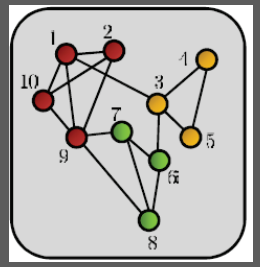

INFERENCE OVER NETWORKS

LECTURE #11: Stability and Long-Term Dynamics

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.

Setting

We established in Lemmas 3.3 and 3.7, for both cases of real and complex data, that the use of a stochastic-gradient algorithm with a decaying step-size sequence of the form $\mu(i) = \tau/(i+1)$ guarantees the almost sure convergence of the iterate $\boldsymbol{w}_i$ to w^o. However, the largest rate of convergence that is attainable under this construction is in the order of $O(1/i)$, namely, for large enough i it holds that

$$\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \;=\; O(1/i) \tag{4.1}$$

Setting: Constant Step-Size

On the other hand, when a constant step-size, μ, is used, we established in Lemmas 3.1 and 3.5 that the stochastic-gradient algorithm is mean-square stable in the sense that the error variance enters a bounded region whose size is in the order of $O(\mu)$, namely, for large enough i it now holds that

$$\limsup_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \;=\; O(\mu) \tag{4.2}$$

Setting: Constant Step-Size

More interestingly, we showed that convergence towards this bounded region occurs at a faster *geometric* rate and is in the order of $O(\alpha^i)$ for some $0 \leq \alpha < 1$. In other words, although some degradation in steady-state performance occurs, the convergence rate is nevertheless exponential.

In this lecture, we will motivate a long-term model that will enable us to assess in the next lecture the size of the fluctuations of of $\boldsymbol{w}_i$ around w^o in steady-state, as $i \to \infty$, for both cases of real and complex data.

Operating Regimes

Definition 4.1 (Operating regimes). The term "steady-state regime" will refer to the operation of the stochastic-gradient implementation after sufficient iterations have elapsed, i.e., as $i \to \infty$. Likewise, the term "slow adaptation regime" will refer to the operation of the stochastic-gradient implementation with a sufficiently small step-size, i.e., as $\mu \to 0$.

"**Steady-state regime:**" operation after sufficient iterations $(i \to \infty)$.

"**Slow adaptation regime:**" operation with small step-size $(\mu \to 0)$.

Smoothness Conditions Real 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.

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

Assumption 4.1 (Conditions on cost function). The cost function $J(w)$ is twice-differentiable and satisfies (4.4) for some positive parameters $\nu \leq \delta$. Condition (4.4) is equivalent to requiring $J(w)$ to be $\nu-$strongly convex and for its gradient vector to be $\delta-$Lipschitz as in (2.14) and (2.17), respectively.

Summary of Conditions

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

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

Conditions on Gradient Noise

Assumption 4.2 (Conditions on gradient noise). It is assumed that the first and second-order conditional moments of the gradient noise process satisfy the following conditions for any $\boldsymbol{w} \in \boldsymbol{\mathcal{F}}_{i-1}$:

$$\mathbb{E}\,[\,\boldsymbol{s}_i(\boldsymbol{w})\,|\,\boldsymbol{\mathcal{F}}_{i-1}\,] = 0 \tag{4.7}$$

$$\mathbb{E}\,\big[\,\|\boldsymbol{s}_i(\boldsymbol{w})\|^2\,|\,\boldsymbol{\mathcal{F}}_{i-1}\,\big] \leq \bar{\beta}^2\,\|\boldsymbol{w}\|^2 + \bar{\sigma}_s^2 \tag{4.8}$$

almost surely, for some nonnegative scalars $\bar{\beta}^2$ and $\bar{\sigma}_s^2$. These conditions were shown in (3.31)–(3.32) to imply that the gradient noise process satisfies for any $\boldsymbol{w}_{i-1} \in \boldsymbol{\mathcal{F}}_{i-1}$:

$$\mathbb{E}\,[\,\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\,|\,\boldsymbol{\mathcal{F}}_{i-1}\,] = 0 \tag{4.9}$$

$$\mathbb{E}\,\big[\,\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2\,|\,\boldsymbol{\mathcal{F}}_{i-1}\,\big] \leq \beta^2\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + \sigma_s^2 \tag{4.10}$$

almost surely, for some nonnegative scalars β^2 and σ_s^2, and where $\widetilde{\boldsymbol{w}}_{i-1} = w^o - \boldsymbol{w}_{i-1}$.

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.

Recall #1 (Example #3.1)

Example 3.1 (LMS adaptation). Let $\boldsymbol{d}(i)$ denote a streaming sequence of zero-mean random variables with variance $\sigma_d^2 = \mathbb{E}\,\boldsymbol{d}^2(i)$. Let $\boldsymbol{u}_i$ denote a streaming sequence of $1 \times M$ independent zero-mean random vectors with covariance matrix $R_u = \mathbb{E}\,\boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{u}_i > 0$. Both processes $\{\boldsymbol{d}(i), \boldsymbol{u}_i\}$ are assumed to be jointly wide-sense stationary. The cross-covariance vector between $\boldsymbol{d}(i)$ and $\boldsymbol{u}_i$ is denoted by $r_{du} = \mathbb{E}\,\boldsymbol{d}(i)\boldsymbol{u}_i^{\mathsf{T}}$. The data $\{\boldsymbol{d}(i), \boldsymbol{u}_i\}$ are assumed to be related via a linear regression model of the form:

$$\boldsymbol{d}(i) \;=\; \boldsymbol{u}_i w^o \;+\; \boldsymbol{v}(i) \tag{3.6}$$

Recall #1 (Example #3.1)

for some unknown parameter vector w^o, and where $\boldsymbol{v}(i)$ is a zero-mean white-noise process with power $\sigma_v^2 = \mathbb{E}\,\boldsymbol{v}^2(i)$ and assumed independent of $\boldsymbol{u}_j$ for all i, j. Observe that we are using parentheses to represent the time-dependency of a scalar variable, such as writing $\boldsymbol{d}(i)$, and subscripts to represent the time-dependency of a vector variable, such as writing $\boldsymbol{u}_i$. This convention will be used throughout this work. In a manner similar to Example 2.1, we again pose the problem of estimating w^o by minimizing the mean-square error cost

$$J(w) = \mathbb{E}\,(\boldsymbol{d}(i) - \boldsymbol{u}_i w)^2 \;\equiv\; \mathbb{E}\,Q(w; \boldsymbol{x}_i) \tag{3.7}$$

Recall #1 (Example #3.1)

$$r_{du} \approx \boldsymbol{d}(i)\boldsymbol{u}_i^{\mathsf{T}}, \quad R_u \approx \boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{u}_i \tag{3.9}$$

