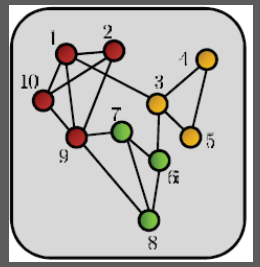


INFERENCE OVER NETWORKS

LECTURE #12: Performance by Single Agents

Professor Ali H. Sayed
UCLA Electrical Engineering

Course EE210B
Spring Quarter 2015

Proc. IEEE, vol. 102, no. 4, pp. 460-497, April 2014.
Foundations and Trends in Machine Learning, vol. 7, no. 4-5, pp. 311-801, July 2014.

Reference

Chapter 4 (Performance of Single Agents, pp. 368-406):

A. H. Sayed, ``Adaptation, learning, and optimization over networks," ***Foundations and Trends in Machine Learning***, vol. 7, issue 4-5, pp. 311-801, NOW Publishers, 2014.

Conditions on Risk Function

We consider the case of real arguments first. Thus, let $J(w) \in \mathbb{R}$ denote the real-valued cost function of a real-valued vector argument, $w \in \mathbb{R}^M$ and consider the same optimization problem (3.1):

$$w^o = \arg\min_w \; J(w) \tag{4.3}$$

We continue to assume that $J(w)$ is twice-differentiable and satisfies (3.2) for some positive parameters $\nu \leq \delta$, namely,

$$0 < \nu I_M \leq \nabla_w^2 J(w) \leq \delta I_M \tag{4.4}$$

Conditions on Risk Function

Assumptions (can be relaxed):

a) $J(w)$ twice-differentiable

b) $J(w)$ is $\nu-$strongly convex $\iff \nabla_w^2 J(w) \geq \nu I_M > 0$

c) $\nabla_w J(w)$ is $\delta-$Lipschitz $\iff \|\nabla_w J(w_2) - \nabla_w J(w_1)\| \leq \delta \|w_2 - w_1\|$

$$\iff \nabla_w^2 J(w) \leq \delta I_M$$

Example: conditions are satisfied by quadratic or logistic risks.

Stochastic Gradient Algorithm

We established in the previous chapter the mean-square-error stability of the following stochastic-gradient recursion for seeking the minimizer w^o in the real data case:

$$\boldsymbol{w}_i \;=\; \boldsymbol{w}_{i-1} \;-\; \mu\,\widehat{\nabla_{w^{\mathsf{T}}} J}(\boldsymbol{w}_{i-1}), \quad i \geq 0 \tag{4.5}$$

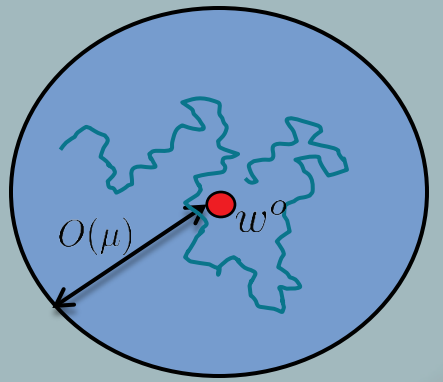
The analysis relied on the conditions in Assumption 3.2 on the gradient noise process, $\boldsymbol{s}_i(\boldsymbol{w}_{i-1})$, which we repeat here for ease of reference. Recall from (3.25) that

$$\boldsymbol{s}_i(\boldsymbol{w}) \;\triangleq\; \widehat{\nabla_{w^{\mathsf{T}}} J}(\boldsymbol{w}) \;-\; \nabla_{w^{\mathsf{T}}} J(\boldsymbol{w}) \tag{4.6}$$

Stability of Error Moments

Lemma 3.1: For small-enough step-sizes, it holds that

$$\begin{cases} \limsup\limits_{i\to\infty} \|\mathbb{E}\,\widetilde{\boldsymbol{w}}_i\| = O(\mu) \\ \limsup\limits_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 = O(\mu) \\ \limsup\limits_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^4 = O(\mu^2) \end{cases}$$



Gradient Noise Covariance

For any $\boldsymbol{w} \in \boldsymbol{\mathcal{F}}_{i-1}$, we let

$$R_{s,i}(\boldsymbol{w}) \triangleq \mathbb{E}\left[\boldsymbol{s}_i(\boldsymbol{w})\boldsymbol{s}_i^{\mathsf{T}}(\boldsymbol{w}) \,|\, \boldsymbol{\mathcal{F}}_{i-1}\right] \tag{4.11}$$

denote the conditional second-order moment of the gradient noise process, which generally depends on i because the statistical distribution of $\boldsymbol{s}_i(\boldsymbol{w})$ can be iteration-dependent.

Gradient Noise Covariance

Note that $R_{s,i}(\boldsymbol{w})$ is a random quantity since it depends on the random iterate $\boldsymbol{w}$. We assume that, in the limit, this covariance matrix tends to a constant value when evaluated at w^o and we denote the limit by

$$R_s \triangleq \lim_{i\to\infty} \mathbb{E}\left[\boldsymbol{s}_i(w^o)\boldsymbol{s}_i^{\mathsf{T}}(w^o) \,|\, \boldsymbol{\mathcal{F}}_{i-1}\right] \tag{4.12}$$

We sometimes refer to the term $\boldsymbol{s}_i(w^o)$ as the *absolute* noise component.

Smoothness Conditions

Assumption 4.3 (Smoothness conditions). It is assumed that the Hessian matrix of the cost function, $J(w)$, and the noise covariance matrix defined by (4.11) are locally Lipschitz continuous in a small neighborhood around $w = w^o$ in the following manner:

$$\left\|\nabla_w^2 J(w^o + \Delta w) - \nabla_w^2 J(w^o)\right\| \leq \kappa_1 \|\Delta w\| \tag{4.18}$$

$$\|R_{s,i}(w^o + \Delta w) - R_{s,i}(w^o)\| \leq \kappa_2 \|\Delta w\|^\gamma \tag{4.19}$$

for small perturbations $\|\Delta w\| \leq \epsilon$ and for some constants $\kappa_1 \geq 0$, $\kappa_2 \geq 0$, and exponent $0 < \gamma \leq 4$.

Weight Error Recursions

$$\widetilde{\boldsymbol{w}}_i = (I_M - \mu \boldsymbol{H}_{i-1})\widetilde{\boldsymbol{w}}_{i-1} + \mu\, \boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \quad (4.64)$$

$$\widetilde{\boldsymbol{w}}'_i = (I_M - \mu H)\widetilde{\boldsymbol{w}}'_{i-1} + \mu\, \boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \quad (4.65)$$

where

$$\boldsymbol{H}_{i-1} \triangleq \int_0^1 \nabla_w^2\, J(w^o - t\widetilde{\boldsymbol{w}}_{i-1})dt \quad (4.38)$$

We introduce the deviation matrix

$$\widetilde{\boldsymbol{H}}_{i-1} \triangleq H - \boldsymbol{H}_{i-1} \quad (4.39)$$

$$H \triangleq \nabla_w^2\, J(w^o)$$

Long-Term Error Dynamics

Lemma 4.3: For small-enough step-sizes, the error dynamics in steady-state is well-approximated by the long-term model:

$$\widetilde{\boldsymbol{w}}_i' \;=\; (I_M - \mu H)\widetilde{\boldsymbol{w}}_{i-1}' \;+\; \mu\, \boldsymbol{s}_i(\boldsymbol{w}_{i-1})$$

Specifically, it holds that:

$$\begin{cases} \limsup\limits_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i - \widetilde{\boldsymbol{w}}_i'\|^2 &= O(\mu^2) \\ \limsup\limits_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 &= \limsup\limits_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i'\|^2 + O(\mu^{3/2}) \\ \limsup\limits_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|_H^2 &= \limsup\limits_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i'\|_H^2 + O(\mu^{3/2}) \end{cases}$$

Performance Metrics

Course EE210B
Spring Quarter 2015

Proc. IEEE, vol. 102, no. 4, pp. 460-497, April 2014.
Foundations and Trends in Machine Learning, vol. 7, no. 4-5, pp. 311-801, July 2014.

Performance Metrics

Two useful metrics for assessing the performance of stochastic gradient algorithms are the mean-square-deviation (MSD) and the excess-risk (ER). We define these two measures below before explaining how the long-term model (4.55) can be used to evaluate their values.

Mean-Square Deviation

To motivate the definition of the MSD, we first remark that we will be establishing further ahead in (4.97) and (4.128) the following two expressions for the limit superior and limit inferior of the error variance:

$$\limsup_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \;=\; \mu\cdot\overline{\mathrm{MSD}} \;+\; o(\mu) \tag{4.83}$$

$$\liminf_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \;=\; \mu\cdot\overline{\mathrm{MSD}} \;-\; o(\mu) \tag{4.84}$$

for some common positive constant $\overline{\mathrm{MSD}}$ whose exact value is not relevant for the current discussion. We explained the meaning of the limit superior operation earlier prior to the statement of Lemma 3.1. We can

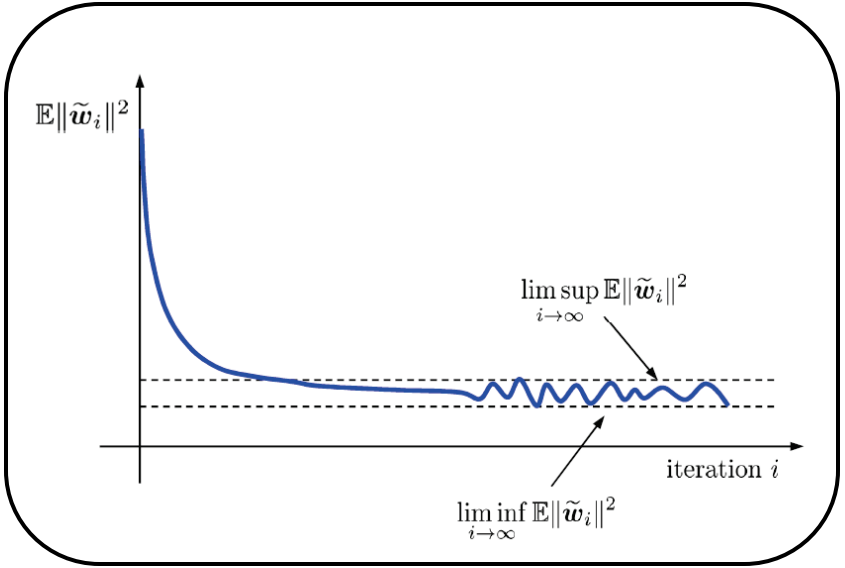
Limits Superior and Inferior

similarly view the *limit inferior* of a sequence as essentially corresponding to the largest lower bound for the limiting behavior of the sequence; this concept is again useful when the sequence is not necessarily convergent but tends towards a small bounded region [89, 144, 202]. A schematic illustration of the limit superior and limit inferior values for the error variance, $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$, is shown in Figure 4.1. If the sequence happens to be convergent, then both its limit superior and limit inferior values will coincide and they will be equal to the regular limiting value of the sequence.

