18 GRADIENT NOISE

The purpose of this chapter is to study the gradient noise process more closely, for both cases of smooth and nonsmooth risk functions, and to derive expressions for its first and second-order moments (i.e., mean and variance). The results will then be exploited in the subsequent chapters to assess how gradient noise affects the convergence behavior of various stochastic approximation algorithms. The presentation in the chapter prepares the ground for the detailed convergence analyses given in the next chapters. Throughout this chapter, we will use the terminology "*smooth*" functions to refer to risks that are at least first-order differentiable everywhere in their domain, and apply the qualification "*non-smooth*" functions to risks that are not differentiable at some points in their domains.

18.1 MOTIVATION

We examined several stochastic optimization algorithms in the previous chapters for the solution of convex optimization problems of the form:

$$w^\star = \operatorname*{argmin}_{w\in\mathbb{R}^M} \; P(w) \tag{18.1}$$

with and without constraints on w, for both smooth and nonsmooth risks, as well as for empirical and stochastic risks, namely,

$$w^\star \;\triangleq\; \operatorname*{argmin}_{w\in\mathbb{R}^M} \left\{ P(w) \;\triangleq\; \frac{1}{N}\sum_{m=0}^{N-1} Q(w;\gamma(m),h_m) \right\} \tag{18.2a}$$

$$w^o \;\triangleq\; \operatorname*{argmin}_{w\in\mathbb{R}^M} \left\{ P(w) \;\triangleq\; \mathbb{E}\, Q(w;\boldsymbol{\gamma},\boldsymbol{h}) \right\} \tag{18.2b}$$

In these expressions, $Q(w,\cdot)$ denotes some convex 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 $\{\boldsymbol{\gamma},\boldsymbol{h}\}$. In most algorithms, the desired gradient or subgradient search direction was approximated by using either instantaneous or mini-batch calculations. For example,

for smooth risk functions $P(w)$, we used approximations of the form:

$$(\textbf{instantaneous}): \ \widehat{\nabla_{w^{\mathsf{T}}} P}(w) = \nabla_{w^{\mathsf{T}}} Q(w; \boldsymbol{\gamma}, \boldsymbol{h}) \tag{18.3a}$$

$$(\textbf{mini-batch}): \ \widehat{\nabla_{w^{\mathsf{T}}} P}(w) = \frac{1}{B} \sum_{b=0}^{B-1} \nabla_{w^{\mathsf{T}}} Q\Big(w; \boldsymbol{\gamma}(b), \boldsymbol{h}_b\Big) \tag{18.3b}$$

where the boldface notation $(\boldsymbol{\gamma}, \boldsymbol{h})$ or $(\boldsymbol{\gamma}(b), \boldsymbol{h}_b)$ refers to data samples selected at random from the dataset $\{\gamma(m), h_m\}$ in empirical risk minimization, or assumed to stream in independently over time in stochastic risk minimization. When $P(w)$ happens to be nonsmooth, the gradient vectors of $Q(w;\cdot)$ are replaced by subgradients, denoted by $s_Q(w; \boldsymbol{\gamma}, \boldsymbol{h})$. The difference between the true gradient and its approximation is *gradient noise* and denoted by

$$\boldsymbol{g}(w) \ \triangleq \ \widehat{\nabla_{w^{\mathsf{T}}} P}(w) - \nabla_{w^{\mathsf{T}}} P(w) \tag{18.4}$$

We explained in Sec. 16.4 that the presence of this noise source alters the dynamics of the optimization algorithms. For example, the following two relations highlight the difference between the original gradient-descent method and its stochastic version for smooth risks:

$$(\textbf{gradient-descent}): w_n = w_{n-1} - \mu \nabla_{w^{\mathsf{T}}} P(w_{n-1}) \tag{18.5a}$$

$$\begin{aligned}(\textbf{stochastic version}): \boldsymbol{w}_n &= \boldsymbol{w}_{n-1} - \mu \widehat{\nabla_{w^{\mathsf{T}}} P}(\boldsymbol{w}_{n-1}) \\ &= \boldsymbol{w}_{n-1} - \mu \nabla_{w^{\mathsf{T}}} P(\boldsymbol{w}_{n-1}) - \mu\, \boldsymbol{g}(\boldsymbol{w}_{n-1})\end{aligned} \tag{18.5b}$$

The gradient noise appears as a driving perturbation in the second recursion. This is illustrated in Fig. 18.1, where the block with z^{-1} represents a unit delay element. The panel on top shows the dynamics of (18.5a), while the panel in the bottom shows the dynamics of the perturbed update (18.5b). The gradient noise seeps into the operation of the algorithm and some degradation in performance is expected. While we were able to show in a previous chapter that the gradient descent implementation (18.5a) converges to the exact minimizer $w^\star$ of $P(w)$ for sufficiently small step-sizes, we will discover in future chapters that the stochastic version (18.5b) can only approach a small neighborhood around $w^\star$ of size $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_n\|^2 = O(\mu)$ as $n \to \infty$.

Example 18.1 (**Gradient noise for quadratic risks**) We illustrate the concept of gradient noise by considering two quadratic risks: one empirical and the other stochastic. Consider first the empirical risk:

$$P(w) \ = \ \rho\|w\|^2 \ + \ \frac{1}{N} \sum_{m=0}^{N-1} (\gamma(m) - h_m^{\mathsf{T}} w)^2, \ \ \rho > 0 \tag{18.6}$$

In this case, the gradient vector and its instantaneous approximation are given by

$$\nabla_w P(w) = 2\rho w - \frac{2}{N} \sum_{m=0}^{N-1} h_m (\gamma(m) - h_m^{\mathsf{T}} w) \tag{18.7a}$$

$$\widehat{\nabla_w P}(w) = 2\rho w - 2\boldsymbol{h}_n (\boldsymbol{\gamma}(n) - \boldsymbol{h}_n^{\mathsf{T}} w) \tag{18.7b}$$

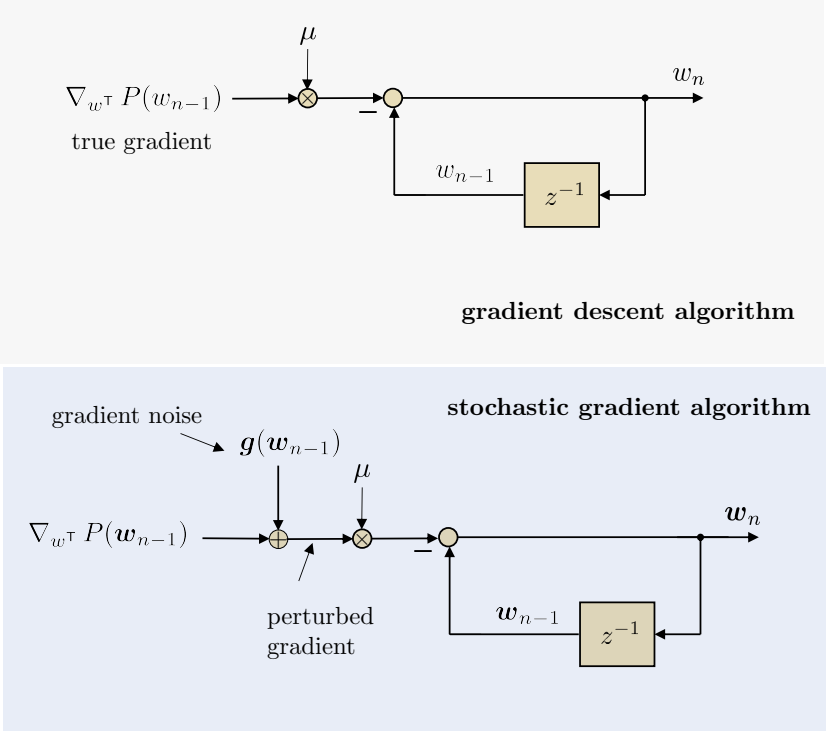

Figure 18.1 The panel on top shows the dynamics of the original gradient-descent recursion (18.5a) while the panel in the bottom shows the dynamics of the stochastic version (18.5b). The true gradient vector is perturbed by gradient noise, which seeps into the operation of the algorithm. The block with z^{-1} represents a unit delay element.

where $\{\boldsymbol{\gamma}(n), \boldsymbol{h}_n\}$ refer to the random sample selected at iteration n by the stochastic gradient implementation. The resulting gradient noise process is then given by

$$\boldsymbol{g}(w) = \frac{2}{N}\sum_{m=0}^{N-1} h_m(\gamma(m) - h_m^\mathsf{T} w) \;-\; 2\boldsymbol{h}_n(\boldsymbol{\gamma}(n) - \boldsymbol{h}_n^\mathsf{T} w) \tag{18.7c}$$

Observe that $\boldsymbol{g}(w)$ depends on the sample $\{\boldsymbol{\gamma}(n), \boldsymbol{h}_n\}$ and, therefore, in principle, we should be writing $\boldsymbol{g}_n(w)$ with a subscript n to highlight its dependency on n.

Consider next the stochastic risk:

$$\begin{aligned} P(w) &= \rho\|w\|^2 \;+\; \mathbb{E}\,(\boldsymbol{\gamma} - \boldsymbol{h}^\mathsf{T} w)^2 \\ &= \sigma_\gamma^2 - 2r_{h\gamma}^\mathsf{T} w + w^\mathsf{T}(\rho I_M + R_h) w \end{aligned} \tag{18.8}$$

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 $R_h = \mathbb{E}\,\boldsymbol{h}\boldsymbol{h}^\mathsf{T}$. The random variables $\{\boldsymbol{\gamma}, \boldsymbol{h}\}$ are assumed to have zero means. In this case, the gradient of $P(w)$ and its instantaneous approximation are given by

$$\nabla_w P(w) = 2\rho w - 2(r_{h\gamma} - R_h w) \tag{18.9a}$$

$$\widehat{\nabla_w P}(w) = 2\rho w - 2\boldsymbol{h}_n(\boldsymbol{\gamma}(n) - \boldsymbol{h}_n^\mathsf{T} w) \tag{18.9b}$$

so that the corresponding gradient noise process is now

$$\boldsymbol{g}(w) = 2(r_{h\gamma} - R_h w) - 2\boldsymbol{h}_n(\boldsymbol{\gamma}(n) - \boldsymbol{h}_n^{\mathsf{T}} w) \tag{18.9c}$$

Observe again that $\boldsymbol{g}(w)$ depends on the streaming sample $\{\boldsymbol{\gamma}(n), \boldsymbol{h}_n\}$.

18.2 SMOOTH RISK FUNCTIONS

To facilitate the analysis and presentation, we will treat smooth and nonsmooth risks separately, although we will end up with the same ultimate conclusion about the gradient noise for both cases. We start with smooth risks and describe the conditions that are normally imposed on the risk and loss functions, $P(w)$ and $Q(w,\cdot)$. The conditions listed here are satisfied by several risk and loss functions of interest, as illustrated in the problems at the end of the chapter.

Empirical risks

Consider smooth empirical risks of the form (18.2a). We will assume that the risk and loss functions satisfy the two conditions listed below. Compared with the earlier conditions (12.12a)–(12.12b) in the gradient-descent case, we see that we now need to take the loss function into consideration since its gradients are the ones used in the stochastic implementation:

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

$$P(w_2) \;\geq\; P(w_1) + (\nabla_{w^{\mathsf{T}}} P(w_1))^{\mathsf{T}}(w_2 - w_1) \;+\; \frac{\nu}{2}\|w_2 - w_1\|^2 \tag{18.10a}$$

for some $\nu > 0$.

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

$$\|\nabla_w Q(w_2; \gamma(k), h_k) - \nabla_w Q(w_1; \gamma(\ell), h(\ell))\| \;\leq\; \delta\,\|w_2 - w_1\| \tag{18.10b}$$

for any $w_1, w_2 \in \mathrm{dom}(Q)$, any $0 \leq k, \ell \leq N-1$, and with $\delta \geq \nu$ (this latter requirement can always be met by enlarging δ). Condition (18.10b) is equivalent to saying the loss function is $\delta-$smooth. It is easy to verify from the triangle inequality of norms that (18.10b) implies that the gradient of $P(w)$ is itself $\delta-$Lipschitz:

$$\|\nabla_w P(w_2) - \nabla_w P(w_1)\| \leq \delta\|w_2 - w_1\| \tag{18.11}$$

Moreover, if it happens that $P(w)$ is twice-differentiable, then we already know from (12.15) that conditions (18.10a) and (18.11) combined are equivalent to:

$$0 < \nu I_M \leq \nabla_w^2 P(w) \;\leq\; \delta I_M \tag{18.12}$$

in terms of the Hessian matrix of $P(w)$.

Stochastic risks

For stochastic risks of the form (18.2b), we continue to assume that $P(w)$ is ν-strongly-convex but that the loss function has gradients that are $\delta-$Lipschitz in the *mean-square sense*:

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

$$P(w_2) \geq P(w_1) + (\nabla_{w^{\mathsf{T}}} P(w_1))^{\mathsf{T}}(w_2 - w_1) + \frac{\nu}{2}\|w_2 - w_1\|^2 \quad (18.13a)$$

for some $\nu > 0$.

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

$$\mathbb{E}\,\|\nabla_w Q(w_2; \boldsymbol{\gamma}, \boldsymbol{h}) - \nabla_w Q(w_1; \boldsymbol{\gamma}, \boldsymbol{h})\|^2 \leq \delta^2 \|w_2 - w_1\|^2 \quad (18.13b)$$

for any $w_1, w_2 \in \mathrm{dom}(Q)$ and with $\delta \geq \nu$. The expectation is over the joint distribution of the random data $\{\boldsymbol{\gamma}, \boldsymbol{h}\}$. Using the fact that for any scalar random variable $\boldsymbol{x}$ it holds that $(\mathbb{E}\,\boldsymbol{x})^2 \leq \mathbb{E}\,\boldsymbol{x}^2$, we conclude from condition (18.13b) that the gradient vectors of the loss function are also $\delta-$Lipschitz on *average*, namely,

$$\mathbb{E}\,\|\nabla_w Q(w_2; \boldsymbol{\gamma}, \boldsymbol{h}) - \nabla_w Q(w_1; \boldsymbol{\gamma}, \boldsymbol{h})\| \leq \delta \|w_2 - w_1\| \quad (18.14)$$

By further applying Jensen inequality (8.77) that $f(\mathbb{E}\,\boldsymbol{x}) \leq \mathbb{E}\,f(\boldsymbol{x})$ for the convex function $f(x) = \|x\|$, we can conclude from (18.14) that the gradients of $P(w)$ are themselves $\delta-$Lipschitz as well:

$$\|\nabla_w P(w_2) - \nabla_w P(w_1)\| \leq \delta \|w_2 - w_1\| \quad (18.15)$$

Proof of (18.15): Note that

$$\begin{aligned}
\|\nabla_w P(w_2) - \nabla_w P(w_1)\| &\triangleq \|\nabla_w \mathbb{E}\,Q(w_2; \boldsymbol{\gamma}, \boldsymbol{h}) - \nabla_w \mathbb{E}\,Q(w_1; \boldsymbol{\gamma}, \boldsymbol{h})\| \\
&\overset{(a)}{=} \Big\|\mathbb{E}\,\Big(\nabla_w Q(w_2; \boldsymbol{\gamma}, \boldsymbol{h}) - \nabla_w Q(w_1; \boldsymbol{\gamma}, \boldsymbol{h})\Big)\Big\| \\
&\leq \mathbb{E}\,\|\nabla_w Q(w_2; \boldsymbol{\gamma}, \boldsymbol{h}) - \nabla_w Q(w_1; \boldsymbol{\gamma}, \boldsymbol{h})\| \\
&\overset{(18.14)}{\leq} \delta \|w_2 - w_1\| \qquad (18.16)
\end{aligned}$$

Step (a) switches the order of the expectation and differentiation operators, which is possible under certain conditions that are generally valid for our cases of interest — recall the explanation in Appendix 16.A on the *dominated convergence theorem*. In particular, the switching is possible when the loss function $Q(w; \cdot, \cdot)$ and its gradient are continuous functions of w.

