EE185524

Quantization and Sampling Theory

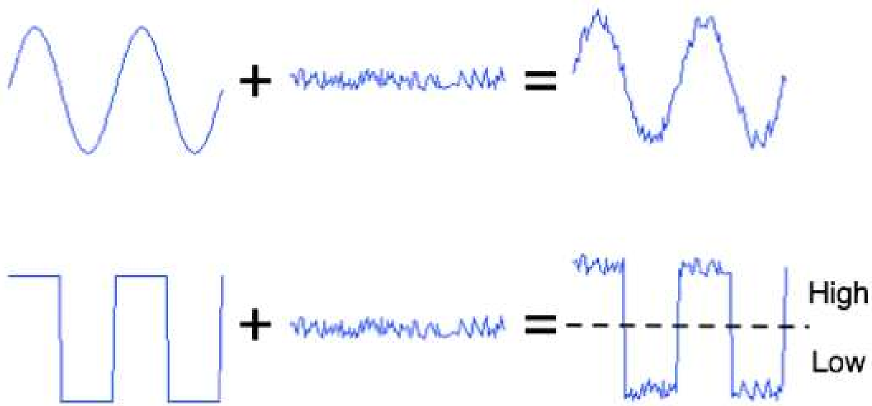
Yurid E. Nugraha
TSP DTE FTEIC ITS

Digital vs analog controller

- A digital control system uses digital electronics hardware, usually in the form of a programmed digital computer
- The evolution of microprocessors / embedded systems allowed their use as control elements:
  - met the stringent performance specifications needed in applications, and
  - have several advantages over their continuous-time counterparts:
- Sampling is thus inherent and may be necessary
- For some control system application, better system performance may be achieved by a digital control system design

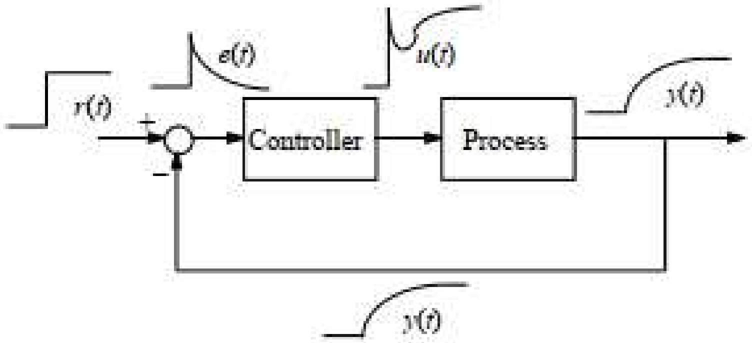
Digital controller

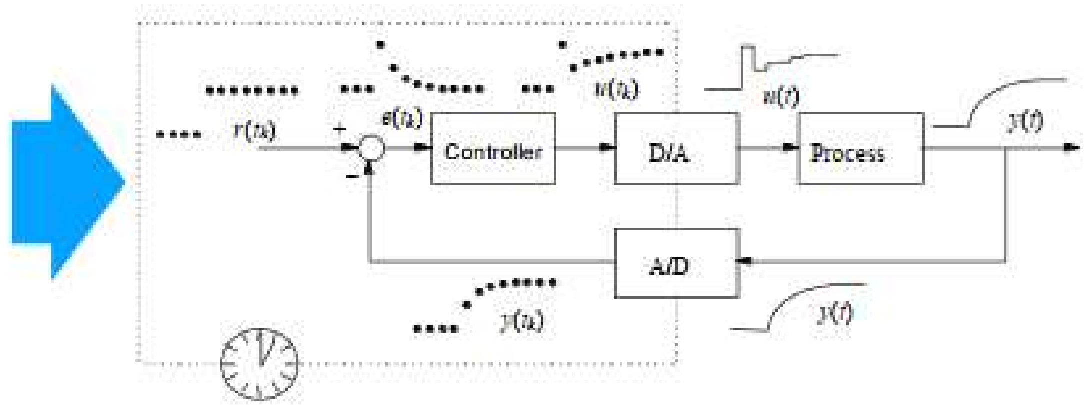
- More reliable due to its improved noise immunity



- Need sampling to convert time-varying signals to discrete-time signals
- Quantization?

