Why Zero Padding Doesn't Increase Frequency Resolution

Hassan Islam B.Eng in Electrical Engineering

1. INTRODUCTION

Frequency resolution measures the ability to discern nearby frequencies. When working with the DFT, bin spacing and frequency resolution are often used interchangeably. But this equivalence can be misleading.

$$\Delta f_{res} = \Delta f_{bin} = \frac{f_s}{N_{samples}}$$

Looking at this equation naively, we might conclude that increasing the number of samples will always increase our resolution, allowing us to resolve frequency components that are nearby. And so, instead of increasing the number of data points measured by a sampling device, we are compelled to zero-pad a signal afterward to increase N artificially.

In this case study, we will observe how zero-padding does not represent an increase in true spectral resolution but only offers extra graphical detail through interpolation.

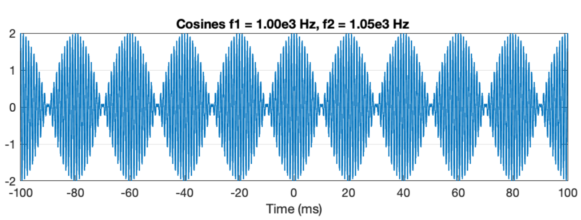
To make the difference clear, we study a simple test case: a sum of two pure sinusoids whose frequencies are only 50 Hz apart.

2. SET UP

Consider a signal $x(t)$ composed of two similar frequencies $f_1 = 1$ kHz, and $f_2 = 1.05$ kHz.

$$x(t) = \cos(2\pi f_1 t) + \cos(2\pi f_2 t),$$

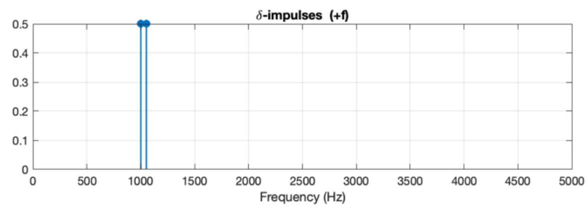
Time domain plot of $x(t)$.



The frequency transform of $x(t)$ is as follows.

$$\mathcal{F}\{x(t)\} =$$
$$X(f) = \frac{1}{2}[\delta(f-f_1) + \delta(f+f_1) + \delta(f-f_2) + \delta(f+f_2)]$$

Plot of the +frequency spectrum of $\mathcal{F}\{x(t)\}$.



In the plot above, we see two impulses corresponding to the frequencies present in $x(t)$.

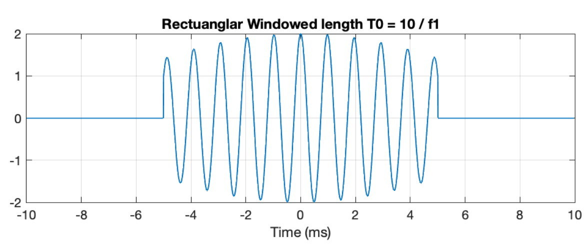
An implicit assumption of the spectrum $\mathcal{F}\{x(t)\}$ is that $x(t)$ is infinite, but in real applications, we can only sample a finite length of a signal. This is mathematically represented by multiplying the signal with the rectangular window function, which takes a snapshot of the signal in time of duration T.

The definition of the rectangular window.

$$w(t) = rect(t/T) = \begin{cases} 1, |t| \leq T/2 \\ 0, otherwise \end{cases}$$

Performing $x(t) \cdot w(t)$ with $T = 10$ ms produces the following plot.

Time domain plot of $x(t) \cdot w(t)$; $T = 10$ ms



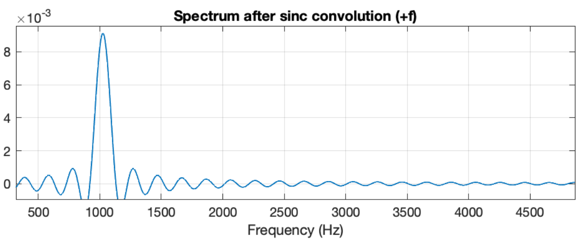
We see that only the information between the time interval of -5 ms and +5 ms is saved—reflecting the capture of only a section (10 ms) of an incoming signal. We can now look at $x(t) \cdot w(t)$ in the frequency domain.

