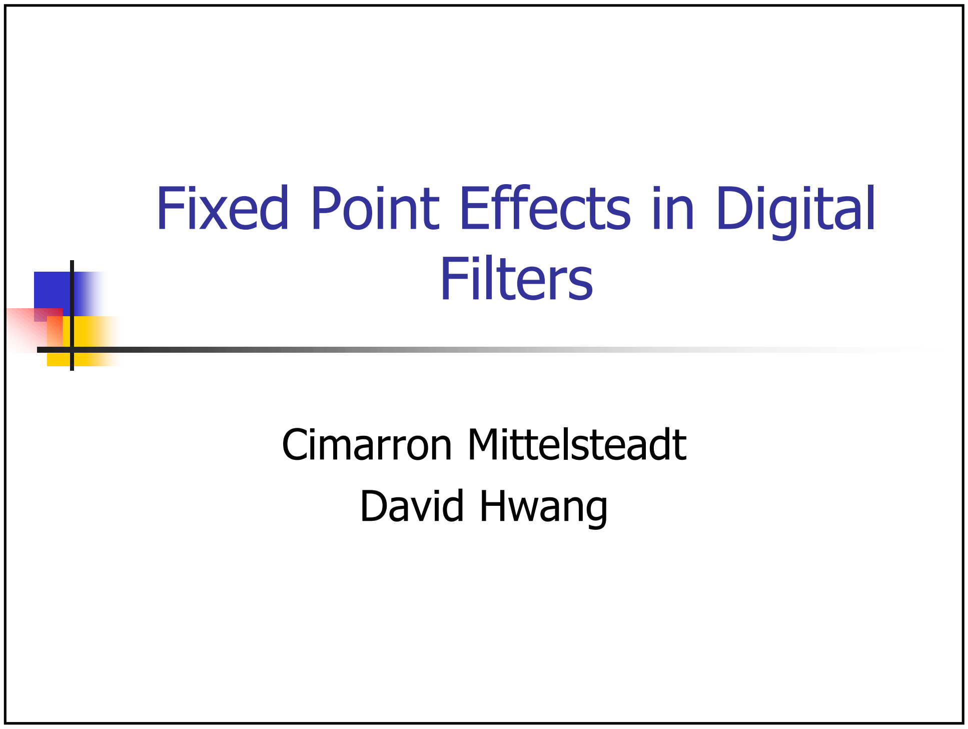
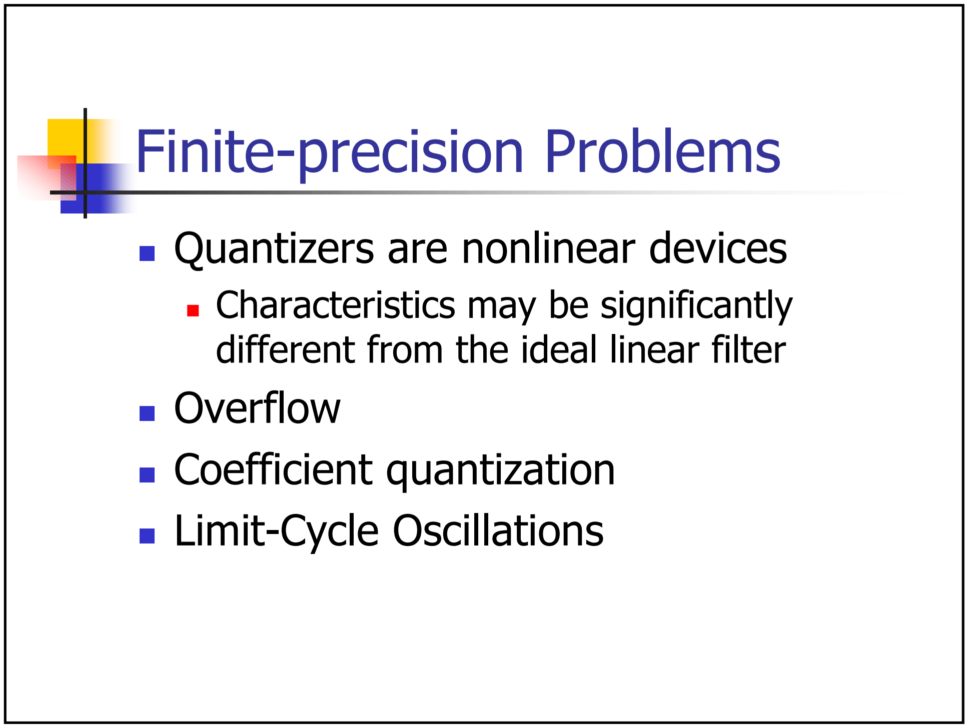


Fixed Point Effects in Digital Filters

Cimarron Mittelsteadt

David Hwang



Finite-precision Problems

- Quantizers are nonlinear devices
  - Characteristics may be significantly different from the ideal linear filter
- Overflow
- Coefficient quantization
- Limit-Cycle Oscillations

