Least Squares

1. What is Least Squares Estimator

least squares는 주어진 데이터 $x[n]$과 노이즈가 없다고 가정한 신호 $s[n]$사이 차이 제곱을 최소화 시키는 방법이다. 따라서 parameter θ에 대한 LSE는 $s[n]$을 $x[n]$에 최대한 가까이 조정하는 역할을 한다. 그리고 해당 추정값은 다음과 같이 구한다.

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

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

Ex) DC Level Signal

$s[n] = A$일 때,

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

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

$$\sum_{n=0}^{N-1} x[n] = NA$$

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

따라서 위 경우, 노이즈가 가우시안 분포를 따른다면 MVUE가 된다.

2. Linear Least Squares

1) scalar case

$s[n] = \theta h[n]$일 때,

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

을 만족하고,

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

이 된다.

따라서 LSE는

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

이 된다.

위 LSE를 오차식에 대입하면,

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

따라서 최소 오차 $J_{\min}$은 $\sum_{n=0}^{N-1} x^2[n]$보다 항상 작거나 같다.

$$0 \le J_{\min} \le \sum_{n=0}^{N-1} x^2[n]$$

2) Vector Case

추정하고자 하는 파라미터가 벡터 파라미터 $\theta \in R^{p \times 1}$인 경우,

$$s = H\theta, \ (s : n \times 1, H : n \times p)$$

이며, 오차는

$$
\begin{aligned}
J(\theta) &= (x - H\theta)^T (x - H\theta) \\
&= x^T x - x^T H\theta - \theta^T H^T x + \theta^T H^T H\theta \\
&= x^T x - 2x^T H\theta + \theta^T H^T H\theta
\end{aligned}
$$

가 된다. 따라서 이를 미분하면

$$\frac{\partial J(\theta)}{\partial \theta} = -2H^T x + 2H^T H\theta = 0$$

이 되어 LSE는

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

가 된다.

이 LSE를 오차식에 넣어 최소 오차를 계산해보면,

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

가 된다. 위 식에서 $I - H(H^T H)^{-1} H^T$은 $A^2 = A$를 만족하는 멱동 행렬(Idempotent matrix)이므로 위와 같이 계산된다.

3) Vector Weighted Case

LSE Criterion에 가중치가 곱해져 있는 경우이다. 여기서는 LSE 식 가운데에 positive definite한 가중치 W가 곱해져 있으며 일반적으로 weighted least squares라고 한다.

$$J(\theta) = (x - H\theta)^T W(x - H\theta)$$

여기서 W가 대각행렬이고 각 성분이 $w_i > 0$이라면 다음과 같이 쓸 수 있다.

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

여기서 $w_n = 1/\sigma^2$라면, 미분 후 0인 값을 찾았을 때 LSE는

$$\hat{A} = \frac{\sum_{n=0}^{N-1} w[n]x[n]}{\sum_{n=0}^{N-1} w[n]} = \frac{\sum_{n=0}^{N-1} \frac{x[n]}{\sigma_n^2}}{\sum_{n=0}^{N-1} \frac{1}{\sigma_n^2}}$$

가 될 것이다. 또한 일반적인 벡터의 형태에서는

$$\hat{\theta} = (H^T WH)^{-1} H^T Wx$$

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

가 되고, 이는 $W = C^{-1}$일 때의 BLUE와 동일하다.

3. Geometrical Interpretation

앞서 풀이한 LS를 기하학적 관점에서 풀어보자.

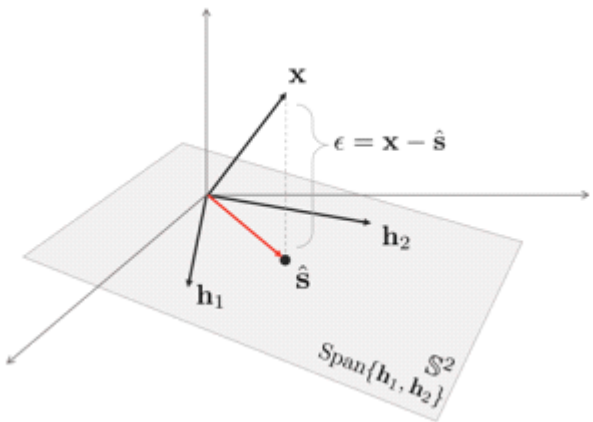
일반적인 신호 모델 $s = H\theta$는 다음과 같다.

