Finite Word Length Effects and Quantisation Noise

Professors A G Constantinides & L R Arnaut

Finite Word Length Effects

- Finite register lengths and A/D converters cause errors at different levels:

  (i) *input*:

  Input quantisation

  (ii) *system*:

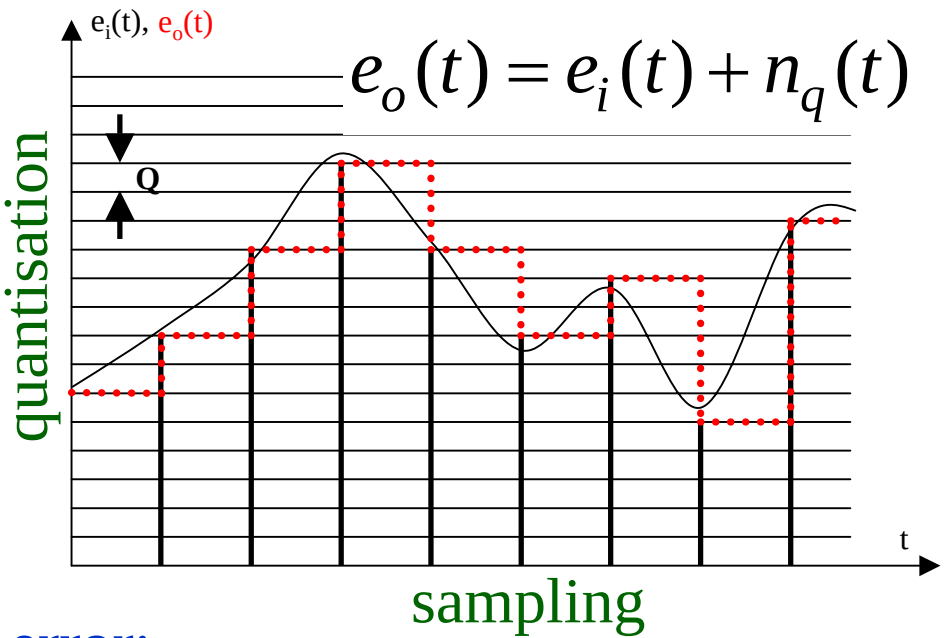
  Coefficient (= multiplier) quantisation

  (iii) *output*:

  Operation (products) truncated or rounded due to finite machine word length

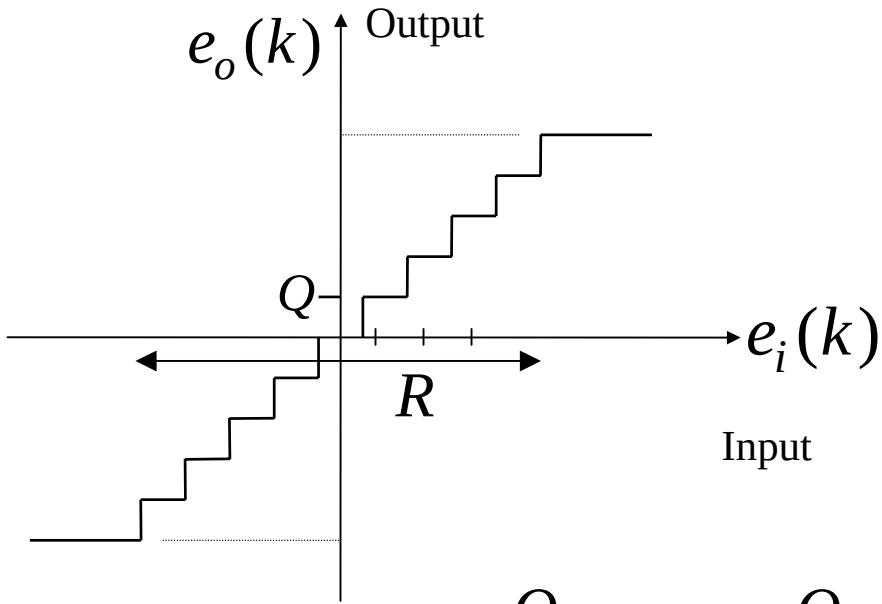
Input Quantisation

- Discretisation of time (sampling) vs. signal values
  - For each sample: search nearest level & round off actual value to this level (S&H) ⇒ finite precision digitized value
  - Most values cannot be represented exactly
    ⇒ *quantisation error e(k)*
  - Rounded value can be above or below actual value: random bounded fluctuation of quantisation error:
    ⇒ conceived as *quantisation noise* $n_q(t)$



Quantisation Error (Linear)

- Quantisation error e for quantisation step Q:



Linear mid-*tread* quantiser

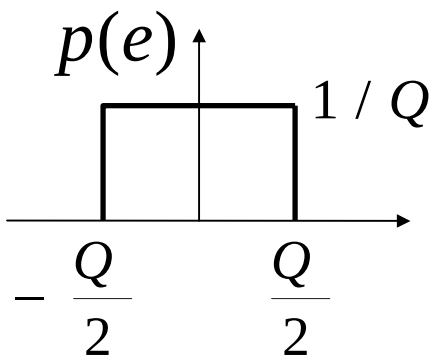
$e_i(k)$ to stay within dynamic range R to avoid clipping of $e_o(k)$ (overflow error)

- Nearest-level rounding $\Rightarrow$ $-\frac{Q}{2} \le e(k) \le \frac{Q}{2}$
- Q = least significant bit (determines precision, resolution)
- Quantisation is irreversible process ($\Rightarrow$ loss of information)

  $\leftrightarrow$ sampling above Nyquist rate

Uniform Quantisation for Rounding

- Ideally, pdf of *e* for rounding assumes uniform distribution between bounds (granular noise)



- Not exactly true for sinusoidal signal, but corrections are generally small for high-amplitude (=spanning range of quantiser), wide-band signals

- Rounding preferred over truncation, because it always yields *unbiased* quantisation error

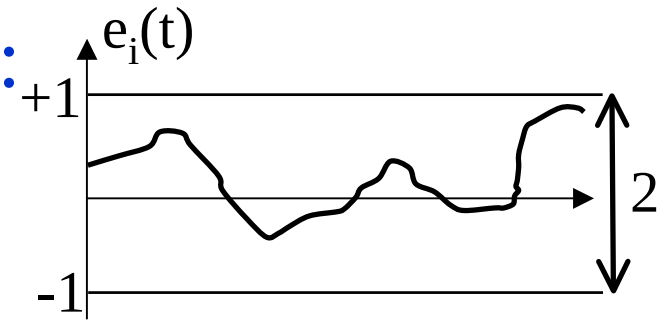
- Quantisation noise power (mean-square error):

$$\sigma^2 = \int_{-Q/2}^{Q/2} e^2 p(e)\,\mathrm{d}e = E\{e^2\} = \frac{(Q/2)^2}{3} = \frac{Q^2}{12}$$

Quantisation Unit-Amplitude Signal

- Let input = signal of unit amplitude (e.g. sine)
- Normalised total signal power:

$$P = \left(\frac{1}{\sqrt{2}}\right)^2 = \frac{1}{2}$$



- If b bits used for binary:

2^b subintervals, so that $Q = 2/2^b$ hence $\sigma^2 = 2^{-2b}/3$

- Thus, $\text{SNR} = P/\sigma^2 = \frac{3}{2}.2^{+2b}$

$$\text{SNR} = 10\log_{10}(1.5) + 10\log_{10}(2^{2b}) = (1.76 + 6b)\,\text{dB}$$

- also called SQNR (signal-to-quantisation-noise ratio)

