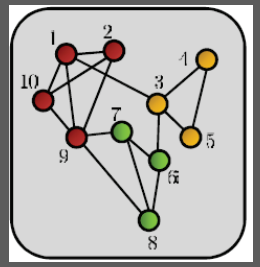


INFERENCE OVER NETWORKS

LECTURE #10: Stochastic Optimization 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 3 (Stochastic Optimization by Single Agents, pp. 338-367):

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

The gradient descent algorithm (2.21) of the previous chapter requires knowledge of the exact gradient vector of the cost function that is being minimized. In the context of adaptation and learning, this information is rarely available beforehand and needs to be approximated. This step is generally achieved by replacing the true gradient by an approximate gradient, thus leading to *stochastic* gradient algorithms. Important challenges and new features arise when the gradient vector is approximated. For instance, the gradient error that is caused by the approximation (and which we shall call *gradient noise*) ends up

Setting

interfering with the operation of the algorithm. It therefore becomes important to assess how much degradation in performance occurs. At the same time, the stochastic approximation step infuses a powerful tracking mechanism into the operation of the gradient descent algorithm; it becomes able to track drifts in the location of the minimizer due to changes in the underlying signal statistics or models. This is because stochastic gradient implementations approximate the gradient vector from streaming data. By doing so, and by relaying on actual data realizations, the drifts in the signal models become reflected in the data and they influence the operation of the algorithm in real-time.

Stochastic Gradient Algorithms

SGAs: among most successful iterative techniques for adaptation and learning:

- "***Learning***": ability to extract information about some unknown parameter from data.
- "***Adaptation***" : ability to track drifts in the parameter.
- Continuous learning and adaptation from "***streaming data***."

Adaptation and Learning: 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 Cost Function

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):

$$w^o = \arg\min_w \; J(w) \tag{3.1}$$

We continue to assume that $J(w)$ is twice-differentiable and satisfies (2.18) for some positive parameters $\nu \leq \delta$, namely,

$$0 < \nu I_M \leq \nabla_w^2 J(w) \leq \delta I_M \tag{3.2}$$

Conditions on Risk Function

Assumption 3.1 (Conditions on cost function). The cost function $J(w)$ is twice-differentiable and satisfies (3.2) for some positive parameters $\nu \leq \delta$. Condition (3.2) 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.

Gradient Descent

We mentioned in the previous chapter that it is common in adaptation and learning applications for the risk function $J(w)$ to be constructed as the expectation of some loss function, $Q(w;\boldsymbol{x})$, say,

$$J(w) = \mathbb{E}\, Q(w;\boldsymbol{x}) \tag{3.3}$$

where the expectation is evaluated over the distribution of $\boldsymbol{x}$. The traditional gradient-descent algorithm for solving (3.1) was described earlier by (2.21), and we repeat it below for ease of reference:

$$w_i = w_{i-1} - \mu \nabla_{w^{\mathsf{T}}} J(w_{i-1}), \quad i \geq 0 \tag{3.4}$$

Gradient Descent

where $i \geq 0$ is an iteration index and $\mu > 0$ is a small step-size parameter. In order to run this recursion, we need to have access to the true gradient vector, $\nabla_{w^\mathsf{T}} J(w_{i-1})$. This information is generally unavailable in most instances involving learning from data. For example, when cost functions are defined as the expectations of certain loss functions as in (3.3), the statistical distribution of the data $\boldsymbol{x}$ may not be known beforehand.

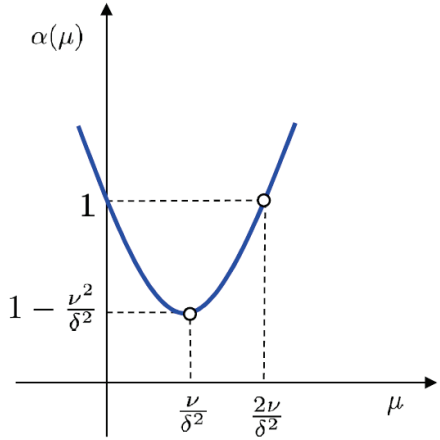
Recall #1: Geometric Convergence

Traditional **gradient-descent** algorithm: $\min_{w} \; J(w)$

$$w_i \;=\; w_{i-1} \;-\; \mu \nabla_{w^{\mathsf{T}}} J(w_{i-1}), \quad i \geq 0$$

$$\widetilde{w}_i = w^o - w_i$$

Lemma 2.1: For small-enough step-sizes, the error converges exponentially as $\|\widetilde{w}_i\|^2 \;\leq\; \alpha \, \|\widetilde{w}_{i-1}\|^2$, where $\alpha \;=\; 1 - 2\mu\nu + \mu^2\delta^2$.



Gradient Descent

In order to run this recursion, we need to have access to the true gradient vector, $\nabla_{w^\mathsf{T}} J(w_{i-1})$. This information is generally unavailable in most instances involving learning from data. For example, when cost functions are defined as the expectations of certain loss functions as in (3.3), the statistical distribution of the data $\boldsymbol{x}$ may not be known beforehand.

Gradient Descent

In that case, the exact form of $J(w)$ will not be known since the expectation of $Q(w;\boldsymbol{x})$ cannot be computed. In such situations, it is necessary to replace the true gradient vector, $\nabla_{w^{\mathsf{T}}} J(w_{i-1})$, by an instantaneous approximation for it, and which we shall denote by $\widehat{\nabla_{w^{\mathsf{T}}} J}(\boldsymbol{w}_{i-1})$. Doing so leads to the following *stochastic-gradient* recursion in lieu of (3.4):

$$\boldsymbol{w}_i = \boldsymbol{w}_{i-1} - \mu \widehat{\nabla_{w^{\mathsf{T}}} J}(\boldsymbol{w}_{i-1}), \quad i \geq 0 \tag{3.5}$$

Boldface Notation

Note that we are using the ***boldface*** notation, $\boldsymbol{w}_i$, for the iterates in (3.5) to highlight the fact that these iterates are randomly perturbed versions of the values $\{w_i\}$ generated by the original recursion (3.4). The random perturbations arise from the use of the approximate gradient vector; different data realizations lead to different realizations for the approximate gradients. The boldface notation is therefore meant to emphasize the random nature of the iterates in (3.5).

Stochastic Gradient Algorithms

Stochastic gradient algorithms are among the most successful iterative techniques for the solution of adaptation and learning problems by stand-alone single agents [190, 207, 243]. We will be using the term "*learning*" to refer broadly to the ability of an agent to extract information about some unknown parameter from streaming data, such as estimating the parameter itself or learning about some of its features. We will be using the term "*adaptation*" to refer broadly to the ability of the learning algorithm to track drifts in the parameter. The

Stochastic Gradient Algorithms

two attributes of learning and adaptation will be embedded simultaneously into the algorithms discussed in this work. We will also be using the term "*streaming data*" regularly because we are interested in algorithms that perform continuous learning and adaptation and that, therefore, are able to improve their performance in response to continuous streams of data arriving at the agent. This is in contrast to off-line algorithms, where the data are first aggregated before being processed for extraction of information.

Stochastic Gradient Algorithms

We illustrate construction (3.5) by considering a scenario from classical adaptive filter theory [107, 206, 262], where the gradient vector is approximated directly from data realizations. The construction will reveal why stochastic-gradient implementations of the form (3.5), using approximate rather than exact gradient information, are naturally endowed with the ability to respond to *streaming* data.

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:

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

$$\rightarrow \quad r_{du} = R_u w^o \qquad \text{(normal equations)}$$

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

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

$$\longrightarrow \quad \nabla_{w^{\mathsf{T}}} J(w) \;=\; 2(R_u w - r_{du})$$

Example #3.1

where the quantities $\{\boldsymbol{d}(i), \boldsymbol{u}_i\}$ represent the random data $\boldsymbol{x}_i$ in the definition of the loss function, $Q(w; \boldsymbol{x}_i)$. Using (3.4), the gradient-descent recursion in this case will take the form:

$$w_i = w_{i-1} - 2\mu \left[R_u w_{i-1} - r_{du}\right], \quad i \geq 0 \tag{3.8}$$

The main difficulty in running this recursion is that it requires knowledge of the moments $\{r_{du}, R_u\}$. This information is rarely available beforehand; the adaptive agent senses instead realizations $\{\boldsymbol{d}(i), \boldsymbol{u}_i\}$ whose statistical distributions have moments $\{r_{du}, R_u\}$. The agent can therefore use these realizations to approximate the moments and the true gradient vector. There are many

Example #3.1

constructions that can be used for this purpose, with different constructions leading to different adaptive algorithms [107, 205, 206, 262]. It is sufficient to illustrate the construction by focusing on one of the most popular adaptive algorithms, which results from using the data $\{\boldsymbol{d}(i), \boldsymbol{u}_i\}$ to compute *instantaneous* approximations for the unavailable moments at every time instant as follows:

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

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

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

Example #3.1

Observe that this construction amounts to replacing the true gradient vector, $\nabla_{w^{\mathsf{T}}} J(w)$, by the gradient vector of the instantaneous loss function itself (which, equivalently, amounts to dropping the expectation operator):

$$\nabla_{w^{\mathsf{T}}} J(w) = \nabla_{\boldsymbol{w}^{\mathsf{T}}} \mathbb{E}\, Q(w; \boldsymbol{x}_i) \tag{3.11}$$

$$\widehat{\nabla_{w^{\mathsf{T}}} J(w)} = \nabla_{\boldsymbol{w}^{\mathsf{T}}} Q(w; \boldsymbol{x}_i) \tag{3.12}$$

Example #3.1

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

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

The LMS algorithm is therefore a stochastic-gradient algorithm. By relying directly on the instantaneous data $\{\boldsymbol{d}(i), \boldsymbol{u}_i\}$, the algorithm is infused with useful tracking abilities. This is because drifts in the model w^o from (3.6) will be reflected in the data $\{\boldsymbol{d}(i), \boldsymbol{u}_i\}$, which are used directly in (3.13).

■

Example #3.2

Example 3.2 (Logistic learner). Let us reconsider the setting of Example 2.2, which dealt with logistic risk functions. Let $\boldsymbol{\gamma}(i)$ be a streaming sequence of binary random variables that assume the values ± 1, and let $\boldsymbol{h}_i$ be a streaming sequence of $M \times 1$ real random (feature) vectors with $R_h = \mathbb{E}\,\boldsymbol{h}_i\boldsymbol{h}_i^{\mathsf{T}} > 0$. We assume the random processes $\{\boldsymbol{\gamma}(i), \boldsymbol{h}_i\}$ are wide-sense stationary. The objective is to seek the vector w that minimizes the following risk function:

$$J(w) \;\triangleq\; \frac{\rho}{2}\|w\|^2 \;+\; \mathbb{E}\left\{\ln\left(1 + e^{-\boldsymbol{\gamma}(i)\boldsymbol{h}_i^{\mathsf{T}} w}\right)\right\} \tag{3.14}$$

Example #3.2

The loss function that is associated with $J(w)$ is

$$Q(w;\boldsymbol{\gamma}(i),\boldsymbol{h}_i) \;\triangleq\; \frac{\rho}{2}\|w\|^2 \;+\; \ln\left(1+e^{-\boldsymbol{\gamma}(i)\boldsymbol{h}_i^{\mathsf{T}}w}\right) \;\equiv\; Q(w;\boldsymbol{x}_i) \qquad (3.15)$$

and the stochastic gradient algorithm for minimizing $J(w)$ then takes the form:

$$\boldsymbol{w}_i = (1-\mu\rho)\boldsymbol{w}_{i-1} \;+\; \mu\boldsymbol{\gamma}(i)\boldsymbol{h}_i\left(\frac{1}{1+e^{\boldsymbol{\gamma}(i)\boldsymbol{h}_i^{\mathsf{T}}\boldsymbol{w}_{i-1}}}\right), \quad i\geq 0 \qquad (3.16)$$

■

Gradient Noise Process

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.

