5. Design of FIR Filters

- *Reference*: Sections 7.2-7.4 of Text

- We want to address in this Chapter the issue of approximating a digital filter with a desired frequency response

$$H_d\left(e^{jw}\right)$$

using filters with finite duration.

5.1 The Windowing Technique

- When the desired frequency response $H_d\left(e^{jw}\right)$ of the system has abrupt transitions (as in the case of an ideal low pass filter), then the impulse response $h_d[n]$ has infinite duration.

- The most obvious way to approximate such a filter (system) is to truncate its impulse response to, say, $M+1$ samples. The impulse response of the new filter (assuming $h_d[n]$ is casual) is thus:

$$h[n] = \begin{cases} h_d[n] & 0 \le n \le M \\ 0 & \text{otherwise} \end{cases}.$$

- The last equation can also be rewritten as

$$h[n] = h_d[n]w_R[n]$$

where

$$w_R[n] = \begin{cases} 1 & 0 \le n \le M \\ 0 & \text{otherwise} \end{cases}$$

is a rectangular windowing function.

- It becomes apparent that one can use different windowing functions to truncate/shape the desired impulse response to a finite duration. Let $w[n]$ represents in general a windowing function of length $M+1$ samples. The truncated impulse response is

$$h[n] = h_d[n]w[n]$$

According to the Windowing Theorem in pp. 2-36 of the lecture notes, the frequency response of this approximation filter is

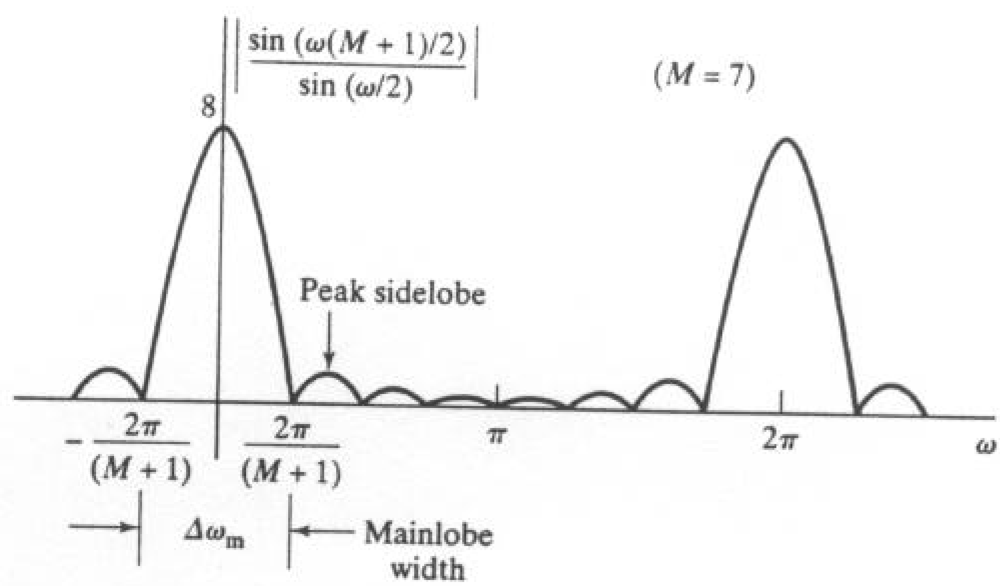
$$\begin{aligned} H\left(e^{j\omega}\right) &= \sum_{n=0}^{M} h[n]e^{-j\omega n} \\ &= \sum_{n=0}^{M} h_d[n]w[n]e^{-j\omega n} \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} H_d\left(e^{j\theta}\right)W\left(e^{j(\omega-\theta)}\right)d\theta \end{aligned}$$

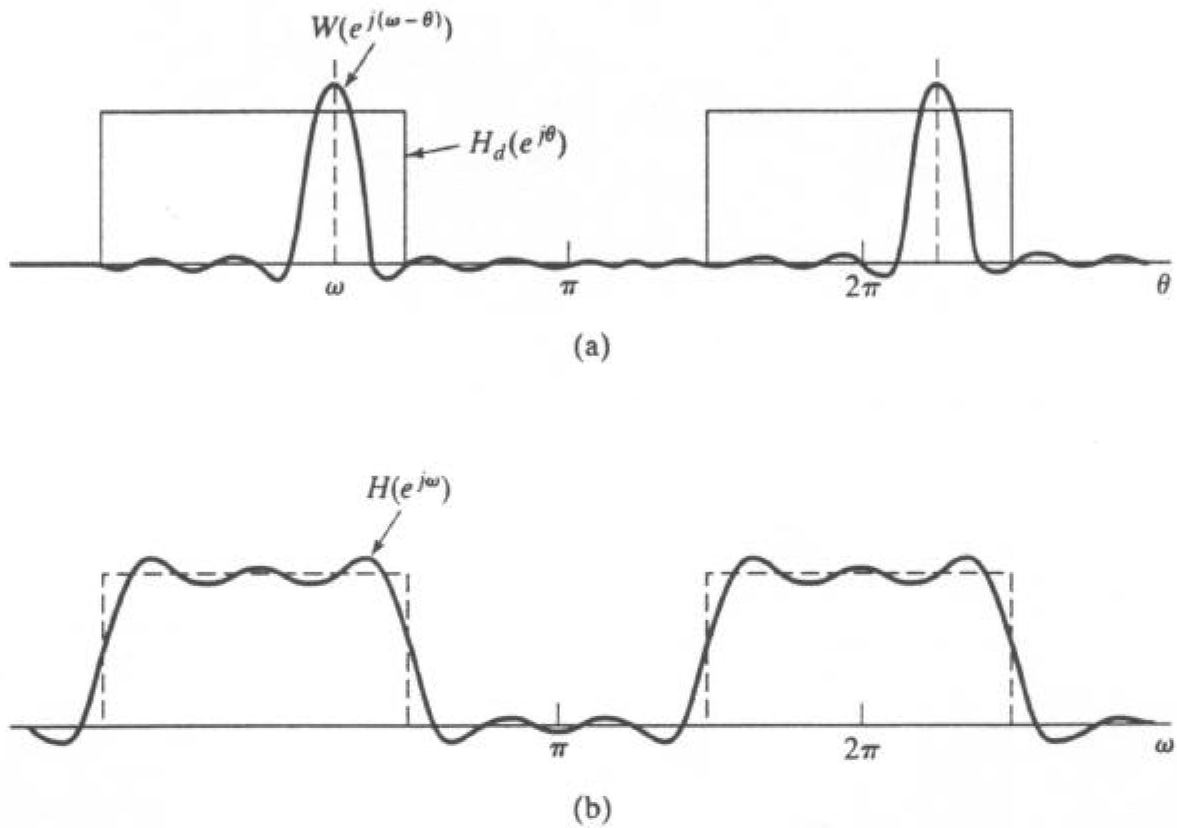
where $W\left(e^{j\omega}\right)$ is the Fourier Transform (FT) of $w[n]$.

- **Example**: When $w[n] = w_R[n]$, i.e. a rectangular window,

$$
\begin{aligned}
W\left(e^{jw}\right) &= \sum_{n=0}^{M} w[n] e^{-jwn} \\
&= \sum_{n=0}^{M} e^{-jwn} \\
&= e^{-jwM/2} \frac{\sin\left[w(M+1)/2\right]}{\sin\left(w/2\right)} \\
&= W_R\left(e^{jw}\right)
\end{aligned}
$$

A plot of $W_R\left(e^{jw}\right)$ when M = 7 is shown below. Also shown is the convolution of an ideal low pass spectrum with $W_R\left(e^{jw}\right)$.





- The convolution in the frequency domain leads to a smearing of the spectrum. The sinusoidal nature of the sinc function in $W_R\left(e^{jw}\right)$ leads to the oscillations in $H\left(e^{jw}\right)$ - the Gibbs phenomenon.

- To reduce the smearing and oscillations, we should use a windowing function whose $W\left(e^{jw}\right)$ has a (relatively)

  1. narrow mainlobe, and
  2. low sidelobes

  Note that

1. the longer the windowing function (in the time domain), the narrower is the main lobe, and
2. the "smoother" the windowing function (in the time domain), the lower the sidelobes.

