Fall 2021 ME424 Modern Control and Estimation

Lecture Note 3
Least Squares and Basic System Identification

- control/robotics: system identification / kalman filter ∈
- signal processing: curve fitting / compressive sensing / face recognition / Fourier transform
- Machine learning: linear regression

Prof. Wei Zhang
Department of Mechanical and Energy Engineering
SUSTech Institute of Robotics
Southern University of Science and Technology

zhangw3@sustech.edu.cn
https://www.wzhanglab.site/

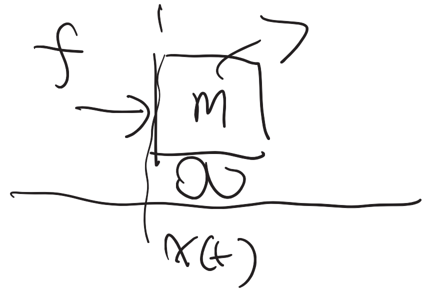
Outline

- **Least-Squares Problem Formulation**
- Solution to Linear Least-Squares Problems
- Linear Least-Squares Examples
- Applications to System ID
- Nonlinear Least Squares

- Last lecture: obtain discrete-time linear state space model from
  - physical process
  - given continuous time state space model
  - given nonlinear state space model
  - given discrete time transfer function
- The goal of this lecture note:
  - learn how to build model based on observed input-output data pairs
    - General case beyond the scope of this course
    - Focus on special case, where first obtain transfer function model from input-output data pairs, and then obtain the corresponding state space model

  - Main method: **Least Squares**

Least-Squares Problem Formulation:



- Measurement Equation:

model

$$y = g(\theta) + v$$

hidden

Data:

$f(t_1)$	$x(t_1)$
$f(t_2)$	$x(t_2)$
⋮	⋮

- $y \in R^m$: measurements data

constraint set $\theta \in [0,1]$ $\theta \geq 0$

- $\theta \in \Theta \subseteq R^n$: parameter to be estimated, where Θ is the constraint set for feasible parameters

→ y, g

$\boxed{\theta = m} > 0$

- $v \in R^m$: unknown measurement noise

- $g: R^n \to R^m$: known (possibly) nonlinear function relates θ with measurement y

$y = 2\theta + v$

$g(\theta) = 2\theta$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \underbrace{\begin{bmatrix} \cos\theta_1 \\ \cos\theta_2 \\ \cos(\theta_1+\theta_2) \end{bmatrix}}_{g(\theta)} + \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}$$

$g: \mathbb{R}^2 \to \mathbb{R}^3$

Least-Squares Problem Formulation:

- **Problem Statement**: Find the best parameter in the constraint set Θ that minimizes the difference between the model and the measured data

measurement

$$\min_{\theta\in\Theta} J(\theta) = \min_{\theta\in\Theta} ||y - g(\theta)||^2$$

model output/prediction

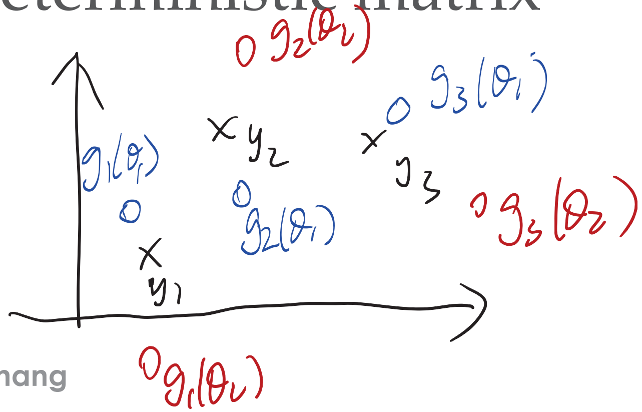
$$\theta = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}, \quad g(\theta) = \begin{bmatrix} g_1(\theta) \\ g_2(\theta) \\ \vdots \\ g_m(\theta) \end{bmatrix}$$

$$= \min_{\theta\in\Theta} \left((y - g(\theta))^T (y - g(\theta)) \right)$$

$$= \min_{\theta\in\Theta} \left[\sum_{i=1}^{m} (y_i - g_i(\theta))^2 \right]$$

- **Linear Least Squares**: $g(\theta) = H\theta$, where $H \in R^{m\times n}$ is a given deterministic matrix



