Digital Signal Processing Lecture 10 - Finite-Precision Implementations (Quantization)

Electrical Engineering and Computer Science
University of Tennessee, Knoxville

September 27, 2013

Overview

Roadmap

- Introduction
- Discrete-time signals
- Discrete-time systems
  - LTI systems: the unit sample response $h[n]$ uniquely characterizes an LTI system
  - Linear constant-coefficient difference equation
  - Frequency response: $H(e^{j\omega})$
  - Fourier transform
- z transform
  - The z-transform, $X(z) = \sum_{n=-\infty}^{\infty} x[n]z^{-n}$
  - Region of convergence - the z-plane
  - System function, $H(z)$
  - Properties of the z-transform
  - The significance of zeros
  - The inverse z-transform, $x[n] = \frac{1}{2\pi j} \oint_C X(z)z^{n-1}dz$: inspection, power series, partial fraction expansion
- Relationships between the n, ω, and z domains

Review - Design structures

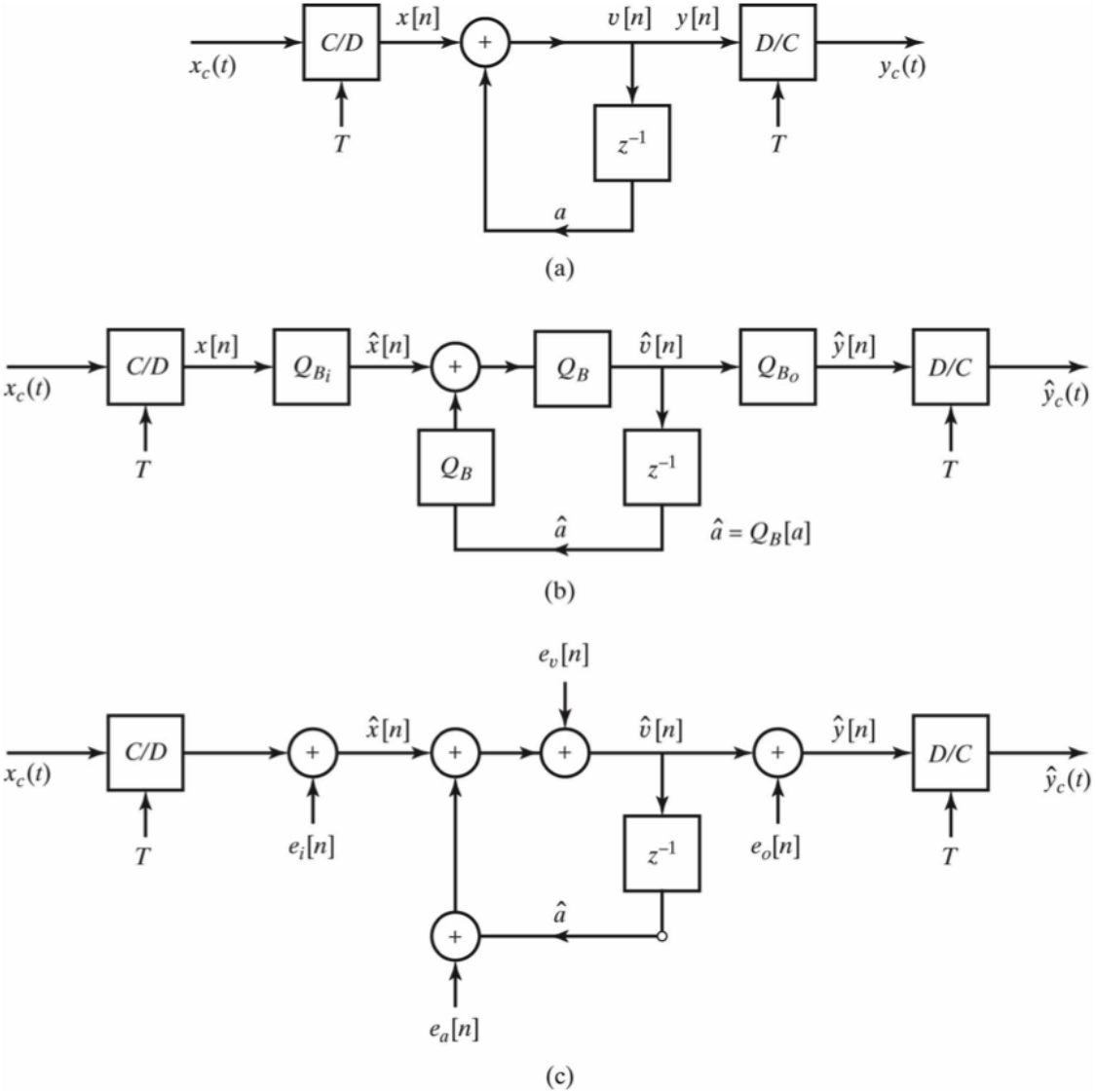
- Different representations of causal LTI systems
  - LCDE with initial rest condition
  - $H(z)$ with $|z| > R_+$
- Block diagram vs. Signal flow graph and how to determine system function (or unit sample response) from the graphs
- Design structures
  - Direct form I (zeros first)
  - Direct form II (poles first) - Canonic structure
  - Transposed form (zeros first)
- IIR: cascade form, parallel form, feedback in IIR (computable vs. noncomputable)
- FIR: direct form, cascade form, parallel form, linear phase FIR
- Metric (why study design structures?)
  - computational resource
  - precision