It is clear that when M is fixed, we have conflicting requirements.

- ***Raised Cosine family of windows***

Many commonly used windowing functions can be written in the general form

$$w[n] = a + b\cos\left(\frac{2\pi n}{M}\right) + c\cos\left(\frac{4\pi n}{M}\right); \qquad 0 \le n \le M .$$

(a) Rectangular window

$$a = 1,\ b = 0,\ c = 0$$

(b) Hanning window

$$a = 0.5,\ b = -0.5,\ c = 0$$

(c) Hamming window

$$a = 0.54,\ b = -0.46,\ c = 0$$

(d) Blackman window

$$a = 0.42,\ b = -0.5,\ c = 0.08$$

- **Example**: The FT of the Hanning window is

$$
\begin{aligned}
W\left(e^{jw}\right) &= \sum_{n=0}^{M} w[n] e^{-jwn} \\
&= \frac{1}{2}\sum_{n=0}^{M} e^{-jwn} + \frac{1}{2}\sum_{n=0}^{M} \cos\left(\frac{2p\,n}{M}\right) e^{-jwn} \\
&= \frac{1}{2}\sum_{n=0}^{M} e^{-jwn} + \frac{1}{4}\sum_{n=0}^{M} \left\{ e^{j2p\,n/M} + e^{-j2p\,n/M} \right\} e^{-jwn} \\
&= \frac{1}{2}\sum_{n=0}^{M} e^{-jwn} + \frac{1}{4}\sum_{n=0}^{M} e^{-j(w-2p/M)n} + \frac{1}{4}\sum_{n=0}^{M} e^{-j(w+2p/M)n} \\
&= \tfrac{1}{2} W_R\left(e^{jw}\right) + \tfrac{1}{4} W_R\left(e^{j(w-2p/M)}\right) + \tfrac{1}{4} W_R\left(e^{j(w+2p/M)}\right) \\
&= \frac{1}{2} e^{-jwM/2} \frac{\sin\left[w(M+1)/2\right]}{\sin\left(w/2\right)} - \frac{1}{4} e^{-jwM/2} \frac{\sin\left[(w-2p/M)(M+1)/2\right]}{\sin\left[(w-2p/M)/2\right]} \\
&\qquad - \frac{1}{4} e^{-jwM/2} \frac{\sin\left[(w+2p/M)(M+1)/2\right]}{\sin\left[(w+2p/M)/2\right]}
\end{aligned}
$$

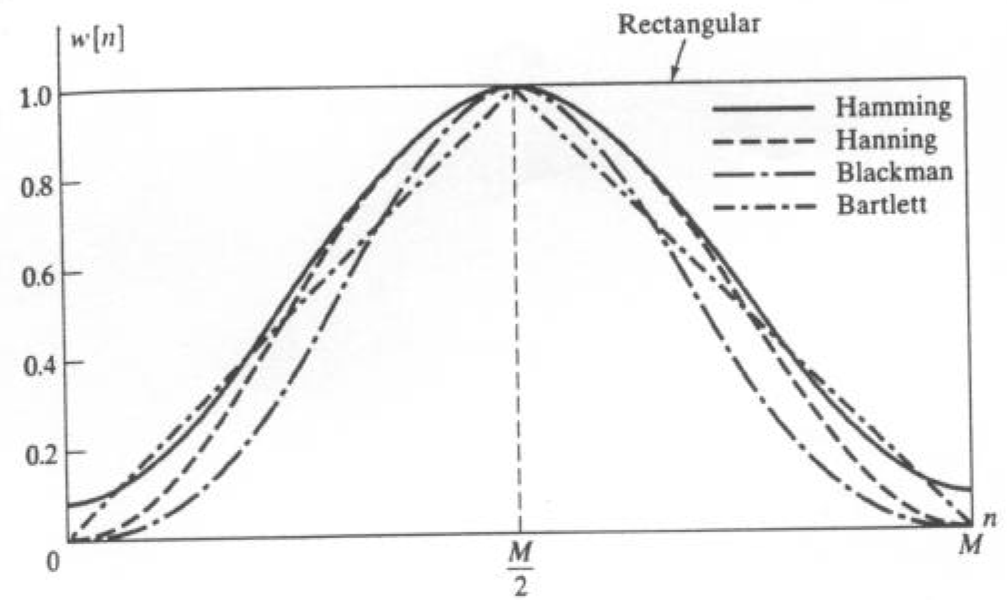
where $W_R\left(e^{jw}\right)$ is the FT of the rectangular window.

- ***Barlett (triangular window):***

$$
w[n] = \begin{cases} 2n/M & 0 \le n \le M/2 \\ 2 - 2n/M & M/2 < n \le M \\ 0 & \text{otherwise} \end{cases}
$$

This can be viewed as the convolution of two identical rectangular windows of half the length.

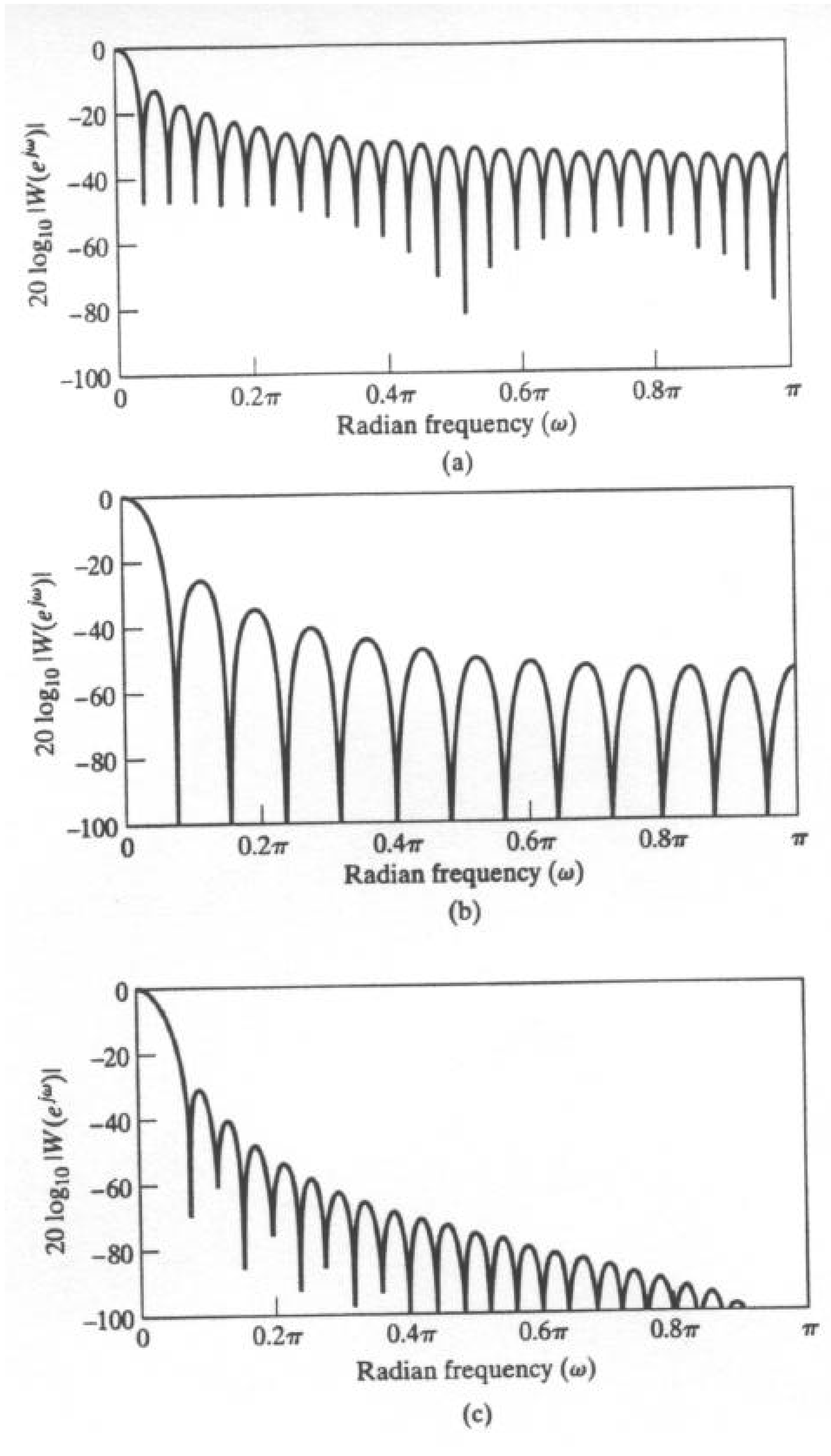
- Plot of different windowing functions in the time domain



