

Exponential Moving Average

Pieter P

Difference equation

The Difference equation of an exponential moving average filter is very simple:

$$y[n] = \alpha x[n] + (1 - \alpha)y[n - 1]$$

In this equation, $y[n]$ is the current output, $y[n - 1]$ is the previous output, and $x[n]$ is the current input; α is a number between 0 and 1. If $\alpha = 1$, the output is just equal to the input, and no filtering takes place

The filter is called 'exponential', because the weighting factor of previous inputs decreases exponentially. This can be easily demonstrated by substituting the previous outputs:

$$\begin{aligned} y[n] &= \alpha x[n] + (1 - \alpha)y[n - 1] \\ &= \alpha x[n] + (1 - \alpha)(\alpha x[n - 1] + (1 - \alpha)y[n - 2]) \\ &= \alpha x[n] + (1 - \alpha)(\alpha x[n - 1] + (1 - \alpha)(\alpha x[n - 2] + (1 - \alpha)y[n - 3])) \\ &\dots \\ &= \alpha \sum_{k=0}^n (1 - \alpha)^k x[n - k] \end{aligned}$$

Impulse and step response

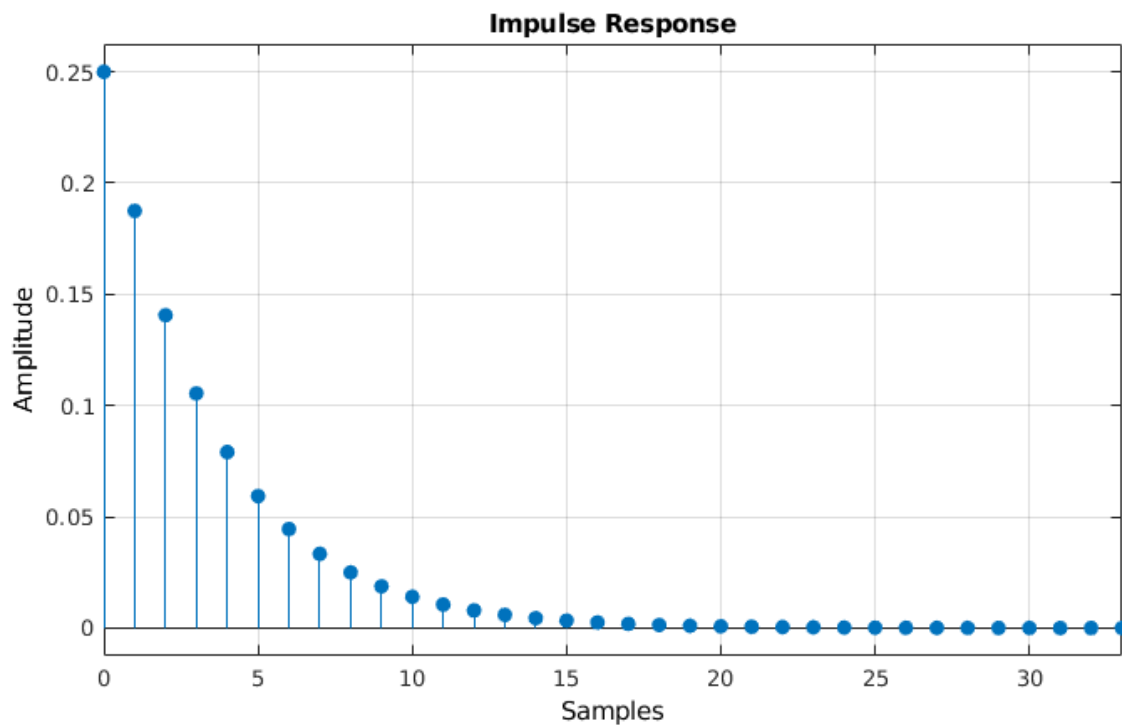
From the previous equation, we can now easily calculate the impulse and step response.

