Let's do some exercises with filters.

Let's start with one of the simplest MA filters, the noncausal, equally weighted, three-point average.

$$y_n = \frac{1}{3}(x_{n-1} + x_n + x_{n+1}) \ .$$

This is obviously linear (sum of inputs at three consecutive times) and time invariant (since there is no explicit time dependence), and thus a filter. It is noncausal since x_{n+1} appears (but is obviously related to the causal filter $y_n = \frac{1}{3}(x_n + x_{n-1} + x_{n-2})$ - how?). It is MA since it contains x_m for various m but no y_m.

First let's find its *impulse response*,i.e., the output h_n when the impulse is the unit impulse (UI) (zero for all times n except $n = 0$, and $x_0 = 1$). At times before $n = -1$ all three terms are zero, and so the impulse response is zero. At time $n = -1$the term $x_{n+1} = x_0 = \frac{1}{3}$ so $h_{-1} = \frac{1}{3}$. Similarly, at time $n = 0$ we have $h_0 = \frac{1}{3}x_0 = \frac{1}{3}$ and at time $n = +1$ we have $h_1 = \frac{1}{3}x_0 = \frac{1}{3}$. So the impulse response is zero for all times except $n = -1, 0, +1$ where it is $\frac{1}{3}$. This shows that this filter is Finite Impulse Response (FIR) (which is always the case for MA filters). Note too that the impulse response of causal filters is always zero for negative n, but since this is a noncausal filter, we can have $h_{-1} \neq 0$.

Before finding the frequency response $H(\omega)$ it is always useful to find two special values, $H(0)$ (the frequency response at DC) and $H(\pi)$ (the frequency response at Nyquist frequency - the highest possible digital frequency.

Why do we call the maximal possible digital frequency $\omega = \pi$? Remember that sampling changes values in the time domain from t (in seconds) into $n = t/t_{sampling}$ (a pure number), and values in the frequency domain from f (in Hertz) into $k = f/f_{sampling}$ (a pure number). From the sampling theorem we know that the maximal frequency is half the sampling frequency, $f_{Nyquist} = \frac{1}{2}f_{sampling}$, and so in the digital domain the highest frequency is $\frac{1}{2}$. (Taking account the negative frequencies, the entire digital spectrum goes from $-\frac{1}{2}$ to $+\frac{1}{2}$, but for real signals the negative frequency part is a mirror reflection of the positive frequency part, so we will only draw from DC to $\frac{1}{2}$.) But we want the angular frequency $\omega = 2\pi f$, so the maximal angular digital frequency is $2\pi\frac{1}{2} = \pi$. (Taking account the negative frequencies, the entire digital angular spectrum has width 2π and goes from $-\pi$ to $+\pi$, but for real signals the negative frequency part is a mirror reflection of the positive frequency part, so we will only draw from DC to π. Also, note that $\omega = -\pi$and $\omega = +\pi$ are actually exactly the same frequency! This can be seen by noting that they are two angles separated by 2π, or alternately by remembering that the frequency axis is actually the zT restricted to the unit circle.)

Also note that we will always compute the frequency response for continuous ω, i.e., $H(\omega)$. But this is *digital* signal processing - why don't we limit ourselves to a finite number of digital frequencies k, i.e., H_k? There are two reasons. First, the input to the filter can be a digital sinusoid of *any* frequency - $x_n = \sin(\omega n)$. Second, we don't know ahead of time how many times n will be known; maybe we will be given 2 values x_0 and x_1 which lead to a digital spectrum with 2 values X_0 and X_1. However, we may be given 4 values, or a million. So, we don't know how many spectral values will be relevant. If we find the frequency response for all ω it is easy to return to a specific k value by sampling. For 2 values the frequencies will be DC and Nyquist; for 4 values DC, half Nyquist, Nyquist, and negative half Nyquist.

To find the frequency response at DC we consider what happens when we enter the DC signal $x_n = 1$ for all n. This always results in $y_n = \frac{1}{3}(1+1+1) = 1$. So the output equals the input and thus their ratio is $H(0) = 1$.