- finiteness of dynamic range affects (reduces) value of 1.76

Applications

- DECT (speech)
  - One requires: SNR > 25 dB, i.e., 4 bits
  - In practice:
    - extra 30 dB of dynamic range required
    - largest signal for nonsinusoidal signal corresponds to 0 dB, not -3 dB
  - Thus, -30-25=-55 dB below maximum; requires minimum 9 bit-encoding
- HiFi (e.g., CD/DVD, in-flight data streaming):
  - Min. 12 bit-encoding (SNR = 78.3 dB); typ. 16-bit (96 dB) because of imperfect A/D devices
- In practice: using companders (compressor+expander):
  - Pre-emphasis of input signal before quantisation
  - Takes advantage of nonuniform probability of analogue values
  - Result: nonuniformly quantised output signal (use small Q when signal is small) $\Rightarrow$ S-shaped (nonlinear) I/O characteristic
  - Noise reduction in audio: Dolby, dbx

Companders

- Comparison:
  - Uniform quantization, 7 bits + 1 sign bit: SNR=44 dB
  - Nonuniform quantisation (nonlinear transformation):
    - μ-255 characteristic (US, Japan) (UNIX sound, JAVA, etc.):
      - Exploits logarithmic characteristic of loudness (human sound perception)
      - Allows for digitization of speech using only 8 bits for SNR=77dB

$$f(x) = \mathrm{sgn}(x)\frac{\ln(1+\mu|x|)}{\ln(1+\mu)}, \quad -1 < x < +1$$

    - A-87.6 characteristic (Europe): international convention

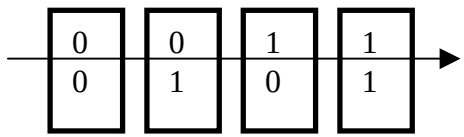
$$f(x) = \begin{cases} \mathrm{sgn}(x)\dfrac{A|x|}{1+\ln(A)}, & 0 \le |x| \le 1/A \\ \mathrm{sgn}(x)\dfrac{1+\ln|Ax|}{1+\ln(A)}, & 1/A \le |x| \le 1 \end{cases}$$

Digital Telephone

- Digital telephone transmission: voice
  - B = 3.4 kHz (analogue)
  - Sampling: F = 2 B = 6.8 kHz (ideal filters)
  - With filter margin: 8 kHz
  - 8 bits/sample, 8k samples/sec $\Rightarrow$ 64 kbps

Multilevel Coding

- Data transmission: Multilevel coding
  - Instead of transmission of single bits: grouping bits, e.g., in pairs:



    - lower symbol rate (i.e., smaller channel bandwidth) or higher bit rate, but decoding more difficult (larger quantisation error, or larger power needed)
    - Illustrates the dilemma “noise vs. bandwidth (speed) vs. accuracy vs. power”
  - Symbol rate (modulation rate) = 1 / duration of 1 pulse
    - Expressed in baud
  - Bit rate = symbol rate × number of bits per symbol
    - Expressed in bps
  - In multilevel coding: symbol rate ≠ bit rate

Coefficient Quantisation: Second-Order Systems

- Consider a simple example of finite precision of the coefficients a,b of a second order system with two complex conjugate poles $\rho e^{\pm j\theta}$:

$$H(z) = \frac{1}{1 + az^{-1} + bz^{-2}} = \frac{1}{1 - 2\rho\cos\theta\, z^{-1} + \rho^2 z^{-2}}$$

where

$$a = -2\rho\cos\theta, \qquad b = \rho^2$$

- Quantisation error of coefficients affects location of poles and zeroes $\Rightarrow$ imperfect frequency response

- Sensitivity of frequency response of filter to quantisation error is minimised if filter is implemented as cascade of 2nd-order filters (can be shown)

Coefficient Quantisation

- For $H(z) = \dfrac{1}{(1 + b_1 z^{-1} + b_2 z^{-2})}$

  instability (oscillation) can occur for $|b_2| \xrightarrow{<} 1$

  i.e., when poles of *H(z)* are either
  - (i) both on unit circle when complex, or
  - (ii) one real pole outside unit circle

