with a finite duration have an infinite bandwidth. Sometimes a great deal is made of this fact.
  - In reality, it is simple to ensure that the “length” of a signal above the noise floor (in either analog or digital) is limited in length.
  - In analog, this happens naturally due to a combination of physics and the properties of things like tubes, transistors, transformers, tape recorders, microphones, etc., all of which have limited bandwidth.
  - The “infinite impulse response” does not mean that the level of the “impulse” stays about the noise floor, in fact physics ensures that signals are soon covered by the noise floor. Additionally the noise floor prevents “perfect detection” of the onset of a signal.
  - In sampled systems, the system bandwidth and noise floor determine the time resolution, JUST LIKE IN ANALOG SYSTEMS. The shape of the system bandwidth can be quite different, but the same rules apply.

Let’s show some pictures.

- The next slide will show pairs of time domain signals and their spectrum. To make this easy to relate to audio systems, all of the spectra are plotted from peak down to -120 dB.
- The top plot in a pair is time domain
- The bottom plot is the amplitude spectrum of the top signal, expressed in dB relative to full scale.
- Examples of both analog and digital signals will be plotted.

- ATTENTION: You must pay attention to the horizontal scale (frequency or time) in each plot. They will vary substantially!

The spectrum of an impulse.



This is the canonical “broadband” signal.

While it is **physically impossible** to create,
it is effectively perfect broadband for any realizable acoustic stimulus that is transmitted via air.

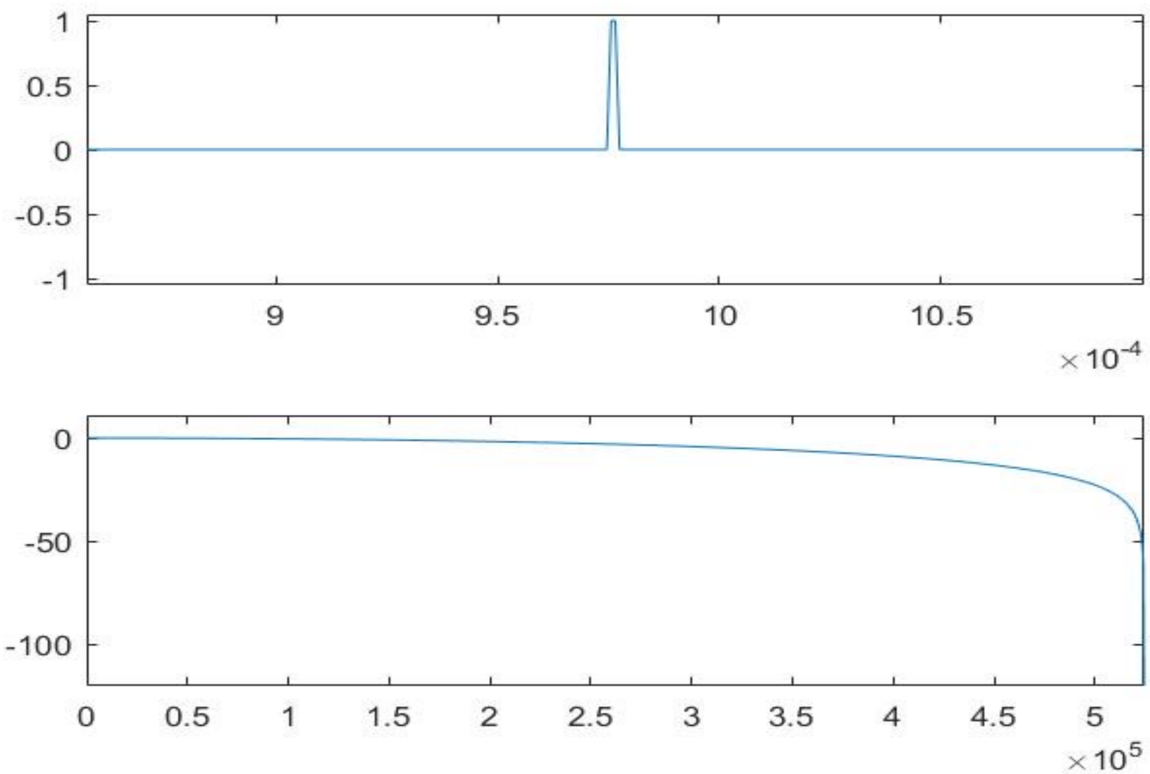
As expected the value of the spectrum is ‘1’ at all frequencies.

Note. For all graphs, the upper axis is time in seconds (horizontal) vs. amplitude (vertical) with ‘1’ being peak amplitude.
The lower axis is frequency (Hz) vs. dB re full scale.

So?

- If the “pulse” that falls completely between samples happens, it has enormous out-of-band energy. It violates the very basics of sampling theory.
- We will see, shortly, that there is a limit to the width of a signal that can be correctly sampled.
- Once that pulse that you think is “missed” is filtered to inside the system bandwidth, it’s no longer smaller than one sample wide!

And now, a .93 microsecond pulse.



As you may note, it's no longer flat.

But that zero in the spectrum is deceptive, because as we will see shortly, it goes back up after that drop at our fs=1024^2 / 2 Nyquist rate..

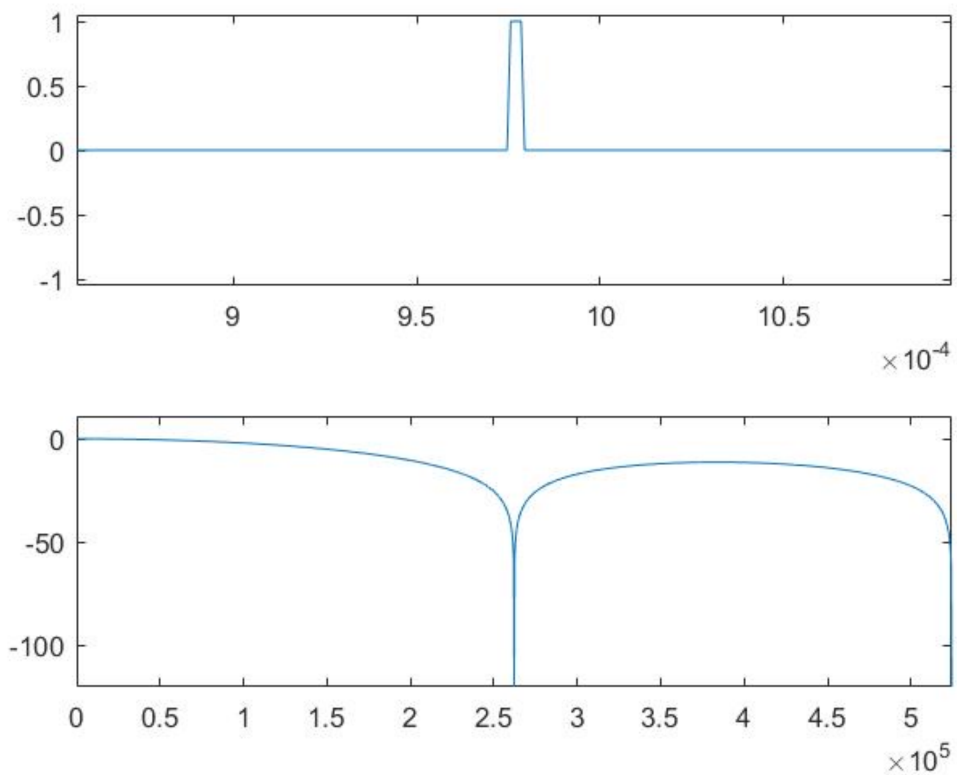
What does this show us?