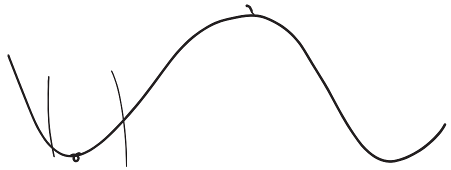
$$(\theta_1 \text{ vs. } \theta_2) \quad ||g(\theta_1) - y||^2 < ||g(\theta_2) - y||^2$$

Outline

- Least-Squares Problem Formulation
- **Solution to Linear Least-Squares Problems**
- Linear Least-Squares Examples
- Applications to System ID
- Nonlinear Least Squares

$df = f'dx$

Optimization of multivariable function



- 1^{st} –order necessary condition for optimality of $J(\theta)$

e.g. 1-dim, $\theta \in \mathbb{R}$ $\quad J(\theta) = J(\theta_0) + \frac{\partial J}{\partial \theta}\Big|_{\theta_0} \cdot (\theta - \theta_0) + H.O.T.$

If θ_0 is a minimizer $\Rightarrow \frac{\partial J}{\partial \theta}\Big|_{\theta_0} = 0$

because if ① $\frac{\partial J}{\partial \theta}\Big|_{\theta_0} > 0$, we can choose $\hat{\theta} = \theta_0 - \varepsilon \Rightarrow J(\hat{\theta}) < J(\theta_0)$

② ... < 0, ... $\hat{\theta} = \theta_0 + \varepsilon \Rightarrow J(\hat{\theta}) < J(\theta_0)$

- Matrix calculus:
  - If $f: \mathrm{R}^n \to R^m$, then $\frac{\partial f}{\partial x}(\mathrm{x}) = Df(x) \Rightarrow \left[\frac{\partial f}{\partial x}\right]_{ij} = \left(\frac{\partial f_i}{\partial x_j}\right)$

$f: \mathbb{R}^2 \to \mathbb{R}^3$ $\quad f(x) = \begin{bmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \\ f_3(x_1, x_2) \end{bmatrix} \Rightarrow df = \left(\frac{\partial f}{\partial x}\right) \cdot dx$

$= \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} \\ \frac{\partial f_2}{\partial x_1} & \vdots \\ \frac{\partial f_3}{\partial x_1} & \end{bmatrix} \begin{bmatrix} dx_1 \\ dx_2 \end{bmatrix}$

Optimization of multivariable function

- **Gradient**: For scaler valued multivariate function $f: R^n \to R$, its gradient is defined as:

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}, \quad df = (\nabla f(x))^T dx$$

$$dx = \begin{bmatrix} 0 \\ \theta - \varepsilon \\ 0 \\ \vdots \end{bmatrix}$$

For L.S. problem $J(\theta)$

$J: \mathbb{R}^n \to \mathbb{R}$

Necessary optimality condition

$$\boxed{\nabla J(\theta^*) = 0}$$

- For $f: R^n \to R$, notational convention $\nabla f(x) = \left(\frac{\partial f}{\partial x}(x)\right)^T$

- Some references use $\frac{\partial f}{\partial x}$ to denote gradient

- Directional derivative: $Df(x; d) = \lim_{\alpha \to 0} \frac{f(x+\alpha d) - f(x)}{h} = \langle \nabla f, d \rangle$

If $f: \mathbb{R}^n \to \mathbb{R}$, $f(x) = a^T x$ $\Downarrow$ $\frac{\partial f}{\partial x} = a^T$, $\nabla f(x) = a$