Limits Superior and Inferior

$$\begin{cases} \limsup\limits_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 &= \mu\cdot\overline{\mathrm{MSD}} + o(\mu) \\ \liminf\limits_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 &= \mu\cdot\overline{\mathrm{MSD}} - o(\mu) \end{cases}$$

for some positive constant $\overline{\mathrm{MSD}}$.



Mean-Square Deviation

Now, comparing the first relation (4.83) with (4.2), it is observed that (4.83) characterizes the size of the coefficient of the first-order term in μ as being equal to $\overline{\text{MSD}}$. Moreover, if we divide both sides of (4.83) and (4.84) by μ and compute the limit as $\mu \to 0$, which corresponds to assuming operation in the slow adaptation regime, then we find that

$$\lim_{\mu\to 0}\left(\limsup_{i\to\infty}\frac{1}{\mu}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2\right)=\lim_{\mu\to 0}\left(\liminf_{i\to\infty}\frac{1}{\mu}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2\right)=\overline{\text{MSD}} \quad (4.85)$$

Mean-Square Deviation

That is, the limiting values of the *scaled* limit superior and limit inferior expressions coincide with each other and they are both equal to $\overline{\text{MSD}}$. This fact indicates that as $\mu \to 0$, the quantity $\frac{1}{\mu}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$ approaches a limiting value after sufficient iterations and, once multiplied by μ, this limiting value can be used to assess the size of the error variance, $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$, in steady-state. For this reason, we shall define the MSD measure as follows:

$$\text{MSD} \triangleq \mu \cdot \left(\lim_{\mu \to 0} \limsup_{i \to \infty} \frac{1}{\mu} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \right) \tag{4.86}$$

Performance Metric: MSD

$$\lim_{\mu\to 0}\left(\limsup_{i\to\infty}\frac{1}{\mu}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2\right)=\lim_{\mu\to 0}\left(\liminf_{i\to\infty}\frac{1}{\mu}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2\right)=\overline{\text{MSD}}$$

- Therefore, as $\mu\to 0$, the quantity $\frac{1}{\mu}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$ approaches a limit.

- Once multiplied by μ, this limit assesses $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$ to first-order in μ:

(simplification)

$$\text{MSD}\triangleq\mu\cdot\left(\lim_{\mu\to 0}\limsup_{i\to\infty}\frac{1}{\mu}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2\right)$$

$$\text{MSD}\triangleq\lim_{i\to\infty}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$$

Mean-Square Deviation

In view of equality (4.85), we could have also defined the MSD by using the lim inf operation in (4.86) instead of the lim sup operation. For uniformity throughout this work, we shall adopt the lim sup notation.

Mean-Square Deviation

Sometimes, with some abuse of notation, we write the definition for the MSD more simply, for sufficiently small step-sizes, as follows:

$$\text{MSD} \triangleq \lim_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \tag{4.87}$$

with the understanding that this limit is computed as in (4.86) since, strictly speaking, the limit on the right-hand side of (4.87) may not exist. Yet, it is useful to note that derivations that assume the validity of (4.87) still lead to the same expression for the MSD to first-order in μ as derivations that rely on the more formal expression (4.86) — this fact can be verified by examining and repeating the proof of Theorem 4.7 further ahead.

Excess Risk

The second useful metric for evaluating the performance of stochastic gradient algorithms relates to the mean excess-cost; which is also called the *excess-risk* (ER) in the machine learning literature [37, 233] and the *excess-mean-square-error* (EMSE) in the adaptive filtering literature [107, 206, 262]. We denote it by the letters ER and, similarly to (4.86), we can motivate the following expression for it:

$$\text{ER} \triangleq \mu \cdot \left(\lim_{\mu \to 0} \limsup_{i \to \infty} \frac{1}{\mu} \mathbb{E}\{J(\boldsymbol{w}_{i-1}) - J(w^o)\} \right) \tag{4.88}$$

Excess Risk

In other words, the ER metric measures the average fluctuation of the cost function around its minimum value in steady-state. Again, we could have used the lim inf operation in (4.88) instead of the lim sup operation. We again adopt the lim sup convention.

Recall #1: Quadratic Increment

Lemma E.2 (Perturbation approximation: Real arguments). Consider the same setting of Lemma E.1 and assume additionally that the Hessian matrix function is locally Lipschitz continuous in a small neighborhood around $z = z^o$ as defined by (E.7). It then follows that the increment in the value of the function $g(z)$ for sufficiently small variations around $z = z^o$ can be well approximated by

$$g(z^o + \Delta z) \;-\; g(z^o) \;\approx\; \Delta z^{\mathsf{T}} \left[\frac{1}{2}\nabla_z^2 g(z^o)\right] \Delta z \qquad \text{(E.10)}$$

where the approximation error is in the order of $O(\|\Delta z\|^3)$.

Recall #2: Jensen's Inequality

There is also a useful stochastic version of Jensen's inequality. If $\boldsymbol{a} \in \mathbb{R}^M$ is a real-valued random variable, then it holds that

$$f(\mathbb{E}\,\boldsymbol{a}) \leq \mathbb{E}\,(f(\boldsymbol{a})) \quad (\text{when } f(x) \in \mathbb{R} \text{ is convex}) \tag{F.29}$$

$$f(\mathbb{E}\,\boldsymbol{a}) \geq \mathbb{E}\,(f(\boldsymbol{a})) \quad (\text{when } f(x) \in \mathbb{R} \text{ is concave}) \tag{F.30}$$

where it is assumed that $\boldsymbol{a}$ and $f(\boldsymbol{a})$ have bounded expectations. We remark that a function $f(x)$ is said to be concave if, and only if, $-f(x)$ is convex.

Excess Risk

Using the smoothness condition (4.20), and result (E.10) from the appendix, we recognize that the mean fluctuation that appears inside (4.88) satisfies:

$$\limsup_{i\to\infty} \mathbb{E}\,\{J(\boldsymbol{w}_{i-1}) - J(w^o)\} = \limsup_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2_{\frac{1}{2}H} + O(\mu^{3/2}) \quad (4.89)$$

in terms of a weighted mean-square-error norm. The appearance of the $O(\mu^{3/2})$ factor in the above expression can be motivated as follows. We note from expression (E.10) in the appendix that the right-most term in (4.89) should be the asymptotic size of $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^3$. We then rely on result (3.67) to note that:

Excess Risk

$$\begin{aligned}\limsup_{i\to\infty}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^3 &\overset{(F.30)}{\leq} \limsup_{i\to\infty}\left(\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^4\right)^{3/4}\\ &\overset{(3.67)}{=} \left(O(\mu^2)\right)^{3/4}\\ &= O(\mu^{3/2}) \qquad (4.90)\end{aligned}$$

where in the first line we called upon Jensen's inequality (F.30) and the fact that the function $f(x) = x^{3/4}$ is concave over the range $x \geq 0$. It follows from (4.88) and (4.89) that we can also evaluate the ER metric by means of the following alternative expression:

Excess Risk

$$\text{ER} \;=\; \mu \cdot \left(\lim_{\mu \to 0} \limsup_{i \to \infty} \frac{1}{\mu} \mathbb{E} \, \|\widetilde{\boldsymbol{w}}_{i-1}\|^2_{\frac{1}{2}H} \right) \tag{4.91}$$

Again, with some abuse in notation, we sometimes write more simply either of the following expressions for sufficiently small step-sizes in place of (4.88) and (4.91):

$$\text{ER} \;=\; \lim_{i \to \infty} \mathbb{E} \, \{J(\boldsymbol{w}_{i-1}) - J(w^o)\} \tag{4.92}$$

$$\text{ER} \;=\; \lim_{i \to \infty} \mathbb{E} \, \|\widetilde{\boldsymbol{w}}_{i-1}\|^2_{\frac{1}{2}H} \tag{4.93}$$

Excess Risk

with the understanding that the limits in the above two expressions are computed as in (4.88) or (4.91) since, strictly speaking, these limits may not exist. Still, it is useful to note that derivations that assume the validity of (4.92)–(4.93) lead to the same expression for the ER to first-order in μ as derivations that rely on the more formal expressions (4.88) or (4.91) — this fact can be verified by examining and repeating the proof of Theorem 4.7. We collect the expressions for the MSD and ER measures in the following statement for ease of reference.

Performance Metrics

Definition 4.2 (Performance measures). The mean-square-deviation (MSD) and excess-risk (ER) performance metrics are defined as follows:

$$\text{MSD} \triangleq \mu \cdot \left(\lim_{\mu \to 0} \limsup_{i \to \infty} \frac{1}{\mu} \mathbb{E} \, \|\widetilde{\boldsymbol{w}}_i\|^2 \right) \tag{4.94}$$

$$\text{ER} \triangleq \mu \cdot \left(\lim_{\mu \to 0} \limsup_{i \to \infty} \frac{1}{\mu} \mathbb{E} \, \{J(\boldsymbol{w}_{i-1}) - J(w^o)\} \right) \tag{4.95}$$

for sufficiently small step-sizes, where the MSD measures the size of the error variance, $\mathbb{E} \, \|\widetilde{\boldsymbol{w}}_i\|^2$, in steady-state, while the ER measures the size of the mean fluctuation, $\mathbb{E} \, \{J(\boldsymbol{w}_{i-1}) - J(w^o)\}$, also in steady state. Under result (3.67), and using the Hessian matrix H from (4.40), the ER expression can also be evaluated as:

$$\text{ER} = \mu \cdot \left(\lim_{\mu \to 0} \limsup_{i \to \infty} \frac{1}{\mu} \mathbb{E} \, \|\widetilde{\boldsymbol{w}}_{i-1}\|^2_{\frac{1}{2}H} \right) \tag{4.96}$$

Performance Metrics

It is noteworthy to observe from (4.94) and (4.96) that both expressions for the MSD and ER involve squared norms of the error vector, $\widetilde{\boldsymbol{w}}_i$, in steady-state. For this reason, in the argument that follows we will focus on evaluating the limit superior of a *weighted* mean-square-error norm of the form $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2_{\Sigma}$, for some positive-definite weighting matrix Σ that we are free to choose. Then, by setting $\Sigma = I_M$ or $\Sigma = \frac{1}{2}H$, we will be able to arrive at the MSD and ER values.

