ELECTRONOTES
1016 Hanshaw Road
Ithaca, NY 14850

APPLICATION NOTE NO. 418

Dec 10, 2014

FITTING A LINE TO A FIXED ERROR CRITERIA

INTRODUCTION:

Among my favorite topics are those that involve curve fitting and interpolation. Typically such "modeling" exercises involve some notion of an error criterion, and a minimization of that thusly defined error – a "best fit" in some well-defined sense. We have recently looked at some issues in this regard [1-3] and these reach back to a 1992 note [4]. Happily this has led to filter designs – another favorite topic of mine. Here we will look at a "fixed error" criterion and will relate this to least squares, exercising our analytical tools as we proceed.

A CURIOUS EXAMPLE

In Fig. 1a, consider the problem of fitting a straight line to the three points x(0), x(1), and x(2). In fact, try this by hand manipulation with a transparent ruler, such that the vertical error is equalized on all three points. Pretty easy to estimate.

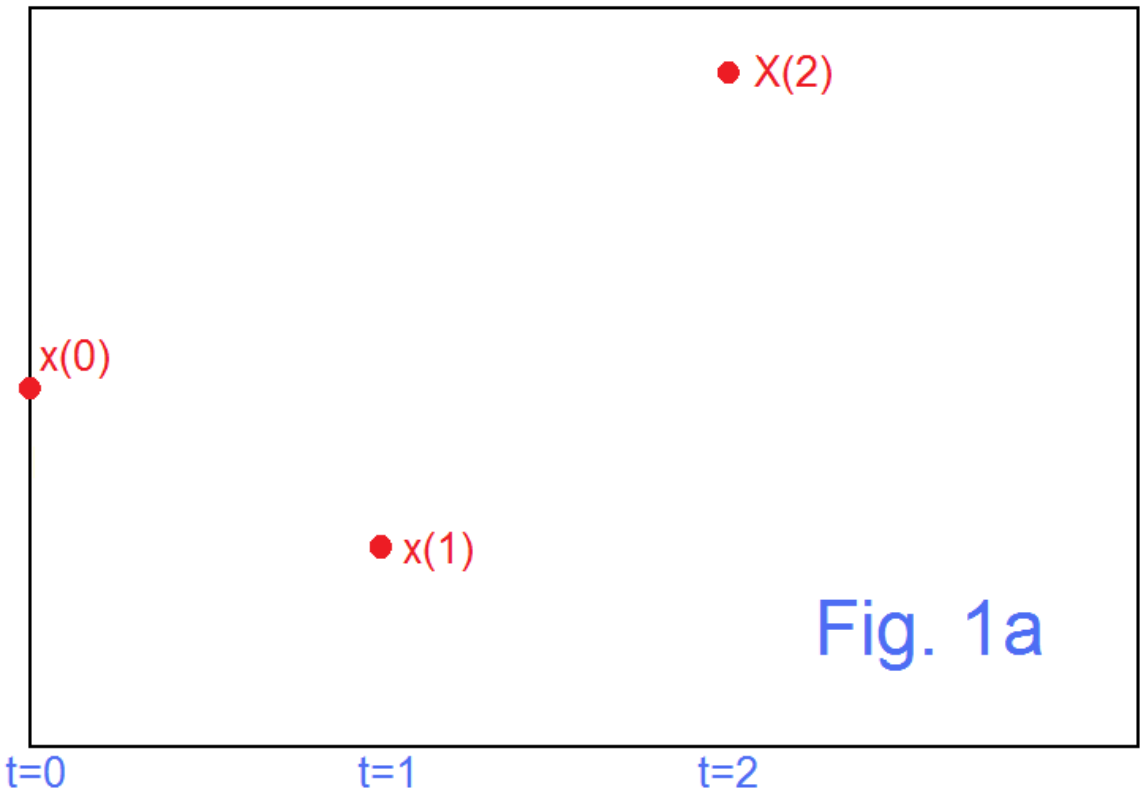
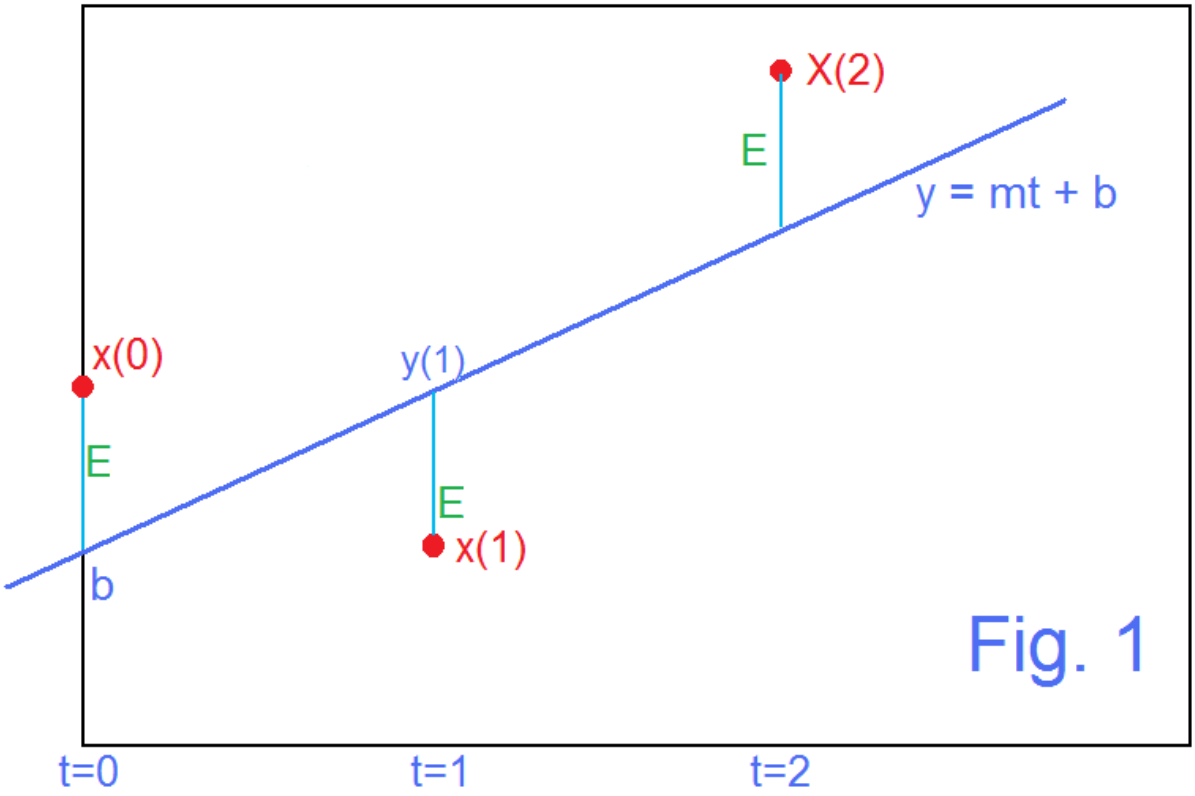