Analog vs digital signals

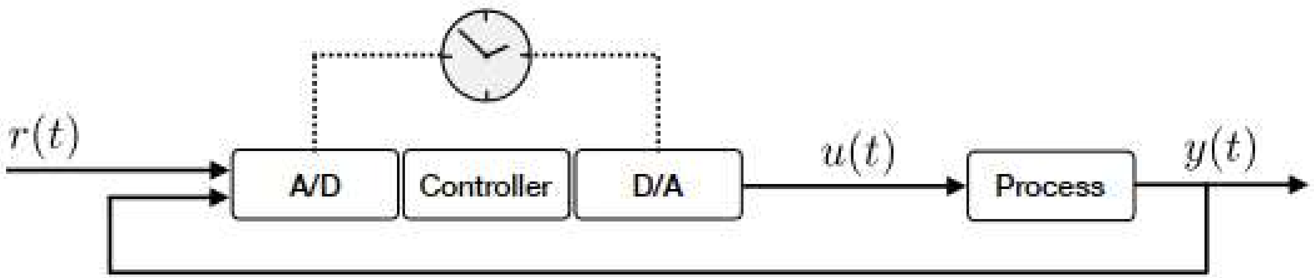




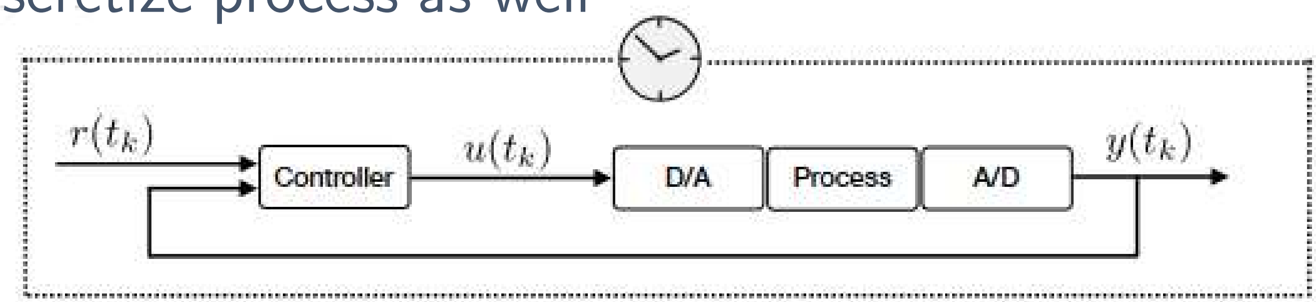
- Reconstruction of analog to digital signal **(and vice versa)** is only an approximation of the actual signal
- Some signal information might be lost or/and delayed in the process

Digital controller design

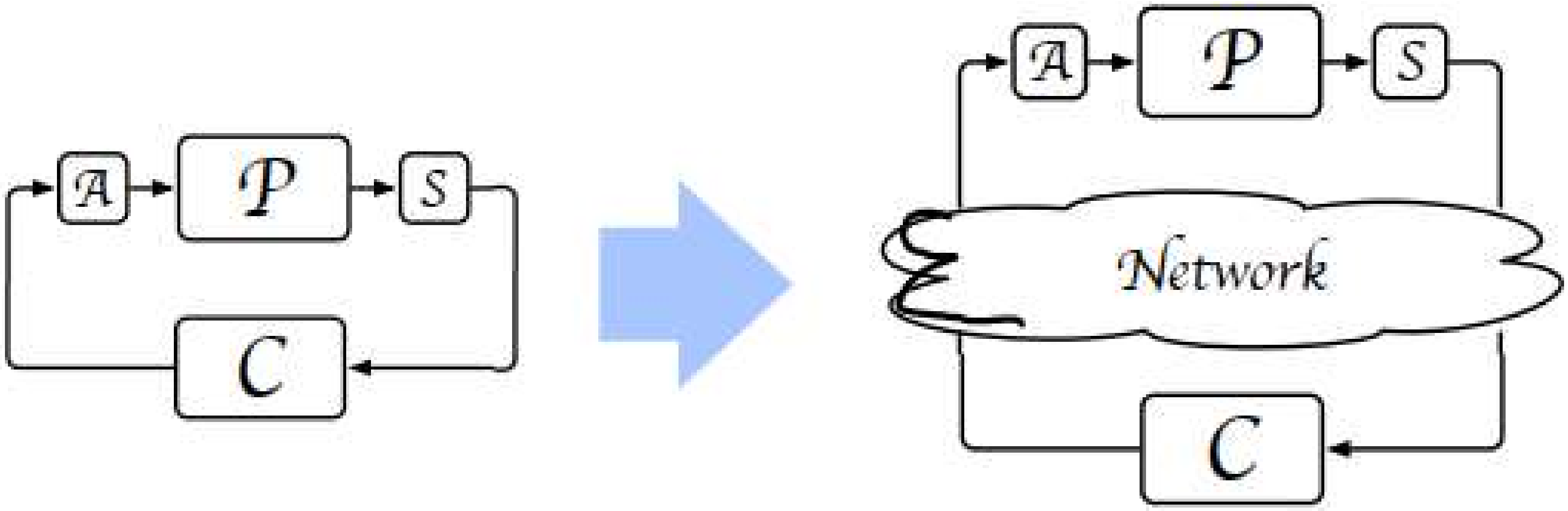
- Discretize analog controller (**more common**)



- Discretize process as well

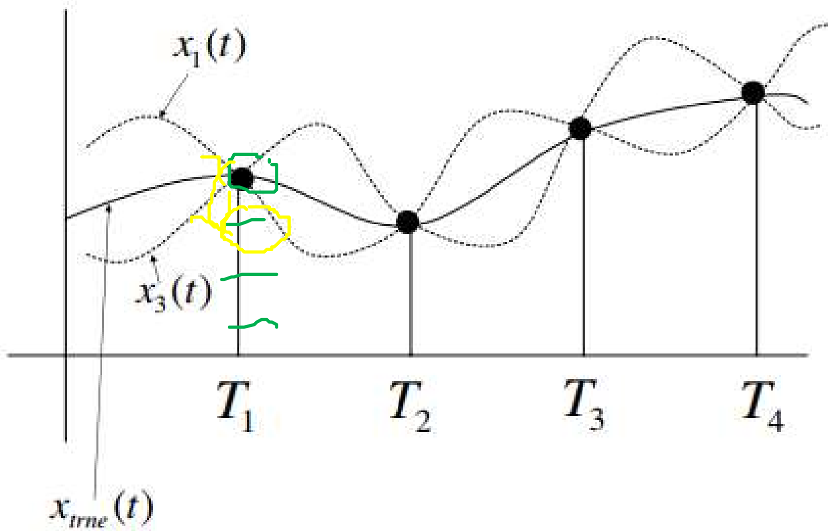


Digital control in networked control



Sampling

- (Usually) takes place in regular intervals, say T_s
- Sampling frequency $f_s = 1/T_s$
- Generally, cannot reconstruct signals fully from samples