Quantizers

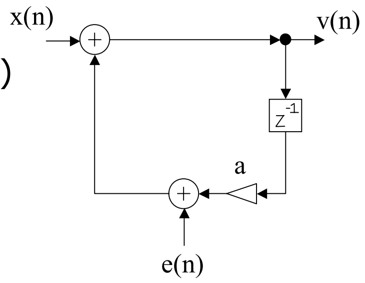
- Nonlinear effects make it extremely difficult to precisely analyze the filter's performance.
- How do we model a fixed-point filter then?
  - Adopt a statistical model of the quantization effects
  - Results in a linear model for the filter

Statistical Characterization

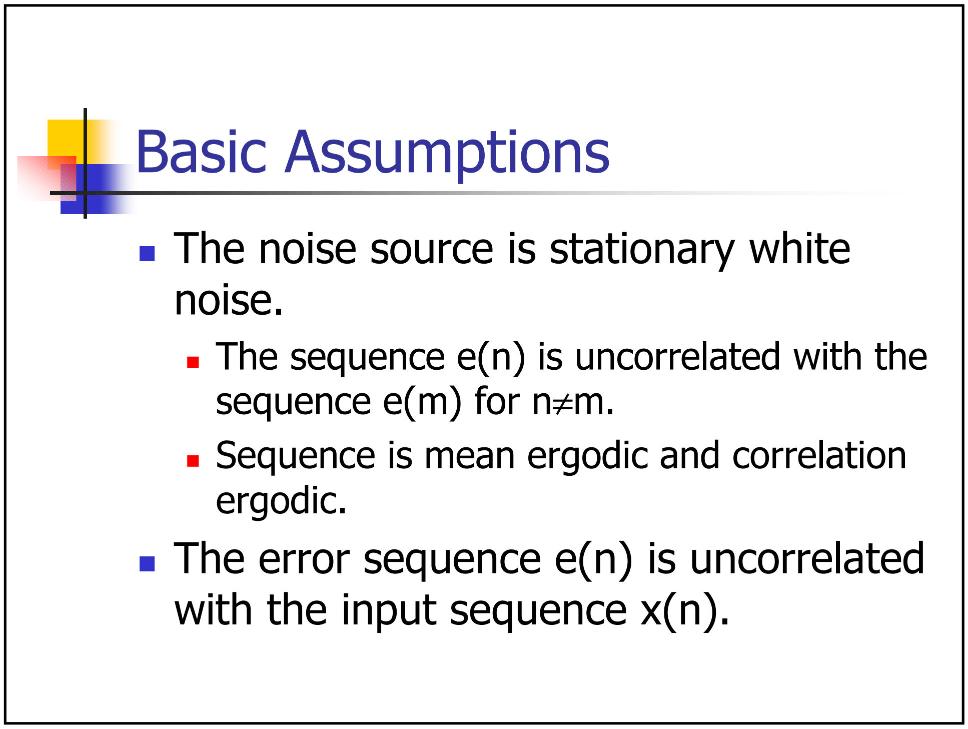
$v(n) = Q[av(n-1)]+x(n)$

$Q[av(n-1)] = av(n-1)+e(n)$

$v(n) = av(n-1)+x(n)+e(n)$

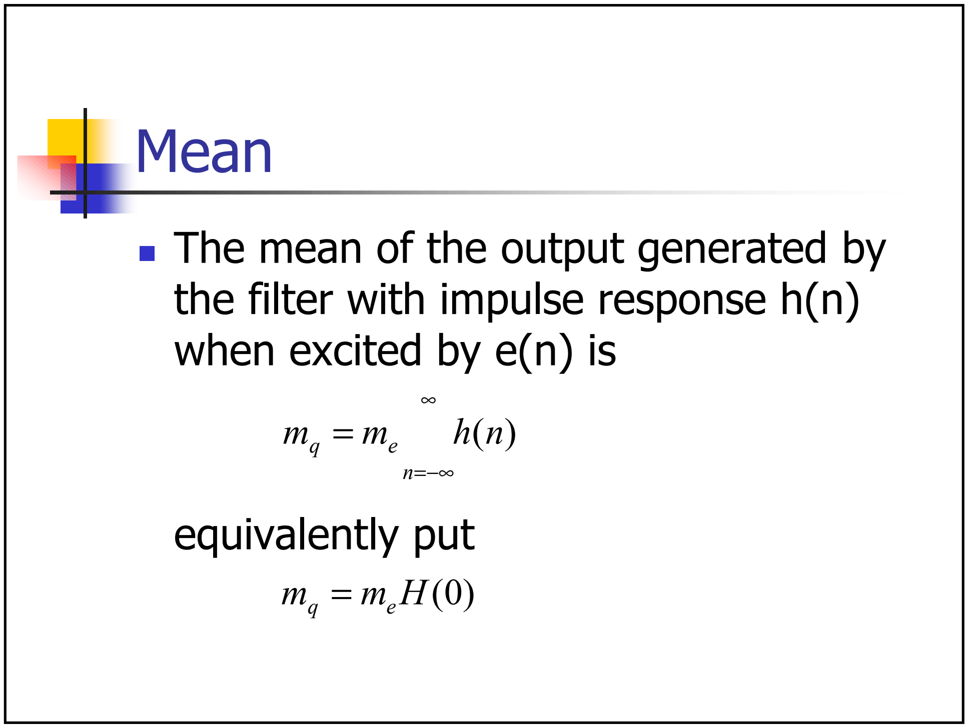


- We can now view the response of the filter as coming from two inputs.