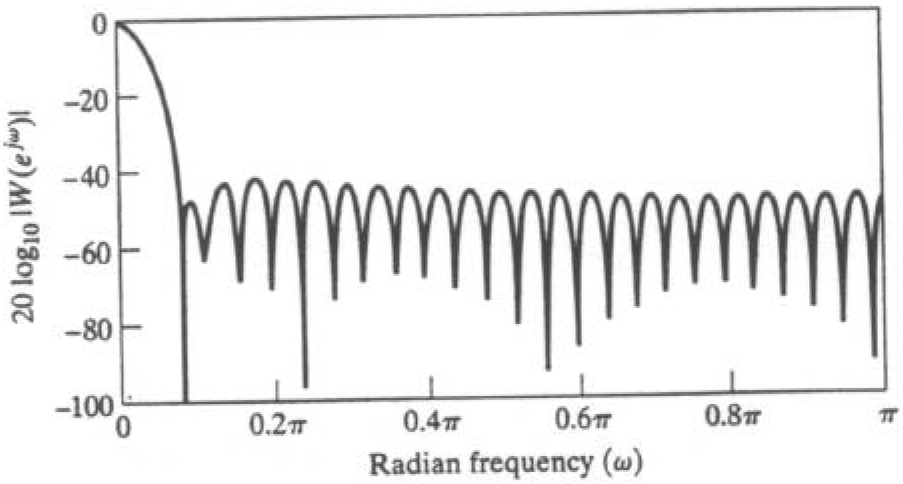
- The spectral characteristics of the different windows are shown in the next two pages. Numerical values of the peak sidelobe and width of the main lobe are summarized below.

TABLE 7.1 COMPARISON OF COMMONLY USED WINDOWS

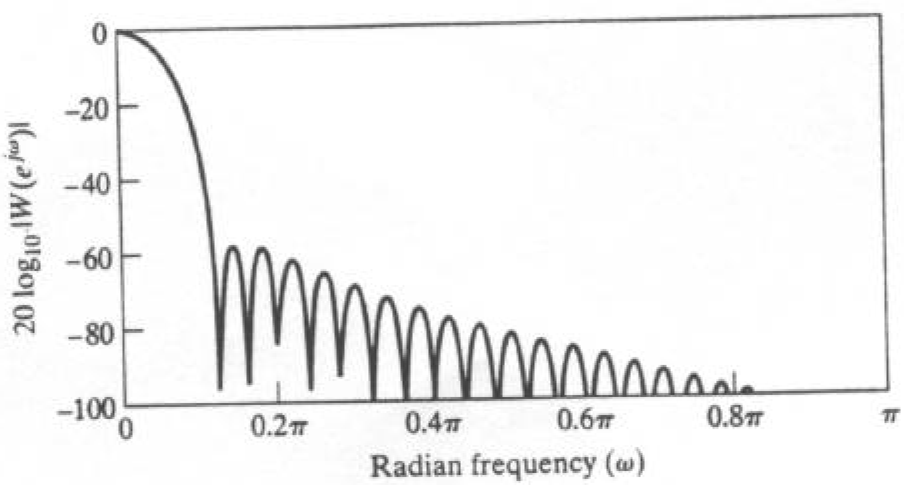
Type of Window	Peak Side-Lobe Amplitude (Relative)	Approximate Width of Main Lobe	Peak Approximation Error, $20\log_{10}\delta$ (dB)	Equivalent Kaiser Window, β	Transition Width of Equivalent Kaiser Window
Rectangular	−13	$4\pi/(M+1)$	−21	0	$1.81\pi/M$
Bartlett	−25	$8\pi/M$	−25	1.33	$2.37\pi/M$
Hanning	−31	$8\pi/M$	−44	3.86	$5.01\pi/M$
Hamming	−41	$8\pi/M$	−53	4.86	$6.27\pi/M$
Blackman	−57	$12\pi/M$	−74	7.04	$9.19\pi/M$



(a) Rectangular window, (b) Barlett window, (c) Hanning window



(d)



(e)

(d) Hamming window, (e) Blackman window

- The spectrum of the rectangular window is characterised by a relatively narrow main lobe but high sidelobes and a slow rate of decay.

- The triangular window is the convolution of 2 rectangular windows of half the size. This means

1. the spectrum of a triangular window has a sinc^2 characteristics, i.e. it decays asymptotically at twice the rate of the spectrum of a rectangular window;

2. the width of the mainlobe and sidelobes are twice that in the spectrum of a rectangular window (because of the "halfing" in the time domain).

- Recall that the spectrum of the Hanning window is the sum of 3 different frequency shifted versions of $W_R\left(e^{jw}\right)$, the spectrum of a rectangular window. Consequently it has a wider main lobe.

  Since this windowing function tapers off to zero very smoothly, it has much lower sidelobes than the rectangular window.

- The main lobe width of the Hamming window is similar to that of the Hanning window for exactly the same reason.

  The sidelobes, although lower than that of the rectangular window, decays very slowly. This is due to the discontinuities at the two edges of the window.

- The spectral mainlobe of the Blackman window is wider than that of the Hamming/Hanning windows because of the additional cosine term in the windowing function.

  Its sidelobes, however, are lower because of the smoother transition to zero.

- All the windows described above have the property that

$$w[n] = \begin{cases} w[M-n] & 0 \le n \le M \\ 0 & \text{otherwise} \end{cases}.$$

This means (assuming M is even)

$$\begin{aligned}
W\left(e^{jw}\right) &= \sum_{n=0}^{M} w[n] e^{-jwn} \\
&= \sum_{n=0}^{M/2-1} w[n]\left\{e^{-jwn} + e^{-jw(M-n)}\right\} + w[M/2] e^{-jwM/2} \\
&= e^{-jwM/2} \sum_{n=0}^{M/2-1} w[n]\left\{e^{jw(M/2-n)} + e^{-jw(M/2-n)}\right\} + w[M/2] e^{-jwM/2} \\
&= 2e^{-jwM/2} \sum_{n=0}^{M/2-1} w[n] \cos\left[w\left(n - \tfrac{M}{2}\right)\right] + w[M/2] e^{-jwM/2} \\
&= e^{-jwM/2} \left\{ w[M/2] + 2\sum_{n=0}^{M/2-1} w[n] \cos\left[w\left(n - \tfrac{M}{2}\right)\right] \right\} \\
&= e^{-jwM/2} W_e\left(e^{jw}\right)
\end{aligned}$$

where

$$W_e\left(e^{jw}\right) = w[M/2] + 2\sum_{n=0}^{M/2-1} w[n] \cos\left[w\left(n - \tfrac{M}{2}\right)\right]$$

is a real and even function in w.

Note that the phase of $W\left(e^{jw}\right)$ is linear.

In most design exercises, we can ignore the phase and focus only on $W_e\left(e^{jw}\right)$.

- If the desired impulse response is of linear phase, i.e.

$$h_d[n] = h_d[M-n],$$

then

$$H_d\left(e^{jw}\right) = H_e\left(e^{jw}\right)e^{-jwM/2}$$

where $H_e\left(e^{jw}\right)$ is a real and even function in w.

- **Exercise**: Shown that if both $w[n]$ and $h_d[n]$ are linear phase, then

$$h[n] = h_d[n]w[n]$$

is also linear phase, i.e.

$$H\left(e^{jw}\right) = A_e\left(e^{jw}\right)e^{-jwM/2}$$

where $A_e\left(e^{jw}\right)$ is a real and even function in w. Consequently during filter design, we can just focus on $A_e\left(e^{jw}\right)$ and $H_e\left(e^{jw}\right)$.