The impulse response is the output of the filter when a Kronecker delta function is applied to the input.

$$\delta[n] = \begin{cases} 1 & n = 0 \\ 0 & n \neq 0 \end{cases}$$

$$y_{\text{impulse}}[n] = h[n] = \alpha \sum_{k=0}^n (1 - \alpha)^k \delta[n - k] = \alpha(1 - \alpha)^n$$

For example, if $\alpha = 0.25$, the impulse response is as follows:

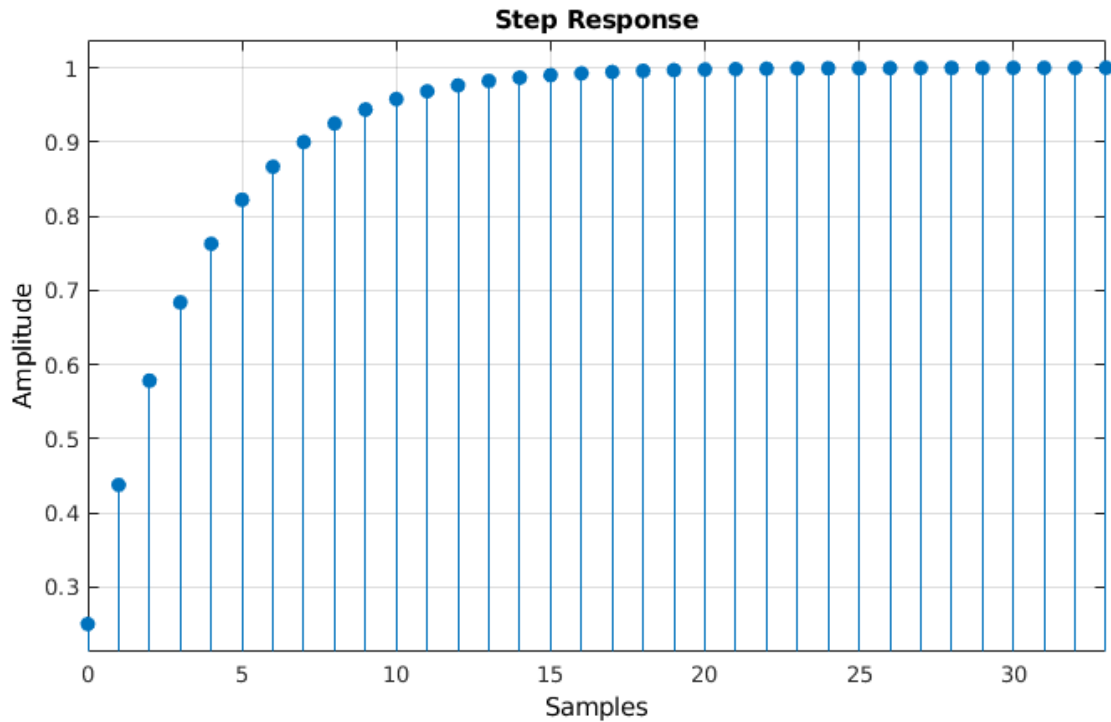


The step response is the output of the filter when a Heaviside step function is applied to the input.

$$H[n] = \begin{cases} 0 & n < 0 \\ 1 & n \geq 0 \end{cases}$$

$$y_{step}[n] = \alpha \sum_{k=0}^n (1 - \alpha)^k H[n - k] = \alpha \sum_{k=0}^n (1 - \alpha)^k$$

For example, if $\alpha = 0.25$, the step response is as follows:



Transfer function

You can express the output of discrete-time linear time-invariant systems, of which the EMA is an example, as the convolution of the input with the impulse response.

This is because any input $x[n]$ can be written as an infinite sum of shifted Kronecker delta functions

$$x[n] = \sum_{k=0}^{\infty} x[k] \delta[k - n]$$

If T is the transformation performed by the DTLTI system, and by using the fact that T is a linear transformation and that it is time-invariant:

$$\begin{aligned}
 y[n] &= T(x[n]) \\
 &= T\left(\sum_{k=0}^{\infty} x[k] \delta[k - n]\right) \\
 &= \sum_{k=0}^{\infty} T(x[k] \delta[k - n]) \\
 &= \sum_{k=0}^{\infty} x[k] T(\delta[k - n]) \\
 &= \sum_{k=0}^{\infty} x[k] h[k - n] \\
 &= x[n] * h[n]
 \end{aligned}$$

Analysis of such systems is usually easier in the Z-domain, in which the convolution is reduced to a simple product.

