Professor Brendan Morris, SEB 3216, brendan.morris@unlv.edu

EE482: Digital Signal Processing Applications

Spring 2014
TTh 14:30-15:45 CBC C222

Lecture 02
Numerical effects
14/01/28

http://www.ee.unlv.edu/~b1morris/ee482/

Outline

- Random Variables
- Fixed-Point Numbers
- Quantization Errors
- Arithmetic Errors

Random Variables

- Function that maps from a sample space to a real value
  - $x : S \to \mathbb{R}$
    - x – random variable (does not have a value)
    - S – sample space
- Cumulative probability function (CDF)
  - $F(X) = P(x \leq X)$
    - E.g. probability $\{x \leq X\}$
- Probability density function
  - $f(X) = \frac{dF(X)}{dX}$
    - $\int_{-\infty}^{\infty} f(X)dX = 1$
    - $P(X_1 < x \leq X_2) = F(X_2) - F(X_1) = \int_{x_1}^{x_2} f(X)dX$
    - For discrete x, takes values $X_i,\ i = 1, 2, 3, \ldots$
      - $p_i = P(x = X_i)$

Uniform Random Variable

- Variable takes on value in a range with equal probability
- $f(X) = \begin{cases} \frac{1}{X_2 - X_1} & X_1 \leq x \leq X_2 \\ 0 & else \end{cases}$

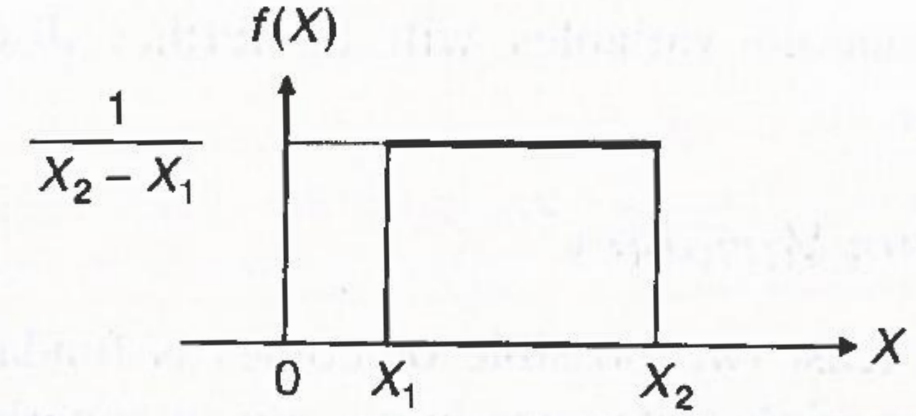


Figure 2.17 The uniform density function

- Be sure you can calculate mean and variance

Statistics of Random Variables

- Expected value (mean)
  - $m_x = E[x]$ expectation operator
    - $m_x = \int_{-\infty}^{\infty} X f(X) dX$ continuous
    - $m_x = \sum_i X_i p_i$ discrete
  - Can be can computed with `mean.m`
- Variance (spread around mean)
  - $\sigma_x^2 = E[(x - m_x)^2] = E[x^2] - m_x^2$
    - $\sigma_x^2 = \int_{-\infty}^{\infty} (X - m_x)^2 f(X) dX$ continuous
    - $\sigma_x^2 = \sum_i p_i (X_i - m_x)^2$ discrete
  - For $m_x = 0$,
    - $\sigma_x^2 = E[x^2] = P_x$ second moment, power

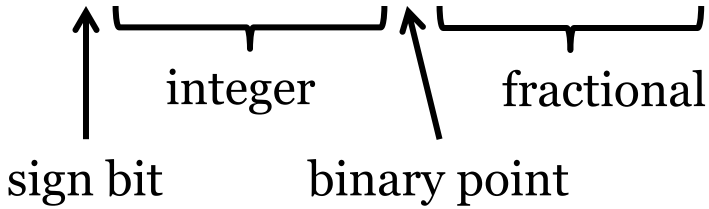
Fixed-Point Numerical Effects

- Fractional numbers are represented in 2's complement with $B = M + 1$ bits
- $x = b_0 . b_1 b_2 \ldots b_{M-1} b_M$

  - $b_0 = \begin{cases} 0 & x \geq 0 \quad \text{positive} \\ 1 & x < 0 \quad \text{negative} \end{cases}$
  - value $= -b_0 + \sum_{m=1}^{M} b_m 2^{-m}$
    - $-1 \leq x \leq (1 - 2^{M})$
    - Unbalanced range with more negative than positive numbers

General Fractional Format Qn.m

- $x = b_0 b_1 b_2 \dots b_n . b_1 b_2 \dots b_M$



- Example 2.25
- $x = 0100\ 1000\ 0001\ 1000b = 0x4818$
- Q0.15
    - $x = 2^{-1} + 2^{-4} + 2^{-11} + 2^{-12} = 0.56323$
- Q2.13
    - $x = 2^{1} + 2^{-2} + 2^{-9} + 2^{-10} = 2.25293$
- Q5.10
    - $x = 2^{4} + 2^{1} + 2^{-6} + 2^{-7} = 18.02344$

Finite Word Length Effects

1. Quantization errors
   - Signal quantization
   - Coefficient quantization
2. Arithmetic errors
   - Roundoff (truncation)
   - Overflow