Recall #3: Gradient Noise

Stochastic Gradient Algorithm

$$\boldsymbol{w}_i \;=\; \boldsymbol{w}_{i-1} \;-\; \mu\,\widehat{\nabla_{w^{\mathsf{T}}} J}(\boldsymbol{w}_{i-1}), \quad i \geq 0 \tag{4.5}$$

Gradient Noise

$$\boldsymbol{s}_i(\boldsymbol{w}) \;\stackrel{\Delta}{=}\; \widehat{\nabla_{w^{\mathsf{T}}} J}(\boldsymbol{w}) \;-\; \nabla_{w^{\mathsf{T}}} J(\boldsymbol{w}) \tag{4.6}$$

Gradient Noise Covariance Matrix

$$R_s \;\stackrel{\Delta}{=}\; \lim_{i\to\infty} \mathbb{E}\left[\boldsymbol{s}_i(w^o)\boldsymbol{s}_i^{\mathsf{T}}(w^o)\,|\,\boldsymbol{\mathcal{F}}_{i-1}\right] \tag{4.12}$$

Recall #4: Long-Term Dynamics

Lemma 4.3 (Long-term error dynamics). Assume the requirements under Assumptions 4.1 and 4.2 and condition (4.18) on the cost function and the gradient noise process hold. After sufficient iterations, $i \gg 1$, the error dynamics of the stochastic-gradient algorithm (4.5) is well-approximated by the following model (as confirmed by future result (4.70)):

$$\widetilde{\boldsymbol{w}}_i' \;=\; (I_M - \mu H)\widetilde{\boldsymbol{w}}_{i-1}' \;+\; \mu\, \boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \tag{4.55}$$

with the iterates denoted by $\widetilde{\boldsymbol{w}}_i'$ using the prime notation.

Recall #5: Size of Approx. Error

Lemma 4.3: For small-enough step-sizes, the error dynamics in steady-state is well-approximated by the long-term model:

$$\widetilde{\boldsymbol{w}}_i' \;=\; (I_M - \mu H)\widetilde{\boldsymbol{w}}_{i-1}' \;+\; \mu\, \boldsymbol{s}_i(\boldsymbol{w}_{i-1})$$

Specifically, it holds that:

$$\begin{cases} \limsup\limits_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i - \widetilde{\boldsymbol{w}}_i'\|^2 \;=\; O(\mu^2) \\ \limsup\limits_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \;=\; \limsup\limits_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i'\|^2 \;+\; O(\mu^{3/2}) \\ \limsup\limits_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|_H^2 \;=\; \limsup\limits_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i'\|_H^2 \;+\; O(\mu^{3/2}) \end{cases}$$

Performance Metrics

Theorem 4.7: For small-enough step-sizes:

$$H \triangleq \nabla_w^2 J(w^o)$$

$$\begin{aligned} \text{MSD} &= \frac{\mu}{2}\,\text{Tr}\left(H^{-1}R_s\right) \\ \text{ER} &= \frac{\mu}{4}\,\text{Tr}\left(R_s\right) \end{aligned}$$

The rate at which $\mathbb{E}\,\|\widetilde{w}_i\|^2$ approaches its steady-state region is approximated to first-order in μ by:

$$\alpha = 1 - 2\mu\lambda_{\min}(H)$$

MSE Performance

Theorem 4.7 (Mean-square-error performance: Real case). Assume the conditions under Assumptions 4.1, 4.2, and 4.4 on the cost function and the gradient noise process hold. Assume further that the step-size is sufficiently small to ensure mean-square stability, as already ascertained by Lemmas 3.1 and 4.4. Then, it holds that

$$\limsup_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \;=\; \frac{\mu}{2}\,\mathrm{Tr}\left(H^{-1}R_s\right) \;+\; O\left(\mu^{1+\gamma_m}\right) \tag{4.97}$$

$$\limsup_{i\to\infty} \mathbb{E}\,\{J(\boldsymbol{w}_{i-1}) - J(w^o)\} \;=\; \frac{\mu}{4}\,\mathrm{Tr}\left(R_s\right) \;+\; O\left(\mu^{1+\gamma_m}\right) \tag{4.98}$$

where

$$\gamma_m \;\stackrel{\Delta}{=}\; \frac{1}{2}\min\{1,\gamma\} > 0 \tag{4.99}$$

MSE Performance

with $0 < \gamma \leq 4$ from (4.19), while R_s and H are defined by (4.12) and (4.40). Consequently, the MSD and ER metrics defined by (4.94) and (4.96) for the stochastic-gradient algorithm (4.5) are given by the following expressions:

$$\text{MSD} = \frac{\mu}{2}\,\text{Tr}\left(H^{-1}R_s\right) \tag{4.100}$$

$$\text{ER} = \frac{\mu}{4}\,\text{Tr}\left(R_s\right) \tag{4.101}$$

Moreover, for $i \gg 1$, the rate at which the error variance, $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$, approaches its steady-state region (4.97) is well-approximated to first-order in μ by

$$\alpha = 1 - 2\mu\lambda_{\min}(H) \tag{4.102}$$

Proof

Proof. We introduce the eigen-decomposition $H = U\Lambda U^{\mathsf{T}}$, where U is orthogonal and Λ is diagonal with positive entries, and rewrite (4.55) in terms of transformed quantities:

$$\overline{\boldsymbol{w}}_i \;=\; (I - \mu\Lambda)\,\overline{\boldsymbol{w}}_{i-1} \;+\; \mu\overline{\boldsymbol{s}}_i(\boldsymbol{w}_{i-1}) \tag{4.103}$$

where $\overline{\boldsymbol{w}}_i = U^{\mathsf{T}}\widetilde{\boldsymbol{w}}'_i$ and $\overline{\boldsymbol{s}}_i(\boldsymbol{w}_{i-1}) = U^{\mathsf{T}}\boldsymbol{s}_i(\boldsymbol{w}_{i-1})$. Since the variables $\{\widetilde{\boldsymbol{w}}'_i, \overline{\boldsymbol{w}}_i\}$ are related to each other via an orthogonal transformation, it is clear that their Euclidean norms are identical and, therefore, $\mathbb{E}\,\|\overline{\boldsymbol{w}}_i\|^2 = \mathbb{E}\,\|\widetilde{\boldsymbol{w}}'_i\|^2$. It follows that we can rely on the mean-square-error of $\overline{\boldsymbol{w}}_i$ to evaluate the mean-square-deviation (MSD) of the long-term model (4.55). We proceed to derive an expression for the MSD by employing energy conservation arguments [6, 205, 206, 269].

Proof

Let Σ denote an arbitrary $M \times M$ diagonal matrix with positive entries that we are free to choose. Then, equating the weighted squared norms of both sides of (4.103) and taking expectations conditioned on the past history $\boldsymbol{\mathcal{F}}_{i-1}$ gives :

$$\mathbb{E}\left[\,\|\overline{\boldsymbol{w}}_i\|_{\Sigma}^2|\boldsymbol{\mathcal{F}}_{i-1}\,\right] \;=\; \|\overline{\boldsymbol{w}}_{i-1}\|_{\Sigma'}^2 \;+\; \mu^2\mathbb{E}\left[\,\|\overline{\boldsymbol{s}}_i(\boldsymbol{w}_{i-1})\|_{\Sigma}^2|\boldsymbol{\mathcal{F}}_{i-1}\,\right] \tag{4.104}$$

where the cross terms are annihilated on the right-hand side because $\mathbb{E}\left[\,\overline{\boldsymbol{s}_i}(\boldsymbol{w}_{i-1})|\boldsymbol{\mathcal{F}}_{i-1}\,\right] = 0$. Moreover, the weighting matrix Σ' is given by

$$\begin{aligned}\Sigma' \;&\stackrel{\Delta}{=}\; (I-\mu\Lambda)\Sigma(I-\mu\Lambda)\\ &=\; \Sigma - 2\mu\Lambda\Sigma \;+\; \mu^2\Lambda\Sigma\Lambda\end{aligned} \tag{4.105}$$

Proof

Taking expectations of both sides of (4.104) gives:

$$\mathbb{E}\,\|\overline{\boldsymbol{w}}_i\|_{\Sigma}^2 \;=\; \mathbb{E}\,\|\overline{\boldsymbol{w}}_{i-1}\|_{\Sigma'}^2 \;+\; \mu^2\,\mathbb{E}\,\|\overline{\boldsymbol{s}}_i(\boldsymbol{w}_{i-1})\|_{\Sigma}^2 \tag{4.106}$$

We now evaluate the two terms that appear on the right-hand side of this expression for $i \gg 1$. With regards to the first term, we use expression (4.105) for Σ' to note that:

$$\mathbb{E}\,\|\overline{\boldsymbol{w}}_{i-1}\|_{\Sigma'}^2 \overset{(4.105)}{=} \mathbb{E}\,\|\overline{\boldsymbol{w}}_{i-1}\|_{\Sigma-2\mu\Lambda\Sigma}^2 \;+\; \mu^2\,\mathbb{E}\,\|\overline{\boldsymbol{w}}_{i-1}\|_{\Lambda\Sigma\Lambda}^2 \tag{4.107}$$

Proof

Now since Σ and Λ are diagonal matrices with positive entries, we observe that the rightmost term satisfies:

$$\begin{aligned} \mathbb{E}\,\|\overline{\boldsymbol{w}}_{i-1}\|^2_{\Lambda\Sigma\Lambda} &\leq \rho(\Lambda^2)\cdot\rho(\Sigma)\cdot\mathbb{E}\,\|\overline{\boldsymbol{w}}_{i-1}\|^2 \\ &\leq \rho(\Lambda^2)\cdot\mathrm{Tr}(\Sigma)\cdot\mathbb{E}\,\|\overline{\boldsymbol{w}}_{i-1}\|^2 \end{aligned} \tag{4.108}$$

where $\rho(A)$ denotes the spectral radius of its matrix argument; obviously, for the matrices Σ and Λ, we have that $\rho(\Lambda)$ is equal to the largest entry in Λ while $\rho(\Sigma)$ is smaller than the trace of Σ. Combining the above result with the fact from (3.39) that the limit superior of $\mathbb{E}\,\|\overline{\boldsymbol{w}}_{i-1}\|^2$ is in the order of $O(\mu)$, we conclude from (4.107) that for $i \gg 1$:

Proof

$$\mathbb{E}\,\|\overline{\boldsymbol{w}}_{i-1}\|^2_{\Sigma'} \;=\; \mathbb{E}\,\|\overline{\boldsymbol{w}}_{i-1}\|^2_{\Sigma-2\mu\Lambda\Sigma} \;+\; \mathrm{Tr}(\Sigma)\cdot O(\mu^3) \qquad (4.109)$$

where we are keeping the factor $\mathrm{Tr}(\Sigma)$ explicit in the rightmost term for later use in (4.129).