■

For ease of reference, we collect in Table 18.1 the main relations and conditions described so far for smooth empirical and stochastic risk minimization.

Table 18.1 Main relations and conditions used for *smooth* empirical and stochastic risk minimization problems.

quantity	empirical risk minimization	stochastic risk minimization
Optimization problem	$w^\star = \underset{w\in\mathbb{R}^M}{\operatorname{argmin}} \left\{ P(w) \triangleq \frac{1}{N}\sum_{m=0}^{N-1} Q(w;\gamma(m),h_m) \right\}$	$w^o = \underset{w\in\mathbb{R}^M}{\operatorname{argmin}} \left\{ P(w) \triangleq \mathbb{E}\, Q(w;\boldsymbol{\gamma},\boldsymbol{h}) \right\}$
True gradient vector	$\nabla_{w^\mathsf{T}} P(w) = \frac{1}{N}\sum_{m=0}^{N-1} \nabla_{w^\mathsf{T}} Q(w;\gamma(m),h_m)$	$\nabla_{w^\mathsf{T}} P(w) = \nabla_{w^\mathsf{T}} \mathbb{E}\, Q(w;\boldsymbol{\gamma},\boldsymbol{h})$
Instantaneous approximation	$\widehat{\nabla_{w^\mathsf{T}} P}(w) = \nabla_{w^\mathsf{T}} Q(w;\boldsymbol{\gamma}(n),\boldsymbol{h}_n)$ $(\boldsymbol{\gamma}(n),\boldsymbol{h}_n)$ selected at random	$\widehat{\nabla_{w^\mathsf{T}} P}(w) = \nabla_{w^\mathsf{T}} Q(w;\boldsymbol{\gamma}(n),\boldsymbol{h}_n)$ $(\boldsymbol{\gamma}(n),\boldsymbol{h}_n)$ streaming in
Mini-batch approximation	$\widehat{\nabla_{w^\mathsf{T}} P}(w) = \frac{1}{B}\sum_{b=0}^{B-1} \nabla_{w^\mathsf{T}} Q(w;\boldsymbol{\gamma}(b),\boldsymbol{h}_b)$ $\{\boldsymbol{\gamma}(b),\boldsymbol{h}_b\}$ selected at random	$\widehat{\nabla_{w^\mathsf{T}} P}(w) = \frac{1}{B}\sum_{b=0}^{B-1} \nabla_{w^\mathsf{T}} Q(w;\boldsymbol{\gamma}(b),\boldsymbol{h}_b)$ $\{\boldsymbol{\gamma}(b),\boldsymbol{h}_b\}$ streaming in
Conditions on risk and loss functions	(18.10a)–(18.10b) $P(w)$ $\nu-$strongly convex $\nabla_{w^\mathsf{T}} Q(w;\gamma,h)$ $\delta-$Lipschitz	(18.13a)–(18.13b) $P(w)$ $\nu-$strongly convex $\nabla_{w^\mathsf{T}} Q(w;\boldsymbol{\gamma},\boldsymbol{h})$ $\delta-$Lipschitz in mean-square sense

18.3 GRADIENT NOISE FOR SMOOTH RISKS

Using the $\delta-$Lipschitz conditions on the gradient of the loss function alone, we will now derive expressions for the first and second-order moments of the gradient noise. For the instantaneous and mini-batch constructions (18.3a)–(18.3b), the gradient noise at iteration n is given by

$$\text{(instantaneous approximation)}$$
$$\boldsymbol{g}(\boldsymbol{w}_{n-1}) = \nabla_{w^{\mathsf{T}}} Q(\boldsymbol{w}_{n-1}; \boldsymbol{\gamma}(n), \boldsymbol{h}_n) - \nabla_{w^{\mathsf{T}}} P(\boldsymbol{w}_{n-1}) \tag{18.17a}$$

for instantaneous gradient approximations, and by

$$\text{(mini-batch approximation)}$$
$$\boldsymbol{g}(\boldsymbol{w}_{n-1}) = \frac{1}{B}\sum_{b=0}^{B-1} \nabla_{w^{\mathsf{T}}} Q(\boldsymbol{w}_{n-1}; \boldsymbol{\gamma}(b), \boldsymbol{h}_b) - \nabla_{w^{\mathsf{T}}} P(\boldsymbol{w}_{n-1}) \tag{18.17b}$$

for mini-batch approximations, where $(\boldsymbol{\gamma}(n), \boldsymbol{h}_n)$ and $\{\boldsymbol{\gamma}(b), \boldsymbol{h}_b\}$ denote the random data samples used at the $n-$th iteration while updating $\boldsymbol{w}_{n-1}$ to $\boldsymbol{w}_n$. It is important to recognize that the gradient noise is *random* in nature because its calculation depends on the random data samples. For this reason, we are denoting it in boldface. Moreover, the gradient noise is dependent on the iteration index n because its calculation depends on $\boldsymbol{w}_{n-1}$ and on the data samples used at that iteration. For added clarify, we will often write $\boldsymbol{g}_n(\boldsymbol{w}_{n-1})$ instead of just $\boldsymbol{g}(\boldsymbol{w}_{n-1})$, with an added subscript n, in order to emphasize that we are referring to the gradient noise computed at iteration n.

The main conclusion of this section (and actually, of this chapter) will be to show that the conditional second-order moment of the gradient noise is bounded as follows:

$$\boxed{\mathbb{E}\left(\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2 \,|\, \boldsymbol{w}_{n-1}\right) \leq \beta_g^2 \|\widetilde{w}_{n-1}\|^2 + \sigma_g^2} \tag{18.18}$$

for some nonnegative constants (β_g^2, σ_g^2) that will be independent of the error $\widetilde{w}_{n-1} = w^\star - w_{n-1}$ (here, we are using $w^\star$ to refer generically to the minimizer of the risk function $P(w)$, whether empirical or stochastic in nature). The conditioning on $\boldsymbol{w}_{n-1}$ in (18.18) could have been written more explicitly as

$$\mathbb{E}\left(\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2 \,|\, \boldsymbol{w}_{n-1} = w_{n-1}\right) \tag{18.19}$$

to indicate that the conditioning is based on an actual realization for $\boldsymbol{w}_{n-1}$. For convenience, we will be using the simpler notation shown in (18.18) throughout our presentation where the conditioning is written relative to the random variable. Result (18.18) shows that the second-order moment of the gradient noise is upper bounded by a quadratic term that involves *two* factors: one factor is dependent on $\|\widetilde{w}_{n-1}\|^2$ and, therefore, gets smaller as the quality of the iterate

w_{n-1} improves, while the second factor is a *constant* term σ_g^2. This latter term is persistent and continues to exist even if $\|\widetilde{w}_{n-1}\|^2$ approaches zero.

It is important to remark that result (18.18) is only dependent on how the approximate gradient vector is constructed; the result does not depend on the particular stochastic approximation algorithm used to update the successive iterates from $\boldsymbol{w}_{n-1}$ to $\boldsymbol{w}_n$. By conditioning on $\boldsymbol{w}_{n-1}$, we are in effect stating that the bound holds regardless of how this iterate is generated. Once its value is given and used to compute the gradient approximation, then the resulting gradient noise will satisfy (18.18).

Before establishing (18.18), it is worth recalling the types of sampling strategies that can be employed by a stochastic approximation algorithm to select itsr random samples.

18.3.1 Sampling Strategies

For *mini-batch* implementations, the B samples can be chosen with or without replacement or they can be streaming in:

(a) (**Sampling with replacement**). In this case, we sample *with replacement* one data point $(\boldsymbol{\gamma}(b), \boldsymbol{h}_b)$ at a time from the $N-$dataset $\{\gamma(m), h_m\}$ until B samples have been selected. In this way, all samples within the mini-batch are selected independently of each other, although some samples may appear repeated.

(b) (**Sampling without replacement**). We can also sample *without replacement*, one data point $(\boldsymbol{\gamma}(b), \boldsymbol{h}_b)$ at a time from the dataset $\{\gamma(m), h_m\}$ until B samples have been selected. Here, the samples within the mini-batch will be different but the selections will not be independent of each other anymore.

(c) (**Streaming data**). For stochastic risk minimization, the samples $\{\boldsymbol{\gamma}(b), \boldsymbol{h}_b\}$ used in the mini-batch will be streaming in independently of each other.

(d) (**Importance sampling**). In this case, a probability value p_m is assigned to each sample $(\gamma(m), h_m)$ in the dataset, and the mini-batch samples are selected at random (with replacement) from the dataset according to this distribution. We explained in Example 16.2 that the approximation for the gradient vector will need to be adjusted to include an additional scaling by $1/Np_b$ — compare with (18.3b):

$$\widehat{\nabla_{w^{\mathsf{T}}} P}(w) = \frac{1}{B}\sum_{b=0}^{B-1} \frac{1}{Np_b} \nabla_{w^{\mathsf{T}}} Q(w; \boldsymbol{\gamma}(b), \boldsymbol{h}_b) \tag{18.20}$$

We clarify in the sequel how this scaling corrects an inherent bias that is present under importance sampling — see argument (18.30).

For implementations with *instantaneous gradient approximations*, the random sample can also be selected with or without replacement or it can stream in:

(a') (**Sampling with replacement**). In this case, the sample $(\boldsymbol{\gamma}(n), \boldsymbol{h}_n)$ at iteration n is selected uniformly at random from the dataset $\{\gamma(m), h_m\}$ with replacement. Some sample points may be selected multiple times.

(b') (**Sampling without replacement**). In this case, the sample $(\boldsymbol{\gamma}(n), \boldsymbol{h}_n)$ at iteration n is selected at random from the same dataset but without replacement.

(c') (**Streaming data**). For stochastic risk minimization, the samples $(\boldsymbol{\gamma}(n), \boldsymbol{h}_n)$ stream in independently of each other.

(d') (**Importance sampling**). In this case, a probability value p_m is assigned to each sample $(\gamma(m), h_m)$ in the dataset, and the sample $(\boldsymbol{\gamma}(n), \boldsymbol{h}_n)$ is selected at random according to this distribution. We also explained in Example 16.2 that the approximation for the gradient vector will need to be adjusted to include an additional scaling by $1/Np_n$ — compare with (18.3a):

$$\widehat{\nabla_{w^{\mathsf{T}}} P}(w) = \frac{1}{Np_n} \nabla_{w^{\mathsf{T}}} Q(w; \boldsymbol{\gamma}(n), \boldsymbol{h}_n) \tag{18.21}$$

where p_n is the probability with which sample $(\boldsymbol{\gamma}(n), \boldsymbol{h}_n)$ is selected. We clarify in the sequel how the scaling corrects the bias that arises under importance sampling — see argument (18.30).

The derivations in the remainder of this section are meant to establish the following main conclusion.

Lemma 18.1. (Gradient noise under smooth risks) *Consider the empirical or stochastic risk optimization problems (18.2a)–(18.2b) and assume the risk and loss functions are first-order differentiable with the gradients of the loss function satisfying the $\delta-$Lipschitz conditions (18.10b) or (18.13b). The first and second-order moments of the gradient noise process will satisfy:*

$$\mathbb{E}\,(\boldsymbol{g}_n(\boldsymbol{w}_{n-1}) \,|\, \boldsymbol{w}_{n-1}) = 0 \tag{18.22a}$$

$$\mathbb{E}\,(\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2 \,|\, \boldsymbol{w}_{n-1}) \leq \beta_g^2 \|\widetilde{w}_{n-1}\|^2 + \sigma_g^2 \tag{18.22b}$$

for some nonnegative constants $\{\beta_g^2, \sigma_g^2\}$ that are independent of $\widetilde{w}_{n-1}$.

Results (18.22a)–(18.22b) 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 (18.22a) 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 analyses in the next chapters. When property (18.22a) does not hold, the convergence argument will need to be adjusted (and becomes more demanding) but will continue to lead to the same conclusion.

To establish properties (18.22a)–(18.22b), we proceed by examining each sampling procedure separately and then show that they all lead to the same result. We consider the zero-mean property (18.22a) first.

18.3.2 First-Order Moment

We verify in this section that for almost all cases of interest, the gradient noise process has zero mean conditioned on the previous iterate, i.e.,

$$\mathbb{E}\,(\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\,|\,\boldsymbol{w}_{n-1}) \;=\; 0 \tag{18.23}$$

Sampling with replacement

Consider first the case of an instantaneous gradient approximation where a single sample is chosen at each iteration n. Let $\boldsymbol{\sigma}$ denote the index of the data sample selected at that iteration so that

$$\mathbb{P}(\boldsymbol{\sigma}=m)=1/N,\quad m\in\{0,1,2,\ldots,N-1\} \tag{18.24}$$

In this case, the approximate search direction is unbiased since, by conditioning on $\boldsymbol{w}_{n-1}$, we get

$$\begin{aligned}\mathbb{E}\,\Big(\widehat{\nabla_{w^{\mathsf{T}}}P}(\boldsymbol{w}_{n-1})\,|\,\boldsymbol{w}_{n-1}\Big) &= \mathbb{E}\,\Big(\nabla_{w^{\mathsf{T}}}Q(\boldsymbol{w}_{n-1};\gamma(\boldsymbol{\sigma}),h_{\boldsymbol{\sigma}})\,|\,\boldsymbol{w}_{n-1}\Big)\\ &= \frac{1}{N}\sum_{\sigma=0}^{N-1}\nabla_{w^{\mathsf{T}}}Q(w_{n-1};\gamma(\sigma),h_\sigma)\\ &\overset{(18.2a)}{=} \nabla_{w^{\mathsf{T}}}P(w_{n-1})\end{aligned} \tag{18.25}$$

where in the second equality we used the fact that the loss function assumes each of the values $Q(w_{n-1};\gamma(\sigma),h_\sigma)$ with probability $1/N$. We conclude that (18.23) holds. This is a reassuring conclusion because it means that, on average, the approximation we are using for the gradient vector agrees with the actual gradient.

The gradient noise process continues to have zero conditional mean in the mini-batch implementation. This is because the approximate search direction is again unbiased:

$$\begin{aligned}\mathbb{E}\,\Big(\widehat{\nabla_{w^{\mathsf{T}}}P}(\boldsymbol{w}_{n-1})\,|\,\boldsymbol{w}_{n-1}\Big) &= \mathbb{E}\left(\frac{1}{B}\sum_{b=0}^{B-1}\nabla_{w^{\mathsf{T}}}Q(\boldsymbol{w}_{n-1};\boldsymbol{\gamma}(b),\boldsymbol{h}_b)\,|\,\boldsymbol{w}_{n-1}\right)\\ &= \frac{1}{B}\sum_{b=0}^{B-1}\mathbb{E}\,\Big(\nabla_{w^{\mathsf{T}}}Q(\boldsymbol{w}_{n-1};\boldsymbol{\gamma}(b),\boldsymbol{h}_b)\,|\,\boldsymbol{w}_{n-1}\Big)\\ &\overset{(a)}{=} \mathbb{E}\,\Big(\nabla_{w^{\mathsf{T}}}Q(\boldsymbol{w}_{n-1};\gamma(\boldsymbol{\sigma}),\boldsymbol{h}_{\boldsymbol{\sigma}})\,|\,\boldsymbol{w}_{n-1}\Big)\\ &= \frac{1}{N}\sum_{\sigma=0}^{N-1}\nabla_{w^{\mathsf{T}}}Q(w_{n-1};\gamma(\sigma),h_\sigma)\\ &= \nabla_{w^{\mathsf{T}}}P(w_{n-1})\end{aligned} \tag{18.26}$$

where in step (a) we used the fact that the data samples $(\boldsymbol{\gamma}(b), \boldsymbol{h}_b)$ are selected independently of each other.

