Low Passband Sensitivity FIR Digital Filter

- We consider here the Type 1 filter as it is the most general linear-phase filter and can realize any type of frequency response
- The frequency response of a Type 1 FIR transfer function $H(z)$ of order N can be expressed as

$$H(e^{j\omega}) = e^{-j\omega N/2}\breve{H}(\omega)$$

where $\breve{H}(\omega)$, a real function of ω, is its amplitude response

Low Passband Sensitivity FIR Digital Filter

- If $H(z)$ is a BR function, then $\breve{H}(\omega) \leq 1$
- Its delay-complementary transfer function $G(z)$ defined by

$$G(z) = z^{-N/2} - H(z)$$

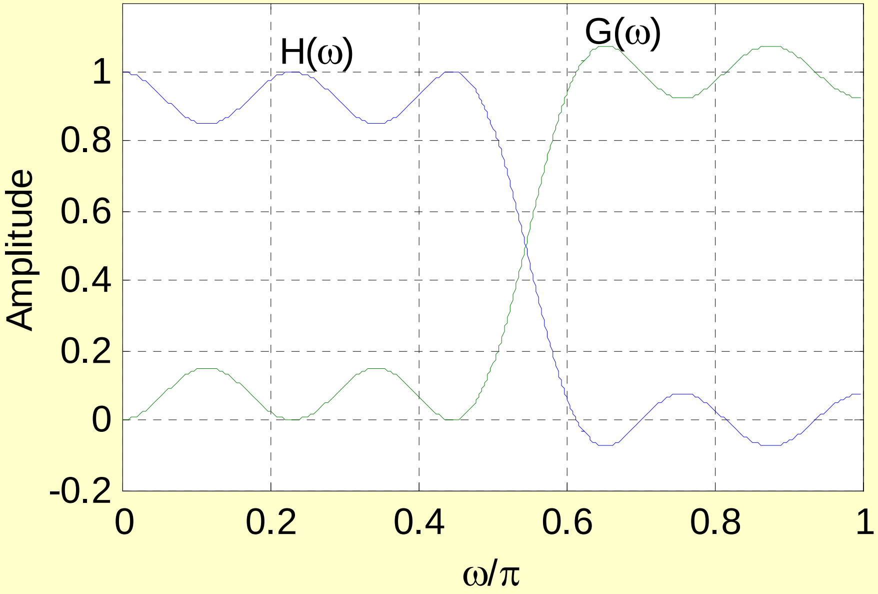
has a frequency response given by

$$G(e^{j\omega}) = e^{-j\omega N/2}[1 - \breve{H}(\omega)] = e^{-j\omega N/2}\breve{G}(\omega)$$

where $\breve{G}(\omega) = 1 - \breve{H}(\omega)$ is its amplitude response

Low Passband Sensitivity FIR Digital Filter

- Amplitude responses of a typical delay-complementary FIR filter pair are shown below



Low Passband Sensitivity FIR Digital Filter

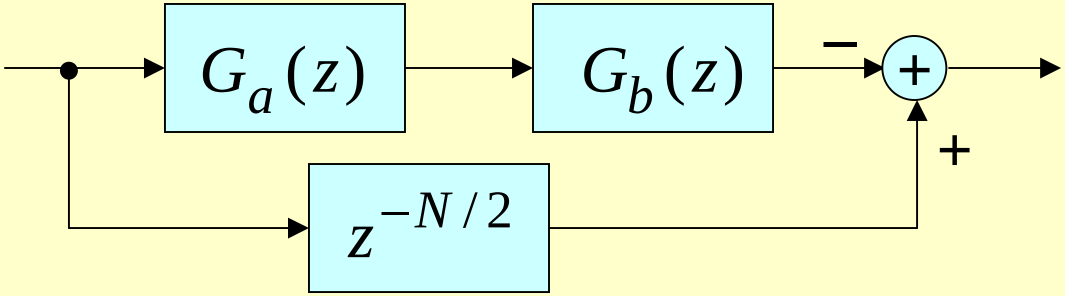
- It follows from the plots of the amplitude responses that at $\omega = \omega_k$, where $|H(e^{j\omega_k})| = 1$ $\breve{G}(\omega)$ has double zeros
- Thus, $G(z)$ can be expressed as

$$G(z) = G_a(z)\prod_{k=1}^{L}(1 - 2\cos\omega_k z^{-1} + z^{-2})^2$$

$$= G_a(z)G_b(z)$$

Low Passband Sensitivity FIR Digital Filter

- A delay-complementary realization of $H(z)$ based on $H(z) = z^{-N/2} - G(z)$ is shown below



- $G_b(z)$ consists of L 4-th order FIR sections with the k-th section having a transfer function $(1 - 2\cos\omega_k z^{-1} + z^{-2})^2$

Low Passband Sensitivity FIR Digital Filter

- If the multiplier coefficient $2\cos\omega_k$ of the k-th section is quantized, its zeros are still double and remain on the unit circle
- Thus, quantization of the coefficients of $G_b(z)$ does not change the sign of the amplitude response $\breve{G}(\omega)$, and in the passband of $H(z)$, $\breve{G}(\omega) \geq 0$

Low Passband Sensitivity FIR Digital Filter

- In addition, $G_a(z)$ has no zeros on the unit circle, and quantization of its coefficients also does not affect the sign of $\breve{G}(\omega)$
- Hence, $\breve{H}(\omega)$ continues to remain bounded above by unity
- The realization of $H(z)$ as indicated remains structurally BR or structurally passive with regard to all coefficients, resulting in a low passband sensitivity realization