Fig. 1a

In AN-417 [2], fitting with least squares, we began by adding a parameter (a free choice of weight w on the center point). By choosing w=1 or w=2, we arrived (happily) at two previous results [4,1]. These were the (respective) cases where we had equal weights (w=1) with our filter ending up as a length-3 moving average h(n) = [1/3 1/3 1/3] and when w=2 we got a different filter: h(n) = [1/4 1/2 1/4]. Here we are doing something that is actually simpler, but less familiar. We are writing equations for what we did when we simply moved a ruler over Fig. 1. We are allowing for an error E that can be the same on all three points. This is NOT least squares. What this does is add an unknown to the fitting process (we have to solve for E). This gives us three equations in three unknowns (m, b, and E) with three input samples specified: x(0), x(1) and x(2), as in Fig. 1b. (Note that we could have chosen to have three different errors, E_0, E_1, and E_2, which would not have been a special case of interest, and would have permitted us to fit any line, most quite ugly to the given data.)



Our three equations are:

$$E = x(0) - b \tag{1a}$$

$$E = m + b - x(1) \tag{1b}$$

$$E = x(2) - 2m - b \tag{1c}$$

which can be put in matrix form as:

$$\begin{bmatrix} x(0) \\ x(1) \\ x(2) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ -1 & 1 & 1 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} E \\ m \\ b \end{bmatrix} \tag{2a}$$

which inverts to:

$$\begin{bmatrix} E \\ m \\ b \end{bmatrix} = \begin{bmatrix} 0.25 & -0.5 & 0.25 \\ -0.5 & 0 & 0.5 \\ 0.75 & 0.5 & -0.25 \end{bmatrix} \begin{bmatrix} x(0) \\ x(1) \\ x(2) \end{bmatrix} \tag{2b}$$

When we take this further, an interesting result occurs. First, the slope is:

$$\mathrm{m} = \frac{[\mathrm{x}(2) - \mathrm{x}(0)]}{2} \tag{3}$$

which is not that surprising. The slope is still determined by the endpoints. What is surprising is that the intercept b is given as:

$$b = \frac{3x(0) + 2x(1) - x(2)}{4} \tag{4}$$

which is exactly the result [equation (5b) of AN-417] for the w=2 case! Unanticipated, but easily understood. By choosing E above and below the line as unconditionally equal, it is the same as having two points above with weight 1, and one below with weight 2 when using least squares. An accident – but a fun result. Only for w=2 does least squares give the same error above and below (see Fig. 2 below).