$$s = [h_1 h_2 \dots h_p] \begin{bmatrix} \theta_1 \\ \theta_2 \\ \dots \\ \theta_3 \end{bmatrix} = \sum_{i=1}^{p} \theta_i h_i$$

그리고 오차식은 다음과 같이 쓸 수 있다.

$$J(\theta) = \|x - \sum_{i=1}^{p} \theta_i h_i\|^2$$

H를 full rank라고 가정한다면, $\sum_{i=1}^{p} \theta_i h_i$는 서로 독립된 H의 열벡터들이 span하는 subspace라고 볼 수 있다. 그리고 해당 오차의 최소값은 x로부터 subspace까지의 최소 거리이니 x에서 subspace로 내린 projection에 해당된다.



따라서

$$(x-\hat{s}) \perp S^2$$

이며

$$(x-\hat{s}) \perp h_1$$
$$(x-\hat{s}) \perp h_2$$

이고, 내적의 성질에 따라

$$(x-\hat{s})^T h_1 = 0$$
$$(x-\hat{s})^T h_2 = 0$$

가 성립한다. 여기서 $\hat{s} = \theta_1 h_1 + \theta_2 h_2$이므로,

$$(x-\theta_1 h_1 - \theta_2 h_2)^T h_1 = 0$$
$$(x-\theta_1 h_1 - \theta_2 h_2)^T h_2 = 0$$

이고, 행렬 형태로 나타내면

$$(x-H\theta)^T h_1 = 0$$
$$(x-H\theta)^T h_2 = 0$$

이다. 두 식을 합치게 되면

$$(x-H\theta)^T H = 0^T$$

이므로 LES는

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

가 된다.

이에 따라 신호모델은 다음과 같다.

$$\begin{aligned} \hat{s} &= H\hat{\theta} \\ &= H(H^T H)^{-1} H^T x \\ &= Px \end{aligned}$$

이 때 projection matrix인 $P = H(H^T H)^{-1} H^T$는 대칭 행렬이며 멱동 행렬이다.

이를 이용해 오차를 표현하면

$$\epsilon = x - \hat{s} = x - Px = (I-P)x = P^{\perp} x$$

가 된다. 여기서 $P^{\perp} = I - P$ 역시 P와 동일한 성질을 가진다.

최종적으로 LS 오차값은 다음과 같다.

$$J_{\min} = ||P^{\perp} x||^2 = ||\epsilon||^2$$

4. Sequential Least Squares

$N-1$개의 데이터에 대한 추정값이 있을 때, N번째 데이터가 들어왔을 때 추정값을 구하는 방법이다.

1) scalar case

$x[n] = A + w[n],\ n = 0, 1, ..., N-1,\ w[n] \sim N(0, \sigma^2)$

위 문제에서 LSE는

$$\hat{A}[N-1] = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$$

와 같다. 여기서 새로운 데이터 $x[N]$이 들어왔을 때,

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

와 같이 LSE가 수정된다. 따라서

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

가 성립한다.

또한, 최소 오차값에 대해서도

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

$$J_{\min}[N] = \sum_{n=0}^{N}(x[n] - \hat{A}[N])^2$$

이므로 여기에 위의 추정값을 넣어 정리하면,

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

가 된다. 데이터의 개수가 늘수록 오차값이 커짐을 확인할 수 있다.

2) Weighted Case

weighted case에 대해서의 추정값은 앞서 구한 것과 같이 다음과 같다.

$$\hat{A}[N-1] = \frac{\sum_{n=0}^{N-1}\frac{x[n]}{\sigma_n^2}}{\sum_{n=0}^{N-1}\frac{1}{\sigma_n^2}}$$

따라서 새로운 데이터 $x[N]$이 들어왔을 때, LSE는 다음과 같다.