Recall#2: Big and Little-O Notation

$$a = O(\mu) \text{ means } |a| \leq c\mu \text{ for some constant } c.$$

$$a = o(\mu) \text{ means that } |a|/\mu \to 0 \text{ as } \mu \to 0.$$

$$\begin{cases} a &= O(\mu) \Longrightarrow |a| \text{ is in the order of } \mu \\ a &= o(\mu) \Longrightarrow |a| \text{ is some higher power in } \mu \end{cases}$$

Gradient Noise Process

Now, the use of an approximate gradient vector in (3.5) introduces perturbations relative to the operation of the original recursion (3.4). We refer to the perturbation as *gradient noise* and define it as the difference:

$$\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \triangleq \widehat{\nabla_{w^{\mathsf{T}}} J}(\boldsymbol{w}_{i-1}) - \nabla_{w^{\mathsf{T}}} J(\boldsymbol{w}_{i-1}) \tag{3.17}$$

which can also be written as

$$\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \triangleq \nabla_{w^{\mathsf{T}}} Q(\boldsymbol{w}_{i-1}; \boldsymbol{x}_i) - \nabla_{w^{\mathsf{T}}} \mathbb{E}\, Q(\boldsymbol{w}_{i-1}; \boldsymbol{x}_i) \tag{3.18}$$

for cost functions of the form (3.3) and where, as in cases (3.7) and (3.15), the $\{\boldsymbol{x}_i\}$ represent the data.

Gradient Noise Process

The presence of the noise perturbation, $\boldsymbol{s}_i(\boldsymbol{w}_{i-1})$, prevents the stochastic iterate, $\boldsymbol{w}_i$, from converging to the minimizer w^o when constant step-sizes are used. Some deterioration in performance occurs since the iterate $\boldsymbol{w}_i$ will instead fluctuate close to w^o in the steady-state regime. We will assess the size of these fluctuations in the next chapter. Here, we argue that they are bounded and that their mean-square-error is in the order of $O(\mu)$ — see (3.39). The next example from [66] illustrates the nature of the gradient noise process (3.17) in the context of mean-square-error adaptation.

Gradient Noise Process

$$\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \triangleq \widehat{\nabla_{w^{\mathsf{T}}} J(\boldsymbol{w}_{i-1})} - \nabla_{w^{\mathsf{T}}} J(\boldsymbol{w}_{i-1})$$

- Gradient noise prevents $\boldsymbol{w}_i$ from converging to w^o.
- $\boldsymbol{w}_i$ will instead fluctuate around w^o in steady-state.
- We will verify that the Mean-Square-Deviation (MSD) is $O(\mu)$.
- SGA converges towards its steady-state MSD level geometrically.

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:

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

Example #3.3

where we introduced the error vector, $\widetilde{\boldsymbol{w}}_i = w^o - \boldsymbol{w}_i$, and used the relations $\boldsymbol{d}(i) = \boldsymbol{u}_i w^o + \boldsymbol{v}(i)$ and $R_u w^o = r_{du}$. Let the symbol $\boldsymbol{\mathcal{F}}_{i-1}$ represent the collection of all possible random events generated by the past iterates $\{\boldsymbol{w}_j\}$ up to time $j \leq i-1$. Formally, $\boldsymbol{\mathcal{F}}_{i-1}$ is the filtration generated by the random process $\boldsymbol{w}_j$ for $j \leq i-1$ (i.e., $\boldsymbol{\mathcal{F}}_{i-1}$ represents the information that is available about the random process $\boldsymbol{w}_j$ up to time $i-1$):

$$\boldsymbol{\mathcal{F}}_{i-1} \stackrel{\Delta}{=} \text{filtration}\,\{\boldsymbol{w}_{-1},\ \boldsymbol{w}_o,\ \boldsymbol{w}_1, \ldots, \boldsymbol{w}_{i-1}\} \tag{3.20}$$

Example #3.3

It follows from the conditions on the random processes $\{\boldsymbol{u}_i, \boldsymbol{v}(i)\}$ in Example 3.1 that

$$\begin{aligned}\mathbb{E}\,[\,\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\,|\boldsymbol{\mathcal{F}}_{i-1}\,] &= 2(R_u - \mathbb{E}\,\boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{u}_i)\widetilde{\boldsymbol{w}}_{i-1} - 2\mathbb{E}\,\boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{v}(i) \\ &= 2(R_u - R_u)\widetilde{\boldsymbol{w}}_{i-1} - 2\left(\mathbb{E}\,\boldsymbol{u}_i^{\mathsf{T}}\right)\left(\mathbb{E}\,\boldsymbol{v}(i)\right) \\ &= 0 \end{aligned} \tag{3.21}$$

and

$$\mathbb{E}\,\left[\,\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2\,|\,\boldsymbol{\mathcal{F}}_{i-1}\,\right] \le 4\,c\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + 4\sigma_v^2\,\mathrm{Tr}(R_u) \tag{3.22}$$

where the constant c is given by

$$c \stackrel{\Delta}{=} \mathbb{E}\,\|R_u - \boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{u}_i\|^2 \tag{3.23}$$

Example #3.3

If we take expectations of both sides of (3.22), we further conclude that

$$\mathbb{E}\,\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2 \;\leq\; 4c\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + 4\sigma_v^2\,\mathrm{Tr}(R_u) \tag{3.24}$$

so that the variance of the gradient noise, $\mathbb{E}\,\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2$, is bounded by the combination of two factors. The first factor depends on the quality of the iterate, $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2$, while the second factor depends on σ_v^2. Therefore, even if the adaptive agent is able to approach w^o with great fidelity so that $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2$ is small, the size of the gradient noise will still depend on σ_v^2.

■

Example #A

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

$$\longrightarrow \nabla_w J(w) = \rho w^{\mathsf{T}} - \mathbb{E}\left\{\boldsymbol{\gamma}\boldsymbol{h}^{\mathsf{T}} \cdot \frac{e^{-\boldsymbol{\gamma}\boldsymbol{h}^{\mathsf{T}}w}}{1+e^{-\boldsymbol{\gamma}\boldsymbol{h}^{\mathsf{T}}w}}\right\}$$

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

Example #A

$\mathcal{F}_{i-1}$: collection of all possible random events generated by past iterates $\{\boldsymbol{w}_j\}$ up to time $i-1$:

$$\boldsymbol{\mathcal{F}}_{i-1} \triangleq \text{filtration}\,\{\boldsymbol{w}_{-1},\ \boldsymbol{w}_o,\ \boldsymbol{w}_1,\ldots,\boldsymbol{w}_{i-1}\}$$

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

$$\mathbb{E}\,\left[\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2|\boldsymbol{\mathcal{F}}_{i-1}\right] \ \leq \ \text{Tr}(R_h)$$

Conditions on Gradient Noise

In order to examine the convergence and performance properties of the stochastic-gradient recursion (3.5), it is necessary to introduce some assumptions on the stochastic nature of the gradient noise process (3.17), whose definition we rewrite more generally as follows for arbitrary vectors $\boldsymbol{w} \in \boldsymbol{\mathcal{F}}_{i-1}$:

$$\boldsymbol{s}_i(\boldsymbol{w}) \triangleq \widehat{\nabla_{w^{\mathsf{T}}} J(\boldsymbol{w})} - \nabla_{w^{\mathsf{T}}} J(\boldsymbol{w}) \tag{3.25}$$

Conditions on Gradient Noise

The conditions that we state below are similar to conditions used earlier in the optimization literature, e.g., in [190, pp. 95–102] and [33, p. 635]; they are also motivated by the conditions we observed in the mean-square-error case in Example 3.3. Following the developments in [66, 70, 277], we assume the gradient noise process satisfies the following conditions.

Conditions on Gradient Noise

Assumption 3.2 (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 $\boldsymbol{w} \in \boldsymbol{\mathcal{F}}_{i-1}$:

$$\mathbb{E}\left[\,\boldsymbol{s}_i(\boldsymbol{w}) \,|\, \boldsymbol{\mathcal{F}}_{i-1}\,\right] = 0 \tag{3.26}$$

$$\mathbb{E}\left[\,\|\boldsymbol{s}_i(\boldsymbol{w})\|^2 \,|\, \boldsymbol{\mathcal{F}}_{i-1}\,\right] \leq \bar{\beta}^2\,\|\boldsymbol{w}\|^2 + \bar{\sigma}_s^2 \tag{3.27}$$

almost surely, for some nonnegative scalars $\bar{\beta}^2$ and $\bar{\sigma}_s^2$.

Conditions on Gradient Noise

Condition (3.26) ensures that the construction of the approximate gradient vector is unbiased. Moreover, using the second condition (3.27), we deduce for any $\boldsymbol{w}_{i-1} \in \boldsymbol{\mathcal{F}}_{i-1}$ that

$$\begin{aligned}
\mathbb{E}\left[\,\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2 \,|\, \boldsymbol{\mathcal{F}}_{i-1}\right] &\leq \bar{\beta}^2\,\|\boldsymbol{w}_{i-1}\|^2 + \bar{\sigma}_s^2 \\
&\overset{(a)}{=} \bar{\beta}^2\,\|\boldsymbol{w}_{i-1} - w^o + w^o\|^2 + \bar{\sigma}_s^2 \\
&\overset{(b)}{\leq} 2\bar{\beta}^2\,\|\boldsymbol{w}_{i-1} - w^o\|^2 + 2\bar{\beta}^2\|w^o\|^2 + \bar{\sigma}_s^2 \\
&\overset{(c)}{\leq} \beta^2\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + \sigma_s^2 \qquad (3.28)
\end{aligned}$$

Conditions on Gradient Noise

where in step (a) we added and subtracted the global minimizer, w^o, and in step (b) we used the inequality $\|x+y\|^2 \leq 2\|x\|^2+2\|y\|^2$ for any vectors x and y, and in step (c) we introduced the nonnegative scalars:

$$\beta^2 \;\stackrel{\Delta}{=}\; 2\bar{\beta}^2 \tag{3.29}$$

$$\sigma_s^2 \;\stackrel{\Delta}{=}\; 2\bar{\beta}^2\|w^o\|^2 \;+\; \bar{\sigma}_s^2 \tag{3.30}$$

Conditions on Gradient Noise

In other words, we conclude from conditions (3.26)–(3.27) that the following conditions also hold:

$$\mathbb{E}\,[\,\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\,|\,\boldsymbol{\mathcal{F}}_{i-1}\,] \;=\; 0 \tag{3.31}$$

$$\mathbb{E}\,\Big[\,\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2\,|\,\boldsymbol{\mathcal{F}}_{i-1}\,\Big] \;\leq\; \beta^2\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 \;+\; \sigma_s^2 \tag{3.32}$$

Conditions on Gradient Noise

in terms of the error vector, $\widetilde{\boldsymbol{w}}_{i-1} = w^o - \boldsymbol{w}_{i-1}$, and for some nonnegative scalars $\beta^2 \geq 0$ and $\sigma_s^2 \geq 0$. We shall use these conditions more frequently in lieu of (3.26)–(3.27). We could have required these conditions directly in the statement of Assumption 3.2. We instead opted to state conditions (3.26)–(3.27) in that manner, in terms of a generic $\boldsymbol{w} \in \boldsymbol{\mathcal{F}}_{i-1}$ rather than $\widetilde{\boldsymbol{w}}_{i-1}$, so that the upper bound in (3.27) is independent of the unknown w^o.

Conditions on Gradient Noise

By further taking expectations of the relations (3.31)–(3.32), we conclude that the gradient noise process also satisfies:

$$\mathbb{E}\, \boldsymbol{s}_i(\boldsymbol{w}_{i-1}) = 0 \tag{3.33}$$

$$\mathbb{E}\, \|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2 \leq \beta^2\, \mathbb{E}\, \|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + \sigma_s^2 \tag{3.34}$$

Example #B

$$\mathbb{E}\,\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2 \;\;\leq\;\; \beta^2\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + \sigma_s^2$$

Quadratic Risks:

$$\sigma_s^2 \to 4\sigma_v^2\,\mathrm{Tr}(R_u), \quad \beta^2 \to 4c$$

Logistic Risks:

$$\sigma_s^2 \to \mathrm{Tr}(R_h), \qquad \beta^2 = 0$$

Second-Order Stability Analysis

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.

