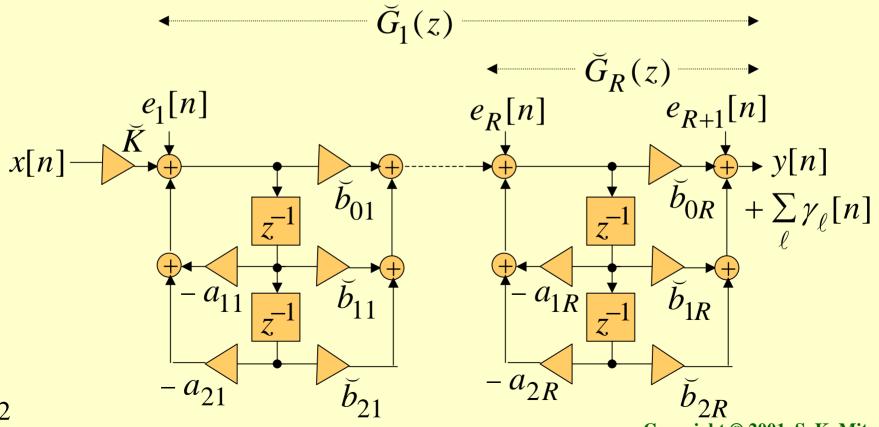
Optimum Ordering and Pole-Zero Pairing of the Cascade Form IIR Digital Filter

- There are many possible cascade realizations of a higher order IIR transfer function obtained by different pole-zero pairings and ordering
- Each one of these realizations will have different output noise power due to product round-offs
- It is of interest to determine the cascade realization with the lowest output noise power

 Consider the scaled cascade structure shown below



 The scaled noise transfer functions are given by

$$\breve{G}_{\ell}(z) = \prod_{i=\ell}^{R} \breve{H}_{i}(z) = \left(\prod_{i=\ell}^{R} \beta_{i}\right) G_{\ell}(z), \quad \ell = 1, 2, \dots, R$$

$$\widetilde{G}_{R+1}(z) = 1$$

where

$$\widetilde{H}_{\ell}(z) = \frac{\widetilde{b}_{0\ell} + \widetilde{b}_{1\ell}z^{-1} + \widetilde{b}_{2\ell}z^{-2}}{1 + a_{1\ell}z^{-1} + a_{2\ell}z^{-2}}$$

$$G_{\ell}(z) = \prod_{i=\ell}^{R} H_{i}(z)$$

 Output noise power spectrum due to product round-off is given by

$$P_{\gamma\gamma}(\omega) = \sigma_o^2 \begin{bmatrix} R+1 \\ \sum_{\ell=1}^{K+1} k_\ell |\breve{G}_\ell(e^{j\omega})|^2 \end{bmatrix}$$

• The output noise variance is thus

$$\sigma_{\gamma}^{2} = \sigma_{o}^{2} \left[\sum_{\ell=1}^{R+1} k_{\ell} \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \breve{G}_{\ell}(e^{j\omega}) \right|^{2} d\omega \right) \right]$$

$$=\sigma_o^2 \left[\sum_{\ell=1}^{R+1} k_\ell \| \breve{G}_\ell \|_2^2 \right]$$

- Note: k_{ℓ} is the total number of multipliers connected to the ℓ -th adder
- If products are rounded before summation

$$k_1 = k_{R+1} = 3$$

 $k_{\ell} = 5, \quad \ell = 2, 3, ..., R$

• If products are rounded after summation

$$k_{\ell} = 1, \quad \ell = 1, 2, ..., R + 1$$

Recall

$$\beta_0 = \frac{1}{\alpha_1}, \quad \beta_r = \frac{\alpha_r}{\alpha_{r+1}}, \quad r = 1, 2, ..., R$$

• Thus,

$$\prod_{i=\ell}^{R} \beta_i = \frac{\alpha_{\ell}}{\alpha_{R+1}} = \frac{\|F_{\ell}\|_p}{\|H\|_p}$$