Sampling without replacement

When the data point $(\boldsymbol{\gamma}(n), \boldsymbol{h}_n)$ is sampled *without* replacement from the dataset $\{\gamma(m), h_m\}$ and used to compute an instantaneous gradient approximation, we find that

$$\begin{aligned}\mathbb{E}\left(\widehat{\nabla_{w^\mathsf{T}} P}(\boldsymbol{w}_{n-1}) \,|\, \boldsymbol{w}_{n-1}\right) &= \mathbb{E}\left(\nabla_{w^\mathsf{T}} Q(\boldsymbol{w}_{n-1}; \gamma(\boldsymbol{\sigma}), h_{\boldsymbol{\sigma}}) \,|\, \boldsymbol{w}_{n-1}\right) \\ &\neq \frac{1}{N}\sum_{m=0}^{N-1} \nabla_{w^\mathsf{T}} Q(w_{n-1}; \gamma(m), h_m) \\ &= \nabla_{w^\mathsf{T}} P(w_{n-1}) \end{aligned} \tag{18.27}$$

where the first line is not equal to the second line because $\boldsymbol{\sigma}$ cannot be selected uniformly with probability $1/N$ when conditioned on $\boldsymbol{w}_{n-1}$. This is due to the fact that knowledge of $\boldsymbol{w}_{n-1}$ carries with it information about the samples that were selected in the previous iterations leading to $\boldsymbol{w}_{n-1}$. As a result, the gradient noise process under random reshuffling is *biased*. For this reason, we will need to adjust the convergence arguments for algorithms employing random reshuffling in comparison to uniform sampling.

A different conclusion holds for mini-batch implementations where the $B > 1$ samples are selected randomly *without* replacement. In this case, the zero mean property for the gradient noise will continue to hold. To see this, observe first that collecting $B-$samples sequentially, one at a time without replacement, is equivalent to choosing B data points at once from the original $N-$long data set. The number of possible choices for this mini-batch of data is given by the combinatorial expression:

$$C_N^B \triangleq \binom{N}{B} = \frac{N!}{B!(N-B)!} \triangleq L \tag{18.28}$$

which we are denoting by L. We number the L possible choices for the mini-batch by $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_L$, and each one of them can be selected with equal probability $1/L$. Assuming that some random mini-batch ℓ is selected at iteration n, we can write

$$\begin{aligned}
\mathbb{E}\left(\widehat{\nabla_{w^\mathsf{T}} P}(\boldsymbol{w}_{n-1}) \,|\, \boldsymbol{w}_{n-1}\right) &= \mathbb{E}\left(\frac{1}{B}\sum_{b\in\mathcal{B}_\ell} \nabla_{w^\mathsf{T}} Q(\boldsymbol{w}_{n-1};\boldsymbol{\gamma}(b),\boldsymbol{h}_b) \,|\, \boldsymbol{w}_{n-1}\right) \\
&= \frac{1}{L}\sum_{\ell=1}^{L}\left(\frac{1}{B}\sum_{b\in\mathcal{B}_\ell} \nabla_{w^\mathsf{T}} Q(\boldsymbol{w}_{n-1};\boldsymbol{\gamma}(b),\boldsymbol{h}_b) \,|\, \boldsymbol{w}_{n-1}\right) \\
&\overset{(a)}{=} \frac{C_{N-1}^{B-1}}{LB}\sum_{m=0}^{N-1} \nabla_{w^\mathsf{T}} Q(w_{n-1};\gamma(m),h_m) \\
&= \frac{1}{N}\sum_{m=0}^{N-1} \nabla_{w^\mathsf{T}} Q(w_{n-1};\gamma(m),h_m) \\
&= \nabla_{w^\mathsf{T}} P(w_{n-1}) \qquad (18.29)
\end{aligned}$$

where the expectation in the first line is relative to the randomness in the mini-batch selections, and step (a) uses result (18.117) from the appendix. Observe that the mini-batches $\{\mathcal{B}_\ell\}$ in the second line will generally contain some common samples. Equality (a) accounts for these repetitions and rewrites the equality only in terms of the original samples within $0 \le m \le N-1$ without any repetitions. We therefore conclude that the gradient noise process continues to have zero mean in this case.

Importance sampling

Under importance sampling, the scaling by $1/Np_n$ of the gradient approximation renders the search directions unbiased since

$$\begin{aligned}
\mathbb{E}\left(\widehat{\nabla_{w^\mathsf{T}} P}(\boldsymbol{w}_{n-1}) \,|\, \boldsymbol{w}_{n-1}\right) &= \mathbb{E}\left(\frac{1}{Np_{\boldsymbol{\sigma}}}\nabla_{w^\mathsf{T}} Q(\boldsymbol{w}_{n-1};\gamma(\boldsymbol{\sigma}),h_{\boldsymbol{\sigma}}) \,|\, \boldsymbol{w}_{n-1}\right) \\
&\overset{(a)}{=} \sum_{\sigma=0}^{N-1} p_\sigma\left(\frac{1}{Np_\sigma}\nabla_{w^\mathsf{T}} Q(w_{n-1};\gamma(\sigma),h_\sigma)\right) \\
&= \frac{1}{N}\sum_{\sigma=0}^{N-1} \nabla_{w^\mathsf{T}} Q(w_{n-1};\gamma(\sigma),h_\sigma) \\
&= \nabla_{w^\mathsf{T}} P(w_{n-1}) \qquad (18.30)
\end{aligned}$$

where in step (a) we used the fact that each $(\gamma(\boldsymbol{\sigma}), h_{\boldsymbol{\sigma}})$ is selected with probability $p_{\boldsymbol{\sigma}}$. The same unbiasedness result holds for the mini-batch version.

Streaming data

Under stochastic risk minimization, the data samples stream in *independently* of each other. As a result, the approximate search direction continues to be unbiased

conditioned on the prior weight iterate since now

$$\begin{aligned}\mathbb{E}\left(\widehat{\nabla_{w^\mathsf{T}} P}(\boldsymbol{w}_{n-1})\,|\,\boldsymbol{w}_{n-1}\right) &= \mathbb{E}\left(\nabla_{w^\mathsf{T}} Q(\boldsymbol{w}_{n-1};\boldsymbol{\gamma}(n),\boldsymbol{h}_n)\,|\,\boldsymbol{w}_{n-1}\right)\\ &\overset{(a)}{=} \nabla_{w^\mathsf{T}}\left(\mathbb{E}\,Q(\boldsymbol{w}_{n-1};\boldsymbol{\gamma}(n),\boldsymbol{h}_n)\,|\,\boldsymbol{w}_{n-1}\right)\\ &\overset{(b)}{=} \nabla_{w^\mathsf{T}}\,\mathbb{E}\,Q(w_{n-1};\boldsymbol{\gamma}(n),\boldsymbol{h}_n)\\ &\overset{(18.2b)}{=} \nabla_{w^\mathsf{T}} P(w_{n-1}) \qquad (18.31)\end{aligned}$$

Step (a) switches the order of the expectation and differentiation operators which, as explained earlier, is possible in most cases of interest since the loss $Q(w;\cdot,\cdot)$ and its gradient will generally be continuous functions of w. Step (b) is because the samples $(\boldsymbol{\gamma}(n),\boldsymbol{h}_n)$ are independent over time and therefore independent of $\boldsymbol{w}_{n-1}$ (which is a function of previous data samples). The conditioning on $\boldsymbol{w}_{n-1}$ that appears in step (a) can therefore be removed in step (b). It follows that the gradient noise has zero mean conditioned on $\boldsymbol{w}_{n-1}$ and result (18.23) continues to hold.

18.3.3 Second-Order Moment

We examine next the second-order moment of the gradient noise process under different sampling procedures and verify that it satisfies

$$\mathbb{E}\left(\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2\,|\,\boldsymbol{w}_{n-1}\right) \le \beta_g^2\|\widetilde{w}_{n-1}\|^2 + \sigma_g^2 \qquad (18.32)$$

for some constants (β_g^2,σ_g^2) independent of $\widetilde{w}_{n-1}$.

Sampling with replacement

Consider first the case of an instantaneous gradient approximation where a single sample is chosen at each iteration n. Let $\boldsymbol{\sigma}$ denote the index of the random data sample selected at that iteration. The squared Euclidean norm of the gradient noise is given by

$$\begin{aligned}\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2 &\triangleq \left\|\widehat{\nabla_{w^\mathsf{T}} P}(\boldsymbol{w}_{n-1}) - \nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-1})\right\|^2\\ &= \left\|\nabla_{w^\mathsf{T}} Q(\boldsymbol{w}_{n-1};\gamma(\boldsymbol{\sigma}),h_{\boldsymbol{\sigma}}) - \nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-1})\right\|^2\\ &\overset{(a)}{=} \|\nabla_{w^\mathsf{T}} Q(\boldsymbol{w}_{n-1}) - \nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-1})\|^2 \qquad (18.33)\end{aligned}$$

we are removing the data argument $(\gamma(\boldsymbol{\sigma}),h_{\boldsymbol{\sigma}})$ from $Q(w;\cdot,\cdot)$ in step (a) to simplify the notation. Adding and subtracting the same term $\nabla_{w^\mathsf{T}} Q(w^\star)$ gives

$$
\begin{aligned}
&\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2 \\
&= \|\nabla_{w^\mathsf{T}} Q(\boldsymbol{w}_{n-1}) - \nabla_{w^\mathsf{T}} Q(w^\star) + \nabla_{w^\mathsf{T}} Q(w^\star) - \nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-1})\|^2 \\
&\overset{(b)}{\leq} 2\,\|\nabla_{w^\mathsf{T}} Q(\boldsymbol{w}_{n-1}) - \nabla_{w^\mathsf{T}} Q(w^\star) - \nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-1})\|^2 + 2\,\|\nabla_{w^\mathsf{T}} Q(w^\star)\|^2 \\
&\overset{(c)}{\leq} 4\,\|\nabla_{w^\mathsf{T}} Q(\boldsymbol{w}_{n-1}) - \nabla_{w^\mathsf{T}} Q(w^\star)\|^2 + 2\,\|\nabla_{w^\mathsf{T}} Q(w^\star)\|^2 + \\
&\quad 4\,\|\nabla_{w^\mathsf{T}} P(w^\star) - \nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-1})\|^2 \\
&\overset{(18.10b)}{\leq} 4\delta^2\|\widetilde{\boldsymbol{w}}_{n-1}\|^2 + 4\delta^2\|\widetilde{\boldsymbol{w}}_{n-1}\|^2 + 2\,\|\nabla_{w^\mathsf{T}} Q(w^\star)\|^2
\end{aligned}
\tag{18.34}
$$

In step (b), we applied Jensen inequality $\|a+b\|^2 \leq 2\|a\|^2+2\|b\|^2$ for any vectors (a,b), and in step (c) we added $\nabla_{w^\mathsf{T}} P(w^\star) = 0$ and applied Jensen inequality again. Conditioning on $\boldsymbol{w}_{n-1}$ and taking expectations over the randomness in data selection, we conclude that (18.32) holds for the following parameters:

$$
\begin{aligned}
\beta_g^2 &= 8\delta^2 && \text{(18.35a)} \\
\sigma_g^2 &= 2\,\mathbb{E}\left(\|\nabla_{w^\mathsf{T}} Q(w^\star;\gamma(\boldsymbol{\sigma}),h_{\boldsymbol{\sigma}})\|^2\right) \\
&\overset{(18.24)}{=} \frac{2}{N}\sum_{\sigma=0}^{N-1}\left\|\nabla_{w^\mathsf{T}} Q(w^\star;\gamma(\sigma),h_\sigma)\right\|^2 && \text{(18.35b)}
\end{aligned}
$$

If desired, the value for β_g^2 can be tightened to $\beta_g^2 = 2\delta^2$ — see Prob. 18.3. It is sufficient for our purposes to know that a bound of the form (18.32) exists; the specific values for the parameters $\{\beta_g^2, \sigma_g^2\}$ are not relevant at this stage.

For mini-batch implementations, the gradient noise process continues to satisfy relation (18.32) albeit with the parameters (β_g^2, σ_g^2) scaled by B. Indeed, since the data points $\{\boldsymbol{\gamma}(b), \boldsymbol{h}_b\}$ are now sampled with replacement and are independent of each other, we have

$$
\begin{aligned}
&\mathbb{E}\left(\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2 \,|\, \boldsymbol{w}_{n-1}\right) \\
&\triangleq \mathbb{E}\left(\left\|\widehat{\nabla_{w^\mathsf{T}} P}(\boldsymbol{w}_{n-1}) - \nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-1})\right\|^2 \,\Big|\, \boldsymbol{w}_{n-1}\right) \\
&= \mathbb{E}\left(\left\|\frac{1}{B}\sum_{b=0}^{B-1}\nabla_{w^\mathsf{T}} Q(\boldsymbol{w}_{n-1};\boldsymbol{\gamma}(b),\boldsymbol{h}_b) - \nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-1})\right\|^2 \,\Big|\, \boldsymbol{w}_{n-1}\right) \\
&= \mathbb{E}\left\|\frac{1}{B}\sum_{b=0}^{B-1}\Big(\nabla_{w^\mathsf{T}} Q(w_{n-1};\boldsymbol{\gamma}(b),\boldsymbol{h}_b) - \nabla_{w^\mathsf{T}} P(w_{n-1})\Big)\right\|^2 \\
&\overset{(a)}{\leq} \frac{1}{B^2}\sum_{b=0}^{B-1}\mathbb{E}\,\|\nabla_{w^\mathsf{T}} Q(w_{n-1};\boldsymbol{\gamma}(b),\boldsymbol{h}_b) - \nabla_{w^\mathsf{T}} P(w_{n-1})\|^2
\end{aligned}
\tag{18.36}
$$

where step (a) follows from the triangle inequality of norms. Using the same bound that would result from argument (18.34) we then get:

$$\mathbb{E}\left(\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2 \,|\, \boldsymbol{w}_{n-1}\right) \le \frac{1}{B^2}\sum_{b=0}^{B-1}\left(\beta_g^2\|\widetilde{w}_{n-1}\|^2 + \sigma_g^2\right)$$
$$= \frac{1}{B}\left(\beta_g^2\|\widetilde{w}_{n-1}\|^2 + \sigma_g^2\right) \tag{18.37}$$

and step (a) uses the same bound that would result from argument (18.34).

