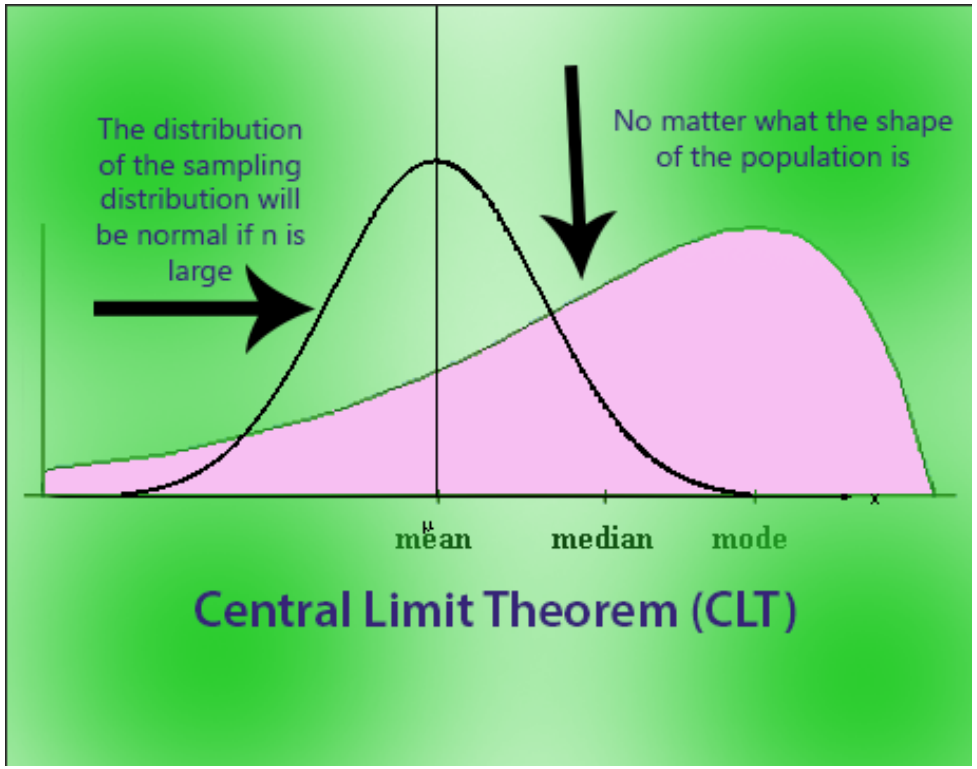
2.160 Identification, Estimation, and Learning

Part 3 Linear System Identification

Lecture 15

Asymptotic Distribution of Parameter Estimates



H. Harry Asada
Department of Mechanical Engineering
MIT

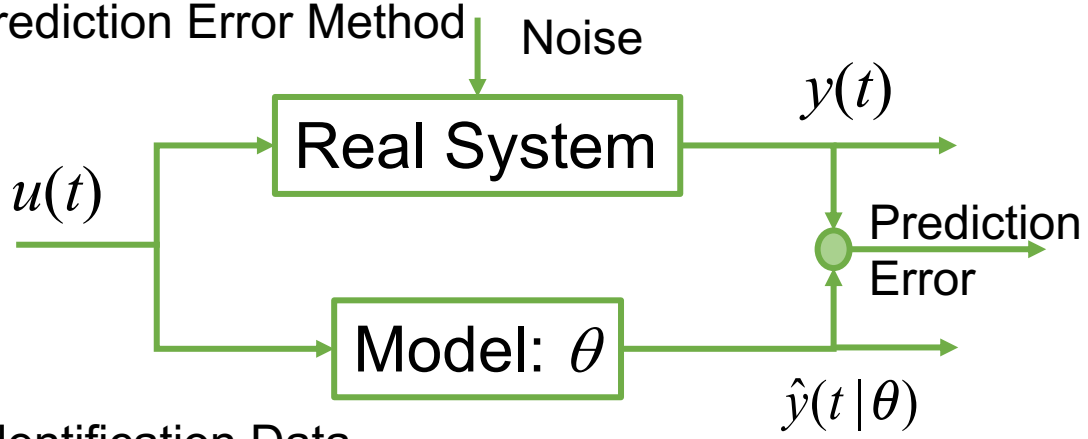
Estimating Model Parameters from Input-Output Data

- Parametric Model

$$\theta = (a_1, \cdots, a_{n_a}, b_1, \cdots, b_{n_b})^T$$

$$G(q;\theta) = \frac{b_1 q^{-1} + b_2 q^{-2} + \cdots + b_{n_b} q^{-n_b}}{1 + a_1 q^{-1} + a_2 q^{-2} + \cdots + a_{n_a} q^{-n_a}}$$

- Prediction Error Method



- Identification Data

$$D = \{(u(t), y(t)) \,|\, t = 1, 2, \cdots, N\}$$

- Apply Least Squares Estimate

$$\hat{\theta}_N = \arg\min_{\theta} \frac{1}{N} \sum (y(t) - \hat{y}(t|\theta))^2$$

Theoretical Questions

- Consistent Estimate

  Does the Least Squares Estimate $\hat{\theta}_N$ approaches the true parameter values as the number of data tends to infinity?

$$\hat{\theta}_N \underset{N\to\infty}{\longrightarrow} \theta_0 \text{ (true)}$$

  This is the same property as "Unbiased Estimate" under the ergodicity assumption.

$$E[\hat{\theta}_N] = \theta_0$$

  This requires 1) correct model structure, 2) correct prediction, and 3) persistently exciting data.

- Convergence Speed: How quickly does $\hat{\theta}_N$ converge?
  - How many data points are required?
  - What are major factors influencing the convergence speed?

Analysis of Asymptotic Distribution of Estimated Parameters

- The major results of the analysis:

1. Convergence rate $\hat{\theta}_N \to \theta^*$

   At a rate proportional to $\frac{1}{\sqrt{N}}$

2. Error distribution

   Converges to a Gaussian distribution

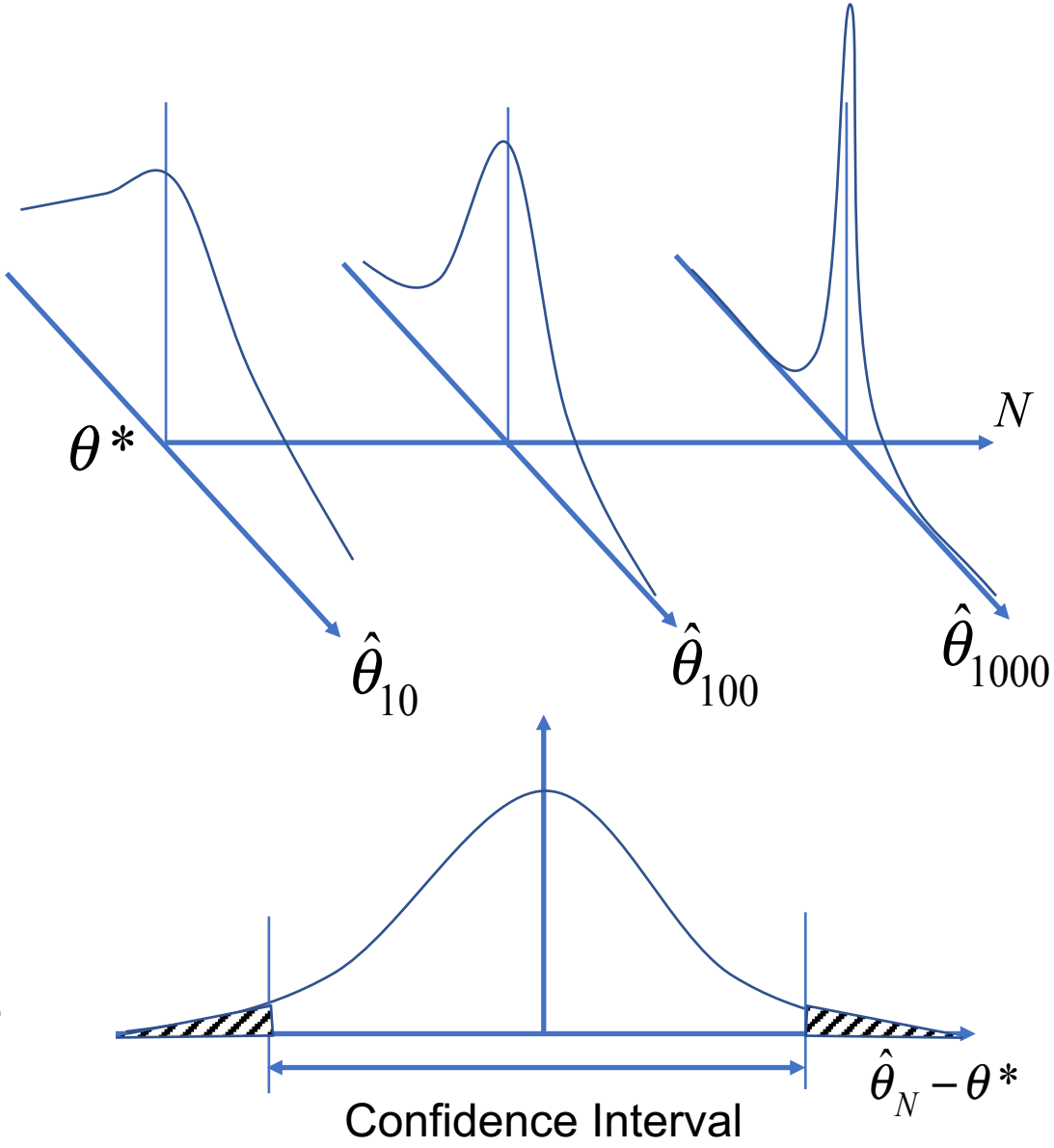
3. Covariance $Q = E[(\hat{\theta}_N - \theta^*)(\hat{\theta}_N - \theta^*)^T]$

