- linear regression models

. linear estimator: the output for the kth input $\underline{x}_k$ is given by

$$y_k = y(\underline{x}_k) = w_0 + \sum_{i=1}^{d} w_i x_{ik}.$$

Let

$\underline{x}_k = [1, x_{1k}, \cdots, x_{dk}]^T$ and $\underline{w} = [w_0, w_1, \cdots, w_d]^T$.

Then,

$$y_k = \underline{w}^T \underline{x}_k.$$

Let d_k be the desired response for the kth input pattern.

Then, the error for the kth pattern is

$$\epsilon_k = d_k - y_k = d_k - \underline{w}^T \underline{x}_k.$$

the error square:

$$\epsilon_k^2 = d_k^2 + \underline{w}^T \underline{x}_k \underline{x}_k^T \underline{w} - 2 d_k \underline{x}_k^T \underline{w}.$$

the mean square error (MSE):

$$E[\epsilon_k^2] = E[d_k^2] + \underline{w}^T E[\underline{x}_k \underline{x}_k^T] \underline{w} - 2E[d_k \underline{x}_k^T] \underline{w}$$

Let

$R = E[\underline{x}_k \underline{x}_k^T]$ (input correlation matrix) and

$P = E[d_k \underline{x}_k]$.

Then, MSE becomes

$E[\epsilon_k^2] = E[d_k^2] + \underline{w}^T R \underline{w} - 2P^T \underline{w}$. (quadratic form)

the derivative of MSE with respect to $\underline{w}$:

$$\nabla E[\epsilon_k^2] = \frac{\partial E}{\partial \underline{w}} = 2R\underline{w} - 2P.$$

The optimal weight vector $\underline{w}^*$ can be determined by the condition of

$\nabla E[\epsilon_k^2]|_{\underline{w} = \underline{w}^*} = 0$.

That is, $2R\underline{w}^* - 2P = 0$.

This implies that the optimal weight is described by

$\underline{w}^* = R^{-1}P$.

In this case, the minimum mean square error (MMSE) becomes

$MMSE = E[d_k^2] + \underline{w}^{*T} R \underline{w}^* - 2P^T \underline{w} = E[d_k^2] - P^T \underline{w}^*$.

Let $\underline{v} = \underline{w} - \underline{w}^*$ and rewrite the MSE as follows:

$$\begin{aligned} E[\epsilon_k^2] &= E[d_k^2] + \underline{w}^T R \underline{w} - 2P^T \underline{w} \\ &= E[d_k^2] - P^T \underline{w}^* + \underline{w}^{*T} R \underline{w}^* + \underline{w}^T R \underline{w} - 2\underline{w}^T R \underline{w}^* \\ &= MMSE + (\underline{w} - \underline{w}^*)^T R (\underline{w} - \underline{w}^*) \\ &= MMSE + \underline{v}^T R \underline{v} \end{aligned}$$

To analyse the MSE, let us set

$$F(\underline{v}) = \underline{v}^T R \underline{v}.$$

That is,

$$E[\epsilon_k^2] = MMSE + F(\underline{v}).$$

The MSE depends on $F(\underline{v})$.

Here, the input correlation matrix R can be decomposed by

$$R = Q\Lambda Q^{-1},\ \ Q = \begin{pmatrix} \\ q_1\, q_2 \cdots q_d \\ \\ \end{pmatrix},\ \text{and}\ \ \Lambda = \begin{pmatrix} \lambda_0 & 0 & \cdots & 0 \\ 0 & \lambda_1 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & \lambda_d \end{pmatrix}$$

where q_i is the column vector representing the eigenvector of R and λ_i is the eigenvalue corresponding to q_i.

Since R is the symmetric matrix, q_is are orthogonal and

$$R = Q\Lambda Q^{-1} = Q\Lambda Q^T.$$

Here, $F(\underline{v})$ can be rewritten as

$$F(\underline{v}) = \underline{v}^T R \underline{v} = \underline{v}^T Q\Lambda Q^T \underline{v}.$$

Let $\underline{v}' = Q^T \underline{v}$, that is, rotation of $\underline{v}$ toward eigenvector axes.

Here, the eigenvectors of R define the principal axes of the error surface. Then,

$$F(\underline{v}') = \underline{v}'^T \Lambda \underline{v}'.\ \text{(ellipsoid)}$$

Let $F(\underline{v}^{'}) = c$ in 2 dimensional space.