- In order for something to even possibly be inside the system bandwidth ***it must be at least 2 samples wide.***

- This shows the absolute minimum that a signal can have, EVER, and be ‘in band’ to any limit whatsoever.

- Effectively at the sample rate of the graph, this is a ‘2 sample’ pulse.

1.4 microsecond pulse (4 samples at 1,057,576 sampling rate)



Now, you see how rolling off to a zero does not mean that the spectrum is done.

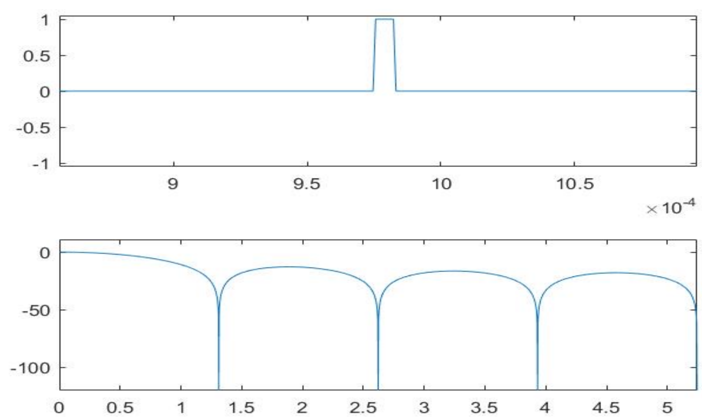
In fact it can never be "done" for a time-limited signal.

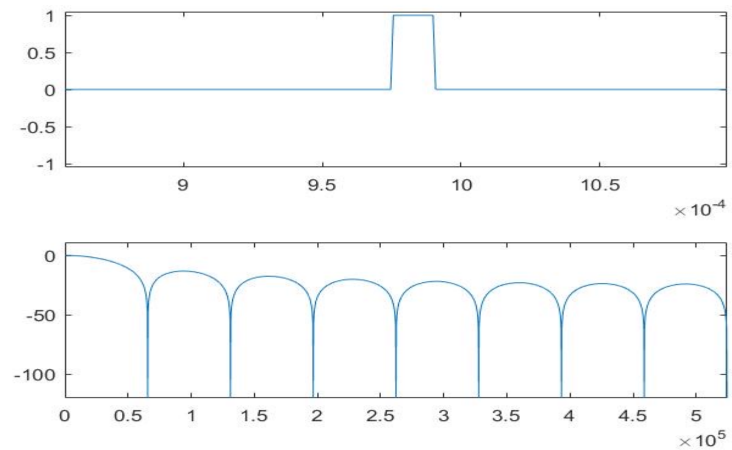
Now we will put up 8, 16, 32 and 64 length pulses.

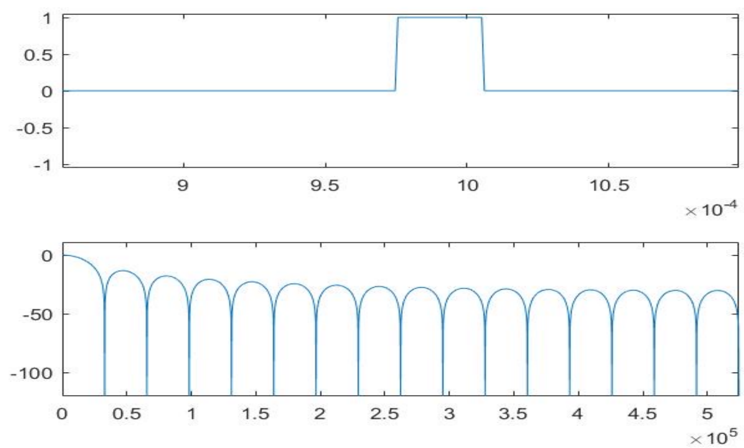
- Now the out of band energy rolls off a little bit. So you get a FEW dB of accurate sampling out of this “system”.

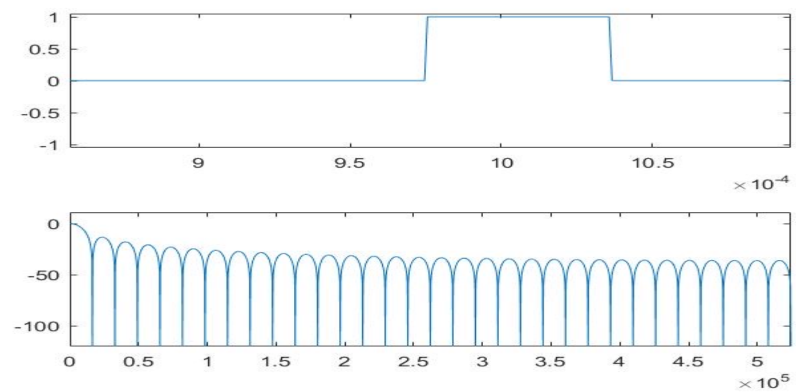
- Making the pulse wider isn’t the way to do this, but we shall see momentarily what a better answer is.

Making that pulse longer, longer, and longer.