The (unilateral) Z-transform is defined as:

$$\mathcal{Z}\{x[n]\} = \sum_{n=0}^{\infty} x[n] z^{-n}$$

If $X(z) = \mathcal{Z}\{x[n]\}$, $Y(z) = \mathcal{Z}\{y[n]\}$ and $H(z) = \mathcal{Z}\{h[n]\}$ are the Z-transforms of the input, output and impulse response respectively, then:

$$\begin{aligned}
 \mathcal{Z}\{y[n]\} &= \mathcal{Z}\{x[n] * h[n]\} \\
 Y(z) &= X(z)H(z) \\
 H(z) &= \frac{Y(z)}{X(z)}
 \end{aligned}$$

$H(z)$ is called the transfer function of the system.

Let's calculate the transfer function of the EMA.

We can take two main approaches: use the difference equation and use some of the properties of the Z-transform to calculate $\frac{Y(z)}{X(z)}$, or apply the definition of the Z-transform to the impulse response $h[n]$ directly.

Using the difference equation

All you have to do is apply the time shifting property of the Z transform:

$$\mathcal{Z}\{y[n - n_0]\} = z^{-n_0}\mathcal{Z}\{y[n]\}$$

Then just rearrange the terms.

$$\begin{aligned} y[n] &= \alpha x[n] + (1 - \alpha)y[n - 1] \\ \mathcal{Z}\{y[n]\} &= \mathcal{Z}\{\alpha x[n] + (1 - \alpha)y[n - 1]\} \\ &= \alpha \mathcal{Z}\{x[n]\} + (1 - \alpha)\mathcal{Z}\{y[n - 1]\} \\ &= \alpha \mathcal{Z}\{x[n]\} + (1 - \alpha)z^{-1}\mathcal{Z}\{y[n]\} \\ Y(z) &= \alpha X(z) + (1 - \alpha)z^{-1}Y(z) \\ Y(z) - (1 - \alpha)z^{-1}Y(z) &= \alpha X(z) \\ \frac{Y(z)}{X(z)} &= \frac{\alpha}{1 - (1 - \alpha)z^{-1}} \\ H(z) &= \frac{\alpha}{1 - (1 - \alpha)z^{-1}} \end{aligned}$$

Using the impulse response

$$\begin{aligned} H(z) &= \mathcal{Z}\{h[n]\} \\ &= \sum_{n=0}^{\infty} h[n]z^{-n} \\ &= \alpha \sum_{n=0}^{\infty} (1 - \alpha)^n z^{-n} \\ &= \alpha \sum_{n=0}^{\infty} \left(\frac{1 - \alpha}{z}\right)^n \\ &= \frac{\alpha}{1 - \frac{1 - \alpha}{z}} \end{aligned}$$

The last step is only valid if the sum converges, this is the case for $|\frac{1 - \alpha}{z}| < 1$. (See [Infinite Geometric Series](#))

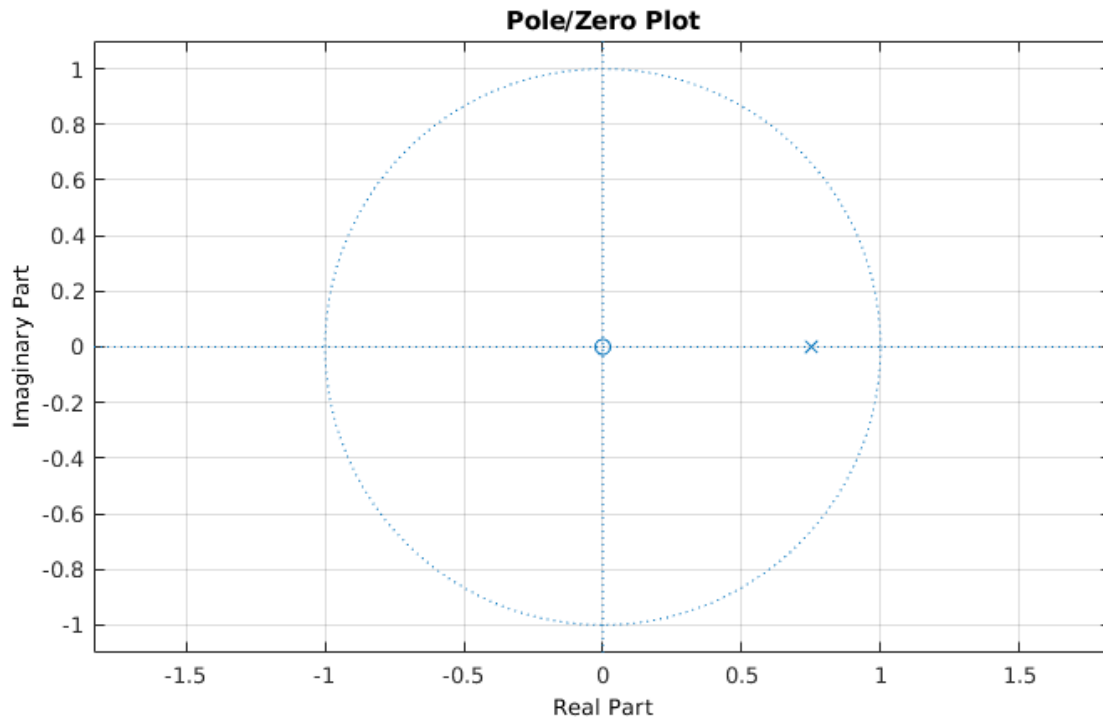
In these expressions, z is a complex variable, and $H(z)$ is a complex function.