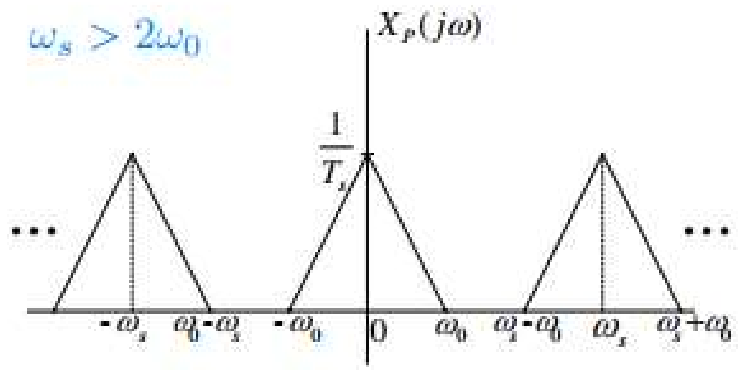


Sampling criterion

- $x_c(t)$ can be uniquely determined by its samples $x_c(nT_s)$ if the sampling angular frequency is at least twice as big as ω_0:

$$\omega_s = \frac{2\pi}{T_s} > 2\omega_0$$

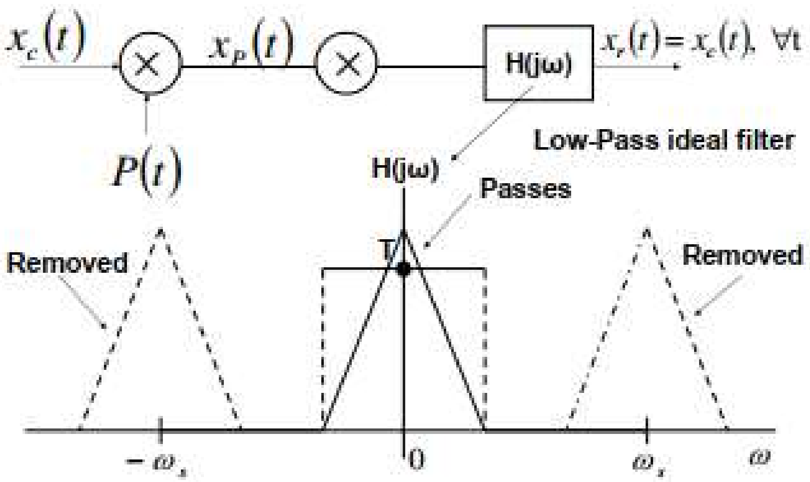
- Nyquist angular frequency: minimum sampling angular frequency



Reconstruction

- D/A conversion
- Requires a low-pass filter w/ cut-off frequency:

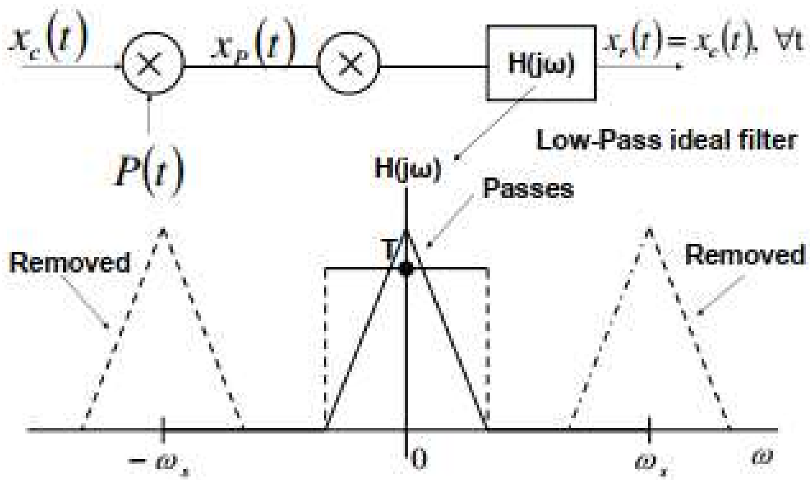
$$\omega_0 < \omega_c < \omega_s - \omega_0$$



Reconstruction

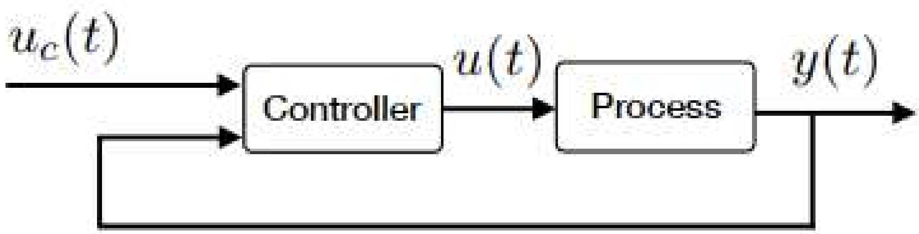
- D/A conversion
- Requires a low-pass filter w/ cut-off frequency:

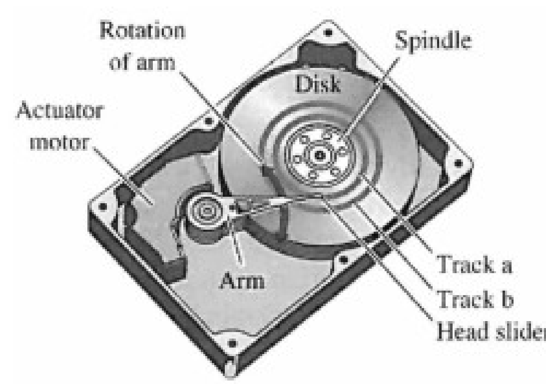
$$\omega_0 < \omega_c < \omega_s - \omega_0$$



Example

- Consider a plant with TF $G(s) = \frac{k}{Js^2}$

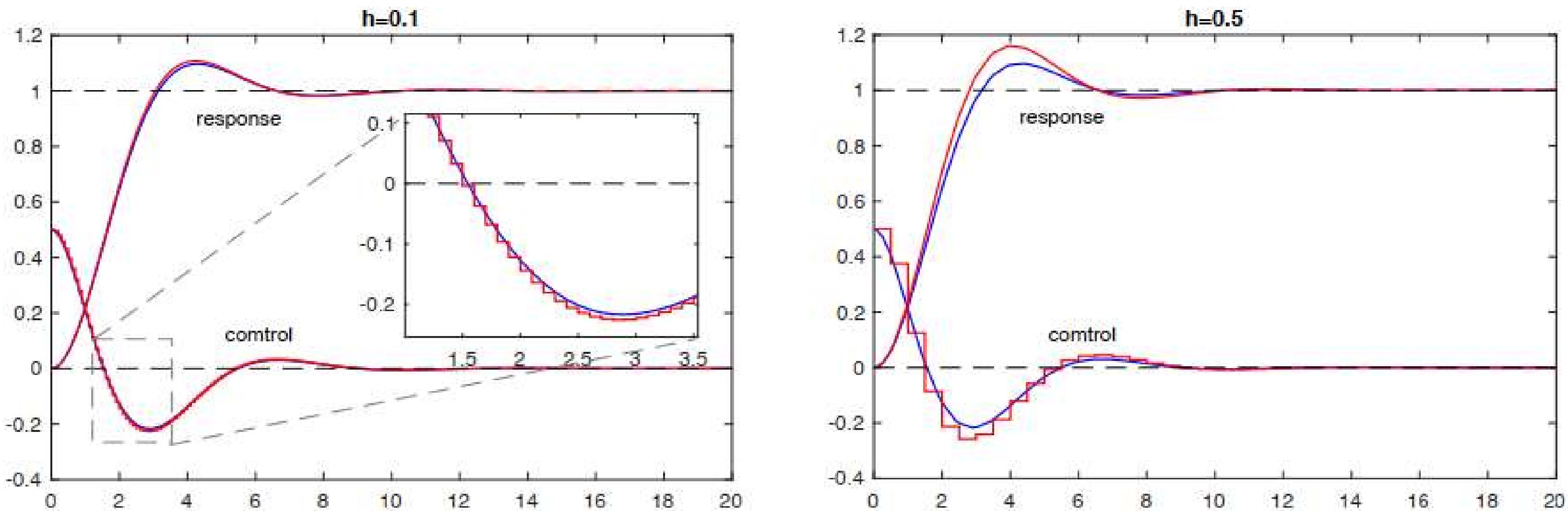




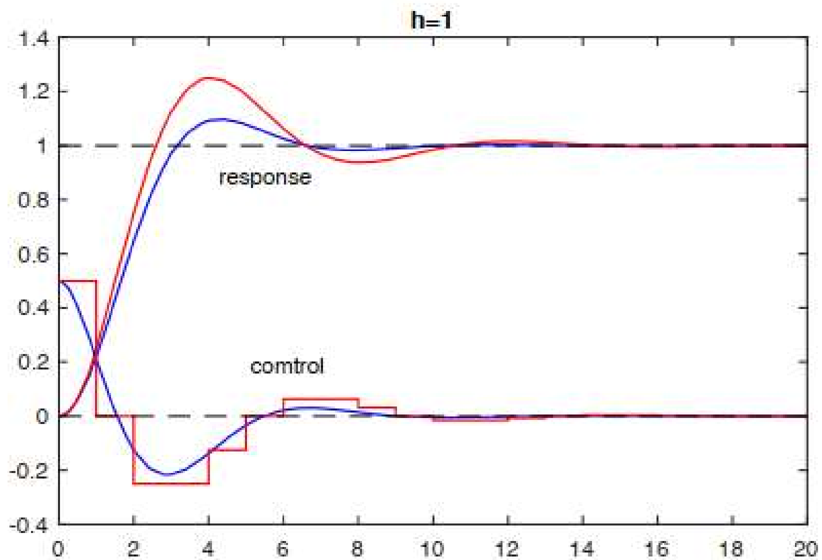
- Suppose controller $U(s) = K\frac{b}{a}U_c(s) - K\frac{(s+b)}{s+a}Y(s)$
- Derivative is approximated w/ a difference