$A = \begin{bmatrix} a_{11} & a_{12} & \cdots \\ & & \end{bmatrix}$

- Some calculus examples:
  - $f(x) = Ax$, $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, $f(x) \in \mathbb{R}^m$

$\frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots \\ \frac{\partial f_2}{\partial x_1} & \vdots & \\ \vdots & & \end{bmatrix} = A$, e.g. $f_1(x) = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n$

$\frac{\partial f_1}{\partial x_1} = a_{11}$, $\frac{\partial f_1}{\partial x_2} = a_{12}$

  - $f(x) = x^T A x = \sum_i \sum_j a_{ij} x_i x_j$

$x \in \mathbb{R}^n$
$A \in \mathbb{R}^{n \times n}$
$f: \mathbb{R}^n \to \mathbb{R}$

$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$

$\frac{\partial f}{\partial x_k} = \frac{\partial}{\partial x_k}\left[\sum_{i \neq k} \sum_{j \neq k} a_{ij} x_i x_j + \sum_{i \neq k} a_{ik} x_i x_k + \sum_{j \neq k} a_{kj} x_k x_j + a_{kk} x_k^2 \right]$

$= \sum_{i=1}^{n} a_{ik} x_i + \sum_{j=1}^{n} a_{kj} x_j$

$\nabla f(x) = A x + A^T x$

if A symmetric: $A = A^T$ $\quad \nabla f(x) = 2Ax$

- **Exercise**: compute $\frac{\partial f}{\partial x}(x)$, where $x \in R^n$, and $f(x) = x^T x \cdot x$ $\Leftarrow$

By definition V.F.Y.

$g(\theta) = H\theta$

- Derivation of $\boxed{\text{linear}}$ least square solutions
  - Normal equation: L.S. $\to \min_\theta J(\theta) = \min_\theta \|y - H\theta\|^2$

$J: \mathbb{R}^n \to \mathbb{R}$

$J(\theta) = \|y - H\theta\|^2 = (y - H\theta)^T (y - H\theta) = y^T y - \theta^T H^T y - y^T H \theta + \theta^T H^T H \theta$

(Note: $\theta^T H^T y = y^T H \theta$, similar to $[3\ 4]\begin{bmatrix}1\\5\end{bmatrix} = [1\ 5]\begin{bmatrix}3\\4\end{bmatrix}$

$= \theta^T H^T H \theta - 2 y^T H \theta + y^T y$

$\Rightarrow \nabla_\theta J(\theta) = 2 H^T H \theta - 2 H^T y + 0 = 0 \Rightarrow \boxed{(H^T H)\theta = H^T y}$

Normal equation

$\theta_{LS} = (H^T H)^{-1} H^T y$

Solution with full rank H :

Recall : $H \in \mathbb{R}^{m \times n}$ assume $m > n$

Normal equation: $(H^T H)\theta = H^T y$

① If H is full rank $(\mathrm{rank}(H) = n)$, $\overset{\text{V.F.Y}}{\Longleftrightarrow}$ $\underbrace{H^T H}_{n \times n}$ is nonsingular

$\Rightarrow \hat{\theta}_{LS} = (H^T H)^{-1} H^T y$ H

② If H is not full rank $\Rightarrow$ $H^T H$ singular

- e.g. $H^T H = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$, $H^T y = \begin{bmatrix} 3 \\ 6 \end{bmatrix}$, $\Rightarrow \hat{\theta}_{LS} = \begin{bmatrix} 3 \\ 0 \end{bmatrix} + \alpha \begin{bmatrix} 2 \\ -1 \end{bmatrix}$

$\mathrm{Null}((H)^T H) = \mathrm{span}\left(\begin{bmatrix} 2 \\ -1 \end{bmatrix}\right)$

e.g 2 : $H^T H = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$, $H^T y = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ $\Rightarrow$ No solution.