To find the frequency response at Nyquist frequency we input a maximal frequency signal $x_n = \ldots -1+1-1+1\ldots$ (positive for even n and negative for odd n). From the *law of filters* y_n must also be of maximal frequency, i.e., $y_n = \ldots - y + y - y + y \ldots$ for some value y. It is easy to do the math. For even n we have $y = \frac{1}{3}(-1+1-1) = -\frac{1}{3}$, and for odd n we similarly find $-y = \frac{1}{3}(1-1+1) = +\frac{1}{3}$. In either case the ratio between y_n and x_n is $-\frac{1}{3}$ and so $|H(Nyquist)| = \frac{1}{3}$ and there is a phase reversal.

We can guess that this filter is a low-pass filter, since it passes DC without attenuation, and attenuates the maximal frequency by a factor of three. To be sure, let's now find the entire frequency response. To do that we need to input a sinusoid of arbitrary frequency ω

$$x_n = e^{\mathrm{i}\omega n}$$

and find the output y_n. From the *law of filters* we knw that we will find:

$$y_n = H(\omega)x_n = H(\omega)e^{\mathrm{i}\omega n}.$$

Substituting a complex exponential sinusoidal input (as usual it is extremely messy use *real* sinusoids!)

$$y_n = \frac{1}{3}\left(e^{\mathrm{i}\omega(n-1)} + e^{\mathrm{i}\omega n} + e^{\mathrm{i}\omega(n+1)}\right) = \frac{1}{3}\left(e^{-\mathrm{i}\omega} + 1 + e^{\mathrm{i}\omega}\right)e^{\mathrm{i}\omega n}$$

we immediately identify $x_n = e^{\mathrm{i}\omega n}$ so we have found

$$y_n = \frac{1}{3}\left(e^{-\mathrm{i}\omega} + 1 + e^{\mathrm{i}\omega}\right)x_n$$

(we *knew* that this was going to happen!)

So

$$H(\omega) = \frac{1}{3}\left(1 + e^{-\mathrm{i}\omega} + e^{\mathrm{i}\omega}\right) = \frac{1}{3}\left(1 + 2\cos(\omega)\right)$$

is the desired frequency response. It is easy to check that indeed $H(0) = 1$ and $H(\pi) = -\frac{1}{3}$.

We usually draw the square of the frequency response (since that gives us what happens to the energy of the components, and ignores the phase).

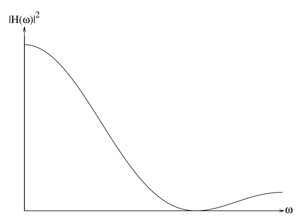

Figure 1: The (squared) frequency response of a simple three-point MA filter.

We see that this system is somewhat low-pass in character (i.e., lower frequencies are passed while higher frequencies are attenuated). However, the attenuation does not increase monotonically with frequency, and in fact the highest possible frequency $\frac{1}{2}f_s$ is not very well attenuated at all!

It turned out that the frequency response was *real*, which means that the angle of $H(\omega)$ is identically zero, implying that there is no phase shift. For every sinusoidal input, the output sinusoid is not only of the same frequency, it has the same phase - it goes up and down and crosses zero in unison with the input. Were we to repeat the exercise with the causal version $y_n = \frac{1}{3}(x_n + x_{n-1} + x_{n-2})$ we would find the same absolute value $|H(\omega)|$, meaning the same attenuation at each frequency, but a non-zero phase shift. In fact, the phase shift is linear in frequency, corresponding to a time shift of one sample. (Try it!) This is a general characteristic of symmetric (or antisymmetric) MA filters.

Let's try another three-point moving average.

$$y_n = \frac{1}{4}x_{n-1} + \frac{1}{2}x_n + \frac{1}{4}x_{n+1}.$$

This formula is fast to compute since it involves no true multiplications (for fixed point the multiplications are actually shifts, and for floating point they are exponent decrements).