   Depends on
   - *Noise intensity*
   - *Sensitivity of predictor to parameter* $\frac{\partial \hat{y}(t|\theta)}{\partial \theta}$

   Larger is better.

4. Confidence Interval

   Quality of parametric system identification



Approach: Central Limit Theorem in Statistics

- We aim to analyze the distribution of estimated parameters by applying the Central Limit Theorem (CLT).
- CLT tells us that the mean of samples of size N drawn from a population with an arbitrary distribution tends towards a normal distribution, as the sample size tends to infinity.
- Let $X_t \in \Re^\ell, t = 1,2,\cdots,$ be multivariate random variables with mean

$$E[X_t] = m \qquad \text{for all } t.$$

- The covariance is given by

$$C_X = E[(X_t - m)(X_t - m)^T] \qquad \text{for all } t.$$

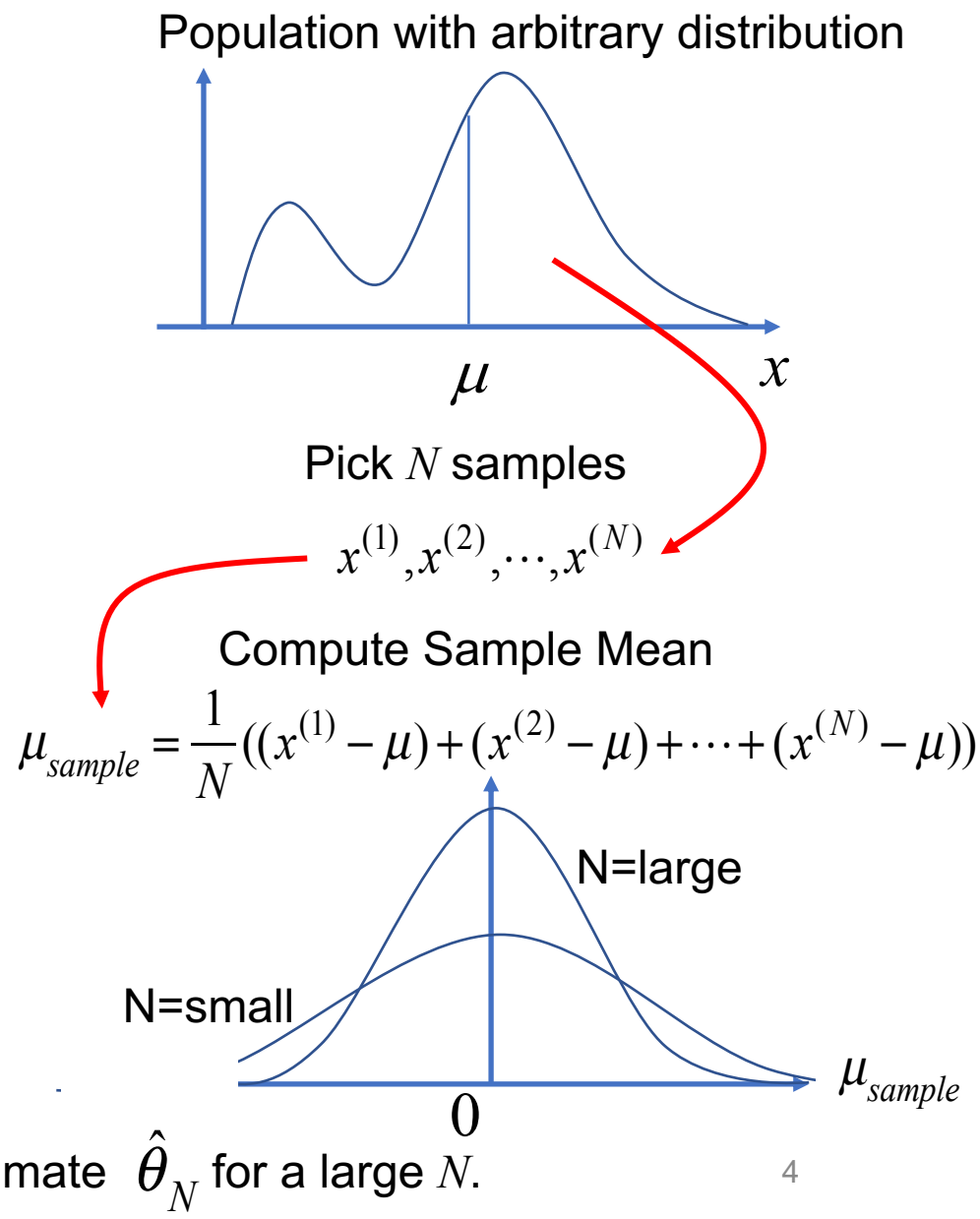
- Consider the sum of $X_t - m$

$$Y_N = \frac{1}{\sqrt{N}} \sum_{t=1}^{N} (X_t - m) = \sqrt{N} \frac{1}{N} \sum_{t=1}^{N} (X_t - m) = \sqrt{N} \times (\text{Sample Mean})$$

- As N tends to infinity, the distribution of Y_N converges to a Gaussian distribution

$$Y = \lim_{N \to \infty} \frac{1}{\sqrt{N}} \sum_{t=1}^{N} (X_t - m) \sim \mathbb{N}(0, C_X)$$

- We will apply this theorem to find the distribution of parameter estimate $\hat{\theta}_N$ for a large N.



Assumptions

1. The predicted output is given as a linear regression:

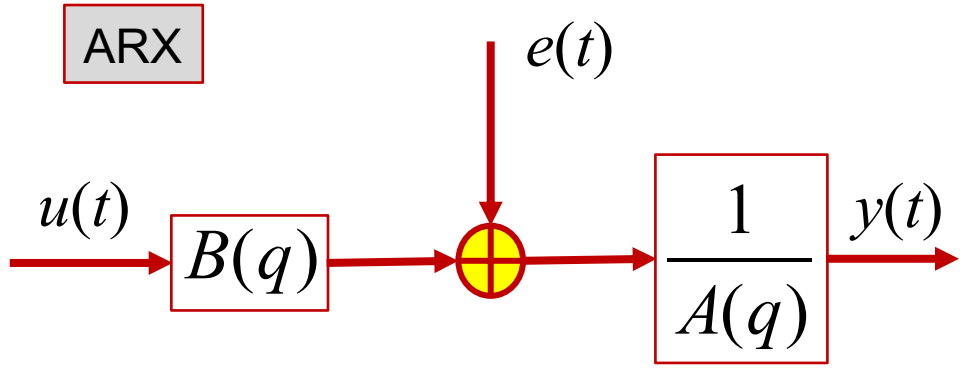
$$\hat{y}(t\,|\,\theta) = \theta^T \varphi(t)$$ - This condition will be relaxed later.