By doing so, the true gradient vector is approximated by:

$$\widehat{\nabla_{w^{\mathsf{T}}} J(w)} = 2\left[\boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{u}_i w - \boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{d}(i)\right] = \nabla_{w^{\mathsf{T}}} Q(w;\boldsymbol{x}_i) \tag{3.10}$$

Substituting (3.10) into (3.8) leads to the well-known least-mean-squares (LMS, for short) algorithm [107, 206, 262]:

$$\boldsymbol{w}_i = \boldsymbol{w}_{i-1} + 2\mu\boldsymbol{u}_i^{\mathsf{T}}[\boldsymbol{d}(i) - \boldsymbol{u}_i\boldsymbol{w}_{i-1}], \quad i \geq 0 \tag{3.13}$$

Recall #2 (Example #3.3)

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:

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

Example #4.1

Example 4.1 (Gradient noise for mean-square-error costs). Let us reconsider the scenario studied in Example 3.3, which dealt with mean-square-error costs of the form $J(w) = \mathbb{E}\,(\boldsymbol{d}(i) - \boldsymbol{u}_i w)^2$. From expression (3.19) we know that

$$\boldsymbol{s}_i(w^o) = -2\boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{v}(i) \tag{4.13}$$

$$R_s = 4\sigma_v^2 R_u \equiv R_{s,i}(w^o), \quad \text{for all } i \tag{4.14}$$

Moreover, from expression (3.19) for $\boldsymbol{s}_i(\boldsymbol{w}_{i-1})$, and from the conditions on the random processes $\{\boldsymbol{u}_i, \boldsymbol{v}(i)\}$ in Example 3.1, we have that

$$R_{s,i}(\boldsymbol{w}_{i-1}) = 4\mathbb{E}\left\{(R_u - \boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{u}_i)\widetilde{\boldsymbol{w}}_{i-1}\widetilde{\boldsymbol{w}}_{i-1}^{\mathsf{T}}(R_u - \boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{u}_i)\,|\,\boldsymbol{\mathcal{F}}_{i-1}\right\} + 4\sigma_v^2 R_u \tag{4.15}$$

Example #4.1

Now since $\boldsymbol{u}_i$ and $\widetilde{\boldsymbol{w}}_{i-1}$ are independent of each other:

$$
\begin{aligned}
&\|R_{s,i}(\boldsymbol{w}_{i-1}) - R_{s,i}(w^o)\| \\
&\quad = 4\left\|\mathbb{E}\left\{(R_u - \boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{u}_i)\widetilde{\boldsymbol{w}}_{i-1}\widetilde{\boldsymbol{w}}_{i-1}^{\mathsf{T}}(R_u - \boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{u}_i) \,|\, \boldsymbol{\mathcal{F}}_{i-1}\right\}\right\| \\
&\quad \le 4\mathbb{E}\left\{\left\|(R_u - \boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{u}_i)\widetilde{\boldsymbol{w}}_{i-1}\widetilde{\boldsymbol{w}}_{i-1}^{\mathsf{T}}(R_u - \boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{u}_i)\right\| \,|\, \boldsymbol{\mathcal{F}}_{i-1}\right\} \\
&\quad \le 4\mathbb{E}\left\{\left\|R_u - \boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{u}_i\right\|^2 \|\widetilde{\boldsymbol{w}}_{i-1}\|^2 \,|\, \boldsymbol{\mathcal{F}}_{i-1}\right\} \\
&\quad = 4\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 \left(\mathbb{E}\left\|R_u - \boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{u}_i\right\|^2\right) \\
&\quad = 4c\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 \qquad (4.16)
\end{aligned}
$$

Example #4.1

with the constant $c = \mathbb{E}\,\|R_u - \boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{u}_i\|^2$. It follows that, by taking expectations,

$$\mathbb{E}\,\|R_{s,i}(\boldsymbol{w}_{i-1}) - R_{s,i}(w^o)\| \;\leq\; 4c\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 \tag{4.17}$$

■

Example #A

Logistic Risks: $$J(w) \triangleq \frac{\rho}{2}\|w\|^2 + \mathbb{E}\left\{\ln\left(1+e^{-\boldsymbol{\gamma}(i)\boldsymbol{h}_i^{\mathsf{T}}w}\right)\right\}$$

$$\boldsymbol{s}(\boldsymbol{w}_{i-1}) = \mathbb{E}\left[\frac{\boldsymbol{\gamma}(i)\boldsymbol{h}_i}{1+e^{\boldsymbol{\gamma}(i)\boldsymbol{h}_i^{\mathsf{T}}\boldsymbol{w}_{i-1}}}\right] - \frac{\boldsymbol{\gamma}(i)\boldsymbol{h}_i}{1+e^{\boldsymbol{\gamma}(i)\boldsymbol{h}_i^{\mathsf{T}}\boldsymbol{w}_{i-1}}}$$

$$\longrightarrow \quad \boldsymbol{s}(w^o) = \rho w^o - \frac{\boldsymbol{\gamma}(i)\boldsymbol{h}_i}{1+e^{\boldsymbol{\gamma}(i)\boldsymbol{h}_i^{\mathsf{T}}\boldsymbol{w}_{i-1}}}$$

Example #B: Noise Covariance

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

$$\|R_{s,i}(\boldsymbol{w}_{i-1}) - R_{s,i}(w^o)\| \leq 2c\,\|\widetilde{\boldsymbol{w}}_{i-1}\|$$

$$c = \mathbb{E}\|\boldsymbol{h}_i\|^3 + [\mathrm{Tr}(R_h)]^{3/2}$$

Smoothness Conditions

Now, in order to pursue a closed form expression for the MSD of the algorithm, we need to introduce two smoothness conditions: one condition is on the cost function and the other condition is on the covariance matrix of the gradient noise process.

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 $0 < \gamma \leq 4$.

Smoothness Conditions

Observe from (4.17) that for mean-square-error costs, the Lipschitz condition (4.19) is satisfied with $\gamma = 2$. Likewise, for mean-square-error costs, the first condition (4.18) is automatically satisfied since the Hessian matrices of quadratic costs are constant and independent of w.