Low Passband Sensitivity FIR Digital Filter

- Example - The filter specifications are length 13 with a normalized passband edge at 0.5 and a normalized stopband edge at 0.6 with equal weights to passband and stopband ripples
- Using the M-file `remez` we determine the transfer function of the lowpass filter $H(z)$ and form its delay-complementary filter

$$G(z) = z^{-6} - H(z)$$

Low Passband Sensitivity FIR Digital Filter

- $G(z)$ has 6 zeros on the unit circle: 2 zeros at $z = 1$, a pair of complex conjugate zeros at $z = -0.26463064626566 \pm j0.9643498437$ and a pair of complex conjugate zeros at
$$z = -0.27683551142484 \pm j0.96091732945$$
- These unit circle zeros constitute
$$G_b(z) = (1 - z^{-1})^2 (1 - 0.52926129 z^{-1} + z^{-2})$$
$$\times (1 - 0.5536710228497 z^{-1} + z^{-2})$$

Low Passband Sensitivity FIR Digital Filter

- By factoring out $G_b(z)$ from $G(z)$ we get

$$G_a(z) = 0.04107997 + 0.051971544z^{-1}$$
$$-0.12094731168z^{-2} - 0.30704562224z^{-3}$$
$$+0.120947311687z^{-4} - 0.0.051971544z^{-5}$$
$$+0.04107997195619z^{-6}$$

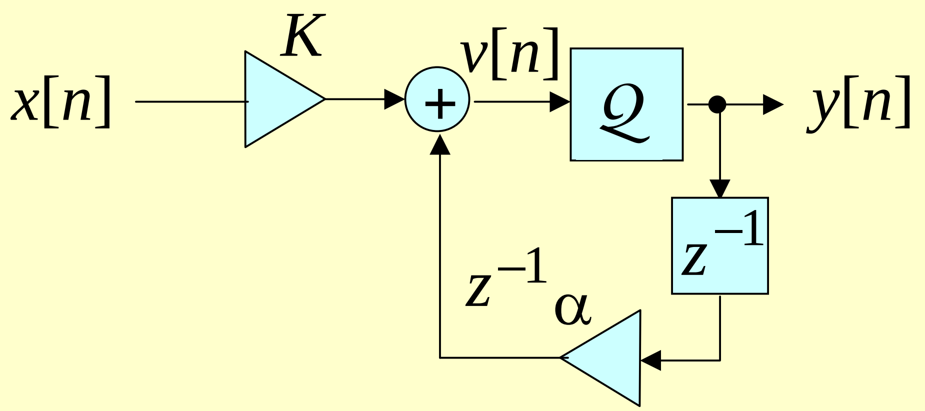
- Next we quantize the coefficients of $G_a(z)$ and $G_b(z)$ by rounding the fractional part to 2 decimal digits

Low Passband Sensitivity FIR Digital Filter

- Finally, from $G(z)$ with quantized coefficients, the delay-complementary transfer function $H(z)$ is determined

First-Order Error-Feedback Structure

- Consider the scaled first-order section

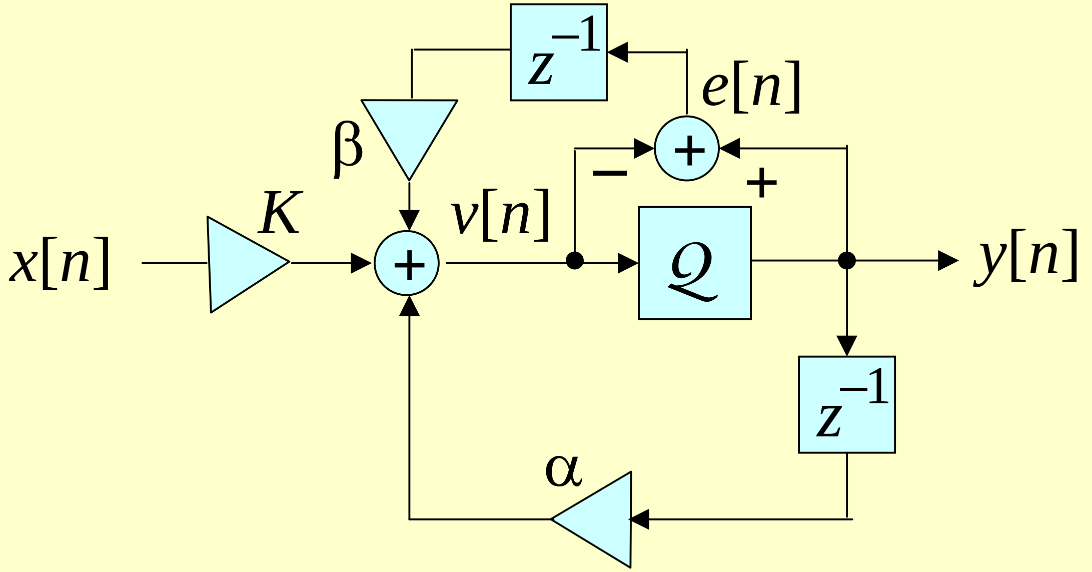


- We assume that all multiplier coefficients are signed $(b + 1)$-bit fractions
- The quantization error signal is given by

$$e[n] = y[n] - v[n]$$

First-Order Error-Feedback Structure

- The first-order section is modified by feeding back the error signal $e[n]$ to the system through a delay and a multiplier β as shown below