The "take-away" here is that the best fit line, evaluated at the center is:

$$y(t = 1) = m + b = \left(\frac{1}{4}\right)x(0) + \left(\frac{1}{2}\right)x(1) + \left(\frac{1}{4}\right)x(2) \tag{5}$$

so the FIR filter has the impulse response:

$$h(n) = \left[\frac{1}{4} \quad \frac{1}{2} \quad \frac{1}{4}\right] \tag{6}$$

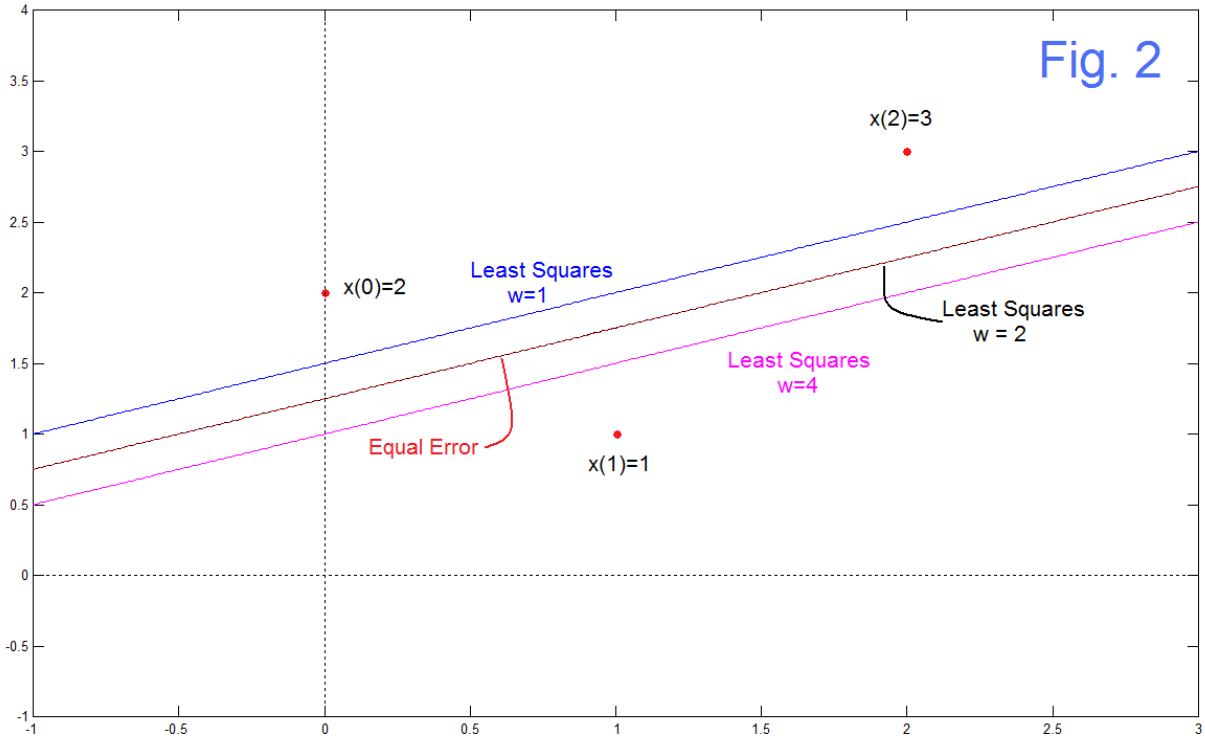
which corresponds to the simple low-pass with a double zero at z=-1. Here also:

$$E = 0.25x(0) - 0.5x(1) + 0.25x(2) \tag{7}$$

FITTING EXAMPLES

Weighted Squared Error Again:

We want to show a plot for various values of the center weight for weighted squared error vs. the equal error case, and this is shown in Fig. 2. Here we have the three input points (arbitrarily chosen) at 2, 1, and 3, and for the least squares



plots, w is the weight on the center point (the weights on the first and last points being 1). As expected when w=1 the straight line fit (blue) is shifted toward the two points x(0) and x(2) on the top. A contrasting case is where w=4 in which case the fit (magenta) is shifted downward toward the x(1) point. The case in the middle is for w=2 and for the equal error case discussed above. It is plotted in black and in red and appears as dotted black and red. Thus the w=2 cases comes out equal error. Is there any other way to do equal error?