We next evaluate the second term on the right-hand side of (4.106). To do so, we shall call upon the results of Lemma 4.1. We start by noting that

$$\begin{aligned}\mathbb{E}\,\|\overline{\boldsymbol{s}}_i(\boldsymbol{w}_{i-1})\|^2_{\Sigma} &= \mathrm{Tr}\left[\Sigma\,\mathbb{E}\,\left(\overline{\boldsymbol{s}}_i(\boldsymbol{w}_{i-1})\,(\overline{\boldsymbol{s}}_i(\boldsymbol{w}_{i-1}))^{\mathsf{T}}\right)\right]\\ &= \mathrm{Tr}\left[U\Sigma U^{\mathsf{T}}\,\mathbb{E}\,\left(\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\,(\boldsymbol{s}_i(\boldsymbol{w}_{i-1}))^{\mathsf{T}}\right)\right] \qquad (4.110)\end{aligned}$$

We showed in last lecture

$$\limsup_{i\to\infty}\left\|\mathbb{E}\,\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\,(\boldsymbol{s}_i(\boldsymbol{w}_{i-1}))^{\mathsf{T}} - R_s\right\| = O(\mu^{\gamma'/2}) \qquad (4.33)$$

Proof

where the covariance matrix $\mathbb{E}\,\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\,(\boldsymbol{s}_i(\boldsymbol{w}_{i-1}))^{\mathsf{T}}$ was already evaluated earlier in (4.33). Using that result, and the sub-multiplicative property of norms, namely, $\|AB\| \leq \|A\|\,\|B\|$, we conclude that:

$$\limsup_{i\to\infty} \left\| U\Sigma U^{\mathsf{T}}\mathbb{E}\,\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\,(\boldsymbol{s}_i(\boldsymbol{w}_{i-1}))^{\mathsf{T}} - U\Sigma U^{\mathsf{T}}R_s \right\| = O(\mu^{\gamma'/2}) \qquad (4.111)$$

where γ' was defined in (4.32) as $\gamma' = \min\{\gamma, 2\}$. Consequently, as stated earlier prior to (4.34), since $|\mathrm{Tr}(X)| \leq c\,\|X\|$ for any square matrix X, we have that:

$$\limsup_{i\to\infty} \left|\, \mathbb{E}\,\|\overline{\boldsymbol{s}}_i(\boldsymbol{w}_{i-1})\|_{\Sigma}^2 \;-\; \mathrm{Tr}(U\Sigma U^{\mathsf{T}}R_s)\,\right| \;=\; O(\mu^{\gamma'/2}) \;\triangleq\; b_1 \qquad (4.112)$$

Proof

in terms of the absolute value of the difference. We are denoting the value of the limit superior by the nonnegative number b_1; we know from (4.112) that $b_1 = O(\mu^{\gamma'/2})$. The same argument that led to (4.26) then leads to

$$\mathrm{Tr}(U\Sigma U^{\mathsf{T}} R_s) - b_o \;\leq\; \mathbb{E}\,\|\overline{\boldsymbol{s}}_i(\boldsymbol{w}_{i-1})\|_{\Sigma}^2 \;\leq\; \mathrm{Tr}(U\Sigma U^{\mathsf{T}} R_s) + b_o \qquad (4.113)$$

for $i \gg 1$ and for some nonnegative constant $b_o = O(\mu^{\gamma'/2})$. It follows from (4.113) that we can also write, for $i \gg 1$:

$$\mathbb{E}\,\|\overline{\boldsymbol{s}}_i(\boldsymbol{w}_{i-1})\|_{\Sigma}^2 \;=\; \mathrm{Tr}(U\Sigma U^{\mathsf{T}} R_s) \;+\; O(\mu^{\gamma'/2}) \qquad (4.114)$$

Proof

Substituting results (4.109) and (4.113) into the variance relation (4.106) we obtain for $i \gg 1$ that:

$$\mathbb{E}\,\|\overline{\boldsymbol{w}}_i\|_{\Sigma}^2 \;\leq\; \mathbb{E}\,\|\overline{\boldsymbol{w}}_{i-1}\|_{\Sigma-2\mu\Lambda\Sigma}^2 + \mu^2\left(\mathrm{Tr}(U\Sigma U^{\mathsf{T}}R_s) + b_o\right) + O(\mu^3) \quad (4.115)$$

$$\mathbb{E}\,\|\overline{\boldsymbol{w}}_i\|_{\Sigma}^2 \;\geq\; \mathbb{E}\,\|\overline{\boldsymbol{w}}_{i-1}\|_{\Sigma-2\mu\Lambda\Sigma}^2 + \mu^2\left(\mathrm{Tr}(U\Sigma U^{\mathsf{T}}R_s) - b_o\right) + O(\mu^3) \quad (4.116)$$

Using the sub-additivity and super-additivity properties of the limit superior and limit inferior operations, namely, for bounded sequences $a(i)$ and $b(i)$ [89, 144, 202]:

Proof

$$\limsup_{i\to\infty}\ (a(i)+b(i)) \quad \leq \quad \limsup_{i\to\infty}\ a(i) \ + \ \limsup_{i\to\infty}\ b(i) \qquad (4.117)$$

$$\liminf_{i\to\infty}\ (a(i)+b(i)) \quad \geq \quad \liminf_{i\to\infty}\ a(i) \ + \ \liminf_{i\to\infty}\ b(i) \qquad (4.118)$$

we conclude from (4.115) and (4.116) that

$$\limsup_{i\to\infty} \mathbb{E}\,\|\overline{\boldsymbol{w}}_i\|_{\Sigma}^2 \quad \leq \quad \limsup_{i\to\infty} \mathbb{E}\,\|\overline{\boldsymbol{w}}_{i-1}\|_{\Sigma-2\mu\Lambda\Sigma}^2 + \mu^2\left(\mathrm{Tr}(U\Sigma U^{\mathsf{T}}R_s)+b_o\right)+O(\mu^3) \qquad (4.119)$$

Proof

and

$$\begin{aligned}\liminf_{i\to\infty}\mathbb{E}\,\|\overline{\boldsymbol{w}}_i\|_{\Sigma}^2 \;\geq\; &\liminf_{i\to\infty}\mathbb{E}\,\|\overline{\boldsymbol{w}}_{i-1}\|_{\Sigma-2\mu\Lambda\Sigma}^2 + \\ &\mu^2\left(\mathrm{Tr}(U\Sigma U^{\mathsf{T}}R_s)-b_o\right)+O(\mu^3) \qquad (4.120)\end{aligned}$$

Grouping terms we get:

$$\limsup_{i\to\infty}\mathbb{E}\,\|\overline{\boldsymbol{w}}_i\|_{2\mu\Lambda\Sigma}^2 \;\leq\; \mu^2\left(\mathrm{Tr}(U\Sigma U^{\mathsf{T}}R_s)+b_o\right)+O(\mu^3) \qquad (4.121)$$

$$\liminf_{i\to\infty}\mathbb{E}\,\|\overline{\boldsymbol{w}}_i\|_{2\mu\Lambda\Sigma}^2 \;\geq\; \mu^2\left(\mathrm{Tr}(U\Sigma U^{\mathsf{T}}R_s)-b_o\right)+O(\mu^3) \qquad (4.122)$$

Proof

and, consequently, by eliminating a common factor μ from all terms and using the fact that the limit inferior of a sequence is upper bounded by its limit superior, we obtain the following inequality relation:

$$\mu\left(\mathrm{Tr}(U\Sigma U^{\mathsf{T}}R_s)-b_o\right)+O(\mu^2) \;\leq\; \liminf_{i\to\infty}\mathbb{E}\,\|\overline{\boldsymbol{w}}_i\|^2_{2\Lambda\Sigma}$$

$$\leq\; \limsup_{i\to\infty}\mathbb{E}\,\|\overline{\boldsymbol{w}}_i\|^2_{2\Lambda\Sigma} \;\leq\; \mu\left(\mathrm{Tr}(U\Sigma U^{\mathsf{T}}R_s)+b_o\right)+O(\mu^2) \tag{4.123}$$

Proof

Recalling that $b_o = O(\mu^{\gamma'/2})$ and $0 < \frac{\gamma'}{2} \leq 1$ so that μb_o dominates $O(\mu^2)$ for small μ, we conclude that the limit superior and limit inferior of the error variance satisfy:

$$\limsup_{i\to\infty} \mathbb{E}\,\|\overline{\boldsymbol{w}}_i\|^2_{2\Lambda\Sigma} = \mu \mathrm{Tr}(U\Sigma U^{\mathsf{T}} R_s) + O\left(\mu^{\min\{2,\,1+\frac{\gamma}{2}\}}\right) \quad (4.124)$$

$$\liminf_{i\to\infty} \mathbb{E}\,\|\overline{\boldsymbol{w}}_i\|^2_{2\Lambda\Sigma} = \mu \mathrm{Tr}(U\Sigma U^{\mathsf{T}} R_s) - O\left(\mu^{\min\{2,\,1+\frac{\gamma}{2}\}}\right) \quad (4.125)$$

Proof

Continuing with (4.124), and since we are free to choose Σ, we let $\Sigma = \frac{1}{2}\Lambda^{-1}$ so that the variance term on the left-hand side of (4.124) becomes $\mathbb{E}\,\|\overline{\boldsymbol{w}}_i\|^2$. Recalling that $\|\overline{\boldsymbol{w}}_i\|^2 = \|\widetilde{\boldsymbol{w}}'_i\|^2$ and noting that $U\Sigma U^{\mathsf{T}} = \frac{1}{2}H^{-1}$, we arrive at

$$\limsup_{i\to\infty}\ \mathbb{E}\,\|\widetilde{\boldsymbol{w}}'_i\|^2 = \frac{\mu}{2}\,\mathrm{Tr}\left(H^{-1}R_s\right)\ +\ O\left(\mu^{\min\left\{2,\,1+\frac{\gamma}{2}\right\}}\right) \tag{4.126}$$

Proof

However, we know from result (4.71) that the error variance of the stochastic-gradient algorithm (4.5) is within $O(\mu^{3/2})$ from the error variance of the long-term model, which is given by the above expression. We therefore need to adjust the exponent of μ inside the big-O term to arrive at the desired expression (4.97) where the factor of 2 is replaced by 3/2 since

$$\min\left\{\frac{3}{2},\, 2,\, 1+\frac{\gamma}{2}\right\} \;=\; \min\left\{\frac{3}{2},\, 1+\frac{\gamma}{2}\right\} \tag{4.127}$$

Proof

Likewise, if we select $\Sigma = \frac{1}{4} I_M$, then a similar argument leads to (4.98). Returning to (4.125), the argument that led to (4.126) would similarly imply that

$$\liminf_{i \to \infty} \mathbb{E} \|\widetilde{\boldsymbol{w}}_i\|^2 = \frac{\mu}{2} \operatorname{Tr}\left(H^{-1} R_s\right) - o(\mu) \tag{4.128}$$

Although this result is unnecessary for the argument in this proof, we nevertheless established it because it was used earlier in (4.84) while motivating the definition of the MSD metric.

