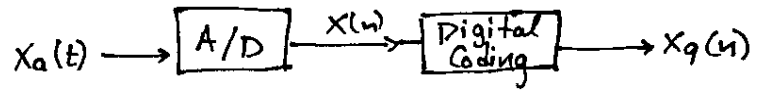
ECE 431 F

Lecture 32

(FINITE PRECISION)

❽ EFFECTS OF ROUNDOFF NOISE IN DSP – APPLICATION IN DIGITAL FILTERS



binary digits representing quantized magnitude of signal

Assumption: $|x_q(n)| < 1$ so that $x_q(n)$ ~~can assume~~ is represented by a binary fraction.

(No loss of generality since a scaling factor can be introduced to relate binary numbers to analog signal amplitudes)

① FIXED POINT REPRESENTATION (usually in two's complement)

$$x_q(n) = B_0 . B_1 B_2 B_3 \ldots B_b$$

(B_0: sign bit)

(infinite precision arithmetic: $b \to \infty$)
(Finite " " : b is finite)

- All numbers have length b bits (plus one bit for sign)
- Result of multiplication of 2 b-bit numbers is b bits long
- Overflow may occur in addition unless we choose long enough accumulators
- Truncation or Rounding might occur in multiplication

② FLOATING POINT REPRESENTATION

- $x_q(n) = M \cdot 2^p$
  - M: mantissa, $\frac{1}{2} \le |M| < 1$
  - p: exponent, positive or negative integer
- M, p assume fixed point representation

③ FIXED VS FLOATING POINT.

FIXED	FLOATING POINT.
– 80% of market	, 20% of market
– Fixed resolution over all dynamic ranges	, Variable resolution across range (i.e. finer resolution for smaller Numbers)
– Need for scaling to avoid overflow	, Larger dynamic range ("no overflows")
– Cheaper and simpler.	, 4 times more current/instruction, bulkier.

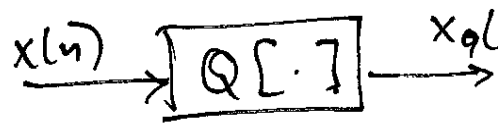
STATISTICAL CHARACTERIZATION OF QUANTIZATION EFFECTS:

Rounding: Numbers are rounded (i.e. choose the closest quantization level) to fit into a finite-length register.

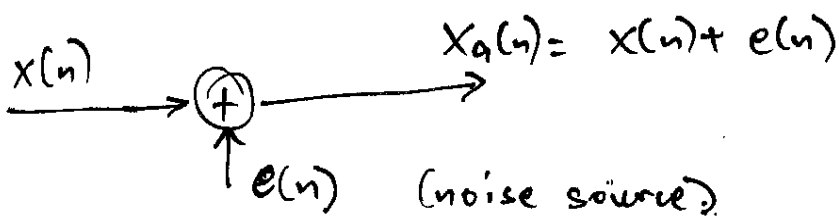
* In practice a precise analysis of rounding (or other effects) is not necessary. Thus an approximate statistical analysis suffices (and difficult)

** Statistical analysis: Introduce a noise source in place of each value quantization. Then, find the power of the total noise introduced. (i.e., effect of all quantization errors)

Quantization of a single value:



or



Multiplication in Fixed point arithmetic: $Q_{Fix}[x_1(n)\,x_2(n)] = x_1(n)\,x_2(n) + e(n)$

Addition and multiplication in Floating point arithmetic:

$$Q_{fl}[x_1(n)x_2(n)] = x_1(n)\,x_2(n)\,[1+\varepsilon_1(n)]$$

$$Q_{fl}[x_1(n)+x_2(n)] = [x_1(n)+x_2(n)]\,[1+\varepsilon_2(n)]$$

Assumptions:

① All errors are white - zero mean noise sequences

② **FIXED POINT AR.**

For register length (b+1) bits (one bit for sign). errors are uniformly distributed in range

$$-\frac{1}{2}2^{-b} \le e(n) \le \frac{1}{2}2^{-b}$$

with zero mean ($m_e = 0$) and variance (power)

$$\sigma_e^2 = \frac{1}{12}2^{-2b}$$

FLOATING POINT ARITHMETIC.