Sampling without replacement

If we repeat the same argument for the implementation with instantaneous gradient approximation, we will similarly find that the same relation (18.32) continues to hold albeit with parameters

$$\beta_g^2 = 8\delta^2 \tag{18.38a}$$
$$\sigma_g^2 = 2\,\mathbb{E}\left(\|\nabla_{w^{\mathsf{T}}} Q(w^\star;\gamma(\boldsymbol{\sigma}),h_{\boldsymbol{\sigma}}\|^2 \,|\, \boldsymbol{w}_{n-1}\right) \tag{18.38b}$$

where the expression for σ_g^2 involves an inconvenient conditioning on $\boldsymbol{w}_{n-1}$. We can remove the conditioning as follows. Let $\boldsymbol{\sigma}$ denote the data index selected at iteration n. We know that $\boldsymbol{\sigma}$ is not necessarily chosen uniformly when we condition on $\boldsymbol{w}_{n-1}$ due to data sampling with replacement. Let us introduce, for the sake of argument, the conditional probabilities:

$$\kappa_m \triangleq \mathbb{P}\left(\boldsymbol{\sigma} = m \,|\, \boldsymbol{w}_{n-1}\right), \quad \kappa_m \ge 0, \quad \sum_{m=0}^{N-1}\kappa_m = 1 \tag{18.39}$$

That is, κ_m is the likelihood of selecting index $\boldsymbol{\sigma} = m$ at iteration n conditioned on knowledge of $\boldsymbol{w}_{n-1}$. Then, substituting into (18.38b), we get

$$\sigma_g^2 = 2\sum_{m=0}^{N-1}\kappa_m\,\|\nabla_{w^{\mathsf{T}}} Q(w^\star;\gamma(m),h_m)\|^2 \;\le\; 2\sum_{m=0}^{N-1}\|\nabla_{w^{\mathsf{T}}} Q(w^\star;\gamma(m),h_m\|^2 \tag{18.40}$$

which is independent of $\boldsymbol{w}_{n-1}$. This result can be used as the expression for σ_g^2 in (18.32).

A similar conclusion holds for the mini-batch gradient approximation where the B samples are randomly selected *without* replacement. To establish this result, we need to appeal to the auxiliary Lemma 18.1 from the appendix. First, for any iterate value w_{n-1}, we introduce the auxiliary vectors:

$$x_m \triangleq \nabla_{w^{\mathsf{T}}} Q\left(w_{n-1};\gamma(m),h_m\right) - \nabla_{w^{\mathsf{T}}} P(w_{n-1}), \quad m = 0,1,\ldots,N-1 \tag{18.41}$$

It is clear from the definition of the empirical risk function (18.2a) that

$$\frac{1}{N}\sum_{m=0}^{N-1} x_m = 0 \tag{18.42}$$

which means that the vectors $\{x_m\}$ satisfy condition (18.129) required by the lemma. At iteration n, the mini-batch implementation selects B vectors $\{\boldsymbol{x}_b\}$ at random without replacement. Let

$$\boldsymbol{g}_n(\boldsymbol{w}_{n-1}) = \frac{1}{B}\sum_{b=0}^{B-1}\boldsymbol{x}_b \tag{18.43}$$

Then, result (18.130) in the appendix implies that

$$\begin{aligned}
&\mathbb{E}\left(\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2 \,|\, \boldsymbol{w}_{n-1}\right) \\
&= \frac{1}{B^2}\frac{B(N-B)}{N(N-1)}\sum_{m=0}^{N-1}\|\nabla_w^{\mathsf{T}} Q(w_{n-1};\gamma(m),h_m) - \nabla_{w^{\mathsf{T}}} P(w_{n-1})\|^2 \\
&\overset{(a)}{\leq} \frac{1}{B}\frac{(N-B)}{N(N-1)}\sum_{m=0}^{N-1}\Big(8\delta^2\|\widetilde{w}_{n-1}\|^2 + 2\,\|\nabla_{w^{\mathsf{T}}} Q(w^\star;\gamma(m),h_m)\|^2\Big) \\
&\overset{(b)}{=} \frac{1}{B}\frac{(N-B)}{(N-1)}\left(\beta_g^2\|\widetilde{w}_{n-1}\|^2 + \sigma_g^2\right)
\end{aligned} \tag{18.44}$$

where step (a) uses the same bound that would result from argument (18.34), and step (b) uses (18.35a)-(18.35b). We can group the results for the mini-batch implementations under sampling with and without replacement into a single statement as follows:

$$\mathbb{E}\left(\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2 \,|\, \boldsymbol{w}_{n-1}\right) \;\leq\; \frac{1}{\tau_B}\left(\beta_g^2\|\widetilde{w}_{n-1}\|^2 + \sigma_g^2\right) \tag{18.45}$$

where the factor τ_B is chosen as:

$$\tau_B \;\triangleq\; \begin{cases} B, & \text{when samples are selected } \textit{with} \text{ replacement} \\ B\dfrac{N-1}{N-B}, & \text{when samples are selected } \textit{without} \text{ replacement} \end{cases} \tag{18.46}$$

Observe from the second line that $\tau_B \approx B$ for N large enough.

Importance sampling

Under importance sampling, the same bound (18.32) holds for both cases of instantaneous and mini-batch gradient approximations, as can be seen from the following argument.

For the instantaneous gradient implementations we have:

$$
\begin{aligned}
&\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2 \qquad (18.47)\\
&\triangleq \left\|\widehat{\nabla_{w^\mathsf{T}} P}(\boldsymbol{w}_{n-1}) - \nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-1})\right\|^2\\
&= \left\|\frac{1}{Np_n}\nabla_{w^\mathsf{T}} Q(\boldsymbol{w}_{n-1};\boldsymbol{\gamma}(n),\boldsymbol{h}_n) - \nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-1})\right\|^2\\
&\overset{(a)}{=} \left\|\frac{1}{Np_n}\nabla_{w^\mathsf{T}} Q(\boldsymbol{w}_{n-1}) - \nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-1})\right\|^2\\
&\overset{(b)}{=} \left\|\frac{1}{Np_n}\nabla_{w^\mathsf{T}} Q(\boldsymbol{w}_{n-1}) - \frac{1}{Np_n}\nabla_{w^\mathsf{T}} Q(w^\star) + \frac{1}{Np_n}\nabla_{w^\mathsf{T}} Q(w^\star) - \nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-1})\right\|^2
\end{aligned}
$$

where step (a) removes the data arguments $(\boldsymbol{\gamma}(n),\boldsymbol{h}_n)$ from $\nabla_{w^\mathsf{T}} Q(w;\cdot,\cdot)$ for convenience, and step (b) adds and subtracts the same quantity $\nabla_{w^\mathsf{T}} Q(w^\star)$. We therefore get

$$
\begin{aligned}
&\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2\\
&\le 2\left\|\frac{1}{Np_n}\nabla_{w^\mathsf{T}} Q(\boldsymbol{w}_{n-1}) - \frac{1}{Np_n}\nabla_{w^\mathsf{T}} Q(w^\star) - \nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-1})\right\|^2 +\\
&\quad 2\left\|\frac{1}{Np_n}\nabla_{w^\mathsf{T}} Q(w^\star)\right\|^2\\
&\overset{(c)}{\le} 4\left\|\frac{1}{Np_n}\nabla_{w^\mathsf{T}} Q(\boldsymbol{w}_{n-1}) - \frac{1}{Np_n}\nabla_{w^\mathsf{T}} Q(w^\star)\right\|^2 +\\
&\quad 4\left\|\nabla_{w^\mathsf{T}} P(w^\star) - \nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-1})\right\|^2 + 2\left\|\frac{1}{Np_n}\nabla_{w^\mathsf{T}} Q(w^\star)\right\|^2\\
&\overset{(18.10b)}{\le} 4\delta^2\|\widetilde{\boldsymbol{w}}_{n-1}\|^2 + 4\left\|\frac{1}{Np_n}\nabla_{w^\mathsf{T}} Q(\boldsymbol{w}_{n-1}) - \frac{1}{Np_n}\nabla_{w^\mathsf{T}} Q(w^\star)\right\|^2 +\\
&\quad 2\left\|\frac{1}{Np_n}\nabla_{w^\mathsf{T}} Q(w^\star)\right\|^2 \qquad (18.48)
\end{aligned}
$$

where in step (c) we applied Jensen inequality $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ for any two vectors (a,b) and added $\nabla_{w^\mathsf{T}} P(w^\star) = 0$. Next, we need to condition on $\boldsymbol{w}_{n-1}$ and take expectations. For that purpose, we note that

$$
\begin{aligned}
\mathbb{E}\left\{2\left\|\frac{1}{Np_n}\nabla_{w^\mathsf{T}} Q(w^\star)\right\|^2 \,\Big|\, \boldsymbol{w}_{n-1}\right\} &= \sum_{m=0}^{N-1} p_m \frac{2}{N^2 p_m^2}\left\|\nabla_{w^\mathsf{T}} Q(w^\star;\gamma(m),h_m)\right\|^2\\
&= \frac{2}{N^2}\sum_{m=0}^{N-1}\frac{1}{p_m}\left\|\nabla_{w^\mathsf{T}} Q(w^\star;\gamma(m),h_m)\right\|^2
\end{aligned} \qquad (18.49)
$$

while

$$\begin{aligned}
&\mathbb{E}\left\{4\left\|\frac{1}{Np_n}\nabla_{w^{\mathsf{T}}}Q(\boldsymbol{w}_{n-1})-\frac{1}{Np_n}\nabla_{w^{\mathsf{T}}}Q(w^\star)\right\|^2\Big|\,\boldsymbol{w}_{n-1}\right\}\\
&=4\sum_{m=0}^{N-1}p_m\frac{1}{N^2p_m^2}\left\|\nabla_{w^{\mathsf{T}}}Q(w_{n-1};\gamma(m),h_m)-\nabla_{w^{\mathsf{T}}}Q(w^\star)\right\|^2\\
&\overset{(18.10b)}{\leq}\frac{4\delta^2}{N^2}\sum_{m=0}^{N-1}\frac{1}{p_m}\|\widetilde{w}_{n-1}\|^2
\end{aligned}\tag{18.50}$$

Substituting into (18.48) we conclude that

$$\mathbb{E}\left(\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2\,|\,\boldsymbol{w}_{n-1}\right)\;\leq\;\beta_g^2\|\widetilde{w}_{n-1}\|^2+\sigma_g^2\tag{18.51}$$

where the parameters $\{\beta_g^2,\sigma_g^2\}$ are given by

$$\beta_g^2=4\delta^2\left(1+\frac{1}{N^2}\sum_{m=0}^{N-1}\frac{1}{p_m}\right)\tag{18.52a}$$

$$\sigma_g^2=\frac{2}{N^2}\sum_{m=0}^{N-1}\frac{1}{p_m}\left\|\nabla_{w^{\mathsf{T}}}Q(w^\star;\gamma(m),h_m)\right\|^2\tag{18.52b}$$

A similar bound holds with the above parameters $\{\beta_g^2,\sigma_g^2\}$ divided by B for the mini-batch version — see Prob. 18.13.

Streaming data

Under stochastic risk minimization, the gradient noise process continues to satisfy relation (18.32) as can be seen from the following sequence of inequalities:

$$\begin{aligned}
&\mathbb{E}\left(\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2\,|\,\boldsymbol{w}_{n-1}\right)\\
&\triangleq\;\mathbb{E}\left(\left\|\widehat{\nabla_{w^{\mathsf{T}}}P}(\boldsymbol{w}_{n-1})-\nabla_{w^{\mathsf{T}}}P(\boldsymbol{w}_{n-1})\right\|^2\Big|\,\boldsymbol{w}_{n-1}\right)\\
&=\mathbb{E}\left(\|\nabla_{w^{\mathsf{T}}}Q(\boldsymbol{w}_{n-1};\boldsymbol{\gamma}(n),\boldsymbol{h}_n)-\nabla_{w^{\mathsf{T}}}P(\boldsymbol{w}_{n-1})\|^2\,|\,\boldsymbol{w}_{n-1}\right)\\
&\overset{(a)}{=}\mathbb{E}\left(\|\nabla_{w^{\mathsf{T}}}Q(\boldsymbol{w}_{n-1})-\nabla_{w^{\mathsf{T}}}P(\boldsymbol{w}_{n-1})\|^2\,|\,\boldsymbol{w}_{n-1}\right)\\
&\overset{(b)}{=}\mathbb{E}\left(\left\|\nabla_{w^{\mathsf{T}}}Q(\boldsymbol{w}_{n-1})-\nabla_{w^{\mathsf{T}}}Q(w^o)+\nabla_{w^{\mathsf{T}}}Q(w^o)-\nabla_{w^{\mathsf{T}}}P(\boldsymbol{w}_{n-1})\right\|^2\,|\,\boldsymbol{w}_{n-1}\right)
\end{aligned}\tag{18.53}$$

In step (a) we removed the data argument $(\boldsymbol{\gamma}(n),\boldsymbol{h}_n)$ from $Q(w;\cdot,\cdot)$ to simplify the notation, and in step (b) we added and subtracted the same quantity

$\nabla_{w^\mathsf{T}} Q(w^o)$. It follows that

$$\begin{aligned}
&\mathbb{E}\,\big(\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2\,|\,\boldsymbol{w}_{n-1}\big)\\
&\overset{(c)}{\leq} 2\,\mathbb{E}\,\big(\|\nabla_{w^\mathsf{T}} Q(\boldsymbol{w}_{n-1}) - \nabla_{w^\mathsf{T}} Q(w^o) - \nabla_{w^\mathsf{T}} P(\boldsymbol{w}_{n-1})\|^2\,|\,\boldsymbol{w}_{n-1}\big) + \\
&\quad 2\,\mathbb{E}\,\big(\|\nabla_{w^\mathsf{T}} Q(w^o)\|^2\big)\\
&\overset{(d)}{\leq} 4\,\mathbb{E}\,\big(\|\nabla_{w^\mathsf{T}} Q(\boldsymbol{w}_{n-1}) - \nabla_{w^\mathsf{T}} Q(w^o)\|^2\,|\,\boldsymbol{w}_{n-1}\big) + 2\,\mathbb{E}\,\big(\|\nabla_{w^\mathsf{T}} Q(w^o)\|^2\big) + \\
&\quad 4\,\|\nabla_{w^\mathsf{T}} P(w^o) - \nabla_{w^\mathsf{T}} P(w_{n-1})\|^2 + \\
&\overset{(18.14)}{\leq} 4\delta^2\|\widetilde{w}_{n-1}\|^2 + 4\delta^2\|\widetilde{w}_{n-1}\|^2 + 2\,\mathbb{E}\,\|\nabla_{w^\mathsf{T}} Q(w^o)\|^2
\end{aligned} \tag{18.54}$$

In step (c) we applied Jensen inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ for any vectors (a, b), and in step (d) we added $\nabla_{w^\mathsf{T}} P(w^o) = 0$ and applied Jensen inequality again. We conclude that relation (18.32) holds with parameters:

$$\beta_g^2 = 8\delta^2 \tag{18.55a}$$

$$\sigma_g^2 = 2\,\mathbb{E}\,\Big(\|\nabla_{w^\mathsf{T}} Q(w^o;\boldsymbol{\gamma},\boldsymbol{h})\|^2\Big) \tag{18.55b}$$

where the expectation in σ_g^2 is over the joint distribution of the data $\{\boldsymbol{\gamma}, \boldsymbol{h}\}$.

Remark 18.1. (Variance-reduced techniques) We will discover through future expressions (19.18a) and (19.26) that the constant factor σ_g^2 in (18.32) is a source of performance degradation. It will prevent the iterates $\boldsymbol{w}_n$ from converging exactly to $w^\star$. In future Chapter 22 we will introduce the class of *variance-reduced* algorithms for empirical risk minimization. These algorithms adjust the gradient approximation in such a way that the constant driving term, σ_g^2, will end up disappearing from the variance expression (18.32). By doing so, we will be able to recover the exact convergence of $\boldsymbol{w}_n$ to $w^\star$.