Straight Line Through Three Points – Other Choices?

When we began with Fig. 1a and the suggestion that the reader try a hand-fit with a ruler, we had in mind that most likely the reader had already glanced at Fig. 1b and/or that the result of Fig. 1b was likely what the reader would choose. There are other possible choices, and we need to show that they are not good choices. Note that Fig. 1b has the "alternation" property in that the line is chosen in a below-above-below manner. This we perhaps favor just on intuition, or on the basis of work in frequency-domain filter design [5].

Probably it seems sensible that if we want a fit a straight line to a cluster of points, the line should go through the cluster. In fact, if we consider the possibility that all three points are on the same side of the line, we get a useless solution (Fig. 3). If the line goes through the point cluster, there are only the usual possibilities where two are on one side of the line and one on the other side. Now, if we further stipulate that the points are in time order, then the single-sided point is either in the middle, or on one (or the other) ends. We have already discussed the case where it is in the middle (the "alternation" case) which results in the w=2 least squares: h(n) = [1/4 1/2 1/4]. Is THE other uniquely distinct case any different?

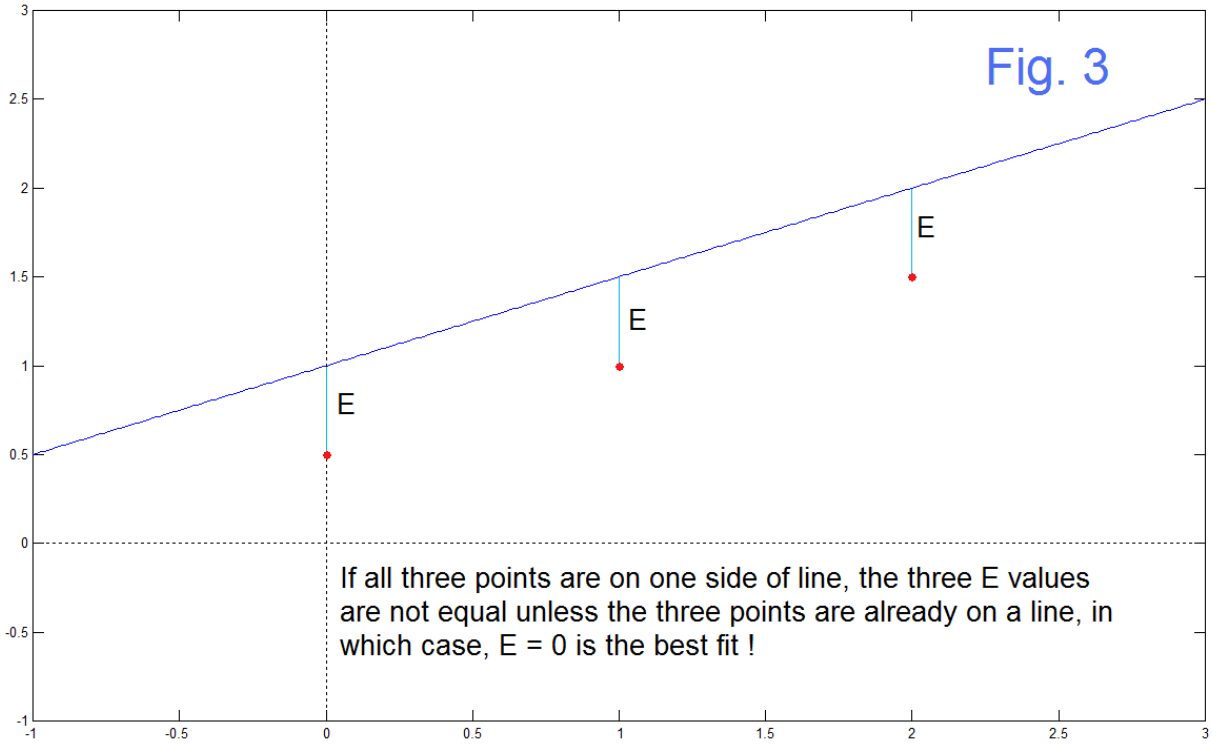


Fig. 3
E
E
E
If all three points are on one side of line, the three E values
are not equal unless the three points are already on a line, in
which case, E = 0 is the best fit !

