FIR filter design from IIR filters

Author: Himanshu Sharma

Our aim is to design FIR filters from IIR filters. Using IIR filters is one method, other methods being sampling the analog filters, and many others. Let us consider an example of lowpass filter and continue with that throughout this article.

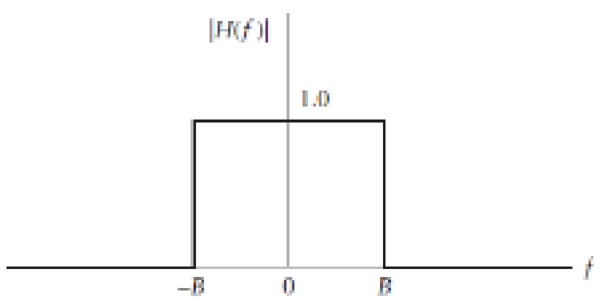

Figure 1: Ideal Lowpass Filter

If we take the inverse Fourier transfrom of this response, we get the impulse response of the filter which is nothing but

$$h(n) = \frac{\sin(Bn)}{\pi n} \tag{1}$$

where, $h(0) = \lim_{n \to 0} \frac{B\cos(Bn)}{\pi} = \frac{B}{\pi}$ . . . L'Hopital Rule

The time domain response is a decaying sinusoid with ∞ samples possible. The problem however is that it is a perfectly non-causal system. Let's see why. Most view the definition of causality as some system whose value does not depend on the future arguments. If I want to calculate $h(-1)$, then I must know $h(1)$, a value that is not currently with me. It is evident from equation 1 that $h(n) = h(-n)$.

Even if you go by the math, then for a causal system

$$h(n) = 0 \;\forall\; n < 0 \tag{2}$$

which by the way is due to the fact that if there is no impulse before our clock has started, the response must also be absent. Here that is not the case with equation 1, because there is no unit-step attached to it. So, even if I am starting my clock at $n = 0$, I must have outputs for time $n < 0$. It is a complete violation of conservation law. So if that happens with a real system then either the user clock hasn't started at the same time as the system clock, or the system is defective. Systems like these are not implementable directly in real time.

A second problem is with the domain of the function $h(n)$ in equation 1. The discrete time can take any value in range $(-\infty, \infty) \in I$. Even if somehow this system is implemented for $n > 0$, I must wait till $n \to \infty$. So, lets forget about the ideal lowpass filter. However, we can still construct a lowpass filter which certainly is not ideal by taking a single portion of this IIR filter. By the way, since it was never mentioned, equation 1 represents an IIR filter.

Now we define a new parameter N which is length of our FIR filter derived from this IIR filter. Ideality is time hungry as we saw. It wants ∞ samples of equation 1 to obtain an ideal lowpass filter. So, it would be better if we take as many samples as we can. In other words, N must be as large as practically possible. However, this transformation from an IIR to FIR is not unique. Two different designs are possible if two different sections of the same IIR filter of length N are taken. So, if both N and the way of slicing equation 1 is fixed, we arrive at a unique solution. The concept of *windows* is a branch off here, but before moving to that, let us consider some math now.

Shown below is the sectioning of the impulse response of an ideal lowpass filter by two different windows. The slicing is symmetric around the y-axis. This will be soon discussed as to why this is done. The scaling of the windows is also different. The peak of the FIR filter when window 2 is used will be of height AC and that with window 1 will be CB. That is not the concern but the problem is with the variable slope of the windows when they are non zero. This will imply different scaling of the sinc function at different times. For example, with the case of window 2, the central lobe will be scaled more than the left and right side of the function, the result will not be an exact copy of the sinc but instead more like a delta pulse if the central lobe is thin. With window 1, opposite will happen.

