19 CONVERGENCE ANALYSIS I: STOCHASTIC GRADIENT ALGORITHMS

We are ready to examine the convergence behavior of the stochastic gradient algorithm for smooth risks under various operation modes. We will consider updates with constant and vanishing step-sizes. We will also consider data sampling with and without replacement, as well as under importance sampling. We will further consider instantaneous and mini-batch gradient approximations. In all cases, the main conclusion will be that the mean-square error $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2$ is guaranteed to approach a small $O(\mu)-$neighborhood, while exact convergence of $\boldsymbol{w}_n$ to $w^\star$ can be guaranteed for some vanishing step-size sequences. These are reassuring results in that the deterioration due to the stochastic gradient approximations remains small, which explains in large part the explosive success of stochastic approximation methods in inference and learning.

19.1 PROBLEM SETTING

We start our exposition by recalling the problem formulation, and the conditions imposed on the risk and loss functions. We also recall the constructions for the gradient approximations, and the first and second-order moment results derived in the last chapter for the gradient noise process.

19.1.1 Risk Minimization Problems

To begin with, in this chapter, we are interested in examining the convergence behavior of the stochastic gradient implementation:

$$\boxed{\boldsymbol{w}_n = \boldsymbol{w}_{n-1} - \mu\,\widehat{\nabla_{w^{\mathsf{T}}}P}(\boldsymbol{w}_{n-1}),\ n \geq 0} \tag{19.1}$$

with constant μ, or even decaying step-sizes $\mu(n)$, for the solution of convex optimization problems of the form:

$$w^\star = \operatorname*{argmin}_{w\in\mathbb{R}^M}\ P(w) \tag{19.2}$$

where $P(w)$ is a first-order differentiable empirical or stochastic risk, i.e., for the solution of:

$$w^\star \triangleq \operatorname*{argmin}_{w\in\mathbb{R}^M} \left\{ P(w) \triangleq \frac{1}{N}\sum_{m=0}^{N-1} Q(w;\gamma(m),h_m) \right\} \tag{19.3a}$$

$$w^o \triangleq \operatorname*{argmin}_{w\in\mathbb{R}^M} \left\{ P(w) \triangleq \mathbb{E}\, Q(w;\boldsymbol{\gamma},\boldsymbol{h}) \right\} \tag{19.3b}$$

Observe that we use $w^\star$ to refer to the minimizer in the empirical case, and w^o for the minimizer in the stochastic case. Often, when there is no room for confusion, we will use $w^\star$ to refer generically to the minimizer of $P(w)$ independent of whether it represents an empirical or stochastic risk. In the above expressions, $Q(w,\cdot)$ denotes the loss function, $\{\gamma(m),h_m\}$ refer to a collection of $N-$data points with $\gamma(m)\in\mathbb{R}$ and $h_m\in\mathbb{R}^M$, and the expectation in the second line is over the joint distribution of $\{\boldsymbol{\gamma},\boldsymbol{h}\}$.

19.1.2 Gradient Vector Approximations

The gradient search direction will be approximated by using either instantaneous or mini-batch calculations, namely,

(*approximations under sampling with and without replacement*)

$$(\textbf{instantaneous}) : \widehat{\nabla_{w^\mathsf{T}} P}(w) = \nabla_{w^\mathsf{T}} Q(w;\boldsymbol{\gamma},\boldsymbol{h}) \tag{19.4a}$$

$$(\textbf{mini-batch}) : \widehat{\nabla_{w^\mathsf{T}} P}(w) = \frac{1}{B}\sum_{b=0}^{B-1} \nabla_{w^\mathsf{T}} Q(w;\boldsymbol{\gamma}(b),\boldsymbol{h}_b) \tag{19.4b}$$

where the boldface notation $(\boldsymbol{\gamma},\boldsymbol{h})$ or $(\boldsymbol{\gamma}(b),\boldsymbol{h}_b)$ refers to data samples selected at random (with or without replacement) from the given dataset $\{\gamma(m),h_m\}$ in empirical risk minimization, or assumed to stream in independently over time in stochastic risk minimization. The difference between the true gradient and its approximation is *gradient noise* and denoted by

$$\boldsymbol{g}(w) \triangleq \widehat{\nabla_{w^\mathsf{T}} P}(w) - \nabla_{w^\mathsf{T}} P(w) \tag{19.5}$$

When the stochastic gradient algorithm operates under importance sampling, the gradient approximations are further scaled by $1/Np_b$, where p_b is the probability of selecting sample $(\boldsymbol{\gamma}(b),\boldsymbol{h}_b)$:

(*approximations under importance sampling*)

$$(\textbf{instantaneous}) : \widehat{\nabla_{w^\mathsf{T}} P}(w) = \frac{1}{Np}\nabla_{w^\mathsf{T}} Q(w;\boldsymbol{\gamma},\boldsymbol{h}) \tag{19.6a}$$

$$(\textbf{mini-batch}) : \widehat{\nabla_{w^\mathsf{T}} P}(w) = \frac{1}{B}\sum_{b=0}^{B-1} \frac{1}{Np_b}\nabla_{w^\mathsf{T}} Q(w;\boldsymbol{\gamma}(b),\boldsymbol{h}_b) \tag{19.6b}$$

19.1.3 Conditions on Risk and Loss Functions

In Sec. 18.2 we introduced the following conditions for empirical risk minimization problems of the form (19.3a):

(A1) (**Strongly convex risk**). $P(w)$ is $\nu-$strongly convex and first-order differentiable, namely, for every $w_1, w_2 \in \mathrm{dom}(P)$:

$$P(w_2) \;\geq\; P(w_1) + (\nabla_{w^{\mathsf{T}}} P(w_1))^{\mathsf{T}}(w_2 - w_1) \;+\; \frac{\nu}{2}\|w_2 - w_1\|^2 \tag{19.7a}$$

(A2) ($\delta-$**Lipschitz loss gradients**). The gradient vectors of $Q(w,\cdot)$ are $\delta-$Lipschitz regardless of the data argument, i.e.,

$$\|\nabla_w Q(w_2;\gamma(k),h_k) - \nabla_w Q(w_1;\gamma(\ell),h_\ell)\| \;\leq\; \delta\,\|w_2 - w_1\| \tag{19.7b}$$

for any $w_1, w_2 \in \mathrm{dom}(Q)$, any $0 \leq k, \ell \leq N-1$, and with $\delta \geq \nu$. We explained that condition (19.7b) implies that the gradient of $P(w)$ is itself $\delta-$Lipschitz:

$$\|\nabla_w P(w_2) - \nabla_w P(w_1)\| \leq \delta\|w_2 - w_1\| \tag{19.8}$$

On the other hand, for stochastic risk minimization problems of the form (19.3b), we continue to assume the strong convexity of $P(w)$ but replace (**A2**) by the requirement that the gradients of the loss are now $\delta-$Lipschitz in the mean-square sense:

(A2') (**Mean-square $\delta-$Lipschitz loss gradients**). The gradient vectors of $Q(w,\cdot)$ satisfy the mean-square bound:

$$\mathbb{E}\,\|\nabla_w Q(w_2;\boldsymbol{\gamma},\boldsymbol{h}) - \nabla_w Q(w_1;\boldsymbol{\gamma},\boldsymbol{h})\|^2 \leq \delta^2\|w_2 - w_1\|^2 \tag{19.9}$$

for any $w_1, w_2 \in \mathrm{dom}(Q)$ and with $\delta \geq \nu$. We further showed that condition (19.9) implies that the gradients of $P(w)$ are $\delta-$Lipschitz as well:

$$\|\nabla_w P(w_2) - \nabla_w P(w_1)\| \leq \delta\|w_2 - w_1\| \tag{19.10}$$

which is the same condition (19.8) under empirical risk minimization.

Note that under conditions (**A1**,**A2**) for empirical risk minimization or (**A1**,**A2'**) for stochastic risk minimization, the following two conditions hold:

(**P1**): $\nu-$strong convexity of $P(w)$ as in (19.7a) (19.11a)

(**P2**): $\delta-$Lipschitz gradients for $P(w)$ as in (19.8) and (19.10) (19.11b)

Moreover, we know from the earlier results (8.29) and (10.20) derived for strongly-convex and $\delta-$smooth functions that conditions **P1** and **P2** imply respectively:

$$(\mathbf{P1}) \;\;\Longrightarrow\;\; \frac{\nu}{2}\|\widetilde{w}\|^2 \leq P(w) - P(w^\star) \leq \frac{1}{2\nu}\|\nabla_w P(w)\|^2 \tag{19.12a}$$

$$(\mathbf{P2}) \;\;\Longrightarrow\;\; \frac{1}{2\delta}\|\nabla_w P(w)\|^2 \leq P(w) - P(w^\star) \leq \frac{\delta}{2}\|\widetilde{w}\|^2 \tag{19.12b}$$

where $\widetilde{w} = w^\star - w$. The bounds in both expressions affirm that whenever we bound $\|\widetilde{w}\|^2$ we will also be automatically bounding the excess risk, $P(w) - P(w^\star)$.

19.1.4 Gradient Noise

We further showed in Sec. 18.3 that, as a result of the Lipschitz conditions (**A2**) or (**A2'**), the first and second-order moments of the gradient noise process satisfy the following two properties denoted by **G1** and **G2** for ease of reference:

$$\textbf{(G1):}\quad \mathbb{E}\,(\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\,|\,\boldsymbol{w}_{n-1}) = 0 \tag{19.13a}$$

$$\textbf{(G2):}\quad \mathbb{E}\,(\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2\,|\,\boldsymbol{w}_{n-1}) \leq \beta_g^2\|\widetilde{w}_{n-1}\|^2 + \sigma_g^2 \tag{19.13b}$$

for some nonnegative constants $\{\beta_g^2, \sigma_g^2\}$ that are independent of $\widetilde{w}_{n-1}$. Condition (**G1**) refers to the zero-mean property of the gradient noise while (**G2**) refers to its bounded "variance." We explained in the previous chapter that results (19.13a)–(19.13b) hold for instantaneous and mini-batch gradient approximations, regardless of whether the samples are streaming in independently of each other, sampled uniformly with replacement, sampled without replacement, or selected under importance sampling. The *only exception* is that the zero-mean property (19.13a) will *not* hold for the instantaneous gradient implementation when the samples are selected without replacement. This exception is not of major consequence for the convergence results in this chapter. When property (19.13a) does not hold, the convergence argument will need to be adjusted (and becomes more demanding) but will continue to lead to the same conclusion.

In summary, we find that conditions (**A1**,**A2**) for empirical risk minimization or (**A1**,**A2**') for stochastic risk minimization imply the validity of conditions **(P1,P2)** on the risk function and conditions **(G1,G2)** on the gradient noise:

$$(\textbf{A1},\textbf{A2})\text{ or }(\textbf{A1},\textbf{A2}') \implies (\textbf{P1},\textbf{P2},\,\textbf{G1},\textbf{G2}) \tag{19.14}$$

Remark 19.1. (Conditions on risk and loss functions) The $\delta-$Lipschitz conditions **A2** and **A2'** on the loss function were shown in the previous chapter to lead to the gradient noise properties **G1,G2**. They also imply the $\delta-$Lipschitz property **P2**. The convergence analyses in the sequel will rely largely on **G2** and **P2**, which relate to the bound on the second-order moment of the gradient noise and to the $\delta-$Lipschitz condition on the gradients of $P(w)$. While the two properties **(G2,P2)** follow from **(A2, A2')**, they can also be introduced on their own as starting assumptions for the convergence analysis.

■

19.2 CONVERGENCE UNDER UNIFORM SAMPLING

We are ready to examine the convergence behavior of the stochastic gradient recursion (19.1) in the mean-square-error sense under conditions (**A1**,**A2**) or (**A1**,**A2**').

19.2.1 Mean-Square Error Convergence

The first result shows that the mean-square error (MSE), denoted by $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2$, does not converge to zero but rather to a small neighborhood of size $O(\mu)$. Specifically, results (19.18b) and (19.18c) below mean that the behavior of $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2$ and the excess-risk $\mathbb{E}\,P(\boldsymbol{w}_n) - P(w^\star)$ can be described by the combined effect of two terms: one term $O(\lambda^n)$ decays exponentially at the rate λ^n and a second term $O(\mu)$ describes the size of the steady-state value that is left after sufficient iterations so that

$$\limsup_{n\to\infty}\ \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2 = O(\mu) \tag{19.15a}$$

$$\limsup_{n\to\infty}\ \Big(\mathbb{E}\,P(\boldsymbol{w}_n) - P(w^\star)\Big) = O(\mu) \tag{19.15b}$$

in terms of the *limit superior* of the variables involved. The limit superior of a sequence corresponds to the smallest upper bound for the limiting behavior of that sequence; this concept is useful when a sequence is not necessarily convergent but tends towards a small bounded region. This situation is illustrated schematically in Fig. 19.1 for $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2$. If a sequence happens to be convergent, then the limit superior will coincide with its normal limiting value.

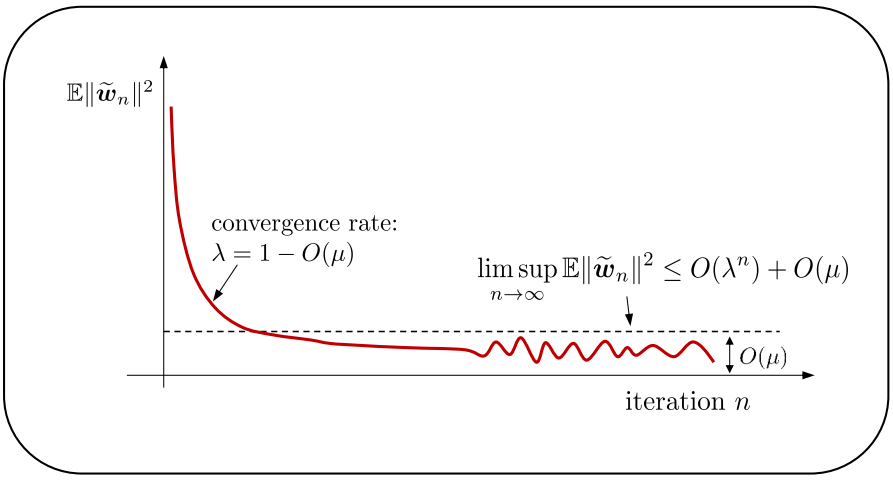

Figure 19.1 Exponential decay of the mean-square error described by expression (19.18a) to a level that is bounded by $O(\mu)$ and at a rate that is on the order of λ^n where $\lambda = 1 - O(\mu)$.

It further follows from the proof of the theorem below that, for sufficiently small step-sizes, the size of the $O(\mu)$ limiting region in the above expressions is actually dependent on σ_g^2 (the absolute noise term) since it will hold that — see

the arguments after (19.26) and (19.28):

$$\limsup_{n\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2 = O(\mu\sigma_g^2/2\nu) \tag{19.16a}$$

$$\limsup_{n\to\infty} \Big(P(\boldsymbol{w}_n) - P(w^\star)\Big) = O(\mu\sigma_g^2/4) \tag{19.16b}$$

where ν is the strong-convexity factor. Thus, observe from the statement of the theorem that the parameters (β_g^2, σ_g^2), which define the bound on the second-order moment of the gradient noise, affect performance in different ways. The value of β_g^2 affects both stability (by defining the bound on μ for convergence) and the rate of convergence λ, whereas σ_g^2 affects the size of the limiting region (i.e., the size of the steady-state error). This observation holds for all convergence results in this chapter.

Theorem 19.1. (MSE convergence under constant step-sizes) *Consider the stochastic gradient recursion (19.1) with the instantaneous gradient approximation (19.4a) under uniform data sampling or streaming data, used to seek the minimizers of empirical or stochastic risks. The risk and loss functions are assumed to satisfy conditions (**A1**,**A2**) or (**A1**,**A2'**). For step-size values satisfying (i.e., for μ small enough):*

$$\mu \;<\; \frac{2\nu}{\delta^2+\beta_g^2} \;\triangleq\; \mu_o \tag{19.17}$$

it holds that $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2$ and the average excess risk, $\mathbb{E}\,P(\boldsymbol{w}_n) - P(w^\star)$, converge exponentially fast according to the recursions:

$$\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2 \;\le\; \lambda\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{n-1}\|^2 \;+\; \mu^2\sigma_g^2 \tag{19.18a}$$

$$\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2 \le O(\lambda^n) + O(\mu) \tag{19.18b}$$

$$\mathbb{E}\,P(\boldsymbol{w}_n) - P(w^\star) \le O(\lambda^n) + O(\mu) \tag{19.18c}$$

where

$$\lambda \;\triangleq\; 1 - 2\nu\mu + (\delta^2+\beta_g^2)\mu^2 \;\in\; [0,1) \tag{19.19}$$

Results (19.18b)–(19.18c) hold for sufficiently small step-sizes.

Proof: We subtract $w^\star$ from both sides of (19.1) and use (19.5) to get

$$\widetilde{\boldsymbol{w}}_n \;=\; \widetilde{\boldsymbol{w}}_{n-1} \;+\; \mu\,\nabla_{w^{\mathsf{T}}}P(\boldsymbol{w}_{n-1}) \;+\; \mu\,\boldsymbol{g}_n(\boldsymbol{w}_{n-1}) \tag{19.20}$$

We will be squaring this expression. In preparation for that step, we first use the fact that $\nabla_w P(w^\star) = 0$ to note that for the first two terms on the right-hand side:

$$\begin{aligned}
&\|\widetilde{\boldsymbol{w}}_{n-1} + \mu\,\nabla_{w^{\mathsf{T}}}P(\boldsymbol{w}_{n-1})\|^2 \\
&= \|\widetilde{\boldsymbol{w}}_{n-1}\|^2 + 2\mu\,\big(\nabla_{w^{\mathsf{T}}}P(\boldsymbol{w}_{n-1})\big)^{\mathsf{T}}\,\widetilde{\boldsymbol{w}}_{n-1} + \mu^2\|\nabla_{w^{\mathsf{T}}}P(\boldsymbol{w}_{n-1})\|^2 \\
&= \|\widetilde{\boldsymbol{w}}_{n-1}\|^2 + 2\mu\,\big(\nabla_{w^{\mathsf{T}}}P(\boldsymbol{w}_{n-1})\big)^{\mathsf{T}}\,\widetilde{\boldsymbol{w}}_{n-1} + \mu^2\|\nabla_{w^{\mathsf{T}}}P(w^\star) - \nabla_{w^{\mathsf{T}}}P(\boldsymbol{w}_{n-1})\|^2 \\
&\overset{\textbf{(P2)}}{\le} \|\widetilde{\boldsymbol{w}}_{n-1}\|^2 + 2\mu\,\big(\nabla_{w^{\mathsf{T}}}P(\boldsymbol{w}_{n-1})\big)^{\mathsf{T}}\,\widetilde{\boldsymbol{w}}_{n-1} + \mu^2\delta^2\|\widetilde{\boldsymbol{w}}_{n-1}\|^2
\end{aligned} \tag{19.21}$$

Next, we appeal to the strong convexity property (19.7a) to get

$$\begin{aligned}(\nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-1}))^\mathsf{T}\,\widetilde{\boldsymbol{w}}_{n-1} &\leq P(w^\star) - P(\boldsymbol{w}_{n-1}) - \frac{\nu}{2}\|\widetilde{\boldsymbol{w}}_{n-1}\|^2 \\ &\overset{(8.23)}{\leq} -\frac{\nu}{2}\|\widetilde{\boldsymbol{w}}_{n-1}\|^2 \;-\; \frac{\nu}{2}\|\widetilde{\boldsymbol{w}}_{n-1}\|^2 \\ &= -\nu\|\widetilde{\boldsymbol{w}}_{n-1}\|^2 \end{aligned} \tag{19.22}$$

Substituting into (19.21) gives

$$\boxed{\|\widetilde{\boldsymbol{w}}_{n-1} + \mu\,\nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-1})\|^2 \leq (1 - 2\mu\nu + \mu^2\delta^2)\|\widetilde{\boldsymbol{w}}_{n-1}\|^2} \tag{19.23}$$

which is a useful intermediate result. Returning to (19.20), squaring both sides, conditioning on $\boldsymbol{w}_{n-1}$, and taking expectations we obtain

$$\begin{aligned}&\mathbb{E}\,\left(\|\widetilde{\boldsymbol{w}}_n\|^2 \,|\, \boldsymbol{w}_{n-1}\right) \\ &= \mathbb{E}\,\Big(\|\widetilde{\boldsymbol{w}}_{n-1} \;+\; \mu\,\nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-1}) \;+\; \mu\,\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2 \,|\, \boldsymbol{w}_{n-1}\Big) \\ &\overset{(a)}{=} \mathbb{E}\,\left(\|\widetilde{\boldsymbol{w}}_{n-1} \;+\; \mu\,\nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-1})\|^2 \,|\, \boldsymbol{w}_{n-1}\right) \;+ \\ &\qquad \mu^2\,\mathbb{E}\,\left(\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2 \,|\, \boldsymbol{w}_{n-1}\right) \\ &\overset{(19.23)}{\leq} (1 - 2\mu\nu + \mu^2\delta^2)\|\widetilde{w}_{n-1}\|^2 \;+\; \mu^2\,\mathbb{E}\,\left(\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2 \,|\, \boldsymbol{w}_{n-1}\right) \\ &\overset{\textbf{(G2)}}{\leq} (1 - 2\mu\nu + \mu^2\delta^2)\|\widetilde{w}_{n-1}\|^2 + \mu^2\left(\beta_g^2\|\widetilde{w}_{n-1}\|^2 + \sigma_g^2\right) \\ &= (1 - 2\mu\nu + \mu^2(\delta^2 + \beta_g^2))\|\widetilde{w}_{n-1}\|^2 + \mu^2\sigma_g^2 \end{aligned} \tag{19.24}$$

where the cross term in (a) is zero because of the zero-mean property **G1**; it is at this step that the zero-mean property of the gradient noise process is used. Taking expectations of both sides again removes the conditioning on $\boldsymbol{w}_{n-1}$ and leads to (19.18a) where λ is defined by (19.19). The same argument used in Fig. 12.2 can be repeated here to show that condition (19.17) ensures $0 \leq \lambda < 1$. By further iterating recursion (19.18a) we obtain

$$\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2 \leq \lambda^{n+1}\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{-1}\|^2 \;+\; \frac{\mu^2\sigma_g^2}{1-\lambda} \tag{19.25}$$

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

$$\limsup_{n\to\infty}\; \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2 \leq \frac{\mu^2\sigma_g^2}{1-\lambda} \;=\; \frac{\mu\sigma_g^2}{2\nu - \mu(\delta^2 + \beta_g^2)} \tag{19.26}$$

It is easy to check that the upper bound does not exceed $\mu\sigma_g^2/\nu$ for any step-size $\mu < \mu_o/2$. If, on the other hand, $\mu \ll \mu_o$ so that the denominator is approximately 2ν, then the upper bound is on the order of $\mu\sigma_g^2/2\nu$. We conclude that (19.18b) holds for sufficiently small step-sizes.