Substituting the above in

$$P_{\gamma\gamma}(\omega) = \sigma_o^2 \begin{bmatrix} R+1 \\ \sum_{\ell=1}^{K+1} k_\ell |\breve{G}_\ell(e^{j\omega})|^2 \end{bmatrix}$$

we get

$$P_{\gamma\gamma}(\omega) = \frac{\sigma_o^2}{\|H\|_p^2} \left[k_{R+1} \|H\|_p^2 + \sum_{\ell=1}^R k_\ell \|F_\ell\|_p^2 |G_\ell(e^{j\omega})|^2 \right]$$
• The output noise variance is now given by

$$\sigma_{\gamma}^{2} = \frac{\sigma_{o}^{2}}{\|H\|_{p}^{2}} \left[k_{R+1} \|H\|_{p}^{2} + \sum_{\ell=1}^{R} k_{\ell} \|F_{\ell}\|_{p}^{2} \|G_{\ell}\|_{2}^{2} \right]$$

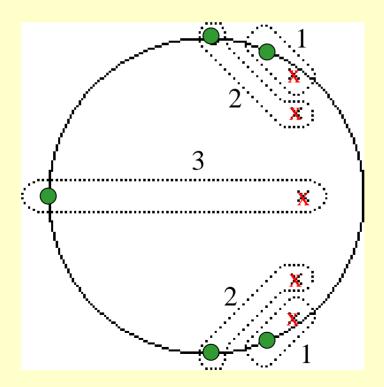
• The scaling transfer function $F_{\ell}(z)$ contains a product of section transfer functions, $H_i(z)$, $i=1,2,\ldots,\ell-1$ whereas, the noise transfer function $G_{\ell}(z)$ contains the product of section transfer functions $H_i(z)$, $i=\ell,\ell+1,\ldots,R$

- Thus every term in the expressions for $P_{\gamma\gamma}(\omega)$ and σ_{γ}^2 includes the transfer functions of all R sections in the cascade realization
- To minimize the output noise power, the norms of $H_i(z)$ should be minimized for all values of i by appropriately pairing the poles and zeros

- Pole-Zero Pairing Rule -
- First, the complex pole-pair closest to the unit circle should be paired with the nearest complex zero-pair
- Next, the complex pole-pair that is closest to the previous set of poles should be matched with its nearest complex zero-pair
- Continue this process until all poles and zeros have been paired

- The suggested pole-zero pairing is likely to lower the peak gain of the section characterized by the paired poles and zeros
- Lowering of the peak gain in turn reduces the possibility of overflow and attenuates the round-off noise

• Illustration of pole-zero pairing



- After the appropriate pole-zero pairings have been made, the sections need to be ordered to minimize the output round-off noise
- A section in the front part of the cascade has its transfer function $H_i(z)$ appear more frequently in the scaling transfer expressions for $P_{\gamma\gamma}(\omega)$ and σ_{γ}^2

- On the other hand, a section near the output end of the cascade has its transfer function H_i(z) appear more frequently in the noise transfer function expressions
- The best locations for $H_i(z)$ obviously depends on the type of norms being applied to the scaling and noise transfer functions

- A careful examination of the expressions for $P_{\gamma\gamma}(\omega)$ and σ_{γ}^2 reveals that if the \mathcal{L}_2 -scaling is used, then ordering of paired sections does not affect too much the output noise power since all norms in the expressions are \mathcal{L}_2 -norms
- This fact is evident from the results of the two examples presented earlier

- If \mathcal{L}_{∞} -scaling is being employed, the sections with the poles closest to the unit circle exhibit a peaking magnitude response and should be placed closer to the output end
- The ordering rule in this case is to place the least peaked section to the most peaked section starting at the input end