■

Remark 18.2. (Bound on second-order moment for gradient noise) The derivation of the bound (18.32) relied almost exclusively on the assumption that the loss function has $\delta-$Lipschitz gradients either in the deterministic sense (18.10b) or in the mean-square sense (18.13b). The convergence analyses in future chapters will continue to hold if one assumes (or imposes) from the start that the gradient noise satisfies (18.32) for some nonnegative constants (β_g^2, σ_g^2).

■

18.4 NONSMOOTH RISK FUNCTIONS

We consider next the case of nonsmooth risk functions where gradient vectors are replaced by subgradients and these are in turn approximated by using either instantaneous or mini-batch versions, say as,

$$(\textbf{instantaneous}):\ \widehat{s}(w) = s_Q(w;\boldsymbol{\gamma},\boldsymbol{h}) \tag{18.56a}$$

$$(\textbf{mini-batch}):\ \widehat{s}(w) = \frac{1}{B}\sum_{b=0}^{B-1} s_Q(w;\boldsymbol{\gamma}(b),\boldsymbol{h}_b) \tag{18.56b}$$

In this notation, $s(w)$ denotes a subgradient construction for $P(w)$, and $s_Q(w;\boldsymbol{\gamma},\boldsymbol{h})$ refers to a subgradient of the loss function at the same location. For example, for the empirical risk minimization case (18.2a), a subgradient construction for $P(w)$ can be chosen as

$$s(w) = \frac{1}{N}\sum_{m=0}^{N-1} s_Q(w;\gamma(m),h_m) \tag{18.57}$$

in terms of individual subgradients of the loss function parameterized by the data points $(\gamma(m),h_m)$. The instantaneous approximation (18.56a) selects one subgradient vector at random, while the mini-batch approximation (18.56b) selects B subgradient vectors at random. The difference between the original subgradient construction and its approximation is again called the *gradient noise*:

$$\boldsymbol{g}(w) \;\triangleq\; \widehat{s}(w) - s(w) \tag{18.58}$$

The following two relations highlight the difference between the original subgradient method and its stochastic version:

$$(\textbf{subgradient method}) : w_n = w_{n-1} - \mu s(w_{n-1}) \tag{18.59a}$$

$$\begin{aligned}(\textbf{stochastic version}) : \boldsymbol{w}_n &= \boldsymbol{w}_{n-1} - \mu\widehat{s}(\boldsymbol{w}_{n-1}) \\ &= \boldsymbol{w}_{n-1} - \mu s(\boldsymbol{w}_{n-1}) - \mu\,\boldsymbol{g}(\boldsymbol{w}_{n-1})\end{aligned} \tag{18.59b}$$

Example 18.2 (**Gradient noise for a nonsmooth quadratic risk**) We illustrate the form of the gradient noise process for two ℓ_1-regularized quadratic risks: an empirical risk and a stochastic risk. Consider first the empirical case:

$$P(w) \;=\; \alpha\|w\|_1 \;+\; \frac{1}{N}\sum_{m=0}^{N-1}(\gamma(m) - h_m^{\mathsf{T}}w)^2 \tag{18.60}$$

Subgradient constructions for $P(w)$ and its loss function can be chosen as

$$s_Q(w;\boldsymbol{\gamma}(n),\boldsymbol{h}_n) = \alpha\,\mathrm{sign}(w) - 2\boldsymbol{h}_n(\boldsymbol{\gamma}(n) - \boldsymbol{h}_n^{\mathsf{T}}w) \tag{18.61a}$$

$$s(w) = \alpha\,\mathrm{sign}(w) - \frac{2}{N}\sum_{m=0}^{N-1} h_m(\gamma(m) - h_m^{\mathsf{T}}w) \tag{18.61b}$$

with the resulting gradient noise vector given by

$$\boldsymbol{g}(w) = \frac{2}{N}\sum_{m=0}^{N-1} h_m(\gamma(m) - h_m^{\mathsf{T}}w) \;-\; 2\boldsymbol{h}_n(\boldsymbol{\gamma}(n) - \boldsymbol{h}_n^{\mathsf{T}}w) \tag{18.62}$$

Observe that $\boldsymbol{g}(w)$ depends on the data $(\boldsymbol{\gamma}(n),\boldsymbol{h}_n)$ and, hence, as explained before, we could have written instead $\boldsymbol{g}_n(w)$ to highlight this dependency. Consider next the stochastic risk

$$\begin{aligned}P(w) &= \alpha\|w\|_1 \;+\; \mathbb{E}\,(\boldsymbol{\gamma} - \boldsymbol{h}^{\mathsf{T}}w)^2 \\ &= \alpha\|w\|_1 + \sigma_\gamma^2 - 2r_{h\gamma}^{\mathsf{T}}w + w^{\mathsf{T}}R_h w\end{aligned} \tag{18.63}$$

Subgradient constructions for $P(w)$ and its loss function can be chosen as

$$s(w) = \alpha\,\mathrm{sign}(w) - 2(r_{h\gamma} - R_h w) \tag{18.64a}$$

$$s_Q(w;\boldsymbol{\gamma}(n),\boldsymbol{h}_n) = \alpha\,\mathrm{sign}(w) - 2\boldsymbol{h}_n(\boldsymbol{\gamma}(n) - \boldsymbol{h}_n^{\mathsf{T}} w) \tag{18.64b}$$

with the resulting gradient noise vector given by

$$\boldsymbol{g}(w) = 2(r_{h\gamma} - R_h w) - 2\boldsymbol{h}_n(\boldsymbol{\gamma}(n) - \boldsymbol{h}_n^{\mathsf{T}} w) \tag{18.65}$$

Observe again that $\boldsymbol{g}(w)$ depends on n.

We are again interested in characterizing the first and second-order moments of the gradient noise process. For this purpose, we describe below the conditions that are normally imposed on the risk and loss functions, $P(w)$ and $Q(w,\cdot)$. The conditions listed here are satisfied by several risk and loss functions of interest, as illustrated in the next example and in the problems at the end of the chapter.

Empirical risks

Consider initially the case of nonsmooth empirical risks of the form (18.2a). We will assume that the risk and loss functions satisfy the two conditions listed below. Compared with the earlier conditions (14.28a)–(14.28b), we see that we are now taking the loss function into consideration since its subgradients are the ones used in the stochastic approximation implementation:

(1) (**Strongly convex risk**). $P(w)$ is $\nu-$strongly convex, namely, for every $w_1, w_2 \in \mathrm{dom}(P)$, there exists a subgradient $s(w_1)$ relative to w^{T} such that

$$P(w_2) \;\geq\; P(w_1) \;+\; (s(w_1))^{\mathsf{T}}(w_2 - w_1) \;+\; \frac{\nu}{2}\|w_2 - w_1\|^2 \tag{18.66a}$$

for some $\nu > 0$.

(2) (**Affine-Lipschitz loss subgradients**). The loss function $Q(w,\cdot)$ is convex over w and its subgradients are affine-Lipschitz, i.e., there exist nonnegative constants $\{\delta, \delta_2\}$ such that, independently of the data samples,

$$\|s_Q(w_2;\gamma(\ell),h_\ell) - s'_Q(w_1;\gamma(k),h_k)\| \leq \delta\|w_2 - w_1\| + \delta_2 \tag{18.66b}$$

for any $w_1, w_2 \in \mathrm{dom}(Q)$, any indexes $0 \leq \ell, k \leq N-1$, and for *any* subgradients:

$$s'_Q(w;\gamma,h) \in \partial_{w^{\mathsf{T}}}\, Q(w;\gamma,h) \tag{18.67}$$

Observe that condition (18.66b) is stated in terms of the *particular* subgradient construction $s_Q(w;\gamma,w)$ used by the stochastic implementation and *any* of the subgradients $s'_Q(w;\gamma,h)$ from the subdifferential set of $Q(w;\gamma,h)$. For later use, it is useful to note that condition (18.66b) implies the following relation, which involves the squared norms as opposed to the actual norms:

$$\|s_Q(w_2;\gamma(\ell),h_\ell) - s'_Q(w_1;\gamma(k),h_k)\|^2 \;\leq\; 2\delta^2\|w_2 - w_1\| + 2\delta_2^2 \tag{18.68}$$

This result follows from the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ for any (a,b).

Based on the explanation given in Example 8.6 on the subdifferential for sums of convex functions, we characterize all subgradients for $P(w)$ by writing

$$s'(w) = \frac{1}{N}\sum_{m=0}^{N-1} s'_Q(w;\gamma(m),h_m) \tag{18.69}$$

in terms of any subdifferential $s'_Q(w,\cdot)$ for $Q(w,\cdot)$. It readily follows from the triangle inequality of norms that subgradient vectors for the risk function $P(w)$ also satisfy affine-Lipschitz conditions, namely, for all $w_1, w_2 \in \mathrm{dom}(P)$:

$$\|s(w_2) - s'(w_1)\| \leq \delta\|w_2 - w_1\| + \delta_2 \tag{18.70a}$$

$$\|s(w_2) - s'(w_1)\|^2 \leq 2\delta^2\|w_2 - w_1\|^2 + 2\delta_2^2 \tag{18.70b}$$

Example 18.3 **(ℓ_2-regularized hinge risk)** We illustrate condition (18.66b) by considering the following ℓ_2-regularized hinge risk, written using a subscript ℓ to index the data $\{\gamma(\ell), h_\ell\}$ instead of m:

$$P(w) = \rho\|w\|^2 \;+\; \frac{1}{N}\sum_{\ell=0}^{N-1} \max\Big\{0, 1-\gamma(\ell)h_\ell^{\mathsf{T}}w\Big\}, \quad w = \mathrm{col}\{w_m\} \in \mathbb{R}^M \tag{18.71}$$

The corresponding loss function is given by

$$Q(w;\gamma(\ell),h_\ell) \;=\; \rho\|w\|^2 \;+\; \max\Big\{0, 1-\gamma(\ell)h_\ell^{\mathsf{T}}w\Big\} \tag{18.72}$$

We know from the earlier results (8.59a) and (8.72a) how to characterize the subdifferential set of $Q(w,\gamma(\ell),h_\ell)$. Let $h_\ell = \mathrm{col}\{h_{\ell,m}\}$ denote the individual entries of h_ℓ. Then, it holds that

$$s'_Q(w;\gamma(\ell),h_\ell) = 2\rho w + \mathrm{col}\Big\{\mathbb{G}_{\gamma(\ell)h_{\ell,m}}(w_m)\Big\} \tag{18.73a}$$

where each $\mathbb{G}_\beta(z)$, for scalars (β, z), is defined by

$$\mathbb{G}_\beta(z) \;=\; \begin{cases} 0, & \beta z > 1 \\ -\beta, & \beta z < 1 \\ [-\beta_\ell, 0], & \beta z = 1, \;\; \beta > 0 \\ [0, -\beta_\ell], & \beta z = 1, \;\; \beta < 0 \end{cases} \tag{18.73b}$$

Moreover, one particular subgradient construction for $Q(w;\cdot)$ is given by

$$s_Q(w;\gamma(\ell),h_\ell) \;=\; 2\rho w - \gamma(\ell)h_\ell\,\mathbb{I}[\gamma(\ell)h_\ell w \leq 1] \tag{18.73c}$$

Using the triangle inequality of norms we get:

$$\begin{aligned}
&\|s_Q(w_2;\gamma(\ell),h_\ell) - s'_Q(w_1;\gamma(k),h_k)\| \\
&\le 2\rho\|w_2-w_1\| \;+\; \Big\|\gamma(\ell)h_\ell\,\mathbb{I}[\gamma(\ell)h_\ell^{\mathsf{T}}w_2 \le 1]\Big\| \;+\; \Big\|\mathrm{col}\Big\{\mathbb{G}_{\gamma(k)h_{k,m}}(w_m)\Big\}\Big\| \\
&\le \underbrace{2\rho}_{\triangleq\,\delta}\|w_2-w_1\| \;+\; \|\gamma(\ell)h_\ell\| + \|\gamma(k)h_k\| \\
&\le \underbrace{2\rho}_{\triangleq\,\delta}\|w_2-w_1\| \;+\; \underbrace{\max_{0\le\ell\le N-1} 2\|\gamma(\ell)h_\ell\|}_{\triangleq\,\delta_2} \\
&= \delta\|w_2-w_1\| + \delta_2
\end{aligned} \tag{18.74}$$

since $\mathbb{I}[a]$ is bounded by one and $\|\mathrm{col}\{\mathbb{G}_{\gamma(k)h_{k,m}}(w_m)\}\|$ is bounded by $\|\gamma(k)h_k\|$. Observe how the factor δ_2 arises from the *non-smooth* component in $P(w)$.

Stochastic risks

For stochastic risks of the form (18.2b), we continue to assume that $P(w)$ is ν-strongly-convex but that the loss function has subgradients that are affine-Lipschitz in the *mean-square sense*:

(1') (**Strongly convex risk**). $P(w)$ is $\nu-$strongly convex, namely, for every $w_1, w_2 \in \mathrm{dom}(P)$ there exists a subgradient vector $s(w_1)$ relative to w^{T} such that

$$P(w_2) \;\ge\; P(w_1) \;+\; (s(w_1))^{\mathsf{T}}(w_2-w_1) \;+\; \frac{\nu}{2}\|w_2-w_1\|^2 \tag{18.75a}$$

for some $\nu > 0$.

(2') (**Mean-square affine-Lipschitz loss gradients**). The loss function $Q(w,\cdot)$ is convex over w and its subgradient vectors satisfy

$$\mathbb{E}\,\|s_Q(w_2;\boldsymbol{\gamma},\boldsymbol{h}) - s'_Q(w_1;\boldsymbol{\gamma},\boldsymbol{h})\|^2 \le \delta^2\|w_2-w_1\|^2 + \delta_2^2 \tag{18.75b}$$

for any $w_1, w_2 \in \mathrm{dom}(Q)$ and for *any*

$$s'_Q(w;\gamma,h) \in \partial_{w^{\mathsf{T}}}\, Q(w;\gamma,h) \tag{18.76}$$

Again, condition (18.75b) is stated in terms of the *particular* subgradient construction $s_Q(w;\cdot,\cdot)$ used by the stochastic optimization algorithm and *any* of the possible subgradients $s'_Q(w;\cdot,\cdot)$ from the subdifferential set of $Q(w;\cdot,\cdot)$. Note from (18.75b) that

$$\begin{aligned}
&\mathbb{E}\,\|s_Q(w_2;\boldsymbol{\gamma},\boldsymbol{h}) - s'_Q(w_1;\boldsymbol{\gamma},\boldsymbol{h})\|^2 \\
&\le \delta^2\|w_2-w_1\|^2 + \delta_2^2 + 2\delta\delta_2\|w_2-w_1\| \\
&= (\delta\|w_2-w_1\| + \delta_2)^2
\end{aligned} \tag{18.77}$$

Now using the fact that for any scalar random variable $\boldsymbol{x}$ it holds that $(\mathbb{E}\,\boldsymbol{x})^2 \le \mathbb{E}\,\boldsymbol{x}^2$, we conclude that the subgradient vectors are also affine Lipschitz on average, namely,

$$\mathbb{E}\,\|s_Q(w_2;\boldsymbol{\gamma},\boldsymbol{h}) - s'_Q(w_1;\boldsymbol{\gamma},\boldsymbol{h})\| \le \delta\|w_2-w_1\| + \delta_2 \tag{18.78}$$

Moreover, the constructions

$$s(w) \stackrel{\Delta}{=} \mathbb{E}\, s_Q(w;\boldsymbol{\gamma},\boldsymbol{h}) \tag{18.79a}$$

$$s'(w) \stackrel{\Delta}{=} \mathbb{E}\, s'_Q(w;\boldsymbol{\gamma},\boldsymbol{h}) \tag{18.79b}$$

correspond to subgradient vectors for the risk function $P(w)$ and we can also conclude that they satisfy similar affine-Lipschitz conditions:

$$\|s(w_2) - s'(w_1)\| \leq \delta\|w_2 - w_1\| + \delta_2 \tag{18.80a}$$

$$\|s(w_2) - s'(w_1)\|^2 \leq \delta^2\|w_2 - w_1\|^2 + \delta_2^2 \tag{18.80b}$$

for any $w_1, w_2 \in \mathrm{dom}(P)$. Expressions (18.79a)–(18.79b) are justified by switching the order of the expectation and sub-differentiation operators to write:

$$\partial_w P(w) = \partial_w \Big(\mathbb{E}\, Q(w;\boldsymbol{\gamma},\boldsymbol{h})\Big) \stackrel{(a)}{=} \mathbb{E}\Big(\partial_w Q(w;\boldsymbol{\gamma},\boldsymbol{h})\Big) \tag{18.81}$$

Step (a) is possible under conditions that are generally valid for our cases of interest — as was already explained in Lemma 16.1. In particular, the switching is possible whenever the loss function $Q(w;\cdot)$ is convex and bounded in neighborhoods where the subgradients are evaluated.

