

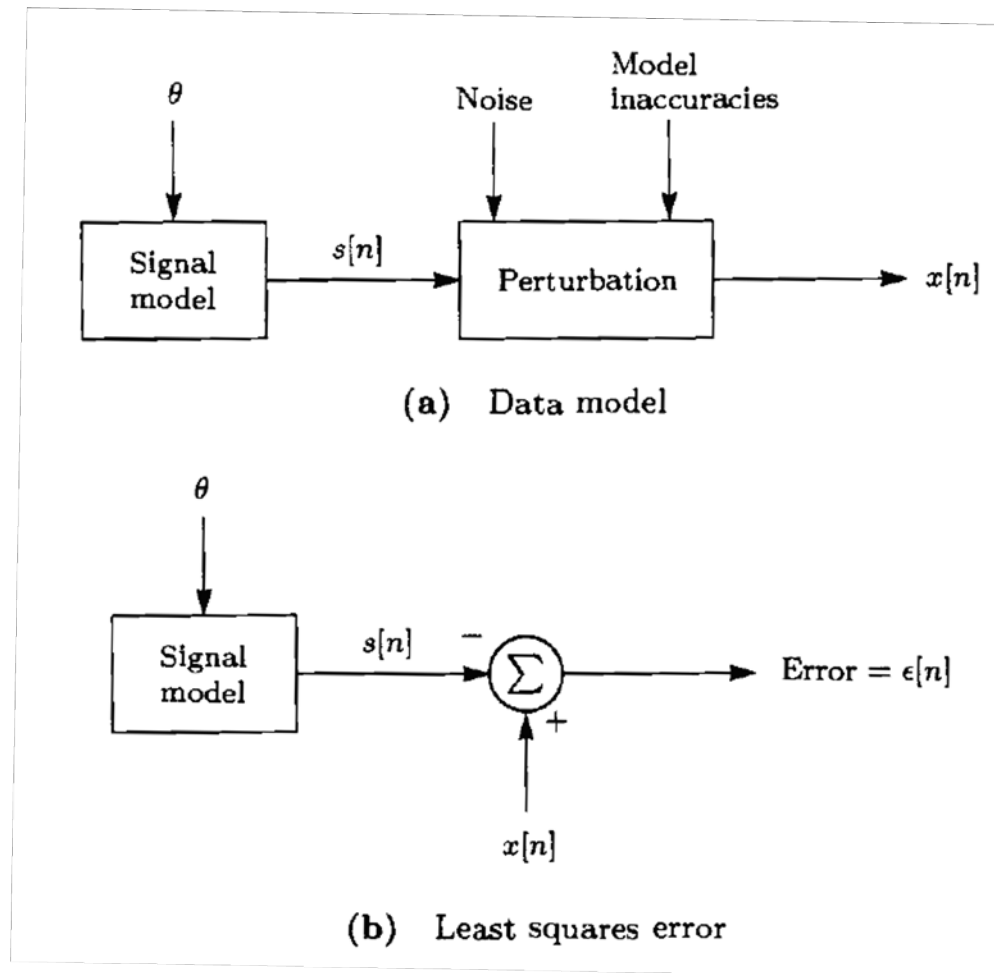
430.714

Estimation Theory

Prof. Songhwai Oh
ECE, SNU

Least Squares

Least Squares Estimator



Least square error criterion:

$$J(\theta) = \sum_{n=0}^{N-1} (x[n] - s[n])^2$$

$$\hat{\theta}_{LS} = \arg \min J(\theta)$$

No probabilistic assumptions

Example:

$$s[n] = A$$

$$J(A) = \sum_{n=0}^{N-1} (x[n] - A)^2$$

$$\begin{aligned} \hat{A} &= \frac{1}{N} \sum_{n=0}^{N-1} x[n] \\ &= \bar{x} \end{aligned}$$

1. Linear least squares
2. Nonlinear least squares

Linear Least Squares

$s[n] = \theta h[n]$ where $h[n]$ is a known sequence.

$$J(\theta) = \sum_{n=0}^{N-1} (x[n] - \theta h[n])^2$$
$$\hat{\theta} = \frac{\sum_{n=0}^{N-1} x[n]h[n]}{\sum_{n=0}^{N-1} h^2[n]}$$

$$\begin{aligned} J_{\min} = J(\hat{\theta}) &= \sum_{n=0}^{N-1} (x[n] - \hat{\theta}h[n])(x[n] - \hat{\theta}h[n]) \\ &= \sum_{n=0}^{N-1} x[n](x[n] - \hat{\theta}h[n]) - \underbrace{\hat{\theta} \sum_{n=0}^{N-1} h[n](x[n] - \hat{\theta}h[n])}_S \\ &= \sum_{n=0}^{N-1} x^2[n] - \hat{\theta} \sum_{n=0}^{N-1} x[n]h[n]. \end{aligned}$$

$$J_{\min} = \sum_{n=0}^{N-1} x^2[n] - \frac{\left(\sum_{n=0}^{N-1} x[n]h[n] \right)^2}{\sum_{n=0}^{N-1} h^2[n]}$$
$$0 \leq J_{\min} \leq \sum_{n=0}^{N-1} x^2[n]$$