2. The assumed model structure is correct.

$$y(t) = \theta_0^T \varphi(t) + e_0(t)$$

where θ_0 is the true parameter values, and

$$E[e_0(t)e_0(s)] = \begin{cases} \lambda; & t = s \\ 0; & t \neq s \end{cases}$$ white



Auto-Regressive model with eXogenous input

Actual data are generated by the correct model with the true parameter values.

3. Ergodicity

$$E[X_t] = \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} X_t$$ (ensemble mean) = (time average)

4. Persistent excitation

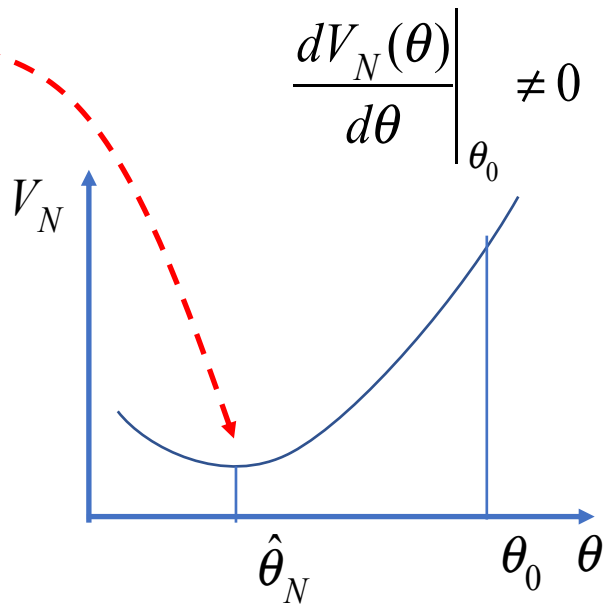
Asymptotic Distribution Analysis

- Under the assumptions described above, input-output data are collected and the Least Square Estimate of parameters is computed.
- Where is the distribution of the parameter estimate converging, as we collect a large number of data, $N \gg 1$? We aim to analyze the distribution of estimated parameters using CLT.

Step 1 If $\hat{\theta}_N$ is the solution to the least squares estimate for a given data set of size N, it minimizes the squared error:

$$\left.\frac{dV_N(\theta)}{d\theta}\right|_{\theta=\hat{\theta}_N} = 0$$

$$V_N = \frac{1}{N}\sum (y(t) - \hat{y}(t\,|\,\theta))^2$$



For a finite N, however,

$$\left.\frac{dV_N(\theta)}{d\theta}\right|_{\theta=\theta_0} \neq 0$$

Asymptotic Distribution Analysis

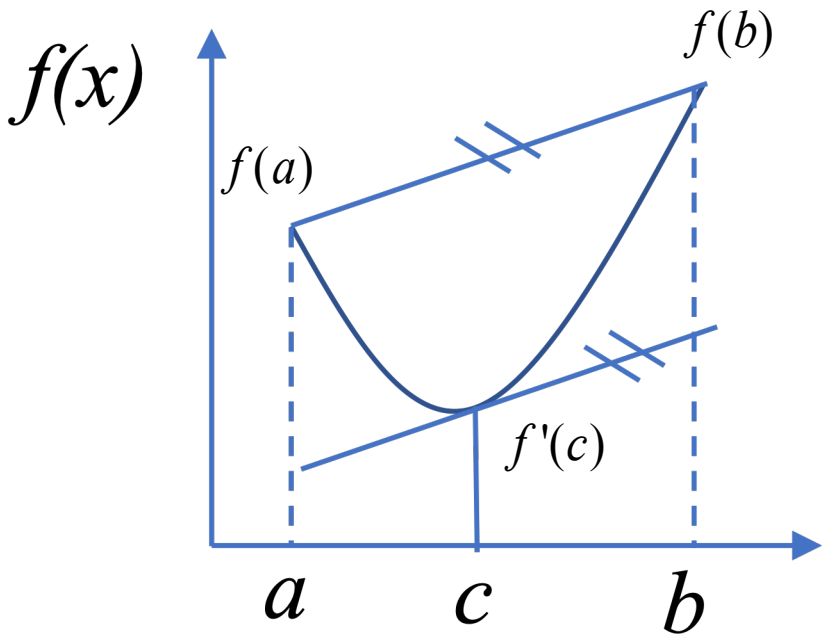
Recall the Mean Value Theorem

Let $f(x)$ be a continuous, differentiable function in $[a,b]$. There exists at least one point at $x = c$, where

$$\frac{f(b)-f(a)}{b-a} = f'(c), \quad a < c < b$$

or $$f(b)-f(a) = f'(c)(b-a)$$

Mean Value Theorem



Asymptotic Distribution Analysis

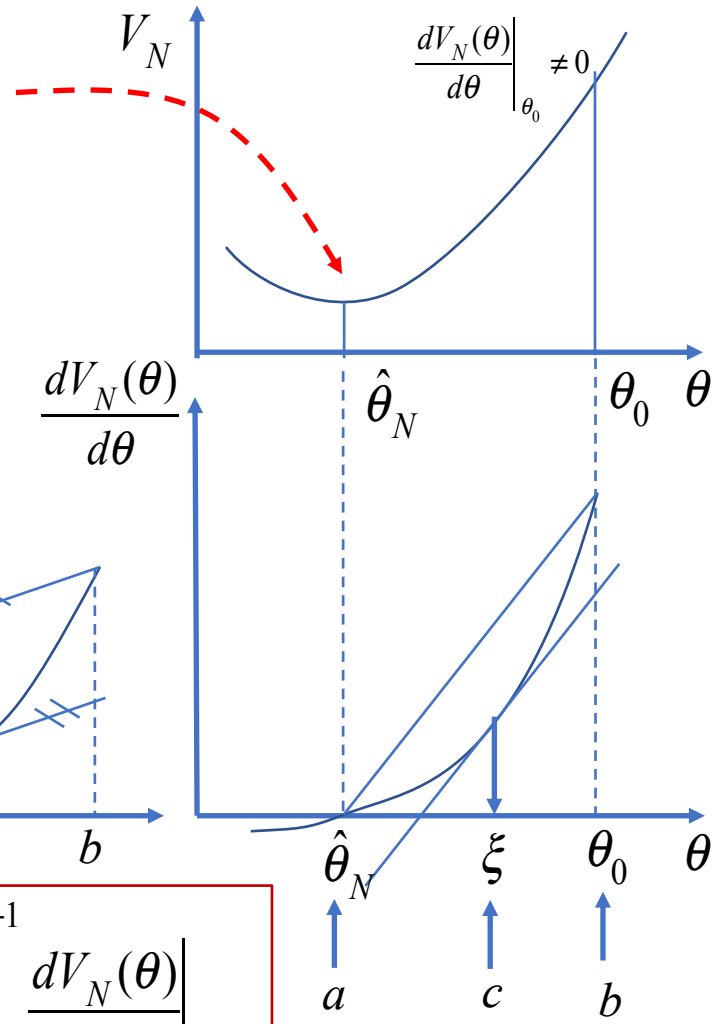
- Under the assumptions described above, input-output data are collected and the Least Square Estimate of parameters is computed. As we collect a large number of data, $N \gg 1$, where is the distribution of the parameter estimate converging? We aim to analyze the distribution of estimated parameters using CLT.