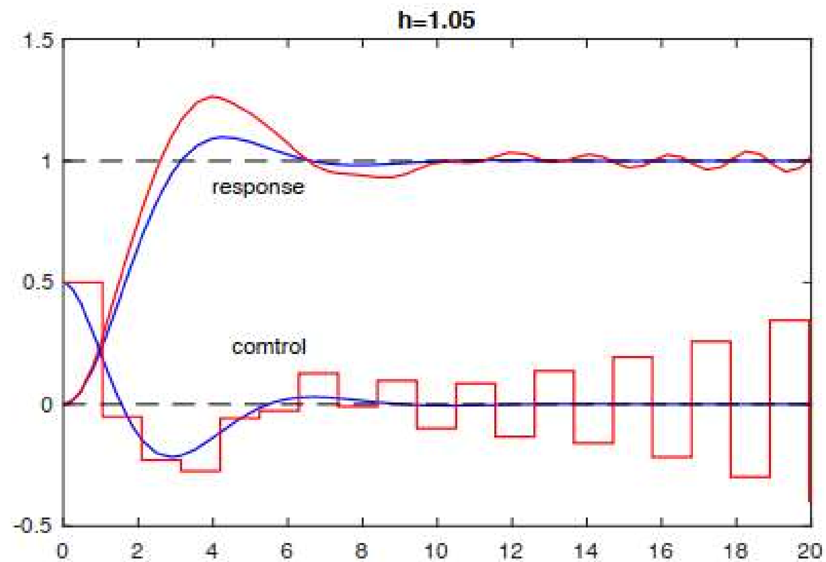
$$\frac{x(t+h)-x(t)}{h} = -ax(t) + (a-b)y(t)$$

Example



Example





Quantization

- The process of mapping input values from a large set (often continuous) to output values in a (countable) smaller set

- **Output**: fixed-point words (8-bit, 16-bit, and 24-bit)

  256 levels | 65536 levels | 16.8 million

- An A/D converter produces these binary representation of the **sampled signals** at each sample time

Fixed-point number representation

- A n-bit fixed-point binary number N

$$N = \sum_{j=-m}^{n-1} b_j 2^j = \underbrace{b_{n-1}2^{n-1} + \cdots + b_0 2^0}_{\text{Integer portion}} + \underbrace{b_{-1}2^{-1} + \cdots + b_{-m}2^{-m}}_{\text{Decimal portion}}$$

$$= (\underset{\substack{\downarrow \\ \text{MSB}}}{b_{n-1}} b_{n-2} \dots b_0 \underset{\substack{\downarrow \\ \text{Binary point}}}{\odot} b_{-1} \dots \underset{\substack{\downarrow \\ \text{LSB}}}{b_{-m}})_2, \qquad b_j \in (0,1)$$

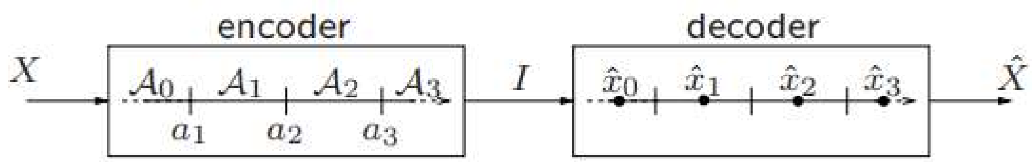
Quantization error

- Depends on the type of arithmetic and type of quantization used
$$-\frac{q}{2} \le e \le \frac{q}{2}$$
where $q \coloneqq 2^{-C}$, C being the number of bits
- More level → lower noise
- Typically, original signal is much larger than LSB

Quantizer

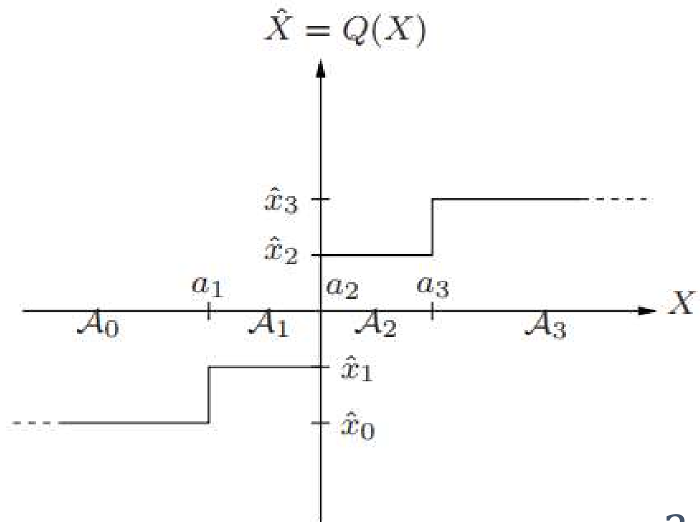
- Samples $\boldsymbol{X} = X_N^1$ into a k-bit index, then produces approximation $\hat{X}_1^N$