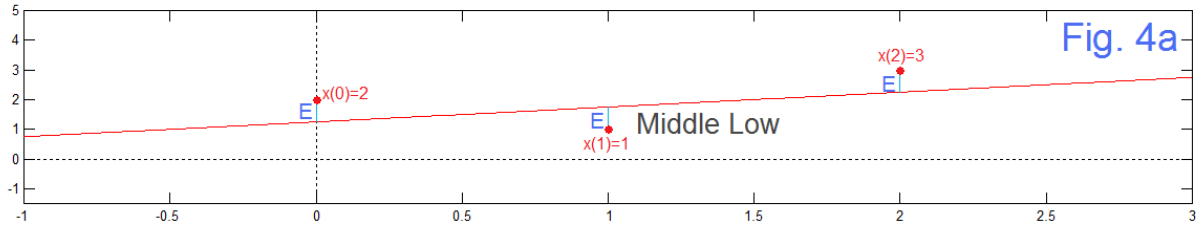


Fig. 4a
x(0)=2
x(1)=1
x(2)=3
E
E
E
Middle Low

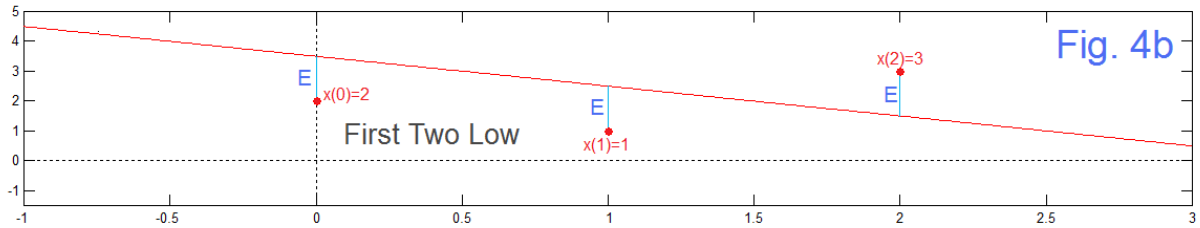


Fig. 4b
E
x(0)=2
E
x(1)=1
x(2)=3
E
First Two Low

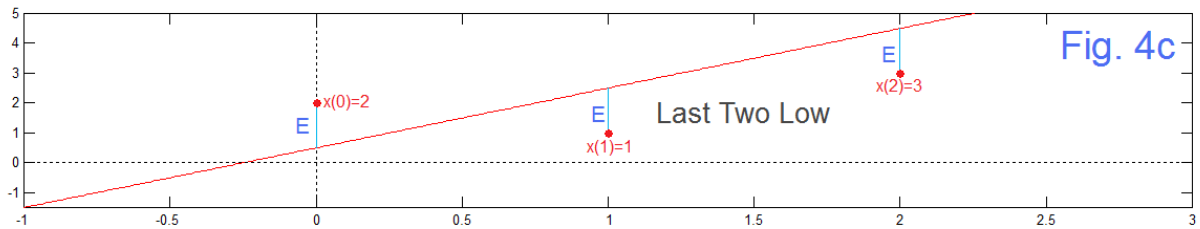


Fig. 4c
E
x(0)=2
E
x(1)=1
E
x(2)=3
Last Two Low

In Fig. 4a we show the (alternation) case where it is the middle point that is on a side of the line by itself [lower in this example of x(0)=2, x(1)=1, and x(2)=3]. This was developed as equations (1) to (7). Fig. 4b and Fig. 4c shows two other cases. We could give other examples, but the two here and any additional ones we try will lead to identical conclusions. For Fig. 4b the line that is fit is parallel to the first two points which are below the line. For Fig. 4c the line that is fit is parallel to the last two points which are also below the line. Let's first develop Fig. 4b.

Corresponding to equation (2a) we have instead:

$$\begin{bmatrix} x(0) \\ x(1) \\ x(2) \end{bmatrix} = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 1 & 1 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} E \\ m \\ b \end{bmatrix} \tag{8a}$$

which inverts to:

$$\begin{bmatrix} E \\ m \\ b \end{bmatrix} = \begin{bmatrix} 0.50 & -1.00 & 0.50 \\ -1.00 & 1.00 & 0 \\ 1.50 & -1.00 & 0.50 \end{bmatrix} \begin{bmatrix} x(0) \\ x(1) \\ x(2) \end{bmatrix} \tag{8b}$$

so now the slope is:

$$m = x(1) - x(0) \tag{9}$$

as determined by first two points. The intercept b is given as:

$$b = \frac{3x(0) - 2x(1) + x(2)}{2} \tag{10}$$

and the error E is:

$$E = 0.5x(0) - x(1) + 0.5x(2) \tag{11}$$

which is exactly twice the error of equation (7), as also looks to be the case from the plot of Fig. 4b. The FIR filter for this case is found from:

$$y(t = 1) = m + b = \left(\frac{x(0) + x(2)}{2} \right) \tag{12}$$

so the FIR filter has the impulse response:

$$h(n) = \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix} \tag{13}$$

This has the disconcerting feature that the estimate of the middle point, y(1), does not seem to involve the original data point x(1) at t=1. So while Fig. 4b looks like a poorer choice just from the figure, there seem to be two deficiencies analytically: <u>twice the error E, and no involvement of x(1).</u> Note that this result does not depend on the actual data points in the example.