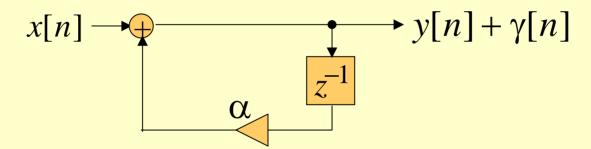
- The ordering is exactly opposite if the objective is to minimize the peak noise and an \mathcal{L}_2 -scaling is used
- Ordering has no effect on the peak noise if \mathcal{L}_{∞} -scaling is used
- The M-file zp2sos can be used to determine the optimum pole-zero pairing and ordering according the above discussed rule

- The output round-off noise variances of unscaled digital filters do not provide a realistic picture of the performances of these structures in practice
- This is due to the fact that introduction of scaling multipliers can increase the number of error sources and the gain for the noise transfer functions

- Therefore the digital filter structure should first be scaled before its round-off noise performance is analyzed
- In many applications, the round-off noise variance by itself is not sufficient, and a more meaningful result is obtained by computing instead the signal-to-noise ratio (SNR) for performance evaluation

- The computation of the SNR for the firstand second-order IIR structures are considered here
- Most conclusions derived from the detailed analysis of these simple structures are also valid for more complex structures
- Methods followed here can be easily extended to the general case

 Consider the causal unscaled first-order filter shown below



 Its output round-off noise variance was computed earlier and is given by

$$\sigma_{\gamma}^2 = \frac{\sigma_o^2}{1 - \alpha^2}$$

- Assume the input x[n] to be a WSS random signal with a uniform probability density function and a variance σ_x^2
- The variance σ_y^2 of the output signal y[n] generated by this input is then

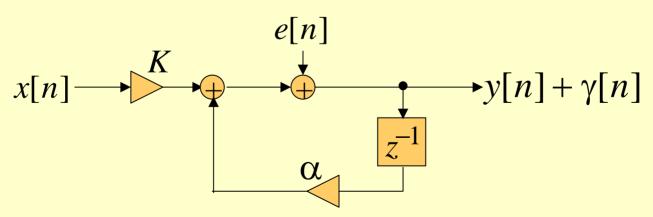
$$\sigma_y^2 = \sigma_x^2 \left(\sum_{n=0}^{\infty} h^2[n] \right) = \sigma_x^2 \left(\frac{1}{1 - \alpha^2} \right)$$

• The SNR of the unscaled filter is then

$$SNR = \frac{\sigma_y^2}{\sigma_\gamma^2} = \frac{\sigma_x^2}{\sigma_o^2}$$
• SNR is independent of α

- However, this is not a valid result since the adder is likely to overflow in an unscaled structure

• The scaled structure is shown below along with its round-off error analysis model assuming quantization after addition of all products



• With the scaling multiplier present, the output signal power becomes

$$\sigma_y^2 = \frac{K^2 \sigma_x^2}{1 - \alpha^2}$$

• The signal-to-noise ratio is then

$$SNR = \frac{K^2 \sigma_x^2}{\sigma_o^2}$$

- Since the scaling multiplier coefficient *K* depends on the pole location and the type of scaling being followed, the SNR will thus reflect this dependence
- The scaling transfer function is given by

$$F(z) = H(z) = \frac{1}{1 - \alpha z^{-1}}$$

with a corresponding impulse response

$$f[n] = Z^{-1}{F(z)} = \alpha^n \mu[n]$$

To guarantee no overflow, we choose

$$K = \frac{1}{\sum_{n=0}^{\infty} |f[n]|}$$

- To evaluate the SNR we need to know the type of input x[n] being applied
- If x[n] is uniformly distributed with $|x[n]| \le 1$ its variance is given by

