ECE 8440 Unit 12

More on finite precision representations (See section 6.7)

Already covered: quantization error due to converting an analog signal to a digital signal.

Other types of errors due to using a finite no. of bits:

- Round-off error due to rounding of products

Example: multiplying two B+1 bit numbers produces a 2B+1 bit product (or 2B+2 bit product), and it may be necessary to round or truncate the product to the closest B+1 bit representation.

- Coefficient quantization errors.

Example: Design of digital filters produces filter coefficients that cannot be represented perfectly using a finite number of bits.

- Overflow errors

Example: In the process of implementing a digital filter, an intermediate term may be generated that is larger than the maximum value that can be accurately represented with the available number of bits.

Number Systems for Binary Representations

- Sign and magnitude (one bit indicates + or -, the remaining bits represent the magnitude.
- One's complement (binary values are negated by changing 0's to 1's and 1's to 0's)
- Two's complement (more often used) (binary values are negated by changing 0's to 1's and 1's to 0's, then adding 1 to the least significant bit position.)

We will use the two's complement system to represent scaled fractions, as shown below:

$$x = X_m\left(-b_0 + \sum_{i=1}^{\infty} b_i 2^{-i}\right) \qquad \text{(equation 6.75)}$$

where b_0 and each b_i of the terms is either 0 or 1. The range of values that can be represented this way is from $-X_m$ to (almost) X_m, where X_m is an arbitrary scale factor.

If b_0=1 and all other b_i terms are 0, then $x = -X_m$. Example: 10000000000. . . 0

If b_0= 0 and all other b_i terms are 1, then $x \to X_m$ as the number of bits $\to \infty$.

Example: 01111111111. . . 1

In general, an infinite number of bits is required for a perfect representation. If the number of bits for a two's complement representation is limited to B+1, then the quantized representation is

$$\hat{x} = Q_B(x) = X_m\left(-b_0 + \sum_{i=1}^{B} b_i 2^{-i}\right) = X_m \hat{x}_B. \qquad \text{(equation 6.76)}$$

The above representation involves an implicit binary point between the upper two bits. For example, $\hat{x}_B$ has the following form:

$\hat{x}_B = b_0.b_1b_2b_3...b_B$

If B = 2, then

1.00 = -1

1.01 = -.75

1.10 = -.5

1.11 = -.25

0.00 = 0

0.01 = .25

0.10 = .5

0.11 = .75

If a total of B bits are used (in addition to b_0), the resolution (smallest difference between values that can be represented) is

$\Delta = X_m 2^{-B}$.

The quantization error is defined as the difference between the desired value and the closest value that can be represented using B+1 bits:

$e = Q_B[x\} - x.$ (equation 6.79)

If values to be quantized are rounded to the closest two-complement representation that uses $B + 1$ bits, then the range of quantization error is

$$-(\Delta / 2) < e \leq (\Delta / 2).$$

If instead the true values are truncated (rounded down) to the next available B+1 bit two's complement representation, the range of quantization error is

$$-\Delta < e \leq 0.$$

Effect of Coefficient Quantization

Designing a digital filter produces a system function H(z) whose parameters are the multipliers of powers of z in the numerator and denominator of H(z).

When these are given an imperfect representation due to using a finite number of bits, these modified coefficient values effectively cause the locations of the poles and zeros of the designed filter to move to modified positions in the z-plane. This, in turn, causes the frequency response of the filter to change.

Example: (12-th order bandpass elliptic filter)

Design specifications:

$$0.99 \leq |H(e^{j\omega})| \leq 1.01, \qquad 0.3\pi \leq \omega \leq 0.4\pi$$

$$|H(e^{j\omega})| \leq 0.01, \qquad \omega \leq 0.29\pi \quad \text{and} \quad 0.41\pi \leq \omega \leq \pi$$

Table 6.1 shows the "unquantized" values (64-bit floating point, 15 decimal digital accuracy) of the a_k and b_k coefficients of the designed filter.

TABLE 6.1 UNQUANTIZED DIRECT-FORM COEFFICIENTS FOR A 12TH-ORDER ELLIPTIC FILTER

k	b_k	a_k
0	0.01075998066934	1.00000000000000
1	-0.05308642937079	-5.22581881365349
2	0.16220359377307	16.78472670299535
3	-0.34568964826145	-36.88325765883139
4	0.57751602647909	62.39704677556246
5	-0.77113336470234	-82.65403268814103
6	0.85093484466974	88.67462886449437
7	-0.77113336470234	-76.47294840588104
8	0.57751602647909	53.41004513122380
9	-0.34568964826145	-29.20227549870331
10	0.16220359377307	12.29074563512827
11	-0.05308642937079	-3.53766014466313
12	0.01075998066934	0.62628586102551

Table 6.2 shows the locations of poles and zeros for the filter with unquantized coefficients.

TABLE 6.2 ZEROS AND POLES OF UNQUANTIZED 12TH-ORDER ELLIPTIC FILTER.

k	$\lvert c_k \rvert$	$\angle c_k$	$\lvert d_k \rvert$	$\angle d_{1k}$
1	1.0	± 1.65799617112574	0.92299356261936	±1.15956955465354
2	1.0	± 0.65411612347125	0.92795010695052	±1.02603244134180
3	1.0	± 1.33272553462313	0.96600955362927	±1.23886921536789
4	1.0	± 0.87998582176421	0.97053510266510	±0.95722682653782
5	1.0	± 1.28973944928129	0.99214245914242	±1.26048962626170
6	1.0	± 0.91475122405407	0.99333628602629	±0.93918174153968

Figure 6.48 provides a plot of the poles and zeros for a direct form (non-factored) Implementation for