$\mathrm{col}(H^T H) = \mathrm{span}\left(\begin{bmatrix} 1 \\ 2 \end{bmatrix}\right)$

$J(\theta) = \|y - H\theta\|^2$

- Geometric interpretation of linear least squares

① For any $\theta \in \mathbb{R}^n$, $H\theta$ is a linear combination of columns of H

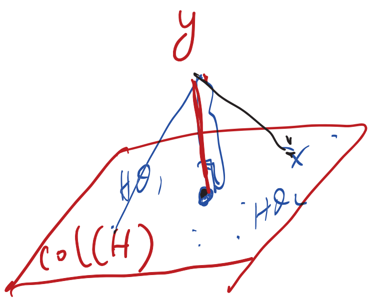
② If $y \in col(H)$, we can find $\hat{\theta}$ such that $y = H\hat{\theta} \Rightarrow J(\hat{\theta}) = 0$

V.F.Y.: If $y \in col(H)$, then $J(\hat{\theta}_{LS}) = 0$ $\leftarrow (H^TH)^{-1}H^Ty$

i.e. $\|y - H\cdot(H^TH)^{-1}H^Ty\|^2 = 0$

③ If $y \notin col(H)$, no exact solution, we have to find the minimum distance solution

Geometrically, e.g. $H = [h_1 \; h_2]$. L.S. tries to find the $\hat{\theta}_{LS}$ to $\min \|y - H\theta\|^2$. Intuitively, $H\hat{\theta}_{LS}$ should be the proj of y onto $col(H)$.



Let's verify our intuition with $\hat{\theta}_{LS}$: error vector $(y - H\hat{\theta}_{LS})$ should be orthogonal to $col(H)$, i.e. $(y - H\hat{\theta}_{LS})^T (H\beta) = 0, \forall \beta$

$$\Rightarrow \left(y - \underline{H(H^T H)^{-1} H^T y}\right)^T H\beta$$

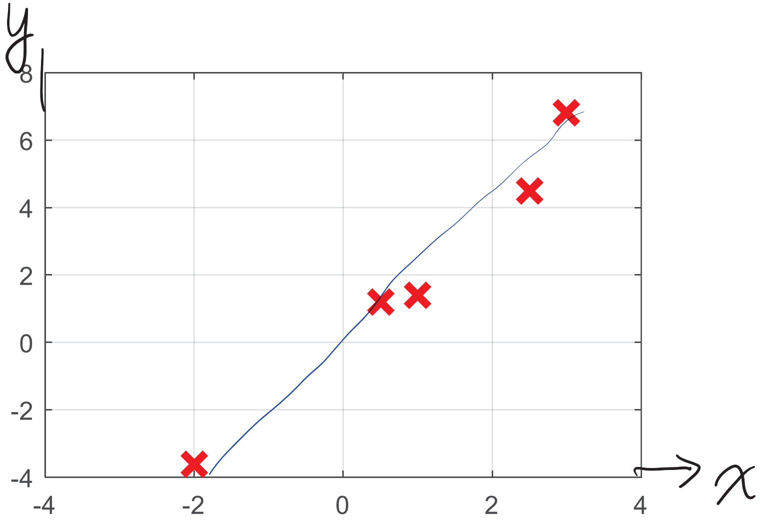
Outline

$$= y^T H \beta - y^T H \cancel{(H^T H)^{-1} H^T} H\beta = 0 \Rightarrow$$

- Least-Squares Problem Formulation
- Solution to Linear Least-Squares Problems
- **Linear Least-Squares Examples**
- Applications to System ID
- Nonlinear Least Squares