- Instability under the "effective pole" model is considered as follows

Effective Pole Model

- In the time domain, from $H(z) = Y(z) \big/ X(z)$:

$$y(n) = x(n) - b_1 y(n-1) - b_2 y(n-2)$$

- With $|b_2| \to 1$, instability issue means $Q[b_2 y(n-2)]$ is indistinguishable from $y(n-2)$, where $Q[\cdot]$ represents quantisation operation

Effective Pole Model

- With rounding,

$$b_2 y(n-2) \pm 0.5 \quad \text{and} \quad y(n-2)$$

are indistinguishable (for integers) if

$$b_2 y(n-2) \pm 0.5 = y(n-2)$$

- Hence

$$y(n-2) = \frac{\pm 0.5}{1 - b_2}$$

- With both positive and negative b_2 :

$$\boxed{y(n-2) = \frac{\pm 0.5}{1 - |b_2|}}$$

$|b_2|=1$ is the effective pole for coefficient quantisation noise (oscillation)

Dead Band – Limit Cycle

- The range of integers $\dfrac{\pm 0.5}{1-|b_2|}$

  constitutes a set of integers that cannot be individually distinguished as separate or distinguished from asymptotic system behaviour.
- The band of integers $\left(-\dfrac{0.5}{1-|b_2|},\ +\dfrac{0.5}{1-|b_2|}\right)$

  is known as the *dead band*.
- In the second order system, under rounding, the output assumes a cyclic set of values of the dead band. This is a *limit cycle*.
- Dead band ≠ hysteresis (no action in dead band)

Effective Pole: Oscillations

- Consider the transfer function

$$G(z) = \frac{1}{(1 + b_1 z^{-1} + b_2 z^{-2})}$$

$$y_k = x_k - b_1 y_{k-1} - b_2 y_{k-2}$$

- If poles are complex then discrete impulse (unit sample) response sequence h_k is

$$h_k = \frac{\rho^k}{\sin\theta}.\sin[(k+1)\theta]$$

with

$$\rho = \sqrt{b_2}, \qquad \theta = \cos^{-1}\left(\frac{-b_1}{2\sqrt{b_2}}\right)$$

Effective Pole: Oscillations

- If $b_2 = 1$ then the response is sinusoidal (oscillatory) with angular frequency

$$\omega = \frac{1}{T}\cos^{-1}\left(-b_1/2\right)$$

- Thus, product quantisation causes instability implying an "effective pole" at $b_2 = 1$.

Limit Cycle of 2nd Order System: Example

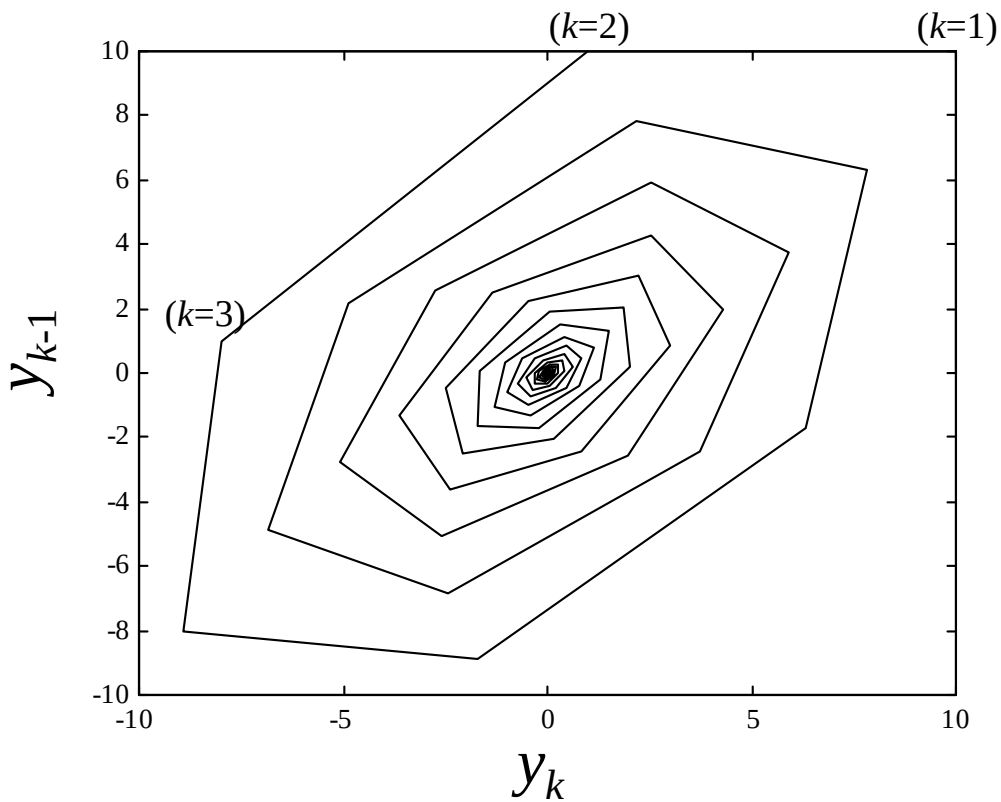
- Consider infinite precision computations for