Basic Assumptions

- The noise source is stationary white noise.
  - The sequence e(n) is uncorrelated with the sequence e(m) for n≠m.
  - Sequence is mean ergodic and correlation ergodic.
- The error sequence e(n) is uncorrelated with the input sequence x(n).



Mean

- The mean of the output generated by the filter with impulse response h(n) when excited by e(n) is

$$m_q = m_e \sum_{n=-\infty}^{\infty} h(n)$$

equivalently put

$$m_q = m_e H(0)$$

Autocorrelation

The autocorrelation is computed to be

$$\gamma_{qq}(n) = \sum_{k=-\infty}^{\infty}\sum_{l=-\infty}^{\infty} h(k)h(l)\gamma_{ee}(k-l+n)$$

This reduces to

$$\sigma_q^2 = \sigma_e^2 \sum_{k=-\infty}^{\infty} h^2(k)$$

By Parseval's theorem $\sigma_q^2 = \dfrac{\sigma_e^2}{2\pi}\int_{-\pi}^{\pi} \left|H(\omega)\right|^2 d\omega$

Types of Quantization

b fractional bits

- Rounding $m_e = 0$ $\sigma_e^2 = \dfrac{2^{-2b}}{12}$
- Truncation $m_e = -\dfrac{2^{-b}}{2}$ $\sigma_e^2 = \dfrac{2^{-2b}}{12}$
- Magnitude Truncation $m_e = 0$ $\sigma_e^2 = \dfrac{2^{-2b}}{3}$

Section Ordering is Important

- Example

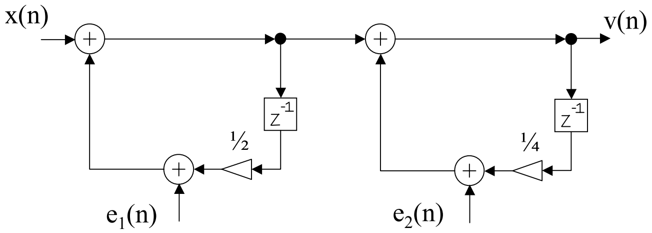
$$H(z) = H_1(z)H_2(z)$$

Where

$$H_1(z) = \frac{1}{1-\frac{1}{2}z^{-1}} \qquad H_2(z) = \frac{1}{1-\frac{1}{4}z^{-1}}$$

and their corresponding impulse responses are given by h(n), h_1(n) and h_2(n)

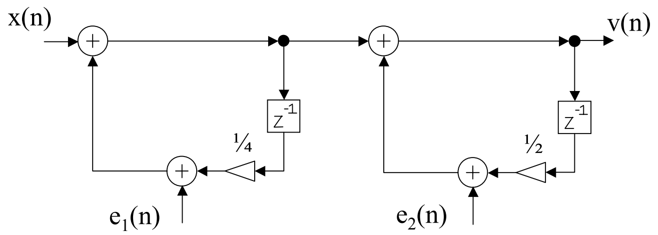
Realization 1



With rounding

$$\sigma_{q1}^2 = \sigma_e^2\left[\,\overset{\infty}{\underset{n=0}{}}\; h^2(n) + \overset{\infty}{\underset{n=0}{}}\; h_2^2(n)\right] \approx 2.90\sigma_e^2$$

Realization 2



With rounding

$$\sigma_{q2}^2 = \sigma_e^2\left[\underset{n=0}{\overset{\infty}{}} h^2(n) + \underset{n=0}{\overset{\infty}{}} h_1^2(n)\right] \approx 3.16\sigma_e^2$$

Section Ordering Comparison

- The overall noises were found to be

$$\sigma_{q1}^2 \approx 2.90\sigma_e^2 \qquad \frac{\sigma_{q2}^2}{\sigma_{q1}^2} \approx 1.09$$

$$\sigma_{q2}^2 \approx 3.16\sigma_e^2$$

- Thus, the second realization leads to 9% more noise power than the first.

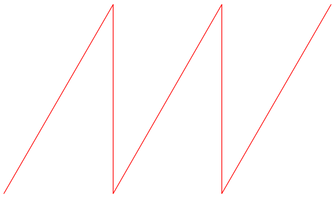
Overflow

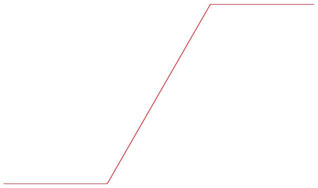
- Wrap Around
  - Ex.

  7 (0111) + 1 (0001) = -8 (1000)

- Saturation
  - Ex.