- Linear Least Squares Example:

i	1	2	3	4	5
x	1	0.5	-2	3	2.5
y	1.4	1.2	-3.6	6.8	4.5



- Assume $y = \alpha x + \beta + v$

$$\theta = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

Here, $y = \begin{bmatrix} 1.4 \\ 1.2 \\ -3.6 \\ 6.8 \\ 4.5 \end{bmatrix}$, $H = \begin{bmatrix} 1 & 1 \\ 0.5 & 1 \\ -2 & 1 \\ 3 & 1 \\ 2.5 & 1 \end{bmatrix}$

$$\begin{bmatrix} y_1 = \alpha x_1 + \beta + v_1 \\ y_2 = \alpha x_2 + \beta + v_2 \\ \vdots \\ y_5 = \alpha x_5 + \beta + v_5 \end{bmatrix} = \underbrace{\begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_5 & 1 \end{bmatrix}}_{H} \underbrace{\begin{bmatrix} \alpha \\ \beta \end{bmatrix}}_{\theta} + v$$

y

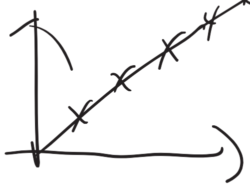
$$\Rightarrow \hat{\theta}_{LS} = \begin{bmatrix} \hat{\alpha}_{LS} \\ \hat{\beta}_{LS} \end{bmatrix} = (H^T H)^{-1} H^T y$$

to be estimated

$$y = H\theta + v$$

known

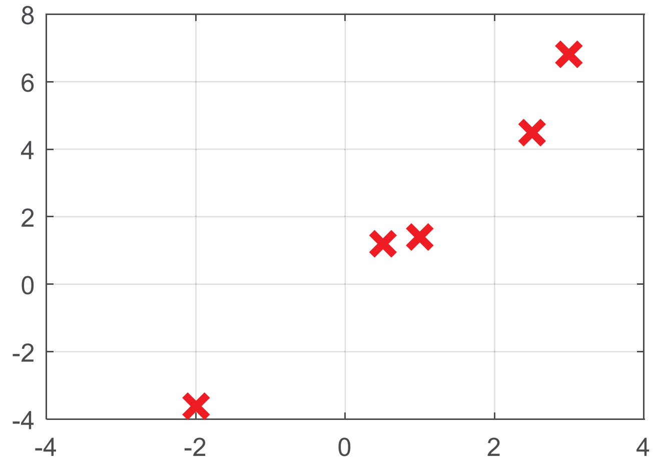
$$g(\theta) = \hat{y}$$

If  $y \in col(H)$,

- Change hypothesis, assume $\underset{\sim}{y \approx be^{ax}}$

use the same data, find the L.S. estimate for (a, b)

$\theta = \begin{bmatrix} a \\ b \end{bmatrix}$, $y \approx be^{ax}$, take log.



$\log y \approx \log b + ax \Rightarrow$

$$\log y_1 \approx \overset{c}{\log b} + a x_1$$
$$\log y_2 \approx \log b + a x_2$$
$$\vdots$$
$$\log y_5 \approx \log b + a x_5$$

defind $c = \log b$

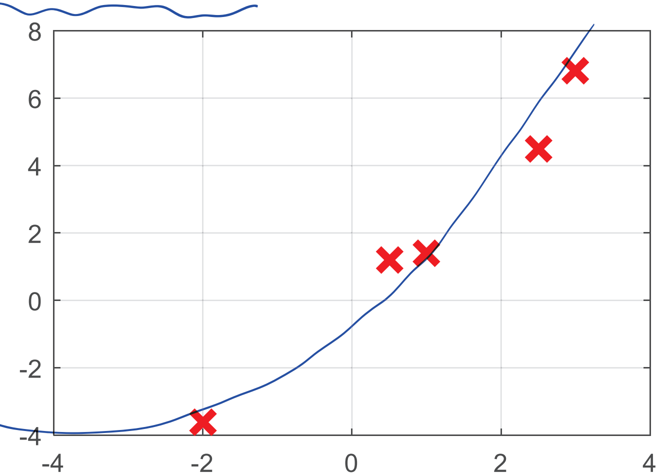
$$\Rightarrow \underbrace{\begin{bmatrix} \log y_1 \\ \log y_2 \\ \vdots \\ \log y_5 \end{bmatrix}}_{\tilde{y}} = \underbrace{\begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \\ x_5 & 1 \end{bmatrix}}_{H} \underbrace{\begin{bmatrix} a \\ c \end{bmatrix}}_{\tilde{\theta}}$$