First-Order Error-Feedback Structure

- In practice, β is chosen to be a simple integer or a power-of-2 fraction, such as ± 1, ± 2, or ± 0.5 so that the multiplication can be performed using a shift operation and will not introduce an additional quantization error

First-Order Error-Feedback Structure

- Analyzing the error-feedback structure we arrive at its transfer function

$$H(z) = \left.\frac{Y(z)}{X(z)}\right|_{E(z)=0} = \frac{K}{1-\alpha z^{-1}}$$

- The noise transfer function $G(z)$ with the error feedback, with $y[n]$ as the output is given by

$$G(z) = \left.\frac{Y(z)}{E(z)}\right|_{X(z)=0} = \frac{1+\beta z^{-1}}{1-\alpha z^{-1}}$$

First-Order Error-Feedback Structure

- The noise transfer function without the error feedback ($\beta = 0$) is given by

$$G_0(z) = \frac{1}{1 - \alpha z^{-1}}$$

- The output noise variance of the error-feedback structure is given by

$$\sigma_\gamma^2 = \left(\frac{1 + 2\alpha\beta + \beta^2}{1 - \alpha^2} \right) \sigma_o^2$$

where σ_o^2 is the variance of $e[n]$

First-Order Error-Feedback Structure

- σ_γ^2 is a minimum when $\beta = -\alpha$
- However, in practice $|\alpha| < 1$
- Hence $\beta = -\alpha$ will introduce an additional quantization noise source, making the analysis resulting in the expression for σ_γ^2 invalid
- Thus, β should be chosen as an integer with a value close to that of $-\alpha$

First-Order Error-Feedback Structure

- For $|\alpha| < 0.5$, $\beta = 0$, implying no error feedback
- However, in this case, the pole of $H(z)$ is far from the unit circle, and as a result, the output noise variance σ_γ^2 is not that high
- For $|\alpha| \geq 0.5$, choose $\beta = (-1)\,\mathrm{sgn}(\alpha)$
- Using this value of β we get

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

First-Order Error-Feedback Structure

- The output noise variance with $\beta = 0$ is

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

- Thus, error feedback has increased the SNR by a factor of

$$-10\log_{10}[2(1-|\alpha|)]\,\text{dB}$$

- This increase in SNR is quite significant if the pole is closer to the unit circle

First-Order Error-Feedback Structure

- For example if $|\alpha| = 0.99$, the improvement is about 17 dB, which is equivalent to about 3 bits of increased accuracy compared to the case without error feedback
- Additional hardware requirements for the error-feedback structure are two new adders and an additional storage register

First-Order Error-Feedback Structure

- The noise transfer function for the error-feedback structure can be expressed as

$$G(z) = (1 + \beta z^{-1}) G_0(z)$$

where $G_0(z)$ is the noise transfer function without error feedback

- The error-feedback circuit is **shaping** the error spectrum by modifying the input quantization noise $E(z)$ to

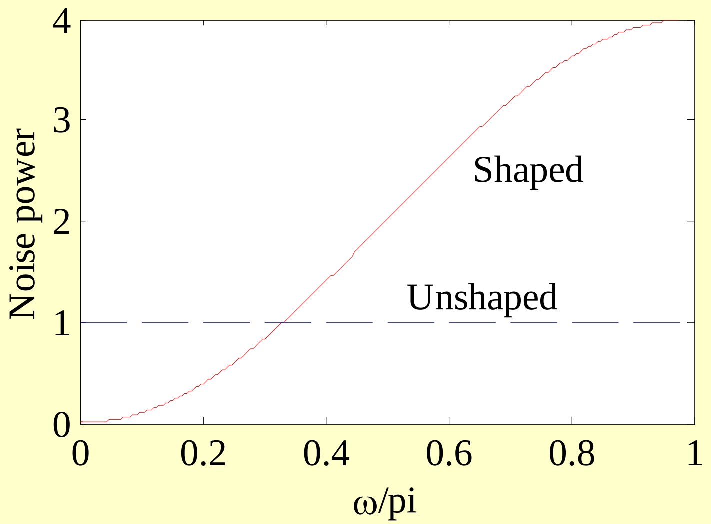
$$E_s(z) = (1 + \beta z^{-1}) E(z)$$

First-Order Error-Feedback Structure

- The output noise is generated by passing $E_s(z)$ through the usual noise transfer function $G_0(z)$
- To illustrate the effect of noise spectrum shaping, consider the case of a narrow-band lowpass first-order filter with $\alpha \to 1$
- We choose $\beta = -1$ and as a result $E_s(z)$ has a zero at $z = 1$ $(\omega = 0)$