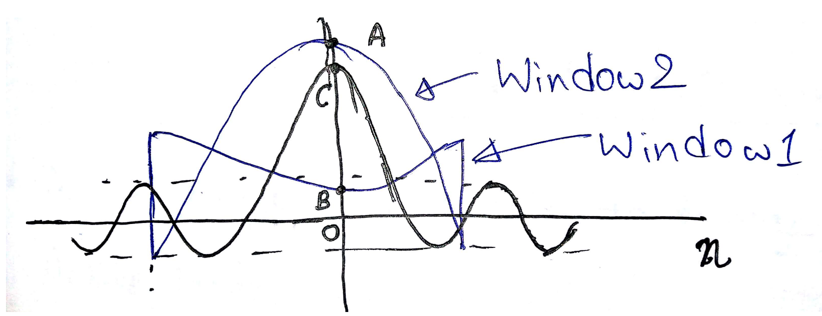

Figure 2: Sectioning the IIR filter

We need a window with zero slope when it is non zero. A rectangular window is what we need. Its height can be anything (height of the frequency response will change to that constant, see figure 1) but that should be constant throughout. With windows 1 and 2, the problem was that the shape of the sinc was not preserved. First of all we were not taking infinite samples as required by the ideal behaviour and now we are not even preserving the shape, that is the only issue. To get the exact frequency response as in figure 1, we move a little further by making the height of the rectangular window equal to 1. We are moving in steps. We first designed a window which is rectangular and then took its height equal to 1. This should be grasped properly because we will be using this in the coming equations. Now let us denote this sliced time domain impulse function as $\hat{h}(n)$. Making rectangular window of height 1 unit ensures that $\hat{h}(n) = h(n)$ within the section. We define an error function $e(n)$ which tracks the deviation from ideal impulse response within our section limits.

$$\Delta(n) = \sum_{n=-\infty}^{\infty} e^2(n) = \sum_{n=-\infty}^{\infty} [\hat{h}(n) - h(n)]^2 \tag{3}$$

Squaring the error has two benefits over taking the absolute of the error.

1. We talk in the terms of energy.

2. A squared function is differentiable at zero.

Since outside the section, $\hat{h}(n) = 0$ (see figure 2), we define split $\Delta(n)$ in three parts.

$$\Delta(n) = \sum_{n<N} e^2(n) + \sum_{n \in N} e^2(n) + \sum_{n>N} e^2(n) =$$
$$\sum_{n<N} h^2(n) + \sum_{n \in N} [\hat{h}(n) - h(n)]^2 + \sum_{n>N} h^2(n)$$

In terms of energy, we have no control over the left and right summations because they are dependent on the energy of the IIR filter. But we have to minimize the overall energy $\Delta(n)$. The only possible way is to make the middle summation amount to 0 and that is achieved when $\hat{h}(n) = h(n)\ \forall\ n \in N$. In other words, within the section limits, the FIR filter must be an exact copy of the IIR filter.

Let's talk about the even symmetry now. We know that there is a unique mapping in time and Fourier domain, i.e., every time domain function has a unique frequency domain function and that is the reason an inverse relation exists. Below, two figures are shown which are the frequency representations of the two sections of the IIR filter.

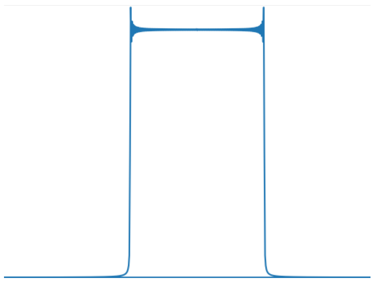

Figure 3: Non-Causal Slicing

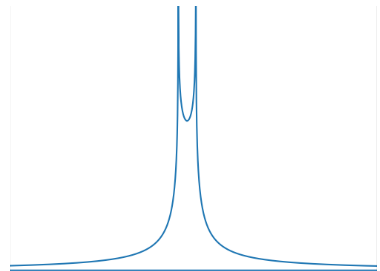

Figure 4: Causal Slicing

As you can see, if $N/2$ samples on both sides are taken, that is, even symmetric slicing is performed then the filter obtained is shown in figure 3. If however, N length samples are taken from the right side ($n \geq 0$), the filter thus obtained is shown in figure 4. The filter shown in figure 3 is more closer to the ideal filter and we would prefer it rather than the filter shown in figure 4. Conslusion is that both sides should be considered while taking the samples from an IIR filter. Both sides mean that $N/2$ samples from both the regions $n > 0$ and $n < 0$ must be taken, then only some degree of ideality will show up in the frequency domain.