Step 1 If $\hat{\theta}_N$ is the solution to the least squares estimate for a given data set of size N, it minimizes the squared error:

$$\left.\frac{dV_N(\theta)}{d\theta}\right|_{\theta=\hat{\theta}_N} = 0$$

For a finite N, however,

$$\left.\frac{dV_N(\theta)}{d\theta}\right|_{\theta=\theta_0} \neq 0$$



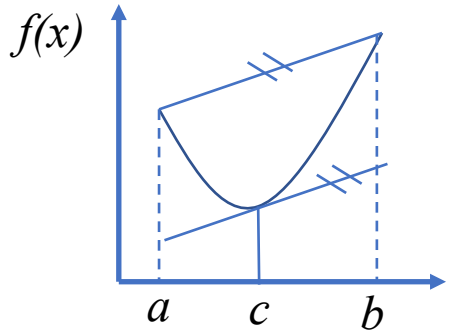
Recall the Mean Value Theorem

Let $f(x)$ be a continuous, differentiable function in $[a,b]$. There exists at least one point at $x = c$, where

$$\frac{f(b)-f(a)}{b-a} = f'(c), \quad a < c < b$$

or $f(b)-f(a) = f'(c)(b-a)$ Applying this to $\frac{dV_N(\theta)}{d\theta}$

Mean Value Theorem



$$\left.\frac{dV_N(\theta)}{d\theta}\right|_{\theta=\theta_0} - \cancel{\left.\frac{dV_N(\theta)}{d\theta}\right|_{\theta=\hat{\theta}_N}} = \left.\frac{d^2V_N(\theta)}{d\theta^2}\right|_{\xi}(\theta_0-\hat{\theta}_N)$$

0

$$\therefore (\hat{\theta}_N-\theta_0) = -\left[\left.\frac{d^2V_N(\theta)}{d\theta^2}\right|_{\xi}\right]^{-1}\left.\frac{dV_N(\theta)}{d\theta}\right|_{\theta=\theta_0} \quad \text{(A)}$$

Step 1. Continued

From the previous page

$$\therefore (\hat{\theta}_N - \theta_0) = -\left[\left.\frac{d^2 V_N(\theta)}{d\theta^2}\right|_{\xi}\right]^{-1} \left.\frac{dV_N(\theta)}{d\theta}\right|_{\theta=\theta}$$

- Compute $\left.\frac{dV_N(\theta)}{d\theta}\right|_{\theta_0}$

$$\left.\frac{dV_N(\theta)}{d\theta}\right|_{\theta_0} = \left.\frac{d}{d\theta}\frac{1}{2N}\sum(y(t)-\hat{y}(t|\theta))^2\right|_{\theta_0} = -\left.\frac{1}{N}\sum(y(t)-\hat{y}(t|\theta))\frac{d}{d\theta}\hat{y}(t|\theta)\right|_{\theta_0} \quad \leftarrow \quad \hat{y}(t|\theta) = \theta^T\varphi(t)$$

$$= -\left.\frac{1}{N}\sum(y(t)-\theta^T\varphi(t))\varphi(t)\right|_{\theta_0} \quad \leftarrow \quad y(t) = \theta_0{}^T\varphi(t) + e_0(t)$$

$$\therefore \left.\frac{dV_N(\theta)}{d\theta}\right|_{\theta_0} = -\frac{1}{N}\sum_{t=1}^{N} e_0(t)\varphi(t) \qquad \text{(B)}$$

- Treat $e_0(t)\varphi(t)$ as X_t in CLT:

$$Y_N = \frac{1}{\sqrt{N}}\sum_{t=1}^{N}(X_t - m)$$

X_t ← $e_0(t)\varphi(t)$

Step 2

Evaluate the mean and covariance of random variable $X_t = e_0(t)\varphi(t)$

- Mean $E[X_t] = E[e_0(t)\varphi(t)] = 0, \quad \forall t$

  Recall $\varphi(t) = \left(-y(t-1),\cdots,-y(t-n_a),u(t-1),\cdots,u(t-n_b)\right)^T \Leftrightarrow e_0(t)$ Uncorrelated

- Covariance $C_X(t,s) = E[(X_t - m)(X_s - m)^T] = E[\varphi(t)e_0(t)\varphi^T(s)e_0(s)]$

  - If $t = s$ $C_X(t,t) = E[\varphi(t)\varphi^T(t)e_0(t)e_0(t)]$

    $= E[\varphi(t)\varphi^T(t)] \cdot E[(e_0(t))^2] = \bar{\boldsymbol{R}} \cdot \lambda$

    Define $\bar{\boldsymbol{R}} \triangleq E[\varphi(t)\varphi^T(t)]$ and $\lambda \triangleq E[(e_0(t))^2]$

  - If $t > s$ $e_0(t) \Leftrightarrow \varphi(t)\varphi^T(s)e_0(s)$ Uncorrelated, independent

    $C_X(t,s) = E[e_0(t)]E[\varphi(t)\varphi^T(s)e_0(s)] = 0$

  - Similarly, if $t < s$ $C_X(t,s) = 0$

$$C_X(t,s) = \begin{cases} \bar{\boldsymbol{R}} \cdot \lambda; & t = s \\ 0; & t \neq s \end{cases}$$

(B')

- Both mean and covariance of $X_t = e_0(t)\varphi(t)$ are uniform over time t.
- Therefore, we can apply CLT to the summation: $Y_N = \frac{1}{\sqrt{N}}\sum_{t=1}^{N}(\varphi(t)e_0(t))$

Step 3

From Step 1, the parameter estimation error is given by $(\hat{\theta}_N - \theta_0) = -\left[\left.\frac{d^2 V_N(\theta)}{d\theta^2}\right|_{\xi}\right]^{-1} \left.\frac{dV_N(\theta)}{d\theta}\right|_{\theta=\theta_0}$ (A)

❑ We compute the second derivative matrix:

$$\left.\frac{d^2 V_N(\theta)}{d\theta^2}\right|_{\theta=\xi} = \frac{d}{d\theta}\left.\frac{dV_N(\theta)}{d\theta}\right|_{\theta=\xi} = -\frac{d}{d\theta}\frac{1}{N}\sum_{t=1}^{N}(y(t)-\theta^T\varphi(t))\varphi(t)\Big|_{\theta=\xi}$$

$$= \frac{1}{N}\sum_{t=1}^{N}\varphi(t)\varphi^T(t)\Big|_{\theta=\xi} \underset{N\to\infty}{\to} \bar{\boldsymbol{R}} \qquad \text{(C)}$$

Note that the ergodicity assumption was used in the last line. $E[\varphi\varphi^T] = \lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}\varphi(t)\varphi^T(t)$

❑ Putting together (A), (B), and (C)

From (A) $\sqrt{N}(\hat{\theta}_N - \theta_0) = -\left[\left.\frac{d^2 V_N(\theta)}{d\theta^2}\right|_{\xi}\right]^{-1} \left.\sqrt{N}\frac{dV_N(\theta)}{d\theta}\right|_{\theta=\theta_0} = \bar{\boldsymbol{R}}^{-1}\frac{1}{\sqrt{N}}\sum_{t=1}^{N}\varphi(t)e_0(t)$

(C) (B)

$Y_N = \frac{1}{\sqrt{N}}\sum_{t=1}^{N}(\varphi(t)e_0(t)) \underset{N\to\infty}{\to} \sim \mathbb{N}(0, C_X)$

Applying CLT, this term converges to Gaussian

Step 3 Continued

$$\sqrt{N}(\hat{\theta}_N - \theta_0) = \bar{\boldsymbol{R}}^{-1} Y_N, \quad Y_N \sim \mathbb{N}(0, \bar{\boldsymbol{R}}\lambda) \quad \text{From (B')} \quad C_X = \bar{\boldsymbol{R}}\lambda$$

- The covariance of Y_N is $\bar{\boldsymbol{R}}\lambda$, but what we want to know is the covariance of $\bar{\boldsymbol{R}}^{-1} Y_N$.