Linear Least Squares: a Vector Case

$$\mathbf{s} = \mathbf{H}\boldsymbol{\theta}, \quad \boldsymbol{\theta} \in \mathbb{R}^p, \mathbf{H} \in \mathbb{R}^{N \times p}$$

$$\begin{aligned} J(\boldsymbol{\theta}) &= \sum_{n=0}^{N-1} (x[n] - s[n])^2 \\ &= (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T (\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) \\ &= \mathbf{x}^T \mathbf{x} - \mathbf{x}^T \mathbf{H}\boldsymbol{\theta} - \boldsymbol{\theta}^T \mathbf{H}^T \mathbf{x} + \boldsymbol{\theta}^T \mathbf{H}^T \mathbf{H} \boldsymbol{\theta} \\ &= \mathbf{x}^T \mathbf{x} - 2\mathbf{x}^T \mathbf{H}\boldsymbol{\theta} + \boldsymbol{\theta}^T \mathbf{H}^T \mathbf{H} \boldsymbol{\theta} \end{aligned}$$

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -2\mathbf{H}^T \mathbf{x} + 2\mathbf{H}^T \mathbf{H} \boldsymbol{\theta}$$

$$\mathbf{H}^T \mathbf{H} \boldsymbol{\theta} = \mathbf{H}^T \mathbf{x} \quad \text{Normal equation}$$

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x}$$

Weighted least squares

$$J(\boldsymbol{\theta}) = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{W} (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})$$

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T \mathbf{W} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{W} \mathbf{x}$$

$$J_{\min} = \mathbf{x}^T (\mathbf{W} - \mathbf{W} \mathbf{H} (\mathbf{H}^T \mathbf{W} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{W}) \mathbf{x}$$

$$\begin{aligned} J_{\min} &= J(\hat{\boldsymbol{\theta}}) \\ &= (\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}})^T (\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}}) \\ &= (\mathbf{x} - \mathbf{H}(\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x})^T (\mathbf{x} - \mathbf{H}(\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x}) \\ &= \mathbf{x}^T (\mathbf{I} - \mathbf{H}(\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T) (\mathbf{I} - \mathbf{H}(\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T) \mathbf{x} \\ &= \mathbf{x}^T (\mathbf{I} - \mathbf{H}(\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T) \mathbf{x}. \\ &= \mathbf{x}^T \mathbf{x} - \mathbf{x}^T \mathbf{H}(\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x} \\ &= \mathbf{x}^T (\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}}). \end{aligned}$$

idempotent matrix,
i.e., $\mathbf{A}^2 = \mathbf{A}$

Geometric Interpretation

$$\mathbf{s} = \mathbf{H}\boldsymbol{\theta}, \quad \boldsymbol{\theta} \in \mathbb{R}^p, \mathbf{H} \in \mathbb{R}^{N \times p}$$

$$\mathbf{s} = \begin{bmatrix} \mathbf{h}_1 & \mathbf{h}_2 & \dots & \mathbf{h}_p \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{bmatrix}$$

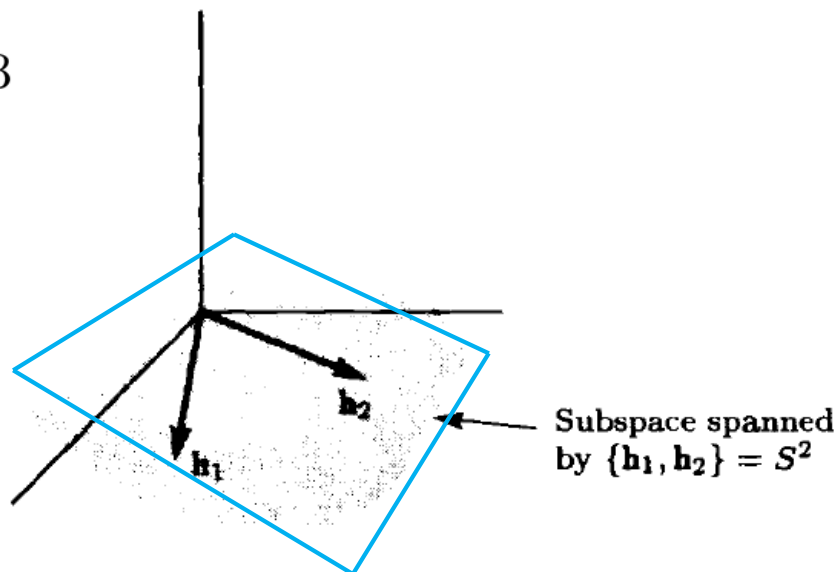
$$= \sum_{i=1}^p \theta_i \mathbf{h}_i \quad ; \text{linear combination of signal vectors}$$

$$J(\boldsymbol{\theta}) = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})$$