Proof of (18.80a)–(18.80b): Note that

$$\begin{aligned}
\|s(w_2) - s'(w_1)\|^2 &= \|\mathbb{E}\, s_Q(w_2;\boldsymbol{\gamma},\boldsymbol{h}) - \mathbb{E}\, s'_Q(w_1;\boldsymbol{\gamma},\boldsymbol{h})\|^2 \\
&\leq \mathbb{E}\,\|s_Q(w_2;\boldsymbol{\gamma},\boldsymbol{h}) - s'_Q(w_1;\boldsymbol{\gamma},\boldsymbol{h})\|^2 \\
&\stackrel{(18.75b)}{\leq} \delta^2\|w_2 - w_1\|^2 + \delta_2^2 \\
&\leq \delta^2\|w_2 - w_1\|^2 + \delta_2^2 + 2\delta\delta_2\|w_2 - w_1\| \\
&\leq (\delta\|w_2 - w_1\| + \delta_2)^2
\end{aligned} \tag{18.82}$$

■

For ease of reference, we collect in Table 18.2 the main relations and conditions described so far for nonsmooth empirical and stochastic risk minimization.

18.5 GRADIENT NOISE FOR NONSMOOTH RISKS

Using the affine-Lipschitz conditions on the subgradients of the convex loss function alone, we will now derive expressions for the first and second-order moments of the gradient noise. For the instantaneous and mini-batch constructions (18.56a)–(18.56b), the gradient noise at iteration n is given by

(**instantaneous approximation**)

$$\boldsymbol{g}(\boldsymbol{w}_{n-1}) = s_Q(\boldsymbol{w}_{n-1};\boldsymbol{\gamma}(n),\boldsymbol{h}_n) - s(\boldsymbol{w}_{n-1}) \tag{18.83a}$$

Table 18.2 Main relations and conditions used for *nonsmooth* empirical and stochastic risk minimization problems.

quantity	empirical risk minimization	stochastic risk minimization
Optimization problem	$w^\star = \underset{w\in\mathbb{R}^M}{\operatorname{argmin}} \left\{ P(w) \triangleq \frac{1}{N}\sum_{m=0}^{N-1} Q(w;\gamma(m),h_m) \right\}$	$w^o = \underset{w\in\mathbb{R}^M}{\operatorname{argmin}} \left\{ P(w) \triangleq \mathbb{E}\, Q(w;\boldsymbol{\gamma},\boldsymbol{h}) \right\}$
subgradient vector	$s(w) = \frac{1}{N}\sum_{m=0}^{N-1} s_Q(w;\gamma(m),h_m)$	$s(w) = \mathbb{E}\, s_Q(w;\boldsymbol{\gamma},\boldsymbol{h})$
Instantaneous approximation	$\widehat{s}(w) = s_Q(w;\boldsymbol{\gamma}(n),\boldsymbol{h}_n)$ $(\boldsymbol{\gamma}(n),\boldsymbol{h}_n)$ selected at random	$\widehat{s}(w) = s_Q(w;\boldsymbol{\gamma}(n),\boldsymbol{h}_n)$ $(\boldsymbol{\gamma}(n),\boldsymbol{h}_n)$ streaming in
Mini-batch approximation	$\widehat{s}(w) = \frac{1}{B}\sum_{b=0}^{B-1} s_Q(w;\boldsymbol{\gamma}(b),\boldsymbol{h}_b)$ $\{\boldsymbol{\gamma}(b),\boldsymbol{h}_b\}$ selected at random	$\widehat{s}(w) = \frac{1}{B}\sum_{b=0}^{B-1} s_Q(w;\boldsymbol{\gamma}(b),\boldsymbol{h}_b)$ $\{\boldsymbol{\gamma}(b),\boldsymbol{h}_b\}$ streaming in
Conditions on risk and loss functions	(18.66a)–(18.66b) $P(w)$ $\nu-$strongly convex, $Q(w,\cdot)$ convex $s_Q(w;\gamma,h)$ affine-Lipschitz	(18.75a)– (18.75b) $P(w)$ $\nu-$strongly convex, $Q(w,\cdot)$ convex $s_Q(w;\boldsymbol{\gamma},\boldsymbol{h})$ affine-Lipschitz in mean-square sense

for instantaneous subgradient approximations or by

(**mini-batch approximation**)

$$\boldsymbol{g}(\boldsymbol{w}_{n-1}) = \frac{1}{B}\sum_{b=0}^{B-1} s_Q(\boldsymbol{w}_{n-1};\boldsymbol{\gamma}(b),\boldsymbol{h}_b) - s(\boldsymbol{w}_{n-1}) \tag{18.83b}$$

for mini-batch approximations, where $(\boldsymbol{\gamma}(n),\boldsymbol{h}_n)$ and $\{\boldsymbol{\gamma}(b),\boldsymbol{h}_b\}$ denote the random data samples used at the $n-$th iteration while updating $\boldsymbol{w}_{n-1}$ to $\boldsymbol{w}_n$. The main conclusion of this section is again to show that the second-order moment of the gradient noise is bounded in the same manner as in (18.18), i.e., as follows:

$$\boxed{\mathbb{E}\left(\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2 \,|\, \boldsymbol{w}_{n-1}\right) \le \beta_g^2\|\widetilde{w}_{n-1}\|^2 + \sigma_g^2} \tag{18.84}$$

for some nonnegative constants $\{\beta_g^2,\sigma_g^2\}$ that will be independent of the error $\widetilde{w}_{n-1} = w^\star - w_{n-1}$. More specifically, the derivations in the remainder of this section are meant to establish the following conclusion, which is similar to the statement of Lemma 18.1 for smooth risks except for the condition on the subgradients of the loss function.

Lemma 18.2. (Gradient noise under nonsmooth risks) *Consider the empirical or stochastic risk optimization problems (18.2a)–(18.2b) and assume the subgradients of the convex loss function satisfy the affine-Lipschitz conditions (18.66b) or (18.75b). The first and second-order moments of the gradient process will satisfy:*

$$\mathbb{E}\,(\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\,|\,\boldsymbol{w}_{n-1}) = 0 \tag{18.85a}$$

$$\mathbb{E}\,(\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2\,|\,\boldsymbol{w}_{n-1}) \le \beta_g^2\|\widetilde{w}_{n-1}\|^2 + \sigma_g^2 \tag{18.85b}$$

for some nonnegative constants $\{\beta_g^2,\sigma_g^2\}$ that are independent of $\widetilde{w}_{n-1}$.

Results (18.85a)–(18.85b) hold for instantaneous or 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. Again, the *only exception* is that the zero-mean property (18.85a) will not hold for the instantaneous gradient implementation when the samples are selected without replacement.

To establish properties (18.85a)–(18.85b), we proceed by examining each sampling procedure separately and then show that they all lead to the same result. We consider the zero-mean property (18.85a) first. Since the arguments are similar to what we have done in the smooth case, we will be brief.

18.5.1 First-Order Moment

We verify again, as was the case with smooth risks, that the gradient noise process has zero mean conditioned on the previous iterate, i.e.,

$$\mathbb{E}\,(\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\,|\,\boldsymbol{w}_{n-1}) \;=\; 0 \tag{18.86}$$

where the expectation is over the randomness in sample selection.

Sampling with replacement

Consider first the case of empirical risk minimization where samples are selected uniformly from the given data set. Let $\boldsymbol{\sigma}$ denote the sample index that is selected at iteration n with

$$\mathbb{P}(\boldsymbol{\sigma} = m) = \frac{1}{N}, \quad m \in \{0, 1, 2, \ldots, N-1\} \tag{18.87}$$

It follows that

$$\begin{aligned}
\mathbb{E}\left(\widehat{s}(\boldsymbol{w}_{n-1})\,|\,\boldsymbol{w}_{n-1}\right) &= \mathbb{E}\left(s_Q(\boldsymbol{w}_{n-1};\gamma(\boldsymbol{\sigma}),h_{\boldsymbol{\sigma}})\,|\,\boldsymbol{w}_{n-1}\right) \\
&= \frac{1}{N}\sum_{m=0}^{N-1} s_Q(w_{n-1};\gamma(m),h_m) \\
&\overset{(18.57)}{=} s(w_{n-1})
\end{aligned} \tag{18.88}$$

where in the first equality we used the fact that the loss function assumes each of the values $s_Q(w_{n-1};\gamma(m),h_m)$ with probability $1/N$. It follows that (18.86) holds for instantaneous subgradient approximations.

The gradient noise process continues to have zero conditional mean in the mini-batch implementation. This is because the approximate search direction is again unbiased:

$$\begin{aligned}
\mathbb{E}\left(\widehat{s}(\boldsymbol{w}_{n-1})\,|\,\boldsymbol{w}_{n-1}\right) &= \mathbb{E}\left(\frac{1}{B}\sum_{b=0}^{B-1} s_Q(\boldsymbol{w}_{n-1};\boldsymbol{\gamma}(b),\boldsymbol{h}_b)\,|\,\boldsymbol{w}_{n-1}\right) \\
&= \frac{1}{B}\sum_{b=0}^{B-1}\mathbb{E}\left(s_Q(\boldsymbol{w}_{n-1};\boldsymbol{\gamma}(b),\boldsymbol{h}_b)\,|\,\boldsymbol{w}_{n-1}\right) \\
&\overset{(a)}{=} \mathbb{E}\left(s_Q(\boldsymbol{w}_{n-1};\gamma(\boldsymbol{\sigma}),\boldsymbol{h}_{\boldsymbol{\sigma}})\,|\,\boldsymbol{w}_{n-1}\right) \\
&= \frac{1}{N}\sum_{\sigma=0}^{N-1} s_Q(w_{n-1};\gamma(\sigma),h_\sigma) \\
&\overset{(18.57)}{=} s(w_{n-1})
\end{aligned} \tag{18.89}$$

where in step (a) we used the fact that the data samples $(\boldsymbol{\gamma}(b),\boldsymbol{h}_b)$ are selected independently of each other.

Sampling without replacement

When samples are selected at random without replacement, we obtain for the instantaneous subgradient approximation:

$$\begin{aligned}\mathbb{E}\left(\widehat{s}(\boldsymbol{w}_{n-1})\,|\,\boldsymbol{w}_{n-1}\right) &= \mathbb{E}\left(s_Q(\boldsymbol{w}_{n-1};\gamma(\boldsymbol{\sigma}),h_{\boldsymbol{\sigma}})\,|\,\boldsymbol{w}_{n-1}\right)\\ &\neq \frac{1}{N}\sum_{m=0}^{N-1} s_Q(w_{n-1};\gamma(m),h_m)\\ &= s(w_{n-1})\end{aligned} \tag{18.90}$$

where the first line is not equal to the second line because, conditioned on $\boldsymbol{w}_{n-1}$, the sample index $\boldsymbol{\sigma}$ cannot be selected uniformly. As a result, the gradient noise process under random reshuffling is *biased* and does *not* have zero mean anymore.

A different conclusion holds for mini-batch implementations where the B samples in the batch are selected randomly *without* replacement. In this case, the zero mean property for the gradient noise continues to hold, as can be verified by repeating the argument that led to (18.29) in the smooth case.

Importance sampling

Under importance sampling, the instantaneous and mini-batch subgradient approximations would be scaled as follows:

$$(\textbf{instantaneous}):\ \widehat{s}(w) = \frac{1}{Np_\sigma}s_Q(w;\gamma(\boldsymbol{\sigma}),h_{\boldsymbol{\sigma}}) \tag{18.91a}$$

$$(\textbf{mini-batch}):\ \widehat{s}(w) = \frac{1}{B}\sum_{b=0}^{B-1}\frac{1}{Np_b}s_Q(w;\boldsymbol{\gamma}(b),\boldsymbol{h}_b) \tag{18.91b}$$

The scalings render these search directions unbiased so that (18.23) continues to hold. The argument is similar to the one used to establish (18.30) in the smooth case.

Streaming data

Under stochastic risk minimization, the data samples stream in *independently* of each other. As a result, the gradient noise process continues to satisfy relation (18.86) since

$$\mathbb{E}\left(s_Q(\boldsymbol{w}_{n-1};\boldsymbol{\gamma}(n),\boldsymbol{h}_n)\,|\,\boldsymbol{w}_{n-1}\right) \overset{(a)}{=} \mathbb{E}\,s_Q(w_{n-1};\boldsymbol{\gamma}(n),\boldsymbol{h}_n) \overset{(18.79a)}{=} s(w_{n-1}) \tag{18.92}$$

Step (a) is because the samples $(\boldsymbol{\gamma}(n),\boldsymbol{h}_n)$ are independent over time and, hence, they are also independent of $\boldsymbol{w}_{n-1}$ (which is a function of all previous data samples).

18.5.2 Second-Order Moment

We examine next the second-order moment of the gradient noise process under different sampling procedures and verify that it satisfies:

$$\mathbb{E}\left(\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2 \,|\, \boldsymbol{w}_{n-1}\right) \;\leq\; \beta_g^2\|\widetilde{w}_{n-1}\|^2 + \sigma_g^2 \tag{18.93}$$

for some constants (β_g^2, σ_g^2) independent of $\widetilde{w}_{n-1}$. The arguments are similar to the ones used in the smooth case and, therefore, we shall be brief again.