First-Order Error-Feedback Structure

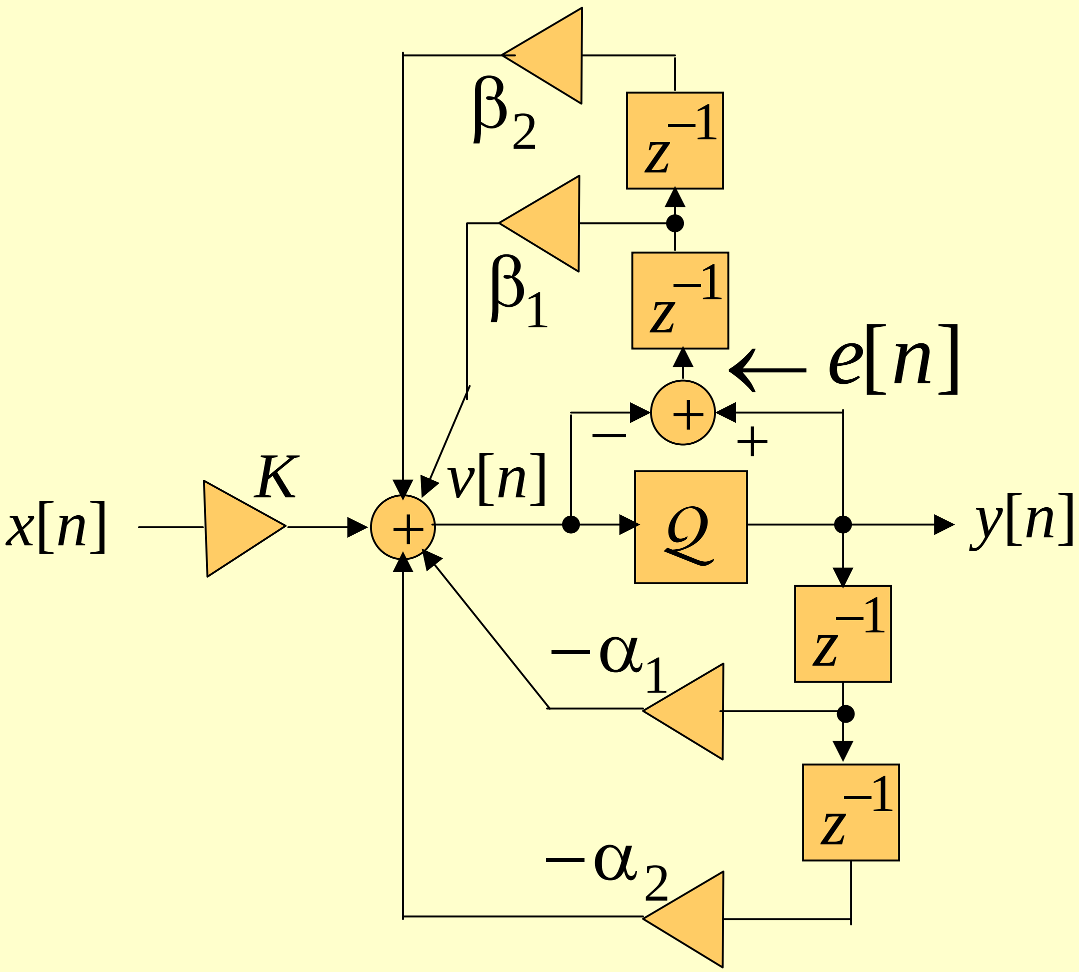
- The power spectral density of the unshaped quantization noise $E(z)$ is σ_o^2, a constant
- The power spectral density of the shaped quantization noise $E_s(z)$ is $4\sin^2(\omega/2)\sigma_o^2$



First-Order Error-Feedback Structure

- The noise shaping redistributes the noise so as to move it mostly into the stopband of the lowpass filter, thus reducing the noise variance
- Because of the noise redistribution caused by the error-feedback, this approach has also been called the **error spectrum shaping method**

Second-Order Error-Feedback Structure



Second-Order Error-Feedback Structure

- The noise transfer function is given by

$$G(z) = \frac{1+\beta_1 z^{-1} + \beta_2 z^{-2}}{1+\alpha_1 z^{-1} + \alpha_2 z^{-2}}$$

- The output round-off noise variance for $\mathcal{L}_2$-scaling is given by

$$\sigma_\gamma^2 = (\|G\|_2)^2 \sigma_o^2$$

- A choice of $\beta_1 = \alpha_1$ and $\beta_2 = \alpha_2$ makes $\|G\|_2 = 1$, yielding $\sigma_\gamma^2 = \sigma_o^2$, an apparent optimal solution

Second-Order Error-Feedback Structure

- However, this choice for the multiplier coefficients in the error-feedback path introduces additional quantization noise sources that invalidates the expression for σ_γ^2
- A more attractive solution is to make β_1 and β_2 integers with values close to α_1 and α_2, respectively

Second-Order Error-Feedback Structure

- For example, for a narrow-band lowpass transfer function, the poles are close to the unit circle and to the real axis, i.e., $r \approx 1$ and $\theta \approx 0$
- Then, α_1 is close to -2 and α_2 is close to 1
- In this case, choose $\beta_1 = -2$ and $\beta_2 = 1$
- Then

$$G(z) = \frac{1 - 2z^{-1} + z^{-2}}{1 + \alpha_1 z^{-1} + \alpha_2 z^{-2}}$$

Second-Order Error-Feedback Structure

- For a very narrowband lowpass filter with $r = 0.995$, $\theta = 0.07\pi$, and $b = 16$, the second-order error-feedback structure has an SNR that is approximately 25 dB higher than that without the error feedback
- The second-order error-feedback structure also provides a noise shaping

Second-Order Error-Feedback Structure

- The error-feedback circuit shapes the error spectrum by modifying the input quantization noise $E(z)$ to