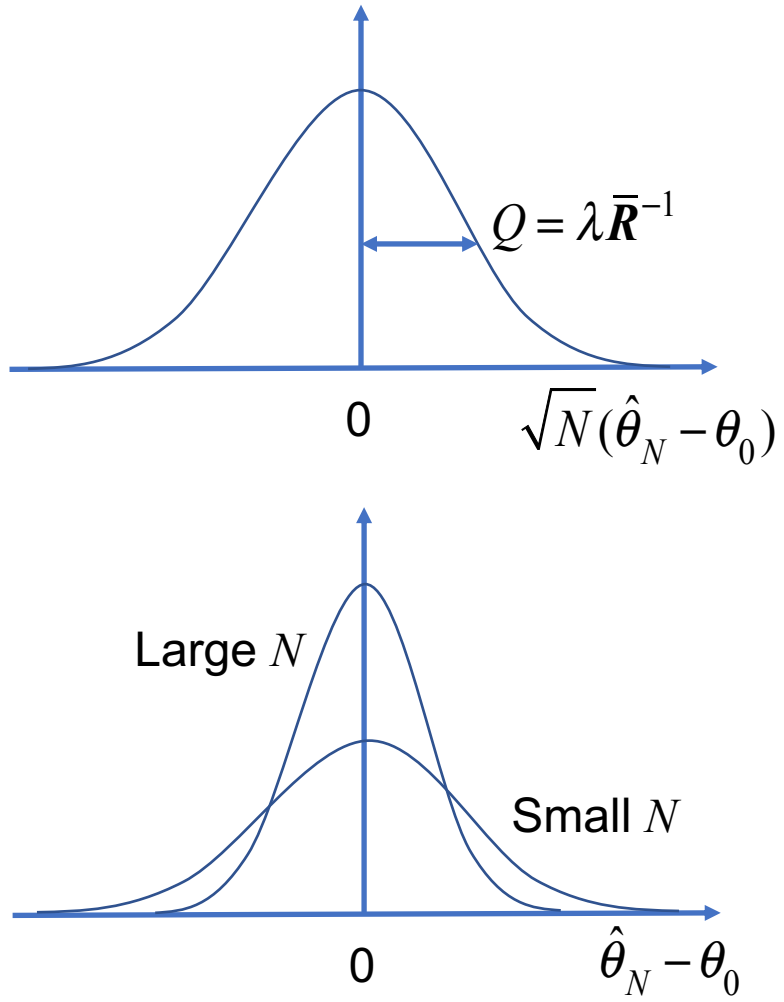
- We apply the following transformation rule:
  - Suppose $y = Ax$ and $C_X = E[xx^T]$
  - Then,
    $$C_Y = E[yy^T] = E[Axx^T A^T] = AE[xx^T]A^T = AC_X A^T$$
  - Replacing $x \leftrightarrow Y_N, \quad C_X \leftrightarrow \bar{\boldsymbol{R}}\lambda, \quad A \leftrightarrow \bar{\boldsymbol{R}}^{-1}, \quad y \leftrightarrow \bar{\boldsymbol{R}}^{-1} Y_N$
  - Covariance of $\bar{\boldsymbol{R}}^{-1} Y_N$ is $Q = \bar{\boldsymbol{R}}^{-1}\bar{\boldsymbol{R}}\lambda\bar{\boldsymbol{R}}^{-1} = \lambda\bar{\boldsymbol{R}}^{-1}$

    Note $\bar{\boldsymbol{R}}$ is symmetric.

- In summary, the distribution of parameter estimate converges to a Gaussian distribution with variance

$$\therefore \sqrt{N}(\hat{\theta}_N - \theta_0) \sim \mathbb{N}(0, \lambda\bar{\boldsymbol{R}}^{-1})$$

- The error $\left|\hat{\theta}_N - \theta_0\right|$ decreases at a rate of $\frac{1}{\sqrt{N}}$.



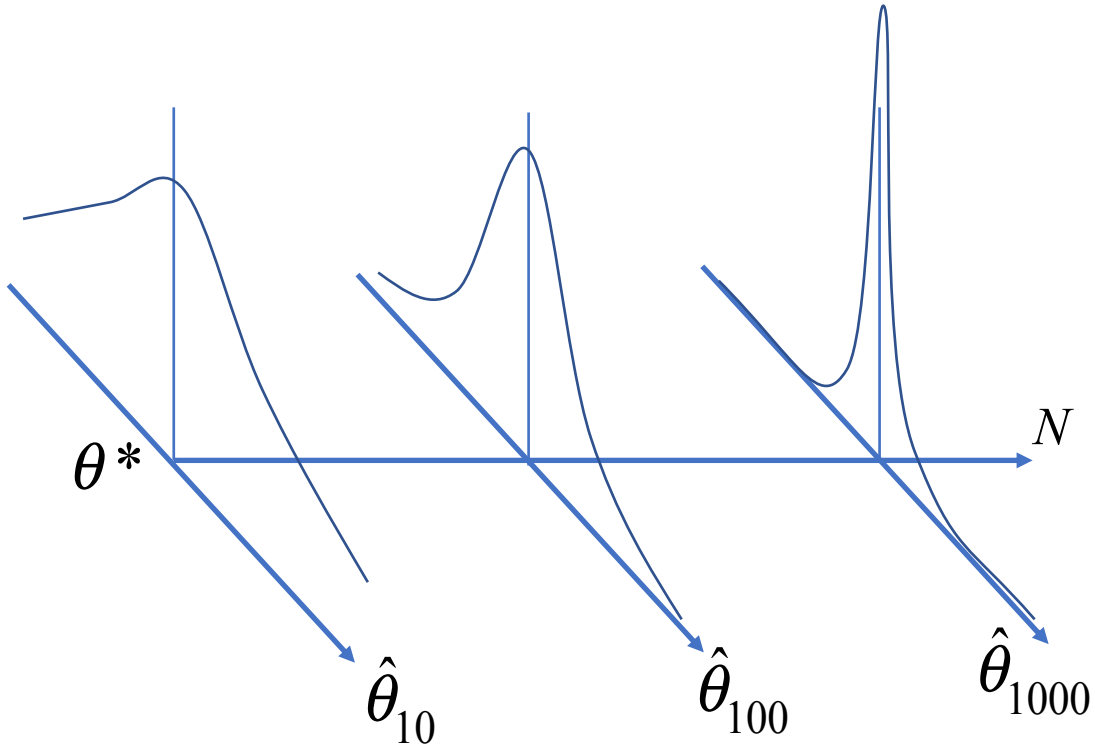
Analysis of Asymptotic Distribution of Estimated Parameters

$$\therefore \sqrt{N}(\hat{\theta}_N - \theta_0) \sim \mathbb{N}(0, \lambda \bar{\boldsymbol{R}}^{-1})$$

- The error $\left|\hat{\theta}_N - \theta_0\right|$ decreases at a rate of $\frac{1}{\sqrt{N}}$.

Question?

A UROP student took data (N = 10) to identify a system; the result had some significant standard deviation. You want to reduce it to 1/10 of the original result.
How many data does the UROP have to take?