(a) the case of unquantized coefficients and

(b) for the case of coefficients represented using 16-bit accuracy

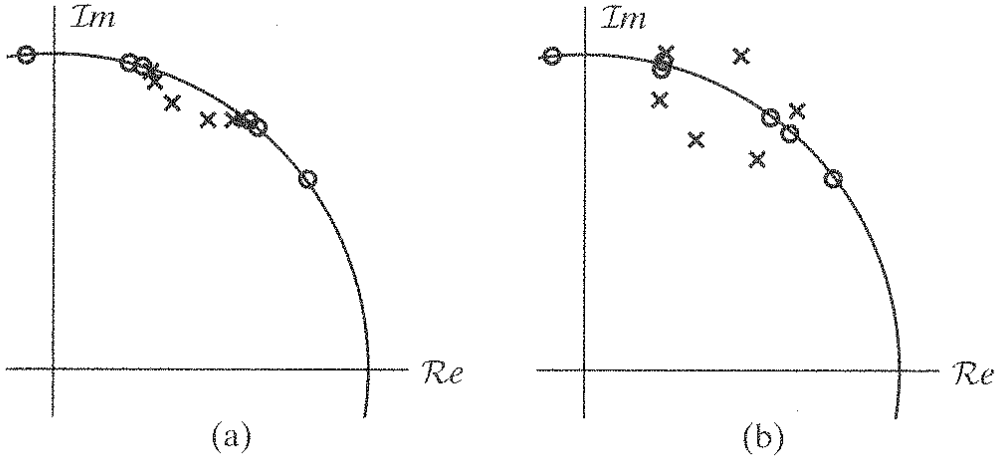


Figure 6.48 IIR coefficient quantization example. (a) Poles and zeros of $H(z)$ for unquantized coefficients. (b) Poles and zeros for 16-bit quantization of the direct form coefficients.

Note that for the case of 16-bit coefficient, the filter has become unstable, poles have moved outside the unit circle in the z-plane.

If the filter is factored into the product of six 2^{-nd} order sections, the unquantized coefficients are shown below

TABLE 6.3 UNQUANTIZED CASCADE-FORM COEFFICIENTS FOR A 12TH-ORDER ELLIPTIC FILTER

k	a_{1k}	a_{2k}	b_{0k}	b_{1k}	b_{2k}
1	0.737904	-0.851917	0.137493	0.023948	0.137493
2	0.961757	-0.861091	0.281558	-0.446881	0.281558
3	0.629578	-0.933174	0.545323	-0.257205	0.545323
4	1.117648	-0.941938	0.706400	-0.900183	0.706400
5	0.605903	-0.984347	0.769509	-0.426879	0.769509
6	1.173028	-0.986717	0.937657	-1.143918	0.937657

If these coefficients are quantized to a 16-bit two-complement representations, they have the following values:

TABLE 6.4 SIXTEEN-BIT QUANTIZED CASCADE-FORM COEFFICIENTS FOR A 12TH-ORDER ELLIPTIC FILTER

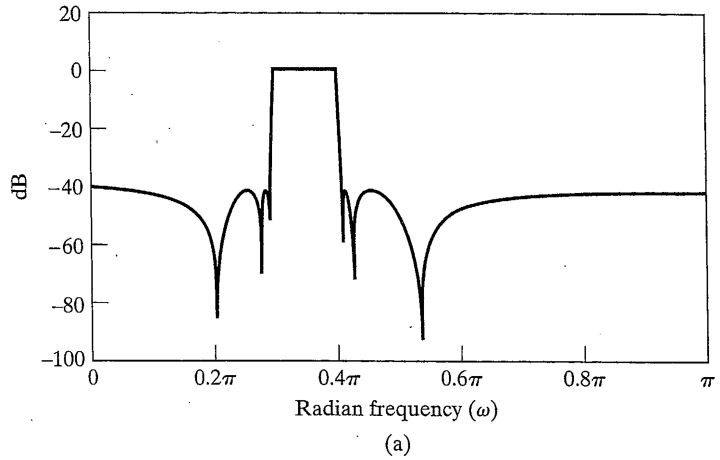
k	a_{1k}	a_{2k}	b_{0k}	b_{1k}	b_{2k}
1	24196×2^{-15}	-27880×2^{-15}	17805×2^{-17}	3443×2^{-17}	17805×2^{-17}
2	31470×2^{-15}	-28180×2^{-15}	18278×2^{-16}	-29131×2^{-16}	18278×2^{-16}
3	20626×2^{-15}	-30522×2^{-15}	17556×2^{-15}	-8167×2^{-15}	17556×2^{-15}
4	18292×2^{-14}	-30816×2^{-15}	22854×2^{-15}	-29214×2^{-15}	22854×2^{-15}
5	19831×2^{-15}	-32234×2^{-15}	25333×2^{-15}	-13957×2^{-15}	25333×2^{-15}
6	19220×2^{-14}	-32315×2^{-15}	15039×2^{-14}	-18387×2^{-14}	15039×2^{-14}

For better comparison, the coefficients for the k = 1 cascade section are shown below for both the "unquantized" case and the 16-bit quantized case:

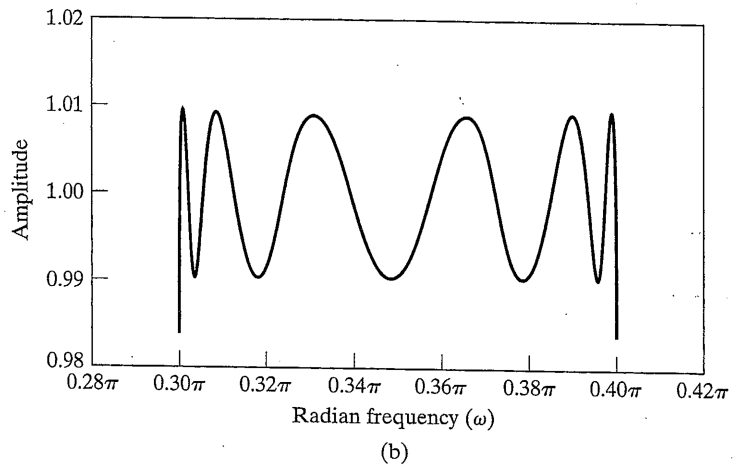
	a_{11}	a_{21}	b_{01}	b_{11}	b_{21}
unquantized	0.737904	-0.851917	0.137493	0.023948	0.137493
16-bit 2's comp.	0.738403	-0.850830	0.135841	0.026268	0.135841