Example #C

Quadratic Risks:

$$\left\|\nabla_w^2 J(\boldsymbol{w}_{i-1}) - \nabla_w^2 J(w^o)\right\| \;\leq\; 0$$

Logistic Risks:

$$\left\|\nabla_w^2 J(\boldsymbol{w}_{i-1}) - \nabla_w^2 J(w^o)\right\| \;\leq\; \left(\mathbb{E}\|\boldsymbol{h}_i\|^3\right) \cdot \|\widetilde{\boldsymbol{w}}_{i-1}\|$$

Example #D

Quadratic Risks:

$$\|R_{s,i}(\boldsymbol{w}_{i-1}) - R_{s,i}(w^o)\| \;\leq\; 4c\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2$$

Logistic Risks:

$$\|R_{s,i}(\boldsymbol{w}_{i-1}) - R_{s,i}(w^o)\| \;\leq\; 2c\,\|\widetilde{\boldsymbol{w}}_{i-1}\|$$

From Local to Global Conditions

Although conditions (4.18)–(4.19) are required to hold only locally in the proximity of $w = w^o$, they actually turn out to imply that similar bounds hold more globally. For example, using result (E.30) from the appendix, it can be verified that condition (4.18) translates into a global Lipschitz property relative to the minimizer w^o, i.e., it will also hold that [278]:

$$\|\nabla_w^2 J(w) - \nabla_w^2 J(w^o)\| \leq \kappa_1' \|w - w^o\| \tag{4.20}$$

for all w and for some constant $\kappa_1' \geq 0$.

From Local to Global Conditions

A similar conclusion follows from (4.19). To see that, let us consider any $\boldsymbol{w} \in \boldsymbol{\mathcal{F}}_{i-1}$ such that $\|w^o - \boldsymbol{w}\| > \epsilon$. This condition corresponds to a situation where the perturbation $\Delta\boldsymbol{w}$ in (4.19) lies outside the disc of radius ϵ. Nevertheless, we can still argue that an upper bound similar to (4.19) still holds, albeit with some adjustment [71] — see expression (4.24). To arrive at this expression, we start by using the triangle inequality of norms to note that

$$\|R_{s,i}(\boldsymbol{w}) - R_{s,i}(w^o)\| \;\leq\; \|R_{s,i}(\boldsymbol{w})\| \;+\; \|R_{s,i}(w^o)\| \qquad (4.21)$$

From Local to Global Conditions

Using the property that $\|A\| \leq \mathrm{Tr}(A)$ for any symmetric nonnegative-definite matrix A (since the trace is the sum of the eigenvalues of the matrix and the $2-$induced norm is its largest eigenvalue), we can bound each term on the right-hand side of (4.21) as follows:

From Local to Global Conditions

$$
\begin{aligned}
\|R_{s,i}(\boldsymbol{w})\| &\leq \operatorname{Tr}\left[R_{s,i}(\boldsymbol{w})\right] \\
&= \operatorname{Tr}\left[\mathbb{E}\left\{\boldsymbol{s}_i(\boldsymbol{w})\boldsymbol{s}_i^{\mathsf{T}}(\boldsymbol{w}) \,|\, \boldsymbol{\mathcal{F}}_{i-1}\right\}\right] \\
&= \mathbb{E}\left\{\operatorname{Tr}\left(\boldsymbol{s}_i(\boldsymbol{w})\boldsymbol{s}_i^{\mathsf{T}}(\boldsymbol{w})\right) \,|\, \boldsymbol{\mathcal{F}}_{i-1}\right\} \\
&= \mathbb{E}\left[\|\boldsymbol{s}_i(\boldsymbol{w})\|^2 \,|\, \boldsymbol{\mathcal{F}}_{i-1}\right] \\
&\overset{(4.10)}{\leq} \beta^2\|w^o - \boldsymbol{w}\|^2 + \sigma_s^2 \qquad (4.22)
\end{aligned}
$$

From Local to Global Conditions

By setting $\boldsymbol{w} = w^o$ we also conclude that $\|R_{s,i}(w^o)\| \leq \sigma_s^2$. Substituting into (4.21) we get

$$\begin{aligned}
\|R_{s,i}(\boldsymbol{w}) - R_{s,i}(w^o)\| &\leq \beta^2\|w^o - \boldsymbol{w}\|^2 + 2\sigma_s^2 \\
&\overset{\text{(a)}}{\leq} \beta^2\|w^o - \boldsymbol{w}\|^2 + 2\sigma_s^2\left(\frac{\|w^o - \boldsymbol{w}\|^2}{\epsilon^2}\right) \\
&= \left(\beta^2 + \frac{2\sigma_s^2}{\epsilon^2}\right)\|w^o - \boldsymbol{w}\|^2 \\
&\triangleq \kappa_3\|w^o - \boldsymbol{w}\|^2 \qquad (4.23)
\end{aligned}$$

From Local to Global Conditions

for some nonnegative constant κ_3 and where in step (a) we used the fact that $\|w^o - \boldsymbol{w}\| > \epsilon$. Combining this result with the localized assumption (4.19) we conclude that the conditional noise covariance matrix satisfies more globally a condition of the following form for *any* $\boldsymbol{w} \in \boldsymbol{\mathcal{F}}_{i-1}$:

$$\begin{aligned} \|R_{s,i}(\boldsymbol{w}) - R_{s,i}(w^o)\| &\leq \max\left\{\kappa_2 \|\widetilde{\boldsymbol{w}}\|^{\gamma},\ \kappa_3 \|\widetilde{\boldsymbol{w}}\|^2\right\} \\ &\leq \kappa_2 \|\widetilde{\boldsymbol{w}}\|^{\gamma} + \kappa_3 \|\widetilde{\boldsymbol{w}}\|^2 \end{aligned} \tag{4.24}$$

where $\widetilde{\boldsymbol{w}} = w^o - \boldsymbol{w}$.

Smoothness Conditions

One useful conclusion that follows from the smoothness condition (4.19) and from (4.24) is that, after sufficient iterations, we can express the covariance matrix of the gradient noise process in terms of the same limiting value R_s defined by (4.12) for the absolute noise component. This fact is established next and will be employed later in the proof of Theorem 4.7.

Recall #3: 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.

Second-Order Moment

Lemma 4.1 (Limiting second-order moment of gradient noise: Real case)**.** Under the smoothness condition (4.19), and for sufficiently small step-sizes, it holds for $i \gg 1$ that:[1]

$$\mathbb{E}\, \boldsymbol{s}_i(\boldsymbol{w}_{i-1})\,(\boldsymbol{s}_i(\boldsymbol{w}_{i-1}))^{\mathsf{T}} \;=\; R_s \;+\; O\left(\mu^{\min\{1,\frac{\gamma}{2}\}}\right) \tag{4.25}$$

where $0 < \gamma \leq 4$ is from (4.19) and R_s is defined by (4.12). Consequently, it holds for $i \gg 1$ that the trace of the covariance matrix satisfies:

$$\mathrm{Tr}(R_s) - b_o \;\leq\; \mathbb{E}\,\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2 \;\leq\; \mathrm{Tr}(R_s) + b_o \tag{4.26}$$

for some nonnegative value $b_o = O\left(\mu^{\min\{1,\frac{\gamma}{2}\}}\right)$.