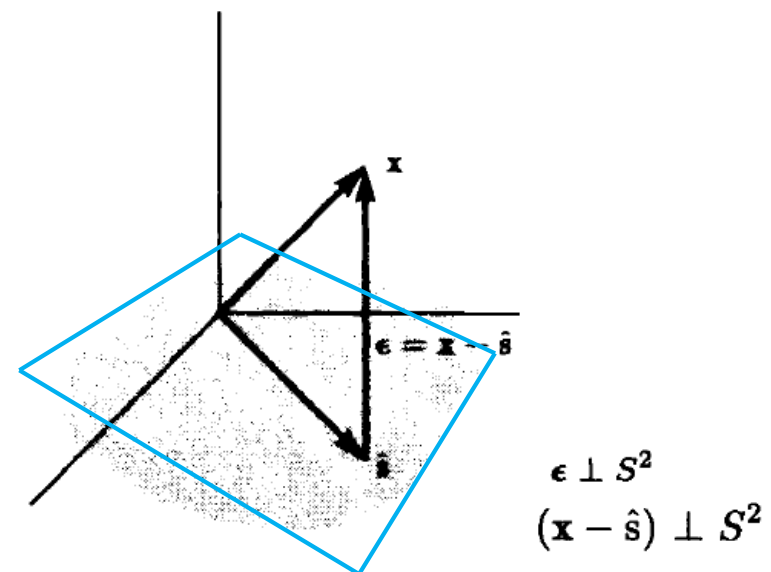
$$\begin{aligned} J(\boldsymbol{\theta}) &= \|\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\|^2 \\ &= \left\| \mathbf{x} - \sum_{i=1}^p \theta_i \mathbf{h}_i \right\|^2 \end{aligned}$$

$\hat{\mathbf{s}} = \mathbf{H}\hat{\boldsymbol{\theta}}$ is the orthogonal projection of $\mathbf{x}$ onto S^2 .

$$\begin{aligned} N &= 3 \\ p &= 2 \end{aligned}$$



(a) Signal subspace

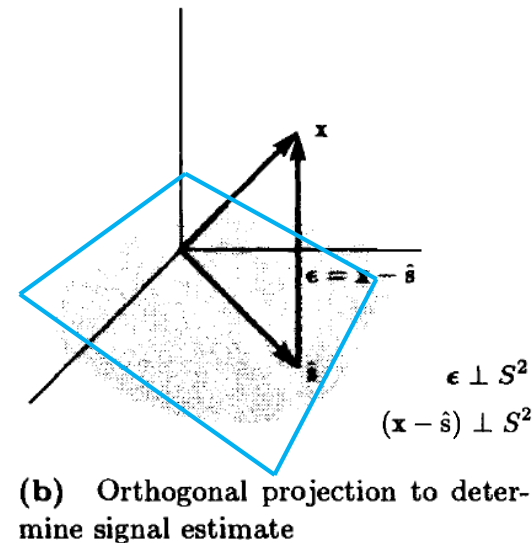
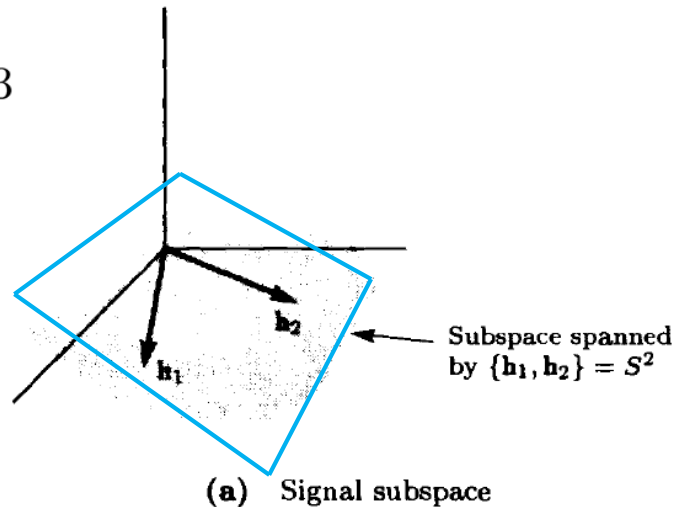