Proof

With regards to the rate at which $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$ approaches its steady-state region (4.97) (and likewise for $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2_{\frac{1}{2}H}$), we refer back to (4.106) and substitute (4.109) and (4.114) to rewrite the former relation as follows for $i \gg 1$:

$$\mathbb{E}\,\|\overline{\boldsymbol{w}}_i\|^2_{\Sigma} = \mathbb{E}\,\|\overline{\boldsymbol{w}}_{i-1}\|^2_{(I_M-2\mu\Lambda)\Sigma} + \mu^2\mathrm{Tr}(\Sigma U^{\mathsf{T}} R_s U) + \mathrm{Tr}(\Sigma)\cdot o(\mu^2) \quad (4.129)$$

where we replaced the approximation error by $o(\mu^2)$ for brevity; it is sufficient to know for the current argument that the power of μ is strictly larger than two. For compactness of notation, we introduce the matrices

$$D \triangleq I_M - 2\mu\Lambda, \qquad Y \triangleq U^{\mathsf{T}} R_s U \quad (4.130)$$

Proof

It is clear that the matrix D is stable for sufficiently small step-sizes and, moreover,

$$\rho(D) \overset{(4.130)}{=} 1 - 2\mu\lambda_{\min}(H) \tag{4.131}$$

where we used the fact that the eigenvalues of Λ coincide with the eigenvalues of H and they are all positive. Therefore, $D^i \to 0$ as $i \to \infty$ and, moreover,

$$\sum_{n=0}^{\infty} D^n = (I_M - D)^{-1} = \frac{1}{2\mu}\Lambda^{-1} \tag{4.132}$$

so that

$$o(\mu^2) \cdot \mathrm{Tr}\left(\sum_{n=0}^{\infty} D^n\right) \overset{(4.132)}{=} o(\mu) \tag{4.133}$$

Proof

These two conclusions are used in the sequel. Indeed, from (4.129) we have that for any $i \gg 1$:

$$\mathbb{E}\,\|\overline{\boldsymbol{w}}_i\|_{\Sigma}^2 \;=\; \mathbb{E}\,\|\overline{\boldsymbol{w}}_{i-1}\|_{D\Sigma}^2 \;+\; \mu^2\mathrm{Tr}(\Sigma Y) \;+\; \mathrm{Tr}(\Sigma)\cdot o(\mu^2) \qquad (4.134)$$

By setting Σ successively equal to the choices $\{I_M,\ D,\ D^2,\ D^3\ \ldots\}$, and by iterating the above recursion, we deduce that

$$\mathbb{E}\,\|\overline{\boldsymbol{w}}_i\|^2 \;=\; \mathbb{E}\,\|\overline{\boldsymbol{w}}_{-1}\|_{D^{i+1}}^2 \;+\; \mu^2\sum_{n=0}^{i}\mathrm{Tr}(D^nY) \;+\; o(\mu^2)\cdot\sum_{n=0}^{i}\mathrm{Tr}(D^n) \qquad (4.135)$$

Proof

The first-term on the right-hand side corresponds to a transient component that dies out with time. The rate of its convergence towards zero determines the rate of convergence of $\mathbb{E}\,\|\overline{\boldsymbol{w}}_i\|^2$ towards its steady-state region. This rate can be characterized as follows. We express the weighted variance of $\overline{\boldsymbol{w}}_{-1}$ as the following trace relation in terms of its un-weighted covariance matrix:

$$\mathbb{E}\,\|\overline{\boldsymbol{w}}_{-1}\|^2_{D^{i+1}} \;=\; \mathbb{E}\,\left(\overline{\boldsymbol{w}}^*_{-1}D^{i+1}\overline{\boldsymbol{w}}_{-1}\right) \;=\; \mathrm{Tr}\left(D^{i+1}\mathbb{E}\,\overline{\boldsymbol{w}}_{-1}\overline{\boldsymbol{w}}^*_{-1}\right) \qquad (4.136)$$

Proof

Then, it is clear that the convergence rate of the transient component is dictated by $\rho(D)$ since this value characterizes the slowest rate at which the transient term dies out. We conclude that the convergence rate of $\mathbb{E}\,\|\overline{\boldsymbol{w}}_i\|^2$ towards the steady-state regime is also dictated by $\rho(D)$, which we can approximate to first-order in μ by expression (4.102).

Proof

Additionally, if desired, computing the limit superior of both sides of (4.135), and using (4.133), we can re-derive the MSD value for the algorithm in an alternative route as follows. Note that

$$\limsup_{i\to\infty} \mathbb{E}\,\|\overline{\boldsymbol{w}}_i\|^2 \;=\; \mu^2\left(\sum_{n=0}^{\infty}\mathrm{Tr}\,(D^nY)\right) \;+\; o(\mu) \qquad (4.137)$$

where the first term on the right-hand side is actually $O(\mu)$ and dominates the second term since

Proof

$$\begin{aligned}
\mu^2\left(\sum_{n=0}^{\infty}\mathrm{Tr}\left(D^nY\right)\right) &= \mu^2\mathrm{Tr}\left[\left(I_M+D+D^2+D^3+\ldots\right)Y\right] \\
&= \mu^2\mathrm{Tr}\left(\left(I_M-D\right)^{-1}Y\right) \\
&\overset{(4.132)}{=} \frac{\mu}{2}\mathrm{Tr}\left(\Lambda^{-1}Y\right) \\
&= O(\mu) \qquad (4.138)
\end{aligned}$$

Proof

If we now use the substitutions $Y = U^{\mathsf{T}} R_s U$, $\Lambda^{-1} = U^{\mathsf{T}} H^{-1} U$, and $\overline{\boldsymbol{w}}_i = U^{\mathsf{T}} \widetilde{\boldsymbol{w}}'_i$, we conclude that

$$\limsup_{i \to \infty} \mathbb{E} \|\widetilde{\boldsymbol{w}}'_i\|^2 = \frac{\mu}{2} \mathrm{Tr}\left(H^{-1} R_s\right) + o(\mu) \tag{4.139}$$

which is in agreement with (4.126).

□

MSE Performance

Results (4.100)–(4.101) are useful expressions that apply to general $\nu-$strongly convex functions $J(w)$ that satisfy Assumptions 4.1 and 4.3. The following example shows that the approximation error in expressions (4.97)–(4.98) can be replaced by $O(\mu^2)$ in the quadratic case.

Example #4.2

Example 4.2 (Quadratic cost functions). When $J(w)$ happens to be quadratic in w, as is the case with the mean-square-error cost of Example 3.1, then the matrices $\boldsymbol{H}_{i-1}$ and H defined by (4.38) and (4.39), respectively, will coincide with each other since the Hessian matrix $\nabla_w^2 J(w)$ will be constant for all w. Thus, in this case $\boldsymbol{H}_{i-1} \equiv H = \nabla_w^2 J(w^o)$. As a result, the perturbation term $\mu \boldsymbol{c}_{i-1}$ in (4.41) will be identically zero and recursions (4.37) and (4.55) will therefore coincide. Both models will then have the same MSD expressions. Therefore, we can rely on expression (4.126) without the need for the adjustment by $O(\mu^{3/2})$. We know from (4.16) that $\gamma = 2$ for mean-square-error costs. Using this value for γ in (4.126), we arrive at

Example #4.2

$$\limsup_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \;=\; \frac{\mu}{2}\,\mathrm{Tr}\left(H^{-1}R_s\right) \;+\; O\left(\mu^2\right) \tag{4.140}$$

with an approximation error in the order of $O(\mu^2)$ rather than the term $O(\mu^{3/2})$ that would result from (4.97)–(4.98). Likewise, we obtain

$$\limsup_{i\to\infty} \mathbb{E}\,\{J(\boldsymbol{w}_{i-1}) - J(w^o)\} \;=\; \frac{\mu}{4}\,\mathrm{Tr}\left(R_s\right) \;+\; O\left(\mu^2\right) \tag{4.141}$$

Example #4.2

In re-deriving this expression for the ER, we called upon expression (E.20) in the appendix where it is shown that for quadratic costs, expression (4.89) is replaced by the exact relation

$$\limsup_{i\to\infty} \mathbb{E}\,\{J(\boldsymbol{w}_{i-1}) - J(w^o)\} = \limsup_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2_{\frac{1}{2}H} \tag{4.142}$$

without the $O(\mu^{3/2})$ correction term that appeared in (4.89).

Example #4.2

The resulting expressions for the MSD and ER performance metrics will continue to be:

$$\text{MSD} = \frac{\mu}{2}\,\text{Tr}\left(H^{-1}R_s\right) \tag{4.143}$$

$$\text{ER} = \frac{\mu}{4}\,\text{Tr}\left(R_s\right) \tag{4.144}$$

Example #4.2

With regards to the convergence rate, we use $\gamma = 2$ (and, hence, $\gamma' = 2$) in (4.114) and recognize that the $o(\mu^2)$ term in (4.129) will be replaced by $O(\mu^3)$. Continuing with the derivation, we will then conclude that the approximation error $o(\mu)$ in (4.137) is replaced by $O(\mu^2)$ and the convergence rate expression (4.102) will still hold in the quadratic case:

$$\alpha = 1 - 2\mu\lambda_{\min}(H) \tag{4.145}$$

■

Recall #6 (LMS Adaptation)

Example 3.3 (Gradient noise). It is clear from the expressions in Examples 2.3 and 3.1 that the corresponding gradient noise process is given by:

$$\begin{aligned}
\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) &= \widehat{\nabla_{w^{\mathsf{T}}} J}(\boldsymbol{w}_{i-1}) \;-\; \nabla_{w^{\mathsf{T}}} J(\boldsymbol{w}_{i-1}) \\
&= 2\left(\boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{u}_i\right)\boldsymbol{w}_{i-1} \;-\; 2\boldsymbol{u}_i^{\mathsf{T}}[\boldsymbol{u}_i w^o + \boldsymbol{v}(i)] \;-\; 2R_u\boldsymbol{w}_{i-1} \;+\; 2R_u w^o \\
&= 2(R_u - \boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{u}_i)\widetilde{\boldsymbol{w}}_{i-1} \;-\; 2\boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{v}(i) \qquad (3.19)
\end{aligned}$$

From expression (3.19) we know that

$$\begin{aligned}
\boldsymbol{s}_i(w^o) &= -2\boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{v}(i) \qquad (4.13) \\
R_s &= 4\sigma_v^2 R_u \;\equiv\; R_{s,i}(w^o), \quad \text{for all } i \qquad (4.14)
\end{aligned}$$

Example #4.3

Example 4.3 (Performance of LMS adaptation). We reconsider the LMS recursion (3.13). We know from Example 3.3 and (4.13) that this situation corresponds to $H = 2R_u$ and $R_s = 4\sigma_v^2 R_u$. Substituting into (4.100)–(4.101) leads to the following well-known expressions for the performance of the LMS filter for sufficiently small step-sizes — see [96, 97, 100, 107, 114, 130, 206, 261, 262]:

$$\text{MSD} = \mu M \sigma_v^2 = O(\mu) \tag{4.146}$$

$$\text{EMSE} = \mu \sigma_v^2 \,\text{Tr}(R_u) = O(\mu) \tag{4.147}$$

where we are replacing ER by the notation EMSE, which is more common in the adaptive filtering literature.