Just to sharpen the point, we also develop Fig. 4c analytically. Corresponding to equation 8a we have instead:

$$\begin{bmatrix} x(0) \\ x(1) \\ x(2) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ -1 & 1 & 1 \\ -1 & 2 & 1 \end{bmatrix} \begin{bmatrix} E \\ m \\ b \end{bmatrix} \tag{14a}$$

which inverts to:

$$\begin{bmatrix} E \\ m \\ b \end{bmatrix} = \begin{bmatrix} 0.50 & -1.00 & 0.50 \\ 0 & -1.00 & 1.00 \\ 0.50 & 1.00 & -0.50 \end{bmatrix} \begin{bmatrix} x(0) \\ x(1) \\ x(2) \end{bmatrix} \tag{14b}$$

so now the slope is:

$$m = \mathrm{x}(2) - \mathrm{x}(1) \tag{15}$$

as determined by last two points. The intercept b is given as:

$$b = \frac{x(0) + 2x(1) - x(2)}{2} \tag{16}$$

and the error E is:

$$E = 0.5x(0) - x(1) + 0.5x(2) \tag{17}$$

which is again exactly twice the error of equation (7). The FIR filter for this case is found from:

$$y(t = 1) = m + b = \left(\frac{x(0) + x(2)}{2} \right) \tag{18}$$

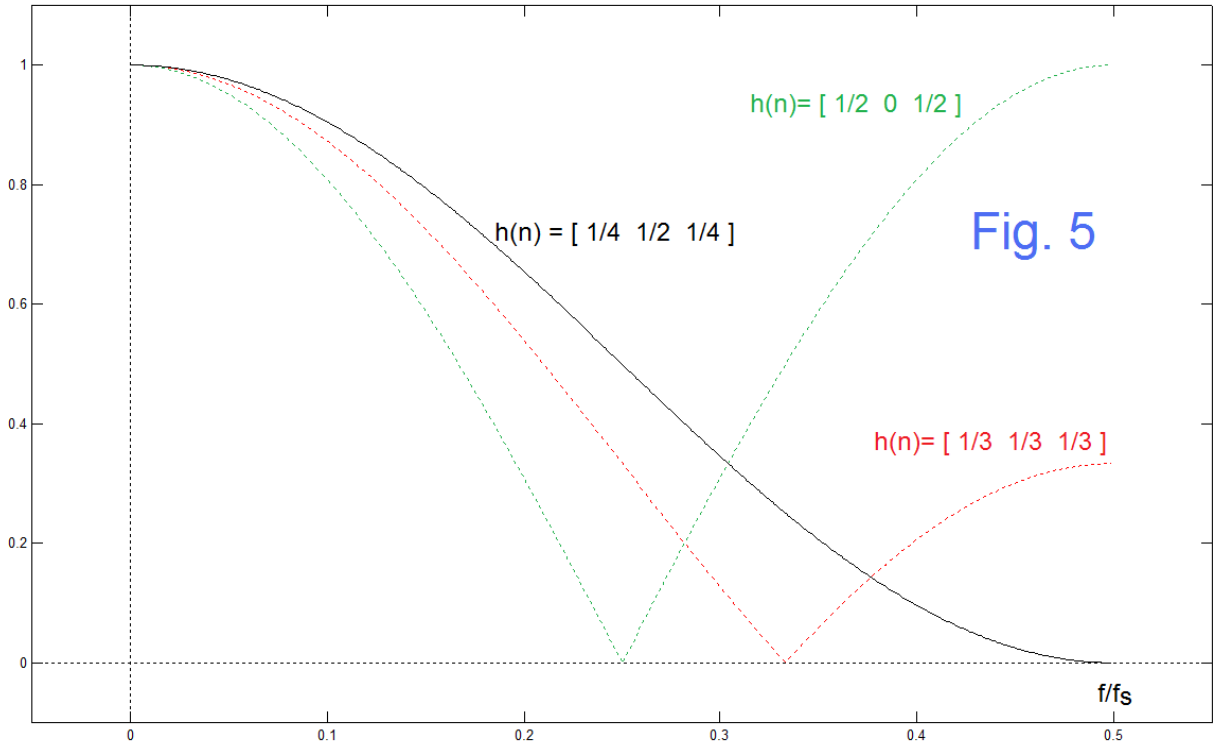
so again the FIR filter has the impulse response:

$$h(n) = \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix} \tag{19}$$

Because of symmetry, the fundamental outcomes of Fig. 4c are the same as for Fig. 4b.