To establish (19.18c), we use (19.12b) to get

$$0 \leq \mathbb{E}\,P(\boldsymbol{w}_n) - P(w^\star) \leq \frac{\delta}{2}\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2 \tag{19.27}$$

so that from (19.26)

$$\limsup_{n\to\infty}\,\Big(\mathbb{E}\,P(\boldsymbol{w}_n) - P(w^\star)\Big) \leq \frac{\delta\mu\sigma_g^2}{2(2\nu - \mu(\delta^2 + \beta_g^2))} \tag{19.28}$$

where the upper bound does not exceed $\mu\delta\sigma_g^2/2\nu$ for any $\mu < \mu_o/2$. If, on the other

hand, $\mu \ll \mu_o$ so that the denominator is approximately 4ν, and since ν and δ are of the same order, then the upper bound is on the order of $\mu\sigma_g^2/4$. Either way, we conclude from (19.27) that $\mathbb{E}\,P(\boldsymbol{w}_n)$ converges towards an $O(\mu)-$neighborhood around $P(w^\star)$ at the same exponential rate as $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2$, which is λ^n.

■

Observe that we can rewrite (19.18a) in the equivalent form

$$\left(\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2 \;-\; \frac{\mu^2\sigma_g^2}{1-\lambda}\right) \;\leq\; \lambda\left(\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{n-1}\|^2 \;-\; \frac{\mu^2\sigma_g^2}{1-\lambda}\right) \tag{19.29}$$

where the steady-state bound is subtracted from both sides. It is clear from this representation that λ determines the rate of decay of the mean-square-error towards its steady-state bound — refer again to Fig. 19.1.

Example 19.1 (**Randomized coordinate-descent**) A similar convergence analysis can be applied to a randomized version of coordinate descent using stochastic gradient approximations — see listing (19.30).

Stochastic randomized coordinate-descent for solving (19.3a) or (19.3b)

given dataset $\{\gamma(m), h_m\}_{m=0}^{N-1}$ or streaming data $(\gamma(n), h_n)$;
start with an arbitrary initial condition $\boldsymbol{w}_{-1}$.
repeat until convergence over $n \geq 0$:
  iterate is $\boldsymbol{w}_{n-1} = \mathrm{col}\{\boldsymbol{w}_{n-1,m}\}_{m=1}^{M}$
  select at random or receive $(\boldsymbol{\gamma}(n), \boldsymbol{h}_n)$;
  select a random index $1 \leq m^o \leq M$;
  $\boldsymbol{w}_{n,m^o} = \boldsymbol{w}_{n-1,m^o} \;-\; \mu\dfrac{\partial Q_u(\boldsymbol{w}_{n-1}; \boldsymbol{\gamma}(n), \boldsymbol{h}_n)}{\partial w_{m^o}}$
  keep $\boldsymbol{w}_{n,m} = \boldsymbol{w}_{n-1,m}$ for all $m \neq m^o$
end
return $w^\star \leftarrow \boldsymbol{w}_n$. (19.30)

The same argument used in the proof of Theorem 19.1 can be repeated to establish convergence conditions for (19.30). We leave the analysis to Prob. 19.7.

19.2.2 Regret Analysis

For empirical risks minimized by stochastic gradient algorithms, the regret value over a window of N iterations is defined as the deviation of the accumulated *loss* from the minimal risk value:

$$\begin{aligned}\mathcal{R}(N) \;&\triangleq\; \frac{1}{N}\sum_{n=0}^{N-1} Q(\boldsymbol{w}_{n-1}; \boldsymbol{\gamma}(n), \boldsymbol{h}_n) \;-\; \min_{w\in\mathbb{R}^M}\left(\frac{1}{N}\sum_{n=0}^{N-1} Q(w; \gamma(n), h_n)\right) \\ &=\; \frac{1}{N}\sum_{n=0}^{N-1} Q(\boldsymbol{w}_{n-1}; \boldsymbol{\gamma}(n), \boldsymbol{h}_n) \;-\; P(w^\star)\end{aligned} \tag{19.31}$$

where the arguments $(\boldsymbol{w}_{n-1}, \boldsymbol{\gamma}(n), \boldsymbol{h}_n)$ are random due to uniform sampling and gradient noise. For this reason, we are denoting the regret variable in boldface to highlight its random nature as well. We already encountered this definition earlier in Sec. 17.2 while discussing the AdaGrad algorithm.

We may compare the above expression with the earlier definition (12.57) used for the gradient-descent case when the actual gradient of $P(w)$ is employed in the update recursion. We observe that in the above expression for the regret, the risk function $P(w)$ from (12.57) is replaced by the loss function $Q(w;\cdot)$. If we evaluate the conditional expectation of $\boldsymbol{\mathcal{R}}(N)$ over the trajectory of weight iterates, and use the unbiasedness property $\mathbb{E}\,Q(w,\boldsymbol{\gamma},\boldsymbol{h}) = P(w)$, we arrive at the following expression for the conditional regret:

$$\begin{aligned}\mathcal{R}(N) &\triangleq \mathbb{E}\left(\boldsymbol{\mathcal{R}}(N)\,|\,\boldsymbol{w}_{-1},\boldsymbol{w}_0,\ldots,\boldsymbol{w}_{N-1}\right)\\ &= \frac{1}{N}\sum_{n=0}^{N-1} P(w_{n-1}) \;-\; P(w^\star)\end{aligned} \tag{19.32}$$

which agrees with the earlier definition (12.57). In stochastic optimization implementations, it is common to employ the regret (19.32) as a performance measure as well. Using (19.18c) and the same argument that led to (12.58), we can readily find that the regret for the stochastic gradient algorithm (19.1) under uniform sampling satisfies — see Prob. 19.5:

$$\mathbb{E}\,\boldsymbol{\mathcal{R}}(N) \;\le\; O(1/N) \;+\; O(\mu) \tag{19.33}$$

19.3 CONVERGENCE OF MINI-BATCH IMPLEMENTATION

Let (β_g^2, σ_g^2) denote the parameters that characterize the bound in (19.13b) for the stochastic gradient algorithm based on *instantaneous* gradient approximations. We showed in (18.45) that these parameters get scaled down by a factor τ_B when a mini-batch implementation is used since the second-order moment of the gradient noise will then satisfy

$$\mathbb{E}\left(\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2\,|\,\boldsymbol{w}_{n-1}\right) \;\le\; \frac{1}{\tau_B}\left(\beta_g^2\|\widetilde{w}_{n-1}\|^2 + \sigma_g^2\right) \tag{19.34}$$

where the value of τ_B depends on whether the mini-batch samples are selected with or without replacement:

$$\tau_B \;\triangleq\; \begin{cases} B, & \text{sampling } \textit{with} \text{ replacement}\\ B\dfrac{N-1}{N-B}, & \text{sampling } \textit{without} \text{ replacement}\end{cases} \tag{19.35}$$

Observe from the second line that $\tau_B \approx B$ for N large enough. The same analysis used to establish Theorem 19.1 will lead to a similar conclusion apart from the scaling of β_g^2 by τ_B — see Prob. 19.1. Observe from the statement below how

the min-batch size B influences the performance expressions. In particular,the size of the steady-state neighborhood is reduced from $O(\mu)$ to $O(\mu/\tau_B)$.

Theorem 19.2. (MSE convergence of mini-batch implementation) *Consider the stochastic gradient recursion (19.1) with the mini-batch gradient approximation (19.4b) under random sampling with or without replacement or streaming data, used to seek the minimizers of empirical or stochastic risks. The risk and loss functions are assumed to satisfy conditions (**A1**,**A2**) or (**A1**,**A2'**). For step-size values satisfying (i.e., for μ small enough):*

$$\mu \;<\; \frac{2\nu}{\delta^2 + \frac{\beta_g^2}{\tau_B}} \;\triangleq\; \mu_o \tag{19.36}$$

it holds that $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2$ and the average excess risk, $\mathbb{E}\,P(\boldsymbol{w}_n) - P(w^\star)$, converge exponentially fast according to the recursions:

$$\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2 \;\leq\; \lambda\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{n-1}\|^2 \;+\; \frac{\mu^2\sigma_g^2}{\tau_B} \tag{19.37a}$$

$$\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2 \leq O(\lambda^n) + O(\mu/\tau_B) \tag{19.37b}$$

$$\mathbb{E}\,P(\boldsymbol{w}_n) - P(w^\star) \leq O(\lambda^n) + O(\mu/\tau_B) \tag{19.37c}$$

where

$$\lambda \;\triangleq\; 1 - 2\nu\mu + \Big(\delta^2 + \frac{\beta_g^2}{\tau_B}\Big)\mu^2 \;\in\; [0,1) \tag{19.38}$$

Results (19.37b) and (19.37c) hold for sufficiently small step-sizes.

19.4 CONVERGENCE UNDER VANISHING STEP-SIZES

We observe from (19.18b) that under constant step-size learning, the mean-square error $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2$ converges to a small neighborhood of size $O(\mu)$; the smaller the value of μ is, the smaller the size of this neighborhood will be. However, small step-sizes affect the convergence rate of the algorithm because they cause the value of λ to approach one. One way to reduce the size of the limiting region to zero is to employ a decaying step-size $\mu(n)$ in place of the constant μ. Doing so, allows us to employ larger step-size values during the initial stages of the algorithm to speed up convergence and smaller step-size values during the latter stages to improve steady-state performance. It is common to choose the sequence $\mu(n) > 0$ to satisfy either of the following two conditions:

$$\text{(\textbf{condition I})} \qquad \sum_{n=0}^{\infty}\mu^2(n) < \infty \quad \text{and} \quad \sum_{n=0}^{\infty}\mu(n) = \infty \tag{19.39a}$$

$$\text{(\textbf{condition II})} \qquad \lim_{n\to\infty}\mu(n) = 0 \quad \text{and} \quad \sum_{n=0}^{\infty}\mu(n) = \infty \tag{19.39b}$$

Clearly, any sequence that satisfies the stronger condition (19.39a) also satisfies (19.39b). Recursion (19.1) would then be replaced by

$$\boxed{\boldsymbol{w}_n \;=\; \boldsymbol{w}_{n-1} \;-\; \mu(n)\,\widehat{\nabla_{w^{\mathsf{T}}}P}(\boldsymbol{w}_{n-1}), \quad n \geq 0} \tag{19.40}$$

The decaying step-size helps annihilate the effect of gradient noise and ensures convergence of $\boldsymbol{w}_n$ to $w^\star$. Specifically, we will show below that $\boldsymbol{w}_n$ will now converge to $w^\star$ in the mean-square sense under both choices (19.39a) or (19.39b), i.e., $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2 \to 0$ as $n \to \infty$. Based on the discussion from Appendix 3.A on the convergence of random variables, this conclusion implies convergence in probability so that

$$\lim_{n\to\infty} \mathbb{P}\left(\|\widetilde{\boldsymbol{w}}_n\|^2 > \epsilon\right) = 0, \quad \text{for any small } \epsilon > 0 \tag{19.41}$$

We will actually establish below the stronger result that under (19.39a), $\boldsymbol{w}_n$ converges to $w^\star$ almost surely, i.e., with probability one:

$$\mathbb{P}\left(\lim_{n\to\infty} \boldsymbol{w}_n = w^\star\right) = 1 \tag{19.42}$$

While decaying step-size sequences of the form (19.39a)–(19.39b) provide favorable convergence properties towards $w^\star$, and assist in countering the effect of gradient noise, they nevertheless force the step-size to approach zero. This is problematic for applications requiring continuous learning from streaming data because the algorithm will update more slowly and become less effective at tracking drifts in the location of $w^\star$ due to changes in the statistical properties of the data.

Theorem 19.3. (Convergence under vanishing step-sizes) *Consider the stochastic gradient recursion (19.1) with the instantaneous gradient approximation (19.4a) under uniform data sampling or streaming data, used to seek the minimizers of empirical or stochastic risks. The risk and loss functions are assumed to satisfy conditions (**A1**,**A2**) or (**A1**,**A2'**). Then, the following convergence properties hold:*

- **(a)** *If the step-size sequence $\mu(n)$ satisfies (19.39a), then $\boldsymbol{w}_n$ converges almost surely to $w^\star$, written as $\boldsymbol{w}_n \to w^\star$ a.s.*
- **(b)** *If the step-size sequence $\mu(n)$ satisfies (19.39b), then $\boldsymbol{w}_n$ converges in the mean-square-error sense to $w^\star$, i.e., $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2 \to 0$, which in turn implies convergence in probability according to (19.41).*

Proof: The same argument leading to (19.24) for constant step-sizes continues to hold giving the inequality:

$$\mathbb{E}\left(\|\widetilde{\boldsymbol{w}}_n\|^2 \,|\, \boldsymbol{w}_{n-1}\right) \leq \lambda(n)\,\|\widetilde{w}_{n-1}\|^2 + \mu^2(n)\sigma_g^2 \tag{19.43}$$

with μ replaced by $\mu(n)$ and where now

$$\lambda(n) \;\triangleq\; 1 - 2\nu\mu(n) + (\delta^2 + \beta_g^2)\mu^2(n) \tag{19.44}$$

We split the term $2\nu\mu(n)$ into the sum of two terms and write

$$\lambda(n) \;=\; 1-\nu\mu(n)-\nu\mu(n)+(\delta^2+\beta_g^2)\mu^2(n) \tag{19.45}$$

Now, since $\mu(n)\to 0$, we conclude that for large enough $n > n_o$, the value of $\mu^2(n)$ is smaller than $\mu(n)$. Therefore, a large enough time index, n_o, exists such that the following two conditions are satisfied:

$$\nu\mu(n) \geq (\delta^2+\beta_g^2)\mu^2(n), \quad 0 \leq \nu\mu(n) < 1, \quad n > n_o \tag{19.46}$$

Consequently,

$$\lambda(n) \;\leq\; 1-\nu\mu(n), \quad n > n_o \tag{19.47}$$

Then, inequalities (19.43) and (19.47) imply that

$$\mathbb{E}\,\left(\|\widetilde{\boldsymbol{w}}_n\|^2 \,|\, \boldsymbol{w}_{n-1}\right) \;\leq\; (1-\nu\mu(n))\,\|\widetilde{w}_{n-1}\|^2 \;+\; \mu^2(n)\sigma_g^2, \quad n > n_o \tag{19.48}$$

Due to the Markovian property of recursion (19.40), where $\boldsymbol{w}_n$ is solely dependent on the most recent iterate $\boldsymbol{w}_{n-1}$, we can also write that

$$\mathbb{E}\,\left(\|\widetilde{\boldsymbol{w}}_n\|^2 \,|\, \boldsymbol{w}_{n-1},\ldots,\boldsymbol{w}_0,\boldsymbol{w}_{-1}\right) \;\leq\; (1-\nu\mu(n))\,\|\widetilde{w}_{n-1}\|^2 \;+\; \mu^2(n)\sigma_g^2, \quad n > n_o \tag{19.49}$$

where the conditioning on the left-hand side is now relative to the entire trajectory. For compactness of notation, let

$$\boldsymbol{u}(n+1) \;\triangleq\; \|\widetilde{\boldsymbol{w}}_n\|^2 \tag{19.50}$$

Then, inequality (19.49) implies

$$\mathbb{E}\left(\boldsymbol{u}(n+1)|\;\boldsymbol{u}(0),\boldsymbol{u}(1),\ldots,\boldsymbol{u}(n)\right) \;\leq\; (1-\nu\mu(n))\;u(n) \;+\; \mu^2(n)\sigma_g^2, \;\; n > n_o \tag{19.51}$$

We now call upon the useful result (19.158) from Appendix 19.A and make the identifications:

$$a(n) \leftarrow \nu\mu(n), \quad b(n) \leftarrow \mu^2(n)\sigma_g^2 \tag{19.52}$$

These sequences satisfy conditions (19.159) in the appendix in view of assumption (19.39a) on the step-size sequence and the second condition in (19.46). We then conclude that $\boldsymbol{u}(n)\to 0$ almost surely and, hence, $\boldsymbol{w}_n \to w^\star$ almost surely.

Finally, taking expectations of both sides of (19.51) leads to

$$\mathbb{E}\,\boldsymbol{u}(n+1) \;\leq\; (1-\nu\mu(n))\,\mathbb{E}\,\boldsymbol{u}(n) \;+\; \mu^2(n)\sigma_g^2, \quad n > n_o \tag{19.53}$$

with the expectation operator appearing on both sides of the inequality. Then, we conclude from the earlier result (14.136), under conditions (19.39b), that $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2 \to 0$ so that $\boldsymbol{w}_n$ converges to $w^\star$ in the mean-square-error sense.

■

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

$$\mu(n) = \frac{\tau}{n+1}, \quad \tau > 0 \tag{19.54}$$

which satisfy both conditions (19.39a) and (19.39b). This particular form for $\mu(n)$ is motivated in the next example. In contrast to the previous result (12.68a) on the convergence rate of gradient-descent algorithms, which was seen to be on the order of $O(1/n^{2\nu\tau})$, the next statement indicates that three rates of convergence are now possible depending on how $\nu\tau$ compares to the value one.

Theorem 19.4. (Rates of convergence under (19.54)) *Consider the stochastic gradient recursion (19.1) with the instantaneous gradient approximation (19.4a) under uniform data sampling or streaming data, used to seek the minimizers of empirical or stochastic risks. The risk and loss functions are assumed to satisfy conditions (**A1**,**A2**) or (**A1**,**A2'**). Assume the step-size sequence is selected according to (19.54). Then, three convergence rates are possible depending on how the factor $\nu\tau$ compares to the value one. Specifically, for large enough n, it holds that:*

$$\begin{cases} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2 \;\leq\; O\left(\frac{1}{n}\right), & \nu\tau > 1 \\ \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2 \;=\; O\left(\frac{\log n}{n}\right), & \nu\tau = 1 \\ \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2 \;=\; O\left(\frac{1}{n^{\nu\tau}}\right), & \nu\tau < 1 \end{cases} \tag{19.55}$$

The fastest convergence rate occurs when $\nu\tau > 1$ (i.e., for large enough τ) and is on the order of $O(1/n)$. The risk values follow a similar convergence behavior as $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2$, namely,

$$\begin{cases} \mathbb{E}\,P(\boldsymbol{w}_n) - P(w^\star) \;\leq\; O\left(\frac{1}{n}\right), & \nu\tau > 1 \\ \mathbb{E}\,P(\boldsymbol{w}_n) - P(w^\star) \;=\; O\left(\frac{\log n}{n}\right), & \nu\tau = 1 \\ \mathbb{E}\,P(\boldsymbol{w}_n) - P(w^\star) \;=\; O\left(\frac{1}{n^{\nu\tau}}\right), & \nu\tau < 1 \end{cases} \tag{19.56}$$

The fastest convergence rate again occurs when $\nu\tau > 1$ and is on the order of $O(1/n)$.

Proof: We use (19.53) and the assumed form for $\mu(n)$ in (19.54) to write

$$\mathbb{E}\,\boldsymbol{u}(n+1) \;\leq\; \left(1 - \frac{\nu\tau}{n+1}\right)\mathbb{E}\,\boldsymbol{u}(n) \;+\; \frac{\tau^2\sigma_g^2}{(n+1)^2}, \quad n > n_o \tag{19.57}$$

This recursion has the same form as (14.136) with the identifications:

$$a(n) \leftarrow \frac{\nu\tau}{n+1}, \quad b(n) \leftarrow \frac{\tau^2\sigma_g^2}{(n+1)^2}, \quad p \leftarrow 1 \tag{19.58}$$

The above rates of convergence then follow from the statement in part (b) of Lemma 14.1 from Appendix 14.A. Result (19.56) follows from (19.27).