(b) Orthogonal projection to determine signal estimate

Geometric Interpretation

$$N = 3$$

$$p = 2$$



$$\begin{aligned} (\mathbf{x} - \hat{\mathbf{s}}) &\perp \mathbf{h}_1 \\ (\mathbf{x} - \hat{\mathbf{s}}) &\perp \mathbf{h}_2 \end{aligned} \quad \longrightarrow \quad \begin{aligned} (\mathbf{x} - \hat{\mathbf{s}})^T \mathbf{h}_1 &= 0 \\ (\mathbf{x} - \hat{\mathbf{s}})^T \mathbf{h}_2 &= 0 \end{aligned}$$

$$\hat{\mathbf{s}} = \theta_1 \mathbf{h}_1 + \theta_2 \mathbf{h}_2 \quad \longrightarrow \quad \begin{aligned} (\mathbf{x} - \theta_1 \mathbf{h}_1 - \theta_2 \mathbf{h}_2)^T \mathbf{h}_1 &= 0 \\ (\mathbf{x} - \theta_1 \mathbf{h}_1 - \theta_2 \mathbf{h}_2)^T \mathbf{h}_2 &= 0 \end{aligned}$$

$$\begin{aligned} (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{h}_1 &= 0 \\ (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{h}_2 &= 0 \end{aligned}$$

$$(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T [\mathbf{h}_1 \quad \mathbf{h}_2] = \mathbf{0}^T$$

$$(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{H} = \mathbf{0}^T$$

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x}$$

$$\epsilon = \mathbf{x} - \mathbf{H}\boldsymbol{\theta}$$

$$\epsilon^T \mathbf{H} = \mathbf{0}^T$$

**Orthogonality
Principle**

Geometric Interpretation

$$\hat{\mathbf{s}} = \mathbf{H}\hat{\boldsymbol{\theta}} = \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}$$

$$\mathbf{P} = \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T \quad \text{(Orthogonal) projection matrix}$$

$$1. \mathbf{P}^T = \mathbf{P}, \text{ symmetric}$$

$$2. \mathbf{P}^2 = \mathbf{P}, \text{ idempotent.}$$

$$\boldsymbol{\epsilon} = \mathbf{x} - \hat{\mathbf{s}} = (\mathbf{I} - \mathbf{P})\mathbf{x}$$

$$\mathbf{P}^\perp = \mathbf{I} - \mathbf{P} \quad \text{Projection matrix onto the subspace complement to P}$$

$$\begin{aligned} J_{\min} &= J(\hat{\boldsymbol{\theta}}) \\ &= (\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}})^T (\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}}) \\ &= (\mathbf{x} - \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x})^T (\mathbf{x} - \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}) \\ &= \mathbf{x}^T (\mathbf{I} - \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T) (\mathbf{I} - \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T) \mathbf{x} \\ &= \mathbf{x}^T (\mathbf{I} - \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T) \mathbf{x}. \\ &= \mathbf{x}^T (\mathbf{I} - \mathbf{P})\mathbf{x} \\ &= \mathbf{x}^T \mathbf{P}^\perp \mathbf{x} \\ &= \mathbf{x}^T \mathbf{P}^{\perp T} \mathbf{P}^\perp \mathbf{x} \\ &= \|\mathbf{P}^\perp \mathbf{x}\|^2 \\ &= \|\boldsymbol{\epsilon}\|^2. \end{aligned}$$

Sequential Least Squares

$$x[n] = A + w[n], \quad n = 0, 1, \dots, N-1, \quad \text{where } w[n] \sim \mathcal{N}(0, \sigma^2)$$

$$\begin{aligned} \text{LSE: } \hat{A}[N-1] &= \frac{1}{N} \sum_{n=0}^{N-1} x[n] & \hat{A}[N] &= \frac{1}{N+1} \sum_{n=0}^N x[n] \\ & & &= \frac{1}{N+1} \left(\sum_{n=0}^{N-1} x[n] + x[N] \right) \\ & & &= \frac{N}{N+1} \hat{A}[N-1] + \frac{1}{N+1} x[N] \end{aligned}$$

$$\hat{A}[N] = \hat{A}[N-1] + \overbrace{\frac{1}{N+1} (x[N] - \hat{A}[N-1])}^{\text{Correction}} \quad J_{\min}[N] = J_{\min}[N-1] + \frac{N}{N+1} (x[N] - \hat{A}[N-1])^2$$

$$x[n] = A + w[n], \quad n = 0, 1, \dots, N-1, \quad \text{where } w[n] \sim \mathcal{N}(0, \sigma_n^2)$$

Estimator Update:

$$\hat{A}[N] = \hat{A}[N-1] + K[N](x[N] - \hat{A}[N-1]) \quad (8.40)$$

where

$$K[N] = \frac{\text{var}(\hat{A}[N-1])}{\text{var}(\hat{A}[N-1]) + \sigma_N^2}. \quad (8.41)$$

Variance Update:

$$\text{var}(\hat{A}[N]) = (1 - K[N])\text{var}(\hat{A}[N-1]). \quad (8.42)$$

Sequential Least Squares for Vector Parameters

$$J = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{C}^{-1} (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})$$

$\mathbf{C}$: covariance matrix of the zero mean noise

$$\mathbf{C}[n] = \text{diag}(\sigma_0^2, \sigma_1^2, \dots, \sigma_n^2)$$

Assumption: $\mathbf{C}$ is diagonal.