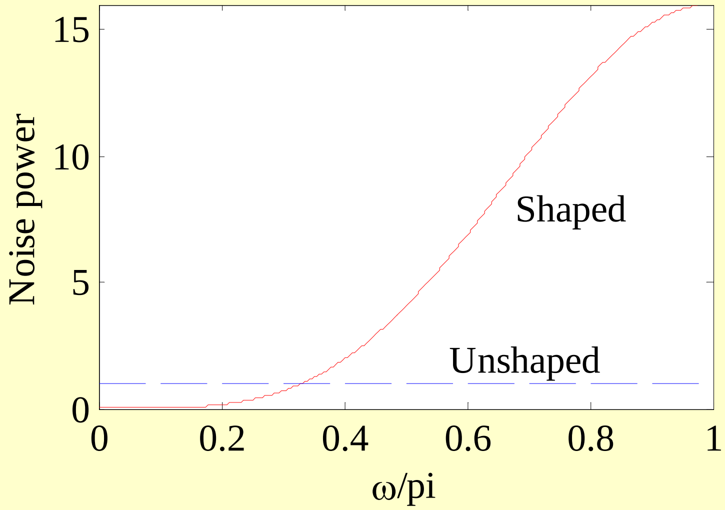
$$E_s(z) = (1 - z^{-1})^2 E(z)$$

- The output noise is generated by passing $E_s(z)$ through the usual noise transfer function

$$G_0(z) = \frac{1}{1 + \alpha_1 z^{-1} + \alpha_2 z^{-2}}$$

Second-Order Error-Feedback Structure

- The power spectral density of the shaped noise source $E_s(z)$ is $16\sin^4(\omega/2)\sigma_o^2$
- The power spectral density of the unshaped noise source is σ_o^2



Limit Cycles in IIR Digital Filters

- So far we have treated the analysis of finite wordlength effects using a linear model of the system
- A practical digital filter is a nonlinear system caused by the quantization of the arithmetic operations
- Such nonlinearities may cause an IIR filter, which is stable under infinite precision, to exhibit an unstable behavior under finite precision arithmetic for specific input signals

Limit Cycles in IIR Digital Filters

- This type of instability usually results in an oscillatory periodic output called a **limit cycle**
- The system remains in this condition until an input of sufficiently large amplitude is applied to move the system into a more conventional operation

Limit Cycles in IIR Digital Filters

- Limit cycles occur in IIR filters due to the presence of feedback
- Such oscillations are absent in FIR filters as they do not have any feedback path
- There are two types of limit cycles

  (1) **Granular limit cycle** is usually of low amplitude

  (2) **Overflow limit cycle** has large amplitudes

Limit Cycles in IIR Digital Filters

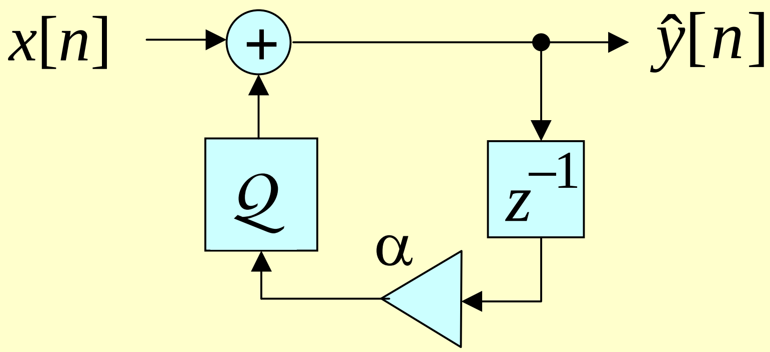
- Two types of granular limit cycles have been observed in IIR digital filters:

  (1) **Inaccessible limit cycle** - can appear only if the initial conditions of the digital filter at the time of starting pertain to that limit cycle

  (2) **Accessible limit cycle** - can appear by starting the digital filter with initial conditions not pertaining to the limit cycle

Granular Limit Cycles

- Consider the first-order IIR filter as shown below

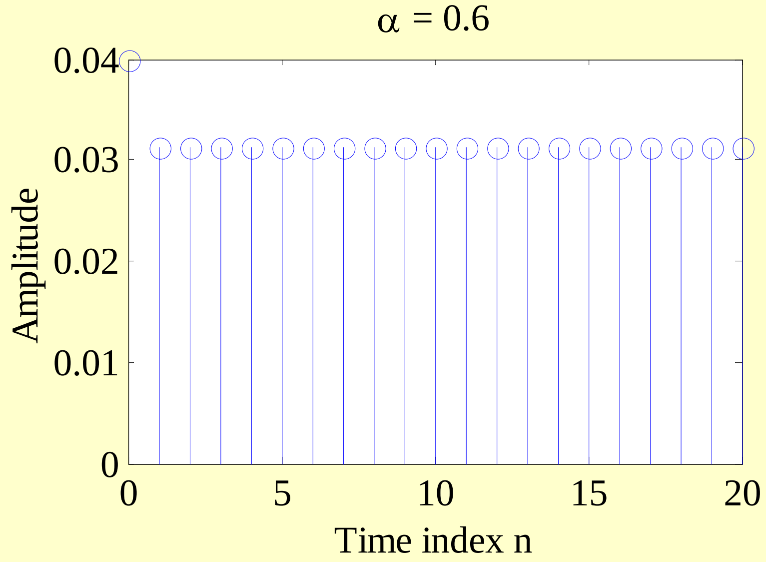


- Assume the quantization operation to be rounding and the filter to be implemented with a signed 6-bit fractional arithmetic
- The nonlinear difference equation characterizing the filter is given by