Discussion

Covariance $Q = \lambda \bar{\boldsymbol{R}}^{-1}$ depends on

- Noise strength: $\lambda = E[(e(t))^2]$
- Input (regressor) signal strength:

$$\bar{\boldsymbol{R}} = E[\varphi(t)\varphi^T(t)]$$

$$\varphi(t) = [-y(t-1), \cdots, u(t-1), \cdots]^T$$

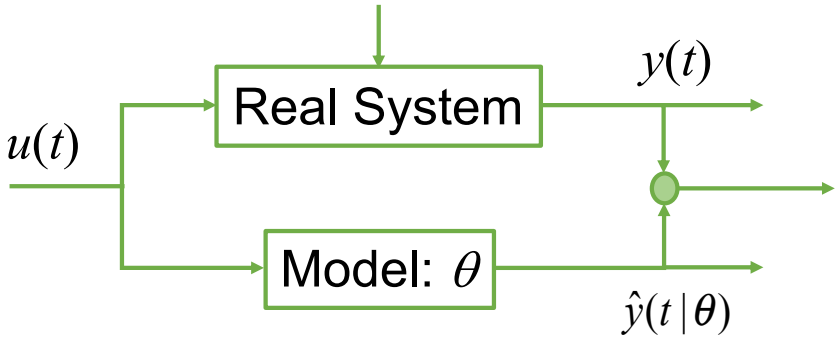
- Recall $\hat{y}(t \mid \theta) = \theta^T \varphi(t)$

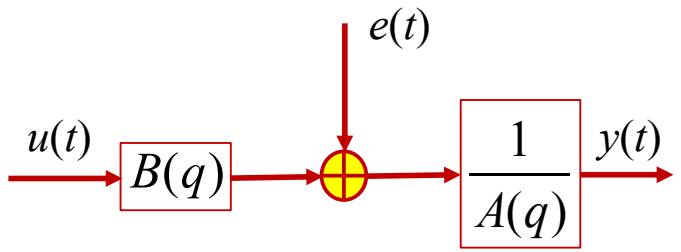
Therefore,

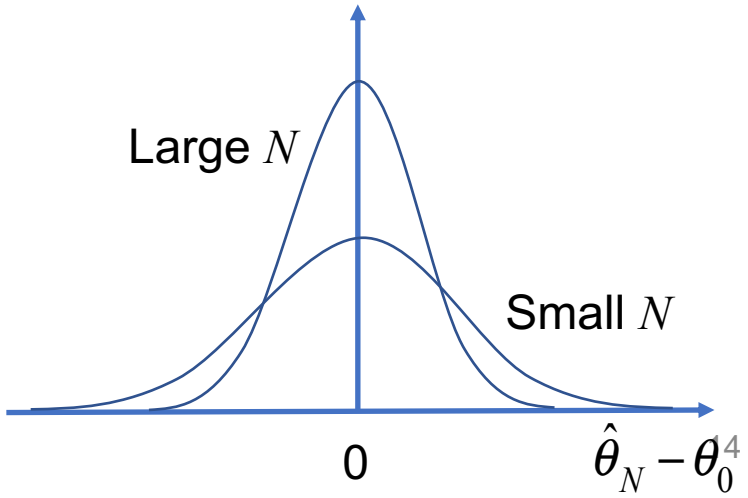
$$\varphi(t) \Rightarrow \frac{d\hat{y}(t \mid \theta)}{d\theta}$$

This implies that the regressor pertains to the sensitivity of the predicted output to parameters

- The higher the sensitivity, the faster the convergence. If an unnecessary parameter is involved in the model, it may have a low sensitivity, which slows down convergence.







Example

Notes on the Computation of Asymptotic Variance from a Model

Consider an ARX model:

$$y(t) = \frac{B(q)}{A(q)}u(t) + \frac{1}{A(q)}e(t) \qquad (1)$$

where $A(q) = 1 + aq^{-1}$ and $B(q) = bq^{-1}$, and $e(t)$ is white noise with variance $E[e(t)^2] = \lambda$. Let $u(t)$ be a white random sequence uncorrelated with $e(t)$. Also, assume E[u(t)2]=μ.

We want to identify the system with two unknown parameters $\theta = [a \quad b]^T$ based on Prediction Error Method. The predictor of this ARX system is given by a linear regression:

$$\hat{y}(t \mid \theta) = \varphi(t)^T \theta \qquad (2)$$

with regressor $\varphi(t) = [-y(t-1), u(t-1)]^T$. Before conducting experiments, we want to find the asymptotic variance of parameter estimation. Using an estimated asymptotic variance one can design experiments; how many data points will be needed for estimating parameters to meet a given confidence interval, etc. The following is to obtain the asymptotic variance at predicted parameter values: $a = \bar{a}, b = \bar{b}$.

From the asymptotic variance analysis based on the Central Limit Theorem, the covariance of estimation error converges to

$$\mathbf{Q} = \lambda \bar{\mathbf{R}}^{-1} \tag{3}$$

where the matrix $\bar{\mathbf{R}}$ is given by

$$\begin{aligned} \bar{\mathbf{R}} &= E\left[\varphi(t)\varphi^T(t)\right] = E\left[\begin{pmatrix} -y(t-1) \\ u(t-1) \end{pmatrix}\begin{pmatrix} -y(t-1) & u(t-1) \end{pmatrix}\right] \\ &= \begin{pmatrix} E\left[y^2(t-1)\right] & -E[u(t-1)y(t-1)] \\ -E[u(t-1)y(t-1)] & E\left[u^2(t-1)\right] \end{pmatrix} = \begin{bmatrix} R_y(0) & -R_{uy}(0) \\ -R_{uy}(0) & \mu \end{bmatrix} \end{aligned} \tag{4}$$

Now the challenge is to find $R_y(0)$ and $R_{uy}(0)$. These can be computed from the assumed model (1). Namely,

$$y(t) + ay(t-1) = bu(t-1) + e(t) \tag{5}$$

$$y(t)+ay(t-1)=bu(t-1)+e(t) \tag{5}$$

From this expression, we can obtain several equations representing the relationships among auto- and cross-correlations. Multiplying $y(t-1)$ to both sides of (5) and taking expectation yield

$$E[y(t)y(t-1)]+aE[y(t-1)^2]=bE[u(t-1)y(t-1)]+E[e(t)y(t-1)]$$

$$\therefore \quad R_y(1)+aR_y(0)=bR_{uy}(0) \tag{6}$$

Multiplying $y(t)$ to both sides and taking expectation yield

$$E[y(t)^2]+aE[y(t)y(t-1)]=bE[u(t-1)y(t)]+E[e(t)y(t)],$$

from which we obtain

$$\therefore \quad R_y(0)+aR_y(1)=bR_{uy}(1)+R_{ey}(0) \tag{7}$$

Similarly, multiplying $e(t)$, $u(t)$, and $u(t-1)$ to (5), respectively, yields,

$$R_{ye}(0)=\lambda \tag{8}$$

$$R_{uy}(0)=0 \tag{9}$$

$$R_{uy}(1)+aR_{uy}(0)=b\mu \tag{10}$$

$$R_y(1) + aR_y(0) = bR_{uy}(0)$$

$$R_y(0) + aR_y(1) = bR_{uy}(1) + R_{ey}(0)$$

$$R_{ye}(0) = \lambda$$

$$R_{uy}(0) = 0$$

$$R_{uy}(1) + aR_{uy}(0) = b\mu$$

Solving these simultaneous equations for $R_y(0)$ and $R_{uy}(0)$, we find

$$R_y(0) = \frac{b^2\mu + \lambda}{1 - a^2}, R_{uy}(0) = 0$$

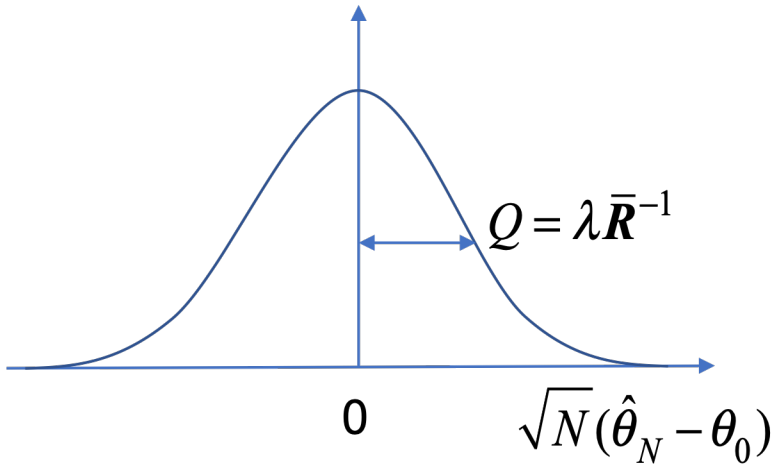
Substituting these into (4) yields

$$\mathbf{Q} = \lambda\bar{\mathbf{R}}^{-1} = \begin{bmatrix} \dfrac{\lambda(1-\bar{a}^2)}{\bar{b}^2\mu + \lambda} & 0 \\ 0 & \dfrac{\lambda}{\mu} \end{bmatrix}$$