$$\mathbf{H}[n] = \begin{bmatrix} \mathbf{H}[n-1] \\ \mathbf{h}^T[n] \end{bmatrix} = \begin{bmatrix} n \times p \\ 1 \times p \end{bmatrix}$$

$$\mathbf{x}[n] = [x[0] \ x[1] \ \dots \ x[n]]^T$$

Estimator Update:

$$\hat{\boldsymbol{\theta}}[n] = \hat{\boldsymbol{\theta}}[n-1] + \mathbf{K}[n] \left(x[n] - \mathbf{h}^T[n] \hat{\boldsymbol{\theta}}[n-1] \right) \quad (8.46)$$

where

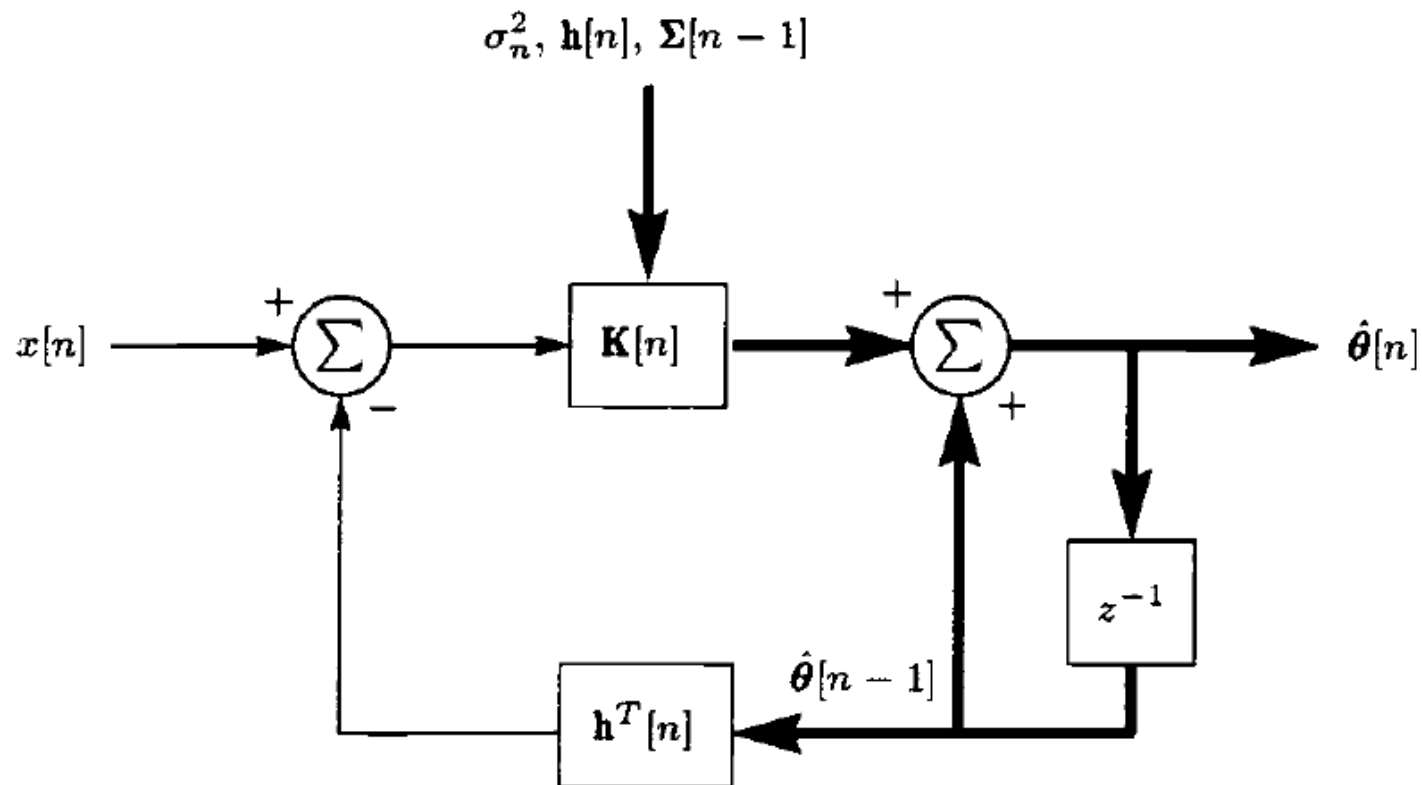
$$\mathbf{K}[n] = \frac{\boldsymbol{\Sigma}[n-1] \mathbf{h}[n]}{\sigma_n^2 + \mathbf{h}^T[n] \boldsymbol{\Sigma}[n-1] \mathbf{h}[n]}. \quad (8.47)$$

Covariance Update:

$$\boldsymbol{\Sigma}[n] = (\mathbf{I} - \mathbf{K}[n] \mathbf{h}^T[n]) \boldsymbol{\Sigma}[n-1]. \quad (8.48)$$