$$\sigma_x^2 = \frac{1}{3}$$

• The *SNR* is then given by

$$SNR = \frac{(1-|\alpha|)^2}{3\sigma_o^2}$$

• For a (b+1)-bit signed fraction with round-off or two's-complement truncation

$$\sigma_o^2 = 2^{-2b} / 12$$

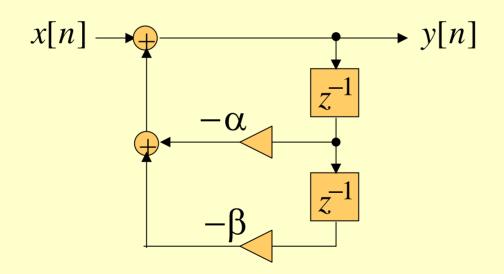
• The *SNR* in dB is given by

$$SNR_{dB} = 20\log_{10}(1-|\alpha|) + 6.02 + 6.02b$$

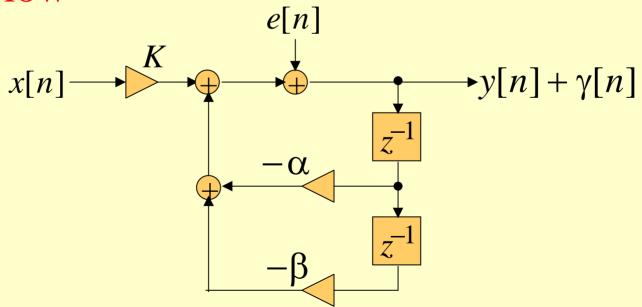
SNR of first-order IIR filters for different inputs with no overflow scaling

Input type	SNR	Typical SNR, dB $(b = 12, \alpha = 0.95)$
WSS, white uniform density	$\frac{(1- \alpha)^2}{3\sigma_o^2}$	52.24
WSS, white Gaussian density $(\sigma_x^2 = 1/9)$	$\frac{(1- \alpha)^2}{9\sigma_o^2}$	47.97
Sinusoid, known frequency	$\frac{(1- \alpha)^2}{2\sigma_o^2}$	69.91

 Consider the causal unscaled second-order IIR filter given below



• Its scaled version along with the round-off noise analysis model, assuming quantization after addition of all products, is shown below



• For a WSS input with a uniform probability density function and a variance σ_x^2 , the signal power at the output is given by

$$\sigma_y^2 = \sigma_x^2 \left(K^2 \sum_{n=0}^{\infty} h^2[n] \right)$$

• The round-off noise power at the output is given by

$$\sigma_{\gamma}^{2} = \sigma_{o}^{2} \begin{pmatrix} \sum_{n=0}^{\infty} h^{2}[n] \\ n=0 \end{pmatrix}$$

• Thus, the signal-to-noise ratio of the scaled structure is given by

$$SNR = \frac{\sigma_y^2}{\sigma_\gamma^2} = \frac{K^2 \sigma_x^2}{\sigma_\gamma^2}$$

• The scaling transfer function F(z) is given by

$$F(z) = H(z) = \frac{1}{1 + \alpha z^{-1} + \beta z^{-2}}$$

• If the poles of H(z) are at $z = re^{\pm j\theta}$, then

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

• Corresponding impulse response is given by

$$f[n] = h[n] = \frac{r^n \sin(n+1)\theta}{\sin \theta} \cdot \mu[n]$$

• The overflow is completely eliminated if

$$K = \frac{1}{\sum_{n=0}^{\infty} |h[n]|}$$

- $\sum_{n=0}^{\infty} |h[n]|$ is difficult to compute analytically
- It is possible to establish some bounds on the summation to provide a reasonable estimate of the value of *K*

 The magnitude response of the unscaled second-order section to a sinusoidal input is given by

$$|H(e^{j\theta})| = \left[\frac{1}{(1-r)^2(1-2r\cos\theta+r^2)}\right]^{1/2}$$

• But, $|H(e^{j\theta})|$ cannot be greater than $\sum_{n=0}^{\infty} |h[n]|$ as the latter is the largest possible value of the output y[n] for an input x[n] with $|x[n]| \le 1$

Signal-to-Noise Ratio in Second-Order IIR Filters

Moreover,

$$\sum_{n=0}^{\infty} |h[n]| = \frac{1}{\sin \theta} \sum_{n=0}^{\infty} r^n |\sin(n+1)\theta|$$

$$\leq \frac{1}{\sin \theta} \sum_{n=0}^{\infty} r^n = \frac{1}{(1-r)\sin \theta}$$

