Stability of the Steepest-Descent Algorithm

Since the SD algorithm can be viewed as a feedback model, it is subject to the possibility of becoming unstable, i.e. $\lim_{n\to\infty}\mathbf{w}(n) \neq \mathbf{w}_o$ and hence $\lim_{n\to\infty} J[\mathbf{w}(n)] \neq J_{\min}$.

We will show that the stability performance of the SD algorithm is determined by two factors: the step-size parameter, μ and the largest eigenvalue λ_{max} of the correlation matrix, **R**.

We begin with the SD recursion (2)

$$\begin{aligned}\mathbf{w}(n+1) &= \mathbf{w}(n) + \mu[\mathbf{p} - \mathbf{R}\mathbf{w}(n)] \\ &= \mathbf{w}(n) + \mu[\mathbf{R}\mathbf{w}_o - \mathbf{R}\mathbf{w}(n)]\end{aligned}.$$

Subtracting the optimal weight vector (translation of coordinate system) from both sides yields

$$\begin{aligned}\mathbf{w}(n+1) - \mathbf{w}_o &= [\mathbf{w}(n) - \mathbf{w}_o] - \mu[\mathbf{R}\mathbf{w}(n) - \mathbf{R}\mathbf{w}_o] \\ &= (\mathbf{I} - \mu\mathbf{R})[\mathbf{w}(n) - \mathbf{w}_o]\end{aligned}.$$

Defining the misalignment vector (weight-error vector) at time *n*

$$\mathbf{c}(n) = \mathbf{w}(n) - \mathbf{w}_o$$

yields

$$\begin{aligned}\mathbf{c}(n+1) &= \mathbf{w}(n) - \mathbf{w}_o + \mu[\mathbf{R}\mathbf{w}_o - \mathbf{R}\mathbf{w}(n)] \\ &= (\mathbf{I} - \mu\mathbf{R})\mathbf{c}(n)\end{aligned}.$$

In order for stability of the algorithm and hence *convergence* of $\mathbf{w}(n)$ to $\mathbf{w}_o$ we need to drive $\mathbf{c}(n)$ to $\mathbf{0}$. To continue the analysis, we employ the diagonalization of **R** to rewrite the above as

$$\mathbf{c}(n+1) = (\mathbf{I} - \mu\Theta\Lambda\Theta^H)\mathbf{c}(n).$$

We apply our unitary transformation (rotation of coordinate system) so that we can decouple the weight error elements from each other. We have

$$\begin{aligned}\mathbf{Q}^H\mathbf{c}(n+1) &= \Theta^H(\mathbf{I} - \mu\Theta\Lambda\Theta^H)\mathbf{c}(n) \\ &= \left(\Theta^H - \mu\underbrace{\Theta^H\Theta}_{\mathbf{I}}\Lambda\Theta^H\right)\mathbf{c}(n) \\ &= (\mathbf{I} - \mu\Lambda)\Theta^H\mathbf{c}(n)\end{aligned}.$$

As before, we denote the translated and rotated (transformed) vectors as

$$\begin{aligned}\nu(n) &= \mathbf{Q}^H\mathbf{c}(n) \\ &= \mathbf{Q}^H[\mathbf{w}(n) - \mathbf{w}_o]\end{aligned}$$

which leads to

$$\nu(n+1) = (\mathbf{I} - \mu\Lambda)\nu(n).$$

We examine a single weight error element, $\nu_k(n)$ (which due to the coordinate transformation and decoupling is only a function of itself)

$$\begin{aligned} \nu_k(n+1) &= (1-\mu\lambda_k)\nu_k(n) \\ &= (1-\mu\lambda_k)^n \nu_k(0) \end{aligned} \quad . \qquad (k\text{th natural mode})$$

We remember that for stability, we wish to drive the misalignment $\mathbf{c}(n)$ to $\mathbf{0}$ and consequently each $\nu_k(n)$ to 0, for $k = 1, ..., M$, as quickly as possible.

Clearly for each mode to die out we require

$$-1 < 1-\mu\lambda_k < 1$$

or

$$0 < \mu < \frac{2}{\lambda_\kappa} .$$

We note that

$$\frac{2}{\lambda_{\mu\alpha\xi}} \le \frac{2}{\lambda_\kappa} \le \frac{2}{\lambda_{\mu\nu}}$$

thus we must select the μ which guarantees all modes die out. That μ is given by

$$\boxed{0 < \mu < \frac{2}{\lambda_{\mu\alpha\xi}}}.$$

We finally note that the mode which takes longest to die out is that associated with $\lambda_{\min}$

$$\nu_k(n+1) = (1-\mu\lambda_{\min})^n \nu_k(0) .$$

We therefore conclude that when the eigenvalues are spread over a large range (or equivalently the condition number, $\chi = \frac{\lambda_{\max}}{\lambda_{\min}}$ is large) the settling time of the SD algorithm is limited by the smallest eigenvalues.