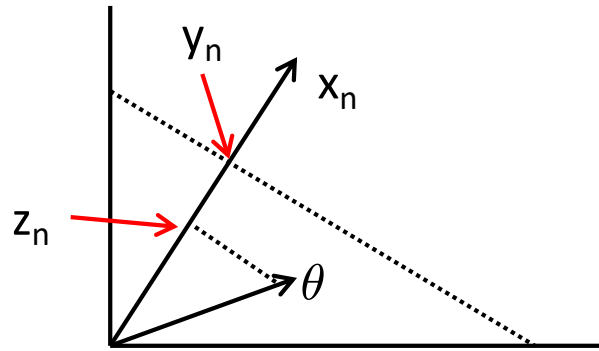
Sequential Least Squares for Vector Parameters

$$\hat{\boldsymbol{\theta}}[n] = \hat{\boldsymbol{\theta}}[n-1] + \mathbf{K}[n] \left(x[n] - \mathbf{h}^T[n] \hat{\boldsymbol{\theta}}[n-1] \right)$$



Least Mean Squares (LMS)

$$y_n = \theta^T x_n + w_n$$



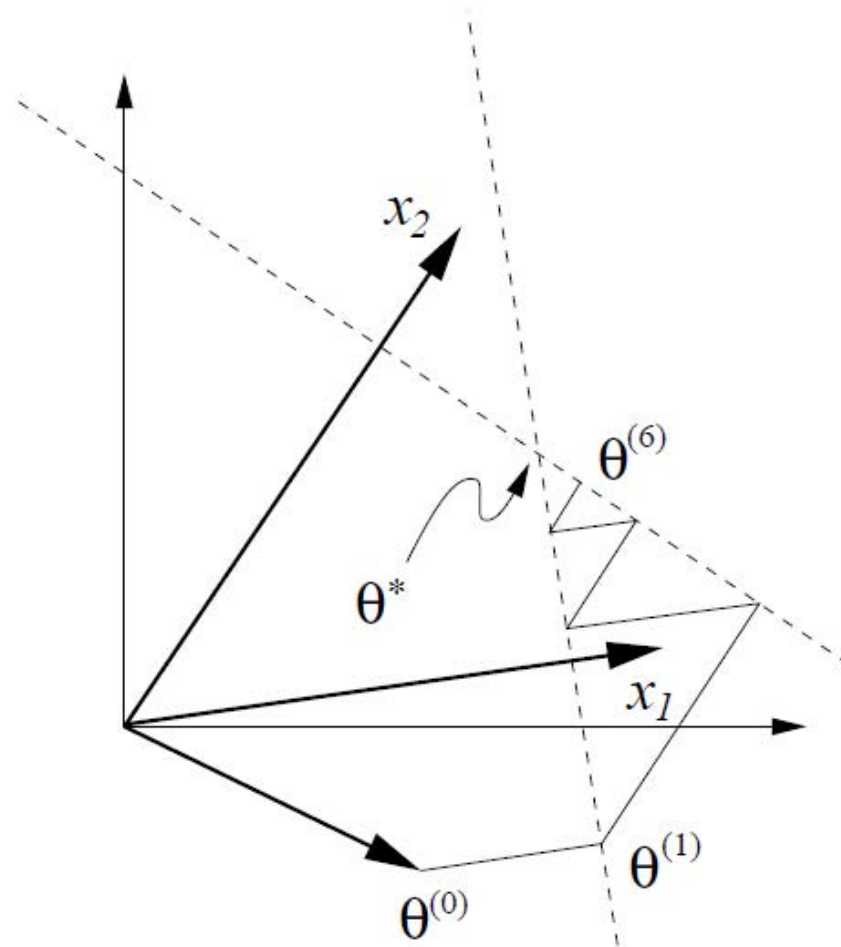
- Assume x_n is a unit vector.
- $z_n = \theta^T x_n$, projection of θ onto x_n .
- error = $(y_n - z_n)$.
- Idea: Update θ to minimize the error. Move along x_n -direction with a small amount, e.g., $(y_n - z_n)$.
- $\theta^{(t+1)} = \theta^{(t)} + (y_n - z_n)x_n = \theta^{(t)} + (y_n - \theta^{(t)T} x_n)x_n$

If x_n is not a unit vector,

- $z_n = \frac{\theta^T x_n}{\|x_n\|}$.
- error = $\frac{y_n}{\|x_n\|} - z_n = \frac{y_n - \theta^T x_n}{\|x_n\|}$.
- unit vector along x_n is $\frac{x_n}{\|x_n\|}$.
- $\theta^{(t+1)} = \theta^{(t)} + \left(\frac{y_n}{\|x_n\|} - z_n\right) \frac{x_n}{\|x_n\|} = \theta^{(t)} + \frac{1}{\|x_n\|^2} (y_n - \theta^{(t)T} x_n) x_n$.
- LMS: $\theta^{(t+1)} = \theta^{(t)} + \mu(y_n - \theta^{(t)T} x_n)x_n$, for some small μ .