Once again this is obviously a noncausal MA filter (why?). Its impulse response is zero for all times except $n = -1, 0, +1$ and $h_{\pm 1} = \frac{1}{4}$, $h_0 = \frac{1}{2}$. Since the filter is symmetric the impulse response is exactly the filter coefficients (in general we need to reverse the coefficients!).

What are our two special values for the frequency response? At DC $x_n = 1$ for all n and so $y = \frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1$. So, once again the gain at DC is one $H(0) = 1$. This time the gain at Nyquist frequency is zero, since when $x_n = \ldots -1 +1 -1 +1 \ldots$ we get (for even n) $y = \frac{1}{4}(-1) + \frac{1}{2}(+1) + \frac{1}{4}(-1) = 0$, and so $H(\pi) = 0$.

Proceeding as before we can find the entire frequency response by substituting a sinusoidal input of arbitrary frequency

$$y_n = \frac{1}{4}e^{\mathrm{i}\omega(n-1)} + \frac{1}{2}e^{\mathrm{i}\omega n} + \frac{1}{4}e^{\mathrm{i}\omega(n+1)} = \left(\frac{1}{4}e^{-\mathrm{i}\omega} + \frac{1}{2} + \frac{1}{4}e^{\mathrm{i}\omega}\right)e^{\mathrm{i}\omega n}$$

We identify $x_n = e^{\mathrm{i}\omega n}$ (we *knew* that this was going to happen!) and so

$$H(\omega) = \left(\frac{1}{4}e^{-\mathrm{i}\omega} + \frac{1}{2} + \frac{1}{4}e^{\mathrm{i}\omega}\right) = \frac{1}{2}\left(1 + \cos(\omega)\right)$$

a formula known as *raised cosine*. We compare the squared frequency responses of our two MA filters in the figure.

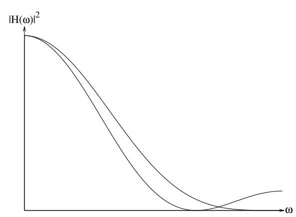

Figure 2: The (squared) frequency responses of two simple three-point average filters. Both responses are clearly low-pass but not ideal. The average with coefficients goes to zero at $\frac{1}{2}f_s$, but is 'wider' than the simple average.

This frequency response is low-pass in character like the previous one, and is more satisfying since it *does* go to zero at $\frac{1}{2}f_s$. However it is far from being an ideal low-pass filter that drops to zero response above some frequency; in fact it is wider than the frequency response of the simple average.

What happens to the frequency response when we average over more signal values? It is straightforward to show (hint: sum the geometric progression) that for equal-weighting simple case

$$y_n = \frac{1}{2L+1}\sum_{l=-L}^{L} x_{n+l}$$

the frequency response is

$$\frac{\sin(\frac{Lx}{2})}{L\sin(\frac{x}{2})}$$

as is depicted in the figure for $L = 3, 5, 7, 9$. We see that as L increases the filter becomes more and more narrow, so that for large L only very low frequencies are passed. However, this is only part of the story, since even for large L the oscillatory behavior persists. Filters with higher L have a narrower main lobe but more sidelobes.

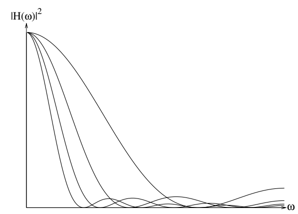

Figure 3: The squared frequency responses of simple averaging filters for $L = 3, 5, 7$ and 9. We see that as L increases the pass-band becomes narrower, but oscillations continue.

By using different coefficients we can get different frequency responses. For example, suppose that we need to pass frequencies below half the Nyquist frequency essentially unattenuated, but need to block those above this frequency as much as possible. We could use a 16-point moving average with the following magically determined coefficients

$$\begin{array}{rrrr} 0.003936, & -0.080864, & 0.100790, & 0.012206, \\ -0.090287, & -0.057807, & 0.175444, & 0.421732, \\ 0.421732, & 0.175444, & -0.057807, & -0.090287, \\ 0.012206, & 0.100790, & -0.080864, & 0.003936 \end{array}$$

the frequency response of which is depicted in Figure 4. While some oscillation exists in both the pass-band and the stop-band, these coefficients perform the desired task relatively well.