$$\hat{y}[n] = \mathrm{Q}(\alpha \cdot \hat{y}[n-1]) + x[n]$$

Granular Limit Cycles

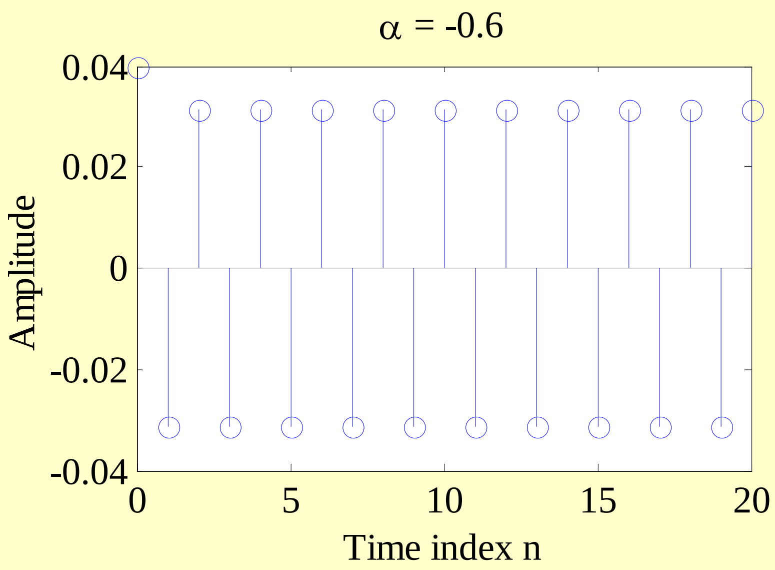
- For $x[n] = 0.04\delta[\mathrm{n}]$, $\hat{y}[-1] = 0$, and $\alpha = 0.6$, the output of the filter is as shown below



- The limit cycle generated has a period of 1

Granular Limit Cycles

- For $x[n] = 0.04\delta[\mathrm{n}]$, $\hat{y}[-1] = 0$, and $\alpha = -0.6$ the output of the filter is as shown below



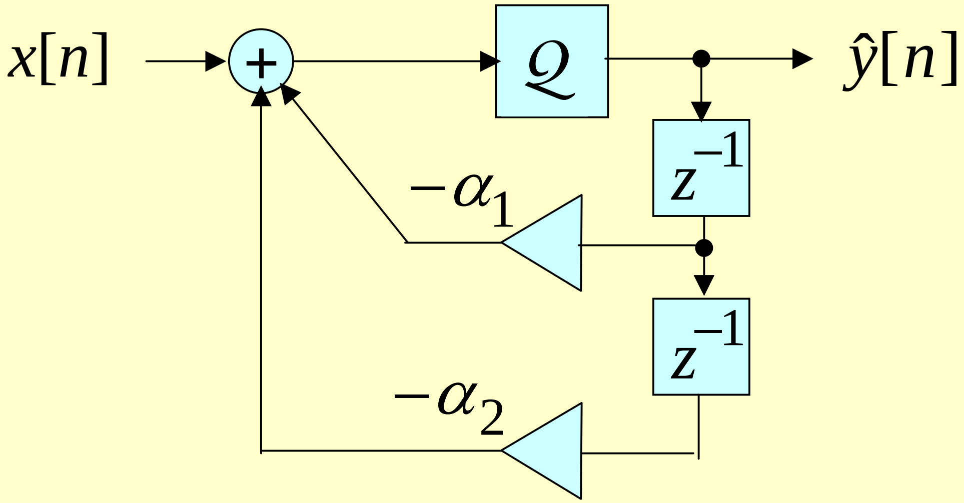
- The limit cycle generated has a period of 2

Overflow Limit Cycles

- Limit-cycle-like oscillations can also result from overflow in digital filters implemented with finite precision arithmetic
- The amplitude of the overflow oscillations can cover the whole dynamic range of the register experiencing the overflow
- Overflow limit cycles are thus much more serious in nature than the granular limit cycles

Overflow Limit Cycles

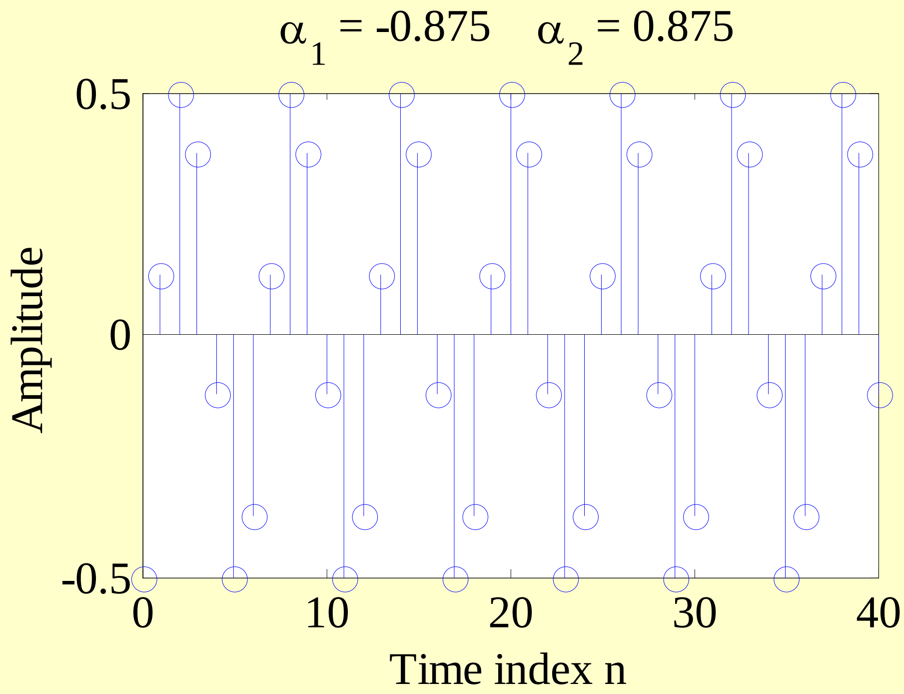
- Consider the causal all-pole second-order IIR digital filter shown below