Then,

$$\lambda_0 v_0^{'2} + \lambda_1 v_1^{'2} = c.$$

the length of $v_0^{'}$ axis = $\sqrt{c/\lambda_0}$, the length of $v_1^{'}$ axis = $\sqrt{c/\lambda_1}$

cf. the equation of ellipse: $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$

. gradient descent method (batch mode)

The weight vector is updated by

$$\begin{aligned}\underline{w}_{k+1} &= \underline{w}_k + \mu(-\nabla F|_{\underline{v} = \underline{v}_k}) \\ &= \underline{w}_k + 2\mu R(\underline{w}^* - \underline{w}_k) \\ &= (I - 2\mu R)\underline{w}_k + 2\mu R\underline{w}^*\end{aligned}$$

This can be redescribed by

$$\underline{v}_{k+1} = (I - 2\mu R)\underline{v}_k = (I - 2\mu Q\Lambda Q^T)\underline{v}_k.$$

Let $\underline{v} = Q\underline{v}^{'}$.

Then,

$Q\underline{v}_{k+1}^{'} = (I - 2\mu Q\Lambda Q^T) Q\underline{v}_k^{'}$, that is, $\underline{v}_{k+1}^{'} = (I - 2\mu\Lambda)\underline{v}_k^{'}$.

The iterative form of $\underline{v}_k^{'}$ is

$$\underline{v}_k^{'} = (I - 2\mu\Lambda)^k \underline{v}_0^{'}.$$

Therefore, the gradient descent algorithm is stable and convergent when

$$\lim_{k \to \infty} (I - 2\mu\Lambda)^k = 0.$$

This implies that the convergence condition is

$$0 < \mu < \frac{1}{tr[R]} \leqq \frac{1}{\lambda_{\max}}$$

where $\lambda_{\max}$ is the maximum eigenvalue of R.

Note that

$$\lambda_{\max} \leqq tr[\Lambda] = tr[R].$$

In this case, the learning curve of MSE is determined as follows:

$$\begin{aligned} E[\epsilon_k^2] &= MMSE + \underline{v}_k^T R \underline{v}_k \\ &= MMSE + \underline{v}_k^{'} \Lambda \underline{v}_k^{'} \\ &= MMSE + [(I - 2\mu\Lambda)^k \underline{v}_0^{'}]^T \Lambda [(I - 2\mu\Lambda)^k \underline{v}_0^{'}] \\ &= MMSE + \underline{v}_0^{'T} (I - 2\mu\Lambda)^{2k} \Lambda \underline{v}_0^{'} \\ &= MMSE + \sum_{n=0}^{d} v_{0n}^{'2} \lambda_n (1 - 2\mu\lambda_n)^{2k} \end{aligned}$$

The learning curve of MSE is dependent upon the geometric ratio

$$r_n^2 \equiv (1 - 2\mu\lambda_n)^2$$

. least mean square (LMS) method (on-line mode)

The weight vector is updated by

$$\begin{aligned} \underline{w}_{k+1} &= \underline{w}_k - \mu \widehat{\nabla}_k \\ &= \underline{w}_k + 2\mu \epsilon_k \underline{x}_k \end{aligned}$$

where

$\epsilon_k = d_k - y_k = d_k - \underline{x}_k^T \underline{w}_k$ and

$$\widehat{\nabla}_k \equiv [\frac{\partial \epsilon_k^2}{\partial w_0}, \frac{\partial \epsilon_k^2}{\partial w_1}, \cdots, \frac{\partial \epsilon_k^2}{\partial w_d}]^T = -2\epsilon_k \underline{x}_k.$$

That is, $\widehat{\nabla}_k$ is not a true gradient, but an estimated gradient from ϵ_k and $\underline{x}_k$.

Is $\widehat{\nabla}_k$ an unbiased estimator?

$$\begin{aligned} E[\widehat{\nabla}_k] &= -2E[\epsilon_k \underline{x}_k] \\ &= -2E[d_k \underline{x}_k - \underline{x}_k \underline{x}_k^T \underline{w}_k] \\ &= 2(R\underline{w}_k - P) \\ &= \nabla E|_{\underline{w} = \underline{w}_k} \end{aligned}$$

The answer is yes. To check the convergence condition of the LMS method, let us consider the expectation of $\underline{w}_{k+1}$:

$$\begin{aligned} E[\underline{w}_{k+1}] &= E[\underline{w}_k] + 2\mu E[\epsilon_k \underline{x}_k] \\ &= E[\underline{w}_k] + 2\mu (E[d_k \underline{x}_k] - E[\underline{x}_k \underline{x}_k^T \underline{w}_k] \\ &= E[\underline{w}_k] + 2\mu (P - RE[\underline{w}_k]) \\ &= (1 - 2\mu R) E[\underline{w}_k] + 2\mu R \underline{w}^* \end{aligned}$$

This can be redescribed by

$$E[\underline{v}_{k+1}] = (I - 2\mu R)E[\underline{v}_k],$$

that is,

$$E[\underline{v}_k^{'}] = (I - 2\mu \Lambda)^k \underline{v}_0^{'}.$$

This implies that the convergence condition is

$$0 < \mu < \frac{1}{tr[R]} \leqq \frac{1}{\lambda_{\max}}$$

where $\lambda_{\max}$ is the maximum eigenvalue of R,

that is, the convergence condition is same as the batch mode.