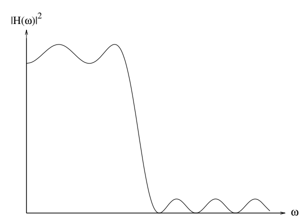

Figure 4: The (squared) frequency responses of a 16-coefficient low-pass filter. With these coefficients the lower frequency components are passed essentially unattenuated, while the higher components are strongly attenuated.

Let's try another MA filter - the first finite difference.

$$y_n = (\hat{\Delta} x)_n = x_n - x_{n-1}$$

This too is obviously a filter and MA, but *is* causal. Since this is a causal filter we know that the impulse response will be zero for all negative n. It is easy to see that $h_0 = +1$ and $h_1 = -1$, which is exactly the coefficients $h_0 = a_0$ and $h_1 = a_1$, and not in reversed order! This is precisely why we use convolutions in the first place (where one index goes up and the other goes down), rather than correlations (where both indexes move in the same direction). Some people prefer using correlations, but then the impulse response is not equal to the coefficients.

As usual, before calculating the full frequency response let's check our two special frequencies. For DC (substitute $x_n = 1$ for all n), $y_n = 0$ for all n, so $H(0) = 0$. For Nyquist (substitute $x_n = -1 + 1 - 1 + 1$), $y_n = \pm 2$, so $H(\pi) = 2$. This is obviously a high-pass filter!

To find the entire frequency response we substituting a complex exponential sinusoidal input.

$$y_n = e^{\mathrm{i}\omega n} - e^{\mathrm{i}\omega(n-1)} = \left(1 - e^{-\mathrm{i}\omega}\right) e^{\mathrm{i}\omega n}$$

and once again we $x_n = e^{\mathrm{i}\omega n}$ (we knew that would happen!). So we have found

$$H(\omega) = \left(1 - e^{-\mathrm{i}\omega}\right) = e^{-\mathrm{i}\omega/2} \left(e^{-\mathrm{i}\omega/2} - e^{-\mathrm{i}\omega/2}\right) = \mathrm{i}e^{-\mathrm{i}\omega/2}\ 2\sin(\omega/2)$$

so that $|H(\omega)| = 2\sin(\omega/2)$ and it is easy to check that indeed $|H(0)| = 0$ and $|H(\pi)| = 2$. The squared frequency response is depicted in the figure.

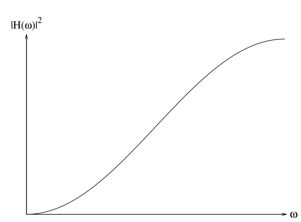

Figure 5: The squared frequency response of a finite difference filter. With these coefficients the lower frequency components are passed essentially unattenuated, while the higher components are strongly attenuated.

We can now see the relationship between the finite difference for digital signals and the derivative for analog signals. What does the true derivative do in the frequency domain? The derivative of $\sin(\omega t)$ is $\omega\cos(\omega t)$, which means that

$|H(\omega)| = \omega$ linear in frequency (with slope 1), and the phase shift is $\pi/2$ for all frequencies. We see in the figure that the finite difference's frequency response is close to being linear with slope 1, but not precisely so.

Let's try one last MA filter

$$y_n = x_{n+1} + x_{n-1}$$

and we'll skip directly to the special values - you should know how to find them by now! We find $H(0) = 1$ and $H(\pi) = 1$. What's happening? This can't be an all-pass filter! In such cases we can try to see what happens at an intermediate frequency, such as half-Nyquist. In the time domain $x_n = \ldots -10+10-10+10 \ldots$ and so we find $y_n = 0$ for all n (note that this trick doesn't always work this well). This means that this is a band-stop filter! I'll leave it as an exercise to find the full frequency response (but I'll draw it in the figure!). Incidentally, this is a special case of the *sinusoid blocker* $y_n = x_{n+1} - 2\cos(\Omega) + x_{n-1}$.