■

Example 19.2 **(Motivating step-size sequences of the form (19.54))** We refer to expression (19.44) for $\lambda(n)$ and notice that the following relation holds whenever $\mu(n) < \nu/(\delta^2 + \beta_g^2)$ (this condition is possible for decaying step-sizes and large enough n):

$$1 - 2\nu\mu(n) + (\delta^2 + \beta_g^2)\mu^2(n) < 1 - \mu(n)\nu \tag{19.59}$$

Then, for large n and sufficiently small $\mu(n)$,

$$\begin{aligned}1-2\nu\mu(n)\nu+(\delta^2+\beta_g^2)\mu^2(n) &< 1-\mu(n)\nu\\ &\leq 1-\mu(n)\nu+\frac{\mu^2(n)\nu^2}{4}\\ &= \left(1-\frac{\mu(n)\nu}{2}\right)^2\\ &\leq 1-\frac{\mu(n)\nu}{2}\end{aligned} \tag{19.60}$$

so that taking expectations of both sides of (19.43):

$$\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2 \leq \left(1-\frac{\mu(n)\nu}{2}\right)\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{n-1}\|^2+\mu^2(n)\sigma_g^2 \tag{19.61}$$

We can select $\mu(n)$ to tighten the upper bound. By minimizing over $\mu(n)$ we arrive at the choice:

$$\mu^o(n)=\frac{\nu}{4\sigma_g^2}\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{n-1}\|^2 \tag{19.62}$$

We now verify that this choice leads to $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{n-1}\|^2 = O(1/n)$ so that the step-size sequence itself satisfies $\mu^o(n) = O(1/n)$. Indeed, substituting into (19.61) gives

$$\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2 \leq \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{n-1}\|^2\left(1-\frac{\nu^2}{16\sigma_g^2}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{n-1}\|^2\right) \tag{19.63}$$

Inverting both sides we obtain a linear recursion for the inverse quantity $1/\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2$:

$$\begin{aligned}\frac{1}{\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2} &\geq \frac{1}{\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{n-1}\|^2}\left(1-\frac{\nu^2}{16\sigma_g^2}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{n-1}\|^2\right)^{-1}\\ &\overset{(a)}{\geq} \frac{1}{\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{n-1}\|^2}\left(1+\frac{\nu^2}{16\sigma_g^2}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{n-1}\|^2\right)\\ &= \frac{1}{\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{n-1}\|^2}+\frac{\nu^2}{16\sigma_g^2}\end{aligned} \tag{19.64}$$

where in step (a) we used the fact that for any small enough scalar $x^2 < 1$, it holds that $1-x^2 \leq 1$ and

$$(1-x)(1+x)\leq 1 \implies (1-x)^{-1} \geq (1+x) \tag{19.65}$$

Iterating (19.64) gives a bound on the value of the mean-square-error:

$$\frac{1}{\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2} \geq \frac{1}{\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{-1}\|^2}+\frac{(n+1)\nu^2}{16\sigma_g^2} \tag{19.66}$$

Substituting into (19.62) we find that

$$\mu^o(n)\leq \frac{\nu}{4\sigma_g^2}\left(\frac{1}{\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{-1}\|^2}+\frac{n\nu^2}{16\sigma_g^2}\right)^{-1} = O(1/n) \tag{19.67}$$

Example 19.3 (**Comparing workloads**) Theorem 19.4 reveals that the stochastic gradient algorithm (19.1) is able to converge to the exact minimizer at the rate of $O(1/n)$. This means that the algorithm will need on the order of $1/\epsilon$ iterations for the average risk value $\mathbb{E}\,P(\boldsymbol{w}_n)$ to get $\epsilon-$close to the optimal value $P(w^\star)$. Since each iteration

requires the computation of a single gradient vector, we say that the workload (or computing time) that is needed is proportional to $1/\epsilon$:

$$\begin{aligned}&\textbf{workload or computing time} =\\ &\text{number of iterations} \times \text{gradient computations per iteration}\end{aligned} \tag{19.68}$$

The result of the theorem is equally applicable to *mini-batch* stochastic implementations — see Prob. 19.2. Therefore, a total of $1/\epsilon$ iterations will again be necessary for $\mathbb{E}\,P(\boldsymbol{w}_n)$ to get $\epsilon-$close to $P(w^\star)$. Now, however, for mini-batches of size B, it is necessary to evaluate B gradient vectors per iteration so that the workload is increased to B/ϵ.

If we were to rely instead on the *full-batch* gradient-descent implementation (12.28) for the minimization of the same empirical risk, then we know from the statement after (12.64a) that a smaller number of $\ln(1/\epsilon)$ iterations will be needed for the risk value $P(w_n)$ to get $\epsilon-$close to $P(w^\star)$. The workload in this case will become $N\ln(1/\epsilon)$. Table 19.1 summarizes the conclusions.

Table 19.1 Computation time or workload needed for the risk value of each algorithm to get $\epsilon-$close to the optimal value.

	algorithm for empirical risk minimization	workload or computing time
1.	stochastic gradient with decaying step-size	$1/\epsilon$
2.	mini-batch stochastic gradient with decaying step-size and mini-batch size B	B/ϵ
3.	full-batch gradient-descent with	$N\ln(1/\epsilon)$

Example 19.4 **(Comparing generalization abilities)** Each of the algorithms considered in the previous example is concerned with the solution of the empirical risk minimization problem (19.3a). For added clarity, in this example alone, we will refer to the risk function by writing $P_{\text{emp}}(w)$, where the subscript is meant to emphasize its empirical nature. Thus, these three algorithms are solving:

$$w^\star \;\triangleq\; \underset{w\in\mathbb{R}^M}{\text{argmin}} \left\{ P_{\text{emp}}(w) \;\triangleq\; \frac{1}{N}\sum_{m=0}^{N-1} Q(w;\gamma(m),h_m) \right\} \tag{19.69}$$

We explained earlier in Sec. 12.1.3 when discussing the concept of "generalization" that, ideally, we would like the solutions by these algorithms to serve as good approximations for the minimizer of the following stochastic optimization problem:

$$w^o \;\triangleq\; \underset{w\in\mathbb{R}^M}{\text{argmin}} \left\{ P(w) \;\triangleq\; \mathbb{E}\,Q(w;\boldsymbol{\gamma},\boldsymbol{h}) \right\} \tag{19.70}$$

Assume we run each algorithm for L iterations and let ϵ denote the resulting gap between the empirical risk at $\boldsymbol{w}_L$ and its optimal value, i.e.,

$$\epsilon \;\triangleq\; \mathbb{E}\left(P_{\text{emp}}(\boldsymbol{w}_L) - P_{\text{emp}}(w^\star)\right), \qquad (\textbf{empirical excess risk}) \tag{19.71}$$

Clearly, the value of ϵ is dependent on the algorithm. We can derive an expression that reveals how close $P(\boldsymbol{w}_L)$ gets to $P(w^o)$, namely, how close the *stochastic* risk at $\boldsymbol{w}_L$

gets to the optimal value. For this purpose, we first note that

$$\begin{aligned}
&\mathbb{E}\left(P(\boldsymbol{w}_L) - P(w^o)\right) \\
&= \mathbb{E}\left(P(\boldsymbol{w}_L) - P_{\text{emp}}(\boldsymbol{w}_L)\right) + \underbrace{\mathbb{E}\left(P_{\text{emp}}(\boldsymbol{w}_L) - P_{\text{emp}}(w^\star)\right)}_{=\,\epsilon} + \\
&\quad \underbrace{\mathbb{E}\left(P_{\text{emp}}(w^\star) - P_{\text{emp}}(w^o)\right)}_{\leq 0} + \mathbb{E}\left(P_{\text{emp}}(w^o) - P(w^o)\right) \\
&\geq \epsilon + \mathbb{E}\left(P(\boldsymbol{w}_L) - P_{\text{emp}}(\boldsymbol{w}_L)\right) + \mathbb{E}\left(P_{\text{emp}}(w^o) - P(w^o)\right) \qquad (19.72)
\end{aligned}$$

so that

$$\boxed{\begin{array}{l}(\textbf{stochastic excess risk}) \\ \mathbb{E}\left(P(\boldsymbol{w}_L) - P(w^o)\right) = \epsilon + O\left(\sqrt{\dfrac{2\ln\ln N}{N}}\right)\end{array}} \qquad (19.73)$$

where we used result (3.226) to approximate the differences between the empirical and true risk values in the last two terms appearing in (19.72). Result (19.73) shows that, for N large enough,

$$\left(\begin{array}{l}\textbf{stochastic} \\ \textbf{excess risk}\end{array}\right) = \left(\begin{array}{l}\textbf{empirical} \\ \textbf{excess risk}\end{array}\right) + O\left(\sqrt{\frac{2\ln\ln N}{N}}\right) \qquad (19.74)$$

This expression reveals how well algorithms generalize. It shows that the stochastic excess risk depends on two factors: the sample size N *and* the empirical excess risk (19.71).

Assume we fix the computational (or workload) budget at a maximum value $\mathcal{C}_{\max}$ for each of the algorithms. Consider first the stochastic gradient implementation (19.1). From the first row in Table 19.1 we know that this algorithm will lead to an empirical excess risk on the order of $\epsilon = 1/\mathcal{C}_{\max}$ and, moreover, this excess is independent of N. Therefore, we conclude from (19.73) that increasing the sample size N will help reduce the stochastic excess risk and improve generalization.

In contrast, consider next the full-batch gradient-descent algorithm (12.28). From the third row in Table 19.1 we know that this algorithm will lead to an empirical excess risk on the order of $\epsilon = e^{-\mathcal{C}_{\max}/N}$, which *depends* on N. In other words, both factors on the right-hand side of (19.73) will now be dependent on N and an optimal choice for N can be selected.

19.5 CONVERGENCE UNDER RANDOM RESHUFFLING

We return to the stochastic gradient algorithm (19.1) with the instantaneous gradient approximation (19.4a) and constant step-size μ. We now examine its convergence behavior under random reshuffling (i.e., sampling without replacement) for *empirical risk minimization*. In this case, the gradient noise process does not have zero mean, and we need to adjust the convergence argument. The analysis will require that we make explicit the multiple epochs (or runs) over

the data. Thus, let k denote the epoch index. Before the start of an epoch, the $N-$size data $\{\gamma(m), h_m\}$ is reshuffled at random. During the run, we select samples sequentially from the reshuffled dataset. We describe the algorithm in listing (19.75).

Stochastic gradient algorithm with random reshuffling for solving the empirical risk minimization problem (19.3a)

given dataset $\{\gamma(m), h_m\}_{m=0}^{N-1}$;
start from an arbitrary initial condition $\boldsymbol{w}_{N-1}^0$.
for each run $k = 1, 2, \ldots, K$:
  set $\boldsymbol{w}_{-1}^k = \boldsymbol{w}_{N-1}^{k-1}$;
  reshuffle the dataset;
  repeat for $n = 0, 1, 2, \ldots, N-1$:
    $\boldsymbol{w}_n^k = \boldsymbol{w}_{n-1}^k - \mu \nabla_{w^\mathsf{T}} Q(\boldsymbol{w}_{n-1}^k; \boldsymbol{\gamma}(n), \boldsymbol{h}_n)$
  end
end
return $w^\star \leftarrow \boldsymbol{w}_{N-1}^K$ (19.75)

In this description, the notation $(\boldsymbol{\gamma}(n), \boldsymbol{h}_n)$ denotes the random sample that is selected at iteration n of the $k-$th run. The initial iterate for each run is the value that is attained at the end of the previous run:

$$\boldsymbol{w}_{-1}^k = \boldsymbol{w}_{N-1}^{k-1} \tag{19.76}$$

Since operation under random reshuffling corresponds to sampling *without* replacement, it is clear that no sample points are repeated during each run of the algorithm. This is in contrast to uniform sampling *with replacement*, where some data points may be repeated during the same run of the algorithm. The argument will show that this simple adjustment to the operation of the algorithm, using data sampling *without* as opposed to *with* replacement, results in performance *improvement*. The mean-square-error, $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2$, will now be reduced to $O(\mu^2)$ in comparison to the earlier $O(\mu)$ value shown in (19.18b). The proof of the following theorem appears in Appendix 19.B.

Theorem 19.5. (Convergence under random reshuffling) *Consider the stochastic gradient recursion (19.1) with the instantaneous gradient approximation (19.4a) under data sampling without replacement, used to seek the minimizers of empirical risks. The risk and loss functions are assumed to satisfy conditions (**A1**,**A2**) or (**A1**,**A2'**). For step-size values satisfying:*

$$\mu < \frac{\nu}{\sqrt{24}N\delta^2} \tag{19.77}$$

it holds that the mean-square error $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n^k\|^2$ converges exponentially over the epoch index k at the rate λ^k where

$$\lambda = 1 - \frac{\mu}{2}\nu N \tag{19.78}$$

and, for any $0 \le n \le N-1$:

$$\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n^k\|^2 \le O(\lambda^k) + O(\mu^2) \tag{19.79a}$$

$$\mathbb{E}\,P(\boldsymbol{w}_n^k) - P(w^\star) \le O(\lambda^k) + O(\mu^2) \tag{19.79b}$$

Observe that the results in the theorem are expressed in terms of the epoch index k tending to $+\infty$. We can also examine the behavior of random reshuffling when an *epoch-dependent* step-size is used, say, of the form

$$\mu(k) = \tau/k, \quad k \ge 1, \quad \tau > 0 \tag{19.80}$$

By repeating the arguments from Appendix 19.B and the technique used to establish Theorem 19.4, we can similarly verify that three convergence rates are possible depending on how the factor $\nu\tau N/2$ compares to the value one. Specifically, for any $0 \le n \le N-1$ and large enough k, it holds that — see Prob. 19.9:

$$\begin{cases} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n^k\|^2 \;\le\; O\left(\frac{1}{k}\right), & \nu\tau N > 2 \\ \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n^k\|^2 \;=\; O\left(\frac{\log k}{k}\right), & \nu\tau N = 2 \\ \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n^k\|^2 \;=\; O\left(\frac{1}{k^{\nu\tau N/2}}\right), & \nu\tau N < 2 \end{cases} \tag{19.81}$$

The fastest convergence rate occurs when $\nu\tau N > 2$ (i.e., for large enough τ) and is on the order of $O(1/k)$. The risk values follow a similar convergence behavior as $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n^k\|^2$, namely,

$$\begin{cases} \mathbb{E}\,P(\boldsymbol{w}_n^k) - P(w^\star) \;\le\; O\left(\frac{1}{k}\right), & \nu\tau N > 2 \\ \mathbb{E}\,P(\boldsymbol{w}_n^k) - P(w^\star) \;=\; O\left(\frac{\log k}{k}\right), & \nu\tau N = 2 \\ \mathbb{E}\,P(\boldsymbol{w}_n^k) - P(w^\star) \;=\; O\left(\frac{1}{k^{\nu\tau N/2}}\right), & \nu\tau N < 2 \end{cases} \tag{19.82}$$

The fastest convergence rate again occurs when $\nu\tau N > 2$ and is on the order of $O(1/k)$.

Example 19.5 (**Simulating random reshuffling**) We compare the performance of the stochastic gradient algorithm (19.1) with constant step-size under both uniform sampling and random reshuffling for the instantaneous gradient approximation (19.4a).

The objective is to illustrate the superior steady-state performance under random reshuffling. According to result (19.79b), if we plot the steady-state deviation value $P(\boldsymbol{w}_n^k) - P(w^\star)$, as $k \to \infty$, versus the step-size parameter μ in a log-log scale, the slope of the resulting line should be at least two since

$$\log_{10}\Big(\mathbb{E}\, P(\boldsymbol{w}_n^k) - P(w^\star)\Big) \;\leq\; 2\log_{10}(\mu), \quad k \to \infty \tag{19.83}$$

In other words, if the step size is reduced by a factor of 10 from μ to $\mu/10$, then the risk deviation should be reduced by at least a factor of 100. In comparison, under uniform sampling, we know from (19.18c) that the risk deviation will be reduced by at least the same factor 10. We illustrate this behavior by means of a simulation. Consider theℓ_2-regularized logistic empirical risk:

$$P(w) = \rho\|w\|^2 + \frac{1}{N}\sum_{m=0}^{N-1} \ln\Big(1 + e^{-\gamma(m)h_m^{\mathsf{T}}w}\Big), \quad w \in \mathbb{R}^M \tag{19.84}$$

with $\rho = 0.1$ and $M = 10$. The step-size parameter is varied between 10^{-4} and 10^{-3}. The simulation generates $N = 1000$ random pairs of data $\{\gamma(m), h_m\}$ according to a logistic model. First, a random parameter model $w^a \in \mathbb{R}^{10}$ is selected, and a random collection of feature vectors $\{h_m\}$ are generated, say, with zero-mean unit-variance Gaussian entries. Then, for each h_m, the label $\gamma(m)$ is set to either $+1$ or -1 according to the following construction:

$$\gamma(m) = +1 \;\text{ if }\; \Big(\frac{1}{1 + e^{-h_m^{\mathsf{T}}w^a}}\Big) \geq 0.5; \;\text{ otherwise } \gamma(m) = -1 \tag{19.85}$$

A total of $K = 2000$ epochs are run over the data. In one simulation, we evaluate the risk value $P(w_{-1}^k)$ at the beginning of each epoch and subsequently average these values over all epochs to approximate the average deviation

$$\mathbb{E}\, P(\boldsymbol{w}_n^k) - P(w^\star) \;\approx\; \frac{1}{K}\sum_{k'=1}^{K} P(w_{-1}^{k'}) - P(w^\star) \tag{19.86}$$

In a second simulation, we use the risk value $P(w_{-1}^K)$ at the last epoch as the approximation for $\mathbb{E}\, P(\boldsymbol{w}_{-1}^k)$, i.e.,

$$\mathbb{E}\, P(\boldsymbol{w}_{-1}^k) - P(w^\star) \;\approx\; P(w_{-1}^K) - P(w^\star) \tag{19.87}$$

Both approximations lead to similar results. We plot the variation of the risk deviations in the logarithmic scale against $\log_{10}(\mu)$ in Fig. 19.2. The plot shows the simulated values for these risk deviations against the step-size parameter. The vertical and horizontal scales are logarithmic. The dotted lines are the fitted regression lines, which provide an estimate of the slope variations for the measurements. The slopes of the lines for uniform sampling and random reshuffling are found in this simulation to be 1.3268 and 2.8512, respectively.

19.6 CONVERGENCE UNDER IMPORTANCE SAMPLING

We now examine the convergence behavior of the stochastic gradient algorithm (19.1) under *importance sampling* for empirical risk minimization. In this implementation, a probability value p_m is assigned to each sample $(\gamma(m), h_m)$ in the dataset, and the samples are selected at random according to this distribution.

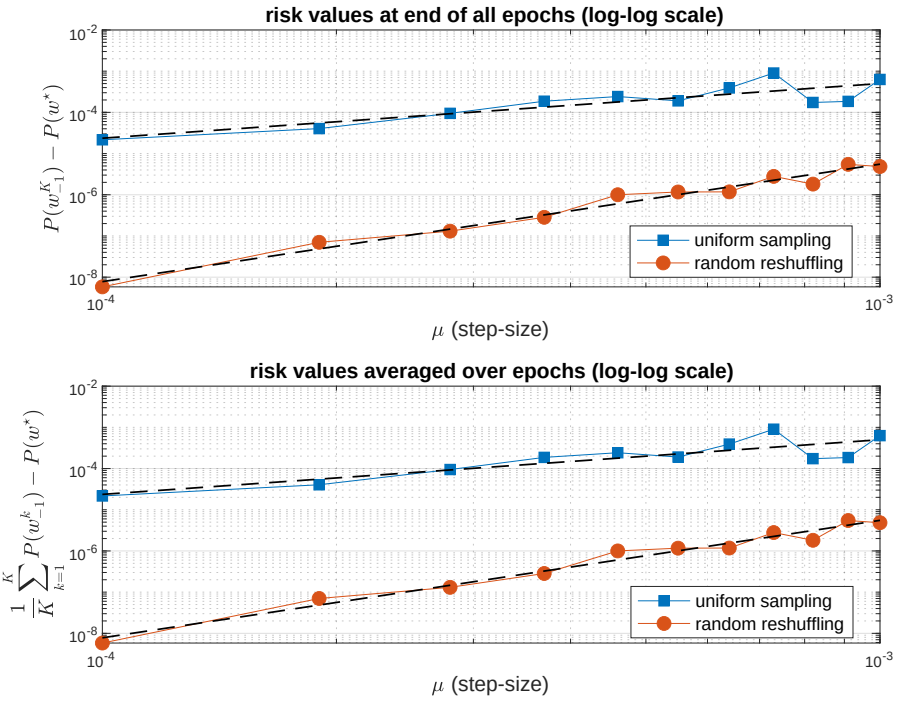

Figure 19.2 Random reshuffling has better risk deviation performance than uniform sampling. The plot shows the simulated values for these risk deviations against the step-size parameter. The vertical and horizontal scales are logarithmic. The dotted lines are the fitted regression lines, which provide estimates for the slopes.

We explained earlier that the approximations for the gradient vector will need to be adjusted and scaled as shown in (19.6a)–(19.6b).

The result of Theorem 19.1 can be extended to the scaled gradient approximations (19.6a)–(19.6b). We will therefore leave the analysis to the problems — see Prob. 19.17, where it is shown that the limiting mean-square error will continue to be $O(\mu\sigma_g^2)$, where the expression for σ_g^2 was derived earlier in (18.52b):

$$\sigma_g^2 = \frac{2}{N^2}\sum_{m=0}^{N-1}\frac{1}{p_m}\left\|\nabla_{w^\mathsf{T}} Q(w^\star;\gamma(m),h_m)\right\|^2 \tag{19.88}$$

for instantaneous gradient approximations. A similar expression holds with σ_g^2 divided by B for the mini-batch version. Observe that the expression for σ_g^2 involves the selection probabilities $\{p_m\}$.

19.6.1 Optimal Importance Sampling

One important question that is relevant for importance sampling implementations is the choice of the selection probabilities $\{p_m\}$. One possibility is to

minimize σ_g^2 over the $\{p_m\}$ in order to reduce the size of the $O(\mu\sigma_g^2)-$limiting neighborhood. Thus, we consider the following constrained optimization problem:

$$\{p_m^o\} \triangleq \underset{\{p_m\}}{\text{argmin}} \left\{ \sum_{m=0}^{N-1} \frac{1}{p_m} \|\nabla_{w^\mathsf{T}} Q(w^\star; \gamma(m), h_m)\|^2 \right\} \tag{19.89a}$$

subject to

$$0 \le p_m \le 1, \ \sum_{m=0}^{N-1} p_m = 1 \tag{19.89b}$$

This problem has a closed-form solution. Let us ignore for the moment the constraint $0 \le p_m \le 1$ and solve the remaining constrained problem by introducing the Lagrangian function:

$$\mathcal{L}(p_m, \alpha) \triangleq \sum_{m=0}^{N-1} \frac{1}{p_m} \|\nabla_{w^\mathsf{T}} Q(w^\star; \gamma(m), h_m)\|^2 + \alpha \left(\sum_{m=0}^{N-1} p_m - 1 \right) \tag{19.90}$$

where α is a Lagrange multiplier. Differentiating relative to p_m and setting the derivative to zero gives an expression for p_m^o in terms of α:

$$p_m^o = \frac{1}{\sqrt{\alpha}} \|\nabla_{w^\mathsf{T}} Q(w^\star; \gamma(m), h_m)\| \tag{19.91}$$

Since the sum of the $\{p_m^o\}$ must be one, we find that

$$\sqrt{\alpha} = \sum_{m=0}^{N-1} \|\nabla_{w^\mathsf{T}} Q(w^\star; \gamma(m), h_m)\| \tag{19.92}$$

and, hence,

$$p_m^o = \frac{\|\nabla_{w^\mathsf{T}} Q(w^\star; \gamma(m), h_m)\|}{\sum_{m=0}^{N-1} \|\nabla_{w^\mathsf{T}} Q(w^\star; \gamma(m), h_m)\|} \tag{19.93}$$

This solution satisfies the constraint $0 \le p_m \le 1$ and leads to an optimal sampling strategy. However, this particular strategy is not practical for two reasons: the values of p_m^o depend on the unknown $w^\star$, and the denominator involves a sum over all N data samples.