  5 (0101) + 4 (0100) = 7 (0111)





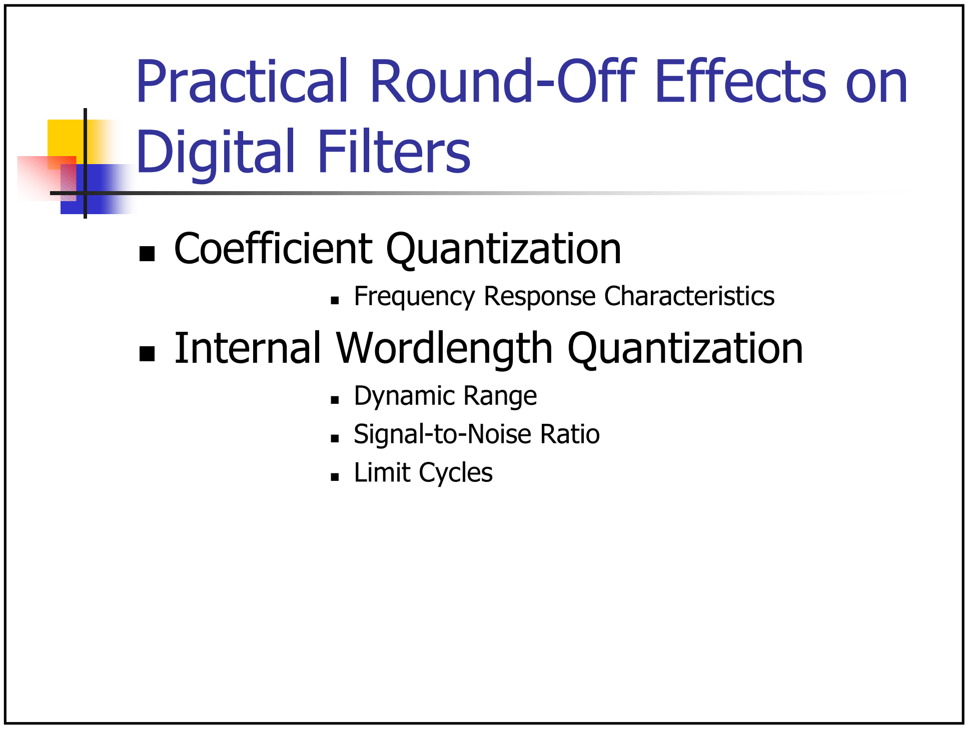
Scaling to Prevent Overflow

- Pessimistic Scaling

$$A_x < \frac{1}{\sum_{m=-\infty}^{\infty} \left| h_k(m) \right|}$$

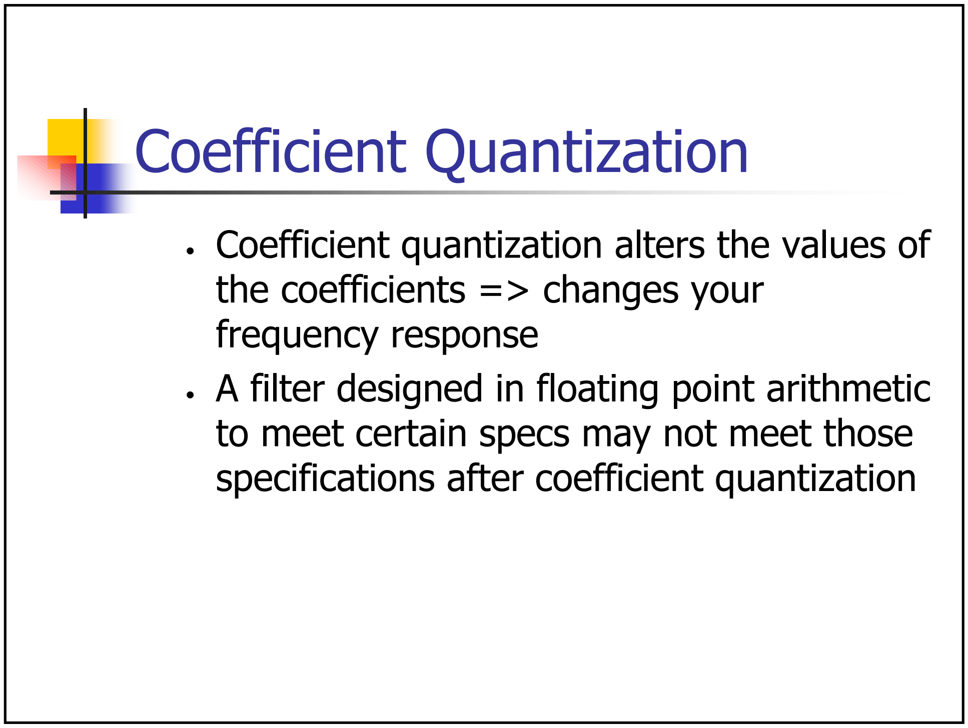
- Narrowband Scaling

$$A_x < \frac{1}{\max_{0 \le \omega \le 2\pi} \left| H_k(\omega) \right|}$$



Practical Round-Off Effects on Digital Filters

- Coefficient Quantization
  - Frequency Response Characteristics
- Internal Wordlength Quantization
  - Dynamic Range
  - Signal-to-Noise Ratio
  - Limit Cycles