Limit Superior

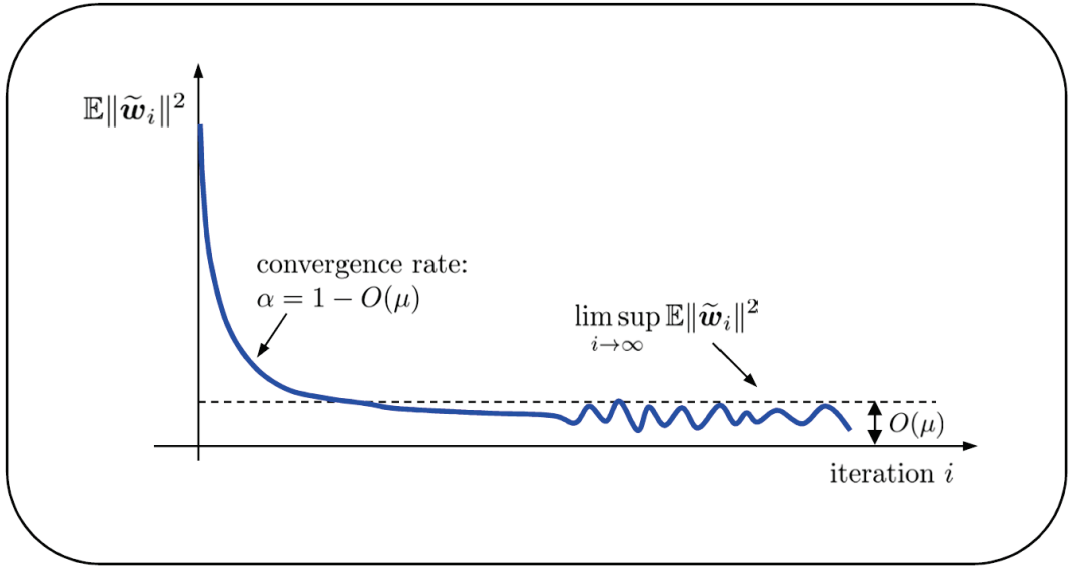


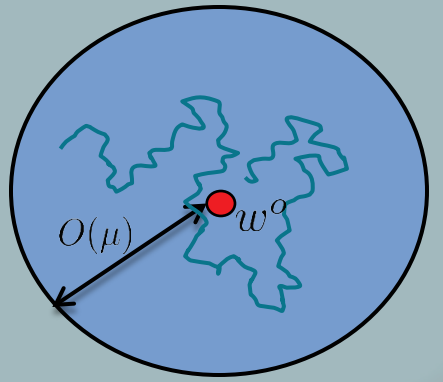
Figure 3.1: Exponential decay of the mean-square error described by (3.37) to a level that is bounded by $O(\mu)$ and at a rate that is in the order of $1 - O(\mu)$.

- **Limit superior:** smallest upper bound for the limiting behavior of the sequence.
- Concept is useful when sequences are not convergent but tend towards bounded regions.

Stability of Error Moments

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

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



Second-Order Moment

We can now examine the convergence of the stochastic-gradient recursion (3.5) in the mean-square-error sense. Result (3.39) below is stated in terms of the *limit superior* of the error variance sequence, $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$. We recall that the limit superior of a sequence essentially corresponds to the smallest upper bound for the limiting behavior of that sequence; this concept is particularly useful when the sequence is not necessarily convergent but tends towards a small bounded region [89, 144, 202]. One such situation is illustrated schematically in Figure 3.1 for the sequence $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$. If the sequence happens to be convergent, then the limit superior will coincide with its regular limiting value.

Mean-Square-Error Stability

Lemma 3.1 (Mean-square-error stability: Real case). Assume the conditions under Assumptions 3.1 and 3.2 on the cost function and the gradient noise process hold, and consider the nonnegative scalars $\{\beta^2, \sigma_s^2\}$ defined by (3.29)–(3.30). For any step-size value, μ, satisfying:

$$\mu < \frac{2\nu}{\delta^2 + \beta^2} \tag{3.36}$$

it holds that $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$ converges exponentially (i.e., at a geometric rate) according to the recursion

Mean-Square-Error Stability

$$\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \;\leq\; \alpha\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 \;+\; \mu^2\sigma_s^2 \tag{3.37}$$

where the scalar α satisfies $0 \leq \alpha < 1$ and is given by

$$\alpha \;=\; 1 - 2\nu\mu + (\delta^2 + \beta^2)\mu^2 \tag{3.38}$$

It follows from (3.37) that, for sufficiently small step-sizes:

$$\limsup_{i\to\infty}\; \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \;=\; O(\mu) \tag{3.39}$$

Proof

Proof. While the result can be established in other ways, we follow the alternative route suggested in the proof of the earlier Lemma 2.1 since this argument is more convenient for extensions to the case of networked agents [66, 69, 70, 277]. We subtract w^o from both sides of (3.5) and use (3.17) to get

$$\widetilde{\boldsymbol{w}}_i \;=\; \widetilde{\boldsymbol{w}}_{i-1} \;+\; \mu\,\nabla_{w^{\mathsf{T}}} J(\boldsymbol{w}_{i-1}) \;+\; \mu\,\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \tag{3.40}$$

Proof

We now appeal to the mean-value relation (D.9) from the appendix to write [190]:

$$\begin{aligned}\nabla_{w^{\mathsf{T}}} J(\boldsymbol{w}_{i-1}) &= -\left(\int_0^1 \nabla_w^2 J(w^o - t\widetilde{\boldsymbol{w}}_{i-1})dt\right)\widetilde{\boldsymbol{w}}_{i-1} \\ &\triangleq -\boldsymbol{H}_{i-1}\widetilde{\boldsymbol{w}}_{i-1} \end{aligned} \tag{3.41}$$

where we are introducing the *symmetric* and *random* time-variant matrix $\boldsymbol{H}_{i-1}$ to represent the integral expression. Substituting into (3.40), we get

$$\widetilde{\boldsymbol{w}}_i = (I_M - \mu\boldsymbol{H}_{i-1})\widetilde{\boldsymbol{w}}_{i-1} + \mu\,\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \tag{3.42}$$

Proof

so that

$$\begin{aligned}\mathbb{E}\left[\|\widetilde{\boldsymbol{w}}_i\|^2 \,|\, \boldsymbol{\mathcal{F}}_{i-1}\right] &\leq \|I_M - \mu \boldsymbol{H}_{i-1}\|^2 \, \|\widetilde{\boldsymbol{w}}_{i-1}\|^2 \; + \\ &\qquad \mu^2 \, \mathbb{E}\left[\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2 \,|\, \boldsymbol{\mathcal{F}}_{i-1}\right] \\ &\overset{(3.32)}{\leq} \|I_M - \mu \boldsymbol{H}_{i-1}\|^2 \, \|\widetilde{\boldsymbol{w}}_{i-1}\|^2 \; + \\ &\qquad \mu^2 \, \left(\beta^2 \|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + \sigma_s^2\right) \end{aligned} \tag{3.43}$$

Proof

Using an argument similar to (2.33) we have

$$\begin{aligned} \|I_M - \mu \boldsymbol{H}_{i-1}\|^2 &= [\rho(I_M - \mu \boldsymbol{H}_{i-1})]^2 \\ &\leq \max\{(1-\mu\delta)^2,\ (1-\mu\nu)^2\} \\ &\leq 1 - 2\mu\nu + \mu^2\delta^2 \end{aligned} \tag{3.44}$$

since $\nu \leq \delta$. Substituting into (3.43) and using the definition (3.38) we obtain

$$\mathbb{E}\left[\|\widetilde{\boldsymbol{w}}_i\|^2 \,|\, \boldsymbol{\mathcal{F}}_{i-1}\right] \leq \alpha \|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + \mu^2\sigma_s^2 \tag{3.45}$$

Proof

Taking expectations of both sides of this inequality we arrive at (3.37). The bound (3.36) on the step-size ensures that $0 \leq \alpha < 1$. Iterating recursion (3.37) gives

$$\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \leq \alpha^{i+1}\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{-1}\|^2 + \frac{\mu^2\sigma_s^2}{1-\alpha} \tag{3.46}$$

which proves that $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$ converges exponentially to a region that is upper bounded by

$$\limsup_{i\to\infty}\ \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \quad \leq \quad \frac{\mu^2\sigma_s^2}{1-\alpha} \quad = \quad \frac{\mu\sigma_s^2}{2\nu-\mu(\delta^2+\beta^2)} \tag{3.47}$$

It is easy to check that the upper bound does not exceed $\mu\sigma_s^2/\nu$ for any step-size $\mu < \nu/(\delta^2+\beta^2)$. We conclude that (3.39) holds for sufficiently small step-sizes.

□

Exponential Decay

Observe that we can rewrite (3.37) in the equivalent form

$$\left(\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \;-\; \frac{\mu^2\sigma_s^2}{1-\alpha}\right) \;\leq\; \alpha\left(\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 \;-\; \frac{\mu^2\sigma_s^2}{1-\alpha}\right) \tag{3.48}$$

where the steady-state bound is subtracted from both sides. It is clear from this representation that α relates to the rate of decay of the mean-square-error towards its steady-state bound — see Figure 3.1.

Limit Superior

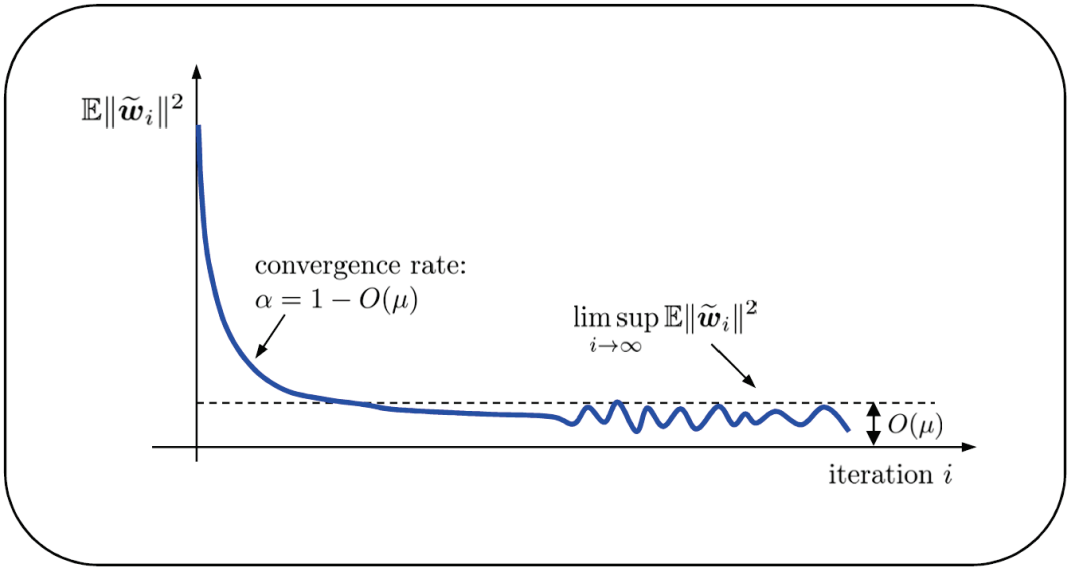


Figure 3.1: Exponential decay of the mean-square error described by (3.37) to a level that is bounded by $O(\mu)$ and at a rate that is in the order of $1 - O(\mu)$.

Fourth-Order Stability Analysis

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.

Fourth-Order Moment

We can also examine the stability of the fourth-order moment of the error vector by showing that the limit superior of $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^4$ tends asymptotically to a region that is bounded by $O(\mu^2)$. The main motivation for establishing this result, in addition to the stability of the second-order moment already established by (3.39), is that these results will be used in the next chapter to derive expressions that quantify the performance of stochastic gradient algorithms to first-order in the step-size parameter.

Fourth-Order Moment

To establish the convergence of the fourth-order moment, $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^4$, to a bounded region, we need to replace Assumption 3.2 by the following condition on the fourth-order moment of the gradient noise process [71, 278].

Conditions on Gradient Noise

Assumption 3.3 (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 $\boldsymbol{w} \in \boldsymbol{\mathcal{F}}_{i-1}$:

$$\mathbb{E}\left[\boldsymbol{s}_i(\boldsymbol{w}) \,|\, \boldsymbol{\mathcal{F}}_{i-1}\right] = 0 \tag{3.49}$$

$$\mathbb{E}\left[\|\boldsymbol{s}_i(\boldsymbol{w})\|^4 \,|\, \boldsymbol{\mathcal{F}}_{i-1}\right] \leq \bar{\beta}^4 \|\boldsymbol{w}\|^4 + \bar{\sigma}_s^4 \tag{3.50}$$

almost surely, for some nonnegative coefficients $\bar{\sigma}_s^4$ and $\bar{\beta}^4$.