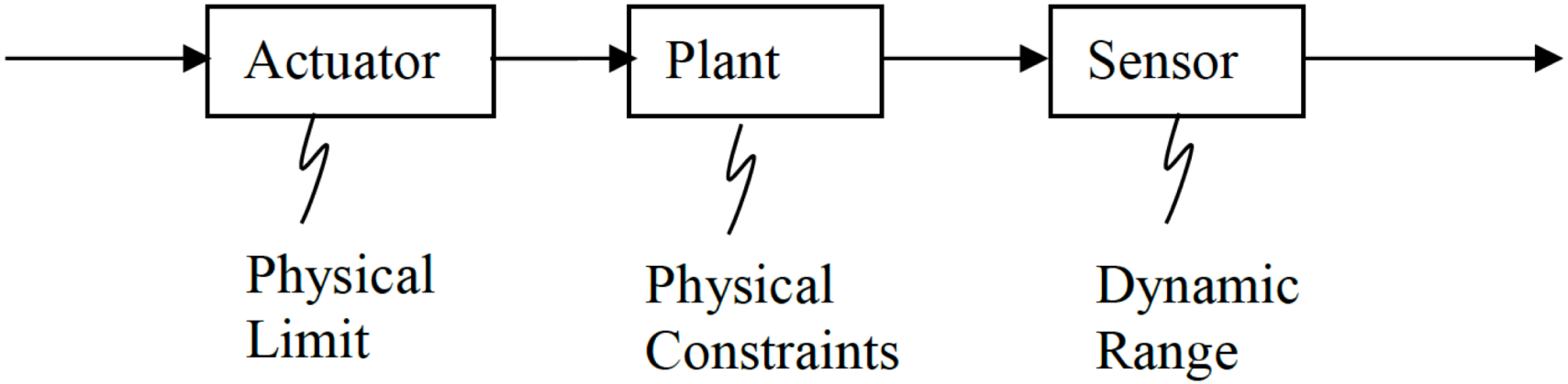
Parameter a

Signal-to-Noise Ratio

Parameter b



Input Design

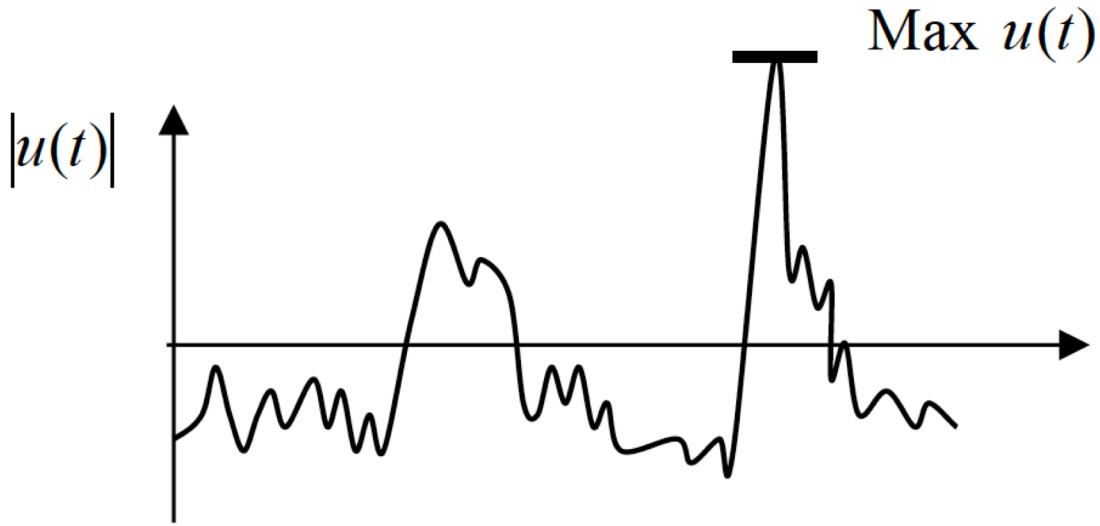


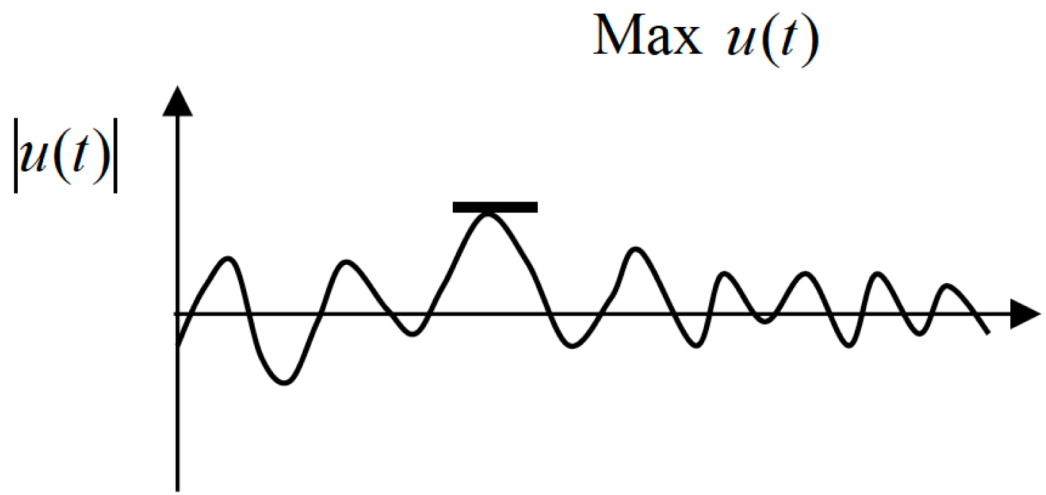
When conducting experiments, various constraints due to physical limits of the process must be satisfied. Among others, the input amplitude limit is a common constraint

$$\underline{u} \le |u(t)| \le \bar{u} \tag{15}$$

where $\underline{u}$ and $\bar{u}$ are lower and upper limits.

In general, the larger the input magnitude $|u(t)|$ becomes, the smaller the asymptotic variance P_θ becomes. Therefore the input sequence should have large amplitude most of the time.





To evaluate this aspect of input signal, the following crest factor is often used for zero-mean signals:

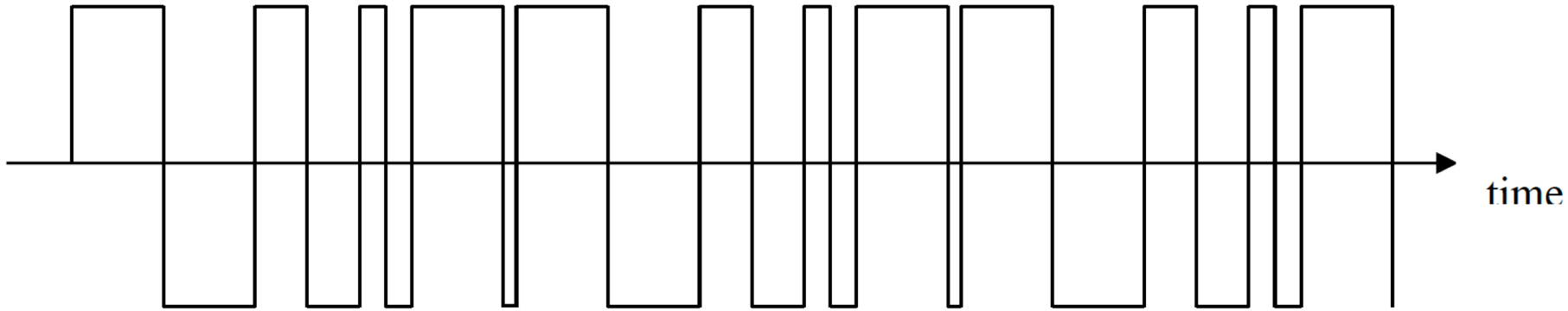
$$C_r^2 \equiv \frac{\max_{1\le t\le N} u^2(t)}{\frac{1}{N}\sum_{t=1}^{N} u^2(t)}, \tag{16}$$