Signal Quantization

- ADC conversion of sampled signals to fixed levels
- Using Q15 and B bits
  - Dynamic range $-1 \leq x < 1$
  - Quantization step
    - $\Delta = \frac{2}{2^B} = 2^{-B+1} = 2^{-M}$
- Quantization error
  - $e(n) = x(n) - x_B(n)$
    - $x_B(n) = Q[x(n)]$
  - $|e(n)| \leq \frac{\Delta}{2} = 2^{-B}$ (rounding)
    - Error dependent on word length B
    - More bits for better resolution, less error (noise)
- Signal to quantization noise (SQNR)
  - $SQNR = \frac{\sigma_x^2}{\sigma_e^2} = 3.2^{2B}\sigma_x^2$
  - $SQNR = 4.77 + 6.02B + 10\log_{10}\sigma_x^2 \ dB$

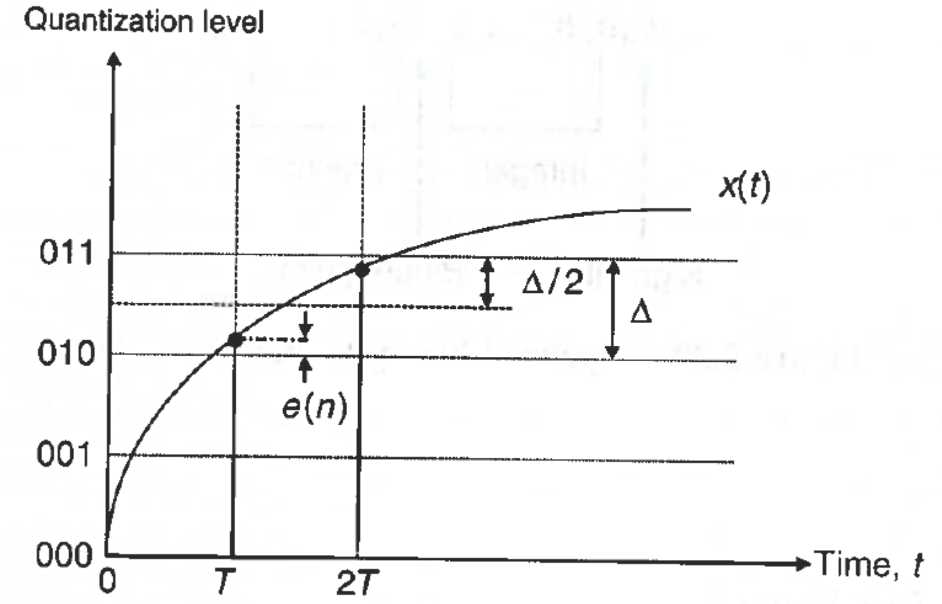


Figure 2.21 Quantization process related to a 3-bit ADC

Coefficient Quantization

- Same error issues as for signals
- Results in movement of the locations of poles/zeros
  - Changes system function polynomials
  - Can lead to instability if poles go outside the unit circle
    - Generally, more a problem with IIR filters
- Can limit coefficient quantization effects by using lower-order filters
  - Use of cascade and parallel filter structures

Roundoff Noise

- A product must be represented in B bits by rounding (truncation)
    - $$\underset{\overset{\uparrow}{2B\text{ bits}}}{y}(n) = \underset{\overset{\uparrow}{B\text{ bits}}}{\alpha}\,\underset{\overset{\nwarrow}{B\text{ bits}}}{x}(n)$$
- Error model
    - $y(n) = Q[\alpha x(n)] = \alpha x(n) + e(n)$
        - $e(n)$ is uniformly distributed zero mean noise (rounding)

Overflow

- $y(n) = x_1(n) + x_2(n)$
  - $-1 \leq x_i(n) < 1$
  - $-1 \leq y(n) < 1$
- Overflow occurs when the sum cannot fit in the word container
- Signals need to be scaled to prevent overflow

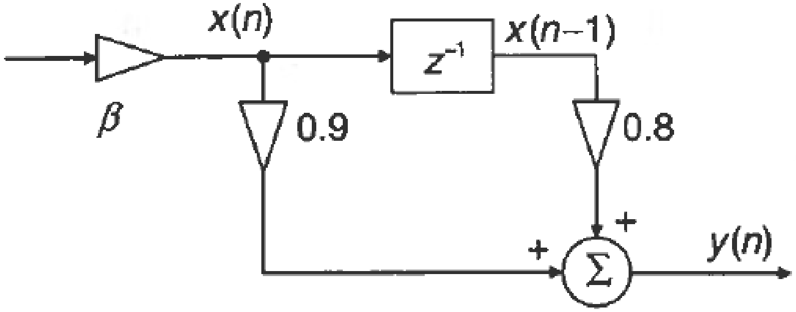


Figure 2.24 Block diagram of simple FIR filter with scaling factor β

- Notice: this reduces the SQNR
  - $SQNR = 10\log_{10}(\frac{\beta^2\sigma_x^2}{\sigma_e^2})$
  - $SQNR = 4.77 + 6.02B + 10\log_{10}\sigma_x^2 + \underbrace{10\log_{10}\beta}_{\text{negative}}\ dB$