Conditions on Gradient Noise

It is straightforward to check that if the above condition on the fourth-order moment holds, then a condition similar to (3.27) on the second-order moment will also hold (while the reverse direction is not necessarily true). Indeed, note that

$$\mathbb{E}\left[\|\boldsymbol{s}_i(\boldsymbol{w})\|^4 \,|\, \boldsymbol{\mathcal{F}}_{i-1}\right] \;\leq\; \left(\bar{\beta}^2\,\|\boldsymbol{w}\|^2 \;+\; \bar{\sigma}_s^2\right)^2 \tag{3.51}$$

so that using the property that $(\mathbb{E}\,\boldsymbol{a})^2 \leq \mathbb{E}\,\boldsymbol{a}^2$ for any real random variable $\boldsymbol{a}$, we conclude that

$$\mathbb{E}\left[\|\boldsymbol{s}_i(\boldsymbol{w})\|^2 \,|\, \boldsymbol{\mathcal{F}}_{i-1}\right] \;\leq\; \bar{\beta}^2\,\|\boldsymbol{w}\|^2 \;+\; \bar{\sigma}_s^2 \tag{3.52}$$

Conditions on Gradient Noise

Therefore, the conditions in Assumption 3.3 continue to ensure the mean-square stability of the stochastic-gradient algorithm, as already established by Lemma 3.1.

Now, for any two vectors a and b, it holds that

$$\begin{aligned} \|a+b\|^4 &= \left\|\frac{1}{2}\cdot 2a + \frac{1}{2}\cdot 2b\right\|^4 \\ &\overset{(a)}{\leq} \frac{1}{2}\|2a\|^4 + \frac{1}{2}\|2b\|^4 \\ &\leq 8\|a\|^4 + 8\|b\|^4 \end{aligned} \tag{3.53}$$

Conditions on Gradient Noise

where in step (a) we called upon Jensen's inequality (F.26) from the appendix and applied it to the convex function $f(x) = \|x\|^4$. Using (3.53), it follows from condition (3.50) that the gradient noise process itself satisfies:

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^4 \,|\, \boldsymbol{\mathcal{F}}_{i-1}\right] &\leq \bar{\beta}^4 \|\boldsymbol{w}_{i-1}\|^4 + \bar{\sigma}_s^4 \\ &= \bar{\beta}^4 \|\boldsymbol{w}_{i-1} - w^o + w^o\|^4 + \bar{\sigma}_s^4 \\ &\leq 8\bar{\beta}^4 \|\widetilde{\boldsymbol{w}}_{i-1}\|^4 + 8\bar{\beta}^4 \|w^o\|^4 + \bar{\sigma}_s^4 \end{aligned} \tag{3.54}$$

Conditions on Gradient Noise

so that the following conditions also hold:

$$\mathbb{E}\left[\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \,|\, \boldsymbol{\mathcal{F}}_{i-1}\right] = 0 \tag{3.55}$$

$$\mathbb{E}\left[\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^4 \,|\, \boldsymbol{\mathcal{F}}_{i-1}\right] \leq \beta_4^4 \|\widetilde{\boldsymbol{w}}_{i-1}\|^4 + \sigma_{s4}^4 \tag{3.56}$$

where we introduced the non-negative parameters:

$$\beta_4^4 \stackrel{\Delta}{=} 8\bar{\beta}^4 \tag{3.57}$$

$$\sigma_{s4}^4 \stackrel{\Delta}{=} 8\bar{\beta}^4\|w^o\|^4 + \bar{\sigma}_s^4 \tag{3.58}$$

Conditions on Gradient Noise

We shall use conditions (3.55)–(3.56) more frequently in lieu of (3.49)–(3.50). By taking expectations of (3.55)–(3.56) we obtain:

$$\mathbb{E}\, \boldsymbol{s}_i(\boldsymbol{w}_{i-1}) = 0 \tag{3.59}$$

$$\mathbb{E}\, \|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^4 \leq \beta_4^4\, \mathbb{E}\, \|\widetilde{\boldsymbol{w}}_{i-1}\|^4 + \sigma_{s4}^4 \tag{3.60}$$

The following example illustrates that the mean-square-error cost considered earlier in Examples 3.1 and 3.2 satisfies the conditions of Assumption 3.3.

Example #3.4

Example 3.4 (Mean-square error costs). Let us consider the same scenario from Example 3.3 where we determined in (3.19) that the gradient noise process is given by

$$\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \;=\; 2(R_u - \boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{u}_i)\widetilde{\boldsymbol{w}}_{i-1} \;-\; 2\boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{v}(i) \tag{3.61}$$

It follows that

$$\begin{aligned}\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^4 &\overset{(3.53)}{\leq} 8\|2(R_u - \boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{u}_i)\widetilde{\boldsymbol{w}}_{i-1}\|^4 \;+\; 8\|2\boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{v}(i)\|^4 \\ &\leq 128\|R_u - \boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{u}_i\|^4\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^4 \;+\; 128\|\boldsymbol{u}_i\|^4\,\|\boldsymbol{v}(i)\|^4\end{aligned} \tag{3.62}$$

Example #3.4

From the conditions on the random processes $\{\boldsymbol{u}_i, \boldsymbol{v}(i)\}$ in Example 3.1, and assuming further that the fourth-order moments of $\{\boldsymbol{v}(i), \boldsymbol{u}_i\}$ are bounded and independent of i, we get

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^4 \,|\, \boldsymbol{\mathcal{F}}_{i-1}\right] &\leq 128\left(\mathbb{E}\,\|R_u - \boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{u}_i\|^4\right)\|\widetilde{\boldsymbol{w}}_{i-1}\|^4 + \\ &\quad 128\left(\mathbb{E}\,\|\boldsymbol{u}_i\|^4\right)\left(\mathbb{E}\,\|\boldsymbol{v}(i)\|^4\right) \\ &\triangleq \beta_4^4\|\widetilde{\boldsymbol{w}}_{i-1}\|^4 + \sigma_{s4}^4 \qquad (3.63)\end{aligned}$$

which is of the same form as (3.60) with

$$\beta_4^4 \triangleq 128\left(\mathbb{E}\,\|R_u - \boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{u}_i\|^4\right) \qquad (3.64)$$

$$\sigma_{s4}^4 \triangleq 128\left(\mathbb{E}\,\|\boldsymbol{u}_i\|^4\right)\left(\mathbb{E}\,\|\boldsymbol{v}(i)\|^4\right) \qquad (3.65)$$

■

Example #C

$$\mathbb{E}\,\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^4 \;\leq\; \beta_4^4\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^4 + \sigma_{s4}^4$$

Quadratic Risks:

$$\beta_4^4 \;\triangleq\; 128\left(\mathbb{E}\,\|R_u - \boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{u}_i\|^4\right)$$

$$\sigma_{s4}^4 \;\triangleq\; 128\left(\mathbb{E}\,\|\boldsymbol{u}_i\|^4\right)\left(\mathbb{E}\,\|\boldsymbol{v}(i)\|^4\right)$$

Logistic Risks:

$$\sigma_{s4}^4 \to 16\,\mathbb{E}\|\boldsymbol{h}_i\|^4, \qquad \beta_4^4 = 0$$

Fourth-Order Stability

Lemma 3.2 (Stability of fourth-order moment: Real case). Assume the conditions under Assumptions 3.1 and 3.3 on the cost function and the gradient noise process hold. Then, for sufficiently small step-sizes, it holds that

$$\limsup_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 = O(\mu) \tag{3.66}$$

$$\limsup_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^4 = O(\mu^2) \tag{3.67}$$

Proof

Proof. We only need to establish (3.67) since (3.66) was established earlier in Lemma 3.1. Following an argument similar to [278], we refer to the error recursion (3.42):

$$\widetilde{\boldsymbol{w}}_i \;=\; (I_M - \mu \boldsymbol{H}_{i-1})\widetilde{\boldsymbol{w}}_{i-1} \;+\; \mu\, \boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \tag{3.68}$$

Using the fact that, for any vectors a and b,

$$\|a+b\|^4 \;=\; \|a\|^4 + \|b\|^4 + 2\|a\|^2\,\|b\|^2 + 4(a^{\mathsf{T}}b)^2 + 4\|b\|^2\,a^{\mathsf{T}}b + 4\|a\|^2\,a^{\mathsf{T}}b \tag{3.69}$$

Proof

we can equate the fourth-order powers of both sides of (3.68) to get

$$\begin{aligned}\|\widetilde{\boldsymbol{w}}_i\|^4 \;=\; & \|(I_M-\mu\boldsymbol{H}_{i-1})\,\widetilde{\boldsymbol{w}}_{i-1}\|^4 \;+\; \mu^4\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^4 \;+\\ & 2\mu^2\|(I_M-\mu\boldsymbol{H}_{i-1})\widetilde{\boldsymbol{w}}_{i-1}\|^2\,\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2 \;+\\ & 4\mu^2\left[\widetilde{\boldsymbol{w}}_{i-1}^{\mathsf{T}}(I_M-\mu\boldsymbol{H}_{i-1})\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\right]^2 \;+\\ & 4\mu^2\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2\,\left[\widetilde{\boldsymbol{w}}_{i-1}^{\mathsf{T}}(I_M-\mu\boldsymbol{H}_{i-1})\mu\boldsymbol{s}(\boldsymbol{w}_{i-1})\right] \;+\\ & 4\|(I_M-\mu\boldsymbol{H}_{i-1})\widetilde{\boldsymbol{w}}_{i-1}\|^2\left[\widetilde{\boldsymbol{w}}_{i-1}^{\mathsf{T}}(I_M-\mu\boldsymbol{H}_{i-1})\mu\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\right]\end{aligned} \tag{3.70}$$

Proof

Applying the Cauchy Schwarz's inequality $(a^{\mathsf{T}}b)^2 \leq \|a\|^2 \|b\|^2$ to the third term on the right-hand side, and using the sub-multiplicative property of norms, we get

$$\begin{aligned}\|\widetilde{\boldsymbol{w}}_i\|^4 \;\leq\; & \|I_M - \mu \boldsymbol{H}_{i-1}\|^4 \|\widetilde{\boldsymbol{w}}_{i-1}\|^4 \;+\; \mu^4 \|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^4 \;+ \\ & 6\mu^2 \|(I_M - \mu \boldsymbol{H}_{i-1})\widetilde{\boldsymbol{w}}_{i-1}\|^2 \|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2 \;+ \\ & 4\mu^2 \|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2 \left[\widetilde{\boldsymbol{w}}_{i-1}^{\mathsf{T}}(I_M - \mu \boldsymbol{H}_{i-1})\mu \boldsymbol{s}_i(\boldsymbol{w}_{i-1})\right] \;+ \\ & 4\|(I_M - \mu \boldsymbol{H}_{i-1})\widetilde{\boldsymbol{w}}_{i-1}\|^2 \left[\widetilde{\boldsymbol{w}}_{i-1}^{\mathsf{T}}(I_M - \mu \boldsymbol{H}_{i-1})\mu \boldsymbol{s}_i(\boldsymbol{w}_{i-1})\right]\end{aligned} \tag{3.71}$$

Proof

Applying further the inequality $2a^{\mathsf{T}}b \leq \|a\|^2 + \|b\|^2$ to the rightmost factor in the third line, and using again the sub-multiplicative property of norms, we get

$$\begin{aligned}\|\widetilde{\boldsymbol{w}}_i\|^4 \;\leq\; & \|I_M - \mu\boldsymbol{H}_{i-1}\|^4 \,\|\widetilde{\boldsymbol{w}}_{i-1}\|^4 \;+\; 3\mu^4\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^4 \;+ \\ & 8\mu^2\|I_M - \mu\boldsymbol{H}_{i-1}\|^2\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2\,\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2 \;+ \\ & 4\|(I_M - \mu\boldsymbol{H}_{i-1})\widetilde{\boldsymbol{w}}_{i-1}\|^2\left[\widetilde{\boldsymbol{w}}_{i-1}^{\mathsf{T}}(I_M - \mu\boldsymbol{H}_{i-1})\mu\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\right]\end{aligned} \tag{3.72}$$