- **Example: Scalar quantizer**



- Encoder: $X \in A_i \to I$, Decoder: $I \to \hat{X} = \hat{x}_i$

Quantizer

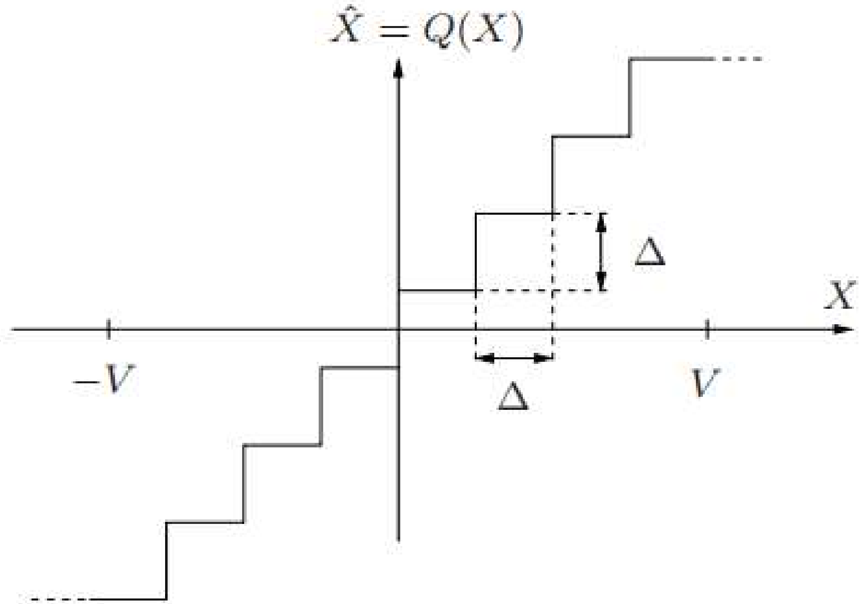


- MSE quantization distortion: $D = \mathbb{E}[(X - \hat{X})^2]$
- Signal-to-quantization-noise ratio: $SQNR = \frac{\mathbb{E}[X^2]}{\mathbb{E}[(X-\hat{X})^2]}$

Uniform quantization

- Step size $\Delta = \frac{2V}{2^k} = 2^{1-k}V$
- Quantization error $\tilde{X} = X - Q(X)$
  - $X \in [-V, V] \rightarrow$ "granular region"
  - $|X| > V \rightarrow$ "overload"
- Quantization noise:

$$D = \underbrace{\int_{-V}^{V} (x - Q(x))^2 f(x) dx}_{\text{Granular distortion}} + \underbrace{\int_{x=|V|} (x - Q(x))^2 f(x) dx}_{\text{Decimal portion}}$$



Quantization for nonuniform distribution

- Distortion is given by

$$D = \sum_{i=0}^{K-1} \int_{A_i} (x - x_i)^2 f(x) dx$$

$$= \sum_{i=0}^{K-1} \int_{a_i}^{a_{i+1}} (x - x_i)^2 f(x) dx$$

- Optimal encoding:

$$a_i = \frac{\hat{x}_{i-1} + \hat{x}_i}{2} \qquad \text{(OE)}$$

- Optimal decoding:

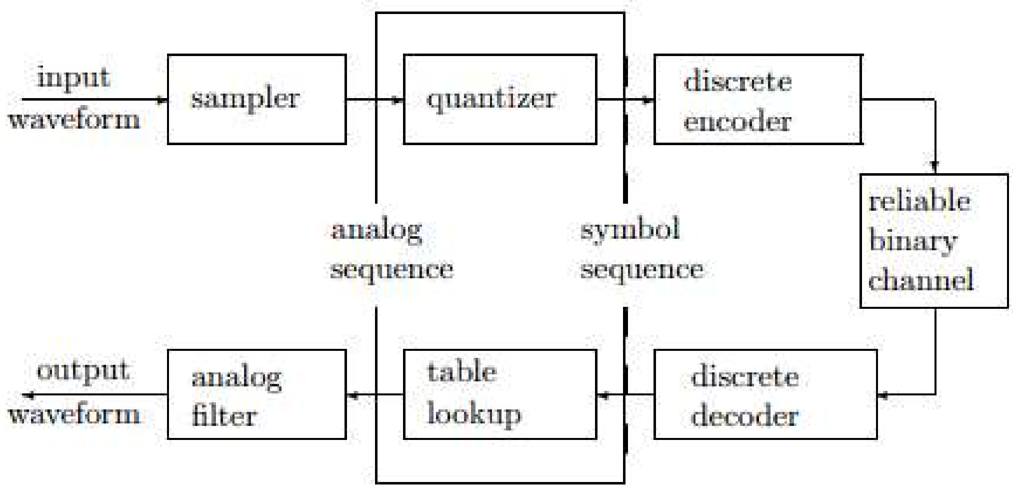
$$\hat{x}_i = \frac{\int_{a_i}^{a_{i+1}} x f(x) dx}{\int_{a_i}^{a_{i+1}} f(x) dx} \qquad \text{(OD)}$$

(assuming pdf known)

Lloyd-Max algorithm

- Iteratively computes quantization variables $\hat{x}_i$ and a_i
- Assumption: $f(x)$ known, $a_0 = -V$, $a_k = V$
- Steps:
  - Step 1: Assume a value for $\hat{x}_0$
  - Step 2: Find a_1 from (OE)
  - Step 3: Find $\hat{x}_2$ from (OD)
  - … etc.

Quantization in communication systems



Sustained oscillations and deadband effects

- When digital controllers are implemented with finite word length, **sustained oscillations** may appear at the controller output
- Consider the controller described by difference equation

$$y[k] = ay[k-1] + x[k]$$

where $a = 0.5,\ x[k] = 0.75\delta[k],\ y[-1] = 0$

$k = 0 \rightarrow 0.75$

$k = 1 \rightarrow 0.75 \times 0.5$

- *If* the controller equation implemented with *infinite word*, then

$$y[k] = 0.75(0.5)^k$$

100

Sustained oscillations and deadband effects

- *If* the controller equation implemented with *3-bit word*, then

$$y_q[k] = Q\left[0.5y_q[k-1]\right] + 0.75\delta[k]$$

- ….. stops at $y_q[k] = 0.125$

Interplay between sampling and quantization error

Quantization error may not be ignored

- For example, consider a controller w/ TF $G(s) = \frac{10^4}{s+1}$
- Discretization w/ **impulse invariant approximation**:

$$G(z) = \frac{10^4}{1 - e^{-h}z^{-1}}$$

- **Unit impulse response**: $g[m] = 10^4 \left(e^{-h}\right)^m, \quad m = 0,1,2,\ldots$
- Thus,

$$\sum_{m=0}^{\infty} g^2[m] = 10^8 \sum_{m=0}^{\infty} e^{-2hm} = 10^8\left(1 + e^{-2h} + \cdots\right) = \frac{10^8}{1 - e^{-2h}}$$