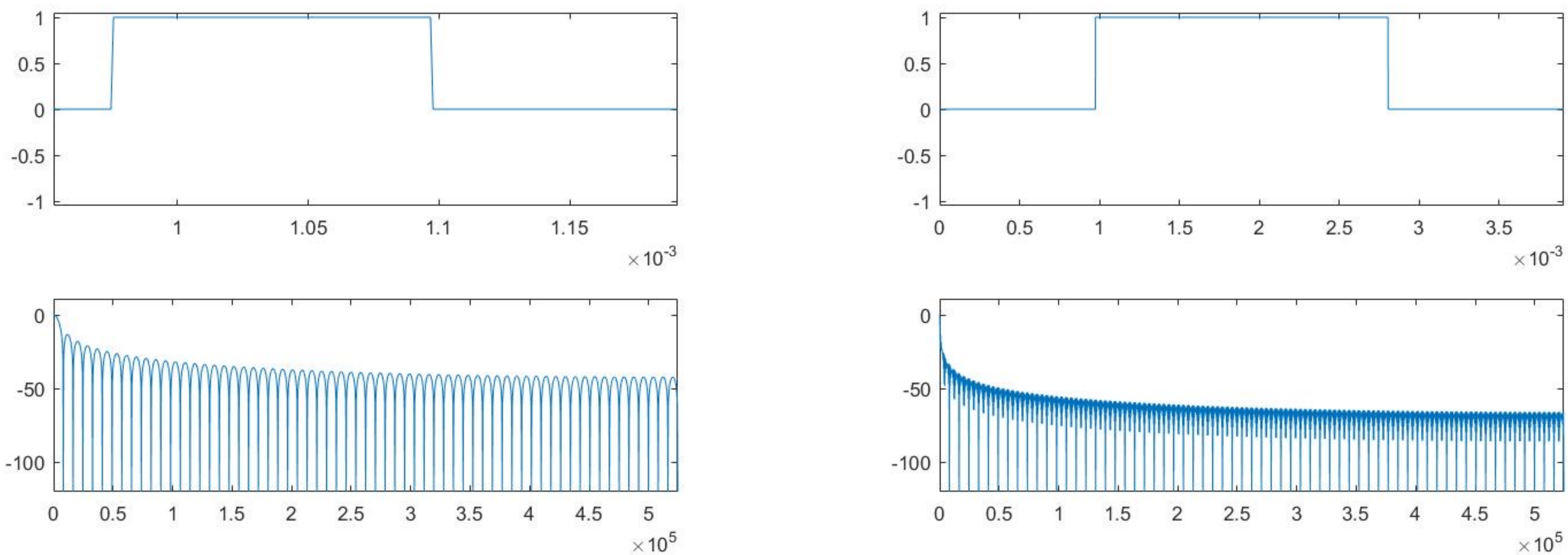




Ok, the “bumps” are getting smaller and smaller, but

- It’s not getting very low, or very fast. What’s the problem?
  - 1) It has sharp edges
  - 2) Sharp edges mean (literally) that the spectrum can roll off slowly, and only slowly. It’s not “the law,” it’s mathematics
  - 3) Let’s keep doing that for a minute
  - 4) But remember “**sharp edges = wide spectrum**” when we talk, later, about DAC outputs
  - 5) For now, we’ll go much, much longer.

“A lot longer”
(please note the change in scale in the time domain!)



So, the “high frequencies keep getting smaller”

- But, not very fast. Again, it’s those ‘edges.’ There is a better way, eliminating “edges” that have a discontinuity.
- What’s this “discontinuity?”
- For these purposes discontinuity is when the continuous time domain signal changes instantly, or the derivative goes to infinity. (This is not a complete definition for the math aware among us!)
- This pulse has a discontinuity in the derivative. That means that the spectrum falls off at a 6dB rate per octave. Not very fast. If there’s a discontinuity in the second derivative, it rolls off at 12dB/octave, and so on.

Derivative, you say?

- I am referring to the rate of change of the signal over time. A very large change in the rate of change that occurs suddenly is effectively the same as an instantaneous change in an analog waveform.

- For instance, a single sample spike (an impulse in the analog sense) has a large positive change followed immediately by an equal negative change). This is called a "doublet" and results in a flat spectrum. Note the doublet is in the derivative!

- A pulse has a sudden positive (or negative) change, followed by little or no change, then by a sudden change in the other direction. This is, in terms of this discussion, a discontinuous derivative, but it is not a doublet. This rolls off in the analog sense at 6dB/octave, and similarly up to half the sample rate In a digital system.

Some more handy rules. If there is:

- One sample peak (unit impulse) in the signal, doublet in the derivative, then the spectrum is flat

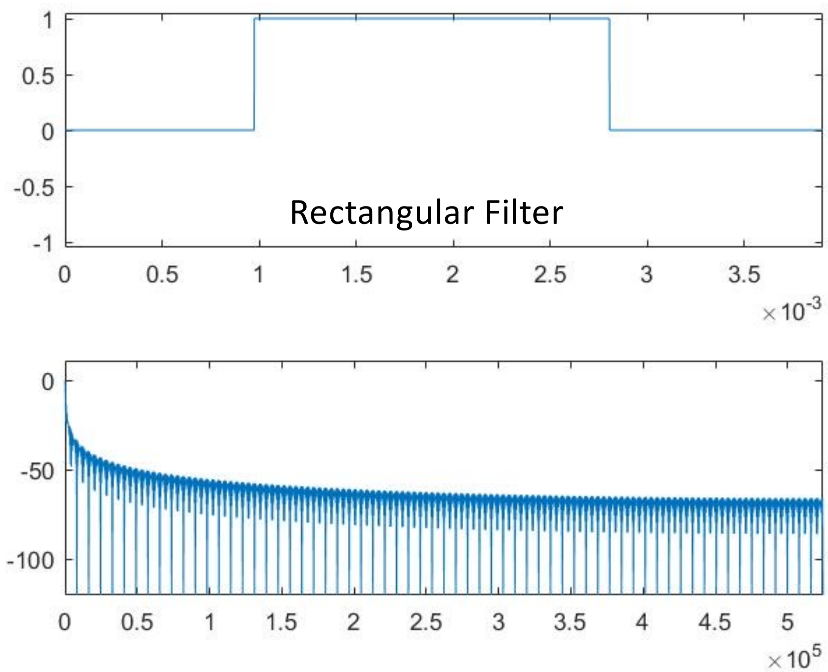
- A unit impulse in the derivative, a jump in the signal, then there is 6dB/octave rolloff about the center frequency