$$\Rightarrow \hat{\theta}_{LS} = (H^T H)^{-1} H^T \tilde{y}$$

$$\rightarrow \begin{bmatrix} a_{LS} \\ c_{LS} \end{bmatrix} \rightarrow \hat{b}_{LS} = e^{\hat{c}_{LS}}$$

- Change hypothesis, assume that $y = \alpha_0 + \alpha_1 x + \alpha_2 x^2$

same data, find L.S. estimate for $\alpha_0, \alpha_1, \alpha_2$



$$\theta = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{bmatrix}, \quad y_1 = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_1^2 + v_1$$

$$y_2 = \alpha_0 + \alpha_1 x_2 + \alpha_2 x_2^2 + v_2$$

$$\vdots$$

$$y_5 = \alpha_0 + \alpha_1 x_5 + \alpha_2 x_5^2 + v_5$$

$$\Rightarrow \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_5 \end{bmatrix}}_{y} = \underbrace{\begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & & \\ 1 & x_5 & x_5^2 \end{bmatrix}}_{H} \underbrace{\begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{bmatrix}}_{\theta} + v$$

$$\Rightarrow \hat{\theta}_{LS} = (H^T H)^{-1} H^T y$$

Outline

- Least-Squares Problem Formulation
- Solution to Linear Least-Squares Problems
- Linear Least-Squares Examples
- **Applications to System ID**
- Nonlinear Least Squares

Application to System Identification for Linear Systems

- ARX(p, q) model :(Autoregressive with exogenous input)

$$y(k) + \alpha_1 y(k-1) + \cdots + \alpha_p y(k-p)$$ ← auto regressive

$$= \beta_0 u(k) + \beta_1 u(k-1) + \cdots + \beta_q u(k-q) + v(k)$$

← moving average

  - $v(k)$: noise signal

If $y(k) + y(k-1) = 2u(k) + \beta_1 u(k-1) + \beta_2 u(k-2)$. $\theta = \begin{bmatrix} 1 \\ 2 \\ \beta_1 \\ \beta_2 \end{bmatrix}$ ✗, $\theta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$ ✓

  - Model parameter: $\theta = [\alpha_1, \dots, \alpha_p, \beta_0, \beta_1, \dots, \beta_q]^T$

  - One-step predictor:

$$\hat{y}(k|\theta) = -\alpha_1 y(k-1) - \cdots - \alpha_p y(k-p)$$
$$+\beta_0 u(k) + \beta_1 u(k-1) + \cdots + \beta_q u(k-q)$$

given parameter θ, use expect to see an output measurement $\hat{y}(k|\theta)$

- System ID problem for ARX model:

  Given data pairs $\{(u(k), y(k))\}_{k \le N}$, find the parameter vector θ that minimizes cost:

  - $J(\theta) = \sum_{k=1}^{N} ||\hat{y}(k|\theta) - y(k)||^2$

$$J(\theta) = \sum_{k=1}^{N} || \hat{y}(k|\theta) - y(k) ||^2$$

- Formulate as least square problem:

  given data set, $(u_1, y_1), (u_2, y_2), \dots, (u_m, y_m)$ ← m = # of data pairs