- Assume implementation using sign-magnitude 4-bit arithmetic with a rounding of the sum of products by a single quantizer

Overflow Limit Cycles

- Let $\alpha_1 = -0.875$, $\alpha_2 = 0.875$, $\hat{y}[-1] = -0.625$ and $\hat{y}[-2] = -0.125$
- Consider $x[n] = 0$ for $n \geq 0$

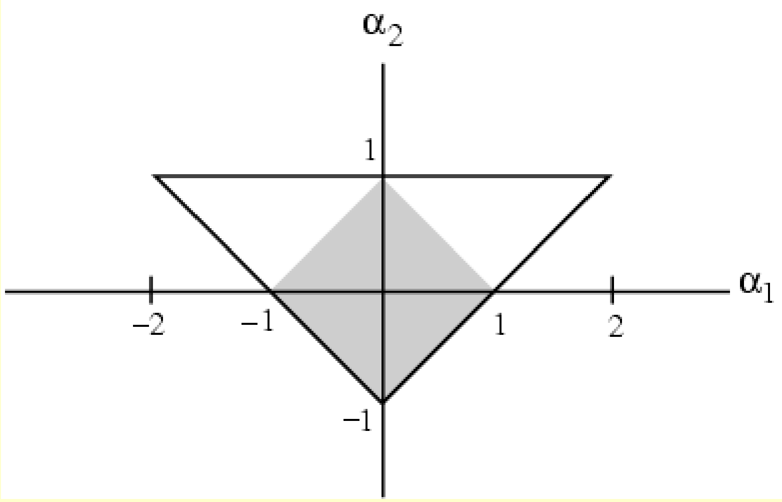


Overflow Limit Cycles

- The second-order direct form IIR structure with multiplier coefficients α_1 and α_2 remains stable if $|\alpha_2| < 1$ and $|\alpha_1| < 1 + \alpha_2$
- However, the structure can still get into a zero-input overflow oscillation mode for a large range of values of the filter constants satisfying the stability constraint when implemented using two's-complement arithmetic with rounding

Overflow Limit Cycles

- It has been shown that overflow limit cycles under zero-input cannot occur if the filter coefficients lie in the shaded region inside the stability triangle shown below



Limit Cycle Free Structures

- Conditions for a digital filter structure to not support limit cycles have been derived in terms of its state transition matrix
- For a second-order causal LTI digital filter, the state-space representation relating the output $y[n]$ to the input $x[n]$ is given by

$$\begin{bmatrix} s_1[n+1] \\ s_2[n+1] \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} s_1[n] \\ s_2[n] \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} x[n]$$

$$y[n] = [c_1 \quad c_2] \begin{bmatrix} s_1[n] \\ s_2[n] \end{bmatrix} + d\, x[n]$$

Limit Cycle Free Structures

- Let $\mathbf{s}[n] = [s_1[n] \quad s_2[n]]^T$

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \mathbf{B} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \mathbf{C} = [c_1 \quad c_2]$$

- The state-space description is then compactly written as

$$\mathbf{s}[n+1] = \mathbf{A}\,\mathbf{s}[n] + \mathbf{B}\,x[n]$$
$$y[n] = \mathbf{C}\,\mathbf{s}[n] + d\,x[n]$$

- **A** is called the **state-transition matrix**
- $\mathbf{s}[n]$ is called the **state-vector**

Limit Cycle Free Structures

- The quantization errors caused by the quantization of the state-transition equation

$$\mathbf{s}[n+1] = \mathbf{A}\,\mathbf{s}[n] + \mathbf{B}\,x[n]$$

go through the feedback loop and are responsible for the generation of limit cycles

- Assume $s_1[n+1]$ and $s_2[n+1]$ are quantized
- Delayed versions of these quantized signals are $s_1[n]$ and $s_2[n]$

Limit Cycle Free Structures

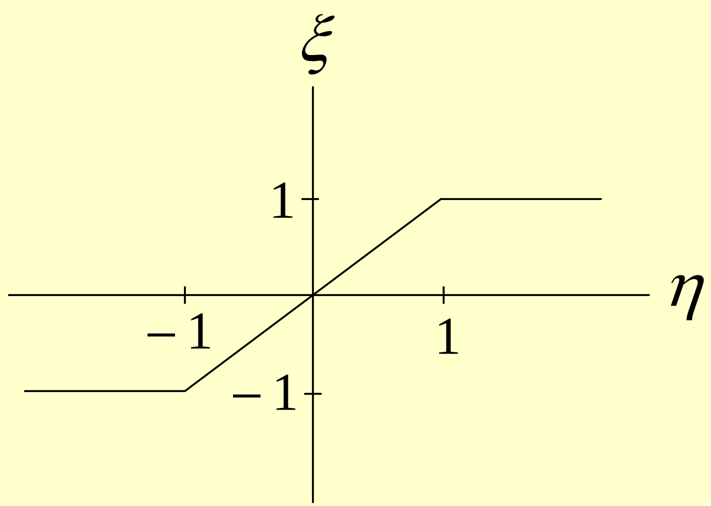
- A quantizer is defined to be **passive** if