$$
\begin{aligned}
\hat{A}[N] &= \frac{\sum_{n=0}^{N} \frac{x[n]}{\sigma_n^2}}{\sum_{n=0}^{N} \frac{1}{\sigma_n^2}} \\
&= \frac{\sum_{n=0}^{N-1} \frac{x[n]}{\sigma_n^2} + \frac{x[N]}{\sigma_N^2}}{\sum_{n=0}^{N} \frac{1}{\sigma_n^2}} \\
&= \frac{(\sum_{n=0}^{N-1} \frac{1}{\sigma_n^2})\hat{A}[N-1]}{\sum_{n=0}^{N} \frac{1}{\sigma_n^2}} + \frac{\frac{x[N]}{\sigma_N^2}}{\sum_{n=0}^{N} \frac{1}{\sigma_n^2}} \\
&= \hat{A}[N-1] - \frac{\frac{1}{\sigma_N^2}\hat{A}[N-1]}{\sum_{n=0}^{N} \frac{1}{\sigma_n^2}} + \frac{\frac{x[N]}{\sigma_N^2}}{\sum_{n=0}^{N} \frac{1}{\sigma_n^2}} \\
&= \hat{A}[N-1] + \frac{\frac{1}{\sigma_N^2}}{\sum_{n=0}^{N} \frac{1}{\sigma_n^2}}(x[N] - \hat{A}[N-1])
\end{aligned}
$$

따라서 정리하면 다음과 같다.

$$
\hat{A}[N] = \hat{A}[N-1] + \frac{\frac{1}{\sigma_N^2}}{\sum_{n=0}^{N} \frac{1}{\sigma_n^2}}(x[N] - \hat{A}[N-1])
$$

여기서 gain factor $K[N] = \frac{\frac{1}{\sigma_N^2}}{\sum_{n=0}^{N} \frac{1}{\sigma_n^2}} = \frac{\frac{1}{\sigma_N^2}}{\frac{1}{\sigma_N^2} + \sum_{n=0}^{N-1} \frac{1}{\sigma_n^2}} = \frac{var(\hat{A}[N-1])}{var(\hat{A}[N-1]) + \sigma_N^2}$ 를 정의하면 분산도 다음과 같이 정리된다.

$$
\begin{aligned}
var(\hat{A}[N]) &= \frac{1}{\sum_{n=0}^{N}\frac{1}{\sigma_n^2}} \\
&= \frac{1}{\sum_{n=0}^{N-1}\frac{1}{\sigma_n^2}+\frac{1}{\sigma_N^2}} \\
&= \frac{1}{\frac{1}{var(\hat{A}[N-1])}+\frac{1}{\sigma_N^2}} \\
&= \frac{var(\hat{A}[N-1])\sigma_N^2}{var(\hat{A}[N-1])+\sigma_N^2} \\
&= (1-\frac{var(\hat{A}[N-1])}{var(\hat{A}[N-1])+\sigma_N^2})var(\hat{A}[N-1]) \\
&= (1-K[N])var(\hat{A}[N-1])
\end{aligned}
$$

3) Vector Case

vector parameter인 경우에 가중치 행렬이 $W=C^{-1}$일 때 앞서 구했듯 LSE와 공분산은 다음과 같다.

$$
\begin{aligned}
J(\theta) &= (x-H\theta)^T C^{-1}(x-H\theta) \\
\hat{\theta} &= (H^T C^{-1} H)^{-1} H^T C^{-1} x \\
C_{\hat{\theta}} &= (H^T C^{-1} H)^{-1}
\end{aligned}
$$

이 때 공분산 행렬이 대각행렬이 아니면 재귀적으로 계산이 불가능하므로 대각행렬이라 가정하면

$$
\begin{aligned}
C[n] &= diag(\sigma_0^2, \sigma_1^2, \sigma_2^2, ..., \sigma_n^2) \\
H[n] &= \begin{bmatrix} H[n-1] \\ h^T[n] \end{bmatrix} = \begin{bmatrix} n\times p \\ 1\times p \end{bmatrix}
\end{aligned}
$$

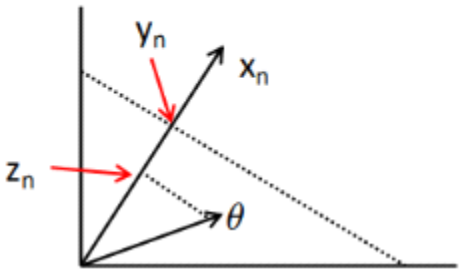
$x[n] = [x[0], x[1], ..., x[n]]^T$

최종적인 sequential LSE는 다음과 같다.