Proof

Conditioning both sides of (3.72) on $\boldsymbol{\mathcal{F}}_{i-1}$ and using (3.55) and (3.56), we obtain

$$\begin{aligned}\mathbb{E}\left[\|\widetilde{\boldsymbol{w}}_i\|^4|\boldsymbol{\mathcal{F}}_{i-1}\right] \;\leq\; & \|I_M-\mu\boldsymbol{H}_{i-1}\|^4\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^4 \;+\\ & 3\mu^4(\beta_4^4\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^4+\sigma_{s4}^4)\;+\\ & 8\mu^2\,\|I_M-\mu\boldsymbol{H}_{i-1}\|^2\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2\,(\beta^2\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2+\sigma_s^2)\end{aligned} \tag{3.73}$$

where the expectation of the last term on the right-hand side of (3.72) is zero since $\mathbb{E}\,[\boldsymbol{s}_i(\boldsymbol{w}_{i-1})|\boldsymbol{\mathcal{F}}_{i-1}]=0$. Using an argument similar to (3.44) we have

$$\begin{aligned}\|I_M-\mu\boldsymbol{H}_{i-1}\|^2 \;&\leq\; 1-2\mu\nu+\mu^2\delta^2\\ &<\; 1+\mu^2\delta^2\end{aligned} \tag{3.74}$$

Proof

and

$$\begin{aligned}\|I_M - \mu \boldsymbol{H}_{i-1}\|^4 &\leq (1 - 2\mu\nu + \mu^2\delta^2)^2 \\ &= 1 - 4\mu\nu + 2\mu^2(2\nu^2 + \delta^2) + \mu^4\delta^4 - 4\mu^3\nu\delta^2 \\ &< 1 - 4\mu\nu + 2\mu^2(2\nu^2 + \delta^2) + \mu^4\delta^4 \qquad (3.75)\end{aligned}$$

Substituting these bounds into (3.73), taking expectations of both sides again to eliminate the conditioning on $\boldsymbol{\mathcal{F}}_{i-1}$, and grouping terms we get

$$\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^4 \leq (1 - a_1)\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^4 + a_2\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + a_3 \qquad (3.76)$$

Proof

where the constants $\{a_1, a_2, a_3\}$ are defined by

$$\begin{aligned} a_1 &= 4\mu\nu - 2\mu^2(2\nu^2 + \delta^2 + 4\beta^2) - \mu^4(\delta^4 + 8\beta^2\delta^2 + 3\beta_4^4) \\ &= O(\mu) \qquad (3.77) \\ a_2 &= 8\mu^2(1+\mu^2\delta^2)\sigma_s^2 \;=\; O(\mu^2) \qquad (3.78) \\ a_3 &= 3\mu^4\sigma_{s4}^4 \;=\; O(\mu^4) \qquad (3.79) \end{aligned}$$

Proof

We can combine (3.76) and the earlier mean-square-error inequality (3.37) into a single linear recursive inequality as follows:

$$\begin{bmatrix} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \\ \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^4 \end{bmatrix} \preceq \begin{bmatrix} \alpha & 0 \\ a_2 & (1-a_1) \end{bmatrix} \begin{bmatrix} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 \\ \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^4 \end{bmatrix} + \begin{bmatrix} \mu^2\sigma_s^2 \\ a_3 \end{bmatrix} \tag{3.80}$$

where the notation $a \preceq b$ means that each entry of vector a is smaller than or equal to the corresponding entry in vector b. We already know from (3.36) that for $\mu < 2\nu/(\delta^2+\beta^2)$, it will hold that $0 \leq \alpha < 1$ so that the mean-square error, $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$, converges asymptotically to a region bounded by $O(\mu)$.

Proof

We can therefore ensure the convergence of recursion (3.80) by showing that a small enough step-size can be chosen to further enforce $|1 - a_1| < 1$ or, equivalently, $0 < a_1 < 2$. Since we know from (3.77) that $a_1 < 4\mu\nu$, then selecting μ according to the following three conditions is sufficient to meet the requirement $0 < a_1 < 2$ (these conditions combined guarantee $\mu\nu < a_1 < 2$):

$$4\mu\nu < 2 \tag{3.81}$$

$$\mu^4(\delta^4 + 8\beta^2\delta^2 + 3\beta_4^4) < \mu^2(2\nu^2 + \delta^2 + 4\beta^2) \tag{3.82}$$

$$\mu^2(2\nu^2 + \delta^2 + 4\beta^2) < \mu\nu \tag{3.83}$$

Proof

or, since $\delta \geq \nu$,

$$\mu < 1/2\delta \tag{3.84}$$

$$\mu < \left(\frac{2\nu^2 + \delta^2 + 4\beta^2}{\delta^4 + 8\beta^2\delta^2 + 3\beta_4^4} \right)^{1/2} \tag{3.85}$$

$$\mu < \frac{\nu}{2\nu^2 + \delta^2 + 4\beta^2} \tag{3.86}$$

Proof

Since the bounds on the right-hand side are positive constants and independent of μ, it is clear that a sufficiently small μ exists that meets all three conditions and leads to $|1-a_1|<1$. For example, the smallest bound among the above three bounds determines an upper limit, μ_o, such that for all $\mu<\mu_o$ we get $0<a_1<2$:

$$\mu_o \;=\; \min\left\{\frac{1}{2\delta},\;\frac{\nu}{2\nu^2+\delta^2+4\beta^2},\;\left(\frac{2\nu^2+\delta^2+4\beta^2}{\delta^4+8\beta^2\delta^2+3\beta_4^4}\right)^{1/2}\right\} \tag{3.87}$$

Proof

It is clear that

$$\frac{\nu}{2\nu^2+\delta^2+4\beta^2} < \frac{\nu}{\delta^2+\beta^2} \tag{3.88}$$

Therefore, any $\mu < \mu_o$ also satisfies $\mu < \nu/(\delta^2+\beta^2)$ and $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$ will be mean-square stable according to (3.36), i.e.,

$$\limsup_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \le b\mu \tag{3.89}$$

for some constant $b > 0$. Computing the limit superior of both sides of (3.76) then gives:

Proof

$$
\begin{aligned}
\limsup_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^4 &\leq \frac{a_2 b\mu + a_3}{a_1} \\
&\overset{\text{(a)}}{\leq} \frac{8\mu^2(1+\mu^2\delta^2)\sigma_s^2 b\mu \;+\; 3\mu^4\sigma_{s4}^4}{\mu\nu} \\
&\leq \left(\frac{8b\sigma_s^2}{\nu}\right)\mu^2 \;+\; \left(\frac{3\sigma_{s4}^4}{\nu}\right)\mu^3 \;+\; \left(\frac{8b\sigma_s^2\delta^2}{\nu}\right)\mu^4 \\
&\overset{\text{(b)}}{\leq} \left(\frac{8b\sigma_s^2}{\nu}\right)\mu^2 + \left(\frac{3\sigma_{s4}^4}{2\nu^2}\right)\mu^2 + \left(\frac{2b\sigma_s^2\delta^2}{\nu^3}\right)\mu^2 \\
&= O(\mu^2) \qquad (3.90)
\end{aligned}
$$

where step (a) is because $a_1 > \mu\nu$ and step (b) is because $\mu < 1/2\nu$.

□

Decaying Step-Sizes

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.

Recall #3: Stochastic Recursion

Lemma F.6 (Stochastic recursion). Let $\boldsymbol{u}(i) \geq 0$ denote a scalar sequence of nonnegative random variables satisfying $\mathbb{E}\,\boldsymbol{u}(0) < \infty$ and the stochastic recursion:

$$\mathbb{E}\,[\boldsymbol{u}(i+1)|\,\boldsymbol{u}(0),\boldsymbol{u}(1),\ldots,\boldsymbol{u}(i)\,] \leq [1-a(i)]\boldsymbol{u}(i) + b(i), \quad i \geq 0 \qquad \text{(F.53)}$$

in terms of the conditional expectation on the left-hand side, and where the scalar and nonnegative deterministic sequences $\{a(i), b(i)\}$ satisfy the five conditions:

$$0 \leq a(i) < 1, \quad b(i) \geq 0, \quad \sum_{i=0}^{\infty} a(i) = \infty, \quad \sum_{i=0}^{\infty} b(i) < \infty, \quad \lim_{i\to\infty} \frac{b(i)}{a(i)} = 0 \qquad \text{(F.54)}$$

Then, it holds that $\lim_{i\to\infty} \boldsymbol{u}(i) = 0$ almost surely, and $\lim_{i\to\infty} \mathbb{E}\,\boldsymbol{u}(i) = 0$.

Recall#4: Deterministic Recursion

Lemma F.5 (Deterministic recursion). Let $u(i) \geq 0$ denote a scalar deterministic (i.e., non-random) sequence that satisfies the inequality recursion:

$$u(i+1) \;\leq\; [1-a(i)]u(i) \;+\; b(i), \quad i \geq 0 \tag{F.49}$$

(a) When the scalar sequences $\{a(i), b(i)\}$ satisfy the four conditions:

$$0 \leq a(i) < 1, \qquad b(i) \geq 0, \qquad \sum_{i=0}^{\infty} a(i) = \infty, \qquad \lim_{i\to\infty} \frac{b(i)}{a(i)} = 0 \tag{F.50}$$

it holds that $\lim_{i\to\infty} u(i) = 0$.

Recall#4: Deterministic Recursion

(b) When the scalar sequences $\{a(i), b(i)\}$ are of the form

$$a(i) = \frac{c}{i+1}, \quad b(i) = \frac{d}{(i+1)^{p+1}}, \quad c > 0, \quad d > 0, \quad p > 0 \tag{F.51}$$

it holds that, for large enough i, the sequence $u(i)$ converges to zero at one of the following rates depending on the value of c:

$$\begin{cases} u(i) \leq \left(\frac{d}{c-p}\right)\frac{1}{i^p} + o\left(1/i^p\right), & c > p \\ u(i) = O\left(\log i/i^p\right), & c = p \\ u(i) = O\left(1/i^c\right), & c < p \end{cases} \tag{F.52}$$

The fastest convergence rate occurs when $c > p$ and is in the order of $1/i^p$.

Decaying Step-Size

If desired, it is also possible to employ iteration-dependent step-size sequences in (3.5) instead of the constant step-size μ, and to require $\mu(i) > 0$ to satisfy either of the following two sets of conditions:

$$\sum_{i=0}^{\infty} \mu(i) = \infty, \qquad \lim_{i \to \infty} \mu(i) = 0 \tag{3.91}$$

or

$$\sum_{i=0}^{\infty} \mu(i) = \infty, \qquad \sum_{i=0}^{\infty} \mu^2(i) < \infty \tag{3.92}$$

Decaying Step-Size

The first set of conditions is the same one we encountered before in (2.38). The second set of conditions is stronger: if a sequence $\mu(i)$ satisfies (3.92) then it also satisfies (3.91). In either case, recursion (3.5) would be replaced by

$$\boldsymbol{w}_i \;=\; \boldsymbol{w}_{i-1} \;-\; \mu(i)\,\widehat{\nabla_{w^{\mathsf{T}}} J}(\boldsymbol{w}_{i-1}), \quad i \geq 0 \tag{3.93}$$

It is well-known [32, 190, 243] that the iterate $\boldsymbol{w}_i$ converges towards w^o in the mean-square sense under (3.91), i.e.,

$$\lim_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 = 0 \qquad \text{(under (3.91))} \tag{3.94}$$

Decaying Step-Size

and it converges to w^o almost surely, i.e., with probability one, under (3.92):

$$\text{Prob}\left(\lim_{i\to\infty} \boldsymbol{w}_i = w^o\right) = 1 \quad \text{(under (3.92))} \tag{3.95}$$

However, as already noted before, conditions (3.91)–(3.92) force the step-size sequence to decay to zero, which is problematic for applications requiring continuous adaptation from streaming data.