There are a couple of interesting values for z : values that result in the numerator becoming zero, called *zeros*, and values that result in the denominator becoming zero, called *poles*.

$$\begin{aligned} H(z) &= \frac{\alpha}{1 - \frac{1 - \alpha}{z}} \\ &= \frac{\alpha z}{z - (1 - \alpha)} \end{aligned}$$

By rewriting the transfer function, we can easily see that $z = 0$ is a zero, and $z = 1 - \alpha$ is a pole.

The poles and zeros determine the overall effect of the transfer function, so pole-zero plots are a very useful tool when describing filters. This is the pole-zero plot of the same example EMA as before, with $\alpha = 0.25$.



The zero in the origin is indicated by an **O**, and the pole at $1 - \alpha = 0.75$ by an **X**.

Frequency response

An important property of discrete-time linear time-invariant systems is that it preserves the pulsation (frequency) of sinusoidal signals, only the phase shift and the amplitude are altered.

This makes it relatively easy to express the frequency response (sometimes called magnitude response) of a filter.

We're interested in the spectrum of frequencies that a signal contains, so it makes sense to decompose it as a sum of sines and cosines. That's exactly what the discrete-time Fourier transform does:

$$X_{2\pi} = \sum_{n=-\infty}^{\infty} x[n]e^{-i\omega n}$$

Note that this is just a special case of the Z-transform, where $z = e^{i\omega}$.

Consider the DTFT of the Kronecker delta. Remember that $\delta[n]$ is always zero, except for $n = 1$. This means that all but one terms of the infinite sum are zero, and what's left is just one.

$$X_{2\pi} = \sum_{n=-\infty}^{\infty} \delta[n]e^{-i\omega n} = \delta[0]e^{-i\omega 0} = 1$$

This is constant in function of pulsation ω , meaning that the Kronecker delta function contains all frequencies with the same amplitude. In other words, it is a very good reference. To determine the frequency response, we apply all frequencies to the input (the Kronecker delta), we measure those frequencies at the output, and compare their amplitudes to the original amplitude of 1.

We already know what the output signal of the system is when applying the Kronecker delta function to the input, it's the impulse response $h[k]$.

We already know the DTFT of the impulse as well, it is the Z-transform of $h[k]$, $H(z)$, where $z = e^{i\omega}$.

$$\begin{aligned} \mathcal{F}_{DTFT}\{h[k]\} &= \sum_{n=-\infty}^{\infty} h[n]e^{-i\omega n} \\ &= H(e^{i\omega}) \\ &= \frac{\alpha}{1 - (1 - \alpha)e^{-i\omega}} \end{aligned}$$

We can now calculate the amplitude of each frequency component in the output by taking the modulus of $H(e^{i\omega})$. For reasons that will become apparent in a minute, we'll calculate the modulus squared. We use Euler's formula for writing the exponential as a sines and cosines.