Example #4.3

Figure 4.2 illustrates this situation numerically. The figure plots the evolution of the ensemble-average learning curve, $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$, over i; the curve is generated by averaging the trajectories $\{\|\widetilde{\boldsymbol{w}}_i\|^2\}$ over 2000 repeated experiments. The label on the vertical axis in the figure refers to the learning curve $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$ by writing MSD(i), with an iteration index i. Each experiment involves running the LMS recursion (3.13) on data $\{\boldsymbol{d}(i), \boldsymbol{u}_i\}$ generated according to the model $\boldsymbol{d}(i) = \boldsymbol{u}_i w^o + \boldsymbol{v}(i)$ with $M = 10$, $\sigma_v^2 = 0.010$, $R_u = 2I_M$, and using $\mu = 0.0025$. The unknown vector w^o is generated randomly and its norm is normalized to one. It is seen in the figure that the learning curve tends to the MSD value predicted by the theoretical expression (4.146).

■

Example #4.3

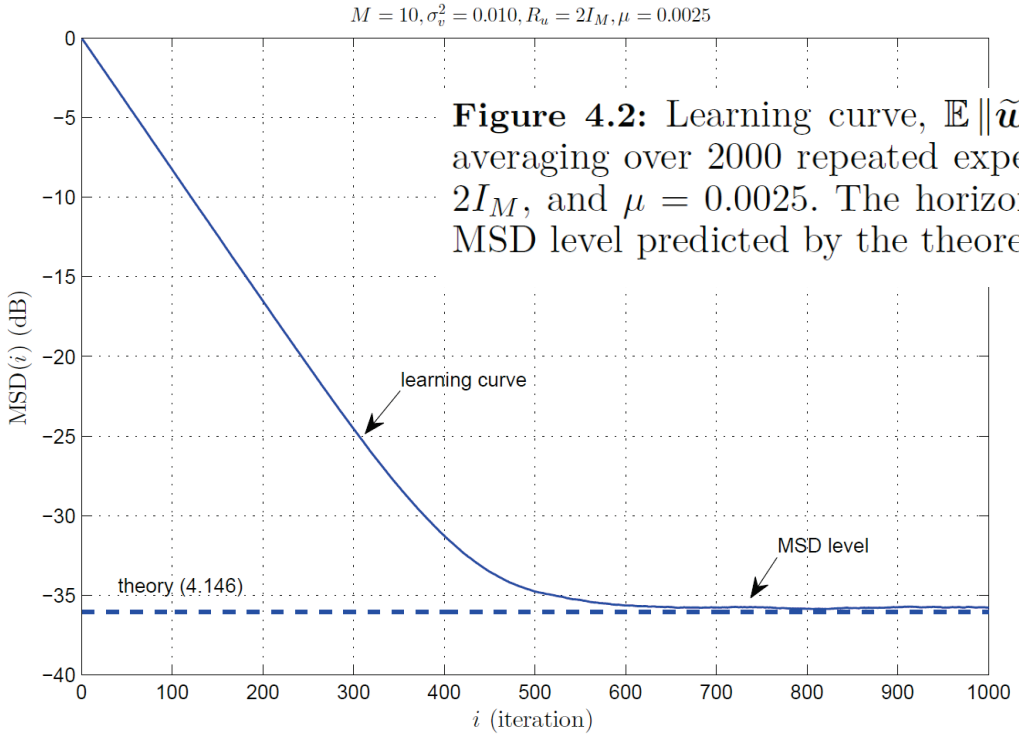


Figure 4.2: Learning curve, $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$, for the LMS rule (3.13) obtained by averaging over 2000 repeated experiments using $M = 10$, $\sigma_v^2 = 0.010$, $R_u = 2I_M$, and $\mu = 0.0025$. The horizontal dashed line indicates the steady-state MSD level predicted by the theoretical expression (4.146).

Example #4.4

Example 4.4 (Performance of logistic learners). We reconsider the stochastic-gradient algorithm (3.16) from Example 3.2 for logistic regression. The absolute component of the gradient noise in that example is given by

$$\boldsymbol{s}_i(w^o) = \rho w^o - \boldsymbol{\gamma}(i)\boldsymbol{h}_i\left(\frac{1}{1+e^{\boldsymbol{\gamma}(i)\boldsymbol{h}_i^{\mathsf{T}}w^o}}\right) \tag{4.148}$$

Example #4.4

with covariance matrix

$$R_s \triangleq \mathbb{E}\left\{\boldsymbol{h}_i\boldsymbol{h}_i^{\mathsf{T}} \cdot \left(\frac{1}{1+e^{\boldsymbol{\gamma}(i)\boldsymbol{h}_i^{\mathsf{T}}w^o}}\right)^2\right\} - \rho^2 w^o (w^o)^{\mathsf{T}} \tag{4.149}$$

Note in particular that $R_s \leq R_h$. Calling upon expression (4.101), we conclude that the excess-risk measure is given by

$$\text{ER} = \frac{\mu}{4}\,\text{Tr}\,(R_s) \leq \frac{\mu}{4}\,\text{Tr}\,(R_h) = O(\mu) \tag{4.150}$$

Example #4.4

Figure 4.3 illustrates this situation numerically. The figure plots the evolution of the ensemble-average excess-risk curve, $\mathbb{E}\{J(\boldsymbol{w}_{i-1}) - J(w^o)\}$, over i; the curve is generated by averaging the curves $\{J(\boldsymbol{w}_{i-1}) - J(w^o)\}$ over 100 repeated experiments. The label on the vertical axis in the figure refers to the learning curve $\mathbb{E}\{J(\boldsymbol{w}_{i-1}) - J(w^o)\}$ by writing $\mathrm{ER}(i)$, with an iteration index i. Each experiment involves running the logistic recursion (3.16) on data $\{\boldsymbol{\gamma}(i), \boldsymbol{h}_i\}$ with $M = 50$, $\rho = 10$, and $\mu = 1 \times 10^{-4}$. The data used for the

Example #4.4

simulation originate from the alpha data set [223]; we use the first 50 features for illustration purposes so that $M = 50$. To generate the trajectories for the experiments in this example, the optimal w^o and the gradient noise covariance matrix, R_s, are first estimated off-line by applying a batch algorithm to all data points. For the data used in this example we have $\mathrm{Tr}(R_s) \approx 131.48$ and $\mathrm{Tr}(R_h) \approx 528.10$. It is seen in the figure that the learning curve tends to the ER value predicted by the theoretical expression (4.150).

■

Example #4.4

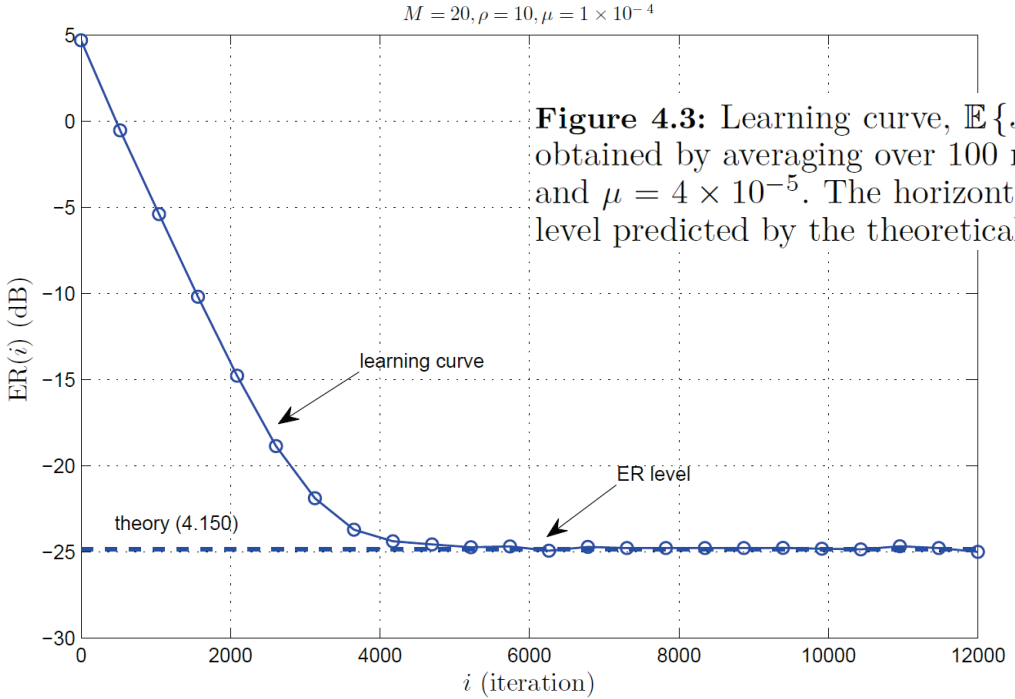


Figure 4.3: Learning curve, $\mathbb{E}\{J(\boldsymbol{w}_{i-1} - J(w^o)\}$, for the logistic rule (3.16) obtained by averaging over 100 repeated experiments using $M = 50$, $\rho = 10$, and $\mu = 4 \times 10^{-5}$. The horizontal dashed line indicates the steady-state ER level predicted by the theoretical expression (4.150).