The following figure shows the frequency of the filter for the following cases:

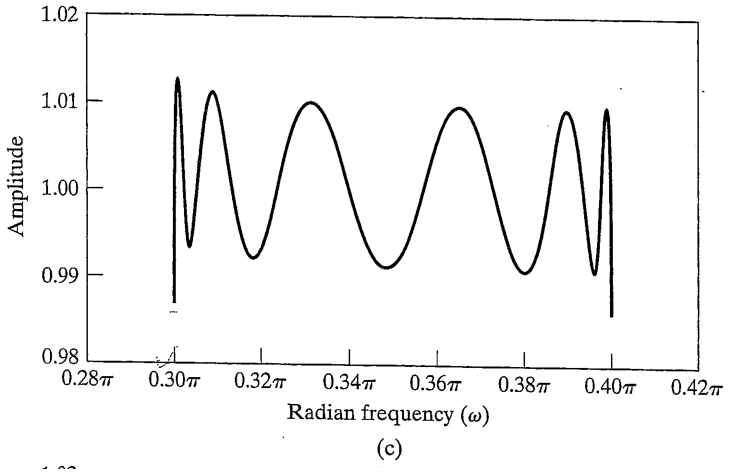
- "unquantized" coefficients (parts a and b of figure)
- 16-bit coefficients, with filter implemented in cascade form (part c of figure)
- 16-bit coefficients, with filter implemented in parallel form (part d of figure)
- 16-bit coefficients, with filter implemented in direct form (part e of figure)



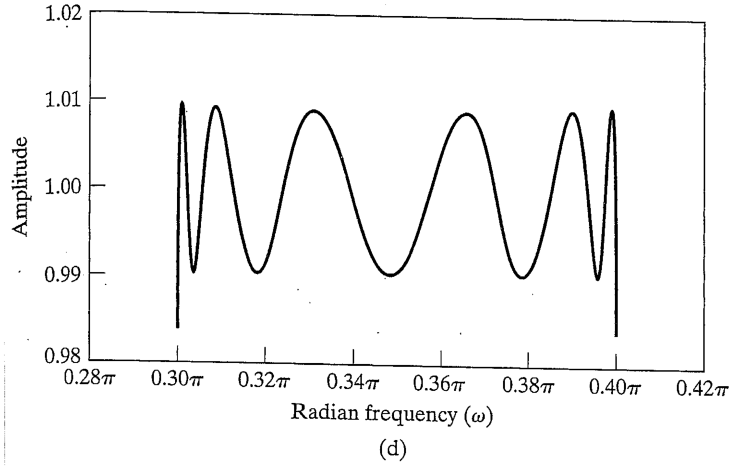
unquantized case



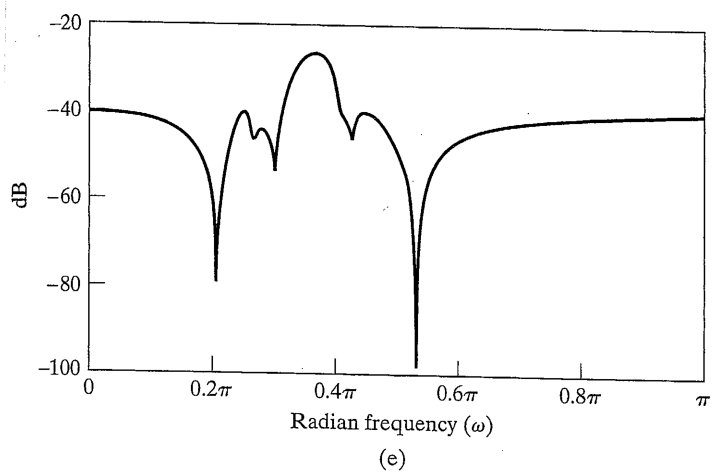
unquantized case



16-bits, cascade form



16-bits, parallel form



16-bits, direct form

Key points from figure:

- Minor degradation in response for 16-bit cascade and 16-bit parallel forms.
- Major degradation in response for 16-bit direct form.

The above is an example of the following general case:

If poles or zeros are tightly clustered, then small errors in the filter coefficients can cause a significant shift in pole or zero positions (and therefore major changes in the frequency response.) Therefore, it is almost always best to implement any IIR filter in the cascade or parallel form.

Note: Because they implement different complex-conjugate poles and zeros independently, the cascade and parallel forms are generally much less sensitive to coefficient quantization errors, as compared to the direct form.

Possible location of poles and zeros using quantized coefficients.

Consider a section order filter with poles at

$z = re^{j\theta}$ and $z = re^{-j\theta}$.