Given the convolution theorem, we see that the effect of performing $x(t) \cdot w(t)$ in the time domain results in a spectrum convolved with a sinc function (the frequency dual of the rect function).

$$Y(f) = X(f) * W(f) =$$
$$\frac{T}{2}[\mathrm{sinc}[(f - f_1)T] + \mathrm{sinc}[(f + f_1)T] + \mathrm{sinc}[(f - f_2)T]$$
$$+ \mathrm{sinc}[(f + f_2)T]$$

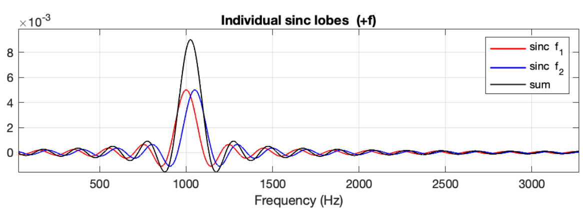
Compared to the spectrum of $x(t)$, the spectrum of $x(t) \cdot w(t)$ replaces each of the impulses with a sinc function—an oscillating waveform with side lobes—centered at the frequencies f_1 and f_2.

Spectrum $\mathcal{F}\{x(t) \cdot \omega(t)\}$; $T = 10\ ms$



Because the two frequencies are similar, it is difficult to distinguish the two sincs; so, the plot below decomposes the spectrum into the individual sincs components.

Decomposed spectrum $\mathcal{F}\{x(t) \cdot \omega(t)\}$; $T = 10$ ms



Again, in real applications, we observe the frequency spectrum in a discrete form (DFT). This can be mathematically represented as the spectrum multiplied by an impulse train, or Dirac comb function.

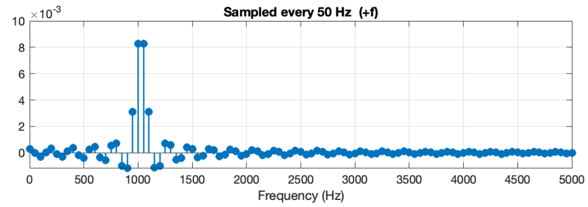
The definition of the Dirac comb function.

$$\Omega(f) = \sum_{k=-\infty}^{\infty} \delta(f - k\Delta f_{bin})$$

Where Δf_{bin} defines the spacing between each of the discrete points.

In the plot below, $\Delta f_{bin} = 50$ Hz.

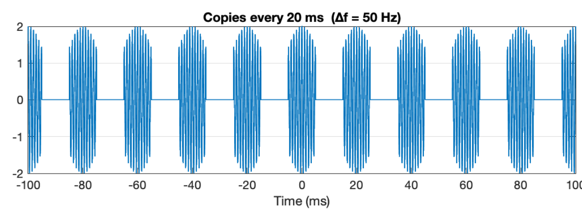
Sampling of spectrum $\mathcal{F}\{x(t) \cdot \omega(t)\}$; $\Delta f_{bin} = 50$ Hz.



We see the smooth continuous function $\mathcal{F}\{x(t) \cdot \omega(t)\}$ replaced by a series of impulses with an envelope that closely resembles the original spectrum. This is how a DFT would present the frequency information.

To observe how this affects the signal in the time domain, we can apply the convolution theorem again. The result is a time domain signal that repeats $x(t) \cdot w(t)$ at an interval of $\frac{1}{\Delta f_{bin}} = 20$ ms (aliasing).

Aliased plot of $x(t) \cdot \omega(t)$; $\Delta f_{bin} = 50$ Hz



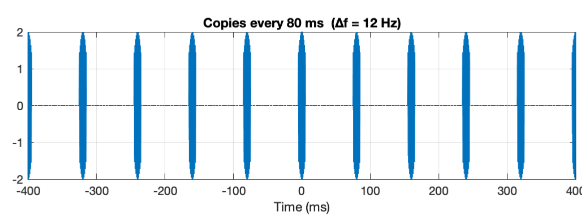
Interestingly, we also observe empty space (zero-padding) in each cycle. The duration of this space is the time difference between T and $\frac{1}{\Delta f_{bin}}$.

The amount of zero-padding is dependent on our Δf_{bin} chosen, and so we can now evaluate whether increasing the amount of zero-padding (by decreasing Δf_{bin}) will resolve the two frequencies in the signal.