Sources of errors

$$y[n] = ay[n-1] + x[n]$$

- Coefficient quantization problem: $a \rightarrow \hat{a}$
- Input quantization error: $x[n] \rightarrow \hat{x}[n] = x[n] + e[n]$
- Product quantization error:
  $v[n] = ay[n-1] \rightarrow \hat{v}[n] = v[n] + e_a[n]$
- Limit cycles: caused by the nonlinearity by the quantization of arithmetic operations. When the input is absent or constant input or sinusoidal input signals are present, the output is in the form of oscillation

Quantization problem in Implementation



Number representations

- The two's complement format

$$x = X_m(-b_0 + \sum_{i=1}^{\infty} b_i 2^{-i})$$

  - X_m: an arbitrary scale factor
  - b_0: the sign bit. $0 \leq x \leq X_m$ if $b_0 = 0$; $-X_m \leq x < 0$ if $b_0 = 1$

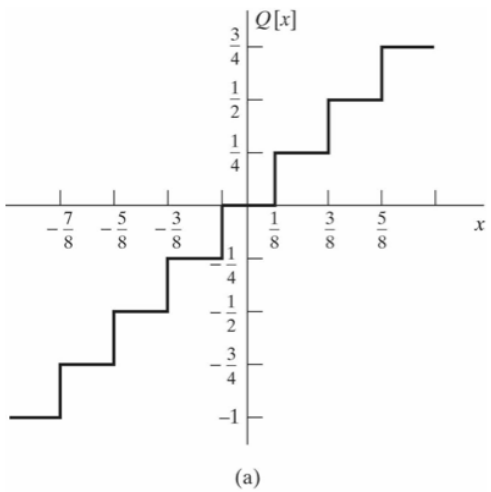
- Fix-point binary numbers

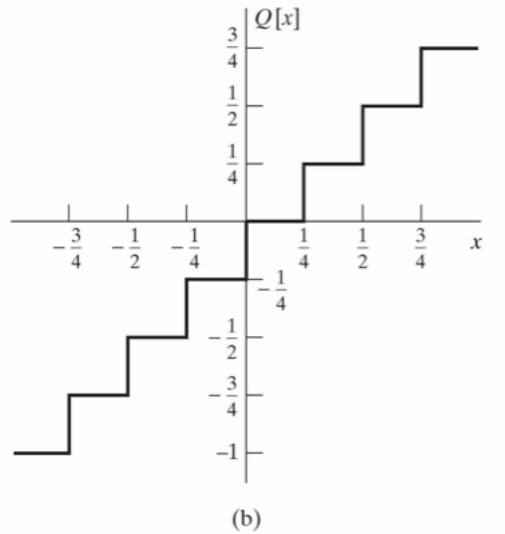
$$\hat{x} = Q_B[x] = X_m(-b_0 + \sum_{i=1}^{B} b_i 2^{-i}) = X_m \hat{x}_B$$