The denominator polynomial can be written as

$$(1 - z^{-1}re^{j\theta})(1 - z^{-1}re^{-j\theta}) = 1 - 2r\cos(\theta)z^{-1} + r^2z^{-2}$$

The direct form implementation of this filter is shown below:

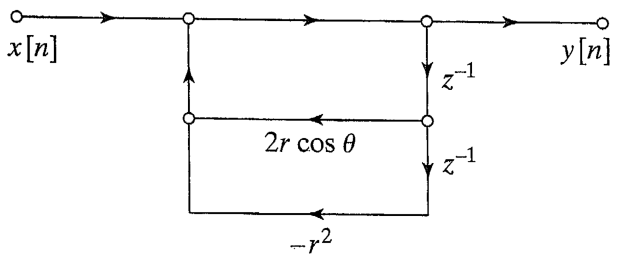
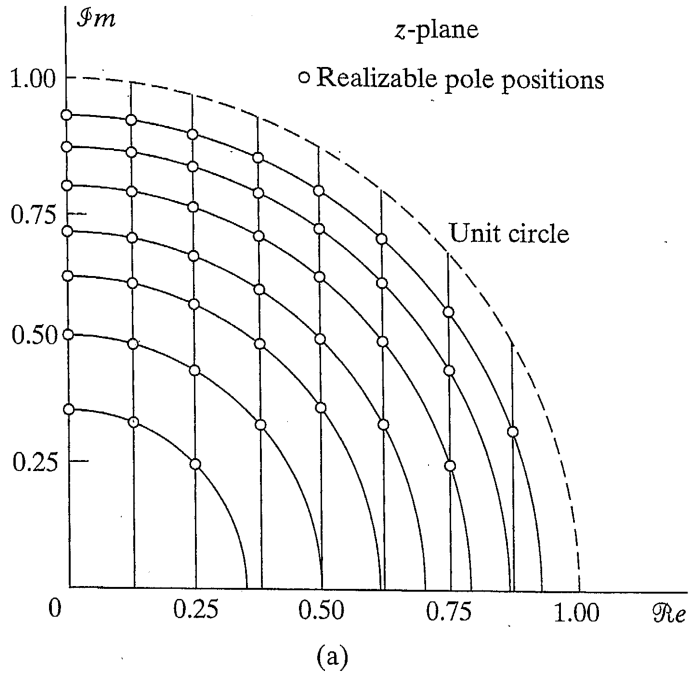


Figure 6.49 Direct-form implementation of a complex-conjugate pole pair.

If the coefficients $-2r\cos(\theta)$ and r^2 are represented using 4-bit accuracy, the possible location of poles in the z-plane are shown below (for the first quadrant in the z-plane):



(a)

In the above plot, note that the spacing of possible pole locations is not uniform. If 7-bit quantization is used to represent the coefficients for this filter, the possible pole locations become much more dense, as shown in the figure below:

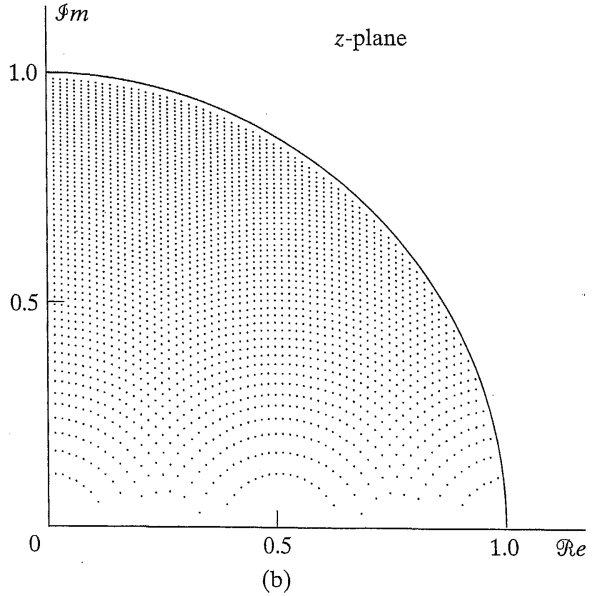


Figure 6.50 Pole-locations for the 2nd-order IIR direct-form system of Figure 6.49. (a) Four-bit quantization of Coeficients (b) Seven-bit quantization

Another implementation of a 2-pole section is shown below:

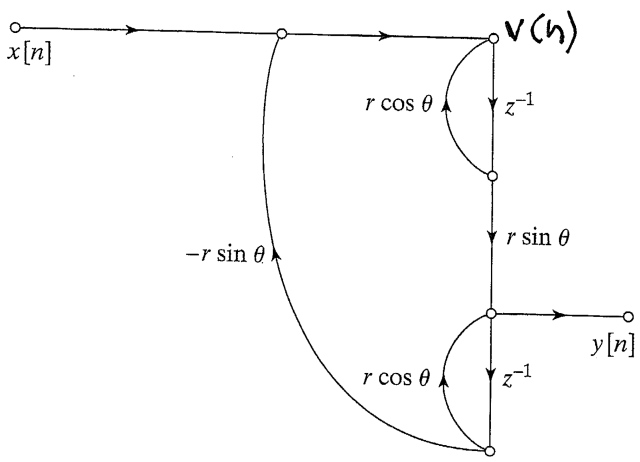


Figure 6.51 Coupled-form implementation of a complex-conjugate pole pair.

This implementation could be implemented using the following two difference equations:

$$v(n) = x(n) - r\sin(\theta)y(n-1) + r\cos(\theta)v(n-1)$$

and

$$y(n) = r\sin(\theta)v(n-1) + r\cos(\theta)y(n-1)$$

As seen above, the coefficients of this implementation are $r\sin(\theta)$ and $r\cos(\theta)$.