5.1.1 Kaiser Windows

- All the windows discussed so far can be approximated by an equivalent Kaiser window of the form

$$w[n] = \begin{cases} \dfrac{I_0\left[b\sqrt{1-\left(\frac{n-a}{a}\right)^2} \right]}{I_0(b)} & 0 \le n \le M \\ 0 & \text{otherwise} \end{cases},$$

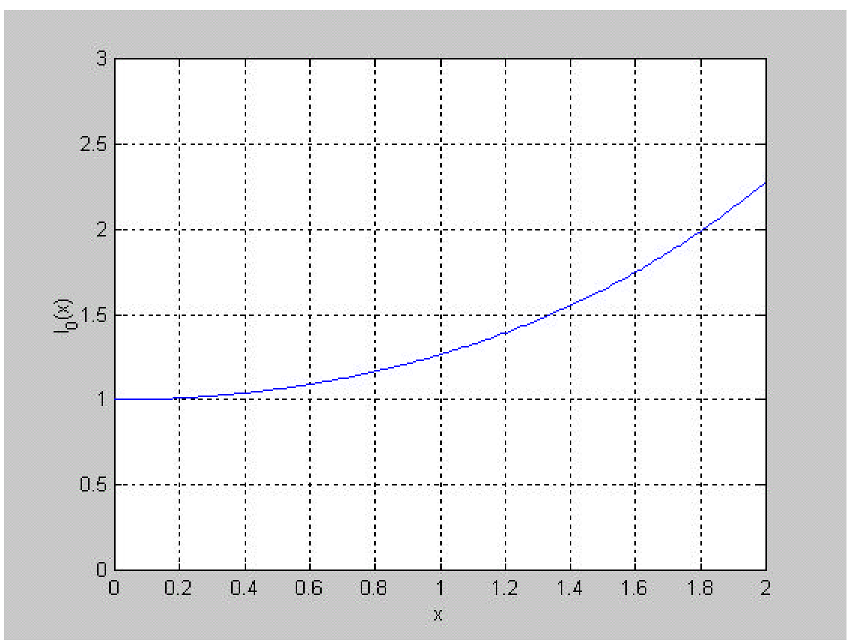
where

$$I_0(x) = \frac{1}{2p}\int_0^{2p} e^{x\cos q}\, dq$$

is the modified zero-th order Bessel function of the first kind,

$$a = M/2$$

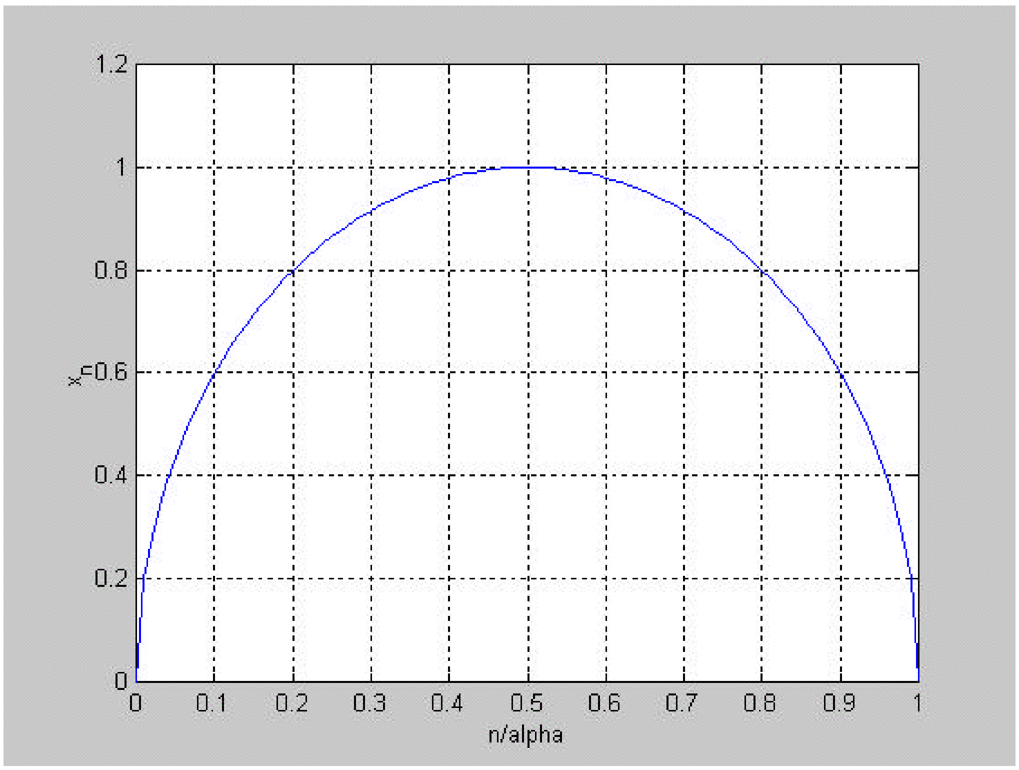
is half the window length, and b is a design parameter.



- Let the square-root term inside one of the bessel functions be denoted by

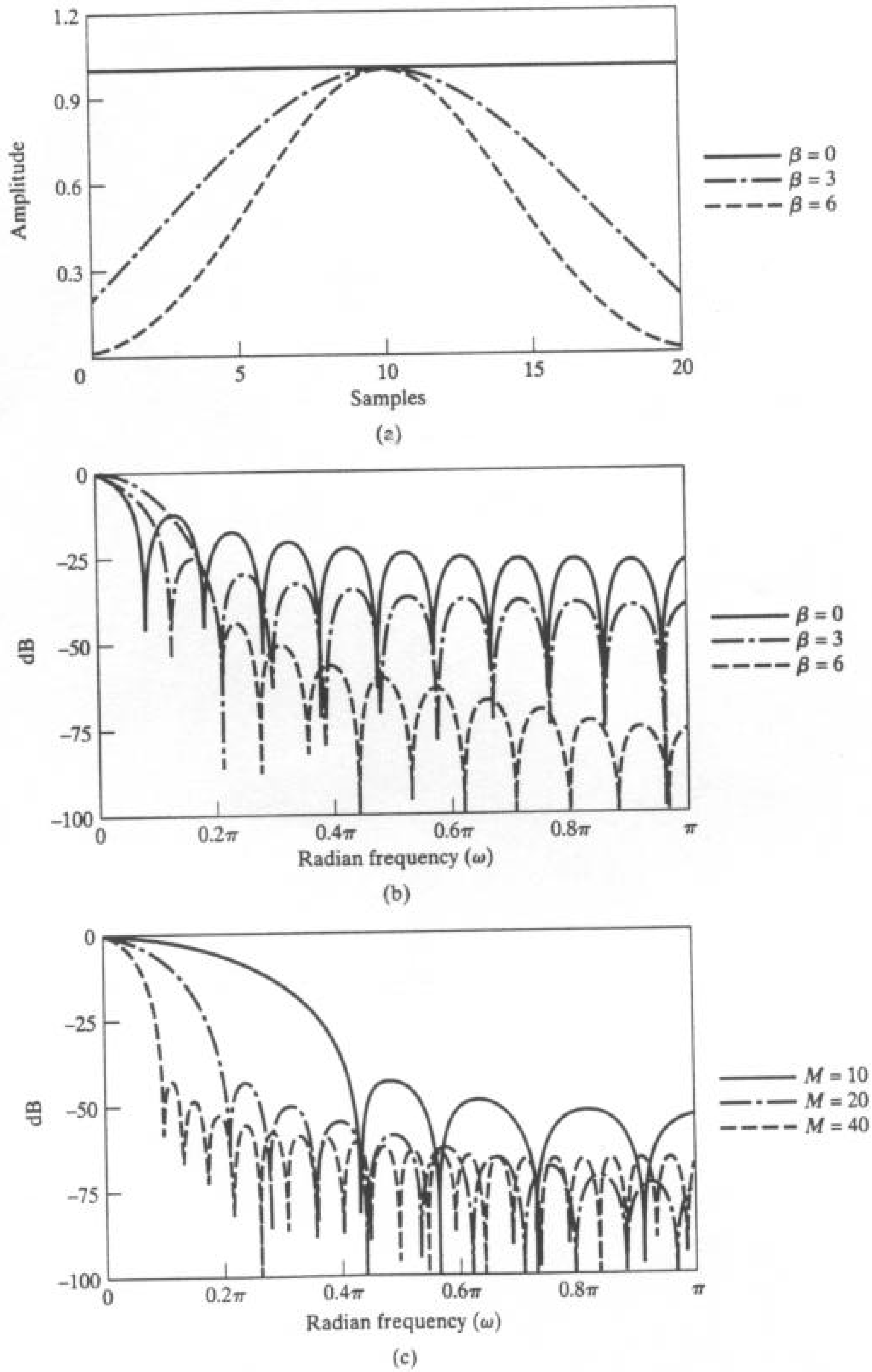
$$x_n = \sqrt{1 - \left[\frac{n-a}{a}\right]^2} \; .$$

As shown in the figure below, x_n increases monotonically from 0 to 1 in the interval $0 \le n \le a$ and decreases monotonically from 1 to 0 in the interval $a \le n \le M$. Consequently, $w[n]$ is largest at $n = a$ but decreases monotonically on ether sides of this maxima.