To check the performance of LMS method, let us consider the following definition of misadjustment:

$$M \equiv \frac{excess\, MSE}{MMSE} = \frac{E[MSE - MMSE]}{MMSE}$$

where the excess MSE is given by

$$\begin{aligned} excess\, MSE &= E[\underline{v}_k^T R \underline{v}_k] \\ &= E[\underline{v}_k^{'T} \Lambda \underline{v}_k^{'}] \\ &= \sum_{n=0}^{d} \lambda_n E[v_{nk}^{'2}] \end{aligned}$$

To analyse the excess MSE, let us set

$$\widehat{\nabla}_k = \nabla_k + \underline{n}_k$$

where $\underline{n}_k$ represents the noise vector associated with $\widehat{\nabla}_k$.

Near $\underline{w}^*$, the noise vector can be approximated as

$$\underline{n}_k \approx \widehat{\nabla}_k = -2\epsilon_k \underline{x}_k.$$

Then, the covariance of $\underline{n}_k$ becomes

$$Cov[\underline{n}_k] = E[\underline{n}_k \underline{n}_k^T] \approx 4E[\epsilon_k^2 \underline{x}_k \underline{x}_k^T].$$

Near $\underline{w}^*$, ϵ_k and $\underline{x}_k$ is uncorrelated, that is,

$$Cov[\underline{n}_k] \approx 4E[\epsilon_k^2]E[\underline{x}_k \underline{x}_k^T] \approx 4MMSE \cdot R$$ and

$$\begin{aligned} Cov[\underline{n}_k^{'}] &= Cov[Q^{-1}\underline{n}_k] \\ &= E[Q^{-1}\underline{n}_k (Q^{-1}\underline{n}_k)^T] \\ &= Q^{-1} Cov[\underline{n}_k] Q \\ &\approx 4MMSE \cdot \Lambda \end{aligned}$$

From the weight update rule,

$$\begin{aligned} \underline{v}_{k+1} &= \underline{v}_k - \mu \widehat{\nabla}_k \\ &= \underline{v}_k - \mu(2R\underline{v}_k + \underline{n}_k) \\ &= (I - 2\mu R)\underline{v}_k - \mu \underline{n}_k \end{aligned}$$

This implies that

$$\underline{v}_{k+1}^{'} = (I - 2\mu \Lambda)\underline{v}_k^{'} - \mu \underline{n}_k^{'}.$$

From this equation, the covariance of $\underline{v}_k^{'}$ is determined by

$$Cov[\underline{v}_k^{'}] = (I - 2\mu\Lambda)^2 Cov[\underline{v}_k^{'}] + \mu^2 Cov[\underline{n}_k^{'}]$$
$$= \frac{\mu}{4}(\Lambda - \mu\Lambda^2)^{-1} Cov[\underline{n}_k^{'}]$$

By substituting the result of $Cov[\underline{n}_k^{'}]$, we get

$$Cov[\underline{v}_k^{'}] \approx \mu MMSE(\Lambda - \mu\Lambda^2)^{-1}\Lambda \approx \mu MMSE \cdot I.$$

This implies that, the excess MSE becomes

$$excess\ MSE = \sum_{n=0}^{d} \lambda_n E[v_{nk}^{'2}] \approx \mu MMSE \sum_{n=0}^{d} \lambda_n.$$

Therefore, the misadjustment becomes

$$M \equiv \frac{excess\ MSE}{MMSE} \approx \mu \cdot tr[R].$$

Here, let us investigate the learning curve when we use the LMS method.

The time constant is defined by the time interval in which the given signal $f(t)$ is reduced by $1/e$.
example. $f(t) = e^{-t/\tau}$, time constant = τ.

the geometric ratio $r^2 \equiv e^{-1/\tau}$, that is, $r^2 \approx 1 - 1/\tau$.
From the previous result,

$$r_n^2 = (1 - 2\mu\lambda_n)^2 \approx 1 - 1/\tau_n, \text{ that is, } 1/\tau_n \approx 1/(4\mu\lambda_n).$$

Here, the trace of R can be redescribed by

$$tr[R] = \sum_{n=0}^{d} \lambda_n \approx \frac{1}{4\mu} \sum_{n=0}^{d} \frac{1}{\tau_n} = \frac{d+1}{4\mu} \frac{1}{\tau_{av}}$$

where $\frac{1}{\tau_{av}} = \frac{1}{d+1} \sum_{n=0}^{d} \frac{1}{\tau_n}$.

This implies that

$\tau_{av} \approx \frac{d+1}{4\mu \cdot tr[R]}$ and

$M \approx \frac{d+1}{4\tau_{av}}$.

These results show that

(1) large μ -> small τ_{av} (fast convergence) -> large M and

(2) small μ -> large τ_{av} (slow convergence) -> small M.

Reference: Adaptive Signal Processing, chapters 3, 4, 5, and 6.