19.6.2 Adaptive Importance Sampling

We can address the first difficulty, at every iteration n, by replacing $w^\star$ by the estimate that is available at the start of that iteration, namely, w_{n-1}. This leads to an *adaptive* importance sampling procedure with:

$$p_{m,n}^o = \frac{\|\nabla_{w^\mathsf{T}} Q(w_{n-1}; \gamma(m), h_m)\|}{\sum_{m=0}^{N-1} \|\nabla_{w^\mathsf{T}} Q(w_{n-1}; \gamma(m), h_m)\|}, \ m = 0, 1, \ldots, N-1 \tag{19.94}$$

where we are adding a subscript n to indicate that the probabilities $\{p_{m,n}^o\}$ are the ones used at iteration n while updating $\boldsymbol{w}_{n-1}$ to $\boldsymbol{w}_n$. Expression (19.94) is

still inefficient because of the sum in the denominator, which involves all data samples and N can be large. We address this second difficulty by devising a recursive scheme to update the denominator.

We introduce an auxiliary vector variable $\psi_n \in \mathbb{R}^N$, whose size is equal to the number of data samples. One entry of $\boldsymbol{\psi}_n$ is updated at each iteration n. Let σ denote the index of the sample $(\boldsymbol{\gamma}, \boldsymbol{h})$ that is selected for use at iteration n. Then, only the $\sigma-$th entry of ψ_n is updated at that iteration:

$$\psi_{n,\sigma} = \beta\psi_{n-1,\sigma} + (1-\beta)\|\nabla_{w^{\mathsf{T}}} Q(w_{n-1}; \gamma(\sigma), h_\sigma)\| \tag{19.95}$$

where $\beta \in (0,1)$ is a design parameter. All other entries of ψ_n stay identical to the entries from the previous instant ψ_{n-1}. We can express this update in vector form as follows. Let D_σ denote the $N \times N$ diagonal matrix with β at location (σ,σ) and ones at all other diagonal entries:

$$D_\sigma = \mathrm{diag}\Big\{1, \ldots, 1, \beta, 1, \ldots, 1\Big\}, \quad (N \times N) \tag{19.96}$$

Let also e_σ denote the basis vector in $\mathbb{R}^N$ with a unit entry at location σ. Then,

$$\boxed{\psi_n = D_\sigma\psi_{n-1} + (1-\beta)\|\nabla_{w^{\mathsf{T}}} Q(w_{n-1}; \gamma(\sigma), h_\sigma)\|\, e_\sigma, \;\; n \geq 0} \tag{19.97}$$

We initialize ψ_{-1} to large positive values. Note that at iteration n, only one entry of ψ_n is updated, and hence this update is computationally inexpensive. Moreover, each entry of index σ in ψ_n corresponds to a smooth running estimate of the norm $\|\nabla_{w^{\mathsf{T}}} Q(w_{n-1}; \gamma(\sigma), h_\sigma)\|$ (which is the quantity that appears in the numerator of (19.94)).

We introduce a second auxiliary scalar quantity, denoted by τ_n for iteration n, in order to keep track of the sum of the entries of ψ; this sum (and, hence, τ) will serve as the approximation for the quantity appearing in the denominator of $p^o_{m,n}$:

$$\tau_n \;\overset{\Delta}{=}\; \|\psi_n\|_1 = \sum_{m=0}^{N-1} \psi_{n,m} = \sum_{m=0}^{N-1} \psi_{n-1,m} + (\psi_{n,\sigma} - \psi_{n-1,\sigma}) \tag{19.98}$$

and, hence, using (19.97):

$$\boxed{\tau_n = \tau_{n-1} + (1-\beta)\Big(\|\nabla_{w^{\mathsf{T}}} Q(w_{n-1}; \gamma(\sigma), h_\sigma)\| - \psi_{n-1,\sigma}\Big), \; n \geq 0} \tag{19.99}$$

with $\tau_{-1} = \|\psi_{-1}\|_1$. Note that each update of τ only requires $O(1)$ operations, which is also inexpensive. This construction leads to a procedure that automatically learns an "optimal" sampling strategy. The algorithm is listed below using instantaneous gradient approximations. In the listing, the vector $r_n \in \mathbb{R}^N$ contains the values of the probabilities $\{p_{m,n}\}$ used at iteration n:

$$r_n \;\overset{\Delta}{=}\; \{p_{m,n}\}, \;\; m = 1, 2, \ldots, M \tag{19.100}$$

Stochastic gradient algorithm with adaptive importance sampling for minimizing the empirical risk (19.3a)

given dataset $\{\gamma(m), h_m\}_{m=0}^{N-1}$;
given a scalar $\beta \in (0,1)$;
start from an arbitrary initial condition $\boldsymbol{w}_{-1} \in \mathbb{R}^M$;
initialize $\boldsymbol{\psi}_{-1} \in \mathbb{R}^N$ to large positive entries;
set $\boldsymbol{\tau}_{-1} = \|\boldsymbol{\psi}_{-1}\|_1$;
set $\boldsymbol{r}_0 = \frac{1}{N}\mathbb{1}_N$; (uniform distribution).
repeat until convergence over $n \geq 0$:

$$
\begin{aligned}
&\text{entries of } \boldsymbol{r}_n \text{ are the probabilities } \{p_{m,n}\}_{m=0}^{N-1} \text{ at iteration } n;\\
&\text{select an index } 0 \leq \sigma \leq N-1 \text{ according to probabilities } \{p_{m,n}\};\\
&\text{let } \boldsymbol{x}_n = \nabla_{w^{\mathsf{T}}} Q(\boldsymbol{w}_{n-1}; \boldsymbol{\gamma}(\sigma), \boldsymbol{h}_\sigma); \quad \text{(approximate gradient)}\\
&\boldsymbol{w}_n = \boldsymbol{w}_{n-1} - \frac{\mu}{N p_{\sigma,n}} \boldsymbol{x}_n\\
&D_\sigma = \mathrm{diag}\{1, \ldots, 1, \beta, 1, \ldots, 1\}; \quad (\beta \text{ at } \sigma\text{−th location})\\
&\boldsymbol{\psi}_n = D_\sigma \boldsymbol{\psi}_{n-1} + (1-\beta)\|\boldsymbol{x}_n\| \, e_\sigma\\
&\boldsymbol{\tau}_n = \boldsymbol{\tau}_{n-1} + (1-\beta)(\|\boldsymbol{x}_n\| - \boldsymbol{\psi}_{n-1,\sigma})\\
&\boldsymbol{r}_{n+1} = \boldsymbol{\psi}_n / \boldsymbol{\tau}_n
\end{aligned}
\tag{19.101}
$$

end
return $w^\star \leftarrow \boldsymbol{w}_n$

Example 19.6 (**Simulating importance sampling**) We illustrate the results by considering the regularized logistic regression problem:

$$
P(w) = \rho\|w\|^2 + \frac{1}{N}\sum_{m=0}^{N-1} \ln\left(1 + e^{-\gamma(m) h_m^{\mathsf{T}} w}\right) \tag{19.102}
$$

where $h_m \in \mathbb{R}^{10}$ and $\gamma(m) \in \{\pm 1\}$. In the simulation, we generate a random dataset $\{h_m, \gamma(m)\}$ with $N = 500$ using the same logistic model from Example 19.5. We set $\rho = 0.01$, $\mu = 0.001$, and $\beta = 0.25$. We also set the initial condition $\psi_{-1} = 1000\,\mathbb{1}_N$. We run algorithm (19.101) over $K = 200$ epochs and compute the risk values $P(w_{-1}^k)$ at the start of each epoch. This leads to a risk deviation curve $P(w_{-1}^k) - P(w^\star)$ over the epoch index k. We repeat this simulation over $L = 100$ trials and average the deviation curves. The results for uniform sampling and importance sampling are shown in Fig. 19.3.

Table 19.2 summarizes the performance results obtained in this chapter for various stochastic gradient algorithms under uniform sampling, importance sampling, random reshuffling, data streaming, and also for mini-batch implementations. Results are shown for strongly-convex risks under both constant and vanishing step-sizes.

Table 19.2 Convergence properties of the stochastic gradient algorithm (19.1) with *constant* and *vanishing* step-sizes, where $P(w)$ denotes either an empirical risk with minimizer $w^\star$ or a stochastic risk with minimizer w^o.

mode of operation	conditions on risk and loss functions	asymptotic convergence property	reference
instantaneous gradient approximation (uniform sampling)	$P(w)$: $\nu-$strongly convex $Q(w)$: $\delta-$Lipschitz gradients conditions (**A1,A2**)	$\mathbb{E}\,P(\boldsymbol{w}_n) - P(w^\star) \leq O(\lambda^n) + O(\mu)$	Thm. 19.1
instantaneous gradient approximation (data streaming)	$P(w)$: $\nu-$strongly convex $Q(w)$: $\delta-$Lipschitz gradients in mean-square sense conditions (**A1,A2'**)	$\mathbb{E}\,P(\boldsymbol{w}_n) - P(w^o) \leq O(\lambda^n) + O(\mu)$	Thm. 19.1
instantaneous gradient approximation (random reshuffling)	$P(w)$: $\nu-$strongly convex $Q(w)$: $\delta-$Lipschitz gradients conditions (**A1,A2**)	$\mathbb{E}\,P(\boldsymbol{w}_n^k) - P(w^\star) \leq O(\lambda^k) + O(\mu^2)$ $k \geq 1$: epoch index $n \geq 0$: iteration index	Thm. 19.5
instantaneous gradient approximation (importance sampling)	$P(w)$: $\nu-$strongly convex $Q(w)$: $\delta-$Lipschitz gradients conditions (**A1,A2**)	$\mathbb{E}\,P(\boldsymbol{w}_n) - P(w^\star) \leq O(\lambda^n) + O(\mu)$	Prob. 19.17 Sec. 19.6
mini-batch gradient approximation (sampling with or without) (replacement or data streaming)	$P(w)$: $\nu-$strongly convex $Q(w)$: $\delta-$Lipschitz gradients in deterministic or mean-square sense conditions (**A1,A2**) or (**A1,A2'**)	$\mathbb{E}\,P(\boldsymbol{w}_n) - P(w^\star) \leq O(\lambda^n) + O(\mu/B)$	Thm. 19.2
instantaneous gradient approximation (uniform sampling or data streaming) (decaying step-size)	$P(w)$: $\nu-$strongly convex $Q(w)$: $\delta-$Lipschitz gradients in deterministic or mean-square sense conditions (**A1,A2**) or (**A1,A2'**) $\mu(n) = \tau/(n+1)$	$\mathbb{E}\,P(\boldsymbol{w}_n) - P(w^\star) \leq O(1/n)$	Thm. 19.4
instantaneous gradient approximation (random reshuffling) (decaying step-size)	$P(w)$: $\nu-$strongly convex $Q(w)$: $\delta-$Lipschitz gradients conditions (**A1,A2**) $\mu(k) = \tau/k$	$\mathbb{E}\,P(\boldsymbol{w}_n^k) - P(w^\star) \leq O(1/k)$	Eq. (19.82)

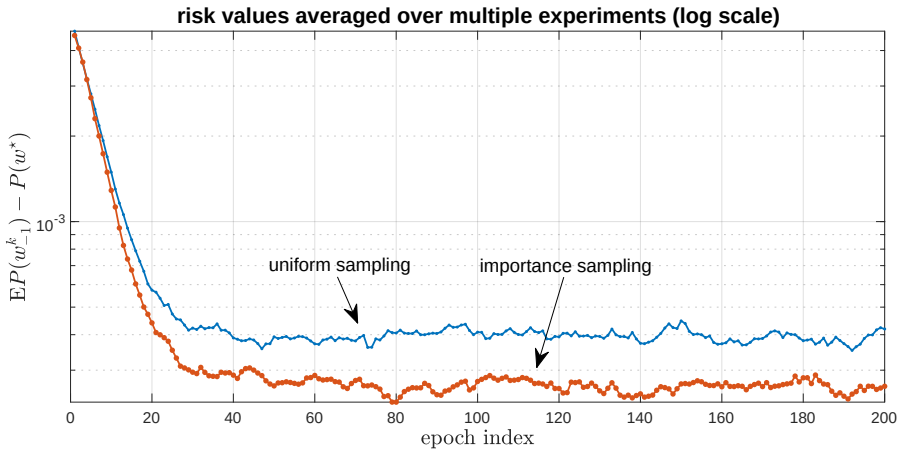

Figure 19.3 Adaptive importance sampling for a regularized logistic regression problem.

19.7 CONVERGENCE OF STOCHASTIC CONJUGATE GRADIENT

We examine in this last section the convergence of the *stochastic* version of the Fletcher-Reeves algorithm listed earlier in (16.102) and repeated in (19.105) for ease of reference. The data are assumed to be sampled uniformly with replacement from a dataset $\{\gamma(m), h_m\}$ for $m = 0, 1, \ldots, N-1$, and the algorithm is used to solve an empirical minimization problem of the form (19.3a).

The convergence analysis given here extends the arguments from Sec. 13.2 to the stochastic case where the gradient of the risk function is replaced by the gradient of the loss function evaluated at a random data point $(\boldsymbol{\gamma}(n), \boldsymbol{h}_n)$. Several of the steps in the argument are similar and we will therefore be brief. Recall that the arguments in Sec. 13.2 dealt with general nonlinear optimization problems *without restricting* $P(w)$ to being convex; they established convergence towards a stationary point. We consider the same scenario.

We assume that a line search procedure is used at each iteration n to select parameters $\{\boldsymbol{\alpha}_n\}$ that satisfy the following variation of the Wolfe conditions for some $0 < \lambda < \eta < 1/2$ — compare with (13.101a)–(13.101b):

$$Q(\boldsymbol{w}_n; \boldsymbol{\gamma}, \boldsymbol{h}) \leq Q(\boldsymbol{w}_{n-1}; \boldsymbol{\gamma}_n, \boldsymbol{h}_n) + \lambda \boldsymbol{\alpha}_n \nabla_w Q(\boldsymbol{w}_{n-1}; \boldsymbol{\gamma}_n, \boldsymbol{h}_n) \boldsymbol{q}_n \quad \text{(19.103a)}$$

$$|\nabla_w Q(\boldsymbol{w}_n; \boldsymbol{\gamma}, \boldsymbol{h}) \boldsymbol{q}_n| \leq \eta \, |\nabla_w Q(\boldsymbol{w}_{n-1}; \boldsymbol{\gamma}_n, \boldsymbol{h}_n) \boldsymbol{q}_n| \quad \text{(19.103b)}$$

Here, the notation $(\boldsymbol{\gamma}(n), \boldsymbol{h}_n)$ refers to the data sample selected at iteration n, and $(\boldsymbol{\gamma}, \boldsymbol{h})$ denotes the random sample for iteration $n+1$. For simplicity, we drop the arguments $(\boldsymbol{\gamma}_n, \boldsymbol{h}_n)$ and $(\boldsymbol{\gamma}, \boldsymbol{h})$ from the loss functions and their gradients

and write $Q(\boldsymbol{w}_{n-1})$ and $Q(\boldsymbol{w}_n)$ instead:

$$Q(\boldsymbol{w}_n) \leq Q(\boldsymbol{w}_{n-1}) + \lambda\boldsymbol{\alpha}_n \nabla_w Q(\boldsymbol{w}_{n-1})\boldsymbol{q}_n \tag{19.104a}$$

$$|\nabla_w Q(\boldsymbol{w}_n)\boldsymbol{q}_n| \leq \eta\,|\nabla_w Q(\boldsymbol{w}_{n-1})\boldsymbol{q}_n| \tag{19.104b}$$

Stochastic Fletcher-Reeves algorithm for minimizing (19.3a)

given dataset $\{\gamma(m), h_m\}_{m=0}^{N-1}$;
start with an arbitrary initial condition $\boldsymbol{w}_{-1}$;
set $\boldsymbol{q}_{-1} = 0$;
repeat until convergence over $n \geq 0$:
  select at random $(\boldsymbol{\gamma}(n), \boldsymbol{h}_n)$;
  $\boldsymbol{r}_{n-1} = -\nabla_{w^{\mathsf{T}}} Q(\boldsymbol{w}_{n-1}; \boldsymbol{\gamma}(n), \boldsymbol{h}_n)$
  if $n = 0$ then $\boldsymbol{\beta}_{-1} = 0$
    else $\boldsymbol{\beta}_{n-1} = \|\boldsymbol{r}_{n-1}\|^2 / \|\boldsymbol{r}_{n-2}\|^2$
  end
  $\boldsymbol{q}_n = \boldsymbol{r}_{n-1} + \boldsymbol{\beta}_{n-1}\,\boldsymbol{q}_{n-1}$
  find $\boldsymbol{\alpha}_n$ using line search: $\min_{\alpha \in \mathbb{R}} Q(\boldsymbol{w}_{n-1} + \alpha \boldsymbol{q}_n)$
  $\boldsymbol{w}_n = \boldsymbol{w}_{n-1} + \boldsymbol{\alpha}_n \boldsymbol{q}_n$
end
return $w^\star \leftarrow \boldsymbol{w}_n$

(19.105)

Lemma 19.1. (Loss descent directions) *Assume the $\{\boldsymbol{\alpha}_n\}$ are selected to satisfy (19.104a)–(19.104b) for $0 < \lambda < \eta < 1/2$, then it holds for any $n \geq 0$ that*

$$-\frac{1}{1-\eta} \;\leq\; \frac{\nabla_w Q(\boldsymbol{w}_{n-1})\boldsymbol{q}_n}{\|\nabla_w Q(\boldsymbol{w}_{n-1})\|^2} \;\leq\; \frac{2\eta - 1}{1-\eta} < 0 \tag{19.106}$$

and, hence, the successive $\{\boldsymbol{q}_n\}$ generated by Fletcher-Reeves are descent directions relative to the loss function $Q(w;\cdot)$.

Proof: The argument is similar to the one used to establish Lemma 13.1. For $n = 0$, we have $\boldsymbol{q}_0 = -\nabla_w Q(\boldsymbol{w}_{-1})$ so that the ratio in the middle is equal to -1 and both sides of the inequality are satisfied. Now suppose the inequality holds for iteration n and let us verify that it holds for iteration $n+1$. It follows that q_n is a descent direction so that condition (19.104b) becomes

$$|\nabla_w Q(\boldsymbol{w}_n)\boldsymbol{q}_n| \;\leq\; -\eta \nabla_w Q(\boldsymbol{w}_{n-1})\boldsymbol{q}_n \tag{19.107}$$

which is equivalent to

$$\eta \nabla_w Q(\boldsymbol{w}_{n-1})\boldsymbol{q}_n \;\leq \nabla_w Q(\boldsymbol{w}_n)\boldsymbol{q}_n \;\leq\; -\eta \nabla_w Q(\boldsymbol{w}_{n-1})\boldsymbol{q}_n \tag{19.108}$$

Now note from recursions (19.105) that

$$\begin{aligned}\frac{\nabla_w Q(\boldsymbol{w}_n)\boldsymbol{q}_{n+1}}{\|\nabla_w Q(\boldsymbol{w}_n)\|^2} &= \frac{\nabla_w Q(\boldsymbol{w}_n)(\boldsymbol{r}_n+\boldsymbol{\beta}_n\boldsymbol{q}_n)}{\|\nabla_w Q(\boldsymbol{w}_n)\|^2}\\ &= -1+\boldsymbol{\beta}_n\frac{\nabla_w Q(\boldsymbol{w}_n)\boldsymbol{q}_n}{\|\nabla_w Q(\boldsymbol{w}_n)\|^2},\ \text{ since } \boldsymbol{r}_n=-\nabla_{w^{\mathsf{T}}}Q(\boldsymbol{w}_n)\\ &= -1+\frac{\|\boldsymbol{r}_n\|^2}{\|\boldsymbol{r}_{n-1}\|^2}\frac{\nabla_w Q(\boldsymbol{w}_n)\boldsymbol{q}_n}{\|\nabla_w Q(\boldsymbol{w}_n)\|^2}\\ &= -1+\frac{\nabla_w Q(\boldsymbol{w}_n)\boldsymbol{q}_n}{\|\nabla_w Q(\boldsymbol{w}_{n-1})\|^2}\end{aligned} \tag{19.109}$$

Using (19.108) we obtain

$$-1+\eta\frac{\nabla_w Q(\boldsymbol{w}_{n-1})\boldsymbol{q}_n}{\|\nabla_w Q(\boldsymbol{w}_{n-1})\|^2} \le -1+\frac{\nabla_w Q(\boldsymbol{w}_n)\boldsymbol{q}_n}{\|\nabla_w Q(\boldsymbol{w}_{n-1})\|^2} \le -1-\eta\frac{\nabla_w Q(\boldsymbol{w}_{n-1})\boldsymbol{q}_n}{\|\nabla_w Q(\boldsymbol{w}_{n-1})\|^2} \tag{19.110}$$

and applying the lower bound from (19.106) to the leftmost and rightmost terms we get

$$-1-\frac{\eta}{1-\eta} \le \frac{\nabla_w Q(\boldsymbol{w}_n)\boldsymbol{q}_{n+1}}{\|\nabla_w Q(\boldsymbol{w}_n)\|^2} \le -1+\frac{\eta}{1-\eta} \tag{19.111}$$

which establishes the validity of (19.106) for $n+1$.

■

The next result extends Zoutendijk condition to the stochastic case. Again, neither the loss nor the risk function are required to be convex in this statement.

Lemma 19.2. (Stochastic Zoutendijk condition) *Consider an empirical risk minimization problem of the form (19.3a), where $P(w)$ is bounded from below (but not necessarily convex) and the loss function $Q(w,\cdot)$ is first-order differentiable with $\delta-$Lipschitz gradients as in (19.7b). The data is sampled uniformly with replacement from $\{\gamma(m), h_m\}$. Assume the $\{\boldsymbol{\alpha}_n\}$ are selected to satisfy (19.104a)–(19.104b) for $0<\lambda<\eta<1/2$ so that the $\{\boldsymbol{q}_n\}$ generated by the stochastic Fletcher–Reeves procedure (19.105) are descent directions relative to $Q(w;\cdot)$ by Lemma 19.1. The iterate $\boldsymbol{w}_{n-1}$ is updated to $\boldsymbol{w}_n=\boldsymbol{w}_{n-1}+\boldsymbol{\alpha}_n\boldsymbol{q}_n$. Let $\boldsymbol{\theta}_n$ denote the angle defined by:*

$$\cos(\boldsymbol{\theta}_n) \triangleq \frac{-\nabla_w Q(\boldsymbol{w}_{n-1})\boldsymbol{q}_n}{\|\nabla_w Q(\boldsymbol{w}_{n-1})\|\,\|\boldsymbol{q}_n\|} \tag{19.112}$$

It then holds that

$$\sum_{n=0}^{\infty}\mathbb{E}\left(\cos^2(\boldsymbol{\theta}_n)\|\nabla_w Q(\boldsymbol{w}_{n-1})\|^2\right) < \infty \tag{19.113}$$

where the expectation is over the randomness of the data. Moreover, $\mathbb{E}\,P(\boldsymbol{w}_n)$ is non-increasing meaning that $\mathbb{E}\,P(\boldsymbol{w}_n) \le \mathbb{E}\,P(\boldsymbol{w}_{n-1})$.