$$
\begin{aligned}
\hat{\theta}[n] &= \hat{\theta}[n-1] + K[n](x[n]-h^T[n]\hat{\theta}[n-1]) \\
\text{where } K[n] &= \frac{\Sigma[n-1]h[n]}{\sigma_n^2 + h^T[n]\Sigma[n-1]h[n]} \\
\Sigma[n] &= (1-K[n])\Sigma[n-1]
\end{aligned}
$$

5. Least Mean Squares (LMS)

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



x_n이 unit vector가 아닐 때,

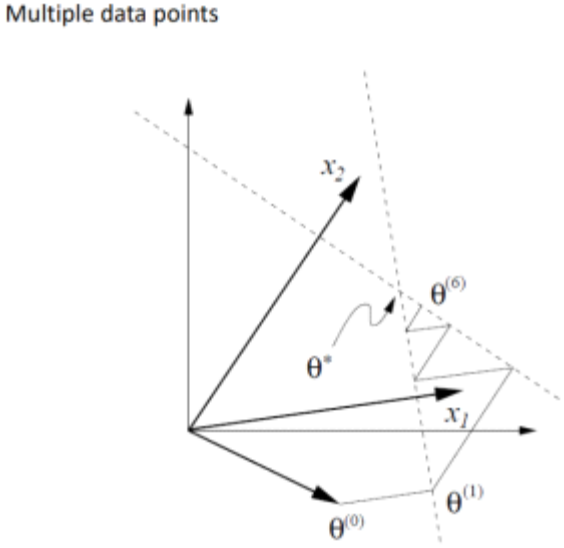
θ의 x_n으로의 projection은 $z_n = \frac{\theta^T x_n}{\|x_n\|}$이다.

따라서 error는 $\frac{y_n}{\|x_n\|} - z_n = \frac{y_n - \theta^T x_n}{\|x_n\|}$로 나타낼 수 있다.

따라서 $\theta^{(t+1)} = \theta^{(t)} + (\frac{y_n}{\|x_n\|} - z_n)\frac{x_n}{\|x_n\|} = \theta^{(t)} + \frac{1}{\|x_n\|^2}(y_n - \theta^{(t)^T} x_n)x_n$이 되고,

어떤 작은 μ에 대해 $\theta^{(t+1)} = \theta^{(t)} + \mu(y_n - \theta^{(t)^T} x_n)x_n$가 성립한다고 말할 수 있다.

multiple data point들에 대해서는 다음 그림과 같이 추정될 수 있다.



유사하게, Steepest Descent algorithm은 다음과 같이 유도할 수 있다.

우선 cost function은 LMS와 유사하게 다음과 같이 정의한다.

$$J(\theta) = \frac{1}{2}\sum_{n=0}^{N}(y_n - \theta^T x_n)^2$$

그리고 그 gradient는 다음과 같다.

$$\nabla_\theta J = -\sum_{n=1}^{N}(y_n - \theta^T x_n)x_n$$

따라서 다음과 같은 점화식이 성립한다.

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

위 식을 전개하면,

$$\begin{aligned}
\theta^{(t+1)} &= \theta^{(t)} + \mu\sum_{n=1}^{N}(y_n - \theta^{(t)^T}x_n)x_n \\
&= \theta^{(t)} + \mu\sum_{n=1}^{N}x_n y_n - \mu\sum_{n=1}^{N}(x_n x_n)\theta^{(t)} \\
&= \theta^{(t)} + \mu X^T y - \mu X^T X\theta^{(t)} \\
&= (I - \mu X^T X)\theta^{(t)} + \mu X^T y \\
&= (I - \mu X^T X)[(I - \mu X^T X)\theta^{(t-1)} + \mu X^T y] + \mu X^T y \\
&= (I - \mu X^T X)^{t+1}\theta^{(0)} + \mu\sum_{i=0}^{t}(I - \mu X^T X)^i X^T y
\end{aligned}$$

와 같이 전개된다. 여기서 $\lim_{t\to\infty}(I - \mu X^T X)^t = 0$이라고 가정한다면

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

와 같이 추정값을 구할 수 있다.