  - Quantizing a number to $B + 1$ bits. Quantization error: $e = Q_b[x] - x$

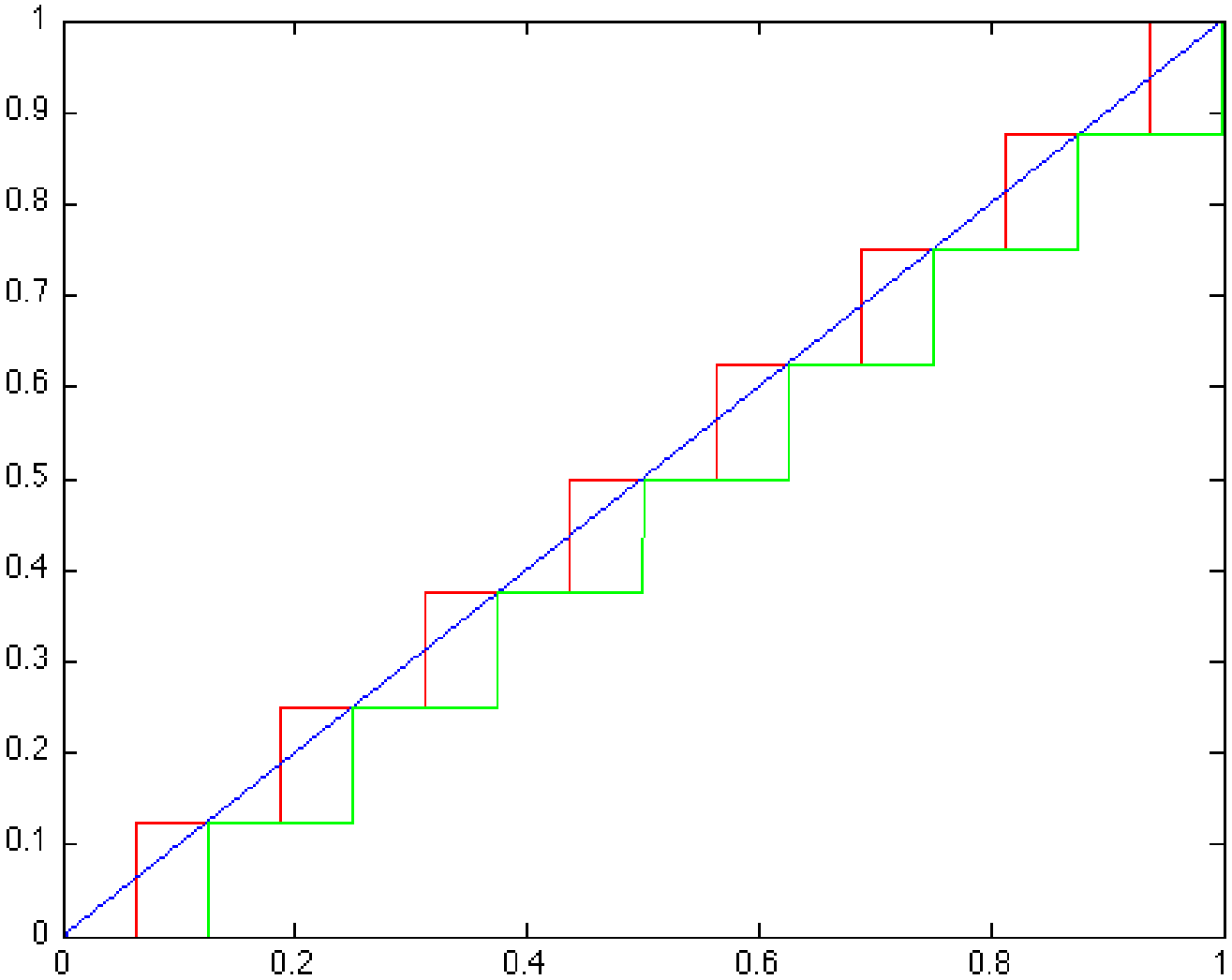
Quantization error

- Rounding: $-\Delta/2 < e \leq \Delta/2$
- Truncating: $-\Delta < e \leq 0$