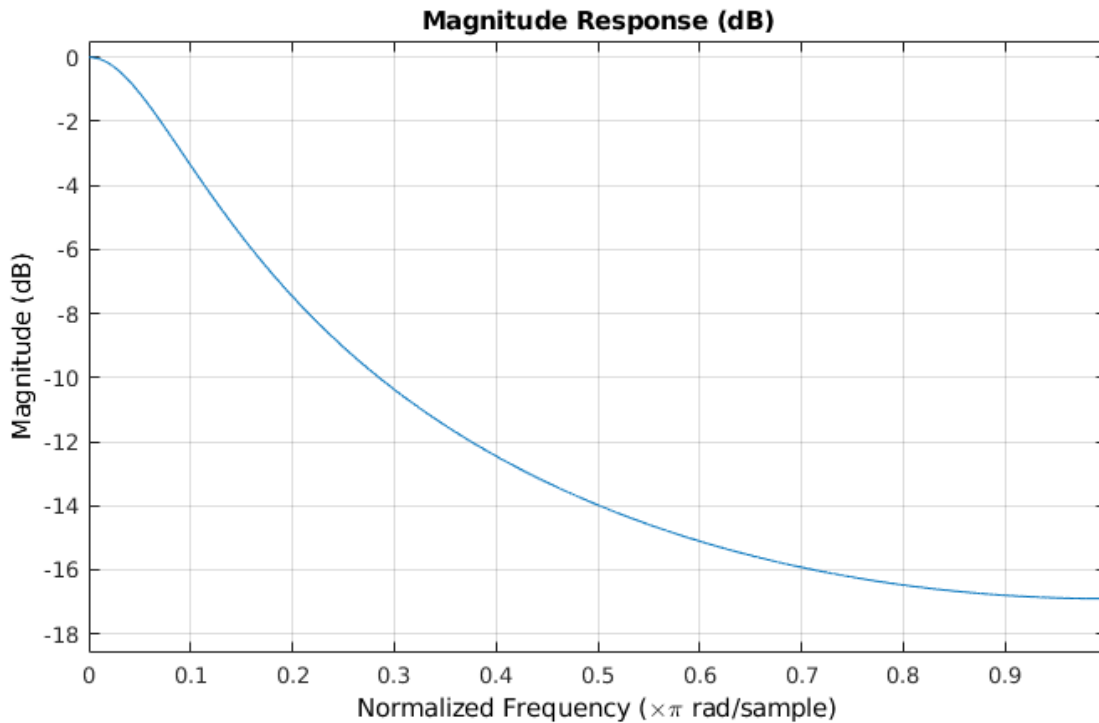
$$\begin{aligned}
 |H(e^{i\omega})|^2 &= \frac{\alpha^2}{|1 - (1 - \alpha)e^{-i\omega}|^2} \\
 &= \frac{\alpha^2}{|1 - (1 - \alpha)(\cos(-\omega) + i \sin(-\omega))|^2} \\
 &= \frac{\alpha^2}{(1 - (1 - \alpha)\cos(\omega))^2 - ((1 - \alpha)\sin(\omega))^2} \\
 &= \frac{\alpha^2}{1 - 2(1 - \alpha)\cos(\omega) + (1 - \alpha)^2 \cos^2(\omega) + (1 - \alpha)^2 \sin^2(\omega)} \\
 &= \frac{\alpha^2}{1 - 2(1 - \alpha)\cos(\omega) + (1 - \alpha)^2}
 \end{aligned}$$

ω is the normalized pulsation in radians per sample. You can substitute it with $\omega = \frac{2\pi f}{f_s}$ where f is the frequency in Hertz, and f_s is the sample frequency of the system in Hertz.

We can now plot the filter's gain in function of the pulsation. These plots often use a logarithmic scale, to show the gain in decibels. In order to calculate the power gain, the amplitude is squared.