참고로 $\lim_{t\to\infty}(I - \mu X^T X)^t = 0$인 경우는 $0 < \mu < \dfrac{2}{\lambda_{\max}(X^T X)}$인 경우이다.

6. Constrained Least Squares

LS 문제에서 제약 조건이 있는 경우이다. 선형 제약 조건이 있는 경우는 쉽게 해결할 수 있다.

우선 오차값은 다음과 같이 정의된다.

$$J_c = (x - H\theta)^T(x - H\theta)$$

여기서 $A\theta = b$라는 제약 조건이 있다고 가정하자. 여기서 lagrangian multiplier λ를 넣어 LS 오차를 다시 쓰면 다음과 같다.

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

그리고 최소를 찾기 위해 이를 미분한다.

$$\frac{\partial J_c}{\partial\theta} = -2H^T x + 2H^T H\theta + A^T\lambda = 0$$

위 식에 따라 LS 추정값은 다음과 같다.

$$\hat{\theta_c} = (H^T H)^{-1}H^T x - \frac{1}{2}(H^T H)^{-1}A^T\lambda = \hat{\theta} - \frac{1}{2}(H^T H)^{-1}A^T\lambda$$

위 식에서 적절한 lagrangian multiplier를 찾기 위해 양변에 A를 곱하면

$A\hat{\theta_c} = A\hat{\theta} - A(H^T H)A^T\dfrac{\lambda}{2} = b$가 되어

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

로 계산된다.

따라서 구하고자 하는 추정값은

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

로 계산된다.

7. Nonlinear Least Squares

일반적인 경우의 Nonlinear LS 문제는 풀기 어렵거나 매우 복잡하다. 따라서 다음과 같은 경우에 대한 풀이법을 살펴보자.

1) transformation of parameters

우선 LS 오차식은 다음과 같다.

$$J = (x - s(\theta))^T (x - s(\theta))$$

그리고 여기서 우리는 다음을 만족하는 $\alpha = g(\theta)$를 찾는다.

$$s(\theta(\alpha)) = s(g^{-1}(\alpha)) = H\alpha$$

그 다음부터는 α에 대한 LS 문제를 푼 후 다음과 같이 비선형 LSE를 구할 수 있다.

$$\hat{\theta} = g^{-1}(\alpha)$$

2) separability of parameters

이번 방법은 비선형 함수 s를 선형인 성분과 비선형인 성분으로 분리하는 방법이다.

추정하는 파라미터 θ가 다음과 같다고 가정하자.

$$\theta = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} (p-q) \times 1 \\ q \times 1 \end{bmatrix}$$

여기서 α는 비선형이고 β는 선형이다. 그러면 오차식은 다음과 같다.

$$J(\alpha, \beta) = (x - H(\alpha)\beta)^T (x - H(\alpha)\beta)$$

먼저 위 식을 바탕으로 β에 대한 LSE를 구하면,

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

가 된다. 위 LSE를 처음 오차식에 대입하면,

$$J(\alpha, \hat{\beta}) = x^T [I - H(\alpha)(H^T(\alpha) H(\alpha))^{-1} H^T(\alpha)] x$$

결론적으로 비선형 LSE를 찾는 문제는 다음 식을 최대화하는 문제가 된다.

$$J(\alpha, \hat{\beta}) \propto arg \max_{\alpha} [H(\alpha)(H^T(\alpha) H(\alpha))^{-1} H^T(\alpha) x]$$

3) 그 외

일반적인 LS 오차식은 다음과 같다.

$J = (x - s(\theta))^T (x - s(\theta))$

최솟값을 찾기 위해 위 식을 미분하면 다음과 같다.

$$\frac{\partial J}{\partial \theta_j} = -2 \sum_{i=0}^{N-1} (x[i] - s[i]) \frac{\partial s[i]}{\partial \theta_j} = 0$$

위 식은 다음 식과 동일하다.

$$\sum_{i=0}^{N-1}(x[i]-s[i])[\frac{\partial s(\theta)}{\partial \theta}]_{ij}=0$$

결국 다음 식을 만족하는 θ를 푸는 문제가 된다.

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

위 문제는 Newton-Raphson method나 Gauss-Newton method로 근사값을 구할 수 있다.