Note $C_r \ge 1$; **Smaller is better.**

19.5 System ID Using Random Signals

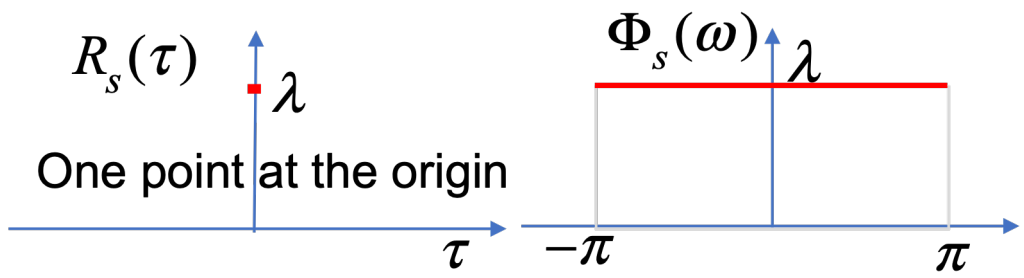
The best crest factor (1, the lowest value) is achieved with zero mean binary signals: $u(t) = \pm\bar{u}$. The signal shown below is an example of random binary sequence with an auto-correlation:

$$R_u(\tau) = E[u(t)u(t-\tau)] = 1 \cdot \delta(\tau) \qquad (17)$$



In system identification, random signals are often used for input sequences because of their superb noise reduction properties. The following is to show this property.

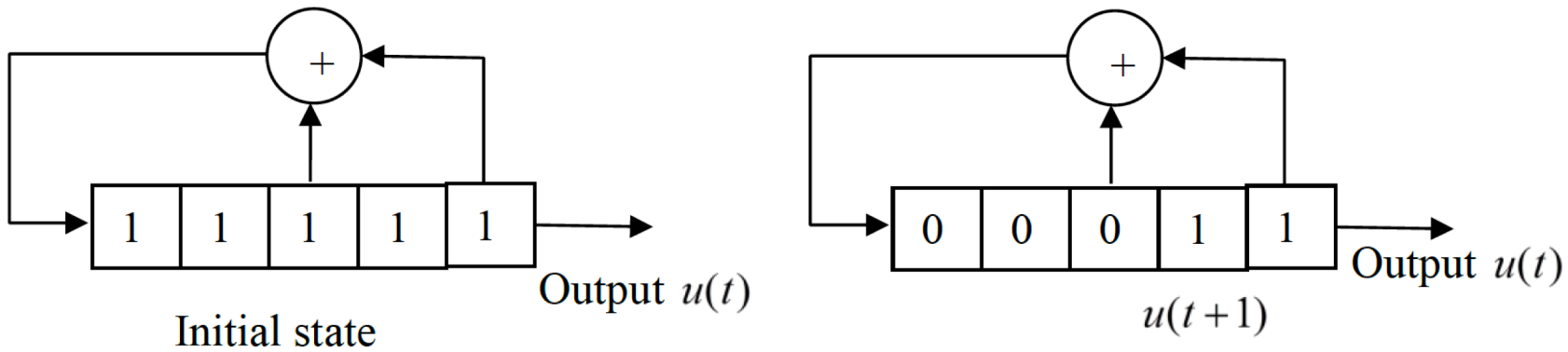
$$C_r^2 \equiv \frac{\max_{1 \le t \le N} u^2(t)}{\frac{1}{N}\sum_{t=1}^{N} u^2(t)} = 1$$



19.6 Pseudo-Random Binary Signal (PRBS)

Now that random binary signals are useful for system identification, how can we generate them? A Pseudo-Random Binary Signal (PRBS) is a periodic, deterministic signal with white noise like properties. It has been widely used for system identification as well as for spread spectrum wireless communication and GPS.

Example: A 5 bit shift register



Binary shift register with feedback

Initial state

1 1 1 1 1 0001101110101000010010110011111000110111010100001001011 00

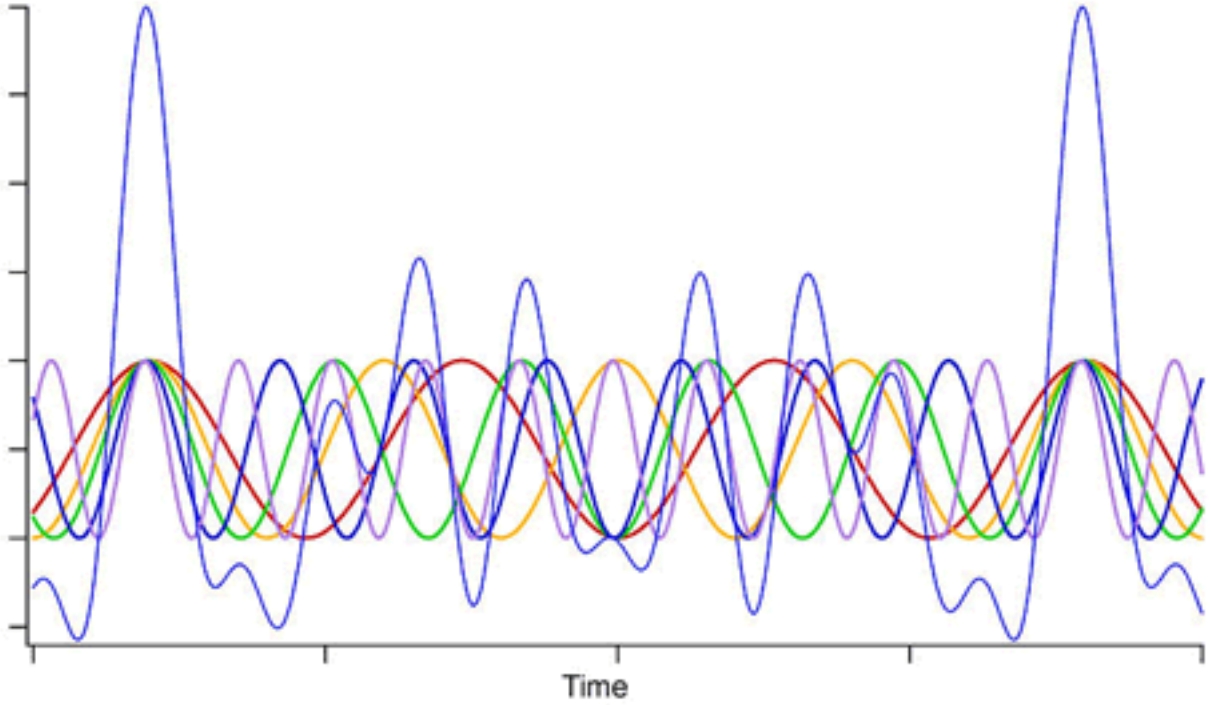
31 bits $=2^5-1$ Repeat the same 31 bits

Multi-Sinewave

Another choice for input design is Sinusoidal signal.

Combining multiple frequencies,

$$u(t) = \sum_{k=1}^{p} a_k \cos(\omega_k t + \phi_k)$$



$$C_r = \sqrt{\frac{\max_{1\le t\le N} u^2(t)}{\frac{1}{N}\sum_{t=1}^{N} u^2(t)}} = \frac{Amplitude \quad of \quad u(t)}{\sqrt{signal \quad power}} = \frac{\sum_{k=1}^{p} a_k}{\sqrt{\sum_{k=1}^{p} \frac{a_k^2}{2}}} \Rightarrow \sqrt{2p}$$

$$for \quad a_k = a, \qquad k = 1, \cdots p$$

If all the sinusoids have the same amplitude, the crest factor is $\sqrt{2p}$, which becomes large as p increases.

Chirp Signal

Continuously varying the frequency between ω_1 and ω_2 over a time period of $0 \le t \le T$ creates a Chirp Signal:

$$u(t) = A\cos\left(\omega_1 t + (\omega_2 - \omega_1)\frac{t^2}{2T}\right)$$

This is one of the most frequently used technique.

The instantaneous frequency, ω_i, is obtained by differentiating $u(t)$

$$\omega_i = \omega_1 + \frac{t}{T}(\omega_2 - \omega_1)$$