When $b = 0$, $w[n] = 1$ for all values of n. In other word, the Kaiser window at this value of b degenerates into a rectangular window.

- Plot of Kaiser windows in the time and frequency domains.



It is observed that:

1. the Kaiser windows have linear phase, and
2. the larger b is, the smoother is the windowing function. This implies a wider mainlobe but lower sidelobes.

The relationships between Kaiser windows and other windows are shown in the table on pp 2-7.

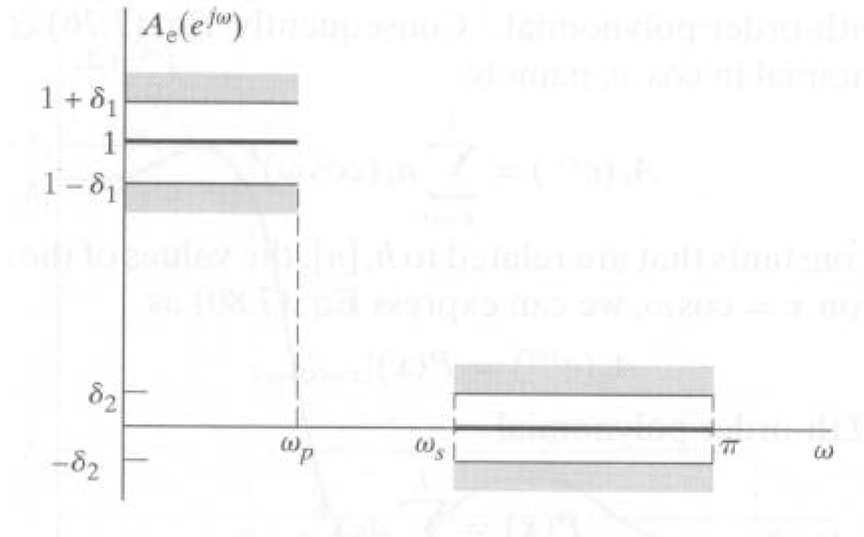
- **Design guidelines using Kaiser windows**

  - Assume a low pass filter .

  - Given that Kaiser windows have linear phase, we will only focus on the function $A_e\left(e^{jw}\right)$ in the approximation filter $H\left(e^{jw}\right)$ and the term $H_e\left(e^{jw}\right)$ in the ideal low pass filter $H_d\left(e^{jw}\right)$.

  - Normalize $H_e\left(e^{jw}\right)$ to unity for $0 \le |w| \le w_c$, where w_c is the cutoff frequency of the ideal low pass filter.

  - Specify $A_e\left(e^{jw}\right)$ in terms of a passband frequency w_p, a stop band frequency w_s, a maximum passband distortion of d_1, and a maximum stopband distortion of d_2. Note that (1) the cutoff frequency w_c of the ideal low pass filter is midway between w_p and w_s, (2) d_1 should equal to d_2 because of the nature of windowing.



- Let

$$\Delta w = w_s - w_p \,,$$

$$d = d_1 = d_2 \,,$$

and

$$A = -20\log_{10} d \;.$$

Then choose b and M according to

$$b = \begin{cases} 0.1102(A-8.7) & A > 50 \\ 0.5842(A-21)^{0.4} + 0.07886(A-21) & 21 \le A \le 50 \\ 0 & A < 21 \end{cases}$$

and

$$M = \frac{A-8}{2.285\Delta w}$$

- **Example**: Determine w_c, b, and M when $w_p = 0.4p$, $w_s = 0.6p$, and $d_1 = d_2 = d = 0.001$.

  Solution:

  Since $A = -20\log_{10} d = 60$, this means

$$b = 0.1102(60 - 8.7) = 5.6533.$$

  Since $\Delta w = w_s - w_p = 0.2p$ and $A = 60$,

$$M = \frac{60 - 8}{2.285 \times 0.2p} = 36.22 \to 37$$

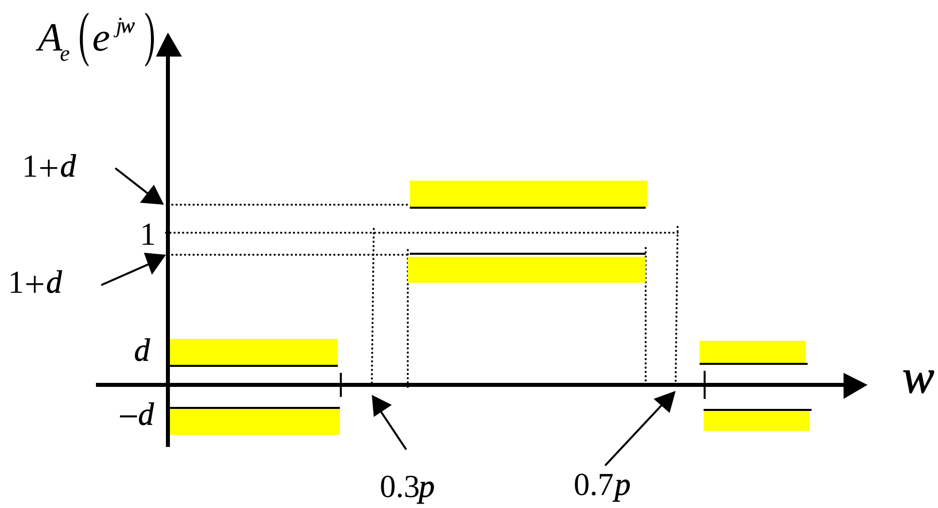
- **Example**: Consider an ideal bandpass filter with a frequency response

$$H_d\left(e^{jw}\right) = \begin{cases} e^{-jwM/2} & 0.3p \le |w| \le 0.7p \\ 0 & \text{otherwise} \end{cases}$$

  The corresponding impulse response, $h_d[n]$, is symmetrical about $a = M/2 = 25$. We want to approximate this ideal filter by multiplying it with a Kaiser window of length $M + 1 = 51$ and with $b = 3.9754$. Determine the width of the transition bands Δw and the maximum distortion d. What are the passband and stopband frequencies?

  Solution:

  - Since the length of the window is $M + 1$ and $h_d[n]$ is symmetrical about $M/2$, the approximation filter $H\left(e^{jw}\right)$ has linear phase and can be rewritten in the form $H\left(e^{jw}\right) = A_e\left(e^{jw}\right)e^{-jwM/2}$, where $A_e\left(e^{jw}\right)$ is a real and even function in w. Consequently we can just focus on the term $A_e\left(e^{jw}\right)$ and have the following specifications:



- In the time domain, the desired impulse response is

$$h_d[n] = \frac{\sin\left(0.7p\,(n-25)\right)}{p\,(n-25)} - \frac{\sin\left(0.3p\,(n-25)\right)}{p\,(n-25)},$$

  which is the difference of two ideal low pass signals.

- When the Kaiser window is applied to any of the two low pass components of $h_d[n]$, there is a peak spectral distortion of l. So the total spectral distortion is $d = 2l$ (a conservative estimate).

