2.160 Identification, Estimation, and Learning

Lecture Notes No. 1

February 8, 2006

Mathematical models of real-world systems are often too difficult to build based on first principles ***alone***.

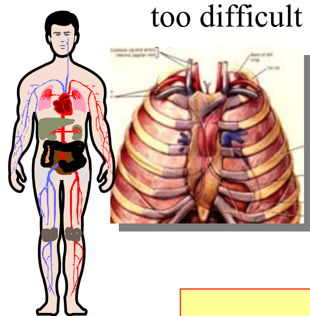


Figure by MIT OCW.

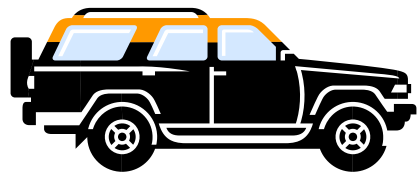


Figure by MIT OCW.

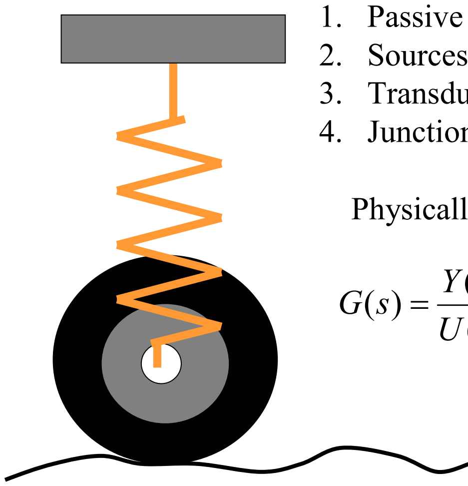
System Identification;
"Let the data speak about the system".

Courtesy of Prof. Asada. Used with permission.

Image removed for copyright reasons.

HVAC

Physical Modeling



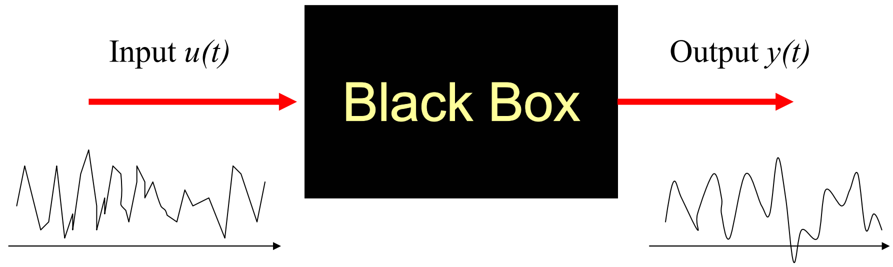
1. Passive elements: mass, damper, spring
2. Sources
3. Transducers
4. Junction structure

Physically meaningful parameters

$$G(s) = \frac{Y(s)}{U(s)} = \frac{b_0 s^m + b_1 s^{m-1} + \cdots + b_m}{s^n + a_1 s^{n-1} + \cdots + a_n}$$

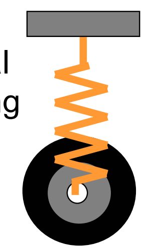
$$a_i = a_i(M, B, K)$$
$$b_i = b_i(M, B, K)$$

System Identification

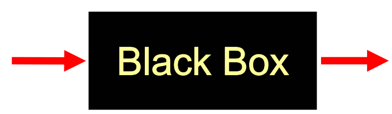


$$G(s) = \frac{Y(s)}{U(s)} = \frac{b_0 s^m + b_1 s^{m-1} + \cdots + b_m}{s^n + a_1 s^{n-1} + \cdots + a_n}$$

Physical modeling



Comparison



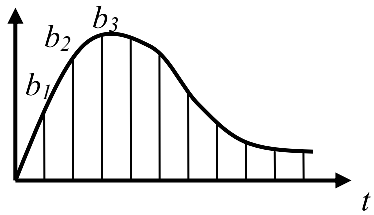
Pros
1. Physical insight and knowledge
2. Modeling a conceived system before hardware is built

Cons
1. Often leads to high system order with too many parameters
2. Input-output model has a complex parameter structure
3. Not convenient for parameter tuning
4. Complex system; too difficult to analyze

Pros
1. Close to the actual input-output behavior
2. Convenient structure for parameter tuning
3. Useful for complex systems; too difficult to build physical model

Cons
1. No direct connection to physical parameters
2. No solid ground to support a model structure
3. Not available until an actual system has been built

Introduction: System Identification in a Nutshell



Finite Impulse Response Model

$$y(t) = b_1 u(t-1) + b_2 u(t-2) + \cdots\cdots + b_m u(t-m)$$

Define $\qquad \theta := [b_1, b_2, \cdots, b_m]^T \in R^m \qquad$ unknown

$$\varphi(t) := [u(t-1), u(t-2), \cdots, u(t-m)]^T \in R^m \qquad \text{known}$$

Vector θ collectively represents model parameters to be identified based on observed data *y(t)* and $\varphi(t)$ for a time interval of $1 \le t \le N$.