FURTHER DETAILS

Fig. 5 shows the frequency response curves of three of the examples above: the three tap moving average (red), the fixed error case which is the w=2 least squares case (black) and the poor choice of straight-line fit (green). We are very familiar with the moving average frequency response, and we see the flattening of the response at



half the sampling rate (due to the second-order zero) for the black curve. The green curve just seems wrong for smoothing, which we would expect to be low-pass. In fact, it is a length-2 moving average for a doubled delay, so had a high-pass as well as a low-pass. So, the fixed-error used here results in a particular special case of the frequency response.

Above we have calculated the impulse responses on the basis of mathematics: equations (6), (13) and (19). Fine – but can we also get the same results by a different procedure, and if possible, one more intuitive than just a healthy respect for mathematics. Yes. Perhaps the refrain that follows is getting old, but it results in a remarkably convincing and intuitive understanding of the results. In particular here there is the curious result that the center term is left out of the cases of equation (13) and (19) and it is less than obvious why. So, once again, "the impulse response of a system is the response of the system to an impulse". To use this idea, we consider an impulse to enter the system in question. Our understanding of the system tells us how it responds to an arbitrary signal, so this understanding applies to a particular input signal (an impulse). In fact, the corresponding output, for this particular time-instance of the input, is often simpler to calculate, the impulse being uniquely simple. Each successive time instance follows, and the impulse response is found point-by-point as is of interest. Here there are three time instances before the input impulse exits.

Fig. 6 shows the case where the impulse response is for the preferred case. This is fit to the three cases where the impulse is in time positions 0, 1, and 2. We fit the straight line to these cases, and evaluate it at t=1. For this case we see the sequence 1/4, 1/2, and 1/4 as outputs, according to the way the line is defined to fit. This is a direct and convincing demonstration. Note also the symmetry.