  - Regressor: at time k

$$\hat{y}(k;\theta) = -\alpha_1 y(k-1) - \alpha_2 y(k-2) \cdots - \alpha_p y(k-p) + \beta_0 u(k) + \beta_1 u(k-1) + \cdots + \beta_q u(k-q)$$

$$= \underbrace{\begin{bmatrix} -y(k-1) & -y(k-2) & \cdots & -y(k-p) & u(k) & u(k-1) & \cdots & u(k-q) \end{bmatrix}}_{\text{is called "regressor"} \triangleq \phi^T(k)} \underbrace{\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \\ \beta_0 \\ \vdots \\ \beta_q \end{bmatrix}}_{\theta}$$

$$\triangleq \phi^T(k)\,\theta$$

Note: For $\phi^T(k)$ to be well-defined, we need $k > \max\{p, q\}$

- Derivation continued

let's denote $k_0 = \max\{p, q\} + 1$

$$y(k_0) = \phi^T(k_0)\,\theta + v(k_0)$$

$$y(k_0+1) = \underbrace{\phi^T(k_0+1)}\,\theta + v(k_0+1)$$

$$\vdots$$

$$\phi^T(k_0+1) = [-y(k_0) \quad -y(k_0-1) \quad \cdots\cdots\ ]$$

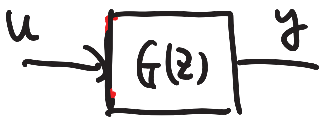
$$\underbrace{\begin{bmatrix} y(k_0) \\ y(k_0+1) \\ \vdots \\ y(m) \end{bmatrix}}_{y} = \underbrace{\begin{bmatrix} \phi^T(k_0) \\ \phi^T(k_0+1) \\ \vdots \\ \phi^T(m) \end{bmatrix}}_{H} \theta + v$$

$$\Rightarrow \quad \hat{\theta}_{LS} = (H^T H)^{-1} H^T y$$

- **System ID Example I:**

$G(z) = \frac{(z^2+b)}{z^3+az}$, find best estimate for a, b,

given data set $(u_1, y_1), (u_2, y_2), \ldots, (u_{20}, y_{20})$



1°: Find ARX model, $G(z) = \frac{z^{-1} + bz^{-3}}{1 + az^{-2}}$

$\Rightarrow Y(z) = G(z)U(z) \Rightarrow (1+az^{-2})Y(z) = (z^{-1} + bz^{-3})U(z)$

Inverse z-transform

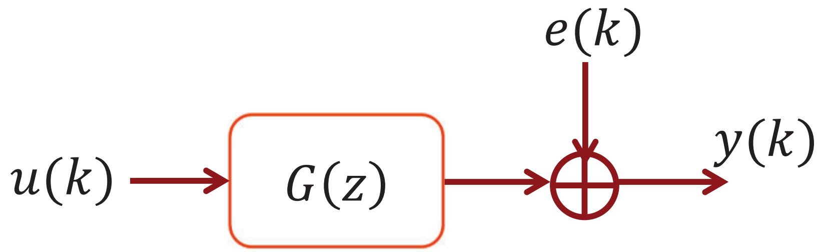
$\Rightarrow \quad y(k) + ay(k-2) = u(k-1) + bu(k-3)$

ARX(2, 3), $k_0 = 3+1 = 4$: $y(4) = -ay(2) + u(3) + bu(1)$

Let $\theta = \begin{bmatrix} a \\ b \end{bmatrix}$ $\quad \Rightarrow y(4) - u(3) = \begin{bmatrix} -y(2) & u(1) \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$

$$\Rightarrow \underbrace{\begin{bmatrix} y(4) - u(3) \\ y(5) - u(4) \\ \vdots \\ y(20) - u(19) \end{bmatrix}}_{y} = \underbrace{\begin{bmatrix} -y(2) & u(1) \\ -y(3) & u(2) \\ \vdots & \\ -y(18) & u(17) \end{bmatrix}}_{H} \underbrace{\begin{bmatrix} a \\ b \end{bmatrix}}_{\theta} \Rightarrow \hat{\theta}_{LS} = (H^T H)^{-1} H^T y$$