Transient Behavior of the Coefficient Vector

We have defined the misalignment vector or weight error vector as

$$\mathbf{c}(n) = \boldsymbol{\omega}(\nu) - \boldsymbol{\omega}_o$$

and the misalignment vector under coordinate rotation (transformed weight error vector) as

$$\begin{aligned} \nu(n) &= \mathbf{Q}^H \mathbf{c}(n) \\ &= \mathbf{Q}^H [\mathbf{w}(n) - \mathbf{w}_o] \end{aligned} .$$

Rearranging, we have

$$\mathbf{Q}^H \mathbf{w}(n) = \mathbf{Q}^H \mathbf{w}_o + \nu(n)$$

which when premultiplied by $\mathbf{Q}$ gives us (since $\mathbf{Q}\mathbf{Q}^H = \mathbf{I}$)

$$\begin{aligned}\mathbf{w}(n) &= \boldsymbol{\omega}_o + \boldsymbol{\Theta}\boldsymbol{\nu}(n) \\ &= \boldsymbol{\omega}_o + [\boldsymbol{\theta}_1 \mid \Lambda \mid \boldsymbol{\theta}_M]\begin{bmatrix}\nu_1(n) \\ \hline \mathrm{M} \\ \hline \nu_M(n)\end{bmatrix}. \\ &= \boldsymbol{\omega}_o + \sum_{k=1}^{M} \boldsymbol{\theta}_k \nu_k(n)\end{aligned}$$

Recall the equation for the *k*th mode (backsolved)

$$\nu_k(n) = (1-\mu\lambda_k)^n \nu_k(0)$$

and consider the *i*th coefficient of **w**(*n*) (substituting the modal equation)

$$w_i(n) = \omega_{oi} + \sum_{k=1}^{M} \theta_{ki}\nu_k(0)(1-\mu\lambda_k)^n .$$

This expression provides a description of the transient behavior of the *i*th coefficient of **w**(*n*). Each coefficient of **w**(*n*) consists of a weighted sum of exponentials. If we choose

$$0 < \mu < \frac{2}{\lambda_{max}}$$

then we guarantee that

$$-1 < (1-\mu\lambda_k) < 1,\ 1 \le k \le M$$

and hence

$$\lim_{n\to\infty}(1-\mu\lambda_k)^n = 0,\ 1 \le k \le M .$$

This of course implies that

$$\lim_{n\to\infty} w_i(n) = \omega_{oi}$$

and

$$\boxed{\lim_{n\to\infty} \mathbf{w}(n) = \mathbf{w}_o}.$$

We note that our coefficient vector converges to the Wiener filter regardless of **w**(0).

Transient Behavior of the Mean-Squared Error

Since our error surface is quadratic with a single, global minimum and $\lim_{n\to\infty} \mathbf{w}(n) = \mathbf{w}_o$ we are guaranteed to have

$$\boxed{\lim_{n\to\infty} J(n) = J_{min}}.$$

This can also be derived independently using our canonical form (Haykin 5.57) of the MSE derived for the Wiener filter

$$J(n) = \vartheta_{\mu\nu} + \sum_{\kappa=1}^{M} \lambda_\kappa \left| \nu_\kappa(\nu) \right|^2$$

$$= \vartheta_{\mu\nu} + \sum_{\kappa=1}^{M} \lambda_\kappa \left| (1-\mu\lambda_\kappa)^{\nu} \nu_\kappa(0) \right|^2$$

$$= \vartheta_{\mu\nu} + \sum_{\kappa=1}^{M} \lambda_\kappa (1-\mu\lambda_\kappa)^{2\nu} \left| \nu_\kappa(0) \right|^2$$

If $0 < \mu < \frac{2}{\lambda_{\mu\alpha\xi}}$ then

$$\lim_{n \to \infty} \vartheta(\nu) = \vartheta_{\mu\nu}$$

as we've already proved.

To get a feel for the progress our search algorithm makes in finding $\mathbf{w}_o$, we typically plot

- $\mathbf{w}(n)$ vs. n [usually $\mathbf{w}(n)$ is length 2] or $\boldsymbol{\nu}(n)$ vs. n. Such a plot is called a weight trajectory or weight error trajectory. Usually we plot $w_0(n)$ and $w_1(n)$ as a function of n along with the loci of $w_0(n)$ and $w_1(n)$ for fixed values of the MSE (contours).
- $J(n)$ vs. n. Such a plot is called the learning curve or simply the MSE plot.