Example #4.5

Example 4.5 (Performance of online learners). More generally, consider a stand-alone learner receiving a streaming sequence of independent data vectors $\{\boldsymbol{x}_i, i \geq 0\}$ that arise from some fixed probability distribution $\mathcal{X}$. The goal is to learn the vector w^o that optimizes some $\nu-$strongly convex risk function $J(w)$ defined in terms of a loss function [236, 252]:

$$w^o \triangleq \arg\min_w J(w) = \arg\min_w \mathbb{E}\, Q(w; \boldsymbol{x}_i) \tag{4.151}$$

The learner seeks w^o by running the stochastic-gradient algorithm:

$$\boldsymbol{w}_i = \boldsymbol{w}_{i-1} - \mu \nabla_{w^\mathsf{T}} Q(\boldsymbol{w}_{i-1}; \boldsymbol{x}_i), \quad i \geq 0 \tag{4.152}$$

Example #4.5

so that the gradient noise vector is given by

$$\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \;=\; \nabla_{w^{\mathsf{T}}} Q(\boldsymbol{w}_{i-1};\boldsymbol{x}_i) \;-\; \nabla_{w^{\mathsf{T}}} J(\boldsymbol{w}_{i-1}) \tag{4.153}$$

Since $\nabla_w J(w^o) = 0$, and since the distribution of $\boldsymbol{x}_i$ is assumed stationary, it follows that the covariance matrix of $\boldsymbol{s}_i(w^o)$ is constant and given by

$$R_s = \mathbb{E}\,\nabla_{w^{\mathsf{T}}} Q(w^o;\boldsymbol{x}_i)\nabla_w Q(w^o;\boldsymbol{x}_i) \tag{4.154}$$

The excess-risk measure that will result from this stochastic implementation is then given by (4.101) so that

$$\mathrm{ER} = \frac{\mu}{4}\mathrm{Tr}(R_s) \tag{4.155}$$

■

Complex Domain

Course EE210B
Spring Quarter 2015

Proc. IEEE, vol. 102, no. 4, pp. 460-497, April 2014.
Foundations and Trends in Machine Learning, vol. 7, no. 4-5, pp. 311-801, July 2014.

Complex Domain

We now extend the performance results of the previous sections to the complex domain in which case the argument $w \in \mathbb{C}^M$ is complex-valued. We explained in Sec. 3.6 that the strongly convex function, $J(w) \in \mathbb{R}$, is now required to satisfy condition (3.114), namely,

$$0 < \frac{\nu}{h} I_{hM} \leq \nabla_w^2 J(w) \leq \frac{\delta}{h} I_{hM} \tag{4.156}$$

Complex Domain

in terms of the data-type variable

$$h \triangleq \begin{cases} 1, & \text{when } w \text{ is real} \\ 2, & \text{when } w \text{ is complex} \end{cases} \tag{4.157}$$

As was the case in the real domain, we continue to assume that the now $2M \times 2M$ Hessian Hessian matrix of $J(w)$ satisfies the local Lipschitz condition (4.18).

Stochastic Gradient Algorithm

We also explained that the constant step-size stochastic gradient recursion is given by

$$\boldsymbol{w}_i \;=\; \boldsymbol{w}_{i-1} \;-\; \mu\,\widehat{\nabla_{w^*} J}(\boldsymbol{w}_{i-1}), \quad i \geq 0 \tag{4.158}$$

and that the gradient noise process is now complex-valued as well, i.e.,

$$\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \;\triangleq\; \widehat{\nabla_{w^*} J}(\boldsymbol{w}_{i-1}) \;-\; \nabla_{w^*} J(\boldsymbol{w}_{i-1}) \tag{4.159}$$

Conditions on Gradient Noise

The first and second-order moments of this noise process are assumed to satisfy the same conditions in Assumption 3.4. The result in Theorem 4.8 further ahead extends the conclusion from Theorem 4.7 to the complex case. Comparing the performance expressions in the lemma below to the earlier expressions in the real case from Theorem 4.7, we observe that in the MSD case, two moment matrices are now involved, and which are denoted by R_s and R_q. These matrices are defined as follows.

Conditions on Gradient Noise

For any $\boldsymbol{w} \in \boldsymbol{\mathcal{F}}_{i-1}$, we introduce the extended gradient noise vector of size $2M \times 1$:

$$\boldsymbol{s}_i^e(\boldsymbol{w}) \triangleq \begin{bmatrix} \boldsymbol{s}_i(\boldsymbol{w}) \\ (\boldsymbol{s}_i^*(\boldsymbol{w}))^{\mathsf{T}} \end{bmatrix} \tag{4.160}$$

where we are using the superscript "e" to denote the extended variable. We then let

$$R_{s,i}^e(\boldsymbol{w}) \triangleq \mathbb{E}\,[\,\boldsymbol{s}_i^e(\boldsymbol{w})\boldsymbol{s}_i^{e*}(\boldsymbol{w})\,|\,\boldsymbol{\mathcal{F}}_{i-1}\,] \tag{4.161}$$

denote the conditional second-order moment of this extended noise process. It is a $2M \times 2M$ matrix whose blocks are given by

Conditions on Gradient Noise

$$R_{s,i}^{e}(\boldsymbol{w}) = \begin{bmatrix} \mathbb{E}\,\boldsymbol{s}_i(\boldsymbol{w})\boldsymbol{s}_i^{*}(\boldsymbol{w}) & \mathbb{E}\,\boldsymbol{s}_i(\boldsymbol{w})\boldsymbol{s}_i^{\mathsf{T}}(\boldsymbol{w}) \\ \mathbb{E}\left(\boldsymbol{s}_i(\boldsymbol{w})\boldsymbol{s}_i^{\mathsf{T}}(\boldsymbol{w})\right)^{*} & \mathbb{E}\left(\boldsymbol{s}_i(\boldsymbol{w})\boldsymbol{s}_i^{*}(\boldsymbol{w})\right)^{\mathsf{T}} \end{bmatrix} \tag{4.162}$$

Compared with the earlier definition (4.11) in the real case, we see that now two moment quantities of the form $\mathbb{E}\,\boldsymbol{s}_i(\boldsymbol{w})\boldsymbol{s}_i^{*}(\boldsymbol{w})$ and $\mathbb{E}\,\boldsymbol{s}_i(\boldsymbol{w})\boldsymbol{s}_i^{\mathsf{T}}(\boldsymbol{w})$ appear in (4.162), with the first one using conjugate transposition and the second one using standard transposition. We assume that, in the limit, these moment matrices tend to constant values when evaluated at w^o and we denote their limits by

Conditions on Gradient Noise

$$R_s \triangleq \lim_{i\to\infty} \mathbb{E}\,[\,\boldsymbol{s}_i(w^o)\boldsymbol{s}_i^*(w^o)\,|\,\boldsymbol{\mathcal{F}}_{i-1}\,] \tag{4.163}$$

$$R_q \triangleq \lim_{i\to\infty} \mathbb{E}\,\left[\,\boldsymbol{s}_i(w^o)\boldsymbol{s}_i^{\mathsf{T}}(w^o)\,|\,\boldsymbol{\mathcal{F}}_{i-1}\,\right] \tag{4.164}$$

Comparing (4.163) with (4.164) we see that $\boldsymbol{s}_i^*(\boldsymbol{w})$ is used in the expression for R_s while $\boldsymbol{s}_i^{\mathsf{T}}(\boldsymbol{w})$ is used in the expression for R_q. The two moment matrices, $\{R_s, R_q\}$, are in general different. It is the first moment, R_s, that is an actual covariance matrix in the complex domain (and is therefore Hermitian and non-negative definite), while the second moment, R_q, is symmetric. Both matrices $\{R_s, R_q\}$ are needed to

Conditions on Gradient Noise

characterize the second-order moment of $\boldsymbol{s}_i(w^o)$ in the complex domain. When $\boldsymbol{s}_i(w^o)$ happens to be real-valued, then R_s and R_q will obviously coincide. Nevertheless, we will continue to use the universal notation R_s (and not R_q) to denote the covariance matrix of $\boldsymbol{s}_i(w^o)$. In other words, whether $\boldsymbol{s}_i(w^o)$ is real or complex-valued, the notation R_s will always denote its limiting covariance matrix:

$$R_s \triangleq \begin{cases} \lim_{i\to\infty} \mathbb{E}\left[\boldsymbol{s}_i(w^o)\boldsymbol{s}_i^{\mathsf{T}}(w^o)\,|\,\boldsymbol{\mathcal{F}}_{i-1}\right] & \text{(for real data)} \\ \lim_{i\to\infty} \mathbb{E}\left[\boldsymbol{s}_i(w^o)\boldsymbol{s}_i^{*}(w^o)\,|\,\boldsymbol{\mathcal{F}}_{i-1}\right] & \text{(for complex data)} \end{cases} \tag{4.165}$$

Conditions on Gradient Noise

Before establishing the next result, we mention that the smoothness condition (4.19) takes the following form in the complex case in terms of the extended covariance matrix:

$$\left\| R^e_{s,i}(w^o + \Delta w) - R^e_{s,i}(w^o) \right\| \leq \kappa_2 \|\Delta w\|^\gamma \qquad (4.166)$$

for small perturbations $\|\Delta w\| \leq \epsilon$, and for some constant $\kappa_2 \geq 0$ and exponent $0 < \gamma \leq 4$.

MSE Performance

Theorem 4.8 (Mean-square-error performance: Complex case). Assume the cost function $J(w)$ satisfies conditions (4.156) and (4.18). Assume further that the gradient noise process satisfies the conditions in Assumption 3.4 and the smoothness condition (4.166), and that the step-size is sufficiently small to ensure mean-square stability, as already ascertained by Lemma 3.5. Then, it holds that

$$\limsup_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \;=\; \frac{\mu}{4}\,\mathrm{Tr}\left(H^{-1}\begin{bmatrix} R_s & R_q \\ R_q^* & R_s^{\mathsf{T}} \end{bmatrix}\right) \;+\; O\left(\mu^{1+\gamma_m}\right) \tag{4.167}$$

$$\limsup_{i\to\infty} \mathbb{E}\,\{J(\boldsymbol{w}_{i-1}) - J(w^o)\} \;=\; \frac{\mu}{2}\,\mathrm{Tr}\left(R_s\right) \;+\; O\left(\mu^{1+\gamma_m}\right) \tag{4.168}$$

where

$$\gamma_m \;\stackrel{\Delta}{=}\; \frac{1}{2}\min\{1,\gamma\} > 0 \tag{4.169}$$

MSE Performance

and $\gamma \in (0, 4]$ is from (4.166). Moreover, $\{R_s, R_q\}$ are defined by (4.163)–(4.164) and $H = \nabla_w^2 J(w^o)$ is $2M \times 2M$. Consequently, the MSD and ER metrics for the complex stochastic-gradient algorithm (4.158) are given by:

$$\text{MSD} = \frac{\mu}{4} \text{Tr}\left(H^{-1} \begin{bmatrix} R_s & R_q \\ R_q^* & R_s^{\mathsf{T}} \end{bmatrix}\right) \tag{4.170}$$

$$\text{ER} = \frac{\mu}{2} \text{Tr}\left(R_s\right) \tag{4.171}$$

Moreover, for $i \gg 1$, the rate at which the error variance, $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$, approaches its steady-state region is well-approximated to first-order in μ by

$$\alpha = 1 - 2\mu\lambda_{\min}(H) \tag{4.172}$$

When $J(w)$ is quadratic in w, the approximation errors in (4.167)–(4.168) are replaced by $O(\mu^2)$.