- Given $b = 3.9754$, it means $A = -20\log_{10} l = 45$, or $l = 5.6234\times10^{-3}$. Consequently

$$d = 2l = 1.3247\times10^{-2}.$$

- The width of each of the two transition bands in $A_e\left(e^{jw}\right)$ is governed by

$$\Delta w = \frac{A-8}{2.285M} = 0.1031p$$

This means the stopband frequencies are

$$w_{s_1} = 0.3p - 0.1031p\ /2 = 0.2485p$$

$$w_{s_2} = 0.7p + 0.1031p\ /2 = 0.7516p$$

and the passband frequencies are

$$w_{p_1} = 0.3p + 0.1031p\ /2 = 0.3516p$$

$$w_{p_2} = 0.7p - 0.1031p\ /2 = 0.6485p$$

- In general, any ideal multiband filter can be expressed as a weighted sum of ideal low pass signals of different frequencies, i.e.

$$h_d[n] = e^{-jwM/2} \sum_{k=1}^{N} w_k \frac{\sin\left(w_k\,(n-M/2)\right)}{p\,(n-M/2)}$$

with real weight coefficients w_k. The total distortion is

$$d = \left(\sum_{k=1}^{N} \left|w_k\right|\right) l$$

where l is the peak distortion when the Kaiser window is applied to each individual ideal low pass signal.

5.2 The Park-McClellan Algorithm

- The windowing method of FIR filter design is straight forward. However, the technique has 2 disadvantages:

  1. distortion in the passband and the stopband are more or less equal, and
  2. distortion is usually largest at the discontinuities of the ideal frequency response.

- Very often, we want to design a filter with different passband and stopband distortions.

  In addition, if the distortion is more evenly spread, we will be able to come up with a shorter FIR filter.

- The Park-McClellan algorithm is an iterative procedure for designing an equi-ripple FIR filter with different distortion in the pass and stop bands.

  We will focus our attention on the design of low pass filters of length

$$M = 2L$$

  using this method.

  The design of low pass filters with an odd value of M, as well as the design of other types of filters (such as bandpass, highpass, etc), require some modifications to the procedure to be described.

- We assume the desired response $H_d\left(e^{jw}\right)$ has linear phase. This is achieved when the desired impulse response, $h_d[n]$, is symmetrical about $n = M/2$, i.e.

$$h_d[n] = h_d[M-n]$$

Given that $H_d\left(e^{jw}\right)$ has linear phase, we will focus on its zero-phase equivalent, $H_e\left(e^{jw}\right)$, in the discussion. This function is a real and even in w.

- We impose the condition that the approximation filter $H\left(e^{jw}\right)$ also has linear phase. This means the approximated impulse response $h[n]$ satisfies

$$h[n] = h[M-n],$$

and the zero-phase equivalent of $H\left(e^{jw}\right)$ is

$$\begin{aligned} A_e\left(e^{jw}\right) &= h[M/2] + 2\sum_{n=0}^{M/2-1} w[n]\cos\left[w\left(n-\tfrac{M}{2}\right)\right] \\ &= h[L] + 2\sum_{m=-L}^{-1} w[m+L]\cos\left[wm\right] \\ &= a_e[0] + 2\sum_{m=-L}^{-1} a_e[m]\cos\left[wm\right] \\ &= a_e[0] + 2\sum_{n=1}^{L} a_e[n]\cos\left[wn\right] \end{aligned}$$

where

$$a_e[n] = h[n+L]$$

is a non-casual signal symmetrical about $n = 0$.

- The term $\cos(wn)$ can be written as a polynomial of degree n in $\cos(w)$. For example, $\cos(2w) = 2\cos^2(w) - 1$, and $\cos(3w) = 4\cos^3(w) - 3\cos(w)$, etc. This means the zero-phase response $A_e(e^{jw})$ can be written as a L-th order polynomial in $\cos(w)$:

$$A_e\left(e^{jw}\right) = \sum_{k=0}^{L} a_k \left(\cos(w)\right)^k$$

where the a_k s are the coefficients in this polynomial respresention.

- The Park-McCelland algorithm allows us to find the optimal $A_e(e^{jw})$ (i.e. the coefficients a_k in the polynomial representation) for fixed L, w_p, w_s, and

$$K = \frac{d_1}{d_2}.$$

The distortion, d_1 (or d_2), however is a variable. Let us define the approximation error function as

$$E(w) = W(w)\left\{H_e\left(e^{jw}\right) - A_e\left(e^{jw}\right)\right\},$$

where

$$W(w) = \begin{cases} 1/K & 0 \le w \le w_p \\ 1 & w_s \le w \le p \end{cases}$$

is a weighting function that normalizes the spectral distortion in the pass and stop bands. (notice that the weighting function is not defined for transition band, i.e. when $w_p \le w \le w_s$).

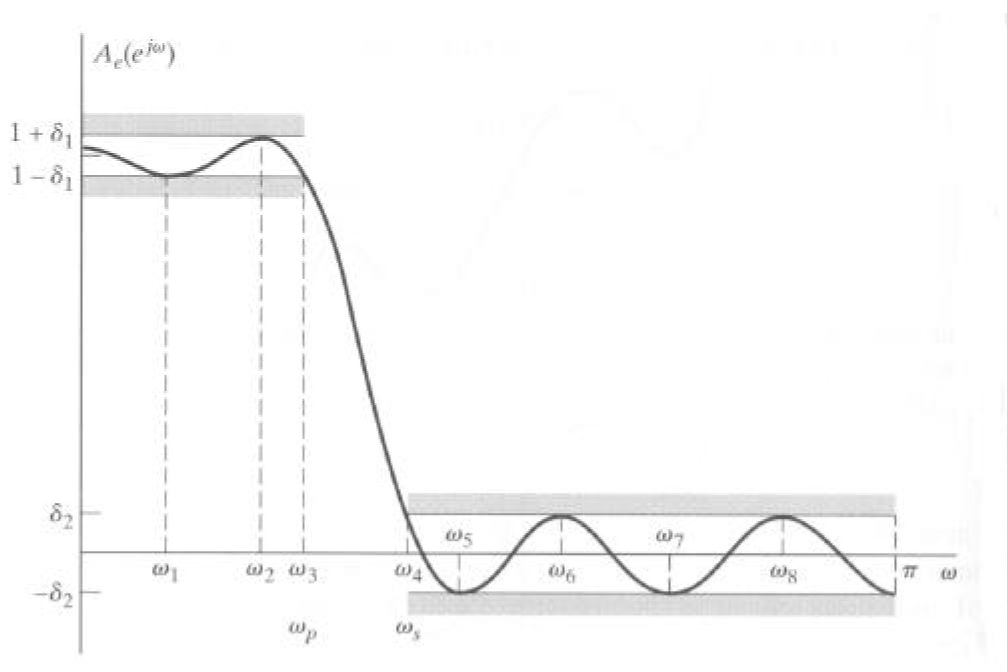
The optimality criterion used by Park and McCelland algorithm is the minimax criterion, i.e. the algorithm finds the set of filter coefficients

$$\left(a_e[0], a_e[1], ..., a[L]\right)$$

or equivalently the set of polynomial coefficients

$$\left(a_0, a_1, ..., a_L\right)$$

that minimizes the maximum of $\left|E(w)\right|$ (normalized peak distortion) over the intervals $0 \le w \le w_p$ and $w_s \le w \le p$. Example of an $A_e\left(e^{jw}\right)$ after optimization is shown below.



- From the so-called Alternation Theorem in polynomial approximation theory, we have the following results:

  There exists a UNIQUE set of polynomial coefficients

$$\left(a_0, a_1, \ldots, a_L\right)$$

and a unique set of *extremal frequencies*

$$\Omega = \left\{w_1, w_2, \ldots, w_{L+2}\right\},$$

with

$$w_1 < w_2 < \cdots < w_{L+2},$$

such that

$$E\left(w_i\right) = \left(-1\right)^{i+1} d \qquad \text{(equiripple)}$$

where d is the minimum peak distortion. Note that both the pass band frequency w_p and the stop band frequency w_s belong to Ω, and if

$$w_k = w_p,$$

then

$$w_{k+1} = w_s.$$

- The Park-McClelland algorithm uses an iterative procedure to find the set of extremal frequencies.