Sampling with replacement

Consider first the case of empirical risk minimization where samples are selected uniformly from the given dataset $\{\gamma(m), h_m\}$. It follows for instantaneous subgradient approximations that:

$$\begin{aligned}
\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2 &\triangleq \|s_Q(\boldsymbol{w}_{n-1};\gamma(\boldsymbol{\sigma}),h_{\boldsymbol{\sigma}}) - s(\boldsymbol{w}_{n-1})\|^2 \\
&\overset{(a)}{=} \|s_Q(\boldsymbol{w}_{n-1}) - s_Q(w^\star) + s_Q(w^\star) - s(\boldsymbol{w}_{n-1})\|^2 \\
&\overset{(b)}{\leq} 2\,\|s_Q(\boldsymbol{w}_{n-1}) - s_Q(w^\star) - s(\boldsymbol{w}_{n-1})\|^2 + 2\,\|s_Q(w^\star)\|^2 \\
&\overset{(c)}{\leq} 4\,\|s_Q(\boldsymbol{w}_{n-1}) - s_Q(w^\star)\|^2 + 2\,\|s_Q(w^\star)\|^2 + \\
&\quad\; 4\,\|s(\boldsymbol{w}_{n-1}) - s'(w^\star)\|^2 \\
&\overset{(18.68)}{\leq} 8\delta^2\|\widetilde{\boldsymbol{w}}_{n-1}\|^2 + 8\delta_2^2 + 8\delta^2\|\widetilde{\boldsymbol{w}}_{n-1}\|^2 + 8\delta_2^2 + 2\,\|s_Q(w^\star)\|^2
\end{aligned} \tag{18.94}$$

In step (a) we removed the data argument $(\gamma(\boldsymbol{\sigma}), h_{\boldsymbol{\sigma}})$ from $s_Q(w;\cdot)$ to simplify the notation and added and subtracted $s_Q(w^\star)$. In step (b) we applied Jensen inequality $\|a+b\|^2 \leq 2\|a\|^2+2\|b\|^2$ for any vectors (a, b), and in step (c) we added and subtracted $s'(w^\star) = 0$. We know from property (8.47) that a subgradient vector $s'(w^\star)$ for $P(w)$ exists that evaluates to zero at the minimizer. We conclude that (18.93) holds with

$$\beta_g^2 \;=\; 16\delta^2 \tag{18.95a}$$

$$\begin{aligned}
\sigma_g^2 &= 16\delta_2^2 + 2\,\mathbb{E}\,\left(\|s_Q(w^\star;\gamma(\boldsymbol{\sigma}),h_{\boldsymbol{\sigma}})\|^2 \,|\, \boldsymbol{w}_{n-1}\right) \\
&= 16\delta_2^2 + 2\,\mathbb{E}\,\|s_Q(w^\star;\gamma(\boldsymbol{\sigma}),h_{\boldsymbol{\sigma}})\|^2 \\
&\overset{(18.87)}{=} 16\delta_2^2 + \frac{2}{N}\sum_{m=0}^{N-1}\|s_Q(w^\star;\gamma(m),h_m)\|^2
\end{aligned} \tag{18.95b}$$

A similar conclusion holds for the mini-batch version with the parameters (β_g^2, σ_g^2) divided by B.

Sampling without replacement

If we repeat the same argument leading to (18.94), we will conclude that the same relation (18.93) holds albeit with parameters:

$$\beta_g^2 = 16\delta^2 \tag{18.96a}$$

$$\sigma_g^2 = 16\delta_2^2 + 2\,\mathbb{E}\left(\|\nabla_{w^\mathsf{T}} Q(w^\star;\gamma(\boldsymbol{\sigma}),h_{\boldsymbol{\sigma}}\|^2 \,|\, \boldsymbol{w}_{n-1}\right) \tag{18.96b}$$

where the expression for σ_g^2 still involves an inconvenient conditioning on $\boldsymbol{w}_{n-1}$. We can remove the conditioning as explained earlier in (18.40) to arrive at

$$\sigma_g^2 \le 16\delta_2^2 + 2\sum_{m=0}^{N-1} \|\nabla_{w^\mathsf{T}} Q(w^\star;\gamma(m),h_m\|^2 \tag{18.97}$$

which is independent of $\boldsymbol{w}_{n-1}$. This result can be used as the expression for σ_g^2 in (18.93).

A similar conclusion holds for the mini-batch subgradient approximation where the B samples are randomly selected *without* replacement. The same argument leading to (18.45) will continue to hold and lead to

$$\mathbb{E}\left(\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2 \,|\, \boldsymbol{w}_{n-1}\right) \;\le\; \frac{1}{\tau_B}\left(\beta_g^2\|\widetilde{w}_{n-1}\|^2 + \sigma_g^2\right) \tag{18.98}$$

where

$$\tau_B \;\triangleq\; \begin{cases} B, & \text{when samples are selected } \textit{with} \text{ replacement} \\ B\dfrac{N-1}{N-B}, & \text{when samples are selected } \textit{without} \text{ replacement} \end{cases} \tag{18.99}$$

We see from the second line that $\tau_B \approx B$ for N large enough.

Importance sampling

Under importance sampling, we have for implementations with instantaneous subgradient approximations:

$$\begin{aligned} &\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2 \\ &\triangleq \|\widehat{s}(\boldsymbol{w}_{n-1}) - s(\boldsymbol{w}_{n-1})\|^2 \\ &= \left\|\frac{1}{Np_n}s_Q(\boldsymbol{w}_{n-1};\boldsymbol{\gamma}(n),\boldsymbol{h}_n) - s(\boldsymbol{w}_{n-1})\right\|^2 \\ &\overset{(a)}{=} \left\|\frac{1}{Np_n}s_Q(\boldsymbol{w}_{n-1}) - s(\boldsymbol{w}_{n-1})\right\|^2 \\ &\overset{(b)}{=} \left\|\frac{1}{Np_n}s_Q(\boldsymbol{w}_{n-1}) - \frac{1}{Np_n}s_Q(w^\star) + \frac{1}{Np_n}s_Q(w^\star) - s(\boldsymbol{w}_{n-1})\right\|^2 \end{aligned} \tag{18.100}$$

where step (a) removes the data arguments $(\boldsymbol{\gamma}(n),\boldsymbol{h}_n)$ from $s_Q(w;\cdot,\cdot)$ for convenience, and step (b) adds and subtracts the same quantity $s_Q(w^\star)$. We therefore

have

$$\begin{aligned}
&\|\boldsymbol{g}(\boldsymbol{w}_{n-1})\|^2 \\
&\leq 2\left\|\frac{1}{Np_n}s_Q(\boldsymbol{w}_{n-1}) - \frac{1}{Np_n}s_Q(w^\star) - s(\boldsymbol{w}_{n-1})\right\|^2 + 2\left\|\frac{1}{Np_n}s_Q(w^\star)\right\|^2 \\
&\overset{(c)}{\leq} 4\left\|\frac{1}{Np_n}s_Q(\boldsymbol{w}_{n-1}) - \frac{1}{Np_n}s_Q(w^\star)\right\|^2 + 2\left\|\frac{1}{Np_n}s_Q(w^\star)\right\|^2 \\
&\quad 4\left\|s'(w^\star) - s(\boldsymbol{w}_{n-1})\right\|^2 \\
&\overset{(18.10b)}{\leq} 4\delta^2\|\widetilde{\boldsymbol{w}}_{n-1}\|^2 + 4\left\|\frac{1}{Np_n}s_Q(\boldsymbol{w}_{n-1}) - \frac{1}{Np_n}s_Q(w^\star)\right\|^2 + 2\left\|\frac{1}{Np_n}s_Q(w^\star)\right\|^2
\end{aligned} \tag{18.101}$$

where in step (c) we applied Jensen inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ for any two vectors (a, b), and used the fact that there exists a subgradient for $P(w)$ such that $s'(w^\star) = 0$. Next, we need to condition on $\boldsymbol{w}_{n-1}$ and take expectations. For that purpose, we note that

$$\mathbb{E}\left\{2\left\|\frac{1}{Np_n}s_Q(w^\star)\right\|^2 \,\Big|\, \boldsymbol{w}_{n-1}\right\} = \frac{2}{N^2}\sum_{m=0}^{N-1}\frac{1}{p_m}\left\|s_Q(w^\star;\gamma(m),h_m)\right\|^2 \tag{18.102}$$

and

$$\mathbb{E}\left\{\frac{4}{N^2p_n^2}\left\|s_Q(\boldsymbol{w}_{n-1}) - s_Q(w^\star)\right\|^2\Big|\,\boldsymbol{w}_{n-1}\right\} \leq \frac{8}{N^2}\sum_{m=0}^{N-1}\frac{1}{p_m}(\delta^2\|\widetilde{w}_{n-1}\|^2 + \delta_2^2) \tag{18.103}$$

Substituting into (18.101) we conclude that

$$\mathbb{E}\left(\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2 \,|\, \boldsymbol{w}_{n-1}\right) \;\leq\; \beta_g^2\|\widetilde{w}_{n-1}\|^2 + \sigma_g^2 \tag{18.104}$$

where the parameters $\{\beta_g^2, \sigma_g^2\}$ are given by

$$\beta_g^2 = 4\delta^2\left(1 + \frac{2}{N^2}\sum_{m=0}^{N-1}\frac{1}{p_m}\right) \tag{18.105a}$$

$$\sigma_g^2 = \frac{2}{N^2}\sum_{m=0}^{N-1}\frac{1}{p_m}\Big(\left\|s_Q(w^\star;\gamma(m),h_m)\right\|^2 + 4\delta_2^2\Big) \tag{18.105b}$$

A similar bound holds with the above parameters $\{\beta_g^2, \sigma_g^2\}$ divided by B for the mini-batch version — see Prob. 18.13.

Streaming data

Under stochastic risk minimization, the data samples stream in *independently* of each other. As a result, the gradient noise process continues to satisfy relation

(18.93) as can be seen from the following sequence of inequalities:

$$
\begin{aligned}
&\mathbb{E}\,\big(\|\boldsymbol{g}(\boldsymbol{w}_{n-1})\|^2\,|\,\boldsymbol{w}_{n-1}\big) \\
&\triangleq\ \mathbb{E}\,\Big(\|s_Q(\boldsymbol{w}_{n-1};\boldsymbol{\gamma}(n),\boldsymbol{h}_n)-s(\boldsymbol{w}_{n-1})\|^2\,|\,\boldsymbol{w}_{n-1}\Big) \\
&\overset{(a)}{=}\ \mathbb{E}\,\big(\|s_Q(\boldsymbol{w}_{n-1})-s_Q(w^o)+s_Q(w^o)-s(\boldsymbol{w}_{n-1})\|^2\,|\,\boldsymbol{w}_{n-1}\big) \\
&\overset{(b)}{\leq}\ 2\,\mathbb{E}\,\big(\|s_Q(\boldsymbol{w}_{n-1})-s_Q(w^o)-s(w_{n-1})\|^2\,|\,\boldsymbol{w}_{n-1}\big)+2\,\mathbb{E}\,\|s_Q(w^o)\|^2 \\
&\overset{(c)}{\leq}\ 4\,\mathbb{E}\,\big(\|s_Q(\boldsymbol{w}_{n-1})-s_Q(w^o)\|^2\,|\,\boldsymbol{w}_{n-1}\big)+2\,\mathbb{E}\,\|s_Q(w^o)\|^2 \\
&\qquad 4\,\|s(w_{n-1})-s'(w^o)\|^2 \\
&\overset{(18.75b)}{\leq}\ 4\delta^2\|\widetilde{w}_{n-1}\|^2+4\delta_2^2+4\delta^2\|\widetilde{w}_{n-1}\|^2+4\delta_2^2+2\,\mathbb{E}\,\|s_Q(w^o;\gamma,h)\|^2
\end{aligned}
\tag{18.106}
$$

In step (a) we removed the argument $(\boldsymbol{\gamma}(n),\boldsymbol{h}_n)$ from $s_Q(w;\cdot,\cdot)$ to simplify the notation and added and subtracted $s_Q(w^o)$. In step (b) we applied Jensen inequality $\|a+b\|^2 \leq 2\|a\|^2+2\|b\|^2$ for any vectors (a,b), and in step (c) we added and subtracted $s'(w^o)=0$. We know from property (8.47) that a subgradient vector $s'(w^o)$ exists for $P(w)$ that evaluates to zero at the minimizer. We conclude that a relation of the form (18.93) continues to hold with

$$\beta_g^2 = 8\delta^2 \tag{18.107a}$$

$$\sigma_g^2 = 8\delta_2^2+2\,\mathbb{E}\,\big(\|s_Q(w^o;\boldsymbol{\gamma},\boldsymbol{h})\|^2\big) \tag{18.107b}$$

where the expectation in σ_g^2 is over the joint distribution of the data $(\boldsymbol{\gamma},\boldsymbol{h})$.

18.6 COMMENTARIES AND DISCUSSION

Moments of gradient noise. We established in the body of the chapter that under certain $\delta-$Lipschitz or affine-Lipschitz conditions on the gradients or subgradients of the loss function, the second-order moment of the gradient noise process in stochastic implementations satisfies the bound

$$\mathbb{E}\,\Big(\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2\,|\,\boldsymbol{w}_{n-1}\Big) \leq \beta_g^2\|\widetilde{w}_{n-1}\|^2+\sigma_g^2 \tag{18.108}$$

for some parameters (β_g^2,σ_g^2) that are independent of the error vector, $\widetilde{w}_{n-1}$. It is verified in the problems that this bound holds for many cases of interest involving risks that arise frequently in inference and learning problems with and without regularization. The bound (18.108) is similar to conditions used earlier in the optimization literature. In Polyak (1987), the term involving $\|\widetilde{w}_{n-1}\|^2$ is termed the "relative noise component," while the term involving σ_g^2 is termed the "absolute noise component." In Polyak and Tsypkin (1973) and Bertsekas and Tsitsiklis (2000), it is assumed instead that the gradient noise satisfies a condition of the type:

$$\mathbb{E}\,\Big(\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2\,|\,\boldsymbol{w}_{n-1}\Big) \leq \alpha\Big(1+\|\nabla_{w^{\mathsf{T}}}P(w_{n-1})\|^2\Big) \tag{18.109}$$

for some positive constant α. One main difference is that (18.109) is introduced as an assumption in these works, whereas we established the validity of (18.108) in the

chapter. We can verify that, for strongly-convex risks, conditions (18.108) and (18.109) are equivalent. One direction follows from the earlier result (10.20) for $\delta-$smooth risks that

$$\frac{1}{2\delta}\|\nabla_w P(w)\|^2 \leq P(w) - P(w^\star) \leq \frac{\delta}{2}\|\widetilde{w}\|^2 \tag{18.110}$$

where $\widetilde{w} = w^\star - w$. Substituting into the right-hand side of (18.109) we get

$$\mathbb{E}\left(\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2 \,|\, \boldsymbol{w}_{n-1}\right) \leq \alpha + \alpha\delta^2\|\widetilde{w}_{n-1}\|^2 \tag{18.111}$$

The other direction follows similarly from property (8.29) for $\nu-$strongly convex functions, namely,

$$\frac{\nu}{2}\|\widetilde{w}\|^2 \leq P(w) - P(w^\star) \leq \frac{1}{2\nu}\|\nabla_w P(w)\|^2 \tag{18.112}$$

Absolute and relative noise terms. The presence of *both* relative and absolute terms in the bound (18.108) is necessary in most cases of interest — see, e.g., Chen and Sayed (2012a) and Sayed (2014a). An example to this effect is treated in Prob. 18.9. Consider the quadratic stochastic risk optimization problem:

$$w^o = \underset{w\in\mathbb{R}^M}{\operatorname{argmin}} \; \left\{\mathbb{E}\,(\boldsymbol{\gamma} - \boldsymbol{h}^{\mathsf{T}}w)^2\right\} \tag{18.113}$$

Assume the streaming data $\{\boldsymbol{\gamma}(n), \boldsymbol{h}_n\}$ arises 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. Let $R_h = \mathbb{E}\,\boldsymbol{h}_n\boldsymbol{h}_n^{\mathsf{T}} > 0$, $r_{h\gamma} = \mathbb{E}\,\boldsymbol{h}_n\boldsymbol{\gamma}(n)$, and $\sigma_v^2 = \mathbb{E}\,\boldsymbol{v}^2(n)$. Moreover, $\boldsymbol{v}(n)$ is a white-noise process that is independent of all other random variables. It is shown in the problem that $w^o = w^\bullet$, which means that the solution to the optimization problem is able to recover the underlying model $w^\bullet$. A stochastic gradient algorithm with instantaneous gradient approximation can then be used to estimate $w^\bullet$ and it is verified in the same problem that the gradient noise process in this case will satisfy

$$\mathbb{E}\left(\|\boldsymbol{g}(\boldsymbol{w}_{n-1})\|^2 | \boldsymbol{w}_{n-1}\right) \leq \beta_g^2\|\widetilde{w}_{n-1}\|^2 + \sigma_g^2 \tag{18.114}$$

where

$$\sigma_g^2 = 4\sigma_v^2\,\mathrm{Tr}(R_h), \quad \beta_g^2 = 4\mathbb{E}\,\|R_h - \boldsymbol{h}_n\boldsymbol{h}_n^{\mathsf{T}}\|^2 \tag{18.115}$$

We observe that even in this case, dealing with a quadratic risk function, the upper bound includes both relative and absolute noise terms.