Coefficient Quantization

- Coefficient quantization alters the values of the coefficients => changes your frequency response
- A filter designed in floating point arithmetic to meet certain specs may not meet those specifications after coefficient quantization

Coefficient Quantization Example

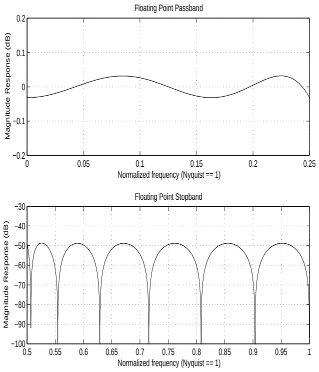
- We used the Parks-McClellan optimal FIR filter algorithm to design a 21-tap filter
- Designed the filter in floating point format
- Took each of the coefficients and rounded to the nearest 8-bit two's complement number
- Ex:

  0.2011929 => .2031250

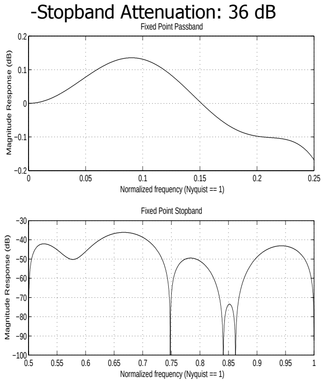
  (00011010 two's complement)
- Lesson: Over-design the filter and/or use an optimization algorithm to meet the spec

Frequency Response

Floating Point
-Passband Ripple: .003 dB
-Stopband Attenuation: 48 dB



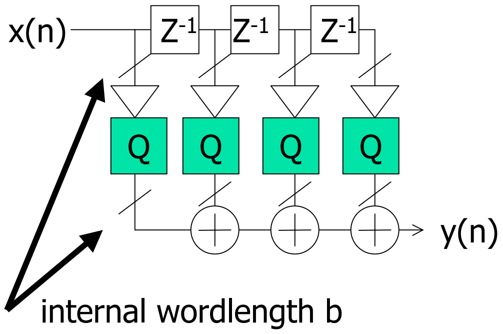
Fixed Point
-Passband Ripple: .14 dB
-Stopband Attenuation: 36 dB



Coefficient Quantization => Frequency Response Degradation

Internal Wordlength Quantization

- Quantization of internal wordlength leads to finite-wordlength effects



1) Dynamic Range
2) Signal-to-Noise Ratio
3) Limit Cycles

Dynamic Range Constraints

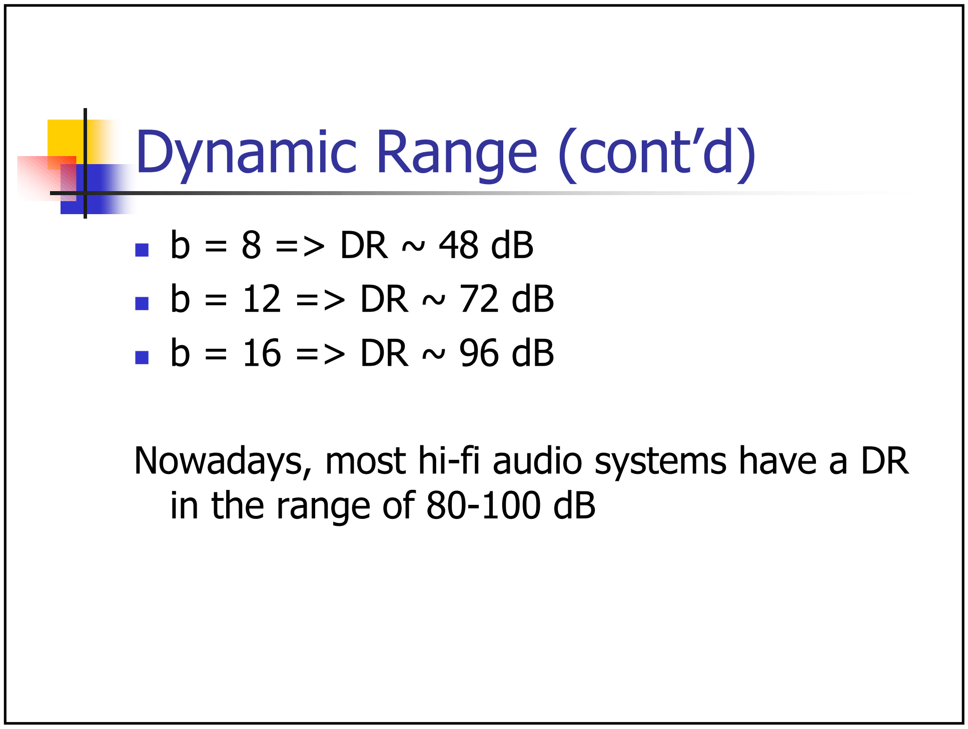
- Dynamic Range is defined as:

$$20 \log_{10} \frac{|\text{ range of representable numbers }|}{|\text{ smallest non-zero representable number }|}$$