1. At the start of the algorithm, guess the the locations of the L+2 extremal frequencies. Call these frequencies $\hat{w}_k;\ k = 1,2,\ldots,L+2$.

2. Define x_k as

$$x_k = \cos\left(\hat{w}_k\right)$$

3. Use (7.101) and (7.102) to obtain $\hat{d}$, an estimate of the normalized peak distortion d.

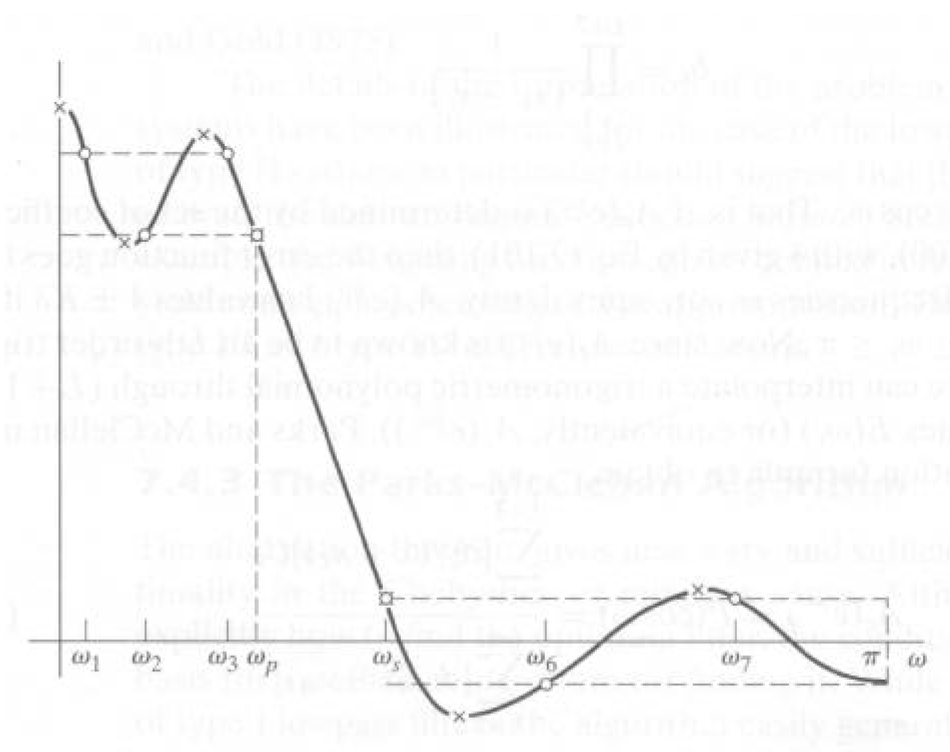
4. Use (7.103) to compute $\hat{A}_e\left(e^{jw}\right)$, an estimate of $A_e\left(e^{jw}\right)$. This function has the characteristics

$$\hat{A}_e\left(e^{jw_k}\right)=\begin{cases}1\pm\hat{d}/K & w_k \le w_p \\ \pm\hat{d} & w_k \ge w_s\end{cases}.$$

5. Locate the local maxima in

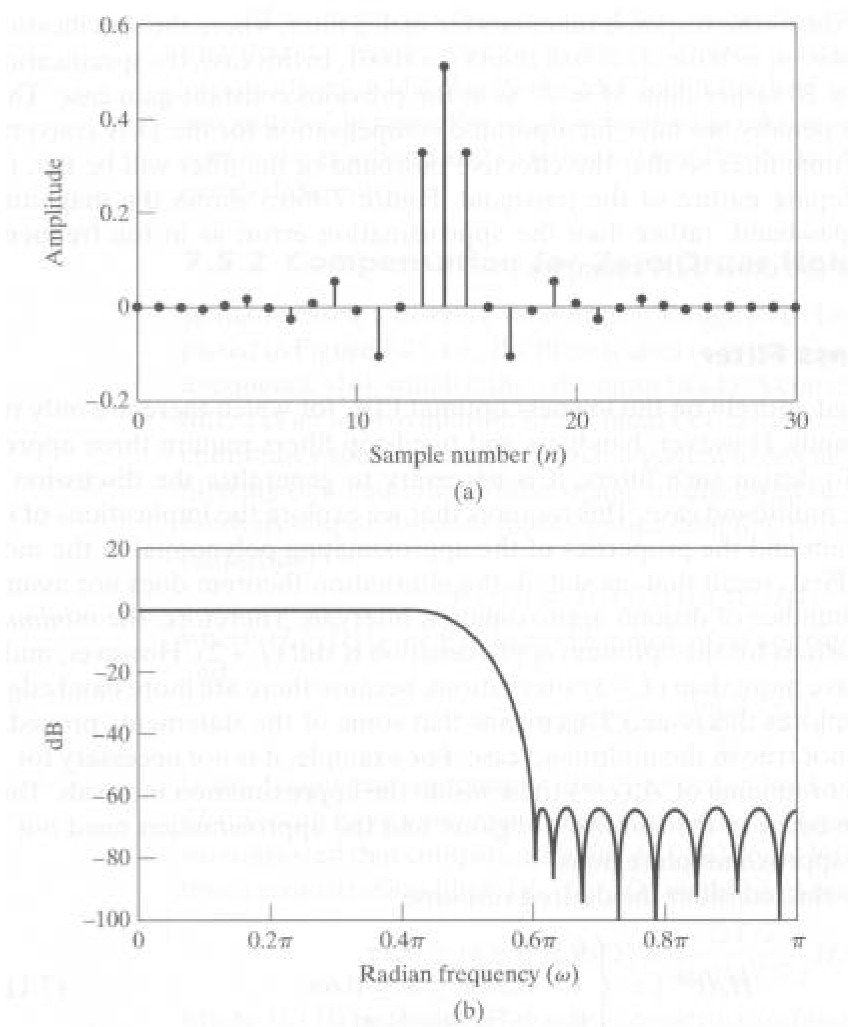
$$\hat{E}(w)=W(w)\left\{H_e\left(e^{jw}\right)-\hat{A}_e\left(e^{jw}\right)\right\}$$

where $\left|\hat{E}(w)\right|\ge\hat{d}$. If there are more than $L+2$ such maxima (w_p and w_s are counted as "maxima" even though the slopes at these two frequencies are not zero), retain only the $L+2$ largest ones. Call these frequencies $\tilde{w}_k;\ k=1,2,\ldots,L+2$.



6. If the $\tilde{w}_k$ s differ “substantianlly” from the $\hat{w}_k$ s, update the $\hat{w}_k$ s using these frequencies and go back to Step 2. If not, declare that the optimal solution has been found. In this case, make $w_k = \hat{w}_k$, $d = \hat{d}$, and $A_e\left(e^{jw}\right) = \hat{A}_e\left(e^{jw}\right)$. Take a M-point IFFT of $A_e\left(e^{jw}\right)$ to obtain $a_e[n]$. Delay $a_e[n]$ by $L = M/2$ samples to obtain $h[n]$.

- The impulse and frequency response of an approximation filter for $w_p = 0.4p$, $w_s = 0.6p$, $K = 10$, and $M = 26$ are shown below. Note that $d = d_1 = 0/0116$ and a FIR filter based on Kaiser window with a similar value of d has a length of M=38.