Proof

Proof. By adding and subtracting the same term, we have

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\left(\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\right)^{\mathsf{T}} \mid \boldsymbol{\mathcal{F}}_{i-1}\right] &= \mathbb{E}\left[\boldsymbol{s}_i(w^o)\left(\boldsymbol{s}_i(w^o)\right)^{\mathsf{T}} \mid \boldsymbol{\mathcal{F}}_{i-1}\right] + \\ &\quad \mathbb{E}\left[\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\left(\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\right)^{\mathsf{T}} \mid \boldsymbol{\mathcal{F}}_{i-1}\right] - \\ &\quad \mathbb{E}\left[\boldsymbol{s}_i(w^o)\left(\boldsymbol{s}_i(w^o)\right)^{\mathsf{T}} \mid \boldsymbol{\mathcal{F}}_{i-1}\right] \\ &\overset{(4.11)}{=} \mathbb{E}\left[\boldsymbol{s}_i(w^o)\left(\boldsymbol{s}_i(w^o)\right)^{\mathsf{T}} \mid \boldsymbol{\mathcal{F}}_{i-1}\right] + \\ &\quad R_{s,i}(\boldsymbol{w}_{i-1}) - R_{s,i}(w^o) \qquad (4.27)\end{aligned}$$

Proof

so that by subtracting the covariance matrix R_s defined by (4.12) from both sides, and computing expectations, we get:

$$\begin{aligned}\mathbb{E}\,\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\,(\boldsymbol{s}_i(\boldsymbol{w}_{i-1}))^{\mathsf{T}} - R_s \;=\; &\mathbb{E}\left(\mathbb{E}\left[\boldsymbol{s}_i(w^o)\,(\boldsymbol{s}_i(w^o))^{\mathsf{T}} \mid \boldsymbol{\mathcal{F}}_{i-1}\right] - R_s\right) + \\ &\mathbb{E}\,(R_{s,i}(\boldsymbol{w}_{i-1}) - R_{s,i}(w^o)) \qquad (4.28)\end{aligned}$$

It then follows from the triangle inequality of norms, and from Jensen's inequality (F.29) in the appendix, that:

Proof

$$
\begin{aligned}
&\left\|\mathbb{E}\,\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\,(\boldsymbol{s}_i(\boldsymbol{w}_{i-1}))^{\mathsf{T}} - R_s\right\| \\
&\qquad \leq \quad \left\|\mathbb{E}\left(\mathbb{E}\left[\boldsymbol{s}_i(w^o)\,(\boldsymbol{s}_i(w^o))^{\mathsf{T}} \mid \boldsymbol{\mathcal{F}}_{i-1}\right] - R_s\right)\right\| + \\
&\qquad\qquad \|\mathbb{E}\,(R_{s,i}(\boldsymbol{w}_{i-1}) - R_{s,i}(w^o))\| \\
&\qquad \overset{(\mathrm{F.29})}{\leq} \quad \mathbb{E}\left\|\mathbb{E}\left[\boldsymbol{s}_i(w^o)\,(\boldsymbol{s}_i(w^o))^{\mathsf{T}} \mid \boldsymbol{\mathcal{F}}_{i-1}\right] - R_s\right\| + \\
&\qquad\qquad \mathbb{E}\,\|R_{s,i}(\boldsymbol{w}_{i-1}) - R_{s,i}(w^o)\| \qquad (4.29)
\end{aligned}
$$

Proof

where the notation $\|X\|$ denotes the 2−induced norm of its matrix argument, X. If we now compute the limit superior of both sides, and recall definition (4.12), we get

$$\limsup_{i\to\infty} \left\| \mathbb{E}\, \boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \left(\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\right)^{\mathsf{T}} - R_s \right\| \leq \limsup_{i\to\infty} \mathbb{E}\, \|R_{s,i}(\boldsymbol{w}_{i-1}) - R_{s,i}(w^o)\| \tag{4.30}$$

The limit superior on the right-hand side can be evaluated by calling upon (4.24) to get:

Proof

$$
\begin{aligned}
&\limsup_{i\to\infty} \mathbb{E}\,\|R_{s,i}(\boldsymbol{w}_{i-1}) - R_{s,i}(w^o)\| \\
&\qquad \leq \quad \limsup_{i\to\infty} \mathbb{E}\,\left\{\kappa_2\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^{\gamma} \;+\; \kappa_3\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2\right\} \\
&\qquad \leq \quad \limsup_{i\to\infty} \left\{\kappa_2\,\mathbb{E}\,\left(\|\widetilde{\boldsymbol{w}}_{i-1}\|^4\right)^{\gamma/4} \;+\; \kappa_3\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2\right\} \\
&\qquad \overset{\text{(a)}}{\leq} \quad \limsup_{i\to\infty} \left\{\kappa_2\,\left(\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^4\right)^{\gamma/4} \;+\; \kappa_3\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2\right\} \\
&\qquad \overset{(3.39)}{=} \quad O(\mu^{\gamma'/2})
\end{aligned}
\tag{4.31}
$$

Proof

where in step (a) we applied Jensen's inequality (F.30) to the function $f(x) = x^{\gamma/4}$; this function is concave over $x \geq 0$ for $\gamma \in (0, 4]$. Moreover, in the last step we called upon results (3.39) and (3.67), namely, that the second and fourth-order moments of $\widetilde{\boldsymbol{w}}_{i-1}$ are asymptotically bounded by $O(\mu)$ and $O(\mu^2)$, respectively. Accordingly, the exponent γ' in the last step is given by

$$\gamma' \stackrel{\Delta}{=} \min\{\gamma, 2\} \tag{4.32}$$

since $O(\mu^{\gamma/2})$ dominates $O(\mu)$ for values of $\gamma \in (0, 2]$ and $O(\mu)$ dominates $O(\mu^{\gamma/2})$ for values of $\gamma \in [2, 4]$. Substituting (4.31) into (4.30) we conclude that

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

Proof

If we denote the difference between R_s and the covariance matrix $\mathbb{E}\,\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\,(\boldsymbol{s}_i(\boldsymbol{w}_{i-1}))^{\mathsf{T}}$ by Δ_i, then result (4.33) implies that, for $i \gg 1$, we have $\|\Delta_i\| = O(\mu^{\gamma'/2})$ and we arrive at (4.25). Moreover, since for any square matrix X, it can be verified that $|\mathrm{Tr}(X)| \leq c\,\|X\|$, for some constant c that is independent of γ', we also conclude from (4.33) that

$$\limsup_{i\to\infty} \left|\, \mathbb{E}\,\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2 - \mathrm{Tr}(R_s) \,\right| \;=\; O(\mu^{\gamma'/2}) \;\;\stackrel{\Delta}{=}\;\; b_1 \qquad (4.34)$$

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 (4.34) that $b_1 = O(\mu^{\gamma'/2})$. The above relation then implies that, given $\epsilon > 0$, there exists an I_o large enough such that for all $i > I_o$ it holds that

$$\left| \mathbb{E}\,\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2 - \mathrm{Tr}(R_s) \right| \;\leq\; b_1 + \epsilon \tag{4.35}$$

If we select $\epsilon = O(\mu^{\gamma'/2})$ and introduce the sum $b_o = b_1 + \epsilon$, then we arrive at the desired result (4.26).