- The larger the dynamic range specification, the larger internal wordlength b required
  - Ex. b-bit number ($x_0. x_1 x_2 x_3 x_4 \ldots x_{b-1}$)
  - DR: $20 \log |2 - 2^{-(b-1)}| \,/\, |\, 2^{-(b-1)} \,|$

    $= 20 \log (2^{-b} - 1)$

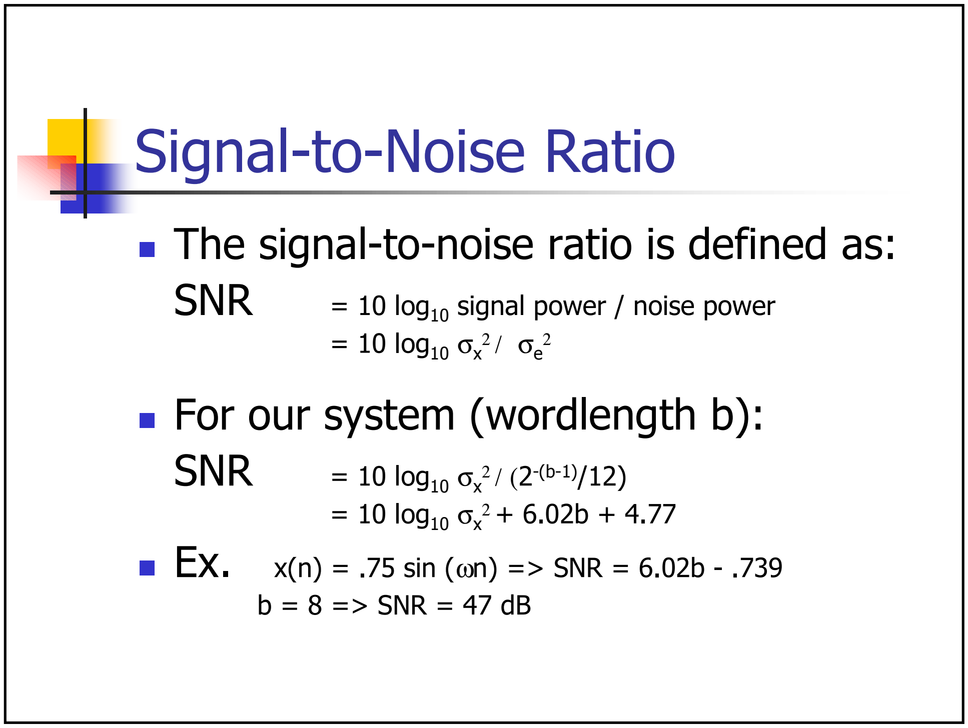
    ~ 6 dB / bit



Dynamic Range (cont'd)

- $b = 8 \Rightarrow DR \sim 48$ dB
- $b = 12 \Rightarrow DR \sim 72$ dB
- $b = 16 \Rightarrow DR \sim 96$ dB

Nowadays, most hi-fi audio systems have a DR in the range of 80-100 dB



Signal-to-Noise Ratio

- The signal-to-noise ratio is defined as:

  SNR $= 10 \log_{10}$ signal power / noise power

  $= 10 \log_{10} \sigma_x^2 / \sigma_e^2$

- For our system (wordlength b):