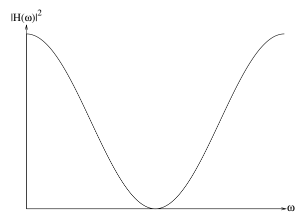

Figure 6: The squared frequency response of a band-stop filter.

Now let's move on to an AR filter. A simple causal AR filter with one delayed output looks like this:

$$y_n = (1-\beta)x_n + \beta y_{n-1} \qquad\qquad 0 \le \beta < 1$$

where $0 \le \beta \le 1$. We will shortly see why we have supplied a gain to the input.

Note that by changing β this AR filter can be set to track rapidly varying signals or to do a better job of removing noise from slowly varying ones. When $\beta = 0$ (corresponding to $L = 0$) the AR filter output y_n is simply equal to the input, no noise is averaged out but no bandwidth lost either. As β increases the past values assume more importance, and the averaging kicks in at the expense of not losing the ability to track the input as rapidly. When $\beta \to 1$ (corresponding to infinite L) the filter paradoxically doesn't look at the current input at all! Unlike a moving average filter, this AR filter never explicitly removes a signal value that it has seen from its consideration. Instead, past values are slowly 'forgotten' (at least for $\beta < 1$). For large β signal values from relatively long ago are still relatively important, while for small β past values lose their influence rapidly.

You can think of this AR filter as being similar to an MA filter operating on L previous values, the times before $n - L$ having been forgotten.

What is the impulse response of this filter? Since the filter is causal we know that $h_n = 0$ for all negative n. When $n = 0$ we easily have $h_0 = (1-\beta)1 + 0 = 1 - \beta$. Next, when $n = 1$ we need to feed back y_0 into the formula, and $h_1 = (1-\beta)0 + \beta y_0 = \beta(1-\beta)$. Continuing by explicitly carrying out the recursion we find that each value is equal to the previous one times β, i.e., $h_{n+1} = \beta h_n$. For $0 < \beta < 1$ this decays to zero, but never actually becomes zero, so this filter is Infinite Impulse Response (IIR).

Let's check the frequency response at our two special frequencies. At DC $y = (1-\beta)1+\beta y$, which means $y = 1$ and $H(0) = 1$ (now you see why we applied that gain to the input?). At Nyquist $y = (1-\beta)1-\beta y$ meaning that $y = (1-\beta)/(1+\beta)$ which is always less than one, and gets smaller and smaller as β increases until finally becoming zero when $\beta = 1$. This implies that we could a low-pass filter.

There are two ways to find the full frequency response. First - the hard way, which involves unraveling the recursion into an infinite convolution.

$$y_n = (1-\beta)x_n + \beta(1-\beta)x_{n-1} + \beta^2(1-\beta)x_{n-2} + \beta^3(1-\beta)x_{n-3} + \dots$$

We see that the coefficient corresponding to x_{n-l} is smaller than that of x_n by a factor of β^l, and so for all practical purposes we can neglect the contributions for times before some l. For example, if $\beta = 0.99$ and we neglect terms that are attenuated by e^{-1}, we need to retain about 100 terms; however for $\beta = 0.95$ only about 20 terms are needed, for $\beta = 0.9$ we are down to ten terms, and for $\beta = 0.8$ to 5 terms. It is not uncommon to use $\beta = 0.5$ where only the x_n and x_{n-1} terms are truly relevant, the x_{n-2} term being divided by 4.