An interesting point to note in figure 3 is that there are ripples at the edges of the square wave. This comes from the convolution theorem. Let us see what does that mean. The frequency response of the rectangular window that we are using for slicing the IIR filter will be a sinc function in the frequency domain and the low pass filter is a rectangular wave from $-B$ to B in the frequency domain. The point about the frequency domain of the rectangular window must be noted carefully. Since a rectangular window in discrete time is a rectangular wave, its frequency domain is a sinc function. It might seem confusing but just to clarify, the time domain representation of a lowpass filter is a sinc and the frequency domain representation of the rectangular window is a sinc. When we say that we are applying a window to an IIR filter to slice its N samples whatwe actually mean is that we are multiplying a window function (which is non-zero for some N samples of an IIR filter) with an IIR filter. That is, in time domain we are multiplying the IIR filter with the rectangular window, and so, in the frequency domain, we convolve their respective frequency responses. The frequency response of the IIR filter is the ideal lowpass filter shown in figure 1 and for the rectangular window, it is again a sinc function for $\omega \in (-\pi, \pi)$. Convolution is nothing but shift and multiply operation. So imagine that our central lobe of the sinc is at far left. We start multiplying it with the low pass filter, shift it to right and again multiply and so on. A point comes when the central lobe lies exactly on the cutoff frequency of the low pass filter, i.e., $\omega = \pm B$, a point of discontinuity in the frequency response of the lowpass filter. A discontinuity has an infinite slope. It acts like a delta pulse for the sinc function and therefore the decaying ripples of the sinc are replicated on the convolution output. This paragraph is a little deviation from what we are discussing.

Let us come back to the implementation problem. Till now we are equipped with the fact that if we need high degree of ideality then N must be as large as possible, the samples of the FIR filter must exactly be the same as that of the IIR filter within the section limit and the even symmetry must be followed while sectioning the IIR filter. There is a problem with the last part. In no way it is practically possible to implement the even symmetry without modifying the phase of the lowpass filter. The ideal lowpass filter has a 0 phase response because $H(e^{j\omega}) = 1 \ \forall \omega \in (-B, B)$, 0 elsewhere. There is no imaginary part here. The phase is zero. Now to accomodate those negative time indices what we do is that we start the system clock at some instant and after $N/2$ samples have passed we initialize our local clock. In this way, we possibly have all the samples which can be termed negative. This is little bit confusing. What this means is that you start the system at some time t_o and then you just sit back and wait until some desired samples have passed. You then start your very own stopwatch for time measurement. According to your stopwatch, whatever samples have been generated by the system till now will occur in negative time and from now on, the time becomes positive. This way we avoid the problem of non-causality but we are at the loss of losing zero phase. Mathematically, it means that if,

$$\hat{h}(n) \leftrightarrow \hat{H}(e^{j\omega})$$

and we start our stopwatch after n_o samples have passed since system epoch, then

$$\hat{h}(n - n_o) \leftrightarrow e^{-j\omega n_o} \hat{H}(e^{j\omega})$$

That is, a linear phase is introduced instead of the original zero phase despite the fact that the time domain FIR representation is technically same. Hence, $\angle \hat{H}(e^{j\omega}) = -\omega n_o \neq 0$. A linear phase is not desired when compared to 0 phase but is better than losing frequency domain behaviour of the lowpass filter. That is, if we start both system clock and user clock at the same time, we preserve 0 phase but loose degree of ideality (see figure 4).