  SNR $= 10 \log_{10} \sigma_x^2 / (2^{-(b-1)}/12)$

  $= 10 \log_{10} \sigma_x^2 + 6.02b + 4.77$

- Ex. $x(n) = .75 \sin(\omega n) \Rightarrow \text{SNR} = 6.02b - .739$

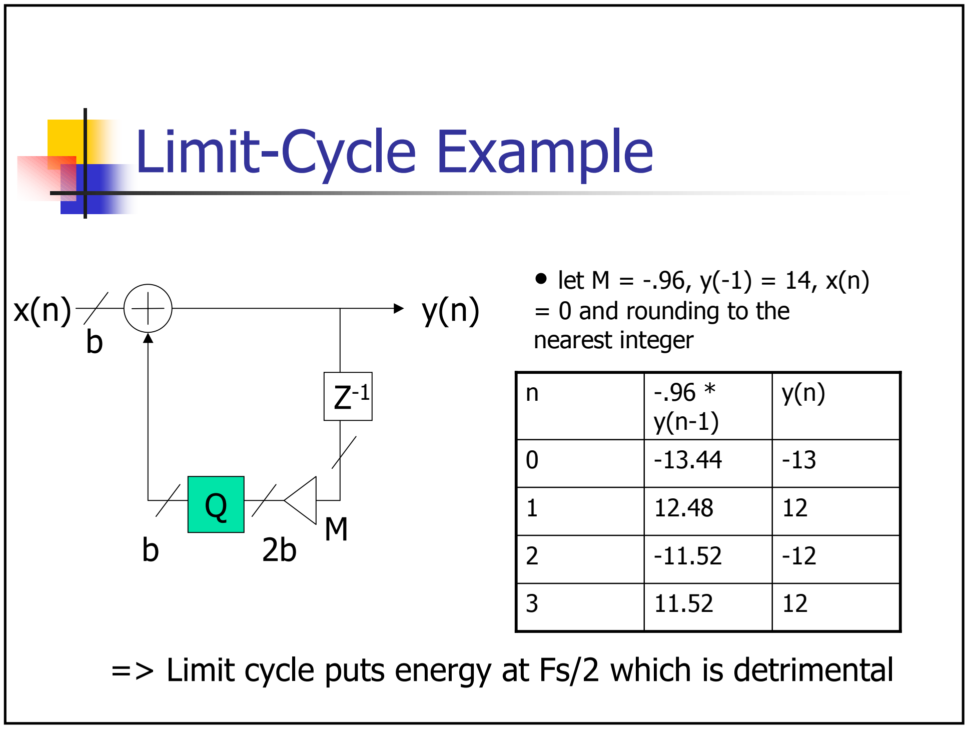
  $b = 8 \Rightarrow \text{SNR} = 47$ dB

Signal-to-Noise Ratio (cont'd)

- Thus, SNR increases by ~ 6 dB / bit
- To maximize SNR, you want to scale the signal x(n) as large as possible (to increase σ_x^2)
- However, there is a tradeoff—you need to keep all internal signals small enough to prevent overflow / saturation (use normalization and scaling techniques)

Limit Cycles

- Limit cycles occur when the output of a digital filter does not decay to zero when the input goes to zero
- Limit cycles occur only in IIR filters and never occur in FIR filters
- They are caused by quantizing the data after a feedback multiplier in a recursive loop