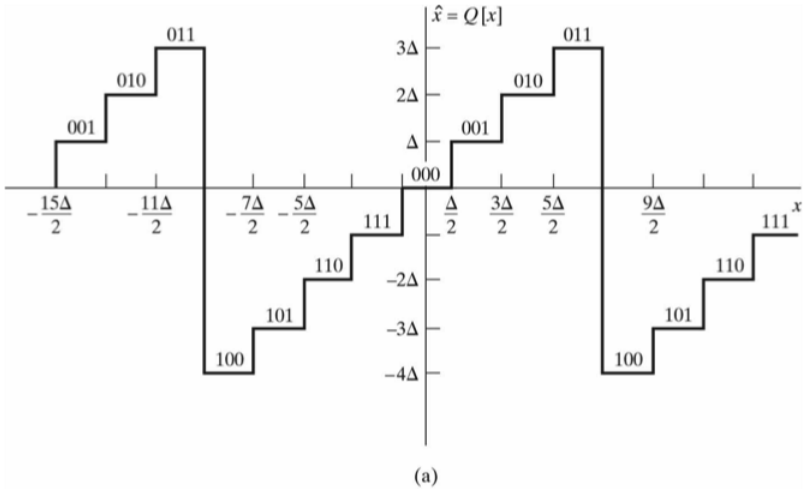


Quantization error (cont')

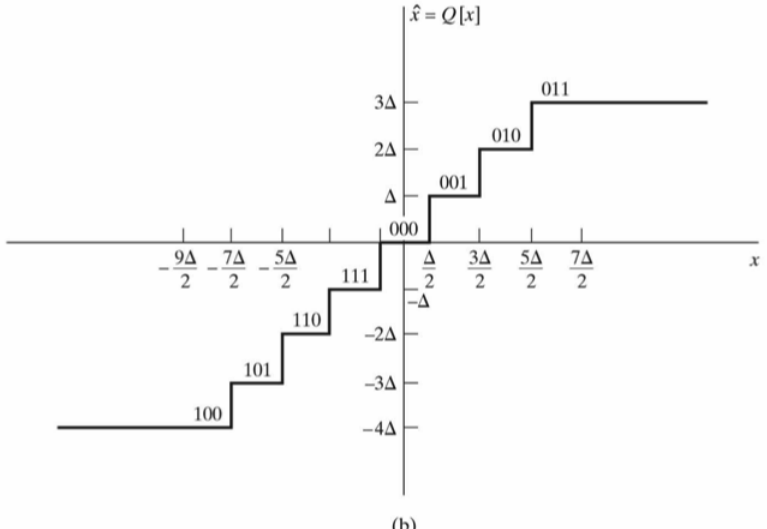


Overflow

- When $x > X_m$
- Saturation overflow (Clipping)



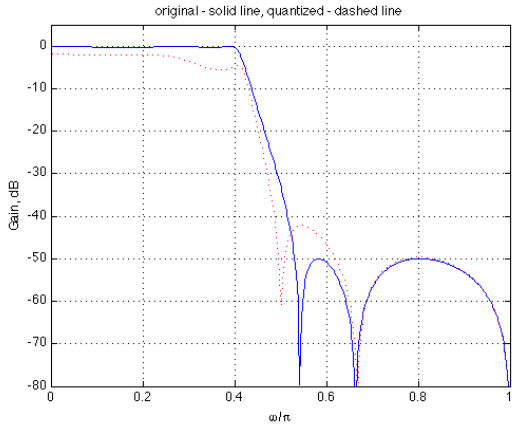
(a)