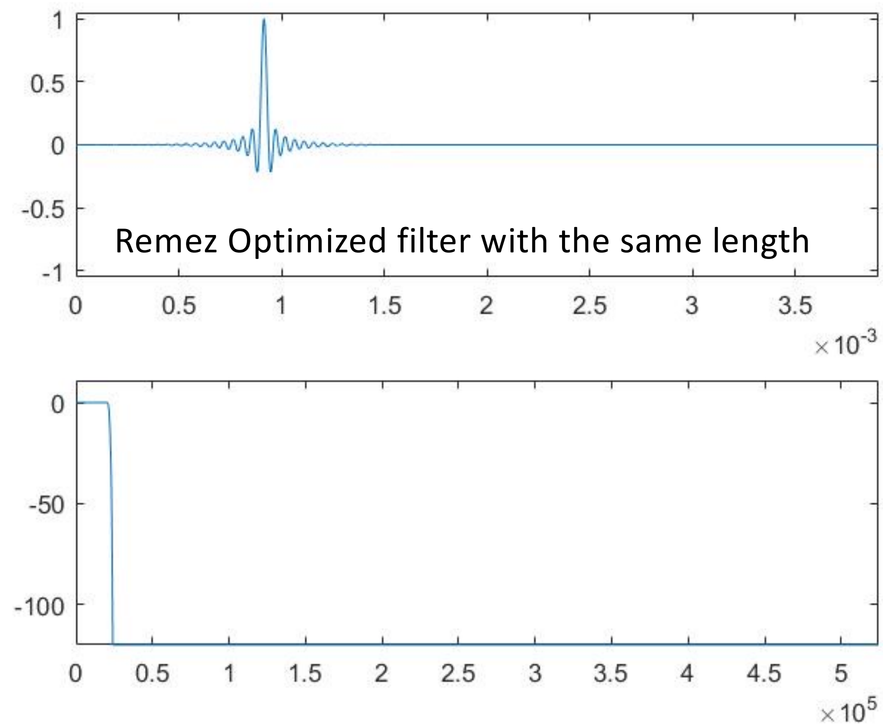
- Unit impulse in the second derivative, square wave in the first, and triangle wave in the signal, then there is 12dB/octave rolloff about the center frequency.

- Other signals may not may not follow similar rules of thumb, depending on the various parts that are effectively convolved together. In general, the lowest order discontinuity found will control the final roll off, but there are ways to trade this off for other properties.

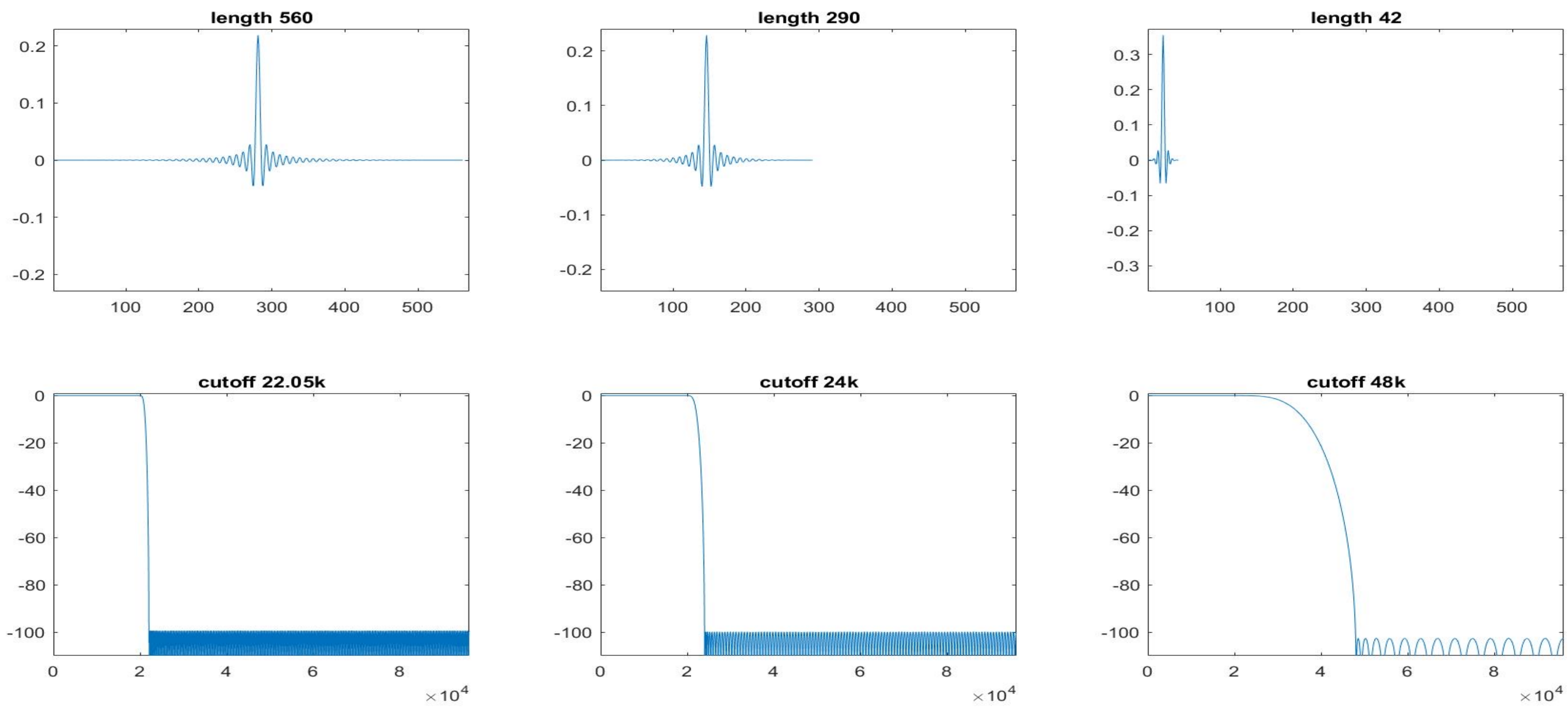
Ok, let's do something different, control those edges!

These two waveforms have EXACTLY THE SAME NON-ZERO LENGTH





Let's put that another way. Antialaising filters for 44, 48, and 96kHz, sampled at 192kHz



So, what happened?

- Simply put, the size of the discontinuity was reduced until it was below -120dB maximum.
- By using a particular kind of filter design, it is also made “flat” in frequency, so the peaks are ‘equiripple,’ each peak in the pass band AND the stop band are identical. In this plot, you can’t see either.

- The point? For audio, this takes you from the threshold of discomfort (way too loud) to below the actual noise of the atmosphere at your ear.

- The length of the filter 2 slides ago was chosen precisely for converting from 1,048,576Hz sampling rate to 48kHz sampling rate, by the way.

Something to consider:

- The width of the center of that filter is just about 42 or 43 samples.