□

Stability of First-Order Error Moment

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.

First-Order Error Moment

Using the Lipschitz property (4.20), we can now examine the mean stability of the error vector, $\widetilde{\boldsymbol{w}}_i$, and show that the limit superior of $\|\mathbb{E}\,\widetilde{\boldsymbol{w}}_i\|$ is bounded by $O(\mu)$.

Indeed, using the fact that $(\mathbb{E}\,\boldsymbol{a})^2 \leq \mathbb{E}\,\boldsymbol{a}^2$, for any real-valued random variable $\boldsymbol{a}$, we note that we may conclude from (3.39) that

$$\limsup_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\| \;=\; O(\mu^{1/2}) \tag{4.36}$$

However, a tighter bound is possible with $\mu^{1/2}$ replaced by μ by appealing to (4.20) and bounding the limiting value of $\|\mathbb{E}\,\widetilde{\boldsymbol{w}}_i\|$.

First-Order Error Moment

Let us reconsider recursion (3.42), namely,

$$\widetilde{\boldsymbol{w}}_i \;=\; (I_M - \mu \boldsymbol{H}_{i-1})\widetilde{\boldsymbol{w}}_{i-1} \;+\; \mu\, \boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \tag{4.37}$$

where

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

We introduce the deviation matrix

$$\widetilde{\boldsymbol{H}}_{i-1} \;\triangleq\; H - \boldsymbol{H}_{i-1} \tag{4.39}$$

First-Order Error Moment

where the constant symmetric and positive-definite matrix H is defined as the value of the Hessian matrix at the minimizer w^o:

$$H \triangleq \nabla_w^2 J(w^o) \tag{4.40}$$

Substituting (4.39) into (4.37) gives

$$\widetilde{\boldsymbol{w}}_i = (I_M - \mu H)\widetilde{\boldsymbol{w}}_{i-1} + \mu\, \boldsymbol{s}_i(\boldsymbol{w}_{i-1}) + \mu \boldsymbol{c}_{i-1} \tag{4.41}$$

in terms of the perturbation term

$$\boldsymbol{c}_{i-1} \triangleq \widetilde{\boldsymbol{H}}_{i-1}\widetilde{\boldsymbol{w}}_{i-1} \tag{4.42}$$

Useful Bound

Note that

$$\begin{aligned}
\|\boldsymbol{c}_{i-1}\| &\overset{(4.42)}{\leq} \|\widetilde{\boldsymbol{H}}_{i-1}\|\,\|\widetilde{\boldsymbol{w}}_{i-1}\| \\
&\overset{(4.38)}{\leq} \|\widetilde{\boldsymbol{w}}_{i-1}\| \int_0^1 \left\|\nabla_w^2\, J(w^o - t\widetilde{\boldsymbol{w}}_{i-1}) - \nabla_w^2 J(w^o)\right\| dt \\
&\overset{(4.20)}{\leq} \kappa_1' \,\|\widetilde{\boldsymbol{w}}_{i-1}\| \int_0^1 \|t\widetilde{\boldsymbol{w}}_{i-1}\| dt \\
&= \frac{\kappa_1'}{2}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 \qquad (4.46)
\end{aligned}$$

$$\Longrightarrow \quad \limsup_{i\to\infty}\ \mathbb{E}\,\|\boldsymbol{c}_{i-1}\| \;=\; O(\mu) \qquad (4.47)$$

Mean-Error Stability

Lemma 4.2 (Mean-error stability: Real case). Assume the requirements under Assumptions 4.1 and 4.2 and condition (4.18) on the cost function and the gradient noise process hold. Then, for sufficiently small step-sizes it holds that

$$\limsup_{i\to\infty} \|\mathbb{E}\,\widetilde{\boldsymbol{w}}_i\| = O(\mu) \tag{4.43}$$

Proof

Proof. Conditioning both sides of (4.41) on $\boldsymbol{\mathcal{F}}_{i-1}$, and using the fact that $\mathbb{E}\,[\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\,|\,\boldsymbol{\mathcal{F}}_{i-1}] = 0$, we conclude that

$$\mathbb{E}\,[\widetilde{\boldsymbol{w}}_i\,|\,\boldsymbol{\mathcal{F}}_{i-1}] \;=\; (I_M - \mu H)\,\widetilde{\boldsymbol{w}}_{i-1} \;+\; \mu \boldsymbol{c}_{i-1} \tag{4.44}$$

Taking expectations again we arrive at the mean recursion

$$\mathbb{E}\,\widetilde{\boldsymbol{w}}_i \;=\; (I_M - \mu H)\,\mathbb{E}\,\widetilde{\boldsymbol{w}}_{i-1} \;+\; \mu \mathbb{E}\,\boldsymbol{c}_{i-1} \tag{4.45}$$

The limit superior of the right-most expectation is bounded by $O(\mu^2)$ for the following reason.

Proof

Now the matrix $(I_M - \mu H)$ is symmetric so that its $2-$induced norm agrees with its spectral radius:

$$\|I_M - \mu H\| \;=\; \rho(I_M - \mu H) \tag{4.48}$$

Moreover, for sufficiently small step-sizes $\mu \ll 1$, it holds that this spectral radius is strictly smaller than one and given by

$$\rho(I_M - \mu H) \;=\; 1 - \mu\lambda_{\min}(H) \tag{4.49}$$

Proof

It then follows from (4.45) that

$$\begin{aligned} \|\mathbb{E}\,\widetilde{\boldsymbol{w}}_i\| &\leq \|I_M - \mu H\|\,\|\mathbb{E}\,\widetilde{\boldsymbol{w}}_{i-1}\| + \mu\|\mathbb{E}\,\boldsymbol{c}_{i-1}\| \\ &\leq (1-\mu\lambda_{\min}(H))\|\mathbb{E}\,\widetilde{\boldsymbol{w}}_{i-1}\| + \mu\mathbb{E}\,\|\boldsymbol{c}_{i-1}\| \end{aligned} \tag{4.50}$$

so that

$$\begin{aligned} \limsup_{i\to\infty} \|\mathbb{E}\,\widetilde{\boldsymbol{w}}_i\| &\leq \frac{1}{1-(1-\mu\lambda_{\min}(H))}\left(\limsup_{i\to\infty}\mu\mathbb{E}\,\|\boldsymbol{c}_{i-1}\|\right) \\ &= O(\mu) \end{aligned} \tag{4.51}$$

as claimed.