Interplay between sampling and quantization error

Quantization error may not be ignored

- Variance of output:

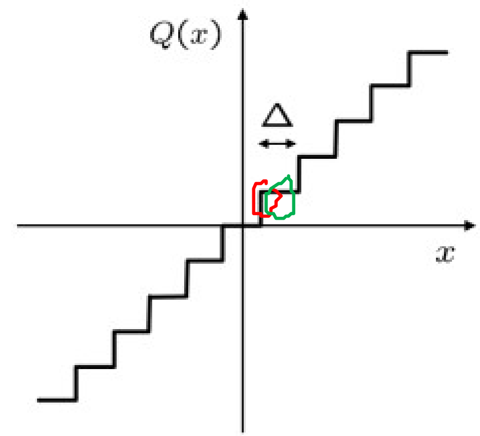
$$var(y_e) = var(e) \sum_{m=0}^{\infty} g^2[m] = \left(\frac{2^{-2C}}{12}\right)\left(\frac{10^8}{1-e^{-2}}\right)$$

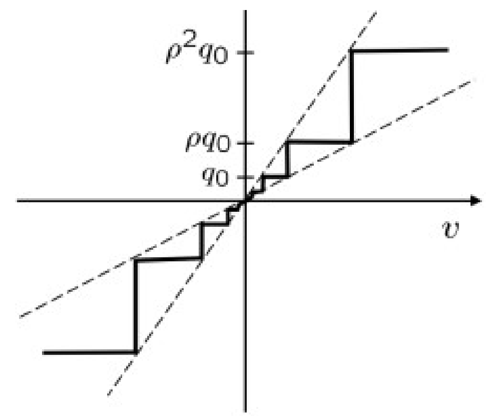
- If C fixed: decreasing h **increases** variance of output noise
- If h fixed: increasing C **decreases** variance of output noise

Which quantizers?

Since a finite bits are transmitted, quantizers should be **designed**

- **Uniform** quantizers: divide space into equal sections
- **Logarithmic** quantizers: provide a finer quantization near the origin
- **Dynamic scaling:** a smaller scaling factor provides a fine quantization near origin; a larger one ensures large number fall within domain of quantization





Into control system…

$$x[k+1] = Ax[k] + Bu[k] + w[k]$$
$$y[k] = Cx[k] + v[k]$$

noise

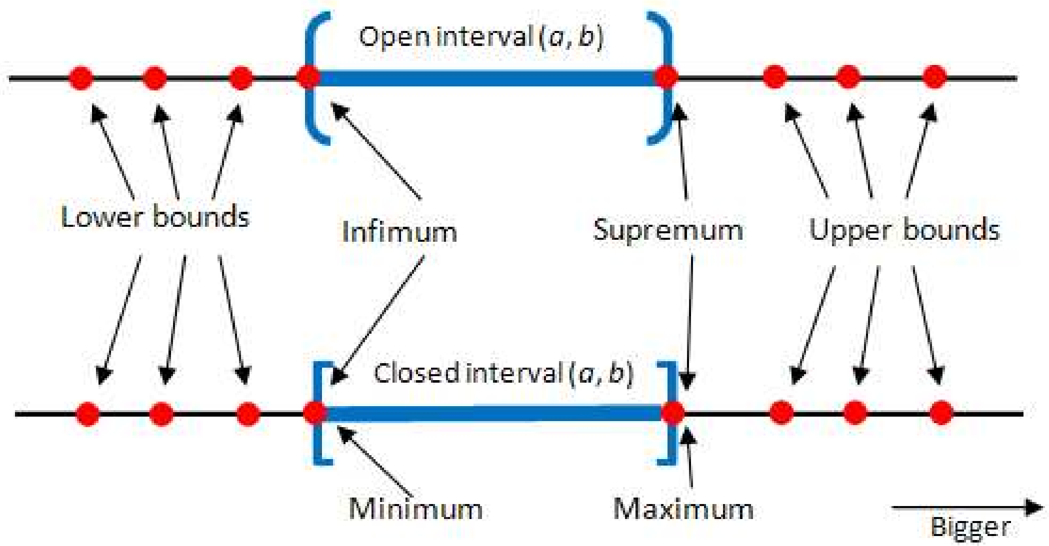
- **Objective**: Identify the trade-off between the unstable modes of the system and the channel's rate to guarantee stability

- **Solution**: Consider second moment stability:

$$\sup_{k \in \mathbb{N}} \mathbb{E}(||X_k||^2) < \infty$$

supremum

Into control system...



Into control system...