• A tighter upper bound on $\sum_{n=0}^{\infty} |h[n]|$ is given by

$$\sum_{n=0}^{\infty} |h[n]| \le \frac{4}{\pi (1-r^2) \sin \theta}$$

Signal-to-Noise Ratio in Second-Order IIR Filters

Therefore

$$(1-r)^2(1-2r\cos\theta+r^2) \ge K^2 \ge \frac{\pi^2(1-r^2)^2\sin^2\theta}{16}$$

• Bounds on the *SNR* for the all-pole secondorder section can also be derived for various types of inputs

Signal-to-Noise Ratio in Second-Order IIR Filters

- As in the case of the first-order section, as the poles move closer to the unit circle $r \rightarrow 1$, the gain of the filter increases
- The input signal then needs to be scaled down to avoid overflow while boosting the output round-off noise
- This type of interplay between the round-off noise and dynamic range is a characteristic of all fixed-point digital filters

Bounded Real Transfer Function

- Bounded Real Transfer Function -
- A causal stable real coefficient transfer function H(z) is defined to be a bounded real (BR) function if it satisfies the following condition:

$$|H(e^{j\omega})| \le 1$$
, for all values of ω

Bounded Real Transfer Function

• If the input and the output to the digital filter are x[n] and y[n], respectively with $X(e^{j\omega})$ and $Y(e^{j\omega})$ denoting their DTFTs, then the equivalent condition for the BR property is given by

 $\left|Y(e^{j\omega})\right|^2 \le \left|X(e^{j\omega})\right|^2$

• Integrating the above from $-\pi$ to π , dividing by 2π , and applying Parseval's relation we get $\sum_{n=0}^{\infty} y^2[n] \le \sum_{n=0}^{\infty} x^2[n]$

Bounded Real Transfer Function

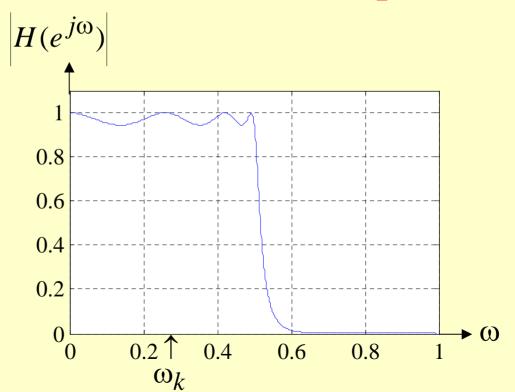
• Thus, a digital filter characterized by a BR transfer function is **passive** as, for all finite-energy inputs,

Output energy ≤ Input energy

- If $H(e^{j\omega}) = 1$ for all values of ω then it is a lossless system
- A causal stable transfer function H(z) with $|H(e^{j\omega})|=1$ is called a lossless bounded real (LBR) transfer function

- A causal stable digital filter with a transfer function H(z) has low coefficient sensitivity in the passband if it satisfies the following conditions:
 - (1) H(z) is a bounded-real transfer function
 - (2) There exists a set of frequencies ω_k at which $|H(e^{j\omega_k})|=1$
 - (3) The transfer function of the filter with quantized coefficients remains bounded real

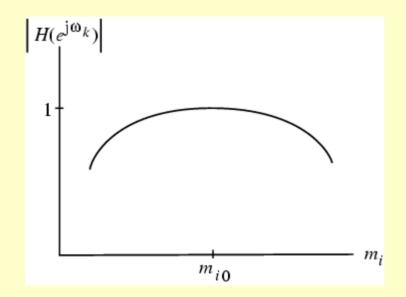
• Since $|H(e^{j\omega})|$ is bounded above by unity, the frequencies ω_k must be in the passband



- Any causal stable transfer function can be scaled to satisfy the first two conditions
- Let the digital filter structure \mathcal{N} realizing the BR transfer function H(z) be characterized by R multipliers with coefficients m_i
- Let the nominal values of these multiplier coefficients assuming infinite precision realization be m_{i0}