If these coefficients are quantized to 4-bit accuracy (part a of figure) or to 7-bit accuracy (part b of figure), the possible pole locations are shown below:

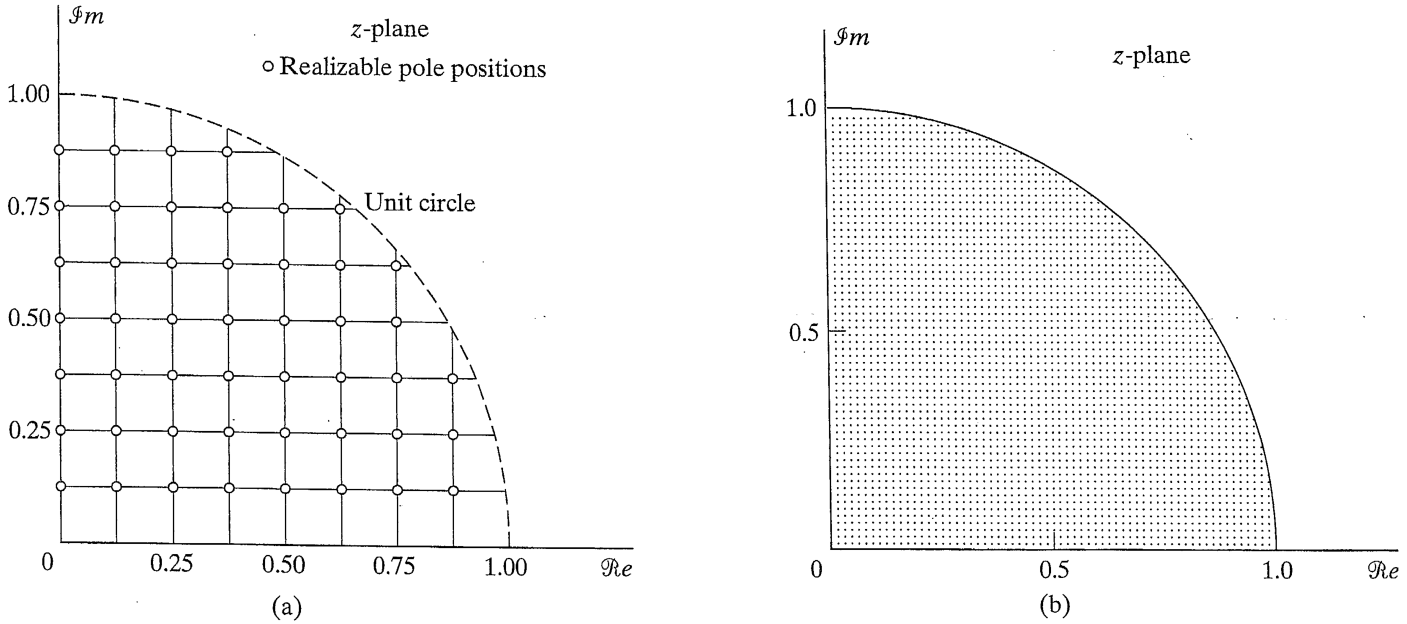


Figure 6.52 Pole locations for coupled-form 2nd-order IIR system of Figure 6.51.
(a) Four-bit quantization of coefficients (b) Seven-bit quantization

Effects of Coefficient Quantization in FIR Filters

In FIR systems, the filter coefficients are one and the same as the h(n) values, since for a causal system M-th order system,

$$y(n) = \sum_{k=0}^{M} b_k x(n-k)$$

and

$$y(n) = \sum_{k=0}^{M} h(k) x(n-k).$$

We may relate the desired h(n) values with the quantized values $\hat{h}(n)$ using

$$\hat{h}(n) = h(n) + \Delta h(n)$$

The system implemented using the quantized coefficients can then be expressed as

$$\hat{H}(z) = \sum_{n=0}^{M} \hat{h}(n) z^{-n} = H(z) + \Delta H(z)$$

where

$$\Delta H(z) = \sum_{k=0}^{M} \Delta h(k) z^{-n}.$$

The following figure provides a block diagram representation of coefficient quantization for the case of FIR filters:

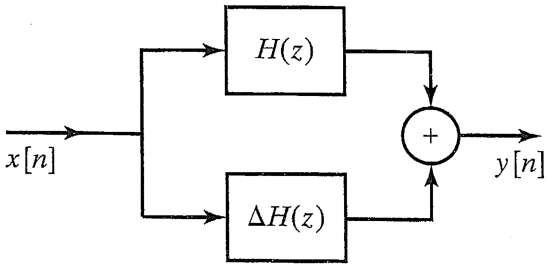


Figure 6.53 representation of coefficient quantization in FIR systems

Research has shown that coefficient quantization in FIR filters has the most effect if zeros of the filter are close together. (This is consistent with the case of zeros and poles of IIR filters.) Note: Since zeros of linear phase FIR filters are typically not as tightly clustered as zeros of IIR filters, it is common practice to implement FIR filters directly, without factoring into lower order sections.

Example: Effect of Quantization of Coefficients in Optimum FIR Lowpass Filter

Design specifications: $0.99 \le |H(e^{j\omega})| \le 1.01, \quad 0 \le \omega \le 0.4\pi$

$|H(e^{j\omega})| \le 0.001, \quad 0.6\pi \le \omega \le \pi$

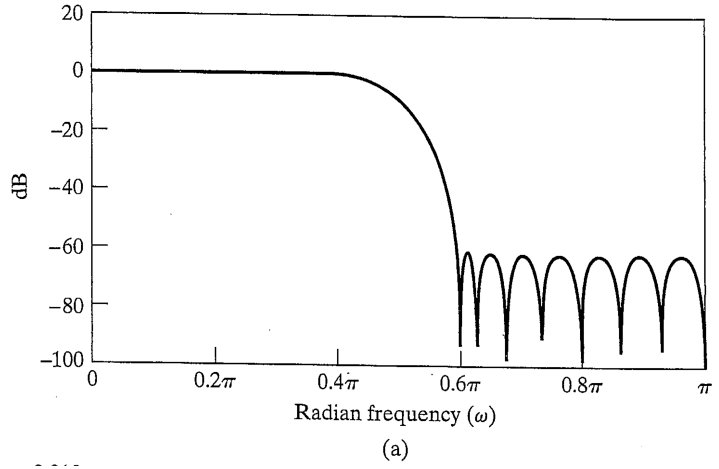
The lowest order FIR filter that satisfies these specification is M = 27.