For mantissa register of length (b+1) bits errors are uniformly distributed in the range

$$-2^{-b} \le \varepsilon_i(n) \le 2^{-b}$$

with zero mean: $m_{\varepsilon_i} = 0$ and with variances (power)

$$\sigma_{\varepsilon_i}^2 = \frac{1}{3}2^{-2b}$$

③ All error sources at any point are uncorrelated with each other and with the signal values (i.e., $E\{e_i(n)\,e_j(n)\} = 0$)

④ Quantization effects due to arithmetic operations are much more severe than quantization of signal values.

PROCEDURE to find the effect of rounting (round-off) noise at the output of a digital filter.

① Model the quantization effect by introducing a noise source at the output of each multiplication (in fixed point) or at the output of each multiplication and addition (in floating point)

② Calculate the mean and variance (m_y and σ_y^2) of noise at the output of the filter

③ Calculate mean and variance of the filter output. (m_y and σ_y^2)

④ Calculate signal to noise ratio $\frac{\sigma_y^2}{\sigma_{y_n}^2}$ or other quantities.....

Lecture 13

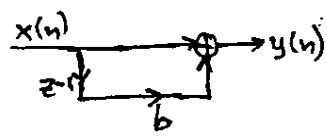
IMPORTANT FORMULAS TO REMEMBER IN APPLICATIONS

Let $e(n)$ the quantization noise error source (random quantity) and
let $h(n)$ be the impulse response of an LSI system
If $n(n) = e(n) * h(n)$ (linear convolution operation)
then

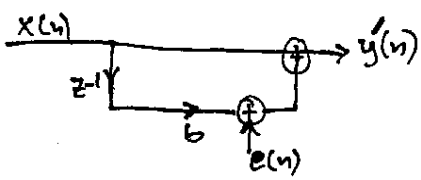
- mean: $E\{n(n)\} = m_n = m_e \cdot \sum_n h(n)$
- variance: $E\{(n(n)-m_n)^2\} = \sigma_n^2 = \sigma_e^2 \cdot \sum_n h^2(n)$

Simple examples with fixed point representation

Ex.1. Let $y(n) = x(n) + bx(n-1)$: DIRECT REALIZ. (no noise)

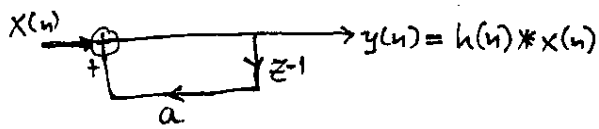


with noise

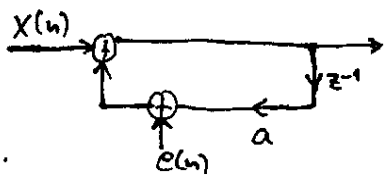


In this case easy to show that
$y'(n) = x(n) + bx(n-1) + e(n) = y(n) + e(n)$
So the total noise at the output is $n(n) = e(n)$
and therefore $m_n = m_e = 0$, $\sigma_n^2 = \sigma_e^2 = \frac{1}{12} 2^{-2b}$

Ex.2 Let $y(n) = a\,y(n-1) + x(n)$: $a < 1$



with noise



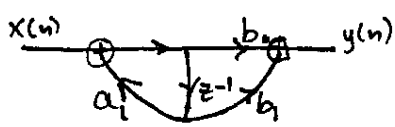
$y'(n) = [x(n) + e(n)] * h(n) = \underbrace{x(n) * h(n)}_{y(n)} + \underbrace{e(n) * h(n)}_{n(n)}$

So $m_n = m_e \sum_n h(n) = 0$

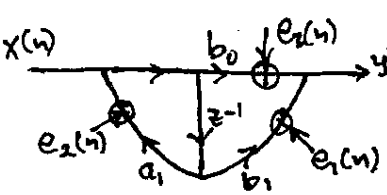
[in our case $h(n) = z^{-1}\left[\frac{1}{1-az^{-1}}\right] = a^n u(n)$]

$\sigma_n^2 = \sigma_e^2 \sum_n h^2(n) = \frac{1}{12} 2^{-2b} \sum_{n=0}^{\infty} a^{2n} = \frac{1}{12} 2^{-2b} \frac{1}{1-a^2}$

Ex.3. $y(n) = a_1 y(n-1) + b_0 x(n) + b_1 x(n-1)$:



with noise



$y'(n) = x(n) * h(n) + e_2(n) * h(n) + e_2(n) + e_3(n)$

So $n(n) = e_2(n) * h(n) + e_2(n) + e_3(n)$
$m_n = 0$
$\sigma_n^2 = \sigma_e^2 \sum_n h^2(n) + 2\sigma_e^2 = \cdots\cdots$