Remote state estimation:

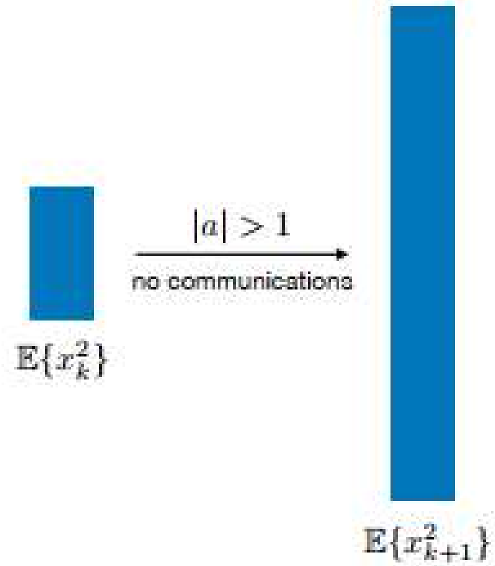
$$x[k+1] = ax[k] + w[k]$$

$w[k]$ Gaussian w/ zero mean and variance σ_w^2:

$$\mathbb{E}\{x_{k+1}^2\} = a^2\mathbb{E}\{x_k^2\} + \sigma_w^2$$

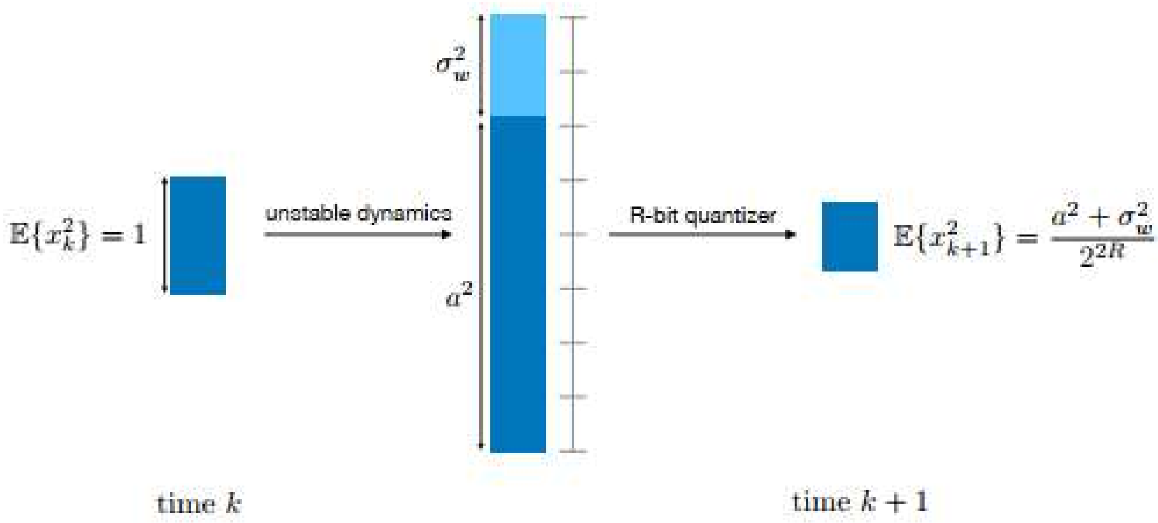
Unstable if $|a| > 1$

Data rate theorem



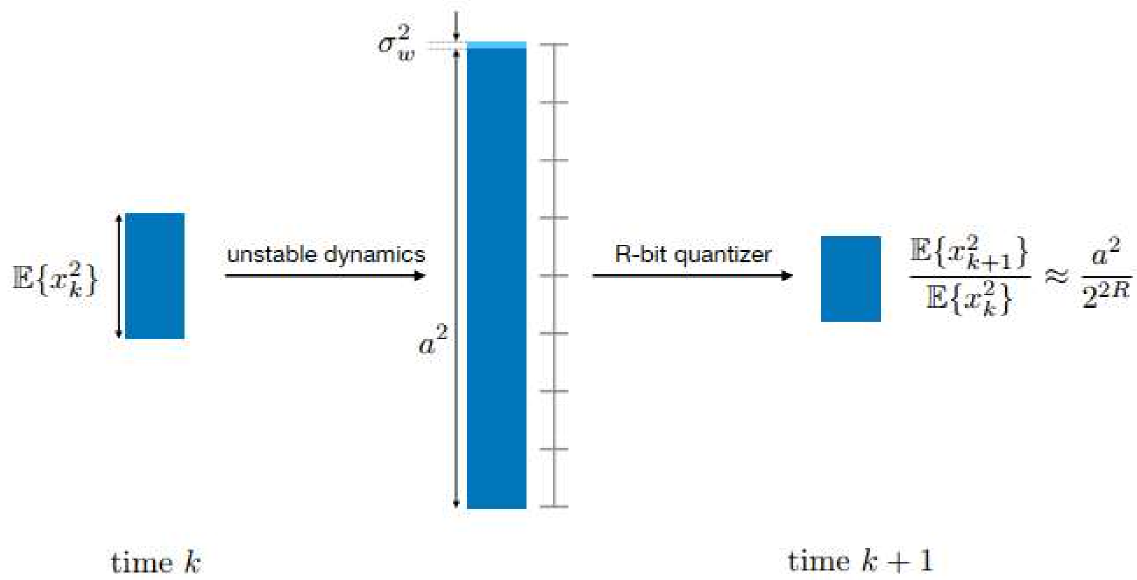
Data rate theorem

- The noise and bandwidth limitations of the channels are captured by modeling channels capable of transmitting only R bits in each time slot
- By transmitting enough bits at each time step, we can ensure the uncertainty decreases



Data rate theorem

- $\mathbb{E}\{x_k^2\}$ grows larger each time
- Thus, (second moment) stability can be achieved if
$$\frac{a^2}{2^{2R}} < 1$$



Conclusion

- Needs to carefully choose the sampling for stability and performance
- The finer the quantization, the better
- How much information is needed to be communicated by the quantizer in order to achieve a certain control objective?