Now we input our complex sinusoid and use the formula for the sum of a geometric series.

$$\begin{aligned}
y_n &= (1-\beta)e^{i\omega n} + \beta(1-\beta)e^{\mathrm{i}\omega(n-1)} + \beta^2(1-\beta)e^{\mathrm{i}\omega(n-2)} + \dots \\
&= (1-\beta)\sum_{k=0}^{\infty}(\beta e^{-\mathrm{i}\omega})^k e^{\mathrm{i}\omega n} \\
&= \frac{(1-\beta)}{(1-\beta e^{-\mathrm{i}\omega})}e^{\mathrm{i}\omega n}
\end{aligned}$$

and we see that $x_n = e^{\mathrm{i}\omega n}$ magically appeared on the right (we *knew* that this was going to happen!). We now identify

$$H(\omega) = \frac{(1-\beta)}{(1-\beta e^{-\mathrm{i}\omega})}$$

as the desired frequency response.

We could have done this much more easily by exploiting the *law of filters*. If $x_n = e^{\mathrm{i}\omega n}$ then we $y_n = H(\omega)e^{\mathrm{i}\omega n}$, and so

$$H(\omega)e^{\mathrm{i}\omega n} = (1-\beta)e^{i\omega n} + \beta H(\omega)e^{\mathrm{i}\omega(n-1)} = (1-\beta)e^{i\omega n} + \beta e^{-\mathrm{i}\omega}H(\omega)e^{\mathrm{i}\omega n}$$

which immediately leads to the previous equation, without the need for summing a geometric series.

We can find the squared frequency response to be

$$|H(\omega)|^2 = \frac{1-2\beta+\beta^2}{1-2\beta\cos(\omega)+\beta^2}$$

which we plot for several values of β in the figure.

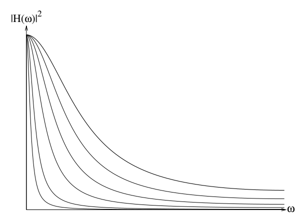

Figure 7: The (squared) frequency response of the simple AR low-pass filter for several different values of β. From top to bottom $\beta = 0.5, 0.6, 0.7, 0.8, 0.9, 0.95$.

We see in the figure that indeed the AR filter is more low-pass for higher β.

As our last example, let's consider an integration filter. For analog signals the integral of $\sin(\omega t)$ is $-1/\omega\cos(\omega t)$, so that $|H(\omega)| = 1/\omega$ (and there is a phase shift of π). For digital signals we define the infinite accumulator (which is the inverse of the first finite difference)

$$y_n = x_n + y_{n-1}$$

which unravels to the following infinite sum:

$$y_n = x_n + x_{n-1} + x_{n-2} + \ldots = \sum_{m=0}^{\infty} x_{n-m} \ .$$

We can write this in terms of the time delay operator

$$y = (1 + \hat{z} + \hat{z}^2 + \ldots)x = \Upsilon x$$

where we have defined the infinite accumulator operator

$$\Upsilon \equiv \sum_{m=0}^{\infty} \hat{z}^m x_n$$

which roughly corresponds to the integration operator for continuous signals. The finite difference $\Delta \equiv (1-\hat{z})$ and the infinite accumulator are related through $\Delta\Upsilon = 1$ and $\Upsilon\Delta = 1$, where 1 is the identity operator.

What happens when the infinite accumulator operates on DC? Since we are summing the same constant over and over again the sum obviously gets larger and larger in absolute value. This is what we previously called *instability*, since the output of the filter grows without limit although the input stays small. Such unstable behavior could never happen with an MA filter; and it is almost always an unwelcome occurrence, since all practical computational devices will eventually fail when signal values grow without limit.

What is the frequency response of the infinite accumulator? If you have gotten this far, it shouldn't be hard for you to see that

$$H(\omega)e^{\mathrm{i}\omega n} = e^{i\omega n} + H(\omega)e^{\mathrm{i}\omega(n-1)}$$

which means that

$$H(\omega) = \frac{1}{1-e^{-\mathrm{i}\omega}} = -\mathrm{i}e^{\mathrm{i}\omega/2}\frac{1}{2\sin(\omega/2)}$$

if you plot this, you will find that it is very close to the frequency response of the analog integral.