- The length of two samples at 48kHz is around 43 samples at this higher sampling frequency. Again, that minimum length of 2 samples shows up.

- The rest of the samples in the filter improve the out-of-band rejection and in-band ripple of the filter. Note, also how much wider the ***PASSBAND*** of the optimum filter is.

- Again, there is no such thing as a signal under 2 samples long that fits in a sampled system bandwidth.

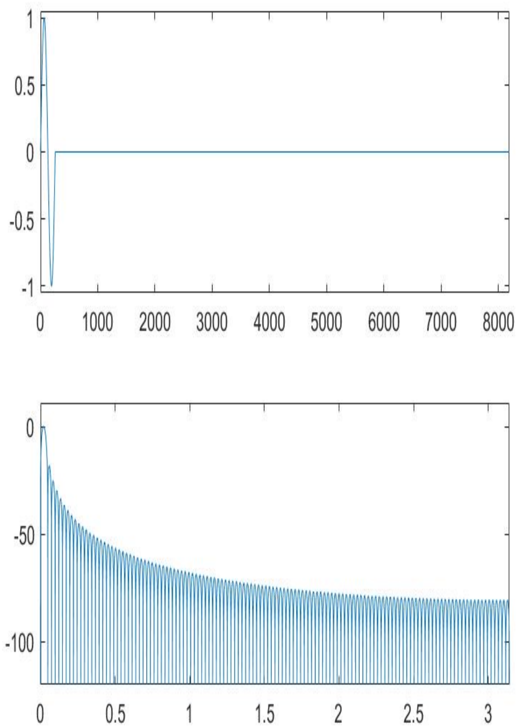
- This is not a “sinc” filter, please note. A sinc filter has a frequency discontinuity at its cutoff frequency, and is therefore infinitely long until windowed. It is not, generally, the most efficient thing to do for standard sampling systems. Any “sinc” filter will have its worst response around the change in amplitude.

What about tone bursts?

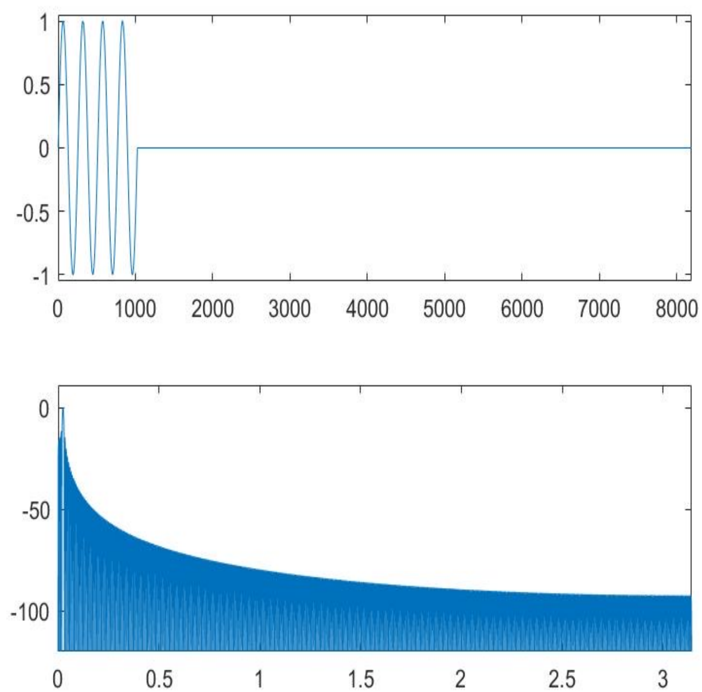
- The next page shows the spectrum of a tone burst of 1, 4, 16 samples inside a long window.
- As you will see, no, it’s not a pure tone by any means at that length. Note, sampling frequency does not matter here, this scales to any frequency you desire.

Examples of tone bursts

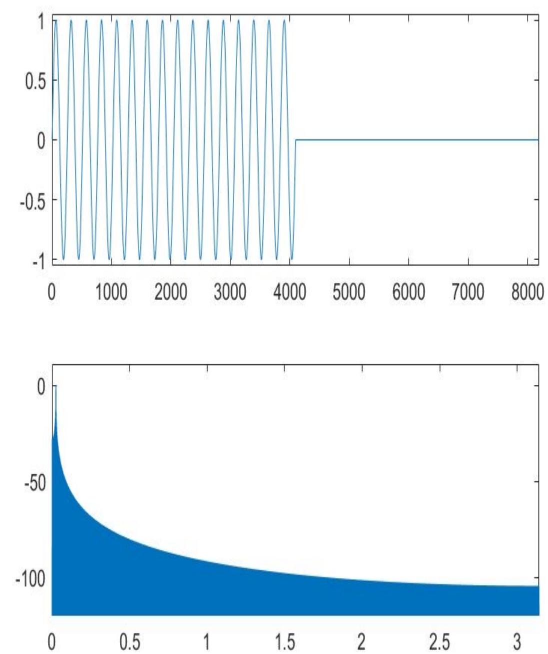
One Cycle



Four Cycles



16 Cycles



What? You can't get that one cycle 20kHz burst through your 48kHz PCM rig?

- No, no, you can't.

- Why?
  - First, it has a substantial energy at 500kHz
  - You CAN get the part that remains under 20kHz into your system
  - In fact, that's ***all*** you should ever get.

My point?

- While any finite length signal is THEORETICALLY “infinite bandwidth,” there are two practical constraints:
- 1) You can’t REALLY make a discontinuous waveform, where the first derivative is infinite. Physics does not approve. (take your choice, infinite energy, infinite force, infinite propagation velocity, or some other infinity, good luck with that!)
- 2) Even when you can (digitally) have such discontinuities, their effect can be managed to a level that physics can not accommodate.
- Therefore, yes, bandwidth CAN be limited.

Ways to limit bandwidth:

- Bandwidth is limited by filtering.
  - Filtering is simply the action of combining the current signal with past parts of the signal.
  - That's what all linear filters do.

- About filters, there are several kinds:
  - There are finite impulse response filters (FIR filters) (like the ones shown previously, yes, the "bad" filters are still filters, even though they are not necessarily useful)
  - There are infinite impulse response filters (IIR filters) that use recursive processes to filter, they also "tail off to nothing" which is to say well below any physically recoverable level. This is true for either analog (system noise) or digital (calculation noise).