- **System ID Example 2:**



  - $G(z) = \frac{z-1}{z-a}$, where a is an unknown scalar
  - Data: $u(1) = 1, u(2) = \frac{1}{2}, u(3) = 1, y(1) = 2, y(2) = 1, y(3) = 2$

Outline

- Least-Squares Problem Formulation
- Solution to Linear Least-Squares Problems
- Linear Least-Squares Examples
- Applications to System ID
- **Nonlinear Least Squares**

- Nonlinear Least Squares:

$$\min_{\theta \in \Theta} J(\theta) = ||y - g(\theta)||^2$$

min $\theta \in \Theta$

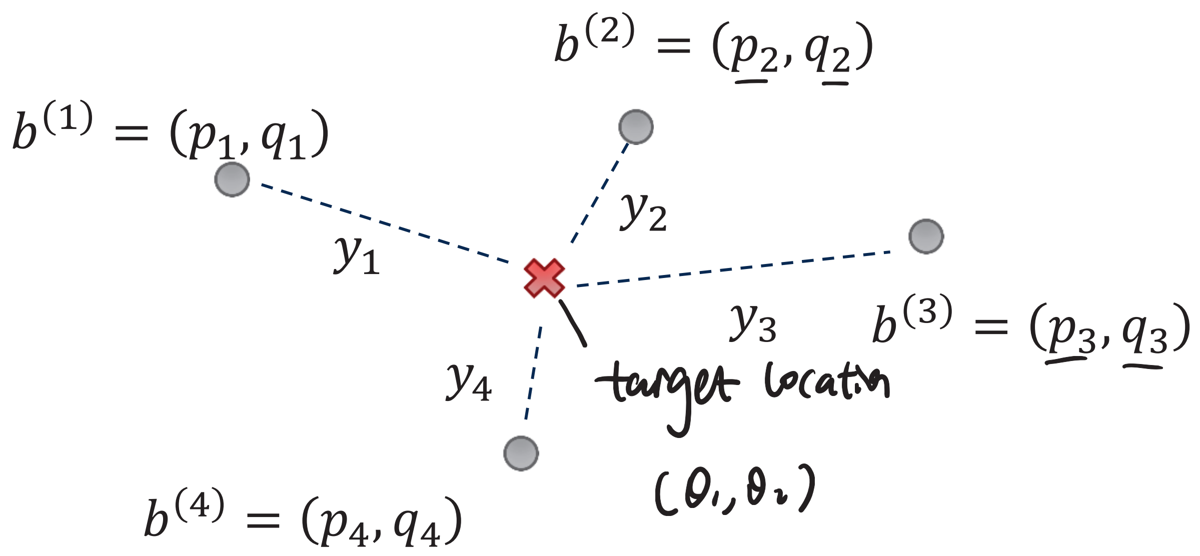
Is a linear L.S. if $g(\theta) = H\theta$

is nonlinear otherwise

  - For general nonlinear function $g(\theta)$, analytical solution to the above optimization is not available
  - Numerical optimization algorithms can be used to find the optimizer

$$\theta^* = argmin_{\theta \in \Theta} J(\theta)$$

- Nonlinear Least Square Example: Navigation by range measurement:



- : beacons with known positions $b^{(i)} = (p_i, q_i)$
- : target with unknown position $\theta = (\theta_1, \theta_2)$
- y_i : known measured distance or range from beacon i:

  typical assumption: $y_i = \|b^{(i)} - \theta\| + v_i$

  $= \sqrt{(p_i - \theta_1)^2 + (q_i - \theta_2)^2} + v_i$

- Given measurements $y_1, y_2, \dots, y_m$, find the best target location θ

- We can choose cost function: $J(\theta) = \sum_{i=1}^{m} \left(y_i - \left\| b^{(i)} - \theta \right\| \right)^2$

$$\hat{\theta}_{MLS} = \underset{\theta \in \Theta}{\operatorname{argmin}} \, J(\theta)$$

- Coding Example