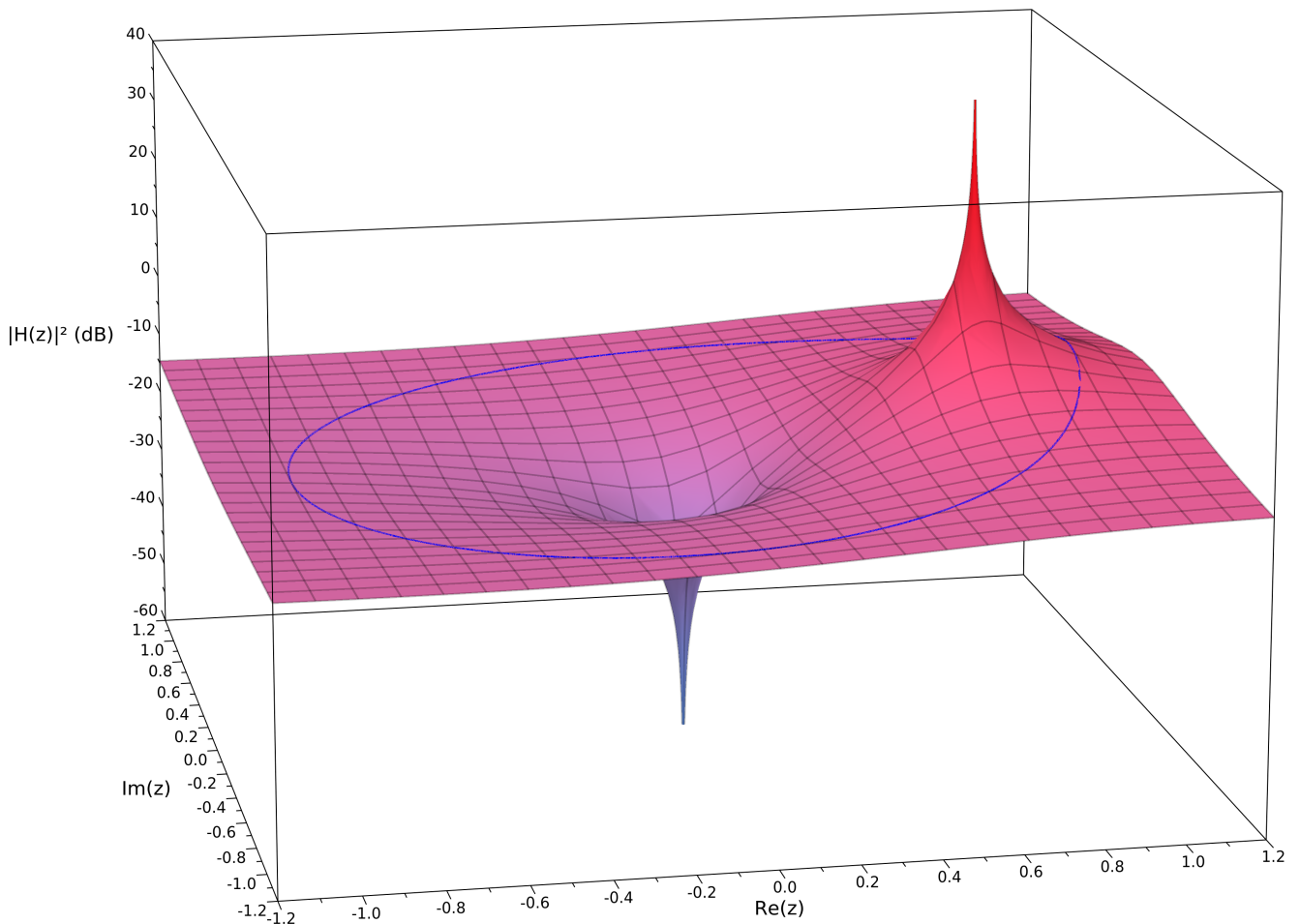
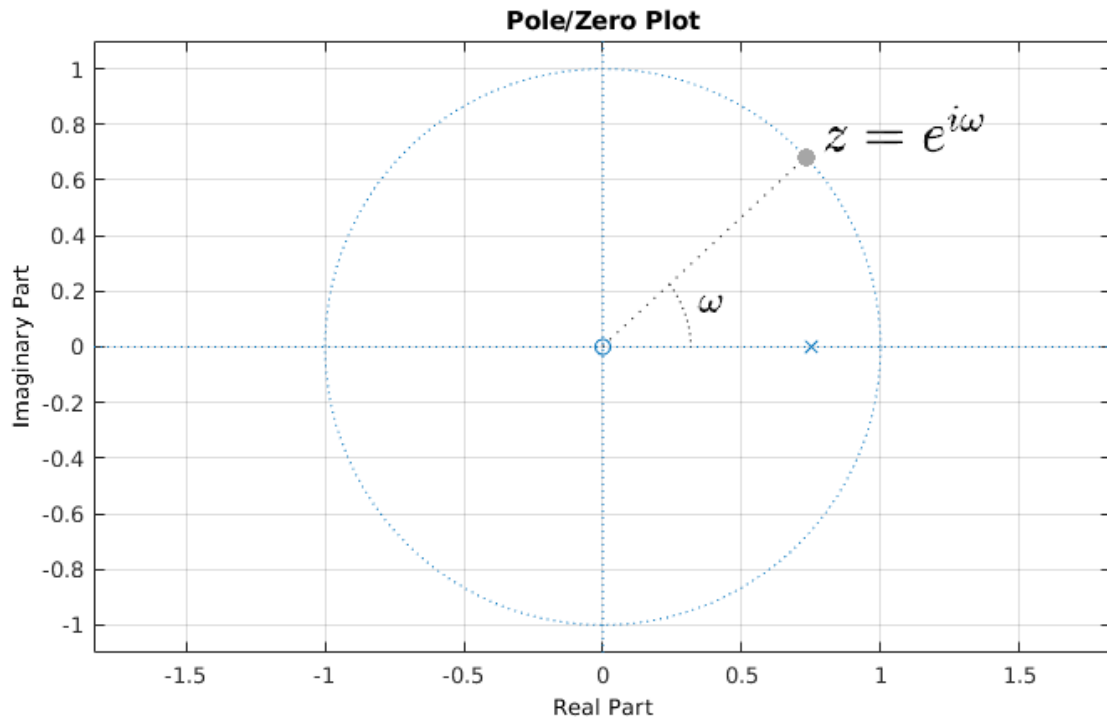
$$A_{dB}(\omega) = 10 \log_{10}|H(e^{i\omega})|^2$$