□

Long-Term Error Dynamics

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.

Long-Term Error Dynamics

Continuing with model (4.41), we can use it to motivate a useful long-term model for the evolution of the error vector $\widetilde{\boldsymbol{w}}_i$ after sufficient iterations, i.e., for $i \gg 1$. For this purpose, we note first that we can deduce from (4.47) that $\|\boldsymbol{c}_{i-1}\| = O(\mu)$ asymptotically with *high probability*. Indeed, let us introduce the nonnegative random variable $\boldsymbol{u} = \|\boldsymbol{c}_{i-1}\|$ and let us recall Markov's inequality [89, 91, 186], which states that for any *nonnegative* random variable $\boldsymbol{u}$ and $\xi > 0$ it holds that

$$\text{Prob}(\boldsymbol{u} \geq \xi) \;\leq\; \mathbb{E}\,\boldsymbol{u}/\xi \tag{4.52}$$

Long-Term Error Dynamics

That is, the probability of the event $\boldsymbol{u} \geq \xi$ is upper bounded by a term that is proportional to $\mathbb{E}\,\boldsymbol{u}$. We employ this result as follows. Let $r_c = n\mu$, for any constant integer $n \geq 1$ that we are free to choose. We then conclude from (4.47) and (4.52) that for $i \gg 1$:

$$\begin{aligned} \text{Prob}\left(\|\boldsymbol{c}_{i-1}\| < r_c\right) &= 1 - \text{Prob}\left(\|\boldsymbol{c}_{i-1}\| \geq r_c\right) \\ &\geq 1 - \left(\mathbb{E}\,\|\boldsymbol{c}_{i-1}\|/r_c\right) \\ &\overset{(4.47)}{\geq} 1 - O\left(1/n\right) \end{aligned} \tag{4.53}$$

Long-Term Error Dynamics

where the term $O(1/n)$ is independent of μ. This result shows that the probability of having $\|\boldsymbol{c}_{i-1}\|$ bounded by r_c can be made arbitrarily close to one by selecting a large enough value for n. Once the value for n has been fixed to meet a desired confidence level, then $r_c = O(\mu)$.

Referring to recursion (4.41), this analysis suggests that we can assess its mean-square performance by examining the following long-term model, which holds with high probability after sufficient iterations:

$$\widetilde{\boldsymbol{w}}_i \;=\; (I_M - \mu H)\widetilde{\boldsymbol{w}}_{i-1} \;+\; \mu\, \boldsymbol{s}_i(\boldsymbol{w}_{i-1}), \qquad i \gg 1 \tag{4.54}$$

Long-Term Error Dynamics

In this model, the perturbation term $\mu \boldsymbol{c}_{i-1}$ that appears in (4.41) is removed. We may also consider an alternative long-term model where $\mu \boldsymbol{c}_{i-1}$ is instead replaced by a constant driving term in the order of $O(\mu^2)$. However, the conclusions that will follow about the performance of the original recursion (4.37) will be the same whether we remove $\mu \boldsymbol{c}_{i-1}$ altogether or replace it by $O(\mu^2)$. We therefore continue our analysis by using model (4.54). Obviously, the iterates $\{\widetilde{\boldsymbol{w}}_i\}$ that are generated by (4.54) are generally different from the iterates that are generated by the original recursion (4.37). To highlight this fact, we rewrite the long-term model (4.54) more explicitly as follows.

Long-Term Error 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.

Long-Term Error Dynamics

Note that the driving process $\boldsymbol{s}_i(\boldsymbol{w}_{i-1})$ in (4.55) continues to be the same gradient noise process from the original recursion (4.37) and is evaluated at $\boldsymbol{w}_{i-1}$. We can view the long-term model (4.55) as a dynamic recursion that is fed by the gradient noise sequence, $\boldsymbol{s}_i(\boldsymbol{w}_{i-1})$. Therefore, assuming both the original system (4.37) and the long-term model (4.55) are launched from the same initial conditions, we observe by iterating (4.55) that $\widetilde{\boldsymbol{w}}'_i$ will still be determined by the past history of the iterates $\{\boldsymbol{w}_j, j \leq i-1\}$ through its dependence on the gradient noise process $\{\boldsymbol{s}_j(\boldsymbol{w}_{j-1}),\ j \leq i\}$. Therefore, it also holds that $\widetilde{\boldsymbol{w}}'_i \in \boldsymbol{\mathcal{F}}_{i-1}$.

Long-Term Error Dynamics

Now working with recursion (4.55) is much more tractable because its dynamics is driven by the constant matrix H as opposed to the random matrix $\boldsymbol{H}_{i-1}$ in the original error recursion (4.37). We shall therefore follow the following route to evaluate the MSD of the stochastic-gradient algorithm (4.5). We shall work with the long-term model (4.55) and evaluate its MSD. Subsequently, we will argue that, under a condition on the fourth-order moment of the gradient noise process, this

Long-Term Error Dynamics

MSD value is within $O(\mu^{3/2})$ from the true MSD expression that would result had we worked directly with the original error recursion (4.37) without the approximation of ignoring $\mu \boldsymbol{c}_{i-1}$ in the long-term. Therefore, the MSD expression that we shall derive based on the long-term model (4.55) will provide an accurate representation for the MSD of the original stochastic-gradient algorithm to first-order in μ.

Long-Term Error Dynamics

We already know from the result of Lemma 3.1 that the original error recursion (4.37) is mean-square stable in the sense that $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$ tends asymptotically to a region that is bounded by $O(\mu)$. We now verify that the long-term model (4.55) is also mean-square stable.