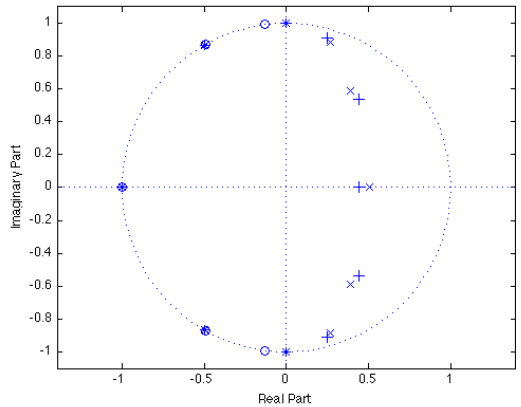


(b)

Coefficient quantization - IIR

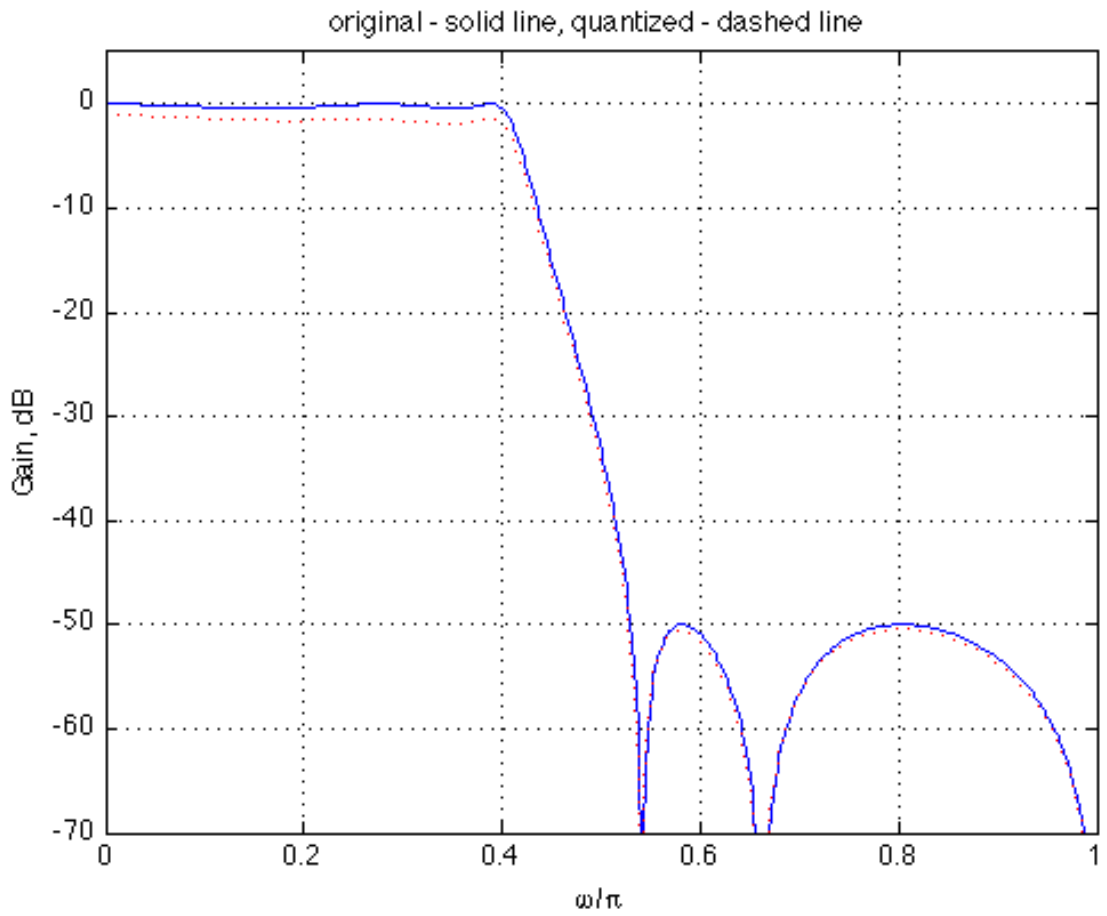
- Effect of coefficient quantization of an IIR digital filter implemented in direct form (5th-order IIR elliptic lowpass filter)