$$|\mathcal{Q}(x)| \le |x|, \quad \text{for all } x$$

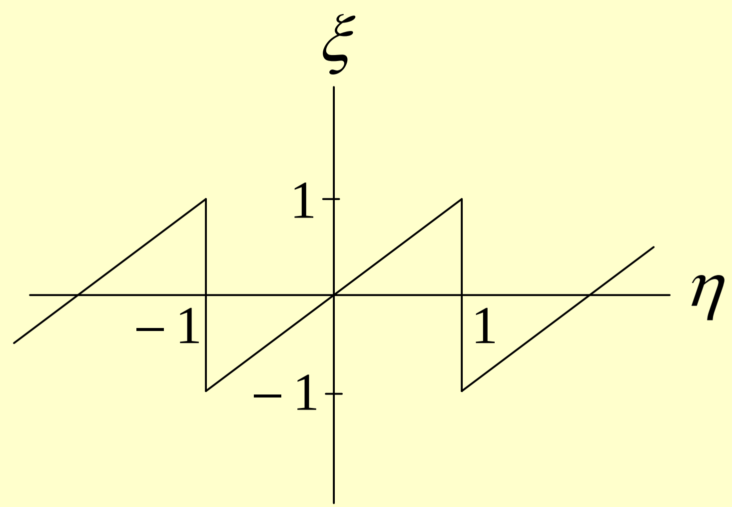
- If x is inside the dynamic range of the system, then for magnitude truncation above inequality holds
- If x is outside the dynamic range, for example by overflow, it must be brought back to the range by following the schemes discussed next

Handling Overflow

- If η, the sum of two fixed-point fractions, exceeds the dynamic range $[-1, 1)$, it is substituted with a number ξ which is within the range using one of the two following schemes



Saturation overflow



Two's-complement overflow

Limit Cycle Free Structures

- Thus, magnitude truncation followed by one of the two overflow handling schemes is again a passive quantizer
- A digital filter structure with a state transition matrix satisfying

$$\mathbf{A}^T\mathbf{A} = \mathbf{A}\mathbf{A}^T$$

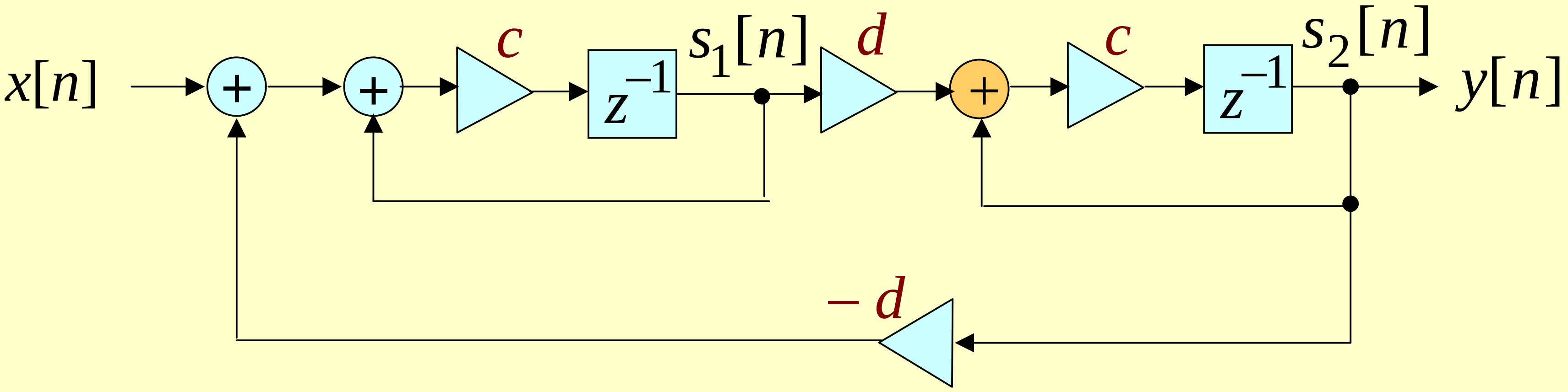
- has been called a **normal form structure**

Limit Cycle Free Structures

- A normal form structure with passive quantizers does not support zero-input limit cycles of either type
- The state transition matrix $\mathbf{A}$ satisfying the condition $\mathbf{A}^T\mathbf{A} = \mathbf{A}\mathbf{A}^T$ and $\|\mathbf{A}\|_2 < 1$ is called a **normal matrix**

Limit Cycle Free Structures

- Example - Consider the digital filter structure shown below



- Analysis yields

$$s_1[n+1] = c\,s_1[n] - cd\,s_2[n] + c\,x[n]$$
$$s_2[n+1] = cd\,s_1[n] + c\,s_2[n]$$

Limit Cycle Free Structures

- The state transition matrix is given by

$$\mathbf{A} = \begin{bmatrix} c & -cd \\ cd & c \end{bmatrix}$$

- The transfer function of the structure is

$$H(z) = \frac{c^2 d z^{-2}}{1 - 2cz^{-1} + c^2(1 + d^2) z^{-2}}$$

Limit Cycle Free Structures

- Comparing the denominator of $H(z)$ with that of a second-order IIR transfer function with poles at $z = re^{\pm j\theta}$ (with $r < 1$ for stability) we obtain $c = r\cos\theta$ and $d = \tan\theta$
- Thus

$$\mathrm{A} = \begin{bmatrix} r\cos\theta & -r\sin\theta \\ r\sin\theta & r\cos\theta \end{bmatrix}$$

- Note: $\mathbf{A}^T\mathbf{A} = \mathbf{A}\mathbf{A}^T = r^2\mathbf{I}$ and $\|\mathbf{A}\|_2 = r < 1$
- $\Longrightarrow$ The filter is a **normal form** structure