Fig. 6

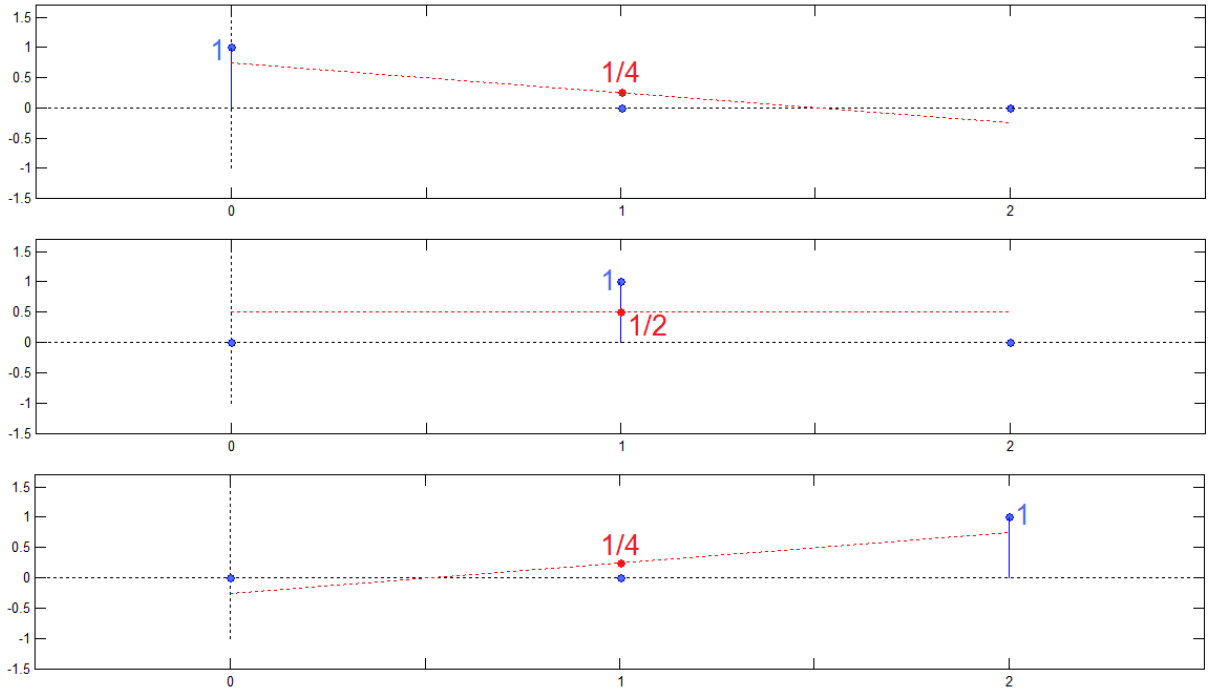


Fig. 7

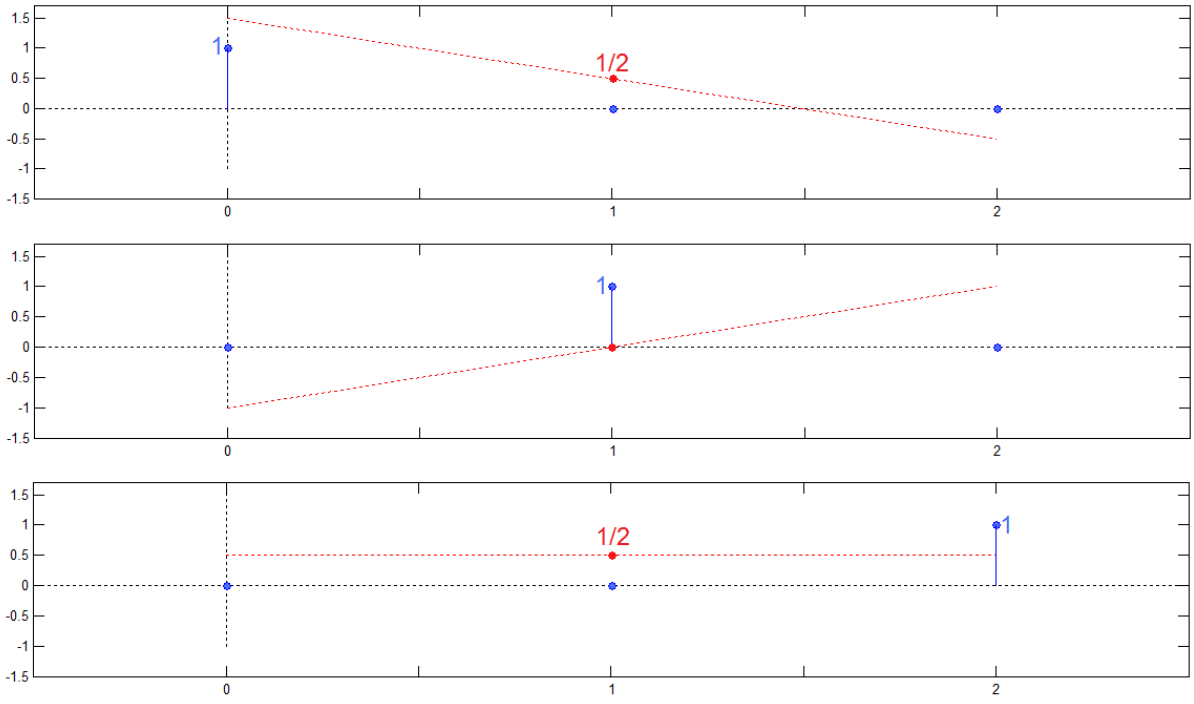


Fig. 7 shows the case where two consecutive samples are on the same side of the line. Here we see clearly (middle panel of Fig. 7) why the middle term of the impulse response is zero. The best fit goes right through 0 at t=1. For the other two times (top and bottom panels of Fig. 7), the best fit line goes through 1/2. We might well have expected the two results yielding 1/2 would be symmetric. They aren't, one of the best fit lines is flat, and the other tips. They do go through 1/2 at the required time points.

REFERENCES

[1] B. Hutchins, "Time-Series Smoothing - A Review " ***Electronotes***, Vol. 23, No. 223, November 2014
http://electronotes.netfirms.com/EN223.pdf

[2] B. Hutchins, "Time Domain Weighted Least Squares" Electronotes Application Note No. 417, Nov 22, 2014
http://electronotes.netfirms.com/AN417.pdf

[3] B. Hutchins, "Savitzky-Golay Smoothing," Electronotes Application Note No. 404, Feb 13, 2014
http://electronotes.netfirms.com/AN404.pdf

[4] B. Hutchins, "Time domain Least Squared Low-Pass Filters," Electronotes App. Note No. 318, Feb. 1992
http://electronotes.netfirms.com/AN318.PDF

[5] B. Hutchins, "Basic Elements of Digital Signal Processing: Filter Elements – Part 2" ***Electronotes***, Vol. 20, No. 198, June 2001
http://electronotes.netfirms.com/EN198.pdf