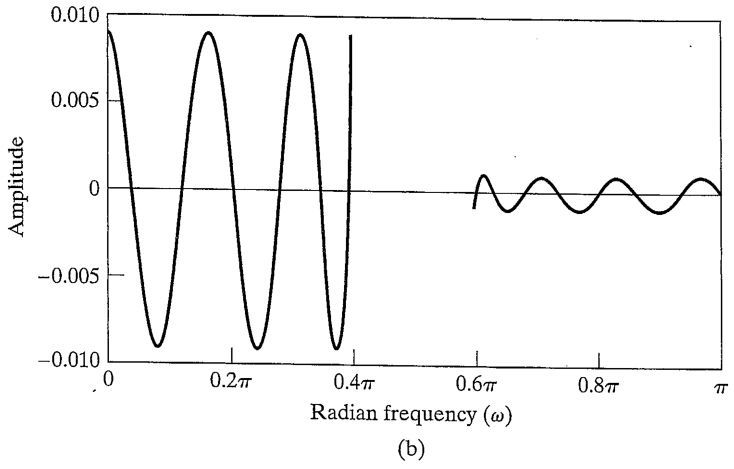
The table below provides a comparison of the "unquantized" coefficients with coefficients quantized to 16 bits, 14 bits, 13 bits, and 8 bits.

TABLE 6.5 UNQUANTIZED AND QUANTIZED COEFFICIENTS FOR AN OPTIMUM FIR LOWPASS FILTER ($M = 27$)

Coefficient	Unquantized	16 bits	14 bits	13 bits	8 bits
$h[0] = h[27]$	1.359657×10^{-3}	45×2^{-15}	11×2^{-13}	6×2^{-12}	0×2^{-7}
$h[1] = h[26]$	-1.616993×10^{-3}	-53×2^{-15}	-13×2^{-13}	-7×2^{-12}	0×2^{-7}
$h[2] = h[25]$	-7.738032×10^{-3}	-254×2^{-15}	-63×2^{-13}	-32×2^{-12}	-1×2^{-7}
$h[3] = h[24]$	-2.686841×10^{-3}	-88×2^{-15}	-22×2^{-13}	-11×2^{-12}	0×2^{-7}
$h[4] = h[23]$	1.255246×10^{-2}	411×2^{-15}	103×2^{-13}	51×2^{-12}	2×2^{-7}
$h[5] = h[22]$	6.591530×10^{-3}	216×2^{-15}	54×2^{-13}	27×2^{-12}	1×2^{-7}
$h[6] = h[21]$	-2.217952×10^{-2}	-727×2^{-15}	-182×2^{-13}	-91×2^{-12}	-3×2^{-7}
$h[7] = h[20]$	-1.524663×10^{-2}	-500×2^{-15}	-125×2^{-13}	-62×2^{-12}	-2×2^{-7}
$h[8] = h[19]$	3.720668×10^{-2}	1219×2^{-15}	305×2^{-13}	152×2^{-12}	5×2^{-7}
$h[9] = h[18]$	3.233332×10^{-2}	1059×2^{-15}	265×2^{-13}	132×2^{-12}	4×2^{-7}
$h[10] = h[17]$	-6.537057×10^{-2}	-2142×2^{-15}	-536×2^{-13}	-268×2^{-12}	-8×2^{-7}
$h[11] = h[16]$	-7.528754×10^{-2}	-2467×2^{-15}	-617×2^{-13}	-308×2^{-12}	-10×2^{-7}
$h[12] = h[15]$	1.560970×10^{-1}	5115×2^{-15}	1279×2^{-13}	639×2^{-12}	20×2^{-7}
$h[13] = h[14]$	4.394094×10^{-1}	14399×2^{-15}	3600×2^{-13}	1800×2^{-12}	56×2^{-7}

The approximation error is shown below for the cases of 16 bits, 14 bits, 13 bits, and 8 bits, along with the approximation error and the frequency response for the unquantized case.





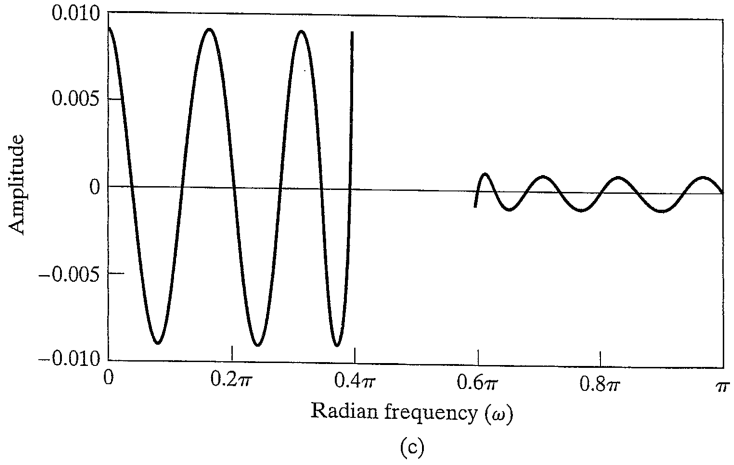
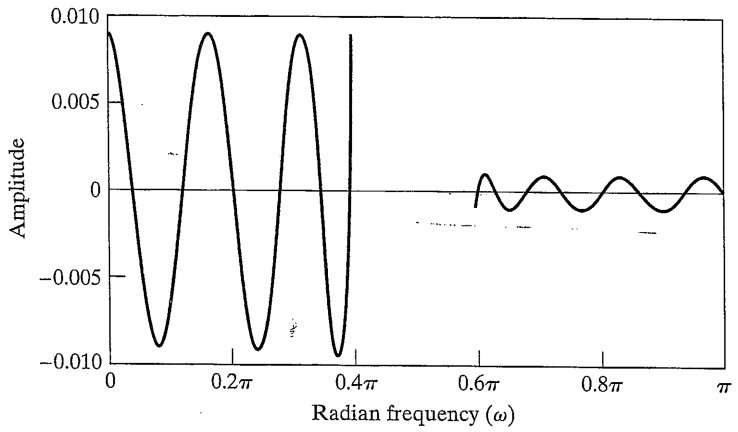