5.3 Implementation Structure of FIR Filters

- *Reference*: Section 6.5 of Text

- The relationship between the input $x[n]$ and the output $y[n]$ of a FIR filter of length $M+1$ is

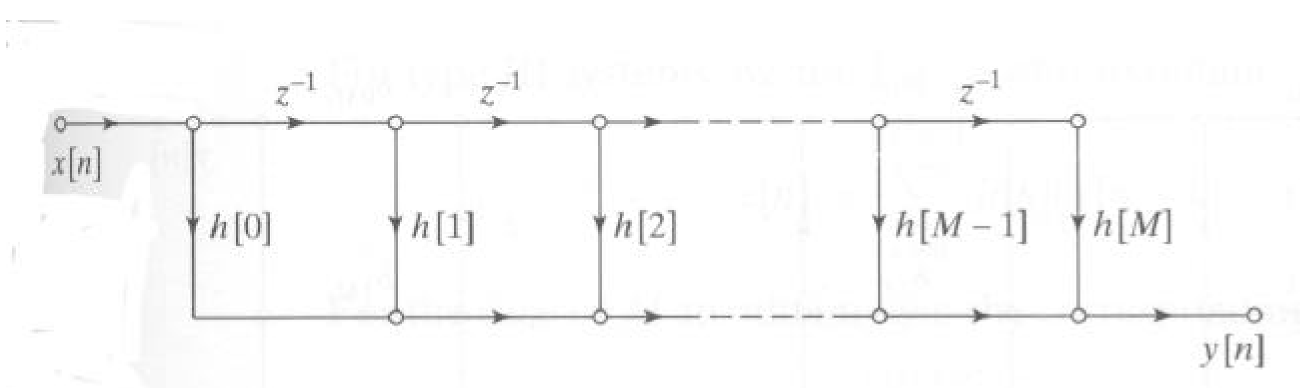
$$y[n] = \sum_{k=0}^{M} h[k]x[n-k],$$

where $h[n]$ is the impulse response of the filter, and

$$H(z) = \sum_{k=0}^{M} h[k]z^{-k}$$

is the transfer function of the filter.

- According to the above equations, a possible implementation structure of the FIR filter is



This is called the *direct form implementation*. Note that a branch in the signal flow graph with a transmittance of z^{-1} represents a delay of 1 sample. On the other hand, a branch with a transmittance of $h[k]$ means the signal at the originating node of that branch is multiplied by the constant $h[k]$. Again, as in any signal flow graph, the signal at a node is the sum of the product of the signal at an originating node and the corresponding branch transmittance.

- The transfer function $H(z)$ of a FIR filter can be expressed in terms of its zeros according to the following equation

$$H(z) = C\prod_{k=1}^{M_1}\left(1 - f_k z^{-1}\right)\prod_{k=1}^{M_2}\left(1 - g_k z^{-1}\right)\left(1 - g_k^* z^{-1}\right)$$

Here the f_k's are the real zeros and the g_k's are the complex zeros. The parameter C is a constant.

The complex zeros will always appear in conjugate pairs (as long as the impulse response is real). Moreover, $\left(1 - g_k z^{-1}\right)\left(1 - g_k^* z^{-1}\right)$ is a 2nd order polynoimal in z^{-1} with real coefficients.

Assuming $M_1 = 2K$, where K is an integer. Then we can group the $1 - f_k z^{-1}$'s into pairs and the product of any pair is always a 2nd order polynoimal in z^{-1} with real coefficients. In other word, the transfer function can always be written in product form as

$$H\left(z\right) = \prod_{k=1}^{M_s}\left(b_{0k} + b_{1k}z^{-1} + b_{2k}z^{-2}\right) = \prod_{k=1}^{M_s} B_k(z),$$

where

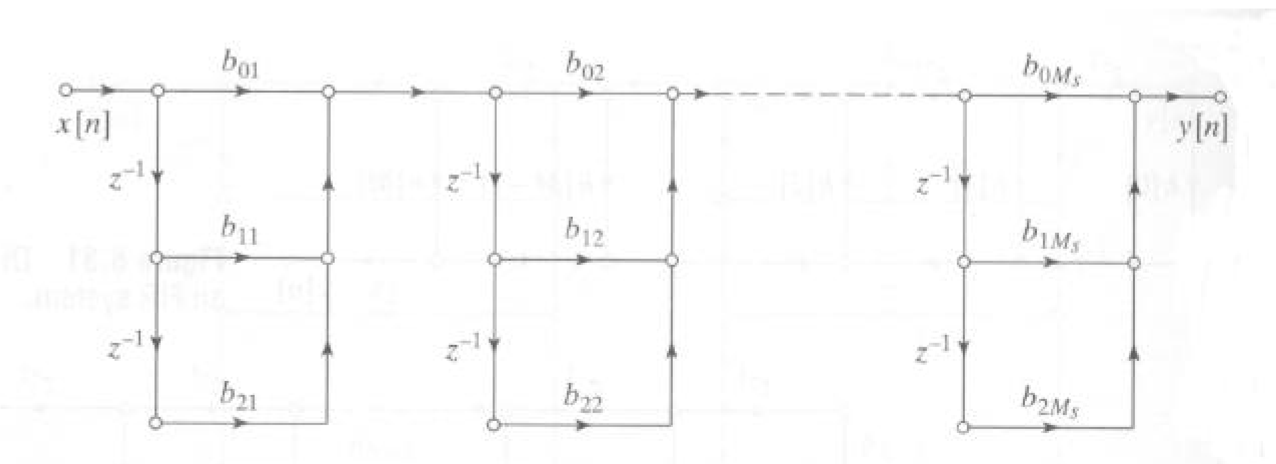
$$B_k(z) = \left(b_{0k} + b_{1k}z^{-1} + b_{2k}z^{-2}\right)$$

and

$$M_s = K + M_2 .$$

This product form suggests that the FIR filter can be considered as the cascade of M_s subsystems, with the transfer function of the k-th subsystem being equal to $B_k(z)$.

The signal flow graph below shows this Cascade form of implementing the FIR filter.



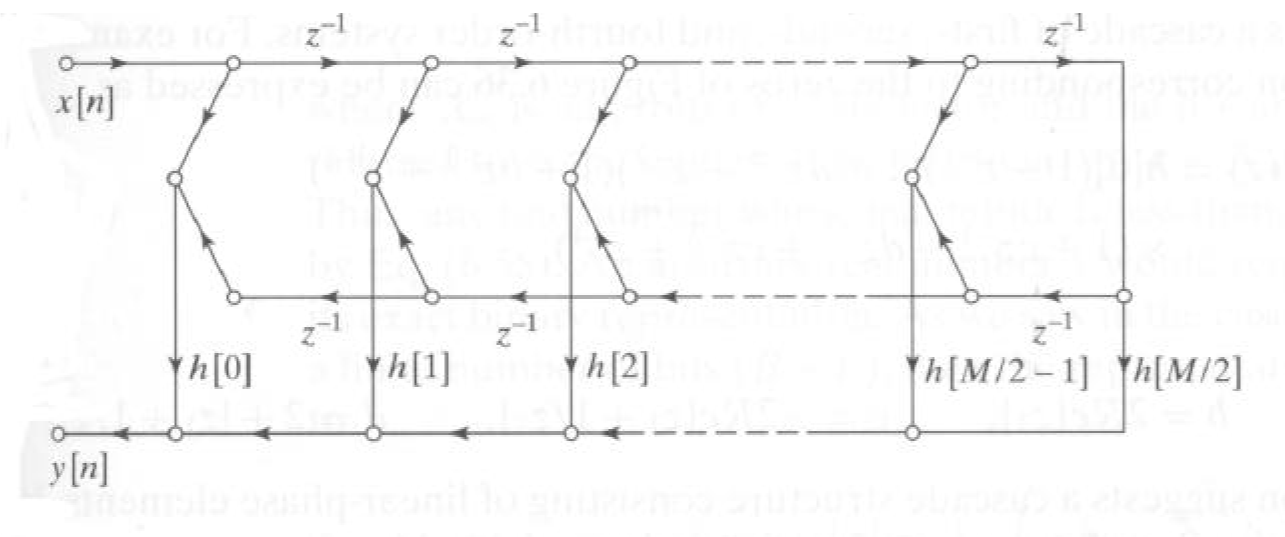
- Compared to the direct form, the cascade form is less sensitive to the quantization of the b_{jk}'s. Specifically, the poles of the transfer function will experience smaller changes.

The computational complexity, as well as the number of delay elements, of the direct and cascade forms are identical.

- All the FIR filters considered in this chapter are assumed to have linear phase. This is a direct result of the symmetry in the impulse response. Specifically

$$h[M-n] = h[n]; \qquad n = 0,1,...,M$$

Because of the symmetry, there is no need to do all the $M+1$ multiplications in the direct form implementation of the FIR filter. We can first add $x[n-k]$ and $x[n+k-M]$ together before multiplying the sum by $h[k]$. The figure below illustrates the direct form structure for a FIR linear phase system when M is an even integer.



- It is also possible to implement a linear phase FIR filter, with reduced complexity, using the cascade form.