A note on FIR vs. IIR filters

- In order for a filter to have “constant delay” (meaning the time for a given frequency to travel through a filter is the same for all frequencies), it must be either
  - Symmetric about the center of its impulse or
  - Antisymmetric about the center of the impulse response AND have zero response at DC
- So, IIR filters have to have phase shift as compared to a constant delay.
- Remember, when phase shift (phi) obeys this form:
  phi=2*pi*f*t, where f is the frequency in question, and t is the time delay in seconds
- THAT means that the filter is a pure delay.
- Any deviation from that is actual phase shift.
- An IIR filter is just another way of implementing a filter impulse response, no more or less. Sometimes it’s very handy, sometimes it’s quite annoying.

Wait? Why is that?

- The only way to get a symmetric IIR filter is to have unstable filters.
- Yes, mathematically you can “do that.”
- No, in practice you do not have infinite mantissa resolution in your arithmetic.
- It is possible to get “mostly pure delay” with a great lot of work, a lot of all pass filters, and a lot of attention to effects of arithmetic accuracy.
- It’s usually not worth it. CPU is cheap nowadays.

Some words on filters:

- This is an extensive subject itself.

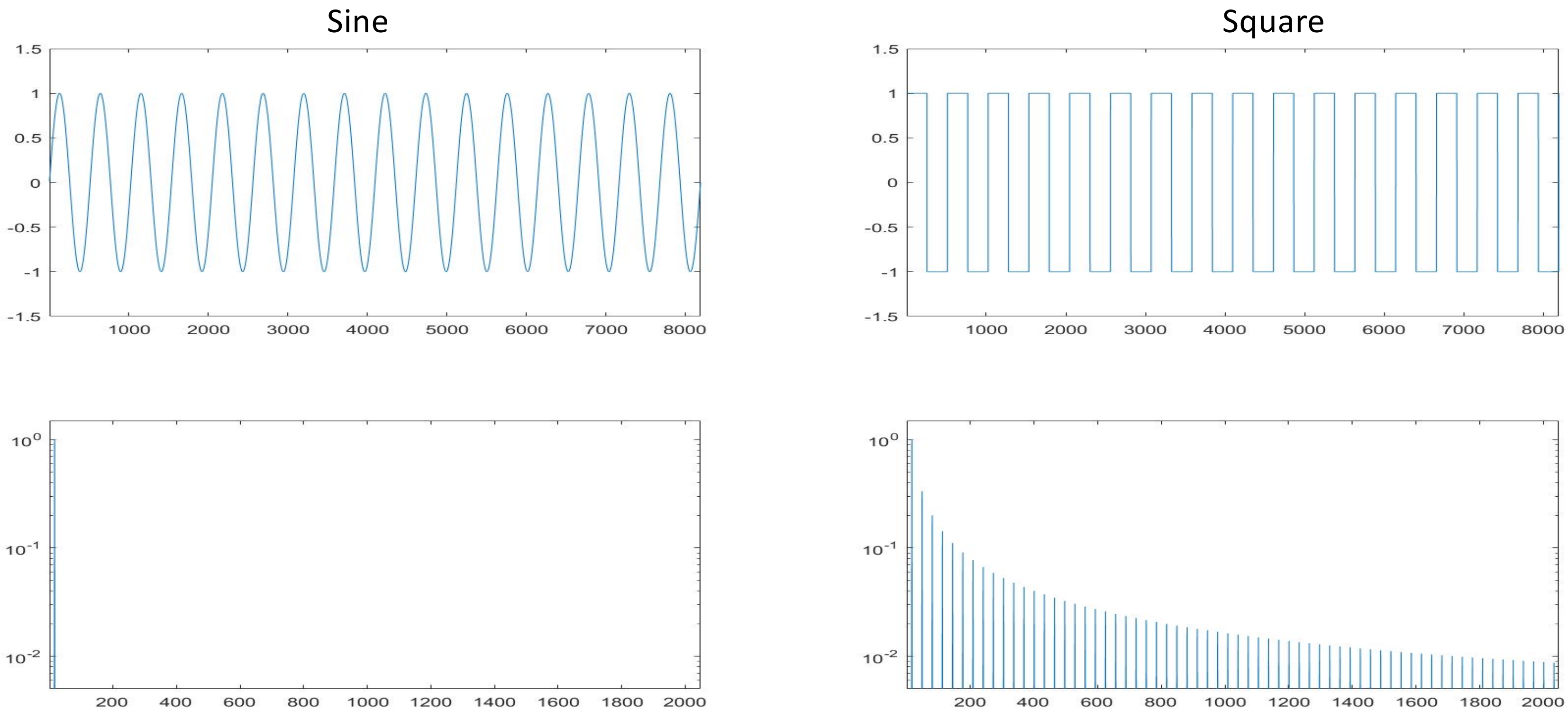
- Filter design tutorial was done at the AES many years ago
- Filter basics was the least attended tutorial EVER for the PNW Section. We drank a lot of coffee!
- It is extensively mathematical.
- There is a lot of established practice.
- There are still unsolved problems.

Why do we limit bandwidth on reconstruction?

- To avoid frequency imaging in capture of analog signals for PCM conversion.
  - Remember, as always, there are two steps, sampling (which is where bandwidth is involved) and quantization (which is an independent process that we may discuss some other day).

- To avoid time confusion on RECONSTRUCTION of PCM signals, and to avoid confusion about “those edges” vs. “time resolution” (which are two different things.

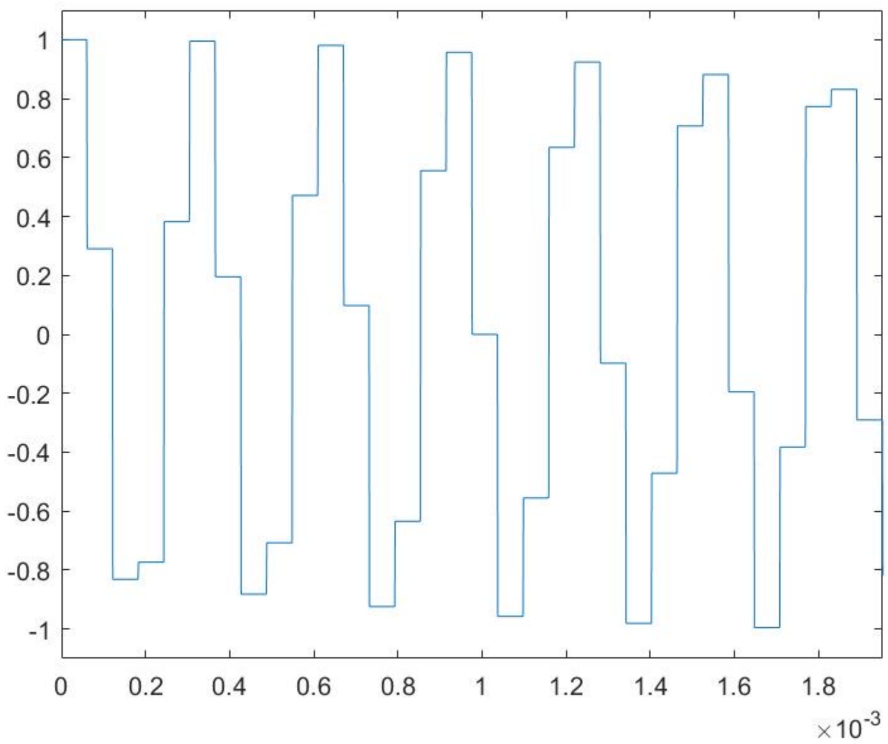
First, let's remind you of a square wave, and its spectrum:

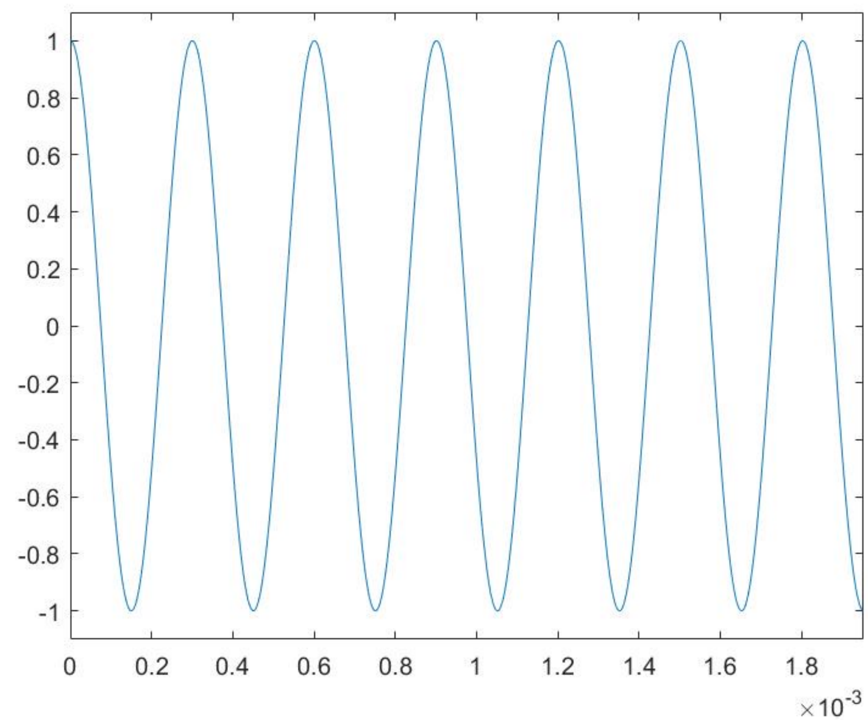


What does a properly reconstructed signal look like?

- First, since it is supposed to be bandwidth limited, there can be no near-discontinuities! Nothing even remotely like a square wave need apply.
- It contains no out-of-band signals!
- Hence, some examples
  - I’ll put up a couple of waveforms and take votes, is this “limited bandwidth”
  - Is it NOT limited bandwidth.
  - The quiz will be at midnight! (KIDDING!)

Here are your choices:





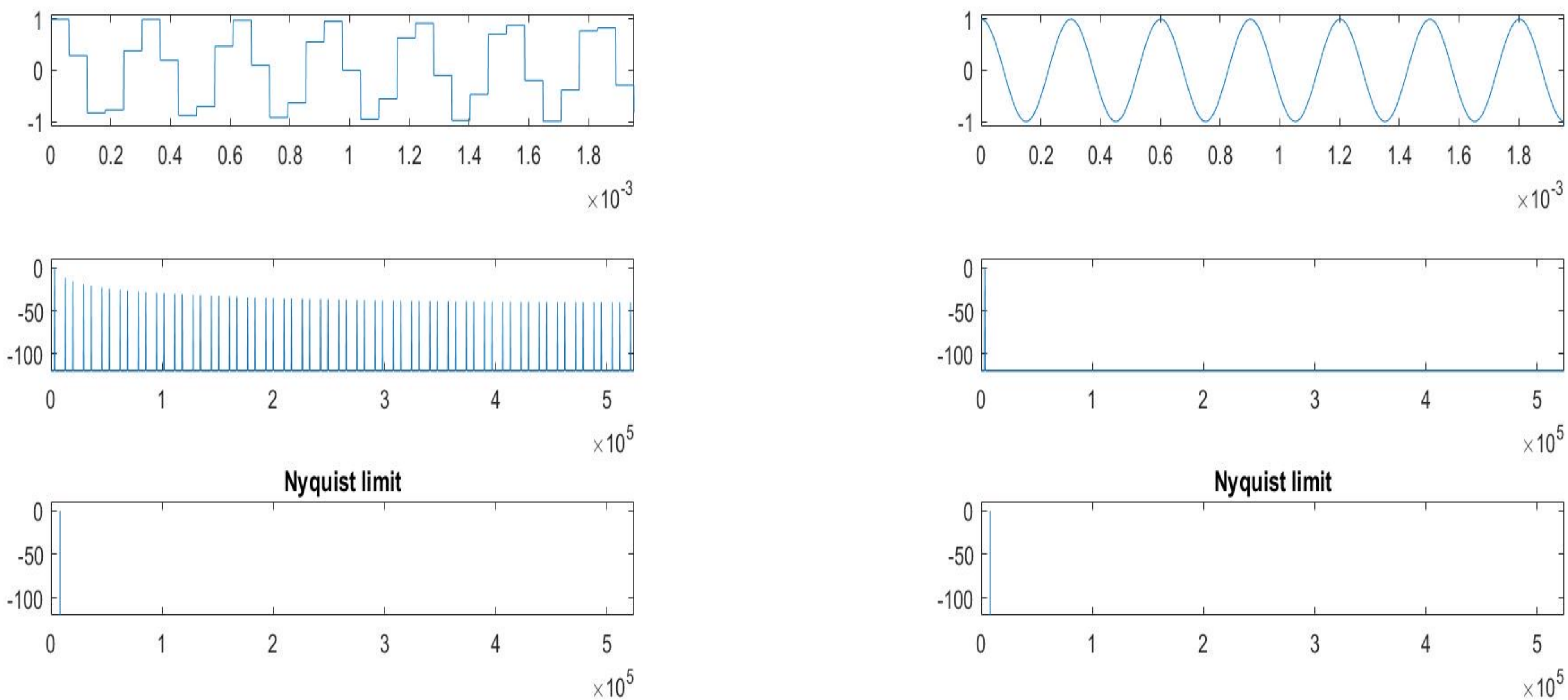
How about it, now?