3. OBSERVATIONS

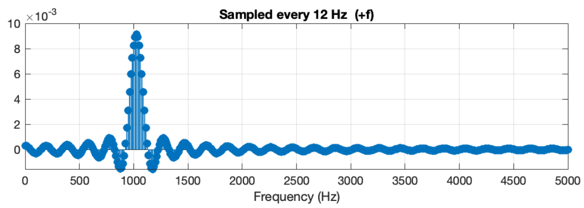
Here we decrease the bin spacing to $\Delta f_{bin} = 12$ Hz

Aliased plot of $x(t) \cdot \omega(t)$; $\Delta f_{bin} = 12$ Hz



We see that we have increased the zero-padding, but when we observe the frequency dual of the signal, we still cannot resolve the two frequencies.

Sampling of spectrum $\mathcal{F}\{x(t) \cdot \omega(t)\}$; $\Delta f_{bin} = 12$ Hz.



In fact, we will never be able to resolve the two frequencies. As $\Delta f_{bin} \to 0$ we converge on the continuous spectrum we started with, where the sinc lobes of the two components still overlap and are difficult to distinguish. It increases the smoothness of the plot by sampling more of the ideal continuous spectrum, but it cannot improve the underlying frequency resolution.

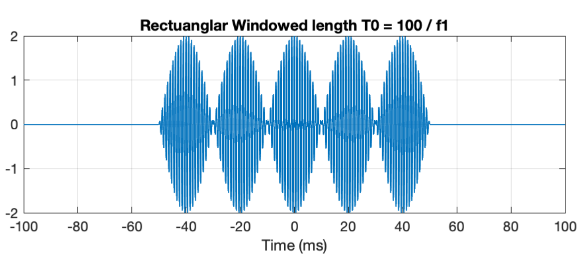
This is the essence of what zero-padding does. It can only interpolate points between the frequencies. It cannot increase the resolution of the signal because no new information is created. What was already present is just revealed by a denser sampling of the spectrum—it will only make the plot prettier and approach a more continuous form.

So, in order to increase the frequency resolution, we must acquire more information about the signal. This is done by increasing the snapshot length *T,* thus capturing more data points.

We can observe the effect of increasing T (increasing our observation window and thereby recording more samples).

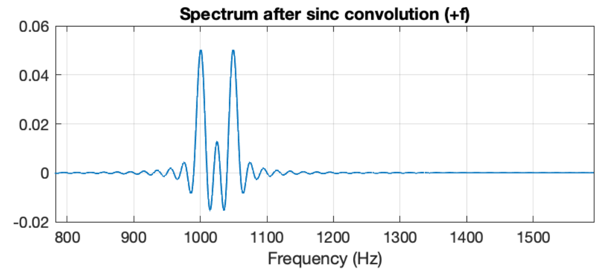
Performing $x(t) \cdot w(t)$ with $T = 100$ ms produces the following plot.

Time domain plot of $x(t) \cdot w(t)$ *with* $T = 100$ ms



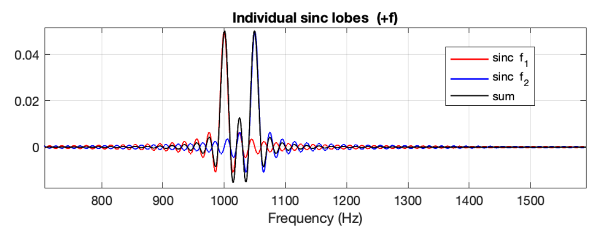
Then we observe the spectrum $\mathcal{F}\{x(t) \cdot \omega(t)\}$ with $T = 100$ ms.

spectrum $\mathcal{F}\{x(t) \cdot \omega(t)\}$; $T = 100$ ms



Immediately we are able to able to resolve the two frequencies. We can still decompose it into its individual sincs to observe the lobe spacing.

Decomposed spectrum $\mathcal{F}\{x(t) \cdot \omega(t)\}$; $T = 100$ ms



From the plot above the sincs have a lobe spacing of about 10 Hz.

Turns out, the lobe spacing is inversely proportional to the length T. Minimizing the lobe spacing, thus maximizing T is what prevents the two lobes from interfering and producing a spectrum that obscures the peaks.

4. CONCLUSION

Zero-padding reduces the DFT bin spacing, but it does not shrink the main-lobe width set by the observation window length T, which is what governs the ability to resolve nearby frequencies. Extra zeros only interpolate between existing spectral samples; only a longer record of actual data points provides gains in frequency resolution.