Mean-Square Stability

Lemma 4.4 (Mean-square stability of long-term model). Assume the conditions under Assumptions 4.1 and 4.2 on the cost function and the gradient noise process hold. Then, for sufficiently small step-sizes, the iterate that is generated by the long-term model (4.55) satisfies:

$$\limsup_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i'\|^2 \;=\; O(\mu) \tag{4.56}$$

Proof

Proof. Note first that since $\widetilde{\boldsymbol{w}}'_{i-1} \in \boldsymbol{\mathcal{F}}_{i-1}$ and $\mathbb{E}\,[\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\,|\,\boldsymbol{\mathcal{F}}_{i-1}] = 0$, we conclude from (4.55) that

$$\mathbb{E}\,\left[\,\|\widetilde{\boldsymbol{w}}'_i\|^2\,|\,\boldsymbol{\mathcal{F}}_{i-1}\right] = \left\|(I_M - \mu H)\widetilde{\boldsymbol{w}}'_{i-1}\right\|^2 \;+\; \mu^2\,\mathbb{E}\,\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\,|\,\boldsymbol{\mathcal{F}}_{i-1}\|^2 \quad (4.57)$$

Taking expectations again, we get

$$\mathbb{E}\,\|\widetilde{\boldsymbol{w}}'_i\|^2 \;=\; \mathbb{E}\,\left\|(I_M - \mu H)\widetilde{\boldsymbol{w}}'_{i-1}\right\|^2 \;+\; \mu^2\,\mathbb{E}\,\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2 \quad (4.58)$$

Proof

Using an argument similar to (2.33) and assuming sufficiently small μ such that $\mu < \nu/\delta^2$, we have:

$$\|I_M - \mu H\|^2 \;\leq\; 1 - 2\mu\nu + \mu^2\delta^2 \;\leq\; 1 - \mu\nu \tag{4.59}$$

and, therefore,

$$\begin{aligned}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}'_i\|^2 &\overset{(4.10)}{\leq} \|I_M - \mu H\|^2\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}'_{i-1}\|^2 + \mu^2\,\left[\beta^2\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + \sigma_s^2\right] \\ &\overset{(4.59)}{\leq} (1-\mu\nu)\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}'_{i-1}\|^2 + \mu^2\,\beta^2\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + \mu^2\sigma_s^2 \end{aligned} \tag{4.60}$$

Proof

We already know from (3.39) that sufficiently small step-sizes ensure the convergence of $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2$ towards a region that is bounded by $O(\mu)$. It follows that

$$\begin{aligned}\limsup_{i\to\infty}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i'\|^2 &\leq \frac{1}{1-(1-\mu\nu)}\left(\mu^2\,\beta^2\cdot O(\mu)\;+\;\mu^2\sigma_s^2\right)\\ &= O(\mu)\end{aligned}\tag{4.61}$$

We therefore conclude that (4.56) holds for sufficiently small step-sizes.

□

Mean-Stability

We can also establish the stability of the mean error for the long-term model (4.55) under the Lipschitz property (4.20).

Lemma 4.5 (Mean stability of long-term model). Assume the requirements under Assumptions 4.1 and 4.2 and condition (4.20) on the cost function and the gradient noise process hold. Then, for sufficiently small step-sizes, the iterates of the long-term model (4.55) are asymptotically zero mean:

$$\lim_{i\to\infty} \mathbb{E}\,\widetilde{\boldsymbol{w}}_i' = 0 \tag{4.62}$$

Proof

Proof. The derivation is similar to the argument used to conclude the proof of Lemma 4.2. Specifically, we first use (4.55) to obtain

$$\mathbb{E}\,\widetilde{\boldsymbol{w}}_i' \;=\; (I_M - \mu H)\mathbb{E}\,\widetilde{\boldsymbol{w}}_{i-1}' \tag{4.63}$$

And since $I_M - \mu H$ is a stable matrix for $\mu \ll 1$, we conclude that (4.62) holds.

□

Size of Approximation Error

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.

Size of Approximation Error

We can also examine how close the trajectories of the original error recursion (4.37) and the long-term model (4.55) are to each other. We reproduce both recursions below, with the state variable for the long-term model denoted by $\widetilde{\boldsymbol{w}}_i'$, namely,

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

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

Size of Approximation Error

Observe that both models are driven by the *same* gradient noise process; in this way, the evolution of the long-term model is coupled to the evolution of the original recursion (but not the other way around). The closeness of the trajectories of both recursions is established under the fourth-order condition (3.50) on the gradient noise process, which we repeat below for ease of reference.

Size of Approximation Error

Assumption 4.4 (Conditions on gradient noise). It is assumed that the first and fourth-order conditional moments of the gradient noise process satisfy the following conditions for any iterates $\boldsymbol{w} \in \boldsymbol{\mathcal{F}}_{i-1}$:

$$\mathbb{E}\,[\,\boldsymbol{s}_i(\boldsymbol{w})\,|\,\boldsymbol{\mathcal{F}}_{i-1}\,] = 0 \tag{4.66}$$

$$\mathbb{E}\,\big[\,\|\boldsymbol{s}_i(\boldsymbol{w})\|^4\,|\,\boldsymbol{\mathcal{F}}_{i-1}\,\big] \leq \bar{\beta}^4\,\|\boldsymbol{w}\|^4 + \bar{\sigma}_s^4 \tag{4.67}$$

almost surely, for some nonnegative coefficients $\bar{\sigma}_s^4$ and $\bar{\beta}^4$. These conditions were shown in (3.55)–(3.56) to imply that the gradient noise process also satisfies for any $\boldsymbol{w}_{i-1} \in \boldsymbol{\mathcal{F}}_{i-1}$:

$$\mathbb{E}\,[\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\,|\,\boldsymbol{\mathcal{F}}_{i-1}] = 0 \tag{4.68}$$

$$\mathbb{E}\,\big[\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^4\,|\,\boldsymbol{\mathcal{F}}_{i-1}\big] \leq \beta_4^4\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^4 + \sigma_{s4}^4 \tag{4.69}$$

almost surely, for some nonnegative coefficients β_4^4 and σ_{s4}^4.

Size of Approximation Error

The next statement establishes two useful facts: (a) it shows that the mean-square difference between the trajectories $\{\widetilde{\boldsymbol{w}}_i, \widetilde{\boldsymbol{w}}'_i\}$ is asymptotically bounded by $O(\mu^2)$, and (b) it shows that the MSD values for the original model (4.64) and the long-term model (4.55) are within $O(\mu^{3/2})$ from each other.

Size of Approximation Error

Lemma 4.6 (Performance error is $O(\mu^{3/2})$). Assume the conditions under Assumptions 4.1, 4.3, and 4.4 on the cost function and the gradient noise process are satisfied. It then holds that, for sufficiently small step-sizes:

$$\limsup_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i - \widetilde{\boldsymbol{w}}'_i\|^2 = O(\mu^2) \tag{4.70}$$

$$\limsup_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 = \limsup_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}'_i\|^2 + O(\mu^{3/2}) \tag{4.71}$$

$$\limsup_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2_H = \limsup_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}'_i\|^2_H + O(\mu^{3/2}) \tag{4.72}$$

where the last line involves weighted norms of $\{\widetilde{\boldsymbol{w}}'_i, \widetilde{\boldsymbol{w}}_i\}$ with weighting matrix equal to H.