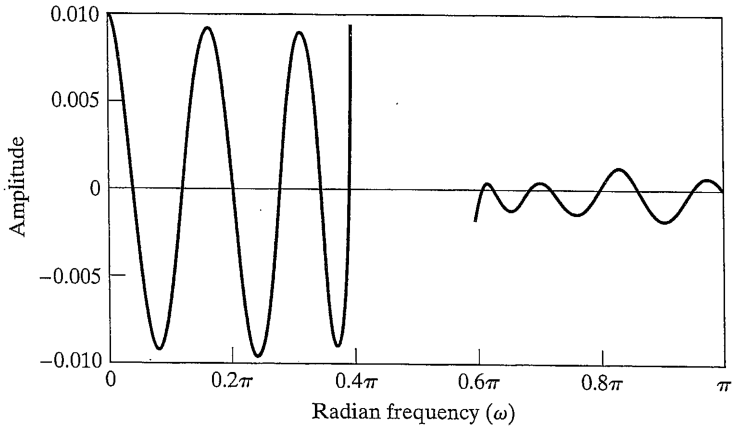
Figure 6.54 FIR quantization example.
(a) Log magnitude for unquantized case.

(b) Approximation error for unquantized case. (Error not defined in transition band.)

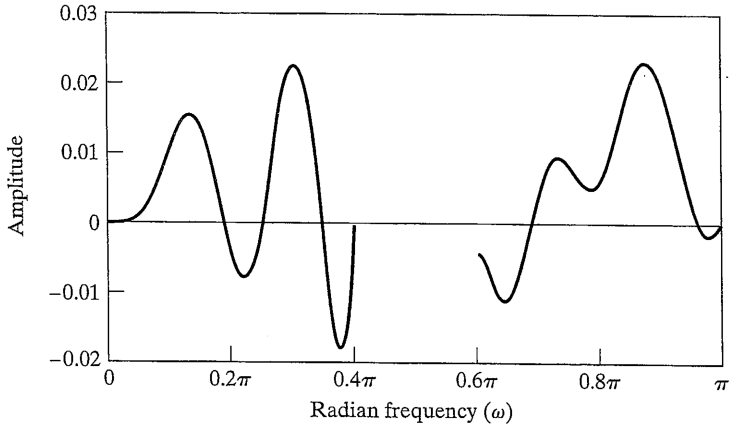
(c) Approximation error for 16-bit quantization.



(d)



(e)



(f)

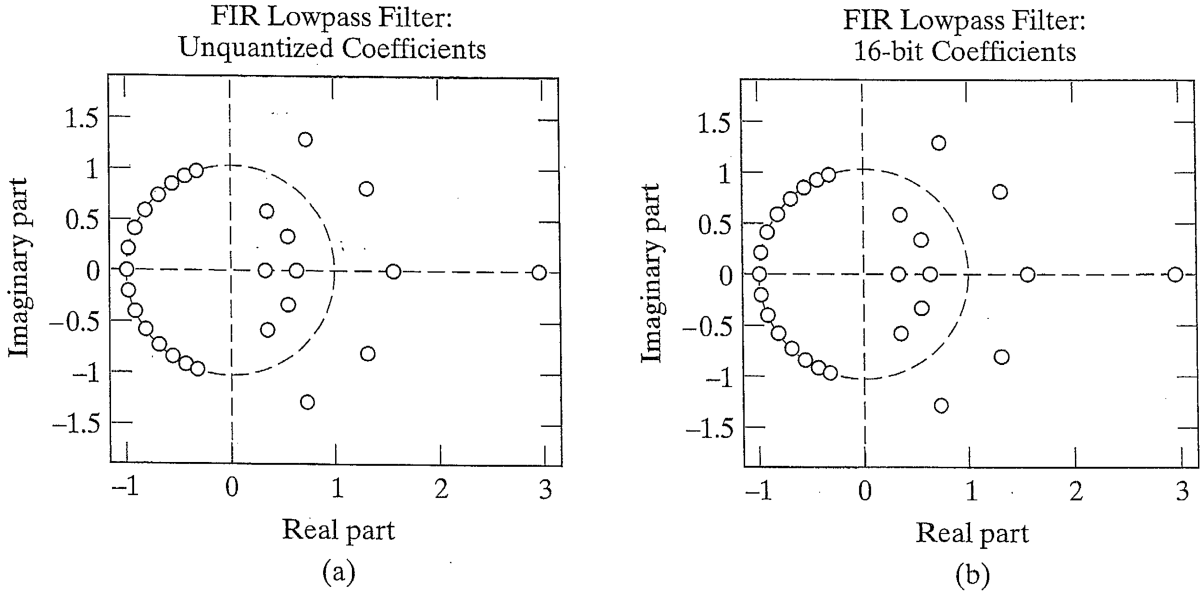
Figure 6.54 (continued)
(d) Approximation error for 14-bit quantization.

(e) Approximation error for 13-bit quantization.

(f) Approximation error for 8-bit quantization.

Note that the filter design specifications are met when 16 bit or 14 bits are used, and are almost satisfied for the case of 13 bits. However, for the 8 bit case, the approximation error exceeds the specification by a factor of approximately 2 in the passband and by a factor of approximately 20 in the stopband.

To relate the observations of this example with previous statements about the relationship of sensitivity to clustering of zeros, the zero locations for the 16-bit case, 13-bit case, and 8-bit case are shown below, along with the zero locations for the unquantized case.