Proof: The argument is similar to the one used to establish Lemma 13.2. Since $\boldsymbol{q}_n$ is

a descent direction relative to the loss function, we conclude from the second Wolfe condition (19.104b) or from (19.108) that

$$\nabla_w Q(\boldsymbol{w}_n)\boldsymbol{q}_n \;\geq\; \eta\,\nabla_w Q(\boldsymbol{w}_{n-1})\boldsymbol{q}_n \tag{19.114}$$

Subtracting $\nabla_w Q(\boldsymbol{w}_{n-1})\boldsymbol{q}_n$ from both sides gives

$$\Big(\nabla_w Q(\boldsymbol{w}_n) - \nabla_w Q(\boldsymbol{w}_{n-1})\Big)\boldsymbol{q}_n \;\geq\; (\eta-1)\nabla_w Q(\boldsymbol{w}_{n-1})\boldsymbol{q}_n \tag{19.115}$$

From the $\delta-$Lipschitz condition (19.7b) on the gradient of $Q(w;\cdot)$ we have by Cauchy-Schwarz:

$$\begin{aligned}\Big(\nabla_w Q(\boldsymbol{w}_n) - \nabla_w Q(\boldsymbol{w}_{n-1})\Big)\boldsymbol{q}_n &\leq \|\nabla_w Q(\boldsymbol{w}_n) - \nabla_w Q(\boldsymbol{w}_{n-1})\|\,\|\boldsymbol{q}_n\| \\ &\leq \delta\|\boldsymbol{w}_n - \boldsymbol{w}_{n-1}\|\,\|\boldsymbol{q}_n\| \\ &= \delta\boldsymbol{\alpha}_n\|\boldsymbol{q}_n\|^2, \quad \text{since } \boldsymbol{w}_n = \boldsymbol{w}_{n-1} + \boldsymbol{\alpha}_n\boldsymbol{q}_n\end{aligned} \tag{19.116}$$

Combining (19.115) and (19.116) shows that $\boldsymbol{\alpha}_n$ is lower-bounded by

$$\boldsymbol{\alpha}_n \geq \frac{(\eta-1)}{\delta}\,\frac{\nabla_w Q(\boldsymbol{w}_{n-1})\boldsymbol{q}_n}{\|\boldsymbol{q}_n\|^2} \tag{19.117}$$

where the term on the right-hand side is positive since $\boldsymbol{q}_n$ is a descent direction and $\eta < 1$. Substituting this conclusion into the first Wolfe condition (19.104a) we find that

$$Q(\boldsymbol{w}_n) \leq Q(\boldsymbol{w}_{n-1}) + \underbrace{\lambda\frac{(\eta-1)}{\delta}}_{\triangleq\, -c}\,\frac{(\nabla_w Q(\boldsymbol{w}_{n-1})\boldsymbol{q}_n)^2}{\|\boldsymbol{q}_n\|^2} \tag{19.118}$$

where we introduced the positive constant $c = \lambda(1-\eta)/\delta$. In other words, in terms of the angles $\{\boldsymbol{\theta}_n\}$ we have that

$$Q(\boldsymbol{w}_n) \leq Q(\boldsymbol{w}_{n-1}) - c\,\cos^2(\boldsymbol{\theta}_n)\,\|\nabla_w Q(\boldsymbol{w}_{n-1})\|^2 \tag{19.119}$$

Taking expectations over the randomness in the data we get

$$\mathbb{E}\,P(\boldsymbol{w}_n) \leq \mathbb{E}\,P(\boldsymbol{w}_{n-1}) - c\,\mathbb{E}\left(\cos^2(\boldsymbol{\theta}_n)\,\|\nabla_w Q(\boldsymbol{w}_{n-1})\|^2\right) \tag{19.120}$$

which shows that $\mathbb{E}\,P(\boldsymbol{w}_m)$ is non-increasing. Summing over n gives

$$\sum_{n=0}^{\infty}\mathbb{E}\left(\cos^2(\boldsymbol{\theta}_n)\,\|\nabla_w Q(\boldsymbol{w}_{m-1})\|^2\right) \;\leq\; \frac{1}{c}\Big(\mathbb{E}\,P(\boldsymbol{w}_{-1}) - \lim_{n\to\infty}\mathbb{E}\,P(\boldsymbol{w}_n)\Big) \tag{19.121}$$

Since, by assumption, the risk function is bounded from below, the term on the right-hand side is bounded by some positive constant and conclusion (19.113) holds.

■

One useful corollary of the previous two lemmas follows if we multiply (19.106) by $\|\nabla_w Q(\boldsymbol{w}_{n-1})\|/\|q_n\|$ and use (19.113) to conclude that the following condition must also hold:

$$\sum_{n=0}^{\infty}\mathbb{E}\left(\frac{\|\nabla_w Q(\boldsymbol{w}_{n-1})\|^4}{\|\boldsymbol{q}_n\|^2}\right) < \infty \tag{19.122}$$

Theorem 19.6. (Convergence of stochastic Fletcher-Reeves) *Consider the same setting of Lemma 19.2 but assume that the loss function has bounded gradients. Then, it holds that*

$$\liminf_{n\to\infty} \|\mathbb{E}\,\nabla_w P(\boldsymbol{w}_n)\| = 0 \tag{19.123}$$

This implies that there exists a subsequence of weight iterates over which the gradient of $P(w)$ converges on average to zero.

Proof: We denote the bound on the gradient of the loss function by

$$\|\nabla_w Q(w)\| \le c_1, \quad \text{for some } c_1 > 0 \tag{19.124}$$

Next, since $\boldsymbol{q}_m$ is a descent direction relative to the loss function, we conclude from the second Wolfe condition (19.104b) that

$$\nabla_w Q(\boldsymbol{w}_n)\boldsymbol{q}_n \;\le\; -\eta\nabla_w Q(\boldsymbol{w}_{n-1})\boldsymbol{q}_n \overset{(19.106)}{\le} \frac{\eta}{\eta-1}\|\nabla_w Q(\boldsymbol{w}_{n-1})\|^2 \tag{19.125}$$

It follows that

$$\begin{aligned}
\|\boldsymbol{q}_{n+1}\|^2 &= \|\boldsymbol{r}_n + \boldsymbol{\beta}_n\boldsymbol{q}_n\|^2 \\
&= \|\nabla_w Q(\boldsymbol{w}_n)\|^2 - 2\boldsymbol{\beta}_n\nabla_w Q(\boldsymbol{w}_n)\boldsymbol{q}_n + \boldsymbol{\beta}_n^2\|\boldsymbol{q}_n\|^2 \\
&= \|\nabla_w Q(\boldsymbol{w}_n)\|^2 - 2\frac{\|\nabla_w Q(\boldsymbol{w}_n)\|^2}{\|\nabla_w Q(\boldsymbol{w}_{n-1})\|^2}\nabla_w Q(\boldsymbol{w}_n)\boldsymbol{q}_n + \boldsymbol{\beta}_n^2\|\boldsymbol{q}_n\|^2 \\
&\overset{(19.125)}{\le} \|\nabla_w Q(\boldsymbol{w}_n)\|^2 + \frac{2\eta}{1-\eta}\|\nabla_w Q(\boldsymbol{w}_n)\|^2 + \boldsymbol{\beta}_n^2\|\boldsymbol{q}_n\|^2 \\
&= \underbrace{\frac{1+\eta}{1-\eta}}_{\triangleq\, c_2>1}\|\nabla_w Q(\boldsymbol{w}_n)\|^2 \;+\; \boldsymbol{\beta}_n^2\|\boldsymbol{q}_n\|^2
\end{aligned} \tag{19.126}$$

Iterating we get

$$\begin{aligned}
\|\boldsymbol{q}_{n+1}\|^2 &\le c_2\|\nabla_w Q(\boldsymbol{w}_n)\|^2 + c_2\frac{\|\nabla_w Q(\boldsymbol{w}_n)\|^4}{\|\nabla_w Q(\boldsymbol{w}_{n-1})\|^2} + c_2\frac{\|\nabla_w Q(\boldsymbol{w}_n)\|^4}{\|\nabla_w Q(\boldsymbol{w}_{n-2})\|^2} + \ldots \\
&= c_2\|\nabla_w Q(\boldsymbol{w}_n)\|^4\left\{\frac{1}{\|\nabla_w Q(\boldsymbol{w}_n)\|^2} + \frac{1}{\|\nabla_w Q(\boldsymbol{w}_{n-1})\|^2} + \ldots\right\} \\
&= c_2\|\nabla_w Q(\boldsymbol{w}_n)\|^4\sum_{j=0}^{n+1}\frac{1}{\|\nabla_w Q(\boldsymbol{w}_{j-1})\|^2}
\end{aligned} \tag{19.127}$$

and, hence,

$$\mathbb{E}\,\|\boldsymbol{q}_{n+1}\|^2 \le c_2 c_1^4\sum_{j=0}^{n+1}\mathbb{E}\left(\frac{1}{\|\nabla_w Q(\boldsymbol{w}_{j-1})\|^2}\right) \tag{19.128}$$

We can now deduce that

$$\liminf_{n\to\infty} \|\mathbb{E}\,\nabla_w Q(\boldsymbol{w}_n)\| = 0 \tag{19.129}$$

We establish its validity by contradiction. Assume the result does not hold. This means that $\mathbb{E}\,\nabla_w Q(\boldsymbol{w})$ is bounded from below for all n, say,

$$\|\mathbb{E}\,\nabla_w Q(\boldsymbol{w}_n)\| \ge c_3, \quad \text{for some } c_3 > 0 \text{ and for all } n > 0 \tag{19.130}$$

For any nonnegative random variable $\boldsymbol{x}$, we know from Jensen inequality that the following relations hold:

$$\mathbb{E}\,|\boldsymbol{x}| \geq |\mathbb{E}\,\boldsymbol{x}|, \quad \mathbb{E}\,\boldsymbol{x}^2 \;\geq\; (\mathbb{E}\,\boldsymbol{x})^2, \quad \text{and} \quad \mathbb{E}\,(1/\boldsymbol{x}) \geq 1/\mathbb{E}\,\boldsymbol{x} \tag{19.131}$$

We then conclude that

$$\mathbb{E}\,\|\nabla_w Q(\boldsymbol{w}_n)\| \geq c_3, \quad \mathbb{E}\,\|\nabla_w Q(\boldsymbol{w}_n)\|^2 \geq c_3^2, \quad \mathbb{E}\left(\frac{1}{\|\nabla_w Q(\boldsymbol{w}_n)\|^2}\right) \geq 1/c_3^2 \tag{19.132}$$

Substituting into (19.128), we find that

$$\mathbb{E}\,\|\boldsymbol{q}_{n+1}\|^2 \leq \frac{c_2 c_1^4}{c_3^2}(n+2) \tag{19.133}$$

This result in turn implies that the series $\{\mathbb{E}\,(1/\|\boldsymbol{q}_j\|^2)\}$ diverges since

$$\sum_{j=0}^{\infty} \mathbb{E}\left(\frac{1}{\|\boldsymbol{q}_j\|^2}\right) \geq \frac{c_3^2}{c_2 c_1^4} \sum_{j=0}^{\infty} \frac{1}{j+1} \tag{19.134}$$

However, this conclusion contradicts Zoutendijk condition (19.113) or its corollary (19.122) since it implies

$$\sum_{n=0}^{\infty} \mathbb{E}\left(\frac{\|\nabla_w Q(\boldsymbol{w}_{n-1})\|^4}{\|\boldsymbol{q}_n\|^2}\right) \geq c_3^4 \sum_{n=0}^{\infty}\left(\frac{1}{\|\boldsymbol{q}_n\|^2}\right) \;\geq\; \frac{c_3^6}{c_2 c_1^4} \sum_{j=0}^{\infty} \frac{1}{j+1} \tag{19.135}$$

which is not bounded. We conclude by contradiction that condition (19.129) is valid, from which (19.123) follows.

■

19.8 COMMENTARIES AND DISCUSSION

Stochastic gradient algorithms. There are extensive works in the literature on stochastic gradient algorithms and their convergence behavior, including by Albert and Gardner (1967), Wasan (1969), Mendel and Fu (1970), Tsypkin (1971), Ljung (1977), Kushner and Clark (1978), Kushner (1984), Polyak (1987), Benveniste, Métivier, and Priouret (1990), Bertsekas and Tsitsiklis (1997,2000), Bottou (1998,2010,2012), Kushner and Yin (2003), Spall (2003), Marti (2005), Sayed (2003,2008,2014a), Shalev-Shwartz and Ben-David (2014), and Bottou, Curtis, and Nocedal (2018). The proof of Theorem 19.1 for operation under constant step-sizes follows the argument from Sayed (2014a). The convergence result (19.158) in the appendix for a stochastic inequality recursion is from Polyak (1987, pp. 49–50). The result is useful in characterizing the convergence rates of stochastic approximation algorithms with diminishing step-sizes, as was shown in the proof of Theorem 19.3 following Polyak (1987). Some of the earlier studies on regret analysis for stochastic optimization algorithms are the works by Gordon (1999) and Zinkevich (2003) — see also the treatment by Shalev-Shwartz (2011) and the references therein. Analysis of the convergence behavior of the stochastic gradient algorithm under Polyak and Nesterov momentum acceleration schemes appear in Yuan, Ying, and Sayed (2016) using the general model described earlier in Prob. 17.12 and arguments similar to those employed in this chapter. The discussion in Example 19.2 motivating the choice $\mu(n) = O(1/n)$ for the decaying step-size sequence is in line with the analysis and conclusions from Robbins and Monro (1951) and Nemirovski *et al.* (2009). The presentation in Examples 19.3–19.4 is motivated by arguments from Bottou, Curtis, and Nocedal (2018). In Sec. 19.7 we examined the convergence of the *stochastic*

version of the Fletcher-Reeves algorithm by extending the derivation from Sec. 13.2 to the stochastic case (19.105). The analysis and proofs follow arguments similar to Zountendijk (1970), Powell (1984), Al-Baali (1985), and more closely the presentation from Nocedal and Wright (2006).

Mean-square deviation and excess risk measures. Theorem 19.1 characterizes the size of the limiting region for the mean-square error, namely, $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2 \rightarrow O(\mu)$ for sufficiently small μ. If desired, one can pursue a more detailed mean-square-error analysis and quantify *more accurately* the size of the constant multiplying μ in the $O(\mu)-$result. Consider a stochastic gradient implementation that is based on instantaneous gradient approximations and uniform sampling of the data. Let $w^\star$ denote the minimizer of the risk function $P(w)$, which can be an empirical or stochastic risk. Let $\boldsymbol{g}_n(w^\star)$ denote the gradient noise at location $w = w^\star$, i.e.,

$$\boldsymbol{g}_n(w^\star) \;=\; \widehat{\nabla_{w^\mathsf{T}} P}(w^\star) - \nabla_{w^\mathsf{T}} P(w^\star) \;=\; \nabla_{w^\mathsf{T}} Q(w^\star;\boldsymbol{\gamma}(n),\boldsymbol{h}_n) \tag{19.136}$$

and denote its steady-state covariance matrix, assumed to exist, by

$$R_g \;\stackrel{\Delta}{=}\; \lim_{n\rightarrow\infty} \mathbb{E}\,\boldsymbol{g}_n(w^\star)\boldsymbol{g}_n^\mathsf{T}(w^\star) \tag{19.137}$$

where the expectation is over the randomness in the data. The above expression assumes that the covariance matrix approaches a stationary value R_g. We know from (19.13b) that $\mathrm{Tr}(R_g) \leq \sigma_g^2$. Assume further that $P(w)$ is twice-differentiable and denote its Hessian matrix at $w = w^\star$ by

$$H \;\stackrel{\Delta}{=}\; \nabla_w^2 P(w^\star) \tag{19.138}$$

Since $P(w)$ is $\nu-$strongly convex, we know that $H \geq \nu I_M$. Then, it can be verified under (**A1**,**A2**) or (**A1**,**A2'**) and by exploiting the bound on the fourth-order moment of the gradient noise process established in Prob. 18.4, that — see the derivation in Sayed (2014a, Ch.4):

$$(\textbf{MSD}):\; \limsup_{n\rightarrow\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2 \;=\; \frac{\mu}{2}\mathrm{Tr}(H^{-1}R_g) + O(\mu^{3/2}) \tag{19.139a}$$

$$(\textbf{ER}):\; \limsup_{n\rightarrow\infty} \Big(\mathbb{E}\,P(\boldsymbol{w}_n) - P(w^\star)\Big) \;=\; \frac{\mu}{4}\mathrm{Tr}(R_g) + O(\mu^{3/2}) \tag{19.139b}$$

where the notation MSD and ER stands for "mean-square deviation" and "excess-risk," respectively. Extensions of these results to sampling *without* replacement appear in Table I of Ying *et al.* (2019). The above expressions are consistent with the bounds derived in the body of the chapter. For example, if we replace the upper bound $H^{-1} \leq \nu I_M$ into (19.139a) we find that $\mathrm{MSD} = O(\mu\sigma_g^2/2\nu)$, which is consistent with (19.16a). Likewise, using $\mathrm{Tr}(R_g) \leq \sigma_g^2$ in (19.139b) we find $\mathrm{ER} = O(\mu\sigma_g^2/4)$, which is consistent with (19.16b).

A simplified justification for (19.139a) is given further ahead under the remarks on the "long-term model." Consider, for illustration purposes, the following special case involving a quadratic stochastic risk:

$$P(w) = \mathbb{E}\,(\boldsymbol{\gamma} - \boldsymbol{h}^\mathsf{T}w)^2 \;=\; \sigma_\gamma^2 - 2r_{h\gamma}^\mathsf{T}w + w^\mathsf{T}R_h w \tag{19.140}$$

which we expanded in terms of the second-order moments $\sigma_\gamma^2 = \mathbb{E}\,\boldsymbol{\gamma}^2$, $r_{h\gamma} = \mathbb{E}\,\boldsymbol{h}\boldsymbol{\gamma}$, and $R_h = \mathbb{E}\,\boldsymbol{h}\boldsymbol{h}^\mathsf{T} > 0$. The random variables $\{\boldsymbol{\gamma},\boldsymbol{h}\}$ are assumed to have zero means. The Hessian matrix of $P(w)$ is $H = 2R_h$ for all w. If we differentiate $P(w)$ relative to w and set the gradient vector to zero, we find that the minimizer occurs at location

$$R_h w^\star = r_{h\gamma} \quad\Longleftrightarrow\quad w^\star = R_h^{-1}r_{h\gamma} \tag{19.141}$$

Assume the streaming data $\{\boldsymbol{\gamma}(n),\boldsymbol{h}_n\}$ arise from a linear regression model of the form $\boldsymbol{\gamma}(n) = \boldsymbol{h}_n^\mathsf{T}w^\bullet + \boldsymbol{v}(n)$, for some model $w^\bullet \in \mathbb{R}^M$, and where $\boldsymbol{h}_n$ and $\boldsymbol{v}(n)$ are zero-mean

uncorrelated processes. Moreover, $\boldsymbol{v}(n)$ is a white-noise process that is independent of all other random variables and has variance denoted by $\sigma_v^2 = \mathbb{E}\,\boldsymbol{v}^2(n)$. We showed in Prob. 18.9 that $w^\star = w^\bullet$, which means that the minimizer $w^\star$ is able to recover the underlying model $w^\bullet$ and, hence, it also holds that

$$\boldsymbol{v}(n) = \boldsymbol{\gamma}(n) - \boldsymbol{h}_n^{\mathsf{T}} w^\star \tag{19.142}$$

The gradient noise for instantaneous gradient approximations is given by

$$\boldsymbol{g}_n(w) = 2(r_{h\gamma} - R_h \boldsymbol{w}_{n-1}) - 2\boldsymbol{h}_n(\boldsymbol{\gamma}(n) - \boldsymbol{h}_n^{\mathsf{T}} w) \tag{19.143}$$

Evaluating at $w = w^\star$ gives

$$\boldsymbol{g}_n(w^\star) = -2\boldsymbol{h}_n(\boldsymbol{\gamma}(n) - \boldsymbol{h}_n^{\mathsf{T}} w^\star) = -2\boldsymbol{h}_n \boldsymbol{v}(n) \tag{19.144}$$

whose covariance matrix is

$$R_g \;\triangleq\; \mathbb{E}\,\boldsymbol{g}_n(w^\star)\boldsymbol{g}_n^{\mathsf{T}}(w^\star) \;=\; 4\sigma_v^2 R_h \tag{19.145}$$

Substituting into (19.139a)–(19.139b) we arrive at the famous expressions for the performance of the least-mean-squares (LMS) algorithm — see, e.g., Widrow and Stearns (1985), Haykin (2001), Sayed (2003,2008):

$$\mathrm{MSD}^{\mathrm{LMS}} \approx \mu M \sigma_v^2 \tag{19.146a}$$

$$\mathrm{ER}^{\mathrm{LMS}} \approx \mu \sigma_v^2 \mathrm{Tr}(R_h) \tag{19.146b}$$

Long-term model. Consider the stochastic gradient algorithm (19.1) and assume an implementation with an instantaneous gradient approximation (similar remarks will hold for the mini-batch version):

$$\begin{aligned} \boldsymbol{w}_n &= \boldsymbol{w}_{n-1} - \mu\,\nabla_{w^{\mathsf{T}}} Q(\boldsymbol{w}_{n-1}; \boldsymbol{\gamma}(n), \boldsymbol{h}_n)) \\ &= \boldsymbol{w}_{n-1} - \mu\,\nabla_{w^{\mathsf{T}}} P(\boldsymbol{w}_{n-1}) - \mu \boldsymbol{g}_n(\boldsymbol{w}_{n-1}) \end{aligned} \tag{19.147}$$

where the second equality is in terms of the true gradient vector and the gradient noise. We analyzed the convergence behavior of this recursion in the body of the chapter and discovered that $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2 \to O(\mu)$ for sufficiently small step-sizes. Recursion (19.147) describes a stochastic system, with its state vector $\boldsymbol{w}_n$ evolving randomly over time due to the randomness in the data samples and the resulting gradient noise. In many instances, it is useful to introduce an approximate model, with constant dynamics, that could serve as a good approximation for the evolution of the state vector for large time instants. We motivate this *long-term model* as follows — see Sayed (2014a,2014b).