- Because of the third condition, regardless of the actual values of m_i in the immediate neighborhood of their design values m_{i0} , the actual transfer function remains BR
- Consider $|H(e^{j\omega_k})|$ which for multiplier values m_{i0} is equal to 1
- The third condition implies that if the coefficient m_i is quantized, then $|H(e^{j\omega_k})|$ can only become less than 1

• A plot of $|H(e^{j\omega_k})|$ will thus appear as indicated below



- Thus the plot of $|H(e^{j\omega_k})|$ will have a zerovalued slope at $m_i = m_{i0}$
- i.e.

$$\frac{\partial |H(e^{j\omega_k})|}{\partial m_i} \Big|_{m_i = m_{i0}} = 0$$

• The first-order sensitivity of $|H(e^{j\omega})|$ with respect to each multiplier coefficient is zero at all frequencies ω_k where $|H(e^{j\omega})|$ assumes its maximum value of unity

• Since all frequencies ω_k , where the magnitude function is exactly equal to unity, are in the passband of the filter and if these frequencies are closely spaced, it is expected that the sensitivity of the magnitude function to be very small at other frequencies in the passband

- A digital filter structure satisfying the conditions for low coefficient sensitivity is called a **structurally bounded** system
- Since the output energy of such a structure is also less than or equal to the input energy for all finite energy input signals, it is also called a **structurally passive** system

- If $|H(e^{j\omega})|=1$, the transfer function H(z) is called a **lossless bounded real** (LBR) function, i.e., a stable allpass function
- An allpass realization satisfying the LBR condition is called a structurally lossless or LBR system implying that the structure remains allpass under coefficient quantization

• Let G(z) be an N-th order causal BR IIR transfer function given by

$$G(z) = \frac{P(z)}{D(z)} = \frac{p_0 + p_1 z^{-1} + \dots + p_N z^{-N}}{1 + d_1 z^{-1} + \dots + d_N z^{-N}}$$

with a power-complementary transfer function H(z) given by

$$H(z) = \frac{Q(z)}{D(z)} = \frac{q_0 + q_1 z^{-1} + \dots + q_N z^{-N}}{1 + d_1 z^{-1} + \dots + d_N z^{-N}}$$

Power-complementary property implies that

$$|G(e^{j\omega})|^2 + |H(e^{j\omega})|^2 = 1$$

- Thus, H(z) is also a BR transfer function
- We determine the conditions under which G(z) and H(z) can be expressed in the form

$$G(z) = \frac{1}{2} \{ A_0(z) + A_1(z) \}$$

$$H(z) = \frac{1}{2} \{ A_0(z) - A_1(z) \}$$

where $A_0(z)$ and $A_1(z)$ are stable allpass functions

•
$$G(z) = \frac{1}{2} \{A_0(z) + A_1(z)\}$$

G(z) must have a symmetric numerator, i.e.,

$$p_n = p_{N-n}$$

- $H(z) = \frac{1}{2} \{A_0(z) A_1(z)\}$
- H(z) must have an anti-symmetric numerator, i.e.,

$$q_n = -q_{N-n}$$

• Symmetric property of the numerator of G(z) implies

$$P(z^{-1}) = z^N P(z)$$

• Likewise, antisymmetric property of the numerator of H(z) implies

$$Q(z^{-1}) = -z^N Q(z)$$

• By analytic continuation, the powercomplementary condition can be rewritten as

$$G(z^{-1})G(z) + H(z^{-1})H(z) = 1$$

• Substituting
$$G(z) = \frac{P(z)}{D(z)}, \qquad H(z) = \frac{Q(z)}{D(z)}$$

in the above we get

$$P(z)P(z^{-1}) + Q(z)Q(z^{-1}) = D(z^{-1})D(z)$$

• Using the relations $P(z^{-1}) = z^N P(z)$ and $Q(z^{-1}) = -z^N Q(z)$ in the previous equation we get