Proof

Proof. We explained in the proof of Lemma 3.5 that results for the complex recursion (4.158) can be recovered by working with the following recursion in terms of an extended $2M \times 1$ *real* variable $\boldsymbol{v}_i$:

$$\boldsymbol{v}_i \;=\; \boldsymbol{v}_{i-1} \;-\; \mu' \,\widehat{\nabla_{v^{\mathsf{T}}} J}(\boldsymbol{v}_{i-1}) \tag{4.173}$$

where $\mu' = \mu/2$ and $\boldsymbol{v}_i = \mathrm{col}\{\boldsymbol{x}_i, \boldsymbol{y}_i\}$ in terms of the real and imaginary parts of $\boldsymbol{w}_i = \boldsymbol{x}_i + j\boldsymbol{y}_i$. The gradient noise process that is associated with this $v-$domain recursion was denoted by

$$\boldsymbol{t}_i(\boldsymbol{v}_{i-1}) \;\triangleq\; \widehat{\nabla_{v^{\mathsf{T}}} J}(\boldsymbol{v}_{i-1}) \;-\; \nabla_{v^{\mathsf{T}}} J(\boldsymbol{v}_{i-1}) \tag{4.174}$$

Proof

and it was shown in (3.150) to be given by

$$\boldsymbol{t}_i(\boldsymbol{v}_{i-1}) \;=\; 2\left[\begin{array}{c}\boldsymbol{s}_{R,i}(\boldsymbol{w}_{i-1})\\ \boldsymbol{s}_{I,i}(\boldsymbol{w}_{i-1})\end{array}\right] \tag{4.175}$$

in terms of the real and imaginary parts of the original gradient noise vector $\boldsymbol{s}_i(\boldsymbol{w}_{i-1})$, defined by (4.159):

$$\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \;\stackrel{\Delta}{=}\; \boldsymbol{s}_{R,i}(\boldsymbol{w}_{i-1}) \;+\; j\boldsymbol{s}_{I,i}(\boldsymbol{w}_{i-1}) \tag{4.176}$$

Therefore, in order to apply the results of Theorem 4.7 to the $v-$domain recursion (4.173) under the conditions in Assumption 3.4, we need to determine two quantities:

Proof

(a) First, we need to determine an expression for the Hessian matrix of the cost function $J(v)$, in the $v-$domain, which will play the role of the matrix H in expressions (4.100)–(4.101).

(b) Second, we need to determine an expression for the second-order moment of the noise component, $\boldsymbol{t}_i(v^o)$, which will play the role of R_s in the same expressions (4.100)–(4.101).

With regards to the Hessian matrix, we recall result (B.26) from the appendix, which relates the Hessian matrix of $J(v)$ in the $v-$domain to the complex Hessian matrix of $J(w)$ in the $w-$domain, and use it to write

$$\nabla_v^2 J(v^o) = D^* \left[\nabla_w^2 J(w^o)\right] D = D^* H D \tag{4.177}$$

Proof

in terms of the matrix D defined by (B.27) and which satisfies $DD^* = 2I_{2M}$. Note that this result also implies that $\nabla_v^2 J(v^o)$ is similar to $2H$ so that

$$\lambda_{\min}\left(\nabla_v^2 J(v^o)\right) = 2\lambda_{\min}(H) \tag{4.178}$$

With regards to the second-order moment of the absolute component of $\boldsymbol{t}_i(\boldsymbol{v}_{i-1})$, we let

$$R_t \triangleq \lim_{i\to\infty} \mathbb{E}\left[\boldsymbol{t}_i(v^o)\boldsymbol{t}_i^{\mathsf{T}}(v^o) \,|\, \boldsymbol{\mathcal{F}}_{i-1}\right] \tag{4.179}$$

Using (4.175), as well as definitions (4.163)–(4.164) for the second-order moments $\{R_s, R_q\}$ associated with the original gradient noise component, $\boldsymbol{s}_i(w^o)$, it can be verified that

Proof

$$
\begin{aligned}
DR_tD^* &= 4\cdot\lim_{i\to\infty}\mathbb{E}\left(\begin{bmatrix} \boldsymbol{s}_i(w^o)\boldsymbol{s}_i^*(w^o) & \boldsymbol{s}_i(w^o)\boldsymbol{s}_i^{\mathsf{T}}(w^o) \\ \big(\boldsymbol{s}_i(w^o)\boldsymbol{s}_i^{\mathsf{T}}(w^o)\big)^* & \big(\boldsymbol{s}_i(w^o)\boldsymbol{s}_i^*(w^o)\big)^{\mathsf{T}} \end{bmatrix}\right) \\
&\triangleq 4\begin{bmatrix} R_s & R_q \\ R_q^* & R_s^{\mathsf{T}} \end{bmatrix} \qquad (4.180)
\end{aligned}
$$

We already know from (3.152)–(3.153) and (3.168) that the second and fourth-order moments of the gradient noise process $\boldsymbol{t}_i(\boldsymbol{v}_{i-1})$ satisfy conditions similar to (4.9)–(4.10) and (4.67) in the real case. Therefore, the results of Theorem 4.7 can be applied to the $v-$domain recursion (4.173). Let

$$
m \triangleq 1+\gamma_m \qquad (4.181)
$$

Proof

We conclude from the expressions in Theorem 4.7 that the limit superior for each of the error variance and the mean fluctuation for the $v-$domain recursion are given by (using $\mu' = \mu/2$)

$$\begin{aligned}
\limsup_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{v}}_i\|^2 &= \frac{\mu'}{2}\,\mathrm{Tr}\left(\left[\nabla_v^2\, J(v^o)\right]^{-1} R_t\right) + O((\mu')^m) \\
&= \frac{\mu}{4}\,\mathrm{Tr}\left(D^{-1}H^{-1}D^{-*}R_t\right) + O(\mu^m) \\
&= \frac{\mu}{4}\,\mathrm{Tr}\left(H^{-1}D^{-*}R_tD^{-1}\right) + O(\mu^m) \\
&= \frac{\mu}{4}\,\mathrm{Tr}\left(H^{-1}\frac{1}{2}DR_t\frac{1}{2}D^{*}\right) + O(\mu^m) \\
&= \frac{\mu}{4}\,\mathrm{Tr}\left(H^{-1}\begin{bmatrix} R_s & R_q \\ R_q^{*} & R_s^{\mathsf{T}} \end{bmatrix}\right) + O(\mu^m) \qquad (4.182)
\end{aligned}$$

Proof

and

$$
\begin{aligned}
\limsup_{i\to\infty} \mathbb{E}\,\{J(\boldsymbol{v}_{i-1}) - J(v^o)\} &= \frac{\mu'}{4}\,\mathrm{Tr}\,(R_t) + O((\mu')^m) \\
&= \frac{\mu}{8}\,\mathrm{Tr}\,\left(D^{-1}DR_t\right) + O(\mu^m) \\
&= \frac{\mu}{8}\,\mathrm{Tr}\,\left(DR_tD^{-1}\right) + O(\mu^m) \\
&= \frac{\mu}{16}\,\mathrm{Tr}\,(DR_tD^*) + O(\mu^m) \\
&= \frac{\mu}{4}\,\mathrm{Tr}\left(\begin{bmatrix} R_s & R_q \\ R_q^* & R_s^{\mathsf{T}} \end{bmatrix}\right) + O(\mu^m) \\
&= \frac{\mu}{2}\,\mathrm{Tr}\,(R_s) + O(\mu^m) \qquad (4.183)
\end{aligned}
$$

Proof

Finally, using (4.172) we conclude that the convergence rate in the $v-$domain is given by the following expression to first-order in μ:

$$\begin{aligned} \alpha &= 1 - 2\mu' \lambda_{\min}(\nabla_v^2 J(v^o)) \\ &= 1 - 2\left(\frac{\mu}{2}\right) 2\lambda_{\min}(H) \\ &\overset{(4.178)}{=} 1 - 2\mu\lambda_{\min}(H) \end{aligned} \tag{4.184}$$

□

Example #4.6

Example 4.6 (Performance of complex LMS adaptation). We reconsider the complex LMS recursion (3.125) from Example 3.4. In this case we have

$$R_s = \sigma_v^2 R_u, \quad H = \begin{bmatrix} R_u & 0 \\ 0 & R_u^{\mathsf{T}} \end{bmatrix}, \qquad G_k = \sigma_{v,k}^2 \begin{bmatrix} R_u & \times \\ \times & R_u^{\mathsf{T}} \end{bmatrix} \tag{4.185}$$

where the block off-diagonal entries of G_k are not be not needed because H_k is block-diagonal. Substituting into (4.170) and (4.171) we find that the MSD and ER performance levels are given by

Example #4.6

$$\text{MSD} = \frac{\mu M \sigma_v^2}{2} \tag{4.186}$$

$$\text{ER} = \frac{\mu \sigma_v^2}{2} \text{Tr}\,(R_s) \tag{4.187}$$

■

Extended Noise Vector

It is useful to remark that the block matrix that appears in expression (4.170) for the MSD is equal to the limiting covariance matrix of the extended gradient noise vector when evaluated at $w = w^o$:

$$\boldsymbol{s}_i^e(w^o) \triangleq \begin{bmatrix} \boldsymbol{s}_i(w^o) \\ (\boldsymbol{s}_i^*(w^o))^{\mathsf{T}} \end{bmatrix} \tag{4.188}$$

Specifically, it holds that

$$\begin{bmatrix} R_s & R_q \\ R_q^* & R_s^{\mathsf{T}} \end{bmatrix} = \lim_{i\to\infty} \mathbb{E}\,[\,\boldsymbol{s}_i^e(w^o)\,(\boldsymbol{s}_i^e(w^o))^*\,|\,\boldsymbol{\mathcal{F}}_{i-1}\,] \triangleq R_s^e \tag{4.189}$$

Extended Noise Vector

If we use R_s^e to denote this extended covariance matrix, then we can rewrite the MSD and ER expressions (4.170)–(4.171) in the equivalent forms:

$$\text{MSD} = \frac{\mu}{4}\,\text{Tr}\left(H^{-1}R_s^e\right) \tag{4.190}$$

$$\text{ER} = \frac{\mu}{4}\,\text{Tr}\left(R_s^e\right) \tag{4.191}$$

End of Lecture

Proc. IEEE, vol. 102, no. 4, pp. 460-497, April 2014.
Foundations and Trends in Machine Learning, vol. 7, no. 4-5, pp. 311-801, July 2014.