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:

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

Proof

Proof. Subtracting recursions (4.64) and (4.65) we get

$$\widetilde{\boldsymbol{w}}_i - \widetilde{\boldsymbol{w}}'_i = (I_M - \mu H)(\widetilde{\boldsymbol{w}}_{i-1} - \widetilde{\boldsymbol{w}}'_{i-1}) \;+\; \mu \boldsymbol{c}_{i-1} \tag{4.73}$$

where, from (4.20), $\boldsymbol{c}_{i-1} = \widetilde{\boldsymbol{H}}_{i-1}\widetilde{\boldsymbol{w}}_{i-1}$. Using again an argument similar to (2.33) and assuming sufficiently small μ such that $\mu < \nu/\delta^2$, we have:

$$\begin{aligned} \|I_M - \mu H\|^2 &\leq 1 - 2\mu\nu + \mu^2\delta^2 \\ &\leq 1 - \mu\nu \\ &\leq 1 - \mu\nu + \frac{\mu^2\nu^2}{4} \\ &= \left(1 - \frac{\mu\nu}{2}\right)^2 \end{aligned} \tag{4.74}$$

Proof

We now call upon Jensen's inequality (F.26) from the appendix and apply it to the convex function $f(x) = \|x\|^2$. Indeed, selecting

$$t = \mu\nu/2 \tag{4.75}$$

and for any small μ that ensures $0 < t < 1$, we can write

Proof

$$
\begin{aligned}
&\left\|(I_M-\mu H)(\widetilde{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{w}}'_{i-1})\;+\;\mu\boldsymbol{c}_{i-1}\right\|^2 \\
&\quad= \quad \left\|(1-t)\frac{1}{1-t}(I_M-\mu H)(\widetilde{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{w}}'_{i-1})\;+\;t\frac{1}{t}\left(\mu\boldsymbol{c}_{i-1}\right)\right\|^2 \\
&\quad\le \quad (1-t)\left\|\frac{1}{1-t}(I_M-\mu H)\right\|^2\|\widetilde{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{w}}'_{i-1}\|^2+t\left\|\frac{1}{t}\left(\mu\boldsymbol{c}_{i-1}\right)\right\|^2 \\
&\quad\overset{(4.74)}{\le} \quad \frac{1}{1-t}\left(1-\frac{\mu\nu}{2}\right)^2\|\widetilde{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{w}}'_{i-1}\|^2\;+\;\frac{1}{t}\|\mu\boldsymbol{c}_{i-1}\|^2 \\
&\quad\overset{(4.75)}{=} \quad \left(1-\frac{\mu\nu}{2}\right)\|\widetilde{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{w}}'_{i-1}\|^2\;+\;\frac{2}{\mu\nu}\|\mu\boldsymbol{c}_{i-1}\|^2 \qquad (4.76)
\end{aligned}
$$

Proof

Using (4.46), we conclude from (4.73) and (4.76) that

$$\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i - \widetilde{\boldsymbol{w}}'_i\|^2 \leq \left(1 - \frac{\mu\nu}{2}\right)\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1} - \widetilde{\boldsymbol{w}}'_{i-1}\|^2 + \frac{\mu(\kappa'_1)^2}{2\nu}\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^4 \quad (4.77)$$

Now using (3.67) we conclude that (4.70) holds. With regards to (4.71), we note that

$$\begin{aligned}
\mathbb{E}\,\|\widetilde{\boldsymbol{w}}'_i\|^2 &= \mathbb{E}\,\|\widetilde{\boldsymbol{w}}'_i - \widetilde{\boldsymbol{w}}_i + \widetilde{\boldsymbol{w}}_i\|^2 \\
&= \mathbb{E}\,\|\widetilde{\boldsymbol{w}}'_i - \widetilde{\boldsymbol{w}}_i\|^2 + \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 + 2\left|\mathbb{E}\,(\widetilde{\boldsymbol{w}}'_i - \widetilde{\boldsymbol{w}}_i)^{\mathsf{T}}\widetilde{\boldsymbol{w}}_i\right| \\
&\leq \mathbb{E}\,\|\widetilde{\boldsymbol{w}}'_i - \widetilde{\boldsymbol{w}}_i\|^2 + \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 + 2\sqrt{\mathbb{E}\,\|\widetilde{\boldsymbol{w}}'_i - \widetilde{\boldsymbol{w}}_i\|^2\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2}
\end{aligned} \quad (4.78)$$

Proof

where in the last step we used the property that $|\mathbb{E}\,\boldsymbol{a}^{\mathsf{T}}\boldsymbol{b}|^2 \le \mathbb{E}\,\|\boldsymbol{a}\|^2\,\mathbb{E}\,\|\boldsymbol{b}\|^2$ for any two real random vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. Therefore, from (3.39) and (4.70) we get

$$\limsup_{i\to\infty}\left(\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i'\|^2 - \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2\right) \;\le\; O(\mu^2) \;+\; \sqrt{O(\mu^3)} \;=\; O(\mu^{3/2}) \qquad (4.79)$$

since $\mu^2 < \mu^{3/2}$ for small $\mu \ll 1$, which establishes (4.71). Similarly, we can write for any two real random vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ and constant symmetric positive-definite matrix H:

$$\begin{aligned} |\mathbb{E}\,\boldsymbol{a}^{\mathsf{T}}H\boldsymbol{b}|^2 &\le \mathbb{E}\,\|\boldsymbol{a}\|^2\,\mathbb{E}\,\|H\boldsymbol{b}\|^2 \\ &= \mathbb{E}\,\|\boldsymbol{a}\|^2\,\mathbb{E}\,\|\boldsymbol{b}\|_{H^2}^2 \\ &\overset{(a)}{\le} \rho^2(H)\,\mathbb{E}\,\|\boldsymbol{a}\|^2\,\mathbb{E}\,\|\boldsymbol{b}\|^2 \end{aligned} \qquad (4.80)$$

Proof

where the notation $\|x\|_A^2$ denotes the weighted quantity $x^{\mathsf{T}}Ax$, and in step (a) we used the Rayleigh-Ritz characterization for the eigenvalues of any symmetric matrix A [104, 113, 263]:

$$\lambda_{\min}(A)\|x\|^2 \;\leq\; x^{\mathsf{T}}Ax \;\leq\; \lambda_{\max}(A)\|x\|^2 \tag{4.81}$$

In particular, by setting $\boldsymbol{b} = \boldsymbol{a}$, it also follows from (4.80) that $\mathbb{E}\,\|\boldsymbol{a}\|_H^2 \leq \rho(H)\mathbb{E}\,\|\boldsymbol{a}\|^2$. Therefore, repeating the argument that led to (4.78) using weighted norms we obtain

$$\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i'\|_H^2 \leq \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|_H^2 + \rho(H)\left[\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i' - \widetilde{\boldsymbol{w}}_i\|^2 + 2\sqrt{\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i' - \widetilde{\boldsymbol{w}}_i\|^2\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2}\right] \tag{4.82}$$

and we arrive at (4.72).

□

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.