Assume $P(w)$ is twice-differentiable and that its Hessian matrix is $\tau-$Lipschitz relative to the minimizer $w^\star$, meaning that

$$\|\nabla_w^2 P(w) - \nabla_w^2 P(w^\star)\| \;\leq\; \tau\,\|\widetilde{w}\|, \;\; \widetilde{w} = w^\star - w \tag{19.148}$$

Since $P(w)$ is $\nu-$strongly convex and its gradient vectors are $\delta-$Lipschitz, we also know that $\nu I_M \leq \nabla_w^2 P(w) \leq \delta I_M$. Using the mean-value theorem (10.8) we can write

$$\nabla_{w^{\mathsf{T}}} P(\boldsymbol{w}_{n-1}) = -\underbrace{\left(\int_0^1 \nabla_w^2 P(w^\star - t\widetilde{\boldsymbol{w}}_{n-1})dt\right)}_{\triangleq\, \boldsymbol{H}_{n-1}} \widetilde{\boldsymbol{w}}_{n-1} \tag{19.149}$$

where $\boldsymbol{H}_{n-1} \leq \delta I_M$ is a symmetric and random matrix changing with the time index n. Subtracting $w^\star$ from both sides of (19.147) and using (19.149) gives

$$\widetilde{\boldsymbol{w}}_n = (I_M - \mu \boldsymbol{H}_{n-1})\widetilde{\boldsymbol{w}}_{n-1} + \mu \boldsymbol{g}_n(\boldsymbol{w}_{n-1}) \tag{19.150}$$

This is a nonlinear stochastic recursion in the error vector. Let $H = \nabla_w^2 P(w^\star)$ denote the Hessian matrix at the minimizer and introduce the deviation relative to it

$$\widetilde{\boldsymbol{H}}_{n-1} \triangleq H - \boldsymbol{H}_{n-1} \tag{19.151}$$

Then, recursion (19.150) can be rewritten as

$$\boldsymbol{c}_{n-1} \triangleq \widetilde{\boldsymbol{H}}_{n-1}\widetilde{\boldsymbol{w}}_{n-1} \tag{19.152a}$$

$$\widetilde{\boldsymbol{w}}_n = (I_M - \mu H)\widetilde{\boldsymbol{w}}_{n-1} + \mu \boldsymbol{g}_n(\boldsymbol{w}_{n-1}) + \mu \boldsymbol{c}_{n-1} \tag{19.152b}$$

Using (19.148), we have that $\|\boldsymbol{c}_{n-1}\| \leq \tau \|\widetilde{\boldsymbol{w}}_{n-1}\|^2$. Now since $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2 \rightarrow O(\mu)$, we conclude that, for large n, the weight-error vector evolves according to the dynamics:

$$\boxed{\widetilde{\boldsymbol{w}}_n = (I_M - \mu H)\widetilde{\boldsymbol{w}}_{n-1} + \mu \boldsymbol{g}_n(\boldsymbol{w}_{n-1}) + O(\mu^2)} \tag{19.153}$$

Working with this long-term model is helpful because its dynamics is driven by the constant matrix H, as opposed to the random matrix $\boldsymbol{H}_{n-1}$. Also, the driving $O(\mu^2)$ term can be ignored for small enough μ. Using this model, we can justify the first term in the MSD expression (19.139a). Indeed, computing the weighted Euclidean norm of both sides of (19.153) using H^{-1} as the weighting matrix we get

$$\begin{aligned}
&\widetilde{\boldsymbol{w}}_n^{\mathsf{T}} H^{-1} \widetilde{\boldsymbol{w}}_n \qquad (19.154)\\
&= \widetilde{\boldsymbol{w}}_{n-1}^{\mathsf{T}}(I - \mu H)H^{-1}(I - \mu H)\widetilde{\boldsymbol{w}}_{n-1} + \mu^2 \boldsymbol{g}_n^{\mathsf{T}}(\boldsymbol{w}_{n-1})H^{-1}\boldsymbol{g}_n(\boldsymbol{w}_{n-1}) + \text{cross term}\\
&\approx \widetilde{\boldsymbol{w}}_{n-1}^{\mathsf{T}}(H^{-1} - 2\mu I_M)\widetilde{\boldsymbol{w}}_{n-1} + \mu^2 \boldsymbol{g}_n^{\mathsf{T}}(\boldsymbol{w}_{n-1})H^{-1}\boldsymbol{g}_n(\boldsymbol{w}_{n-1}) + \text{cross term}
\end{aligned}$$

where we are ignoring the term $\mu^2 \widetilde{\boldsymbol{w}}_{n-1}^{\mathsf{T}} H \widetilde{\boldsymbol{w}}_{n-1}$, which is on the order of μ^3 as $n \rightarrow \infty$. Under expectation, the cross-term is zero since $\mathbb{E}\,(\boldsymbol{g}_n(\boldsymbol{w}_{n-1})|\boldsymbol{w}_{n-1}) = 0$. Taking expectations and letting $n \rightarrow \infty$ we get

$$\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2 \rightarrow \frac{\mu}{2}\mathrm{Tr}(H^{-1}R_g) \tag{19.155}$$

as claimed.

Random reshuffling. We established in the body of the chapter (see, e.g., the summary in Table 19.2) that the performance of stochastic gradient algorithms differs under sampling *with* and *without* replacement. In the first case, the steady-state mean-square error, $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2$, approaches a neighborhood of size $(O\mu)$, while in the second case under random reshuffling the neighborhood size is reduced to $O(\mu^2)$, where μ is the small step-size parameter. This is a remarkable conclusion showing that the manner by which the same data points are processed by the algorithm can have a nontrivial effect on performance. It has been noted in several studies, e.g., by Bottou (2009), Recht and Re (2012), Gürbüzbalaban, Ozdaglar, and Parrilo (2015b), and Shamir (2016) that incorporating random reshuffling into the operation of a stochastic gradient algorithm helps improve performance. The last three works pursued justifications for the enhanced behavior of the algorithm by examining the convergence rate of the learning process under vanishing step-sizes. Some of the justifications rely on loose bounds or their conclusions are dependent on the sample size. Also, some of the results only establish that random reshuffling will not degrade performance relative to uniform sampling. In the body of the chapter, and specifically the arguments used in Appendix 19.B, we followed the approach by Ying *et al.* (2019). The contribution in this work provided a detailed analysis justifying analytically the improved performance from $O(\mu)$ to $O(\mu^2)$ under *constant* step-size operation.

Importance sampling The derivation of the optimal and adaptive sampling strategies in Sec. 19.6 follows the approach proposed by Yuan *et al.* (2016). There are of course other

sampling strategies in the literature. For example, in some works, condition (19.7b) on the loss function is stated instead in the form:

$$\|\nabla_w Q(w_2;\gamma(m),h_m) \;-\; \nabla_w Q(w_1;\gamma(m),h_m)\| \;\leq\; \delta_m \|w_2 - w_1\| \tag{19.156}$$

with a separate Lipschitz constant δ_m for each sample $m = 0, 1, \ldots, N-1$. One sampling strategy proposed by Needell, Ward, and Srebro (2014) and Zhao and Zhang (2015) measures the importance of each sample according to its Lipschitz constant and selects the assignment probabilities according to

$$p_m = \frac{\delta_m}{\sum_{m=0}^{N-1} \delta_m} \tag{19.157}$$

This construction is not the result of an optimized design and it requires knowledge of the Lipschitz constants, which are generally not available in advance. One feature of the adaptive sampling strategy described in (19.101) is that it relies solely on the available data.

PROBLEMS

19.1 Repeat the steps in the proof of Theorem 19.1 to establish Theorem 19.2 for the mini-batch stochastic gradient implementation.

19.2 Extend Theorem 19.4 for decaying step-sizes to the mini-batch stochastic gradient implementation.

19.3 Consider a stochastic gradient implementation with instantaneous gradient approximations. Assume an empirical risk minimization problem where the $N-$data points $\{\gamma(m), h_m\}$ are randomly reshuffled at the start of each run. Let $(\boldsymbol{\gamma}(n), \boldsymbol{h}_n)$ denote generically the sample that is selected at iteration n in the $k-$th run. Let $\boldsymbol{\sigma}(0:n-1)$ denote the history of all sample selections before the $n-$th iteration.

(a) Show that, conditioned on $\boldsymbol{w}_{n-1}$ and $\boldsymbol{\sigma}(0:n-1)$, it holds that

$$\mathbb{E}\left(\|\boldsymbol{g}(\boldsymbol{w}_{n-1})\|^2 \,|\, \boldsymbol{w}_{n-1}, \boldsymbol{\sigma}(0:n-1)\right) \;\leq$$
$$8\delta^2\|\widetilde{w}_{n-1}\|^2 + 2\,\mathbb{E}\left(\|\nabla_{w^{\mathsf{T}}} Q(w^\star;\boldsymbol{\gamma}(n),\boldsymbol{h}_n\|^2 \,|\, \boldsymbol{w}_{n-1}, \boldsymbol{\sigma}(0:n-1)\right)$$

(b) Conclude that $\mathbb{E}\left(\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2 \,|\, \boldsymbol{w}_{n-1}\right) \leq \beta_g^2\|\widetilde{w}_{n-1}\|^2 + \sigma_g^2$, where $\beta_g^2 = 8\delta^2$ and $\sigma_g^2 = \max_{0\leq m\leq N-1} \|\nabla_{w^{\mathsf{T}}} Q(w^\star;\gamma(m),h_m)\|^2$.

19.4 Refer to the intermediate result (19.23). Show that it also holds

$$\|\widetilde{\boldsymbol{w}}_{n-1} + \mu\,\nabla_{w^{\mathsf{T}}} P(\boldsymbol{w}_{n-1})\|^2 \leq \left(1 - \frac{\mu\nu}{2}\right)^2 \|\widetilde{\boldsymbol{w}}_{n-1}\|^2$$

19.5 Establish bound (19.33) on the average regret for the stochastic gradient algorithm with constant step-size.

19.6 How would the convergence rates shown in (19.55) change for step-size sequences of the form $\mu(n) = \tau/(n+1)^q$ for $\frac{1}{2} < q \leq 1$ and $\tau > 0$?

19.7 Extend the proof of Theorem 19.1 to the stochastic coordinate descent recursion (19.30) to derive conditions on the step-size for convergence. Assess the limiting behavior of the algorithm; its convergence rate and limiting mean-square-error performance.

19.8 Assume the search direction in a stochastic gradient implementation is scaled by a diagonal positive-definite matrix A as follows:

$$\boldsymbol{w}_n = \boldsymbol{w}_{n-1} - \mu\, A^{-1}\,\nabla_{w^{\mathsf{T}}} Q(\boldsymbol{w}_{n-1};\boldsymbol{\gamma}(n),\boldsymbol{h}_n), \quad n \geq 0$$

where $A \triangleq \mathrm{diag}\{a(1), a(2), \ldots, a(M)\}$, $0 < a(m) \leq 1$, and $\mu > 0$.

(a) Extend the result of Theorem 19.1 to this case.

(b) Extend the result of Theorem 19.3 to this case when μ is replaced by $\mu(n)$.

19.9 Establish result (19.81).

19.10 Refer to the stochastic gradient recursion (19.1) and assume that the step-size is a random parameter with mean $\mathbb{E}\,\boldsymbol{\mu} = \bar{\mu}$ and variance σ_μ^2. Assume $\boldsymbol{\mu}$ is independent of all other random variables. Follow the arguments used in the proof of Theorem 19.1 and show how the results of the theorem would need to be adjusted. *Remark.* For a related discussion, the reader may refer to Zhao and Sayed (2015a,b) and Sayed and Zhao (2018).

19.11 Refer to the stochastic gradient recursion (19.1) and assume that the step-size $\boldsymbol{\mu}$ is a Bernoulli random variable that is equal to μ with probability p and zero with probability $1-p$. That is, the recursion is active p fraction of the times. Assume $\boldsymbol{\mu}$ is independent of all other random variables. Follow the arguments used in the proof of Theorem 19.1 and show how the results of the theorem would need to be adjusted. *Remark.* For a related discussion, the reader may refer to Zhao and Sayed (2015a,b) and Sayed and Zhao (2018).

19.12 Assume $P(w)$ is only convex (but not necessarily strongly-convex) with a loss function whose gradients are $\delta-$Lipschitz satisfying (18.10b). Consider the stochastic-gradient recursion (19.1). Show that

$$\frac{1}{N}\sum_{n=0}^{N-1}\|\widetilde{\boldsymbol{w}}_{n-1}\|^2 \;\geq\; \frac{1}{N\mu^2\delta^2}\left(\|\widetilde{\boldsymbol{w}}_{N-1}\|^2 - \|\widetilde{\boldsymbol{w}}_{-1}\|^2\right)$$

How would the result change if $P(w)$ is $\nu-$strongly-convex?

19.13 This problem extends the result of Prob. 12.13 to the stochastic gradient scenario. Thus, refer to the stochastic gradient recursion (19.1) and assume $P(w)$ is only convex (but not necessarily strongly-convex) with a loss function whose gradients are $\delta-$Lipschitz satisfying (18.10b). Let $\mu < 1/\delta$.

(a) Use property (11.120) for convex functions with $\delta-$Lipschitz gradients to argue that the average risk value, $\mathbb{E}\,P(\boldsymbol{w}_n)$, increases by at most $O(\mu^2)$ per iteration. Specifically, verify that $\mathbb{E}\,P(\boldsymbol{w}_n) \leq \mathbb{E}\,P(\boldsymbol{w}_{n-1}) - \frac{\mu}{2}\mathbb{E}\,\|\nabla_w P(\boldsymbol{w}_{n-1})\|^2 + \frac{1}{2}\mu^2\delta\sigma_g^2$.

(b) Show that

$$\mathbb{E}\,P(\boldsymbol{w}_n) - P(w^\star) \leq \frac{1}{2\mu}\left(\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{n-1}\|^2 - \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2\right) + \mu\left(\beta_g^2\mathbb{E}\,\|\nabla_w P(\boldsymbol{w}_{n-1})\|^2 + \sigma_g^2\right)$$

(c) Conclude that $\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\,P(\boldsymbol{w}_k) - P(w^\star) \leq O(1/n) + O(\mu)$.

19.14 Refer to the stochastic gradient algorithm (19.75) under random reshuffling and assume an epoch-dependent step-size $\mu(k) = \tau/k$, for $k \geq 1$, is used. Repeat the arguments in Appendix 19.B and the technique used to derive Lemma 19.4 to establish the convergence rates (19.81)–(19.82).

19.15 The proof technique used to establish the convergence properties in Theorem 19.1 exploits the fact that the gradient noise process has zero mean conditioned on the past iterate $\boldsymbol{w}_{n-1}$. Motivated by the arguments used in Appendix 19.B, assume we follow now a similar proof technique to avoid the reliance on the zero-mean property for the gradient noise.

(a) Let $0 < t < 1$ be any scalar that we are free to choose. Subtract $w^\star$ from both sides of (19.1) and establish the result

$$\|\widetilde{\boldsymbol{w}}_n\|^2 \leq \frac{1}{t}\left(1 - 2\mu\nu + \mu^2\delta^2\right)\|\widetilde{\boldsymbol{w}}_{n-1}\|^2 \;+\; \frac{\mu^2}{1-t}\,\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2$$

(b) Verify that $1 - 2\mu\nu + \mu^2\delta^2 \leq \left(1 - \frac{\mu\nu}{2}\right)^2$ for $\mu < \nu/\delta^2$. Select $t = 1 - \frac{\mu\nu}{2}$ and show that $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2 \leq \lambda\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{n-1})\|^2 + 2\mu\sigma_g^2/\nu$, where $\lambda = 1 - \mu(\frac{\nu}{2} - \frac{2\beta_g^2}{\nu})$. Does $\lambda \in (0,1)$?

(c) Are you able to conclude from the recursion in part (b) that the mean-square deviation $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2$ approaches a neighborhood of size $O(\mu)$?

19.16 Show that the convergence rates in Lemma 19.4 continue to hold for the mini-batch stochastic gradient implementation.

19.17 Consider an empirical risk minimization problem and apply a stochastic gradient algorithm with importance sampling using either (19.6a) or (19.6b) to approximate the gradient direction. Extend the proof of Theorem 19.1 to show that the limiting mean-square-error region will continue to be $O(\mu\sigma_g^2)$.

19.18 Probs. 19.18–19.20 are motivated by the discussion in Bottou, Curtis, and Nocedal (2018). Refer to the stochastic-gradient recursion (19.1) and assume the risk function $P(w)$ has $\delta-$Lipschitz gradients as in (19.8) or (19.10). Use property (10.13) for $\delta-$smooth functions to establish the following inequality regardless of how the stochastic gradient is constructed:

$$\begin{aligned}&\mathbb{E}\left(P(\boldsymbol{w}_n)|\boldsymbol{w}_{n-1}\right) - P(w_{n-1})\\ &\leq -\mu\Big(\nabla_{w^{\mathsf{T}}}P(w_{n-1})\Big)^{\mathsf{T}}\,\mathbb{E}\,\widehat{\nabla_{w^{\mathsf{T}}}P}(w_{n-1}) + \frac{\mu^2\delta}{2}\mathbb{E}\,\|\widehat{\nabla_{w^{\mathsf{T}}}P}(w_{n-1})\|^2\end{aligned}$$

where the expectation operator $\mathbb{E}$ is over the statistical distribution of the data $\{\boldsymbol{\gamma},\boldsymbol{h}\}$ conditioned on the past iterate $\boldsymbol{w}_{n-1}$. Conclude that if the gradient approximation is unbiased then

$$\mathbb{E}\left(P(\boldsymbol{w}_n)|\boldsymbol{w}_{n-1}\right) - P(w_{n-1}) \leq -\mu\|\nabla_{w^{\mathsf{T}}}P(w_{n-1})\|^2 + \frac{\mu^2\delta}{2}\mathbb{E}\,\|\widehat{\nabla_{w^{\mathsf{T}}}P}(w_{n-1})\|^2$$

19.19 Continuing with Prob. 19.18, assume the stochastic gradient approximation satisfies the following three conditions in terms of the squared Euclidean norm:

$$\begin{aligned}&i)\ \Big(\nabla_{w^{\mathsf{T}}}P(w_{n-1})\Big)^{\mathsf{T}}\,\mathbb{E}\,\widehat{\nabla_{w^{\mathsf{T}}}P}(w_{n-1}) \geq a\,\|\nabla_{w^{\mathsf{T}}}P(w_{n-1})\|^2\\ &ii)\ \|\mathbb{E}\,\widehat{\nabla_{w^{\mathsf{T}}}P}(w_{n-1})\| \leq b\,\|\nabla_{w^{\mathsf{T}}}P(w_{n-1})\|\\ &iii)\ \mathrm{var}\Big(\widehat{\nabla_{w^{\mathsf{T}}}P}(w_{n-1})\Big) \leq \alpha + \beta\|\nabla_{w^{\mathsf{T}}}P(w_{n-1})\|^2\end{aligned}$$

for some constants $b \geq a > 0$ and $\alpha, \beta \geq 0$ and where, by definition,

$$\mathrm{var}\Big(\widehat{\nabla_{w^{\mathsf{T}}}P}(w_{n-1})\Big) \;\triangleq\; \mathbb{E}\,\|\widehat{\nabla_{w^{\mathsf{T}}}P}(w_{n-1})\|^2 - \|\mathbb{E}\,\widehat{\nabla_{w^{\mathsf{T}}}P}(w_{n-1})\|^2$$

(a) Let $\beta_1 = \beta + b^2$. Verify that

$$\mathbb{E}\,\|\widehat{\nabla_{w^{\mathsf{T}}}P}(w_{n-1})\|^2 \leq \alpha + \beta_1\|\nabla_{w^{\mathsf{T}}}P(w_{n-1})\|^2$$

(b) Conclude that

$$\mathbb{E}\left(P(\boldsymbol{w}_n)|\boldsymbol{w}_{n-1}\right) - P(w_{n-1}) \leq -\Big(a - \frac{1}{2}\delta\mu\beta_1\Big)\mu\|\nabla_{w^{\mathsf{T}}}P(w_{n-1})\|^2 + \frac{\mu^2}{2}\alpha\delta$$

19.20 Continuing with Prob. 19.19, assume now that $P(w)$ is a $\nu-$strongly convex risk that is bounded from below. Verify that for $\mu \leq a/\delta\beta_1$ we have

$$\mathbb{E}\,P(\boldsymbol{w}_n) - P(w^\star) \;\leq\; \frac{\mu\delta\alpha}{2\nu a} + (1-\mu\nu\alpha)^{n+1} \times \Big(P(w_{-1}) - P(w^\star) - \frac{\mu\delta\alpha}{2\nu a}\Big)$$

and conclude that

$$\limsup_{n\to\infty}\Big(\mathbb{E}\,P(\boldsymbol{w}_n) - P(w^\star)\Big) \leq \frac{\mu\delta\alpha}{2\nu a} \;=\; O(\mu)$$

19.21 The stochastic gradient algorithm can be implemented with Polyak-Ruppert averaging as shown earlier in (16.52), i.e.,

$$\begin{cases} \boldsymbol{w}_n = \boldsymbol{w}_{n-1} - \mu\,\nabla_{w^{\mathsf{T}}}\,Q(\boldsymbol{w}_{n-1};\boldsymbol{\gamma}(n),\boldsymbol{h}_n), \quad n \geq 0 \\ \bar{\boldsymbol{w}}_n = \bar{\boldsymbol{w}}_{n-1} + \dfrac{1}{n+2}(\boldsymbol{w}_n - \bar{\boldsymbol{w}}_{n-1}) \end{cases}$$

Extend the result of Theorem 19.1 to this case. *Remark.* For more discussion on this technique, the reader may refer to Ruppert (1988) and Polyak and Juditsky (1992).