$$z^{-N}[P(z)P(z)-Q(z)Q(z)] = D(z^{-1})D(z)$$

• or, equivalently

$$P^{2}(z) - Q^{2}(z) = [P(z) + Q(z)][P(z) - Q(z)]$$
$$= z^{N} D(z)D(z^{-1})$$

- Using the relations $P(z) = z^{-N} P(z^{-1})$ and $Q(z) = -z^{-N} Q(z^{-1})$ we can write $P(z) Q(z) = z^{-N} [P(z^{-1}) + Q(z^{-1})]$
- Let $z = \xi_k$, $1 \le k \le N$, denote the zeros of [P(z) + Q(z)]
- Then $z = \frac{1}{\xi_k}$, $1 \le k \le N$, are the zeros of [P(z) Q(z)]

- [P(z)-Q(z)] is the mirror image of the polynomial [P(z)+Q(z)]
- From $P(z) Q(z) = z^{-N} [P(z^{-1}) + Q(z^{-1})]$ it follows that the zeros of [P(z) + Q(z)]inside the unit circle are also zeros of D(z), and the zeros of [P(z) + Q(z)] outside the unit circle are zeros of $D(z^{-1})$, since G(z)and H(z) are assumed to be stable functions

- Let $z = \xi_k$, $1 \le k \le r$, be the r zeros of [P(z) + Q(z)] inside the unit circle, and the remaining N r zeros, $z = \xi_k$, $r + 1 \le k \le N$, be outside the unit circle
- Then the N zeros of D(z) are given by

$$z = \begin{cases} \xi_k, & 1 \le k \le r \\ \frac{1}{\xi_k}, & r+1 \le k \le N \end{cases}$$

• To identify the above zeros of D(z) with the appropriate allpass transfer functions $A_0(z)$ and $A_1(z)$ we observe that these allpass functions can be expressed as

$$A_0(z) = G(z) + H(z) = \frac{P(z) + Q(z)}{D(z)}$$

$$A_1(z) = G(z) - H(z) = \frac{P(z) - Q(z)}{D(z)}$$

 Therefore, the two allpass transfer functions can be expressed as

$$A_{0}(z) = \prod_{k=r+1}^{N} \left(\frac{z^{-1} - (\xi_{k}^{*})^{-1}}{1 - \xi_{k}^{-1} z^{-1}} \right)$$

$$A_{1}(z) = \prod_{k=1}^{r} \left(\frac{z^{-1} - \xi_{k}^{*}}{1 - \xi_{k} z^{-1}} \right)$$

- In order to arrive at the above expressions, it is necessary to determine the transfer function H(z) that is power-complementary to G(z)
- Let U(z) denote the 2N-th degree polynomial

$$U(z) = P^{2}(z) - z^{-N}D(z^{-1})D(z) = \sum_{n=0}^{2N} u_{n}z^{-n}$$

Then

$$Q^{2}(z) = \sum_{n=0}^{2N} u_{n} z^{-n}$$

• Solving the above equation for the coefficients q_k of Q(z) we get

$$\begin{aligned} q_0 &= \sqrt{u_0} \\ q_1 &= \frac{u_1}{2q_0} \\ q_k &= q_{N-k} = \frac{u_k - \sum_{\ell=1}^{k-1} q_\ell q_{n-\ell}}{2q_0}, \ k \ge 2 \end{aligned}$$

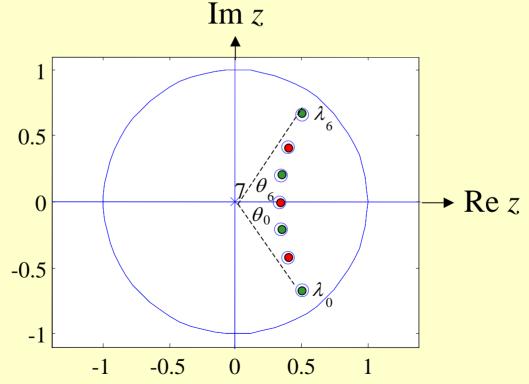
- After Q(z) has been determined, we form the polynomial [P(z) + Q(z)], find its zeros $z = \xi_k$, and then determine the two allpass functions $A_0(z)$ and $A_1(z)$
- It can be shown that IIR digital transfer functions derived from analog Butterworth, Chebyshev and elliptic filters via the bilinear transformation can be decomposed into the sum of allpass functions