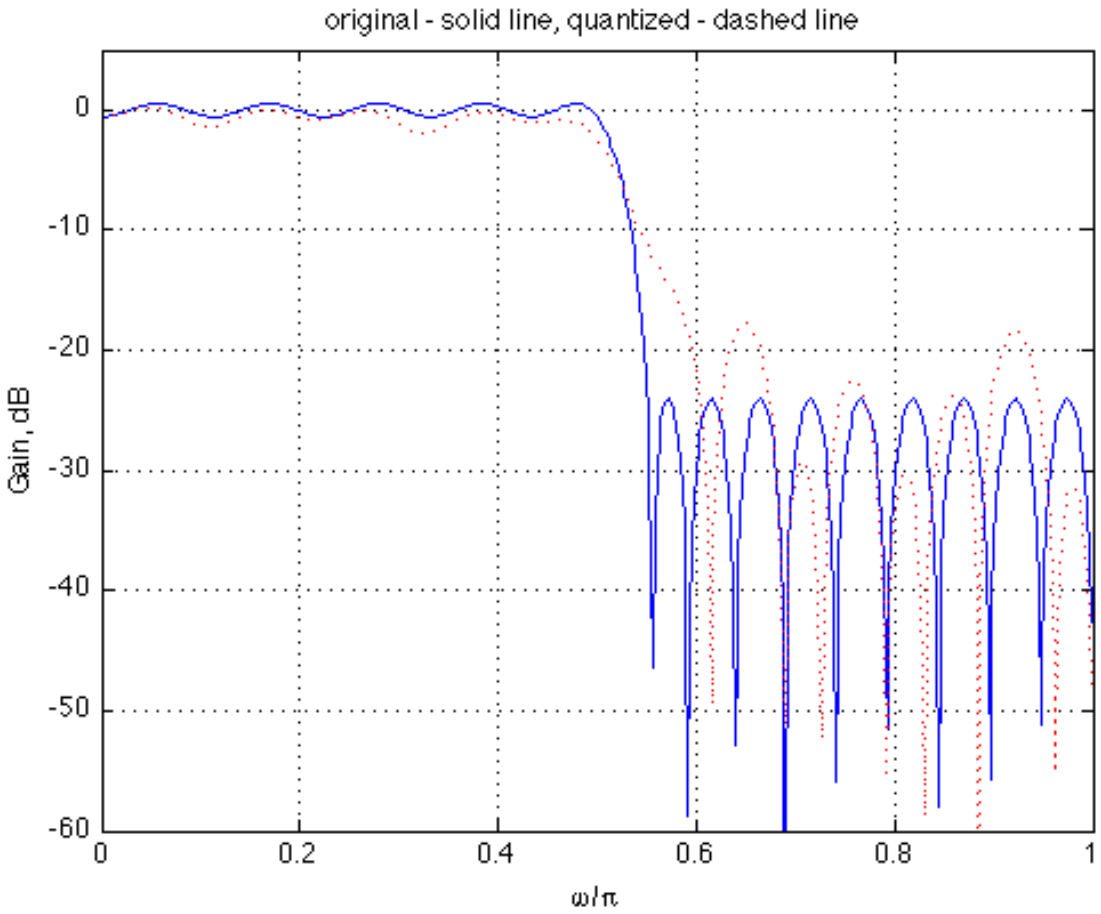
Coefficient quantization - IIR (cont')

- Effect of coefficient quantization of an IIR digital filter implemented in cascade form (5th-order IIR elliptic lowpass filter)