Least Mean Squares (LMS)

Multiple data points



Steepest Descent and Stochastic Gradient Algorithms

$$y_n = \theta^T x_n + w_n$$

- Cost function: $J(\theta) = \frac{1}{2} \sum_{n=1}^N (y_n - \theta^T x_n)^2$.
- $\nabla_{\theta} J = - \sum_{n=1}^N (y_n - \theta^T x_n) x_n$.
- Steepest descent:

$$\begin{aligned} \theta^{(t+1)} &= \theta^{(t)} - \mu \nabla_{\theta} J \\ &= \theta^{(t)} + \mu \sum_{n=1}^N (y_n - \theta^{(t)T} x_n) x_n \end{aligned}$$

- Stochastic gradient (=LMS):

$$\theta^{(t+1)} = \theta^{(t)} + \mu (y_n - \theta^{(t)T} x_n) x_n$$

Convergence Analysis

$$\begin{aligned}
 \theta^{(t+1)} &= \theta^{(t)} + \mu \sum_{n=1}^N (y_n - \theta^{(t)T} x_n) x_n \\
 &= \theta^{(t)} + \mu \sum_{n=1}^N x_n y_n - \mu \sum_{n=1}^N (x_n x_n^T) \theta^{(t)} \\
 &= \theta^{(t)} + \mu X^T y - \mu X^T X \theta^{(t)} \\
 &= (I - \mu X^T X) \theta^{(t)} + \mu X^T y.
 \end{aligned}$$

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_N^T \end{bmatrix}$$

$$\begin{aligned}
 \theta^{(t+1)} &= (I - \mu X^T X) \theta^{(t)} + \mu X^T y \\
 &= (I - \mu X^T X) \left[(I - \mu X^T X) \theta^{(t-1)} + \mu X^T y \right] + \mu X^T y \\
 &= (I - \mu X^T X)^{t+1} \theta^{(0)} + \mu \sum_{i=0}^t (I - \mu X^T X)^i X^T y.
 \end{aligned}$$

Assume that $\lim_{t \rightarrow \infty} (I - \mu X^T X)^t = 0$. Then

$$\begin{aligned}
 \theta^{(\infty)} &= \mu \sum_{i=0}^{\infty} (I - \mu X^T X)^i X^T y \\
 &= \mu (\mu X^T X)^{-1} X^T y \\
 &= (X^T X)^{-1} X^T y,
 \end{aligned}$$

Neumann series

$\lim_{t \rightarrow \infty} (I - \mu X^T X)^t = 0$ if

$$\begin{aligned}
 0 &< \lambda_{\max}(\mu X^T X) < 2 \\
 0 &< \mu < 2 / \lambda_{\max}(X^T X).
 \end{aligned}$$

Constrained Least Squares

Least square cost function: $J(\boldsymbol{\theta}) = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})$

Constraint: $\mathbf{A}\boldsymbol{\theta} = \mathbf{b}$

Using a Lagrangian multiplier:

$$J_c = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) + \boldsymbol{\lambda}^T(\mathbf{A}\boldsymbol{\theta} - \mathbf{b})$$

$$J_c = \mathbf{x}^T \mathbf{x} - 2\boldsymbol{\theta}^T \mathbf{H}^T \mathbf{x} + \boldsymbol{\theta}^T \mathbf{H}^T \mathbf{H} \boldsymbol{\theta} + \boldsymbol{\lambda}^T \mathbf{A} \boldsymbol{\theta} - \boldsymbol{\lambda}^T \mathbf{b}$$

$$\frac{\partial J_c}{\partial \boldsymbol{\theta}} = -2\mathbf{H}^T \mathbf{x} + 2\mathbf{H}^T \mathbf{H} \boldsymbol{\theta} + \mathbf{A}^T \boldsymbol{\lambda}$$

$$\begin{aligned}\hat{\boldsymbol{\theta}}_c &= (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x} - \frac{1}{2} (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{A}^T \boldsymbol{\lambda} \\ &= \hat{\boldsymbol{\theta}} - (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{A}^T \frac{\boldsymbol{\lambda}}{2}\end{aligned}$$

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x}$$

$$\mathbf{A}\boldsymbol{\theta}_c = \mathbf{A}\hat{\boldsymbol{\theta}} - \mathbf{A}(\mathbf{H}^T \mathbf{H})^{-1} \mathbf{A}^T \frac{\boldsymbol{\lambda}}{2} = \mathbf{b} \quad \frac{\boldsymbol{\lambda}}{2} = [\mathbf{A}(\mathbf{H}^T \mathbf{H})^{-1} \mathbf{A}^T]^{-1} (\mathbf{A}\hat{\boldsymbol{\theta}} - \mathbf{b})$$

$$\hat{\boldsymbol{\theta}}_c = \hat{\boldsymbol{\theta}} - (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{A}^T [\mathbf{A}(\mathbf{H}^T \mathbf{H})^{-1} \mathbf{A}^T]^{-1} (\mathbf{A}\hat{\boldsymbol{\theta}} - \mathbf{b})$$