$$y_k = x_k + y_{k-1} - 0.9y_{k-2}$$

$$x_0 = 10$$

$$x_k = 0;\ k \neq 0$$

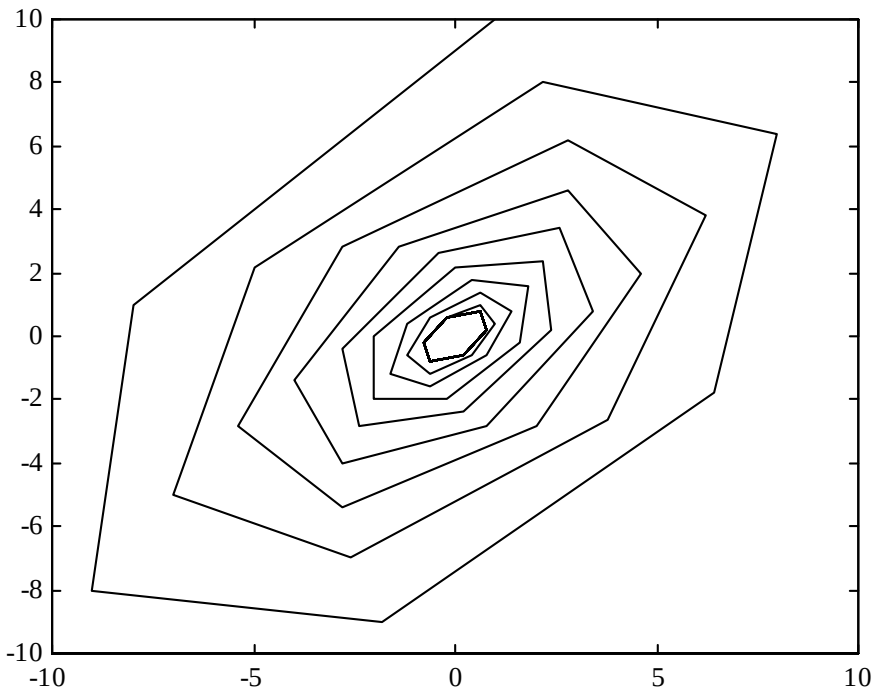
$$y_k = 0;\ k < 0$$



response converges to the origin without limit

Limit Cycle of 2nd Order System: Example

- Now the same operation with integer precision



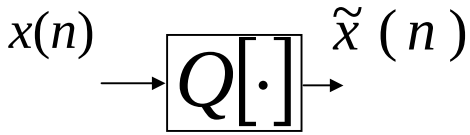
-Reponse does not converge to the origin, but assumes cyclically a set of values with nondecreasing quantisation error: the *Limit Cycle*