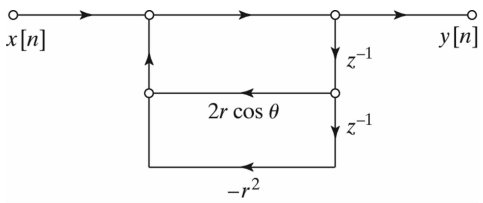
Coefficient quantization - FIR

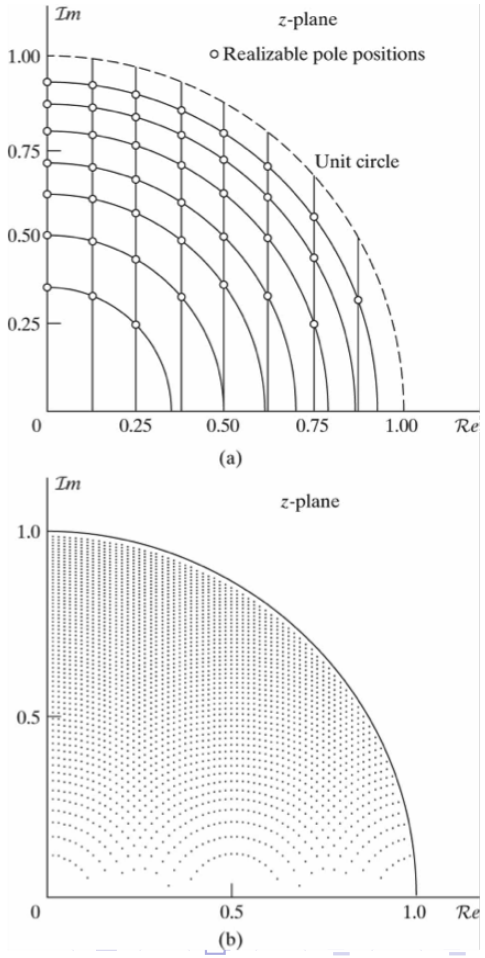
- Effect of coefficient quantization of an FIR digital filter implemented in direct form (39th-order FIR equiripple lowpass filter)



Pole sensitivity of second-order structures (Product quantization)

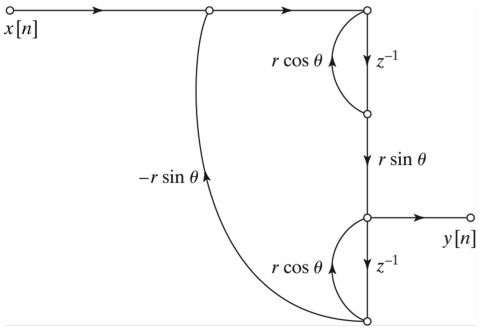
- The direct form structure exhibits high pole sensitivity with poles closer to the real axis and low pole sensitivity with poles closer to $z = \pm j$

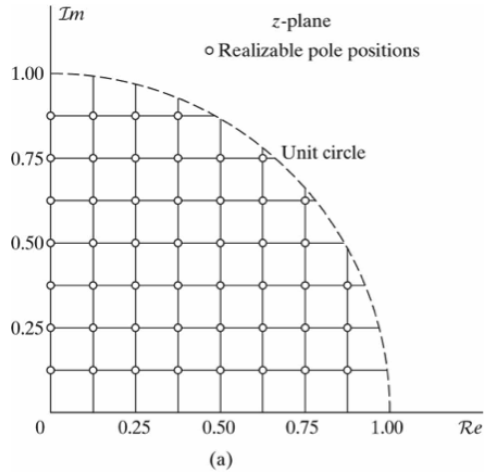




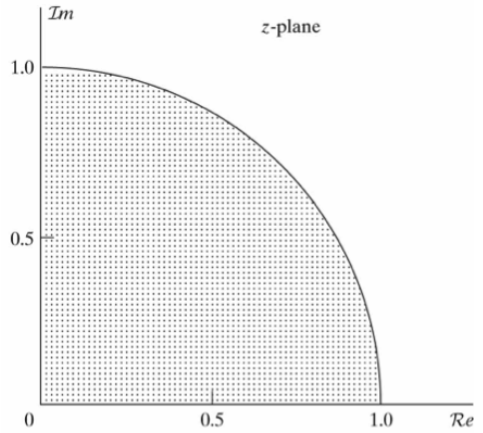
Pole sensitivity of second-order structures (cont')

- The coupled form structure is more suitable for implementing any type of second-order transfer function.





(a)



(b)