Observed data: $y(1), \cdots\cdots, y(N)$

$\longrightarrow$ Estimate θ $\qquad$ Estimation $\hat{y}(t) = \varphi(t)^T \theta$

This predicted output may be different from the actual *y(t)*.

Find θ that minimize $V_N(\theta)$

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

$$\hat{\theta} = avg \min_{\theta} V_N(\theta)$$

$$\left.\begin{aligned} &\frac{dV_N(\theta)}{d\theta} = 0 \\ &V_N(\theta) = \frac{1}{N} \sum_{t=1}^{N} (y(t) - \varphi^T(t)\theta)^2 \end{aligned}\right\}$$

$$\frac{2}{N} \sum_{t=1}^{N} (y(t) - \varphi^T(t)\theta)(-\varphi) = 0$$

$$\sum_{t=1}^{N} y(t)\varphi(t) = \sum_{t=1}^{N} (\varphi^T(t)\theta)\varphi(t)$$

$$\underbrace{\left[\sum_{t=1}^{N}(\varphi(t)\varphi^{T}(t))\right]}_{\substack{\parallel \\ R_N}}\theta=\sum_{t=1}^{N}y(t)\varphi(t)$$

$$\therefore \quad \hat{\theta}_N=R_N^{-1}\sum_{t=1}^{N}y(t)\varphi(t)$$

<u>Question1</u> What will happen if we repeat the experiment and obtain $\hat{\theta}_N$ again?

Consider the expectation of $\hat{\theta}_N$ when the experiment is repeated many times? Average of $\hat{\theta}_N$

Would that be the same as the true parameter θ_0?

Let's assume that the actual output data are generated from

$$y(t)=\varphi^{T}(t)\theta_0+e(t)$$

θ_0 is considered to be the true value.

Assume that the noise sequence *{e(t)}* has a zero mean value, i.e. E[e(t)]=0, and has no correlation with input sequence *{u(t)}*.

$$\hat{\theta}_N=R_N^{-1}\sum_{t=1}^{N}y(t)\varphi(t)=R_N^{-1}\sum_{t=1}^{N}\left[\left(\varphi^{T}(t)\theta_0+e(t)\right)\varphi(t)\right]$$

$$=R_N^{-1}\underbrace{\left(\sum_{t=1}^{N}\varphi(t)\varphi^{T}(t)\right)}_{R_N}\theta_0+R_N^{-1}\sum_{t=1}^{N}\varphi(t)e(t)$$

$$\therefore \quad \hat{\theta}_N-\theta_0=R_N^{-1}\sum_{t=1}^{N}\varphi(t)e(t)$$

Taking expectation

$$E[\hat{\theta}_N-\theta_0]=E\left[R_N^{-1}\sum_{t=1}^{N}\varphi(t)e(t)\right]=R_N^{-1}\sum_{t=1}^{N}\varphi(t)\cdot E[e(t)]=0$$

<u>Question2</u> Since the true parameter θ_0 is unknown, how do we know how close $\hat{\theta}_N$ will be to θ_0? How many data points, *N*, do we need to reduce the error $\hat{\theta}_N-\theta_0$ to a certain level?

Consider the variance (the covariance matrix) of the parameter estimation error.

$$
\begin{aligned}
P_N &= E[(\hat{\theta}_N - \theta_0)(\hat{\theta}_N - \theta_0)^T] \\
&= E\left[R_N^{-1} \sum_{t=1}^{N} \varphi(t) e(t) \cdot \left(R_N^{-1} \sum_{s=1}^{N} \varphi(s) e(s) \right)^T \right] \\
&= E\left[R_N^{-1} \sum_{t=1}^{N} \sum_{s=1}^{N} \varphi(t) e(t) e(s) \varphi^T(s) R_N^{-1} \right] \\
&= R_N^{-1} \left[\sum_{t=1}^{N} \sum_{s=1}^{N} \varphi(t) E[e(t)e(s)] \varphi^T(s) \right] R_N^{-1}
\end{aligned}
$$