Almost-Sure Convergence

Lemma 3.3 (Almost-sure convergence: Real case). Assume the conditions under Assumptions 3.1 and 3.2 on the cost function and the gradient noise process hold. Then, the following convergence properties hold for (3.93):

(a) If the step-size sequence $\mu(i)$ satisfies (3.92), then $\boldsymbol{w}_i$ converges almost surely to w^o, written as $\boldsymbol{w}_i \to w^o$ a.s.

(b) If the step-size sequence $\mu(i)$ satisfies (3.91), then $\boldsymbol{w}_i$ converges in the mean-square-error sense to w^o, i.e., $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \to 0$.

Proof

Proof. We again subtract w^o from both sides of (3.93) to get

$$\widetilde{\boldsymbol{w}}_i \;=\; \widetilde{\boldsymbol{w}}_{i-1} \;+\; \mu(i)\,\nabla_{w^{\mathsf{T}}} J(\boldsymbol{w}_{i-1}) \;+\; \mu(i)\,\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \tag{3.96}$$

We then use the mean-value relation (D.7) from the appendix to note that

$$\nabla_{w^{\mathsf{T}}} J(\boldsymbol{w}_{i-1}) \;=\; \underbrace{\left(\int_0^1 \nabla_w^2\; J(w^o - t\widetilde{\boldsymbol{w}}_{i-1})dt\right)}_{\triangleq\, \boldsymbol{H}_{i-1}} \widetilde{\boldsymbol{w}}_{i-1} \tag{3.97}$$

Proof

where we are introducing the *symmetric* and *random* time-variant matrix $\boldsymbol{H}_{i-1}$, which is defined in terms of the Hessian of the cost function; note that this matrix depends on the random error vector $\widetilde{\boldsymbol{w}}_{i-1}$. Substituting the above relation into (3.96), we get the recursion

$$\widetilde{\boldsymbol{w}}_i \;=\; (I_M - \mu(i)\boldsymbol{H}_{i-1})\widetilde{\boldsymbol{w}}_{i-1} \;+\; \mu(i)\,\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \tag{3.98}$$

Proof

It then follows that

$$\begin{aligned}
\mathbb{E}\left[\|\widetilde{\boldsymbol{w}}_i\|^2 \mid \boldsymbol{\mathcal{F}}_{i-1}\right] &\leq \|I_M - \mu(i)\boldsymbol{H}_{i-1}\|^2 \|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + \\
&\quad \mu^2(i)\,\mathbb{E}\left[\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2 | \boldsymbol{\mathcal{F}}_{i-1}\right] \\
&\overset{(a)}{\leq} (1 - 2\mu(i)\nu + \delta^2\mu^2(i))\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + \\
&\quad \beta^2\mu^2(i)\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + \mu^2(i)\sigma_s^2 \qquad (3.99)
\end{aligned}$$

Proof

where step (a) uses an argument similar to (3.44). Therefore, it holds that:

$$\mathbb{E}\,\left[\,\|\widetilde{\boldsymbol{w}}_i\|^2\,|\,\boldsymbol{\mathcal{F}}_{i-1}\,\right]\;\leq\;\alpha(i)\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2\;+\;\mu^2(i)\sigma_s^2 \tag{3.100}$$

where

$$\alpha(i)\;\stackrel{\Delta}{=}\;1-2\nu\mu(i)+(\delta^2+\beta^2)\mu^2(i) \tag{3.101}$$

Now note that we can split the term $2\nu\mu(i)$ in the above expression for $\alpha(i)$ into the sum of two terms and write

$$\alpha(i)\;=\;1-\nu\mu(i)-\nu\mu(i)+(\delta^2+\beta^2)\mu^2(i) \tag{3.102}$$

Proof

And since $\mu(i) \to 0$, we conclude that for large enough $i > i_o$, the sequence $\mu^2(i)$ will assume smaller values than $\mu(i)$. Therefore, a large enough time index, i_o, exists such that the following two conditions are satisfied:

$$\nu\mu(i) \geq (\delta^2 + \beta^2)\mu^2(i), \qquad 0 \leq \nu\mu(i) < 1, \qquad i > i_o \tag{3.103}$$

Consequently,

$$\alpha(i) \;\leq\; 1 - \nu\mu(i), \qquad i > i_o \tag{3.104}$$

Then, inequalities (3.100) and (3.104) imply that

$$\mathbb{E}\left[\, \|\widetilde{\boldsymbol{w}}_i\|^2 \mid \boldsymbol{\mathcal{F}}_{i-1} \,\right] \;\leq\; (1 - \nu\mu(i))\, \|\widetilde{\boldsymbol{w}}_{i-1}\|^2 \;+\; \mu^2(i)\sigma_s^2, \qquad i > i_o \tag{3.105}$$

Proof

For convenience of notation, let

$$\boldsymbol{u}(i+1) \stackrel{\Delta}{=} \|\widetilde{\boldsymbol{w}}_i\|^2 \tag{3.106}$$

Then, inequality (3.105) implies that

$$\mathbb{E}\,[\boldsymbol{u}(i+1)|\,\boldsymbol{u}(0),\boldsymbol{u}(1),\ldots,\boldsymbol{u}(i)\,] \leq (1-\nu\mu(i))\,\boldsymbol{u}(i) + \mu^2(i)\sigma_s^2, \quad i > i_o \tag{3.107}$$

We now call upon the useful result (F.53) from the appendix and make the identifications

$$a(i) = \nu\mu(i), \quad b(i) = \mu^2(i)\sigma_s^2 \tag{3.108}$$

Proof

These sequences satisfy conditions (F.54) in the appendix in view of assumption (3.92) on the step-size sequence and the second condition in (3.103). We then conclude that $\boldsymbol{u}(i) \to 0$ almost surely and, hence, $\boldsymbol{w}_i \to w^o$ almost surely.

Finally, taking expectations of both sides of (3.107) leads to

$$\mathbb{E}\,\boldsymbol{u}(i+1) \;\leq\; (1-\nu\mu(i))\;\mathbb{E}\,\boldsymbol{u}(i) \;+\; \mu^2(i)\sigma_s^2, \quad i > i_o \tag{3.109}$$

with the expectation operator appearing on both sides of the inequality. Then, we conclude from result (F.49) in the appendix, under conditions (3.91), that $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \to 0$ so that $\boldsymbol{w}_i$ converges to w^o in the mean-square-error sense.

□

Rates of Convergence

We can be more specific and quantify the rate at which the variance $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$ converges towards zero for step-size sequences of the form:

$$\mu(i) = \frac{\tau}{i+1}, \quad \xi > 0 \tag{3.110}$$

which satisfy both conditions (3.91) and (3.92). In contrast to the result of Lemma 2.2 on the convergence rate of gradient descent algorithms, which was seen to be in the order of $O(1/i^{2\nu\tau})$, the next statement indicates that now three rates of convergence are possible depending on how $\nu\tau$ compares to the value one.

Rates of Convergence

Lemma 3.4 (Rates of convergence for a decaying step-size)**.** Assume the conditions under Assumptions 3.1 and 3.2 on the cost function and the gradient noise process hold. Assume further that the step-size sequence is selected according to (3.110). Then, three convergence rates are possible depending on how the factor $\nu\tau$ compares to the value one. Specifically, for large enough i, it holds that:

$$\begin{cases} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \leq \left(\frac{\tau^2\sigma_s^2}{\nu\tau-1}\right)\frac{1}{i} + o\left(\frac{1}{i}\right), & \nu\tau > 1 \\ \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 = O\left(\frac{\log i}{i}\right), & \nu\tau = 1 \\ \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 = O\left(\frac{1}{i^{\nu\tau}}\right), & \nu\tau < 1 \end{cases} \tag{3.111}$$

The fastest convergence rate occurs when $\nu\tau > 1$ (i.e., for large enough τ) and is in the order of $O(1/i)$.

Proof

Proof. We use (3.109) and the assumed form for $\mu(i)$ in (3.110) to write

$$\mathbb{E}\,\boldsymbol{u}(i+1) \;\leq\; \left(1-\frac{\nu\tau}{i+1}\right)\mathbb{E}\,\boldsymbol{u}(i) \;+\; \frac{\tau^2\sigma_s^2}{(i+1)^2}, \quad i > i_o \tag{3.112}$$

This recursion has the same form as recursion (F.49) in the appendix with the identifications

$$a(i) = \frac{\nu\tau}{i+1}, \quad b(i) = \frac{\tau^2\sigma_s^2}{(i+1)^2}, \quad p = 1 \tag{3.113}$$

The above rates of convergence then follow from the statement in part (b) of Lemma F.5 in the appendix.

□

Adaptation and Learning: 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.

Recall#5: Real vs Complex

Consider a function $J(w)$, with $w = x + jy$, $v = \text{col}\{x, y\}$:

$$
\begin{aligned}
H(v) &= \nabla_v^2 J(v) = D^* \left[\nabla_w^2 J(w)\right] D \\
\frac{1}{2}\left[\nabla_v J(v)\right] D^* &= \left[\begin{array}{cc} \nabla_w J(w) & (\nabla_{w^*} J(w))^{\mathsf{T}} \end{array}\right] \\
D &= \left[\begin{array}{cc} I_M & jI_M \\ I_M & -jI_M \end{array}\right] \\
DD^* &= 2I_{2M}
\end{aligned}
$$

$$
\underbrace{\left[\begin{array}{cc} I_M & jI_M \\ I_M & -jI_M \end{array}\right]}_{=D} \underbrace{\left[\begin{array}{c} x \\ y \end{array}\right]}_{=v} = \left[\begin{array}{c} w \\ (w^*)^{\mathsf{T}} \end{array}\right]
$$

Recall#6: Real vs Complex

Consider a $\nu-$strongly convex function $J(w)$:

$$J(\alpha w_1 + (1-\alpha)w_2) \leq \alpha J(w_1) + (1-\alpha)J(w_2) - \frac{\nu}{2}\alpha(1-\alpha)\|w_1 - w_2\|^2$$

$$\Longleftrightarrow \begin{cases} w \text{ real}: & \nabla_w^2 J(w) \geq \nu I_M > 0 \\ w \text{ complex}: & \nabla_w^2 J(w) \geq \dfrac{\nu}{2} I_{2M} > 0 \end{cases}$$

Recall#7: Real vs Complex

Consider a convex function $J(w)$:

$$\begin{cases} w \text{ real}: & \nabla_w^2 J(w) \leq \delta I_M \iff \|\nabla J(w_1) - \nabla J(w_2)\| \leq \delta \|w_1 - w_2\| \\ w \text{ complex}: & \nabla_w^2 J(w) \leq \dfrac{\delta}{2} I_{2M} \iff \|\nabla J(w_1) - \nabla J(w_2)\| \leq \dfrac{\delta}{2} \|w_1 - w_2\| \end{cases}$$

Complex Domain

We now extend the previous results to the case in which the argument $w \in \mathbb{C}^M$ is complex-valued. As was explained earlier in Sec. 2.5, the strongly-convex function, $J(w) \in \mathbb{R}$, is required to satisfy condition (2.62), namely,

$$0 < \frac{\nu}{h} I_{hM} \leq \nabla_w^2 J(w) \leq \frac{\delta}{h} I_{hM} \tag{3.114}$$

in terms of the data-type variable

$$h \triangleq \begin{cases} 1, & \text{when } w \text{ is real} \\ 2, & \text{when } w \text{ is complex} \end{cases} \tag{3.115}$$

Complex Domain

Condition (3.114) captures the requirements that $J(w)$ is twice-differentiable, $\nu-$strongly convex, and has a $\delta-$Lipschitz gradient vector function. The condition is also applicable to both cases of real and complex data. In this section, we are interested in the case $h = 2$ corresponding to complex data. The previous sections studied the case $h = 1$.

Complex Domain

In the complex domain, the stochastic gradient recursions (3.4) and (3.93) are replaced by

$$\boldsymbol{w}_i \;=\; \boldsymbol{w}_{i-1} \;-\; \mu\,\widehat{\nabla_{w^*} J}(\boldsymbol{w}_{i-1}), \quad i \geq 0 \tag{3.116}$$

and

$$\boldsymbol{w}_i \;=\; \boldsymbol{w}_{i-1} \;-\; \mu(i)\,\widehat{\nabla_{w^*} J}(\boldsymbol{w}_{i-1}), \quad i \geq 0 \tag{3.117}$$

respectively, where the second form employs an iteration-dependent step-size sequence. Comparing with (3.4) and (3.93) we see that transposition of the approximate gradient vector is replaced by complex

Gradient Noise Process

conjugation. We again denote the approximation error by the gradient noise model:

$$\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \triangleq \widehat{\nabla_{w^*} J(\boldsymbol{w}_{i-1})} - \nabla_{w^*} J(\boldsymbol{w}_{i-1}) \tag{3.118}$$

This noise process is now complex-valued.