Note that when a frequency is not present in the output signal, the gain will be $-\infty$ dB. If a frequency has an amplitude of one in the output signal, the gain will be 0 dB.



You can clearly see the low-pass behavior of the EMA: low frequencies have a near-unit gain, and high frequencies are attenuated.

To get a better understanding of where this curve comes from, we can plot the entire $|H(z)|$ surface in the Z-domain. As mentioned above, the DTFT is just a special case of the Z-transform, where $z = e^{i\omega}$, i.e. the unit circle in the complex plane. Remember that the point $z = e^{i\omega}$ is a point with a distance of 1 to the origin, and with an angle of ω rad between its position vector and the positive x axis.



The image of the unit circle is shown in blue. Notice that this is the same curve as the blue curve in the magnitude response graph above: close to 0 when $\omega \rightarrow 0$ (the right half of the circle) and negative when $\omega \rightarrow \pi$ (the left half of the circle).

You can clearly see the effect of the pole at $0.75 + 0i$. The zero in the origin has no effect on the frequency response, because it doesn't alter the image of the unit circle, since $|e^{i\omega}| = 1$ or $\log(|e^{i\omega}|) = 0$.

Cutoff frequency

The cutoff frequency is defined as the frequency of the half-power point, where the power gain is a half. It's often called the -3 dB -point, because $10 \log_{10} \left(\frac{1}{2} \right) \approx -3.01 \text{ dB}$.

To find it, just solve the following equation:

$$|H(e^{i\omega_c})|^2 = \frac{1}{2}$$

$$\frac{\alpha^2}{1 - 2(1 - \alpha)\cos(\omega_c) + (1 - \alpha)^2} = \frac{1}{2}$$

$$\omega_c = \arccos\left(\frac{\alpha^2 + 2\alpha + 2}{2\alpha + 2}\right)$$

For example, if $\alpha = 0.25$, then $\omega_c = \arccos\left(\frac{23}{24}\right) \approx 0.2897 \frac{\text{rad}}{\text{sample}}$.

To convert it to a frequency in Hertz, you can multiply ω by $\frac{f_s}{2\pi}$, with f_s the sample frequency.

For example, if the sample frequency is $f_s = 1000 \frac{\text{samples}}{\text{s}}$, then $f = \arccos\left(\frac{23}{24}\right) \frac{f_s}{2\pi} \approx 46.12 \text{ Hz}$.