- The left is what many people imagine comes out of a DAC.

- The right is a sine (well, cosine) wave.

- So, is one, both, or none of these within fs/2?

Now the spectra as well.



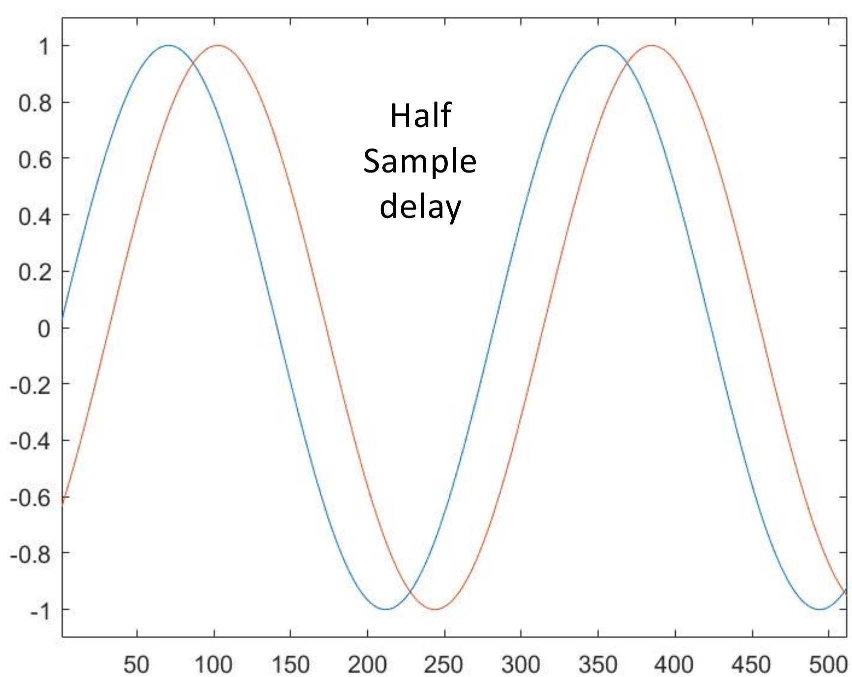
All that “squareness”

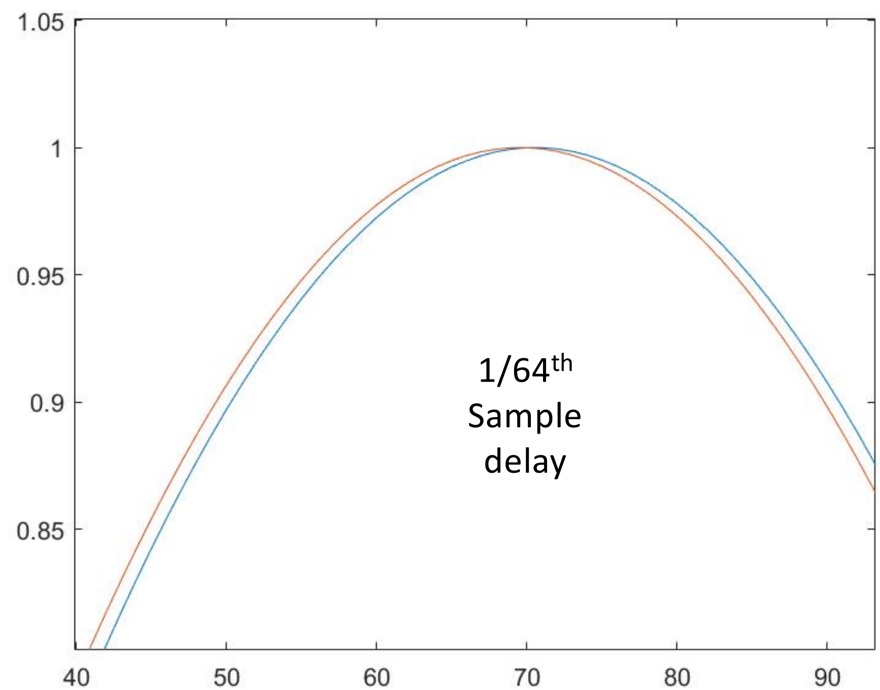
- While it’s not harmonics, like a square wave, it’s all due to out of band content.
- You really shouldn’t see that, no, no, you shouldn’t.
- Unfortunately many programs plot a digital signal (which is, mathematically, a series of impulses) as something kind of squared off.
- That’s easy to plot, though, but don’t let it fool you.

Discontinuities

- Yes, those “steps” from the DAC are evidence of a faulty process.
- A DAC must, absolutely, positively, also have an anti-imaging filter.
- The sine wave in the second graph is nothing more than the FILTERED, yes FILTERED stepped data from the first, with filter delay (constant delay filter) removed.
- All those “stairsteps” are ***OUT OF BAND*** components of the sampling process.
- If you see those stairsteps, you need an anti-imaging filter on your DAC.

Just for grins, let's demonstrate "sub sample time resolution"





Yes, those are actual output waveforms.

- The time resolution of Redbook, 48k Sampled audio, 96kHz sampled signal, if they are quantized to 16 bits, is:

  1/(2 * pi * bandwidth * 2^16)

- Yes, the time resolution is a function of two things:
  - 1) Bandwidth
  - 2) Level resolution
  - No, I did NOT include sampling rate, I said ***BANDWIDTH***

- Something to think about there, isn't there?

What’s the point?

- As is shown by the plots on the previous page, the time resolution of PCM is much, much smaller than one sample period.
- Claims otherwise based on “step functions” are a fundamental misunderstanding.

- It’s time for that myth to go away and stay away.

In summary:

- 1) Bandwidth limiting is required for conventional sampling.
- 2) No signal shorter than 2 samples has any chance whatsoever of being "bandlimited" properly. Even then, the amount of bandlimiting is very, very poor.
- 3) Nothing can "happen between samples." In the analog domain, you broke the rules of sampling. In the digital domain THERE IS NO "between samples" (although you can easily calculate half-sample delays, etc., if one wishes).
- "When does a tone burst become a "pure tone"? Strictly speaking never, but the longer it lasts the closer it gets. (But see previous talk by Bob Smith and I on "Windowing,")
- Bob Smith will now talk about what some commercial software will show you.