- For lowpass-highpass filter pairs, the order N of the transfer function must be odd with the orders of $A_0(z)$ and $A_1(z)$ differing by 1
- For bandpass-bandstop filter pairs, the order N of the transfer function must be even with the orders of $A_0(z)$ and $A_1(z)$ differing by 2

- A simple approach to identify the poles of the two allpass functions for odd-order digital Butterworth, Chebyshev, and elliptic lowpass or highpass transfer functions is as follows
- Let $z = \lambda_k, 0 \le k \le N 1$ denote the poles of G(z) or H(z)

- Let θ_k denote the angle of the pole λ_k
- Assume that the poles are numbered such that $\theta_k < \theta_{k+1}$
- Then the poles of $A_0(z)$ are given by λ_{2k} and the poles of $A_1(z)$ are given by λ_{2k+1}

• The pole interlacing property is illustrated below



- Example Consider the parallel allpass realization of a 5-th order elliptic lowpass filter with the specifications: $\omega_p = 0.4\pi$, $\alpha_p = 0.5 \, \mathrm{dB}$, and $\alpha_s = 40 \, \mathrm{dB}$
- The transfer function obtained using MATLAB is given by

$$G(z) = \frac{0.528 + 0.0797z^{-1} + 0.1295z^{-2}}{(1 - 0.491z^{-1})(1 - 0.7624z^{-1} + 0.539z^{-2})}$$

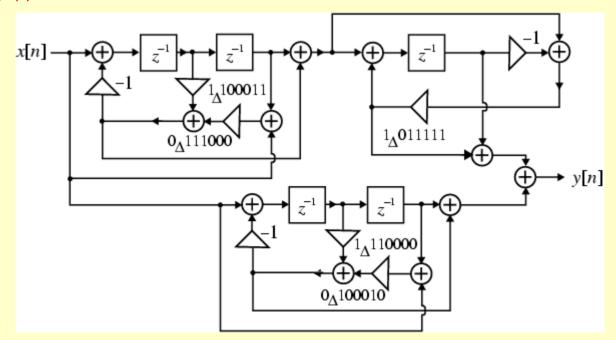
$$\times (1 - 0.557z^{-1} + 0.8828z^{-2})$$

 Its parallel allpass decomposition is given by

$$G(z) = \frac{1}{2} \left[\frac{(0.491 - z^{-1})(0.8828 - 0.557z^{-1} + z^{-2})}{(1 - 0.491z^{-1})(1 - 0.557z^{-1} + 0.8828z^{-2})} \right]$$

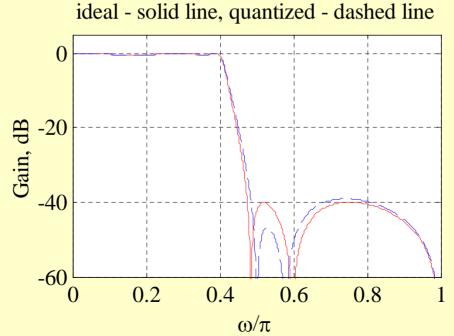
$$+\frac{0.539-0.762z^{-1}+z^{-2}}{1-0.762z^{-1}+0.539z^{-2}}]$$

• A 5-multiplier realization using a signed 7-bit fractional sign-magnitude representation of each multiplier coefficient is shown below



• The gain response of the filter with infinite precision multiplier coefficients and that with quantized coefficients are shown





 Passband details of the parallel allpass realization and a direct form realization