-System can only be driven out of its limit cycle only if new & *sufficiently large* input is applied

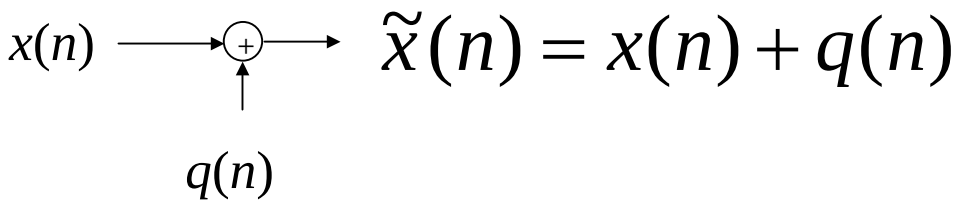
-*Truncation* (as opposed to rounding) can eliminate most limit cycles. However, truncation can cause biased quantisation error

Output Quantisation

- Linear modelling of product quantisation



modelled as

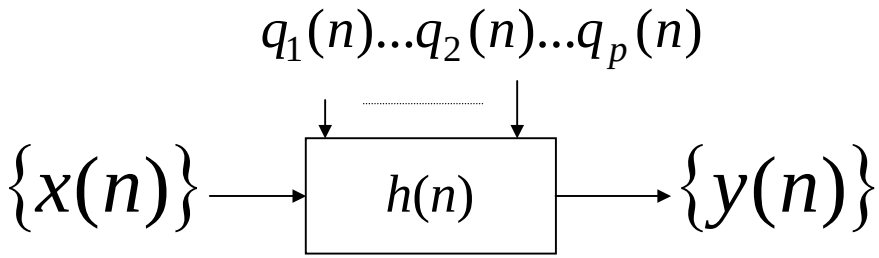


Now interested in output of multiplier

Output Quantisation

- Recall: for rounding operations, $q(n)$ is uniformly distributed between $-\frac{Q}{2}$ and $\frac{Q}{2}$, where Q is the quantisation step (i.e. in a word length of b bits with sign+magnitude representation mod 2, $Q = 2^{-b}$)
- A discrete-time system with quantisation at the output of each multiplier may be considered as a multi-input linear system

Output Quantisation



- Each $q_\lambda(n)$ contributes to output accuracy
- Then

$$y(n) = \sum_{r=0}^{\infty} x(r).h(n-r) + \sum_{\lambda=1}^{p}\left[\sum_{r=0}^{\infty} q_\lambda(r).h_\lambda(n-r)\right]$$

where $h_\lambda(n)$ is the impulse response of the system from λ th output of the multiplier to $y(n)$.

Output Quantisation

- Avoid output quantification error (clipping) by avoiding overflow in output caused by q
- For zero input (free running), i.e., $x(n) = 0, \forall n$:

$$|y(n)| \le \sum_{\lambda=1}^{p} |\hat{q}_\lambda| . \sum_{r=0}^{\infty} |h_\lambda(n-r)|$$

where $|\hat{q}_\lambda|$ is an upper bound for $|q_\lambda(r)|, \ \forall \lambda, r$

This bound is the maximum error $\frac{Q}{2}$

Hence $$|y(n)| \le \frac{Q}{2} . \sum_{\lambda=1}^{p} \left[\sum_{n=0}^{\infty} |h_\lambda(n-r)| \right]$$

Output Quantisation

- However, error does not exceed unit step input:

$$\sum_{n=0}^{\infty} |h_\lambda(n)| \le \sum_{n=0}^{\infty} |h(n)|$$

hence

$$\boxed{|y(n)| \le \frac{pQ}{2} . \sum_{n=0}^{\infty} |h(n)|} \quad \text{(conservative)}$$

Thus, we can estimate the maximum range at the output to avoid clipping from the system parameters $h(n)$ and quantisation level Q