Example #3.5

Example 3.5 (LMS adaptation in the complex domain). We extend the formulation of Examples 3.1 and 3.3 to the complex case. Thus, let $\boldsymbol{d}(i)$ denote a streaming sequence of zero-mean (now complex-valued) random variables with variance $\sigma_d^2 = \mathbb{E}\,|\boldsymbol{d}(i)|^2$. Let $\boldsymbol{u}_i$ denote a streaming sequence of $1 \times M$ independent zero-mean (now complex-valued) random vectors with covariance matrix $R_u = \mathbb{E}\,\boldsymbol{u}_i^*\boldsymbol{u}_i > 0$. 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 $r_{du} = \mathbb{E}\,\boldsymbol{d}(i)\boldsymbol{u}_i^*$. The data $\{\boldsymbol{d}(i), \boldsymbol{u}_i\}$ are assumed to be related via the same linear regression model

$$\boldsymbol{d}(i) \;=\; \boldsymbol{u}_i w^o \;+\; \boldsymbol{v}(i) \tag{3.119}$$

Example #3.5

for some unknown parameter vector w^o, and where $\boldsymbol{v}(i)$ is a zero-mean white-noise process with power $\sigma_v^2 = \mathbb{E}\,|\boldsymbol{v}(i)|^2$ and assumed independent of $\boldsymbol{u}_j$ for all i, j. In a manner similar to Example 2.1, we again pose the problem of estimating w^o by minimizing the mean-square error cost

$$\begin{aligned} J(w) &= \mathbb{E}\,|\boldsymbol{d}(i) - \boldsymbol{u}_i w|^2 \\ &= \sigma_d^2 - r_{du}^* w - w^* r_{du} + w^* R_u w \\ &\equiv \mathbb{E}\,Q(w; \boldsymbol{x}_i) \end{aligned} \tag{3.120}$$

where the quantities $\{\boldsymbol{d}(i), \boldsymbol{u}_i\}$ represent the random data $\boldsymbol{x}_i$ in the definition of $Q(w; \boldsymbol{x}_i)$. Using (2.66), the gradient-descent recursion in this case will take the form:

Example #3.5

$$w_i \;=\; w_{i-1} \;-\; \mu\,[R_u w_{i-1} - r_{du}]\,, \quad i \geq 0 \tag{3.121}$$

Observe that the factor of 2 that used to appear multiplying μ in (3.8) in the real case is not needed here since now

$$\nabla_{w^*} J(w_{i-1}) = R_u w_{i-1} - r_{du} \tag{3.122}$$

Again, the main difficulty in running (3.121) is that it requires knowledge of the moments $\{r_{du}, R_u\}$. Using the *instantaneous* approximations:

$$r_{du} \;\approx\; \boldsymbol{d}(i)\boldsymbol{u}_i^*, \qquad R_u \;\approx\; \boldsymbol{u}_i^*\boldsymbol{u}_i \tag{3.123}$$

Example #3.5

we can replace the true gradient vector by the approximation:

$$\widehat{\nabla_{w^*} J}(w) = [\boldsymbol{u}_i^* \boldsymbol{u}_i w - \boldsymbol{u}_i^* \boldsymbol{d}(i)] \;=\; \nabla_{w^*} Q(w; \boldsymbol{x}_i) \tag{3.124}$$

Substituting (3.124) into (3.121) leads to the complex form of the least-mean-squares (LMS) algorithm [107, 206, 262]:

$$\boldsymbol{w}_i \;=\; \boldsymbol{w}_{i-1} \;+\; \mu \boldsymbol{u}_i^* [\boldsymbol{d}(i) - \boldsymbol{u}_i \boldsymbol{w}_{i-1}], \quad i \geq 0 \tag{3.125}$$

It can be verified from the construction of the approximate gradient vector that the corresponding gradient noise process is now given by

$$\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \;=\; (R_u - \boldsymbol{u}_i^* \boldsymbol{u}_i)\widetilde{\boldsymbol{w}}_{i-1} \;-\; \boldsymbol{u}_i^* \boldsymbol{v}(i) \tag{3.126}$$

Example #3.5

It can be verified from the construction of the approximate gradient vector that the corresponding gradient noise process is now given by

$$\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \;=\; (R_u - \boldsymbol{u}_i^*\boldsymbol{u}_i)\widetilde{\boldsymbol{w}}_{i-1} \;-\; \boldsymbol{u}_i^*\boldsymbol{v}(i) \tag{3.126}$$

in terms of $\widetilde{\boldsymbol{w}}_i = w^o - \boldsymbol{w}_i$. If we again let $\boldsymbol{\mathcal{F}}_{i-1}$ represent filtration generated by the random process $\boldsymbol{w}_j$ for $j \leq i-1$, we readily obtain that

$$\mathbb{E}\,[\,\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\,|\boldsymbol{\mathcal{F}}_{i-1}\,] \;=\; 0 \tag{3.127}$$

$$\mathbb{E}\,\left[\,\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2 \,|\, \boldsymbol{\mathcal{F}}_{i-1}\,\right] \;\leq\; c\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 \;+\; \sigma_v^2\,\mathrm{Tr}(R_u) \tag{3.128}$$

Example #3.5

where the constant c is given by

$$c \triangleq \mathbb{E}\,\|R_u - \boldsymbol{u}_i^* \boldsymbol{u}_i\|^2 \tag{3.129}$$

If we take expectations of both sides of (3.128), we further conclude that

$$\mathbb{E}\,\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2 \;\leq\; c\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + \sigma_v^2\,\mathrm{Tr}(R_u) \tag{3.130}$$

so that the variance of the gradient noise, $\mathbb{E}\,\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2$, is again bounded by the combination of two factors. The first factor depends on the quality of the iterate, $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2$, while the second factor depends on σ_v^2.

■

Conditions on Gradient Noise

Assumption 3.4 (Conditions on gradient noise: Complex case)**.** It is assumed that the first and second-order conditional moments of the gradient noise process satisfy the following conditions for any $\boldsymbol{w} \in \boldsymbol{\mathcal{F}}_{i-1}$:

$$\mathbb{E}\,[\,\boldsymbol{s}_i(\boldsymbol{w})\,|\,\boldsymbol{\mathcal{F}}_{i-1}\,] \;=\; 0 \tag{3.131}$$

$$\mathbb{E}\,\big[\,\|\boldsymbol{s}_i(\boldsymbol{w})\|^2\,|\,\boldsymbol{\mathcal{F}}_{i-1}\,\big] \;\leq\; \big(\bar{\beta}/h\big)^2\,\|\boldsymbol{w}\|^2 \;+\; \bar{\sigma}_s^2 \tag{3.132}$$

almost surely, for some nonnegative scalars $\bar{\beta}^2$ and $\bar{\sigma}_s^2$.

Conditions on Gradient Noise

In a manner similar to the derivation of (3.31)–(3.32) in the real case, we can again verify that the above two conditions lead to the following forms, which we shall use frequently:

$$\mathbb{E}\,[\,\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\,|\,\boldsymbol{\mathcal{F}}_{i-1}\,] \;=\; 0 \tag{3.133}$$

$$\mathbb{E}\left[\,\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2\,|\,\boldsymbol{\mathcal{F}}_{i-1}\,\right] \;\leq\; (\beta/h)^2\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 \;+\; \sigma_s^2 \tag{3.134}$$

and where the scalars $\{\beta^2, \sigma_s^2\}$ are defined by

$$\beta^2 \;\triangleq\; 2\bar{\beta}^2 \tag{3.135}$$

$$\sigma_s^2 \;\triangleq\; 2(\bar{\beta}/h)^2\|w^o\|^2 \;+\; \bar{\sigma}_s^2 \tag{3.136}$$

Complex Domain

By taking expectations of (3.133)–(3.134), we conclude that the gradient noise process also satisfies:

$$\mathbb{E}\, \boldsymbol{s}_i(\boldsymbol{w}_{i-1}) = 0 \tag{3.137}$$

$$\mathbb{E}\, \|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2 \leq (\beta/h)^2\, \mathbb{E}\, \|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + \sigma_s^2 \tag{3.138}$$

It is straightforward to verify from Example 3.5 that the gradient noise process in the mean-square-error case satisfies conditions (3.133)–(3.134). Note in particular from (3.130) that we can make the identifications

$$\sigma_s^2 \to \sigma_v^2\, \mathrm{Tr}(R_u), \quad \beta^2 \to 4c \tag{3.139}$$

Second-Order Stability

Lemma 3.5 (Mean-square-error stability: Complex case). Assume the cost function $J(w)$ satisfies (3.114) and the gradient noise process satisfies the conditions in Assumption 3.4, and consider the nonnegative scalars $\{\beta^2, \sigma_s^2\}$ defined by (3.135)–(3.136). If the step-size parameter is chosen to satisfy

$$\frac{\mu}{h} < \frac{2\nu}{\delta^2 + \beta^2} \tag{3.140}$$

Then, it holds that for any initial condition, $\boldsymbol{w}_{-1}$, the mean-square error, $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$, converges exponentially (i.e., at a geometric rate) according to the recursion:

$$\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \;\leq\; \alpha\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 \;+\; \mu^2\sigma_s^2 \tag{3.141}$$

Second-Order Stability

where

$$\alpha \;=\; 1 - 2\nu\left(\frac{\mu}{h}\right) \;+\; (\delta^2+\beta^2)\left(\frac{\mu}{h}\right)^2 \tag{3.142}$$

It follows from (3.141) that, for sufficiently small step-sizes:

$$\limsup_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \;=\; O(\mu) \tag{3.143}$$

Proof

Proof. We apply the result of Lemma 3.1 to the $v-$domain recursion:

$$\boldsymbol{v}_i \;=\; \boldsymbol{v}_{i-1} \;-\; \mu' \widehat{\nabla_{v^{\mathsf{T}}} J}(\boldsymbol{v}_{i-1}) \tag{3.144}$$

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 $\boldsymbol{w}_i = \boldsymbol{x}_i + j\boldsymbol{y}_i$. We already know from (E.39) in the appendix that $J(v)$ is $\nu-$strongly convex since $J(w)$ is $\nu-$strongly convex. We also know from from (E.22) and (E.56) in the same appendix that the gradient vector function of $J(v)$ is $\delta-$Lipschitz. Therefore, the equivalent function $J(v)$, defined in terms of the real-valued argument v, satisfies the conditions stated in Lemma 3.1.