Nonlinear Least Squares

$$J = (\mathbf{x} - \mathbf{s}(\boldsymbol{\theta}))^T (\mathbf{x} - \mathbf{s}(\boldsymbol{\theta})) \quad ; \text{ s is not linear in } \theta$$

1. Transformation of parameters

- Let $\alpha = g(\theta)$ such that g^{-1} exists.
- Find g such that $s(\theta(\alpha)) = s(g^{-1}(\alpha)) = H\alpha$.
- Then $\hat{\theta}_{LSE} = g^{-1}(\hat{\alpha}_{LSE})$ where $\hat{\alpha}_{LSE} = (H^T H)^{-1} H^T x$.

2. Separability of parameters

- Separable signal model: $s = H(\alpha)\beta$, where $\theta = [\alpha, \beta]^T$. Linear in β but nonlinear in α .
- Minimizing $J(\alpha, \beta) = (x - H(\alpha)\beta)^T (x - H(\alpha)\beta)$, we get $\hat{\beta} = (H^T(\alpha)H(\alpha))^{-1} H^T(\alpha)x$.
- Resulting least square error is

$$J(\alpha, \hat{\beta}) = x^T \left(I - H(\alpha)(H^T(\alpha)H(\alpha))^{-1} H^T(\alpha) \right) x.$$

- Now we maximize the following with respect to α .

$$x^T H(\alpha)(H^T(\alpha)H(\alpha))^{-1} H^T(\alpha)x$$

Nonlinear Least Squares

$$J = (\mathbf{x} - \mathbf{s}(\boldsymbol{\theta}))^T (\mathbf{x} - \mathbf{s}(\boldsymbol{\theta})) \quad ; \text{ s is not linear in } \theta$$

Necessary conditions for optimality:

$$\frac{\partial J}{\partial \theta_j} = -2 \sum_{i=0}^{N-1} (x[i] - s[i]) \frac{\partial s[i]}{\partial \theta_j} = 0 \quad \text{for } j = 1, 2, \dots, p.$$

$$\sum_{i=0}^{N-1} (x[i] - s[i]) \left[\frac{\partial \mathbf{s}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right]_{ij} = 0 \quad j = 1, 2, \dots, p \quad \left[\frac{\partial \mathbf{s}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right]_{ij} = \frac{\partial s[i]}{\partial \theta_j}$$

$$\frac{\partial \mathbf{s}(\boldsymbol{\theta})^T}{\partial \boldsymbol{\theta}} (\mathbf{x} - \mathbf{s}(\boldsymbol{\theta})) = 0$$

Newton-Raphson method:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

Newton-Raphson iteration:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \left(\frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^{-1} \mathbf{g}(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_k} \quad \text{where} \quad \mathbf{g}(\boldsymbol{\theta}) = \frac{\partial \mathbf{s}(\boldsymbol{\theta})^T}{\partial \boldsymbol{\theta}} (\mathbf{x} - \mathbf{s}(\boldsymbol{\theta}))$$

Nonlinear Least Squares

$$J = (\mathbf{x} - \mathbf{s}(\boldsymbol{\theta}))^T (\mathbf{x} - \mathbf{s}(\boldsymbol{\theta})) \quad ; \text{ s is not linear in } \boldsymbol{\theta}$$

Newton-Raphson method

- using the necessary condition of optimality

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \left(\mathbf{H}^T(\boldsymbol{\theta}_k) \mathbf{H}(\boldsymbol{\theta}_k) - \sum_{n=0}^{N-1} \mathbf{G}_n(\boldsymbol{\theta}_k) (x[n] - [\mathbf{s}(\boldsymbol{\theta}_k)]_n) \right)^{-1} \cdot \mathbf{H}^T(\boldsymbol{\theta}_k) (\mathbf{x} - \mathbf{s}(\boldsymbol{\theta}_k))$$
$$[\mathbf{H}(\boldsymbol{\theta})]_{ij} = \left[\frac{\partial \mathbf{s}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right]_{ij} = \frac{\partial s[i]}{\partial \theta_j}$$
$$[\mathbf{G}_n(\boldsymbol{\theta})]_{ij} = \frac{\partial^2 s[n]}{\partial \theta_i \partial \theta_j}$$

Gauss-Newton method

- using the linearization of the signal model

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + (\mathbf{H}^T(\boldsymbol{\theta}_k) \mathbf{H}(\boldsymbol{\theta}_k))^{-1} \mathbf{H}^T(\boldsymbol{\theta}_k) (\mathbf{x} - \mathbf{s}(\boldsymbol{\theta}_k))$$