Assume that *{e(t)}* is stochastically independent

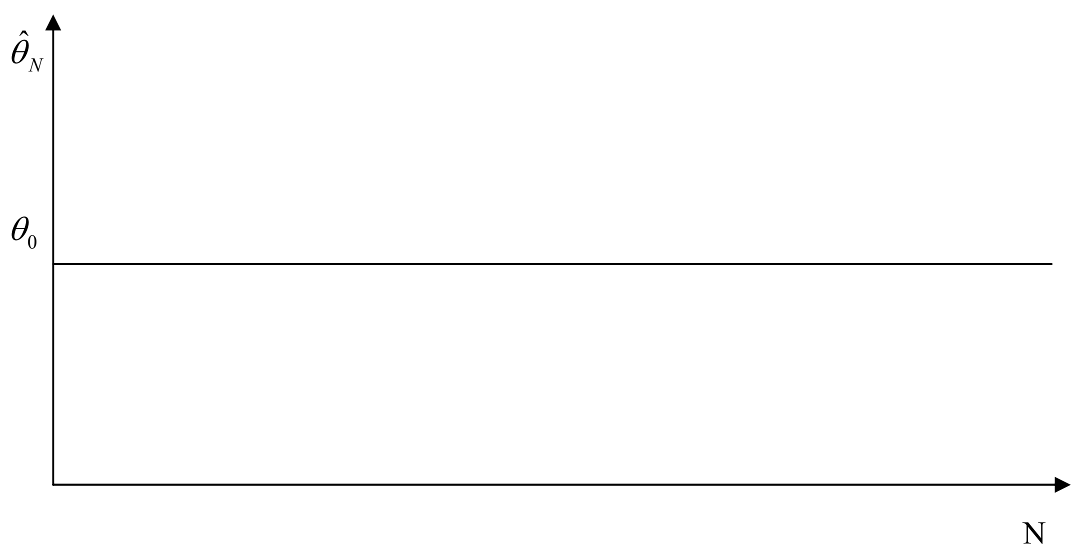
$$
E[e(t)e(s)] = \begin{cases} E[e(t)e(s)] = 0 & t \neq s \\ E[e^2(t)] = \lambda & t = s \end{cases}
$$

Then $P_N = R_N^{-1} \left[\sum_{t=1}^{N} \varphi(t) \lambda \varphi^T(t) \right] R_N^{-1} = \lambda R_N^{-1}$

As N increases, R_N tends to blow out, but R_N/N converges under mild assumptions.

$$
\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} \varphi(t) \varphi^T(t) = \lim_{N\to\infty} \frac{1}{N} R_N = \overline{R}
$$

For large N, $R_N \cong N\overline{R}$, $R_N^{-1} \cong \frac{1}{N} \overline{R}$



$P_N = \frac{\lambda}{N} \overline{R}^{-1}$ for large N.

I. The covariance P_N decays at the rate *1/N*.

$\longrightarrow$ Parameters approach he limiting value at the rate of $\dfrac{1}{\sqrt{N}}$

II. The covariance is inversely proportional to

$$P_N \propto \frac{\lambda}{magnitude\overline{R}}$$

$$R_N = \begin{bmatrix} r_{11} & \cdots & r_{1m} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{mm} \end{bmatrix}$$

$$r_{ij} = \sum_{t=1}^{N} u(t-i)u(t-j)$$

III. The convergence of $\hat{\theta}_N$ to θ_0 may be accelerated if we design inputs such that $\overline{R}$ is large.

IV. The covariance does not depend on the average of the input signal. Only the second moment

What will be addressed in 2.160?

A) How to best estimate the parameters

What type of input is maximally informative?

- Informative data sets
- Persistent excitation
- Experiment design
- Pseudo Random Binary signals, Chirp sine waves, etc.

How to best tune the model / best estimate parameters

How to best use each data point

- Covariance analysis
- Recursive Least Squares
- Kalman filters
- Unbiased estimate
- Maximum Likelihood

B). How to best determine a model structure

How do we represent system behavior? How do we parameterize the model?

i. Linear systems
   - FIR, ARX, ARMA, BJ,…..
   - Data compression: Laguerre series expansion

ii. Nonlinear systems
   - Neural nets
   - Radial basis functions

iii. Time-Frequency representation
   - Wavelets

Model order: Trade-off between accuracy/performance and reliability/robustness

- Akaike's Information Criterion
- MDL