19.22 A variation of the Polyak-Ruppert averaging algorithm of Prob. 19.21 is to generate $\bar{\boldsymbol{w}}_n$ by means of a convex combination, say

$$\begin{aligned} \boldsymbol{w}_n &= \boldsymbol{w}_{n-1} - \mu\,\nabla_{\boldsymbol{w}^{\mathsf{T}}}\,Q(\boldsymbol{w}_{n-1};\boldsymbol{\gamma}(n),\boldsymbol{h}_n) \\ \bar{\boldsymbol{w}}_n &= \beta\bar{\boldsymbol{w}}_{n-1} + (1-\beta)\boldsymbol{w}_n, \;\; \bar{\boldsymbol{w}}_{-1} = \boldsymbol{w}_{-1} = 0 \end{aligned}$$

where $\beta \in [0,1]$. Extend the result of Theorem 19.1 to this case.

19.23 Refer to the stochastic Nesterov momentum method (17.73) and examine its convergence properties. *Remark.* For a related discussion, refer to Yu, Jin, and Yang (2019).

19.24 In this problem we seek to re-derive the AdaGrad algorithm (17.13) by relying on the same mean-square-error analysis used in the proof of Theorem 19.1. Thus, consider a stochastic-gradient recursion of the form

$$\boldsymbol{w}_n = \boldsymbol{w}_{n-1} \;-\; \mu A^{-1}\nabla_{w^{\mathsf{T}}}Q(\boldsymbol{w}_{n-1};\boldsymbol{\gamma}(n),\boldsymbol{h}_n)$$

with a constant step-size, μ and a scaling symmetric and positive-definite matrix A^{-1}. Let $\sigma_{\max}(A^{-1})$ denote the maximum singular value of A. If A is restricted to being diagonal with positive entries, then $\sigma_{\max}(A^{-1}) = 1/a_{\min}$ where $a_{\min}$ is the smallest entry in A. Introduce the gradient noise vector

$$\boldsymbol{g}(w) \;\triangleq\; A^{-1}\widehat{\nabla_{w^{\mathsf{T}}}P}(w) \;-\; A^{-1}\nabla_{w^{\mathsf{T}}}\,P(w)$$

(a) Verify that under uniform sampling:

$$\begin{aligned} &\mathbb{E}\left(\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\,|\,\boldsymbol{w}_{n-1}\right) = 0, \;\; \mathbb{E}\left(\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|_A^2\,|\,\boldsymbol{w}_{n-1}\right) \leq \beta_g^2\|\widetilde{w}_{n-1}\|_A^2 + \sigma_g^2 \\ &\beta_g^2 = 8\delta^2/\sigma_{\min}^2(A) \\ &\sigma_g^2 = \frac{2}{N}\sum_{m=0}^{N-1}\|\nabla_{w^{\mathsf{T}}}\,Q(w^\star;\gamma(m),h_m)\|_{A^{-1}}^2 \end{aligned}$$

(b) Repeat the argument leading to (19.24) and verify that the relation now becomes

$$\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|_A^2 \leq \Big(1 - \frac{2\mu\nu}{\sigma_{\max}(A)} + \frac{9\mu^2\delta^2}{\sigma_{\min}^2(A)}\Big)\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{n-1}\|_A^2 + \mu^2\sigma_g^2$$

(c) Argue that selecting A to minimize σ_g^2 under the condition $\mathrm{Tr}(A) \leq c$, leads to the same optimization problem (17.35) obtained from the regret analysis.

19.25 Refer to the stochastic Fletcher-Reeves algorithm (19.105). Assume the parameters $\boldsymbol{\beta}_n$ are bounded for all n, say, $\boldsymbol{\beta}_n \leq \beta$ for some $\beta > 0$.

(a) Use an argument similar to (13.124) to show that

$$\mathbb{E}\,\|\boldsymbol{q}_{n+1}\|^2 \leq \beta^2\mathbb{E}\,\|\boldsymbol{q}_n\|^2 + \frac{(1+\eta)}{1-\eta}\mathbb{E}\,\|\nabla_w P(\boldsymbol{w}_{n-1})\|^2$$

(b) Let $c = (1+\eta)/(1-\eta)$. Iterate part (a) to conclude that

$$\mathbb{E}\,\|\boldsymbol{q}_{n+1}\|^2 \leq c\beta^{n+1}\left(\frac{1-\beta^{n+3}}{1-\beta}\right)\mathbb{E}\,\|\nabla_w P(\boldsymbol{w}_{-1})\|^2$$

(c) Assume the $\{\boldsymbol{\alpha}_m\}$ are limited to the bounded interval $\boldsymbol{\alpha}_m = (\alpha_\ell, \alpha_u)$ where $0 < \alpha_\ell < \alpha_u$. Assume each loss term $Q(w;\cdot)$ is $\nu-$strongly convex and has $\delta-$Lipschitz gradients. Show that the average excess risk evolves according to

$$\mathbb{E}\, P(\boldsymbol{w}_n) - P(w^\star) \;\leq\; \rho^n \Big(\mathbb{E}\, P(\boldsymbol{w}_{n-1}) - P(w^\star\Big)$$

for some positive factor $\rho < 1$.

Remark. The reader may refer to Jin *et al.* (2019) for a related discussion.

19.26 Refer to the stochastic Fletcher-Reeves algorithm (19.105). Assume the parameters $\{\boldsymbol{\alpha}_n\}$ are generated as follows:

$$\boldsymbol{\alpha}_n = -\rho \times \frac{\nabla_w Q(\boldsymbol{w}_{n-1})\boldsymbol{q}_n}{\boldsymbol{q}_n^{\mathsf{T}}\Sigma_n \boldsymbol{q}_n}$$

where $\rho \in (0, \nu_{\min}/\delta)$ and Σ_n is a given deterministic sequence of matrices satisfying $\nu_{\min}\|x\|^2 \leq x^{\mathsf{T}}\Sigma_n x \leq \nu_{\max}\|x\|^2$ for any x and where $\nu_{\min}$ and $\nu_{\max}$ are positive. Assume $P(w)$ is $\nu-$strongly convex with $\delta-$Lipschitz gradients. Establish that $\liminf_{n\to\infty} \|\mathbb{E}\,\nabla_w P(\boldsymbol{w}_n)\| = 0$. *Remark.* See the work by Sun and Zhang (2001) for a related discussion in the non-stochastic case.

19.A STOCHASTIC INEQUALITY RECURSION

The following useful result from Polyak (1987, p.49) is originally from Gladyshev (1965) and deals with the convergence of stochastic inequality recursions; it is the stochastic analogue of the earlier deterministic recursion (14.136).

Lemma 19.1. (Stochastic recursion) *Let $\boldsymbol{u}(n) \geq 0$ denote a scalar sequence of non-negative random variables satisfying $\mathbb{E}\,\boldsymbol{u}(0) < \infty$ and consider the stochastic recursion:*

$$\mathbb{E}\left(\boldsymbol{u}(n+1)|\,\boldsymbol{u}(0),\boldsymbol{u}(1),\ldots,\boldsymbol{u}(n)\,\right) \;\leq\; (1-a(n))\boldsymbol{u}(n) \;+\; b(n), \quad n \geq 0 \tag{19.158}$$

where the scalar deterministic sequences $\{a(n), b(n)\}$ satisfy the five conditions:

$$0 \leq a(n) < 1, \quad b(n) \geq 0, \quad \sum_{n=0}^{\infty} a(n) = \infty, \quad \sum_{n=0}^{\infty} b(n) < \infty, \quad \lim_{n\to\infty} \frac{b(n)}{a(n)} = 0 \tag{19.159}$$

Then, it holds that

$$\lim_{n\to\infty} \boldsymbol{u}(n) = 0, \quad \textit{almost surely} \tag{19.160a}$$

$$\lim_{n\to\infty} \mathbb{E}\,\boldsymbol{u}(n) = 0 \tag{19.160b}$$

Proof: For completeness, we establish the lemma by following the same argument from Polyak (1987, pp. 49–50). First, observe by taking expectations of both sides of (19.158) that the recursion reduces to the same form covered by Lemma 14.1, namely,

$$\mathbb{E}\,\boldsymbol{u}(n+1) \leq (1-a(n))\mathbb{E}\,\boldsymbol{u}(n) \;+\; b(n) \tag{19.161}$$

and, therefore, $\mathbb{E}\,\boldsymbol{u}(n) \to 0$ as $n \to \infty$. Next, introduce the auxiliary variable:

$$\boldsymbol{s}(n) \;\triangleq\; \boldsymbol{u}(n) \;+\; \sum_{j=n}^{\infty} b(j) \tag{19.162}$$

We know from the conditions $0 \leq a(n) < 1$ and $b(n) \geq 0$ that $\boldsymbol{s}(n) \geq 0$. Moreover, we also get $\mathbb{E}\,\boldsymbol{s}(0) < \infty$ since

$$\mathbb{E}\,\boldsymbol{s}(0) = \mathbb{E}\,\boldsymbol{u}(0) \;+\; \sum_{j=0}^{\infty} b(j) \;< \infty \tag{19.163}$$

Computing the conditional expectation of $\boldsymbol{s}(n+1)$ relative to $\{\boldsymbol{s}(0), \boldsymbol{s}(1), \ldots, \boldsymbol{s}(n)\}$ we get

$$\begin{aligned}
&\mathbb{E}\left(\boldsymbol{s}(n+1)|\,\boldsymbol{s}(0), \boldsymbol{s}(1), \ldots, \boldsymbol{s}(n)\right) \\
&= \mathbb{E}\left(\boldsymbol{u}(n+1)|\boldsymbol{u}(0), \boldsymbol{u}(1), \ldots, \boldsymbol{u}(n)\right) \;+\; \sum_{j=n+1}^{\infty} b(j) \\
&\leq (1 - a(n))\boldsymbol{u}(n) + b(n) \;+\; \sum_{j=n+1}^{\infty} b(j) \\
&= (1 - a(n))\boldsymbol{u}(n) \;+\; \sum_{j=n}^{\infty} b(j) \\
&\leq \boldsymbol{u}(n) \;+\; \sum_{j=n}^{\infty} b(j) \\
&= \boldsymbol{s}(n)
\end{aligned} \tag{19.164}$$

In other words, we established that

$$\mathbb{E}\left(\boldsymbol{s}(n+1)|\,\boldsymbol{s}(0), \boldsymbol{s}(1), \ldots, \boldsymbol{s}(n)\right) \leq \boldsymbol{s}(n) \tag{19.165}$$

This property means that $\boldsymbol{s}(n) \geq 0$ is a semi-martingale process, which also satisfies $\mathbb{E}\,\boldsymbol{s}(0) < \infty$. For such processes, it is known that there exists a random variable $\boldsymbol{s} \geq 0$ such that $\boldsymbol{s}(n) \to \boldsymbol{s}$ almost surely (see, e.g., Lipster and Shiryayev (1989), Williams (1991), and He, Wang, and Yan (1992)). Now note that, by construction,

$$\boldsymbol{u}(n) = \boldsymbol{s}(n) \;-\; \sum_{j=n}^{\infty} b(j) \tag{19.166}$$

so that, as $n \to \infty$,

$$\begin{aligned}
\mathbb{P}\left(\lim_{n\to\infty} \boldsymbol{u}(n) = \boldsymbol{s}\right) &= \mathbb{P}\left(\lim_{n\to\infty} \boldsymbol{s}(n) \;-\; \sum_{j=n}^{\infty} b(j) = \boldsymbol{s}\right) \\
&= \mathbb{P}\left(\lim_{n\to\infty} \boldsymbol{s}(n) = \boldsymbol{s}\right) \\
&= 1
\end{aligned} \tag{19.167}$$

and we conclude that $\boldsymbol{u}(n)$ also tends almost surely to $\boldsymbol{s} \geq 0$. We showed earlier that $\mathbb{E}\,\boldsymbol{u}(n) \to 0$ as $n \to \infty$. It follows that $\boldsymbol{s} = 0$ so that $\boldsymbol{u}(n)$ converges in probability to zero.

■

19.B PROOF OF THEOREM 19.5

In this appendix we follow the derivation from Ying *et al.* (2019) to establish the performance results (19.79a)–(19.79b) for operation under random reshuffling.

To begin with, note that recursion (19.75) shows how to move from one iterate to another within the same run k. The argument below will deduce from this recursion a similar relation that shows how to move from the initial iterate $\boldsymbol{w}_{-1}^{k-1}$ for run $k-1$ to the initial iterate $\boldsymbol{w}_{-1}^{k}$ for run k. That is, we first transform the description of the algorithm from iterations within the same run to iterations across epochs. Doing so will enable us to exploit a useful property of the random reshuffling mechanism, as explained below in (19.170). Once this new recursion across epochs is derived, we will then use it to establish (19.79a)–(19.79b).

Proof: Subtracting $w^\star$ from both sides of (19.75) gives

$$\widetilde{\boldsymbol{w}}_n^k = \widetilde{\boldsymbol{w}}_{n-1}^k + \mu \nabla_{w^{\mathsf{T}}} Q(\boldsymbol{w}_{n-1}^k; \boldsymbol{\gamma}(n), \boldsymbol{h}_n) \tag{19.168}$$

where the notation $(\boldsymbol{\gamma}(n), \boldsymbol{h}_n)$ denotes the random sample that is selected at iteration n of the $k-$th epoch. Iterating gives, where we are now dropping the data samples as arguments for $Q(w;\cdot,\cdot)$ for simplicity (we will restore them when necessary):

$$\begin{aligned}
\widetilde{\boldsymbol{w}}_{-1}^{k+1} &\triangleq \widetilde{\boldsymbol{w}}_{N-1}^k \\
&= \widetilde{\boldsymbol{w}}_{-1}^k + \mu \sum_{n=0}^{N-1} \nabla_{w^{\mathsf{T}}} Q(\boldsymbol{w}_{n-1}^k) \\
&\overset{(a)}{=} \widetilde{\boldsymbol{w}}_{-1}^k + \mu \sum_{n=0}^{N-1} \nabla_{w^{\mathsf{T}}} Q(\boldsymbol{w}_{n-1}^k) + \mu \sum_{n=0}^{N-1} \nabla_{w^{\mathsf{T}}} Q(\boldsymbol{w}_{-1}^k) - \mu \sum_{n=0}^{N-1} \nabla_{w^{\mathsf{T}}} Q(\boldsymbol{w}_{-1}^k) \\
&\overset{(b)}{=} \widetilde{\boldsymbol{w}}_{-1}^k + \mu N \nabla_{w^{\mathsf{T}}} P(\boldsymbol{w}_{-1}^k) + \mu \sum_{n=0}^{N-1} \Big(\nabla_{w^{\mathsf{T}}} Q(\boldsymbol{w}_{n-1}^k) - \nabla_{w^{\mathsf{T}}} Q(\boldsymbol{w}_{-1}^k) \Big)
\end{aligned} \tag{19.169}$$

where in step (a) we added and subtracted the same quantity, and in step (b) we used the fact that under random reshuffling:

$$\frac{1}{N} \sum_{n=0}^{N-1} Q(\boldsymbol{w}_{-1}^k; \boldsymbol{\gamma}(n), \boldsymbol{h}_n) = \frac{1}{N} \sum_{m=0}^{N-1} Q(\boldsymbol{w}_{-1}^k; \gamma(m), h_m) \;=\; P(\boldsymbol{w}_{-1}^k) \tag{19.170}$$

The first equality in (19.170) is because each data pair is sampled once under random reshuffling. Observe that this property would not hold under uniform sampling *with* replacement.

Now, let $0 < t < 1$ be any scalar that we are free to choose. Continuing with (19.169),

we square both sides and note that

$$\begin{aligned}
\|\widetilde{\boldsymbol{w}}_{-1}^{k+1}\|^2 &= \bigg\| \frac{t}{t}\Big(\widetilde{\boldsymbol{w}}_{-1}^{k} + \mu N \nabla_{w^{\mathsf{T}}} P(\boldsymbol{w}_{-1}^{k})\Big) \;+ \\
&\qquad \frac{1-t}{1-t}\,\mu \sum_{n=0}^{N-1}\Big(\nabla_{w^{\mathsf{T}}} Q(\boldsymbol{w}_{n-1}^{k}) - \nabla_{w^{\mathsf{T}}} Q(\boldsymbol{w}_{-1}^{k})\Big)\bigg\|^2 \\
&\overset{(a)}{\leq} t\left\|\frac{1}{t}\Big(\widetilde{\boldsymbol{w}}_{-1}^{k} + \mu N \nabla_{w^{\mathsf{T}}} P(\boldsymbol{w}_{-1}^{k})\Big)\right\|^2 \;+ \\
&\qquad (1-t)\left\|\frac{\mu}{1-t}\sum_{n=0}^{N-1}\Big(\nabla_{w^{\mathsf{T}}} Q(\boldsymbol{w}_{n-1}^{k}) - \nabla_{w^{\mathsf{T}}} Q(\boldsymbol{w}_{-1}^{k})\Big)\right\|^2 \\
&= \frac{1}{t}\left\|\widetilde{\boldsymbol{w}}_{-1}^{k} + \mu N \nabla_{w^{\mathsf{T}}} P(\boldsymbol{w}_{-1}^{k})\right\|^2 \;+ \\
&\qquad \frac{\mu^2}{1-t}\left\|\sum_{n=0}^{N-1}\Big(\nabla_{w^{\mathsf{T}}} Q(\boldsymbol{w}_{n-1}^{k}) - \nabla_{w^{\mathsf{T}}} Q(\boldsymbol{w}_{-1}^{k})\Big)\right\|^2 \\
&\overset{(b)}{\leq} \frac{1}{t}\left\|\widetilde{\boldsymbol{w}}_{-1}^{k} + \mu N \nabla_{w^{\mathsf{T}}} P(\boldsymbol{w}_{-1}^{k})\right\|^2 \;+ \\
&\qquad \frac{\mu^2 N}{1-t}\sum_{n=0}^{N-1}\left\|\nabla_{w^{\mathsf{T}}} Q(\boldsymbol{w}_{n-1}^{k}) - \nabla_{w^{\mathsf{T}}} Q(\boldsymbol{w}_{-1}^{k})\right\|^2
\end{aligned} \tag{19.171}$$

where step (a) uses Jensen inequality (8.76) and step (b) uses the same inequality again to justify the following property for any vectors $\{x_n\}$:

$$\left\|\sum_{n=1}^{N-1} x_n\right\|^2 = N^2\left\|\sum_{n=1}^{N-1}\frac{1}{N}x_n\right\|^2 \overset{(8.76)}{\leq} N\sum_{n=1}^{N-1}\|x_n\|^2 \tag{19.172}$$

Let us now examine the two terms on the right-hand side of (19.171). First note that

$$\begin{aligned}
&\left\|\widetilde{\boldsymbol{w}}_{-1}^{k} + \mu N \nabla_{w^{\mathsf{T}}} P(\boldsymbol{w}_{-1}^{k})\right\|^2 \\
&= \|\widetilde{\boldsymbol{w}}_{-1}^{k}\|^2 + 2\mu N\left(\nabla_{w^{\mathsf{T}}} P(\boldsymbol{w}_{-1}^{k})\right)^{\mathsf{T}}\widetilde{\boldsymbol{w}}_{-1}^{k} + \mu^2 N^2\|\nabla_{w^{\mathsf{T}}} P(\boldsymbol{w}_{-1}^{k})\|^2 \\
&= \|\widetilde{\boldsymbol{w}}_{-1}^{k}\|^2 + 2\mu N\left(\nabla_{w^{\mathsf{T}}} P(\boldsymbol{w}_{-1}^{k})\right)^{\mathsf{T}}\widetilde{\boldsymbol{w}}_{-1}^{k} + \mu^2 N^2\|\underbrace{\nabla_{w^{\mathsf{T}}} P(w^{\star})}_{=0} - \nabla_{w^{\mathsf{T}}} P(\boldsymbol{w}_{-1}^{k})\|^2 \\
&\overset{\textbf{(P2)}}{\leq} \|\widetilde{\boldsymbol{w}}_{-1}^{k}\|^2 + 2\mu N\left(\nabla_{w^{\mathsf{T}}} P(\boldsymbol{w}_{-1}^{k})\right)^{\mathsf{T}}\widetilde{\boldsymbol{w}}_{-1}^{k} + \mu^2 N^2\delta^2\|\widetilde{\boldsymbol{w}}_{-1}^{k}\|^2
\end{aligned} \tag{19.173}$$

Next, we appeal to the strong convexity property (18.10a) to find that

$$\begin{aligned}
\left(\nabla_{w^{\mathsf{T}}} P(\boldsymbol{w}_{-1}^{k})\right)^{\mathsf{T}}\widetilde{\boldsymbol{w}}_{-1}^{k} &\leq P(w^{\star}) - P(\boldsymbol{w}_{-1}^{k}) - \frac{\nu}{2}\|\widetilde{\boldsymbol{w}}_{-1}^{k}\|^2 \\
&\overset{(8.23)}{\leq} -\frac{\nu}{2}\|\widetilde{\boldsymbol{w}}_{-1}^{k}\|^2 - \frac{\nu}{2}\|\widetilde{\boldsymbol{w}}_{-1}^{k}\|^2 \\
&= -\nu\|\widetilde{\boldsymbol{w}}_{-1}^{k}\|^2
\end{aligned} \tag{19.174}$$

Substituting into (19.173) gives

$$\|\widetilde{\boldsymbol{w}}_{-1}^{k} + \mu N\,\nabla_{w^{\mathsf{T}}} P(\boldsymbol{w}_{-1}^{k})\|^2 \leq (1 - 2\mu\nu N + \mu^2\delta^2 N^2)\|\widetilde{\boldsymbol{w}}_{-1}^{k}\|^2 \tag{19.175}$$

Note that for

$$\mu < \frac{2\nu}{3N\delta^2} \tag{19.176}$$

we have

$$\begin{aligned}1-2\mu\nu N+\mu^2\delta^2N^2 &\le 1-\frac{4\mu\nu N}{3}\\ &\le 1-\frac{4\mu\nu N}{3}+\frac{4\mu^2\nu^2N^2}{9}\\ &\le \left(1-\frac{2\mu\nu N}{3}\right)^2\end{aligned} \tag{19.177}$$

which has the form of a perfect square. It follows from (19.175) that

$$\|\widetilde{\boldsymbol{w}}_{-1}^k+\mu N\,\nabla_{w^\mathsf{T}}P(\boldsymbol{w}_{-1}^k)\|^2 \le \left(1-\frac{2\mu\nu N}{3}\right)^2\|\widetilde{\boldsymbol{w}}_{-1}^k\|^2 \tag{19.178}$$