Proof

All that remains to check is to identify the nature of the gradient noise associated with the modified recursion (3.144) and to verify that this noise satisfies conditions of the same form required by Assumption 3.2. Let us denote the gradient noise of the above recursion in the $v-$domain by

$$\boldsymbol{t}_i(\boldsymbol{v}_{i-1}) \triangleq \widehat{\nabla_{v^{\mathsf{T}}} J}(\boldsymbol{v}_{i-1}) - \nabla_{v^{\mathsf{T}}} J(\boldsymbol{v}_{i-1}) \tag{3.145}$$

We now express $\boldsymbol{t}_i(\cdot)$ in terms of the original gradient noise $\boldsymbol{s}_i(\boldsymbol{w}_{i-1})$ from the $w-$domain given by (3.118). To begin with, recursion (3.144) is equivalent to

$$\boldsymbol{v}_i = \boldsymbol{v}_{i-1} - \frac{\mu}{2}\nabla_{v^{\mathsf{T}}} J(\boldsymbol{v}_{i-1}) - \frac{\mu}{2}\boldsymbol{t}_i(\boldsymbol{v}_{i-1}) \tag{3.146}$$

Proof

Multiplying (3.146) from the left by the matrix D from (B.27) in the appendix and using (C.32), we can transform the above recursion into the following form in terms of the original variables $\boldsymbol{w}_i$:

$$\begin{bmatrix} \boldsymbol{w}_i \\ (\boldsymbol{w}_i^*)^{\mathsf{T}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{w}_{i-1} \\ (\boldsymbol{w}_{i-1}^*)^{\mathsf{T}} \end{bmatrix} - \mu \begin{bmatrix} \nabla_{w^*} J(\boldsymbol{w}_{i-1}) \\ \nabla_{w^{\mathsf{T}}} J(\boldsymbol{w}_{i-1}) \end{bmatrix} - \frac{\mu}{2} D\, \boldsymbol{t}_i(\boldsymbol{v}_{i-1}) \quad (3.147)$$

If we instead start from (3.117), then we would obtain

$$\begin{bmatrix} \boldsymbol{w}_i \\ (\boldsymbol{w}_i^*)^{\mathsf{T}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{w}_{i-1} \\ (\boldsymbol{w}_{i-1}^*)^{\mathsf{T}} \end{bmatrix} - \mu \begin{bmatrix} \nabla_{w^*} J(\boldsymbol{w}_{i-1}) \\ \nabla_{w^{\mathsf{T}}} J(\boldsymbol{w}_{i-1}) \end{bmatrix} - \mu \begin{bmatrix} \boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \\ (\boldsymbol{s}_i^*(\boldsymbol{w}_{i-1}))^{\mathsf{T}} \end{bmatrix} \quad (3.148)$$

Proof

Comparing (3.147) and (3.148) we conclude that the processes $\boldsymbol{t}_i(\cdot)$ and $\boldsymbol{s}_i(\cdot)$ are related as follows:

$$\frac{1}{2}\,D\,\boldsymbol{t}_i(\boldsymbol{v}_{i-1}) \;=\; \begin{bmatrix} \boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \\ (\boldsymbol{s}_i^*(\boldsymbol{w}_{i-1}))^{\mathsf{T}} \end{bmatrix} \tag{3.149}$$

from which, using the fact that $D^*D = 2I_{2M}$ from (B.28) in the appendix, we can solve for $\boldsymbol{t}_i(\boldsymbol{v}_{i-1})$ and find that

$$\boldsymbol{t}_i(\boldsymbol{v}_{i-1}) \;=\; 2\begin{bmatrix} \boldsymbol{s}_{R,i}(\boldsymbol{w}_{i-1}) \\ \boldsymbol{s}_{I,i}(\boldsymbol{w}_{i-1}) \end{bmatrix} \tag{3.150}$$

Proof

in terms of the real and imaginary parts of the gradient noise vector:

$$\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \stackrel{\Delta}{=} \boldsymbol{s}_{R,i}(\boldsymbol{w}_{i-1}) + j\boldsymbol{s}_{I,i}(\boldsymbol{w}_{i-1}) \tag{3.151}$$

Now since $\boldsymbol{s}_i(\boldsymbol{w}_{i-1})$ satisfies conditions (3.133)–(3.134), it follows that

$$\mathbb{E}\,[\,\boldsymbol{t}_i(\boldsymbol{v}_{i-1}) \,|\, \boldsymbol{\mathcal{F}}_{i-1}\,] = 0 \tag{3.152}$$

and

Proof

$$\begin{aligned}
\mathbb{E}\left[\,\|\boldsymbol{t}_i(\boldsymbol{v}_{i-1})\|^2 \mid \boldsymbol{\mathcal{F}}_{i-1}\,\right] &\overset{(3.150)}{=} 4\,\mathbb{E}\left[\,\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2 \mid \boldsymbol{\mathcal{F}}_{i-1}\,\right] \\
&\overset{(3.138)}{\leq} 4\left(\frac{\beta}{h}\right)^2 \|\widetilde{\boldsymbol{w}}_{i-1}\|^2 \;+\; 4\,\sigma_s^2 \\
&= \beta^2\,\|\widetilde{\boldsymbol{w}}_{i-1})\|^2 \;+\; 4\,\sigma_s^2 \qquad (3.153)
\end{aligned}$$

Proof

where we used $h = 2$ for complex data. Therefore, the gradient noise process $\boldsymbol{t}_i(\boldsymbol{v}_{i-1})$ satisfies conditions similar to (3.34) and the result of Lemma 3.1 is then immediately applicable to the $v-$domain recursion (3.144). Specifically, we know from the statement of that lemma that the stochastic gradient recursion (3.146) converges in the mean-square sense when $\mu' < 2\nu/(\delta^2 + \beta^2)$, which is equivalent to (3.140). Moreover, from (3.37) we get

$$\begin{aligned} \mathbb{E}\,\|\widetilde{\boldsymbol{v}}_i\|^2 &\leq \alpha\,\mathbb{E}\,\|\widetilde{\boldsymbol{v}}_{i-1}\|^2 + (\mu')^2\,(4\sigma_s^2) \\ &= \alpha\,\mathbb{E}\,\|\widetilde{\boldsymbol{v}}_{i-1}\|^2 + \mu^2\sigma_s^2 \end{aligned} \tag{3.154}$$

Proof

where

$$\begin{aligned}\alpha &= 1-2\nu\mu'+(\mu')^2(\delta^2+\beta^2)\\ &= 1-\nu\mu+\frac{\mu^2}{4}(\delta^2+\beta^2) \qquad (3.155)\end{aligned}$$

and, therefore, from (3.154):

$$\limsup_{i\to\infty}\ \mathbb{E}\,\|\widetilde{\boldsymbol{v}}_i\|^2 \ \le\ \frac{\mu\sigma_s^2}{\nu-\frac{\mu}{4}(\delta^2+\beta^2)} \qquad (3.156)$$

It is easy to check that the upper bound does not exceed $2\mu\sigma_s^2/\nu$ for any μ satisfying $\mu < 2\nu(\delta^2+\beta^2)$. We conclude that (3.143) holds for sufficiently small step-sizes.

□

Fourth-Order Stability

Assumption 3.5 (Conditions on gradient noise: Complex case)**.** It is assumed that the first and fourth-order conditional moments of the gradient noise process satisfy the following conditions for any iterates $\boldsymbol{w} \in \boldsymbol{\mathcal{F}}_{i-1}$:

$$\mathbb{E}\,[\,\boldsymbol{s}_i(\boldsymbol{w})\,|\,\boldsymbol{\mathcal{F}}_{i-1}\,] \;=\; 0 \tag{3.157}$$

$$\mathbb{E}\,[\,\|\boldsymbol{s}_i(\boldsymbol{w})\|^4\,|\,\boldsymbol{\mathcal{F}}_{i-1}\,] \;\leq\; \big(\bar{\beta}/h\big)^4\,\|\boldsymbol{w}\|^4 \;+\; \bar{\sigma}_s^4 \tag{3.158}$$

almost surely, for some nonnegative coefficients $\bar{\sigma}_s^4$ and $\bar{\beta}^4$.

Conditions on Gradient Noise

In a manner similar to the derivation of (3.55)–(3.56) in the real case, we can again verify that the above two conditions lead to the following forms:

$$\mathbb{E}\left[\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \,|\, \boldsymbol{\mathcal{F}}_{i-1}\right] = 0 \tag{3.159}$$

$$\mathbb{E}\left[\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^4 \,|\, \boldsymbol{\mathcal{F}}_{i-1}\right] \leq \beta_4^4 \,\|\widetilde{\boldsymbol{w}}_{i-1}\|^4 + \sigma_{s4}^4 \tag{3.160}$$

in terms of the nonnegative parameters:

Conditions on Gradient Noise

$$\beta_4^4 \;\triangleq\; 8\bar{\beta}^4 \tag{3.161}$$

$$\sigma_{s4}^4 \;\triangleq\; 8(\bar{\beta}/h)^4\|w^o\|^4 \;+\; \bar{\sigma}_s^4 \tag{3.162}$$

By taking expectations of (3.159)–(3.160) we obtain:

$$\mathbb{E}\,\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \;=\; 0 \tag{3.163}$$

$$\mathbb{E}\,\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^4 \;\leq\; (\beta_4/h)^4\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^4 + \sigma_{s4}^4 \tag{3.164}$$

Fourth-Order Stability

Lemma 3.6 (Stability of fourth-order moment: Complex case)**.** Assume the conditions under Assumptions 3.1 and 3.5 on the cost function and the gradient noise process hold. Then, for sufficiently small step-sizes, it again holds that

$$\limsup_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 = O(\mu) \tag{3.165}$$

$$\limsup_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^4 = O(\mu^2) \tag{3.166}$$

Proof

Proof. We apply Lemma 3.2 to the $v-$domain recursion

$$\boldsymbol{v}_i \;=\; \boldsymbol{v}_{i-1} \;-\; \mu' \,\widehat{\nabla_{v^{\mathsf{T}}} J}(\boldsymbol{v}_{i-1}) \tag{3.167}$$

where $\mu' = \mu/2$ after noting that the gradient noise process $\boldsymbol{t}_i(\boldsymbol{v}_{i-1})$ satisfies a fourth-order condition of the same form as (3.60) since

Proof

$$\begin{aligned}
\mathbb{E}\,\big[\,\|\boldsymbol{t}_i(\boldsymbol{v}_{i-1})\|^4\,\big|\,\boldsymbol{\mathcal{F}}_{i-1}\,\big] &= \mathbb{E}\,\Big[\,\big(\|\boldsymbol{t}_i(\boldsymbol{v}_{i-1})\|^2\big)^2\,\big|\,\boldsymbol{\mathcal{F}}_{i-1}\,\Big] \\
&\overset{(3.150)}{=} \mathbb{E}\,\Big[\,\big(4\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2\big)^2\,\big|\,\boldsymbol{\mathcal{F}}_{i-1}\,\Big] \\
&= 16\mathbb{E}\,\big[\,\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^4\,\big|\,\boldsymbol{\mathcal{F}}_{i-1}\,\big] \\
&\overset{(3.164)}{\leq} \beta_4^4\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^4 \;+\; 16\,\sigma_{s4}^4 \qquad (3.168)
\end{aligned}$$

using $h = 2$.

□

Decaying Step-Size

Lemma 3.7 (Almost-sure convergence: Complex case). Assume the cost function $J(w)$ satisfies (3.114) and the gradient noise process satisfies the conditions in Assumption 3.4. Then, the following convergence properties hold for (3.117):

(a) If the step-size sequence $\mu(i)$ satisfies (3.92), then $\boldsymbol{w}_i$ converges almost surely to w^o, written as $\boldsymbol{w}_i \to w^o$ a.s.

(b) If the step-size sequence $\mu(i)$ satisfies (3.91), then $\boldsymbol{w}_i$ converges in the mean-square-error sense to w^o, i.e., $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \to 0$.

Proof

Proof. We apply the result of Lemma 3.3 to the $v-$domain recursion:

$$\boldsymbol{v}_i \;=\; \boldsymbol{v}_{i-1} \;-\; \mu'(i)\,\widehat{\nabla_{v^{\mathsf{T}}}\,J}(\boldsymbol{v}_{i-1}) \tag{3.169}$$

where $\mu'(i) = \mu(i)/2$.

□

Rates of Convergence

Lemma 3.8 (Rates of convergence for a decaying step-size). Assume the cost function $J(w)$ satisfies (3.114) and the gradient noise process satisfies the conditions in Assumption 3.4. Assume further that the step-size sequence is selected according to (3.110). Then, three convergence rates are possible depending on how the factor $\nu\tau/h$ compares to the value one. Specifically, for large enough i, it holds that:

$$\begin{cases} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \leq \left(\frac{\tau^2\sigma_s^2}{\nu\tau/h-1}\right)\frac{1}{i} + o\left(\frac{1}{i}\right), & \nu\tau/h > 1 \\ \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 = O\left(\frac{\log i}{i}\right), & \nu\tau/h = 1 \\ \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 = O\left(\frac{1}{i^{\nu\tau/h}}\right), & \nu\tau/h < 1 \end{cases} \tag{3.170}$$

where $h = 2$ for complex data and $h = 1$ for real data. The fastest convergence rate occurs when $\nu\tau/h > 1$ (i.e., for large enough τ) and is in the order of $O(1/i)$.

Proof

Proof. Apply the result of Lemma 3.4 to (3.169) noting that

$$\mu'(i) = \frac{\tau/2}{i+1} \tag{3.171}$$

so that τ is replaced by $\tau/2$ and, from (3.153), σ_s^2 is replaced by $4\sigma_s^2$.

□

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.