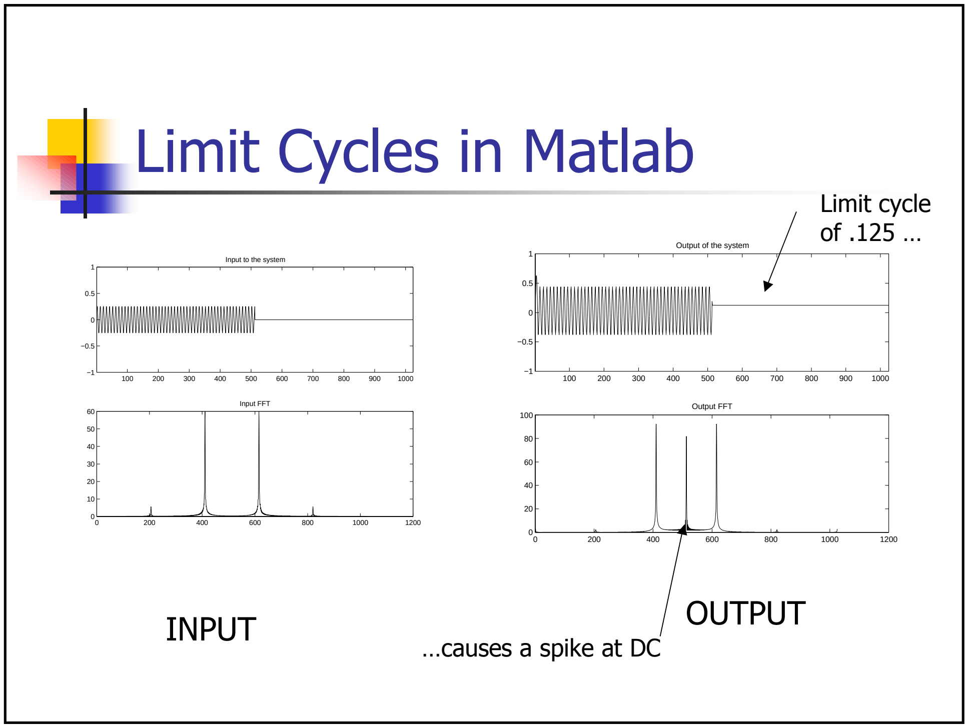


Limit-Cycle Example

- let M = -.96, y(-1) = 14, x(n) = 0 and rounding to the nearest integer

n	-.96 * y(n-1)	y(n)
0	-13.44	-13
1	12.48	12
2	-11.52	-12
3	11.52	12

=> Limit cycle puts energy at Fs/2 which is detrimental



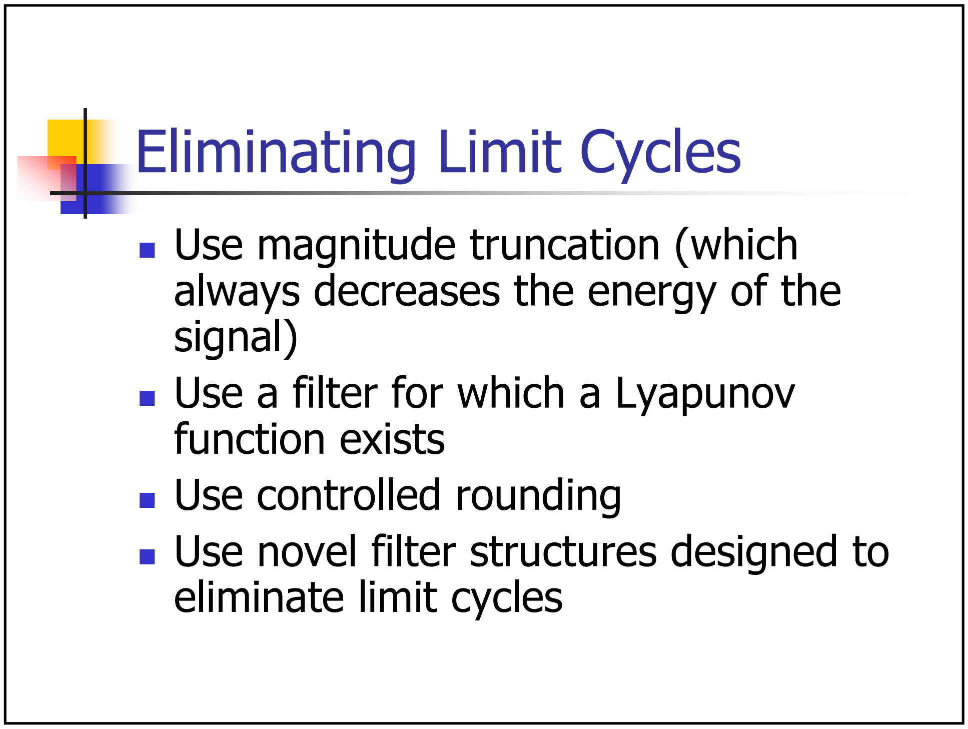
Limit Cycles in Matlab

Limit cycle of .125 …

INPUT

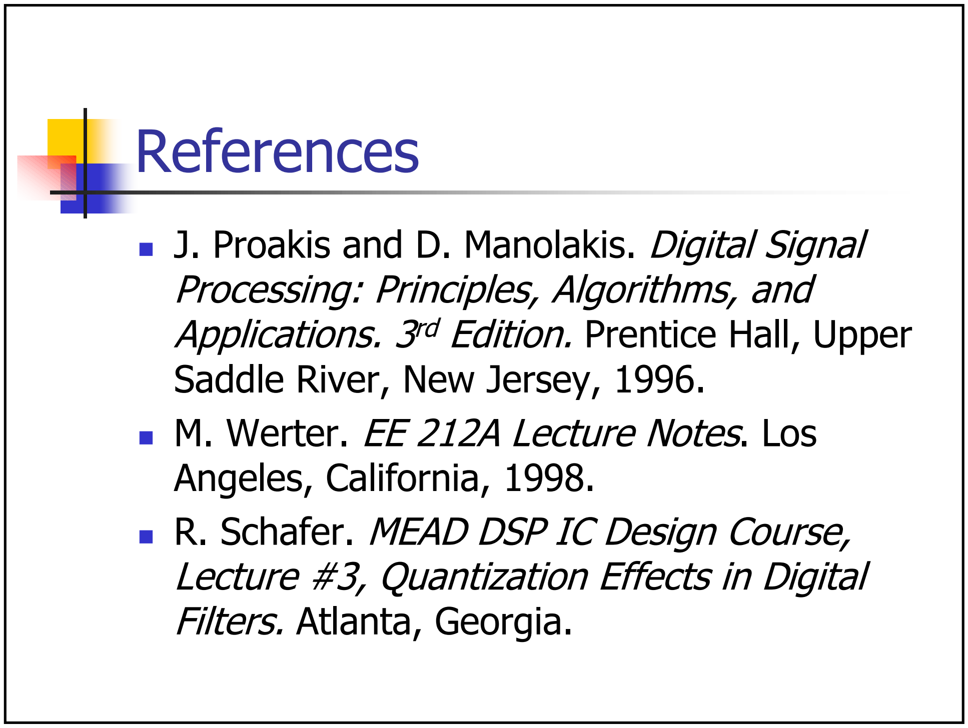
OUTPUT

…causes a spike at DC



Eliminating Limit Cycles

- Use magnitude truncation (which always decreases the energy of the signal)
- Use a filter for which a Lyapunov function exists
- Use controlled rounding
- Use novel filter structures designed to eliminate limit cycles



References

- J. Proakis and D. Manolakis. *Digital Signal Processing: Principles, Algorithms, and Applications. 3rd Edition.* Prentice Hall, Upper Saddle River, New Jersey, 1996.
- M. Werter. *EE 212A Lecture Notes*. Los Angeles, California, 1998.
- R. Schafer. *MEAD DSP IC Design Course, Lecture #3, Quantization Effects in Digital Filters.* Atlanta, Georgia.