Consider now the second term on the right-hand side of (19.171) and note that

$$\begin{aligned}\sum_{n=0}^{N-1}\left\|\nabla_{w^\mathsf{T}}Q(\boldsymbol{w}_{n-1}^k)-\nabla_{w^\mathsf{T}}Q(\boldsymbol{w}_{-1}^k)\right\|^2 &\overset{(18.10b)}{\le} \delta^2\sum_{n=0}^{N-1}\left\|\boldsymbol{w}_{n-1}^k-\boldsymbol{w}_{-1}^k\right\|^2\\ &\overset{(a)}{=} \delta^2\sum_{n=0}^{N-1}\left\|\sum_{m=0}^{n-1}\boldsymbol{w}_m^k-\boldsymbol{w}_{m-1}^k\right\|^2\\ &\overset{(19.172)}{\le} \delta^2\sum_{n=0}^{N-1}n\sum_{m=0}^{n-1}\left\|\boldsymbol{w}_m^k-\boldsymbol{w}_{m-1}^k\right\|^2\\ &\overset{(b)}{=} \delta^2\sum_{m=0}^{N-2}\left\|\boldsymbol{w}_m^k-\boldsymbol{w}_{m-1}^k\right\|^2\left(\sum_{n=m+1}^{N-1}n\right)\\ &\overset{(c)}{\le} \frac{\delta^2N^2}{2}\sum_{m=0}^{N-2}\left\|\boldsymbol{w}_m^k-\boldsymbol{w}_{m-1}^k\right\|^2\end{aligned} \tag{19.179}$$

where step (a) uses the telescoping sum

$$\boldsymbol{w}_{n-1}^k-\boldsymbol{w}_{-1}^k \;=\; \sum_{m=0}^{n-1}\boldsymbol{w}_m^k-\boldsymbol{w}_{m-1}^k \tag{19.180}$$

Step (b) uses the easily-verified property

$$\sum_{n=0}^{N-1}\sum_{m=0}^{n-1}a_{nm} \;=\; \sum_{m=0}^{N-2}\sum_{n=m+1}^{N-1}a_{nm} \tag{19.181}$$

and step (c) uses

$$\sum_{n=m+1}^{N-1}n \;\le\; \sum_{n=0}^{N-1}n \;=\; \frac{N(N-1)}{2} \;\le\; \frac{N^2}{2} \tag{19.182}$$

Continuing with (19.179), we appeal to the stochastic gradient recursion to observe

that for each term in the sum:

$$\begin{aligned}\left\|\boldsymbol{w}_m^k-\boldsymbol{w}_{m-1}^k\right\|^2 &\overset{(19.75)}{=} \mu^2\left\|\nabla_{w^{\mathsf{T}}}Q(\boldsymbol{w}_{m-1}^k)\right\|^2\\ &= \mu^2\left\|\nabla_{w^{\mathsf{T}}}Q(\boldsymbol{w}_{m-1}^k)+\nabla_{w^{\mathsf{T}}}Q(w^\star)-\nabla_{w^{\mathsf{T}}}Q(w^\star)\right\|^2\\ &\le 2\mu^2\left\|\nabla_{w^{\mathsf{T}}}Q(\boldsymbol{w}_{m-1}^k)+\nabla_{w^{\mathsf{T}}}Q(w^\star)\right\|^2+2\mu^2\left\|\nabla_{w^{\mathsf{T}}}Q(w^\star)\right\|^2\\ &\overset{(18.10b)}{\le} 2\mu^2\delta^2\left\|\widetilde{\boldsymbol{w}}_{m-1}^k\right\|^2+2\mu^2\left\|\nabla_{w^{\mathsf{T}}}Q(w^\star)\right\|^2\end{aligned} \tag{19.183}$$

Therefore, we have

$$\begin{aligned}\left\|\boldsymbol{w}_m^k-\boldsymbol{w}_{m-1}^k\right\|^2 &\le 2\mu^2\delta^2\left\|w^\star-\boldsymbol{w}_{-1}^k+\boldsymbol{w}_{-1}^k-\boldsymbol{w}_{m-1}^k\right\|^2+2\mu^2\left\|\nabla_{w^{\mathsf{T}}}Q(w^\star)\right\|^2\\ &\le 4\mu^2\delta^2\left\|\widetilde{\boldsymbol{w}}_{-1}^k\right\|^2+4\mu^2\delta^2\|\boldsymbol{w}_{-1}^k-\boldsymbol{w}_{m-1}^k\|^2+2\mu^2\left\|\nabla_{w^{\mathsf{T}}}Q(w^\star)\right\|^2\end{aligned} \tag{19.184}$$

Introduce the average loss value

$$Q_{\text{av}}\triangleq\frac{1}{N}\sum_{m=0}^{N-1}\left\|\nabla_{w^{\mathsf{T}}}Q(w^\star;\gamma(m),h_m)\right\|^2 \tag{19.185}$$

We know from (18.40) that

$$Q_{\text{av}}=O(\sigma_g^2) \tag{19.186}$$

i.e., it is on the order of the factor σ_g^2 that bounds the second-order moment of the gradient noise process. Adding (19.184) over m gives

$$\begin{aligned}&\sum_{m=0}^{N-1}\left\|\boldsymbol{w}_m^k-\boldsymbol{w}_{m-1}^k\right\|^2\\ &\le 4\mu^2\delta^2N\left\|\widetilde{\boldsymbol{w}}_{-1}^k\right\|^2+2\mu^2NQ_{\text{av}}+4\mu^2\delta^2\sum_{m=0}^{N-1}\|\boldsymbol{w}_{m-1}^k-\boldsymbol{w}_{-1}^k\|^2\\ &\overset{(a)}{=} 4\mu^2\delta^2N\left\|\widetilde{\boldsymbol{w}}_{-1}^k\right\|^2+2\mu^2NQ_{\text{av}}+4\mu^2\delta^2\sum_{m=0}^{N-1}\left\|\sum_{n=0}^{m-1}\boldsymbol{w}_n^k-\boldsymbol{w}_{n-1}^k\right\|^2\\ &\overset{(b)}{\le} 4\mu^2\delta^2N\left\|\widetilde{\boldsymbol{w}}_{-1}^k\right\|^2+2\mu^2NQ_{\text{av}}+4\mu^2\delta^2\sum_{m=0}^{N-1}\sum_{n=0}^{m-1}m\left\|\boldsymbol{w}_n^k-\boldsymbol{w}_{n-1}^k\right\|^2\\ &\overset{(c)}{=} 4\mu^2\delta^2N\left\|\widetilde{\boldsymbol{w}}_{-1}^k\right\|^2+2\mu^2NQ_{\text{av}}+4\mu^2\delta^2\sum_{n=0}^{N-2}\left\|\boldsymbol{w}_n^k-\boldsymbol{w}_{n-1}^k\right\|^2\left(\sum_{m=n+1}^{N-1}m\right)\\ &\overset{(d)}{\le} 4\mu^2\delta^2N\left\|\widetilde{\boldsymbol{w}}_{-1}^k\right\|^2+2\mu^2NQ_{\text{av}}+2\mu^2\delta^2N^2\sum_{n=0}^{N-2}\left\|\boldsymbol{w}_n^k-\boldsymbol{w}_{n-1}^k\right\|^2\\ &\le 4\mu^2\delta^2N\left\|\widetilde{\boldsymbol{w}}_{-1}^k\right\|^2+2\mu^2NQ_{\text{av}}+2\mu^2\delta^2N^2\sum_{n=0}^{N-1}\left\|\boldsymbol{w}_n^k-\boldsymbol{w}_{n-1}^k\right\|^2\end{aligned} \tag{19.187}$$

where in step (a) we used again a telescoping sum representation, in step (b) we used property (19.172), in step (c) we appealed again to (19.181), and in step (d) we used

(19.182). In the last step, we increased the upper limit on the summation on the right-hand side to $N-1$. It follows that

$$\sum_{m=0}^{N-1}\left\|\boldsymbol{w}_m^k-\boldsymbol{w}_{m-1}^k\right\|^2\leq\frac{1}{1-2\mu^2\delta^2N^2}\left(4\mu^2\delta^2N\,\|\widetilde{\boldsymbol{w}}_{-1}^k\|^2\;+\;2\mu^2NQ_{\rm av}\right) \tag{19.188}$$

Combining (19.178), (19.179), and (19.188) into (19.171), we arrive at

$$\begin{aligned}\|\widetilde{\boldsymbol{w}}_{-1}^{k+1}\|^2\leq\frac{1}{t}\left(1-\frac{2\mu\nu N}{3}\right)^2\|\widetilde{\boldsymbol{w}}_{-1}^k\|^2+\\ \frac{\mu^2\delta^2N^3}{2(1-t)(1-2\mu^2\delta^2N^2)}\left(4\mu^2\delta^2N\,\|\widetilde{\boldsymbol{w}}_{-1}^k\|^2\;+\;2\mu^2NQ_{\rm av}\right)\end{aligned} \tag{19.189}$$

We select $t=1-\frac{2\mu\nu N}{3}$ so that

$$\begin{aligned}\|\widetilde{\boldsymbol{w}}_{-1}^{k+1}\|^2&\leq\left(1-\frac{2\mu\nu N}{3}\right)\|\widetilde{\boldsymbol{w}}_{-1}^k\|^2+\\&\qquad\frac{3\mu\delta^2N^2}{4\nu(1-2\mu^2\delta^2N^2)}\left(4\mu^2\delta^2N\,\|\widetilde{\boldsymbol{w}}_{-1}^k\|^2\;+\;2\mu^2NQ_{\rm av}\right)\\&\leq\left(1-\frac{2\mu\nu N}{3}+\frac{3\mu^3\delta^4N^3}{\nu(1-2\mu^2\delta^2N^2)}\right)\|\widetilde{\boldsymbol{w}}_{-1}^k\|^2+\\&\qquad\frac{3\mu^3\delta^2N^3}{2\nu(1-2\mu^2\delta^2N^2)}Q_{\rm av}\end{aligned} \tag{19.190}$$

Assume again that μ is small enough such that

$$1-2\mu^2\delta^2N^2>\frac{3}{4}\iff\mu<\frac{1}{\sqrt{8}N\delta} \tag{19.191}$$

Since $\nu\leq\delta$, this condition is met by any

$$\mu<\frac{\nu}{\sqrt{8}N\delta^2} \tag{19.192}$$

Then, we have

$$\|\widetilde{\boldsymbol{w}}_{-1}^{k+1}\|^2\leq\left(1-\frac{2\mu\nu N}{3}+\frac{4\mu^3\delta^4N^3}{\nu}\right)\|\widetilde{\boldsymbol{w}}_{-1}^k\|^2+\frac{2\mu^3\delta^2N^3}{\nu}Q_{\rm av} \tag{19.193}$$

Assume further that μ is small enough such that

$$1-\frac{2\mu\nu N}{3}+\frac{4\mu^3\delta^4N^3}{\nu}<1-\frac{\mu}{2}\nu N \tag{19.194}$$

which is equivalent to

$$\mu<\frac{\nu}{\sqrt{24}N\delta^2} \tag{19.195}$$

Conditions (19.176), (19.192), and (19.195) are met by (19.77). Then, it follows that

$$\|\widetilde{\boldsymbol{w}}_{-1}^{k+1}\|^2\leq\left(1-\frac{\mu}{2}\nu N\right)\|\widetilde{\boldsymbol{w}}_{-1}^k\|^2+\frac{2\mu^3\delta^2N^3}{\nu}Q_{\rm av} \tag{19.196}$$

or, by taking expectations of both sides,

$$\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{-1}^{k+1}\|^2 \le \Big(1-\frac{\mu}{2}\nu N\Big)\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{-1}^{k}\|^2 + \frac{2\mu^3\delta^2N^3}{\nu}Q_{\rm av} \tag{19.197}$$

and, hence,

$$\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{-1}^{k}\|^2 \;\le\; O(\lambda^k) + O(\mu^2) \tag{19.198}$$

with $\lambda = 1-\frac{\mu\nu N}{2}$. Finally, note that for any n we have

$$\begin{aligned}
\|\widetilde{\boldsymbol{w}}_n^k\|^2 &= \|\widetilde{\boldsymbol{w}}_n^k - \widetilde{\boldsymbol{w}}_{-1}^k + \widetilde{\boldsymbol{w}}_{-1}^k\|^2 \\
&\le 2\|\widetilde{\boldsymbol{w}}_{-1}^k\|^2 + 2\|\widetilde{\boldsymbol{w}}_n^k - \widetilde{\boldsymbol{w}}_{-1}^k\|^2 \\
&= 2\|\widetilde{\boldsymbol{w}}_{-1}^k\|^2 + 2\|\boldsymbol{w}_n^k - \boldsymbol{w}_{-1}^k\|^2 \\
&\overset{(a)}{=} 2\|\widetilde{\boldsymbol{w}}_{-1}^k\|^2 + 2\Big\|\sum_{m=0}^{n}\boldsymbol{w}_m^k - \boldsymbol{w}_{m-1}^k\Big\|^2 \\
&\overset{(19.172)}{\le} 2\|\widetilde{\boldsymbol{w}}_{-1}^k\|^2 + 2(n+1)\sum_{m=0}^{n}\Big\|\boldsymbol{w}_m^k - \boldsymbol{w}_{m-1}^k\Big\|^2 \\
&\overset{(b)}{\le} O(\lambda^k) + O(\mu^2) + O(\mu^2), \quad \text{large } k \\
&= O(\lambda^k) + O(\mu^2)
\end{aligned} \tag{19.199}$$

where in step (a) we used a telescoping series representation and in step (b) we used (19.188) and (19.198). We therefore arrive at (19.79a). To establish (19.79b), we use (19.12b) to note that

$$0 \le \mathbb{E}\,P(\boldsymbol{w}_n^k) - P(w^\star) \le \frac{\delta}{2}\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n^k\|^2 \tag{19.200}$$

■

REFERENCES

Al-Baali, M. (1985), "Descent property and global convergence of the Fletcher-Reeves method with inexact line search," *IMA J. Numerical Analysis*, vol. 5, pp. 121–124.

Albert, A. E. and L. A. Gardner (1967), *Stochastic Approximation and Nonlinear Regression*, MIT Press, Cambridge, MA.

Benveniste, A., M. Métivier, and P. Priouret (1987), *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, NY.

Bertsekas, D. P. and J. N. Tsitsiklis (1997), *Parallel and Distributed Computation: Numerical Methods*, Athena Scientific, Singapore.

Bertsekas, D. P. and J. N. Tsitsiklis (2000), "Gradient convergence in gradient methods with errors," *SIAM J. Optim.*, vol. 10, no. 3, pp. 627–642.

Bottou, L. (2009), "Curiously fast convergence of some stochastic gradient descent algorithms," in *Proc. Symposium on Learning and Data Science*, pp. 1–4, Paris.

Bottou, L. (2010), "Large-scale machine learning with stochastic gradient descent," *Proc. International Conference on Computational Statistics*, pp. 177–186, Paris, France.

Bottou, L., F. E. Curtis, and J. Nocedal (2018), "Optimization methods for large-scale machine learning," *SIAM Review*, vol. 60, no. 2, pp. 223–311.

Gladyshev, E. G. (1965), "On stochastic approximations," *Theory of Probability and its Applications*, vol. 10, pp. 275–278.

Gordon, G. J. (1999), "Regret bounds for prediction problems," *Proc. Annual Conference on Computational Learning Theory* (COLT), pp. 29–40, 1999.

Gürbüzbalaban, M., A. Ozdaglar, and P. Parrilo (2015a), "A globally convergent incremental Newton method," *Mathematical Programming*, vol. 151, no. 1, pp. 283–313.

Haykin, S. (2001), *Adaptive Filter Theory*, 4th edition, Prentice Hall, NJ.

He, S., J. Wang, and J. Yan (1992), *Semimartingale Theory and Stochastic Calculus*, CRC Press.

Jin. X.-B., X.-Y. Zhang, K. Huang, and G.-G. Geng (2019), "Stochastic conjugate gradient algorithm with variance reduction," *IEEE Trans. Neural Networks and Learning Systems*, vol. 30, no. 5, pp. 1360–1369.

Kushner, H. J. and D. S. Clark (1978), *Stochastic Approximation for Constrained and Unconstrained Systems*, Springer-Verlag, NY.

Kushner, H. J. and G. G. Yin (2003), *Stochastic Approximation and Recursive Algorithms and Applications*, Springer, NY.

Lipster, R. and A. N. Shiryayev (1989), *Theory of Martingales*, Springer, NY.

Ljung, L. (1977), "Analysis of recursive stochastic algorithms," *IEEE Trans. Automat. Contr.*, vol. 22, pp. 551–575.

Marti, K. (2005), *Stochastic Optimization Methods*, Springer, NY.

Mendel, J. M. and K. S. Fu (1970), *Adaptive, Learning, and Pattern Recognition Systems: Theory and Applications*, Academic Press, NY.

Needell, D., R. Ward, and N. Srebro (2014), "Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm," *Proc. Advances Neural Information Processing Systems* (NIPS), pp. 1017–1025, Montreal, Canada.

Nemirovski, A. S., A. Juditsky, G. Lan, and A. Shapiro (2009), "Robust stochastic approximation approach to stochastic programming," *SIAM J. Optm.*, vol. 19, no. 4, pp. 1574–1609.

Nocedal, J. and S. J. Wright (2006), *Numerical Optimization*, Springer, NY.

Polyak, B. T. (1987), *Introduction to Optimization*, Optimization Software, NY.

Polyak, B. T. and A. Juditsky (1992), "Acceleration of stochastic approximation by averaging," *SIAM J. Control and Optim.*, vol. 30, no. 4, pp. 838–855.

Powell, M. J. D. (1985), "Convergence properties of algorithms for nonlinear optimization," *Report DAMTP 1985*, Department of Applied Mathematics and Theoretical Physics, Cambridge University, England.

Recht, B. and C. Re (2012), "Toward a noncommutative arithmetic-geometric mean inequality: Conjectures, case-studies, and consequences," in *Proc. Conference on Learning Theory* (COLT), pp. 1–11, Edinburgh, Scotland.

Robbins, H. and S. Monro (1951), "A stochastic approximation method," *Ann. Math. Stat.*, vol. 22, pp. 400–407.

Ruppert, D. (1988), *Efficient Estimation From a Slowly Convergent Robbins-Monro Process*, Technical Report 781, Cornell University, School of Operations Research and Industrial Engineering.

Sayed, A. H. (2003), *Fundamentals of Adaptive Filtering*, Wiley, NJ.

Sayed, A. H. (2008), *Adaptive Filters*, Wiley, NJ.

Sayed, A. H. (2014a), *Adaptation, Learning, and Optimization over Networks,* Foundations and Trends in Machine Learning, NOW Publishers, vol. 7, no. 4–5, pp. 311–801.

Sayed, A. H. (2014b), "Adaptive networks," *Proc. IEEE*, vol. 102, no. 4, pp. 460–497, 2014.

Sayed, A. H. and X. Zhao (2018), "Asynchronous adaptive networks," in *Cooperative and Graph Signal Processing*, P. Djuric and C. Richard, *Eds.*, pp. 3–68, Elsevier, 2018. Also available online at https://arxiv.org/abs/1511.09180.

Shalev-Shwartz, S. (2011), "Online learning and online convex optimization," *Foundations and Trends in Machine Learning*, vol. 4, no. 2, pp. 107–194.

Shalev-Shwartz, S. and S. Ben-David (2014), *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press.

Shamir, O. (2016), "Without-replacement sampling for stochastic gradient methods: Convergence results and application to distributed optimization," *Proc. Advances Neural Information Processing Systems* (NIPS), D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (*Eds.*), pp. 46–54, Curran Associates.

Spall, J. C. (2003), *Introduction to Stochastic Search and Optimization: Estimation, Simulation and Control*, Wiley, NJ.

Sun, J. and J. Zhang (2001), "Global convergence of conjugate gradient methods," *Annals of Operations Research*, vol. 103, pp. 161–173.

Tsypkin, Y. Z. (1971), *Adaptation and Learning in Automatic Systems*, Academic Press, NY.

Wasan, M. T. (1969), *Stochastic Approximation*, Cambridge University Press, London.

Widrow, B. and S. D. Stearns (1985), *Adaptive Signal Processing*, Prentice Hall, NJ.

Williams, D. (1991), *Probability with Martingales*, Cambridge University Press.

Ying, B., K. Yuan, S. Vlaski, and A. H. Sayed (2019), "Stochastic learning under random reshuffling with constant step-sizes," *IEEE Trans. Signal Processing*, vol. 67, no. 2, pp. 474–489.

Yu, H., R. Jin, and S. Yang (2019), "On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization," *Proceedings of Machine Learning Research* (PMLR), vol. 97, pp. 7184–7193.

Yuan, K., B. Ying, S. Vlaski, and A. H. Sayed (2016), "Stochastic gradient descent with finite sample sizes," *Proc. IEEE International Workshop on Machine Learning for Signal Processing* (MLSP), pp. 1–6, Salerno Italy.

Zhao, X. and A. H. Sayed (2015a), "Asynchronous adaptation and learning over networks – Part I: Modeling and stability analysis," *IEEE Trans. Signal Process.*, vol. 63, no. 4, pp. 811–826.

Zhao, X. and A. H. Sayed (2015b), "Asynchronous adaptation and learning over networks – Part II: Performance analysis," *IEEE Trans. Signal Process.*, vol. 63, no. 4, pp. 827–842.

Zhao, P. and T. Zhang (2015), "Stochastic optimization with importance sampling for regularized loss minimization," in *Proc. International Conference on Machine Learning* (ICML), Lille, France, pp. 1355–1363.

Zinkevich, M. (2003), " Online convex programming and generalized infinitesimal gradient ascent," *Proc. Intern. Conference on Machine Learning* (ICML), pp. 928–936, Washington, DC.

Zountendijk, G. (1970), "Nonlinear programming, computational methods," in *Integer and Nonlinear Programming*, J. Abadie, *Editor*, pp. 37–86, North-Holland, Amsterdam.