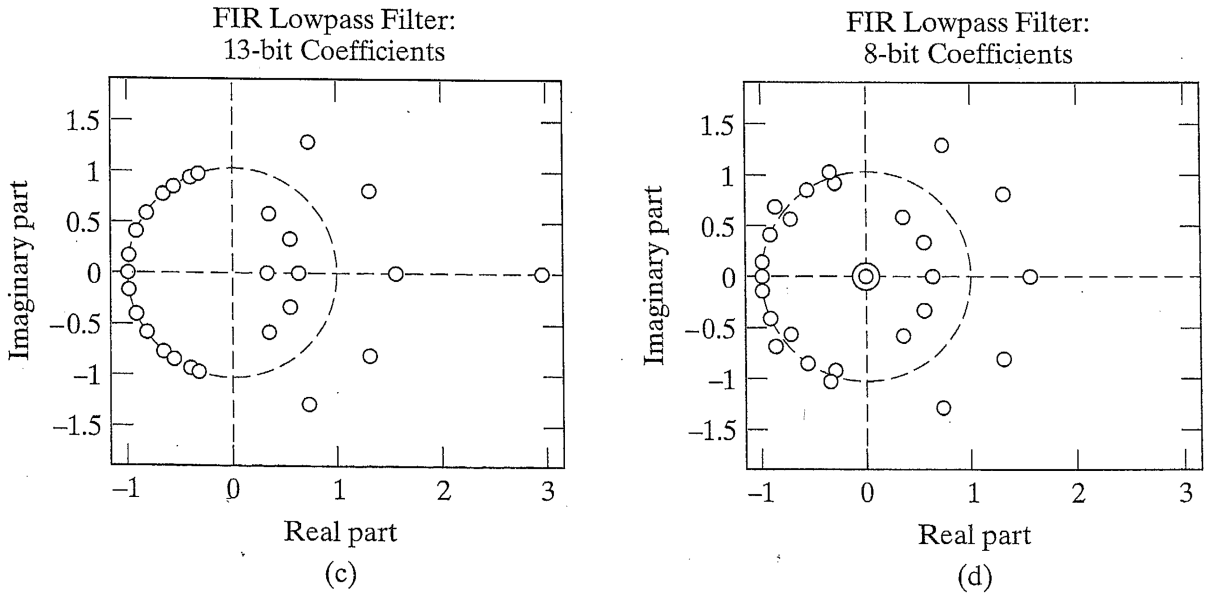


Figure 6. 55 Effect of impulse response quantization on zeros of H(z)
(a) Unquantized (b) 16-bit quantization
(c) 13-bit quantization
(d) 8-bit quantization

In the above figure, note the increased clustering of zeros on the unit circle for the 13 bit case, and especially for the 8-bit case.

Maintaining Generalized Linear Phase in FIR filters, Even with Quantization of Coefficients.

If an FIR filter is implemented directly (without factoring), then generalized linear phase is automatically maintained after quantization, since the symmetry conditions that guarantee generalized linear phase are not affected by quantization.

Recall that the symmetry conditions are $h(n) = h(M-n)$ for Types I and II filters and $h(n) = -h(M-n)$ for Types III and IV filters. Quantization affects $h(n)$ and $h(M-n)$ exactly the same, therefore preserving symmetric conditions.

If it is desired to implement an FIR filter in factored form, we can preserve generalized linear phase even though quantization is present, by recalling that zeros of a generalized phase filters occur in special patterns, each of which is considered below:

Group of 2 consisting of 2 complex-conjugate zeros on unit circle

- The corresponding second-order factor of H(z) has the form

$$(1 - z^{-1}e^{j\theta})(1 - z^{-1}e^{-j\theta}) = 1 - z^{-1}2\cos(\theta) + 1$$

Any quantization error in representing the value of $2\cos(\theta)$ changes only the angle, not the radius of the zeros. Thus, they will still be complex conjugates on the unit circle.

Group of 4 consisting of 2 complex-conjugate zeros not on unit circle and their reciprocals

- The corresponding fourth order factor of H(z) has the form

$$(1-z^{-1}re^{j\theta})(1-z^{-1}re^{-j\theta})(1-z^{-1}\frac{1}{r}e^{j\theta})(1-z^{-1}\frac{1}{r}e^{-j\theta})$$

$$=(1-z^{-1}2r\cos(\theta)+z^{-2}r^2)(1-z^{-1}\frac{2}{r}\cos(\theta)+\frac{z^{-2}}{r^2})$$

$$=(1-z^{-1}2r\cos(\theta)+z^{-2}r^2)\frac{1}{r^2}(r^2-z^{-1}2r\cos(\theta)+z^{-2})$$

Since both pairs of zeros have the same coefficients, $2r\cos(\theta)$ and r^2, any quantization error will affect both zero-pairs the same way, and the conjugate-reciprocal property of the 4 zeros will be preserved.

Group of two real zeros not on unit circle

$$(1-z^{-1}a)(1-z^{-1}\frac{1}{a})=1-z^{-1}(a+\frac{1}{a})+z^{-2}$$

If quantization error is present in the representation of the multiplier $(a+\frac{1}{a})$, the two zeros involved will still be real reciprocals, as shown below:

First, let $c=(a+\frac{1}{a})$.

The zeros of $1-cz^{-1}+z^{-2}$ are

$$\frac{c+\sqrt{c^2-4}}{2} \quad \text{and} \quad \frac{c-\sqrt{c^2-4}}{2}$$

Now show that the reciprocal of the first term above is in fact the second term:

$$\frac{2}{c+\sqrt{c^2-4}} = \frac{2(c-\sqrt{c^2-4})}{(c+\sqrt{c^2-4})(c-\sqrt{c^2-4})}$$

$$\frac{2(c-\sqrt{c^2-4})}{c^2-(c^2-4)} = \frac{c-\sqrt{c^2-4}}{2}$$

<u>zeros at 1 or -1</u>

Factors of the form $(1-z^{-1})$ and $(1+z^{-1})$ are not subject to quantization error.