Affine Lipschitz conditions. For nonsmooth risks, the affine-Lipschitz conditions (18.66b)–(18.75b) are from the work by Ying and Sayed (2018). It is customary in the literature to use a more restrictive condition that assumes the subgradient vectors $s_Q(w,\cdot)$ are uniformly bounded either in the absolute sense or in the mean-square sense depending on whether one is dealing with empirical or stochastic minimization problems — see, e.g., Bertsekas (1999), Nemirovski *et al.* (2009), Nedic and Ozdaglar (2009), Ram, Nedic, and Veeravalli (2010), Srivastava and Nedic (2011), and Agarwal *et al.* (2012). That is, in these works, it is generally imposed that

$$\|s_Q(w;\gamma,h)\| \leq G \quad \text{or} \quad \mathbb{E}\,\|s_Q(w;\boldsymbol{\gamma},\boldsymbol{h})\|^2 \leq G \tag{18.116}$$

for some constant $G \geq 0$ and for all subgradients in the subdifferential set of $Q(w,\cdot)$. We know from the result of Prob. 14.3 that the bounded subgradient assumption, $\|s_Q(w;\gamma,h\cdot)\| \leq G$, is in conflict with the $\nu-$strong convexity assumption on $P(w)$; the latter condition implies that the subgradient norm cannot be bounded. One common way to circumvent the difficulty with the bounded requirement on the subgradients

is to restrict the domain of $P(w)$ to some bounded convex set, say, $w \in \mathcal{W}$, in order to bound its subgradient vectors, and then employ a projection-based subgradient implementation. This approach can still face challenges. First, projections onto $\mathcal{W}$ may not be straightforward to perform unless the set $\mathcal{W}$ is simple enough and, second, the bound G that results on the subgradient vectors by limiting w to $\mathcal{W}$ can be loose. In our presentation, we established and adopted the more relaxed affine-Lipschitz conditions (18.66b) or (18.75b).

PROBLEMS

18.1 Consider the ℓ_2-regularized quadratic and logistic losses defined by

$$Q(w;\boldsymbol{\gamma},\boldsymbol{h}) = \begin{cases} \rho\|w\|^2 \;+\; (\boldsymbol{\gamma} - \boldsymbol{h}^{\mathsf{T}}w)^2, & \text{(quadratic)} \\ \rho\|w\|^2 \;+\; \ln\left(1 + e^{-\boldsymbol{\gamma}\boldsymbol{h}^{\mathsf{T}}w}\right), & \text{(logistic)} \end{cases}$$

Verify that these losses satisfy the mean-square $\delta-$Lipschitz condition (18.13b) for zero-mean random variables $\{\boldsymbol{\gamma},\boldsymbol{h}\}$.

18.2 Verify that the mini-batch gradient approximation (18.20) is unbiased under importance sampling conditioned on $\boldsymbol{w}_{n-1}$.

18.3 If desired, we can tighten the bound in (18.32) to $\beta_g^2 = 2\delta^2$ as follows. Use the fact that, for any scalar random variable $\boldsymbol{x}$, we have $\mathbb{E}\,(\boldsymbol{x} - \mathbb{E}\,\boldsymbol{x})^2 \leq \mathbb{E}\,\boldsymbol{x}^2$ to show that

$$\mathbb{E}\,(\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2 \,|\, \boldsymbol{w}_{n-1}) \leq \mathbb{E}\left(\|\nabla_{w^{\mathsf{T}}}\, Q(\boldsymbol{w}_{n-1})\|^2 \,|\, \boldsymbol{w}_{n-1}\right)$$

where we are not showing the data arguments of $Q(w,\cdot)$ for convenience. Conclude that $\mathbb{E}\,(\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2 \,|\, \boldsymbol{w}_{n-1}) \leq 2\delta^2\|\widetilde{w}_{n-1}\|^2 + \sigma_g^2$.

18.4 Repeat the argument that led to the second-order moment bound (18.32) for both cases of empirical and stochastic risks and establish that the fourth-order moment of the gradient noise process satisfies a similar relation, namely,

$$\mathbb{E}\left(\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^4 \,|\, \boldsymbol{w}_{n-1}\right) \;\leq\; \beta_{g4}^4\|\widetilde{w}_{n-1}\|^4 + \sigma_{g4}^4$$

for some nonnegative constants $(\beta_{g4}^4, \sigma_{g4}^4)$. Show further that if the above bound on the fourth-order moment of the gradient noise process holds, then it automatically implies that the following bound on the second-order moment also holds:

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

where $\beta_g^2 = (\beta_{g4}^4)^{1/2}$ and $\sigma_g^2 = (\sigma_{g4}^4)^{1/2}$.

18.5 Assume the bound given in Prob. 18.4 holds for the fourth-order moment of the gradient noise process generated by a stochastic gradient algorithm with instantaneous gradient approximation. Consider instead a mini-batch implementation for smooth risks. Show that the gradient noise satisfies

$$\mathbb{E}\left(\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^4 \,|\, \boldsymbol{w}_{n-1}\right) \;\leq\; \frac{C_B}{B^2}\left(\beta_{g4}^4\|\widetilde{w}_{n-1}\|^4 + \sigma_{g4}^4\right)$$

where $C_B = 3 - \frac{2}{B} \leq 3$. Conclude that a B^2-fold decrease occurs in the mean-fourth moment of the gradient noise.

18.6 Consider a stochastic gradient implementation with instantaneous gradient approximation using data sampling without replacement. Show that $\frac{1}{N}\sum_{n=0}^{N-1}\boldsymbol{g}_n(w) = 0$.

18.7 Consider a stochastic gradient implementation with instantaneous gradient approximation. Assume multiple epochs are run using random reshuffling at the start of each epoch. Show that the conditional mean of the gradient noise at the *beginning* of every $k-$th epoch satisfies $\mathbb{E}\left(\boldsymbol{g}_0(\boldsymbol{w}_{-1}^k)\,|\,\boldsymbol{w}_{-1}^k\right)=0$.

18.8 Let $\boldsymbol{\gamma}(n)$ be a streaming sequence of binary random variables assuming the values ± 1, and let $\boldsymbol{h}_n\in\mathbb{R}^M$ be a streaming sequence of real random vectors with $R_h=\mathbb{E}\,\boldsymbol{h}_n\boldsymbol{h}_n^{\mathsf{T}}>0$. Assume the random processes $\{\boldsymbol{\gamma}(n),\boldsymbol{h}_n\}$ are jointly wide-sense stationary and zero mean. Consider the regularized logistic risk function:

$$P(w)\;=\;\frac{\rho}{2}\|w\|^2\;+\;\mathbb{E}\,\ln\left(1+e^{-\boldsymbol{\gamma}\boldsymbol{h}^{\mathsf{T}}w}\right)$$

(a) Write down the expression for the gradient noise process, $\boldsymbol{g}_n(\boldsymbol{w}_{n-1})$, that would result from using a constant step-size stochastic gradient algorithm with instantaneous gradient approximation.

(b) Verify from first principles that this noise process satisfies

$$\begin{aligned}&\mathbb{E}\,(\,\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\,|\,\boldsymbol{w}_{n-1})=0\\&\mathbb{E}\,\left(\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^2\,|\,\boldsymbol{w}_{n-1}\right)\le\beta_g^2\,\|\widetilde{w}_{n-1}\|^2\;+\;\sigma_g^2\end{aligned}$$

for some nonnegative constants β_g^2 and σ_g^2.

(c) Verify also that the fourth-order moment of the gradient noise process satisfies

$$\mathbb{E}\,\left(\|\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\|^4\,|\,\boldsymbol{w}_{n-1}\right)\le\beta_{g4}^4\,\|\widetilde{w}_{n-1}\|^4\;+\;\sigma_{g4}^4$$

for some nonnegative constants β_{g4}^4 and σ_{g4}^4. What conditions on the moments of the data are needed to ensure this result?

(d) Define $R_{g,n}(\boldsymbol{w})=\mathbb{E}\,\left(\boldsymbol{g}_n(\boldsymbol{w})\boldsymbol{g}_n^{\mathsf{T}}(\boldsymbol{w})\,|\,\boldsymbol{w}_{n-1}\right)$, which denotes the conditional second-order moment of the gradient noise process. Show that

$$\begin{aligned}\left\|\nabla_w^2\;P(w^o+\Delta w)-\nabla_w^2\;P(w^o)\right\|&\le\kappa_1\,\|\Delta w\|\\\|R_{g,n}(w^o+\Delta w)-R_{g,n}(w^o)\|&\le\kappa_2\,\|\Delta w\|^{\alpha}\end{aligned}$$

for small perturbations $\|\Delta w\|\le\epsilon$ and for some constants $\kappa_1\ge 0$, $\kappa_2\ge 0$, and positive exponent α. What conditions on the moments of the data are needed to ensure these results?

18.9 Consider the quadratic stochastic risk optimization problem:

$$w^o=\underset{w\in\mathbb{R}^M}{\operatorname{argmin}}\;\mathbb{E}\,(\boldsymbol{\gamma}-\boldsymbol{h}^{\mathsf{T}}w)^2$$

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 parameter $w^{\bullet}\in\mathbb{R}^M$ and where $\boldsymbol{h}_n$ and $\boldsymbol{v}_n$ are zero-mean uncorrelated processes. Let $R_h=\mathbb{E}\,\boldsymbol{h}_n\boldsymbol{h}_n^{\mathsf{T}}>0$, $r_{h\gamma}=\mathbb{E}\,\boldsymbol{h}_n\boldsymbol{\gamma}(n)$, and $\sigma_v^2=\mathbb{E}\,\boldsymbol{v}^2(n)$. Moreover, $\boldsymbol{v}(n)$ is a white-noise process that is independent of all other random variables.

(a) Show that $w^o=w^{\bullet}$. That is, show that the optimal solution w^o is able to recover the underlying model $w^{\bullet}$.

(b) Verify that the gradient noise is $\boldsymbol{g}_n(\boldsymbol{w}_{n-1})=2(R_h-\boldsymbol{h}_n\boldsymbol{h}_n^{\mathsf{T}})\widetilde{\boldsymbol{w}}_{n-1}-2\boldsymbol{h}_n\boldsymbol{v}(n)$.

(c) Show that $\mathbb{E}\,(\boldsymbol{g}_n(\boldsymbol{w}_{n-1})\,|\,\boldsymbol{w}_{n-1})=0$.

(d) Show that $\mathbb{E}\,(\|\boldsymbol{g}(\boldsymbol{w}_{n-1})\|^2|\boldsymbol{w}_{n-1})\le\beta_g^2\|\widetilde{w}_{n-1}\|^2+\sigma_g^2$ where $\sigma_g^2=4\sigma_v^2\,\mathrm{Tr}(R_h)$ and $\beta_g^2=4\mathbb{E}\,\|R_h-\boldsymbol{h}_n\boldsymbol{h}_n^{\mathsf{T}}\|^2$.

18.10 Consider the ℓ_1-regularized quadratic and logistic losses defined by

$$Q(w;\boldsymbol{\gamma},\boldsymbol{h})=\begin{cases}\alpha\|w\|_1\;+\;(\boldsymbol{\gamma}-\boldsymbol{h}^{\mathsf{T}}w)^2, & \text{(quadratic)}\\ \alpha\|w\|_1\;+\;\ln\left(1+e^{-\boldsymbol{\gamma}\boldsymbol{h}^{\mathsf{T}}w}\right), & \text{(logistic)}\end{cases}$$

Verify that these losses satisfy the mean-square affine-Lipschitz condition (18.75b) for some (δ,δ_2). *Remark.* For a related discussion, see Ying and Sayed (2018).

18.11 Consider the ℓ_2-regularized hinge loss

$$Q(w;\boldsymbol{\gamma},\boldsymbol{h}) = \rho\|w\|^2 \;+\; \max\{0, 1-\boldsymbol{\gamma}\boldsymbol{h}^{\mathsf{T}}w\}$$

Verify that this loss satisfies the mean-square affine-Lipschitz condition (18.75b) for some (δ,δ_2). *Remark.* For a related discussion, see Ying and Sayed (2018).

18.12 Consider the quadratic, Perceptron, and hinge losses defined by

$$Q(w;\gamma(m),h_m) = \begin{cases} q(w) \;+\; (\gamma(n)-h_m^{\mathsf{T}}w)^2, & \text{(quadratic)} \\ q(w) \;+\; \max\{0,-\gamma(m)h_m^{\mathsf{T}}w\}, & \text{(Perceptron)} \end{cases}$$

Show that these losses satisfy the affine Lipschitz condition (18.66b) under ℓ_1, ℓ_2, or elastic-net regularization. Determine for each case the respective values for $\{\delta,\delta_2\}$.

18.13 Refer to the bound (18.104) derived for a stochastic gradient implementation under importance sampling. Repeat the derivation assuming instead a mini-batch implementation where the B samples are selected with replacement. Show that the same bound holds with $\{\beta_g^2,\sigma_g^2\}$ divided by B.

18.14 Refer to the statement of Lemma 18.1 and let β be any nonnegative constant. Verify that

$$\mathbb{E}\,\Big\|\sum_{j=1}^{B}\beta^{B-j}\boldsymbol{x}_{\boldsymbol{\sigma}(j)}\Big\|^2 \;=\; \frac{1}{N(N-1)}\times\left(N\sum_{j=0}^{B-1}\beta^{2j}-\Big(\sum_{j=0}^{B-1}\beta^j\Big)^2\right)\times\sum_{n=0}^{N-1}\|x_n\|^2$$

18.A AVERAGING OVER MINI-BATCHES

In this appendix, we establish the validity of the third step in the argument leading to conclusion (18.29). To do so, we need to validate the equality:

$$\sum_{\ell=1}^{L}\left(\frac{1}{B}\sum_{b\in\mathcal{B}_\ell}\nabla_{w^{\mathsf{T}}}Q(w;\gamma(b),h_b)\right) \;=\; \frac{C_{N-1}^{B-1}}{B}\sum_{m=0}^{N-1}\nabla_{w^{\mathsf{T}}}Q(w;\gamma(m),h_m) \tag{18.117}$$

where C_{N-1}^{B-1} is the combinatorial coefficient for choosing $B-1$ data points out of $N-1$ total samples. We simplify the notation and denote the data point $(\gamma(m),h_m)$ by the letter x_m. We also introduce the symbols

$$q(w;x_m) \;\triangleq\; \nabla_{w^{\mathsf{T}}}Q(w;x_m) \tag{18.118a}$$

$$q^{\mathcal{B}_\ell}(w) \;\triangleq\; \frac{1}{B}\sum_{b\in B_\ell}\nabla_{w^{\mathsf{T}}}Q(w;x_b) \;=\; \frac{1}{B}\sum_{b\in B_\ell}q(w;x_b) \tag{18.118b}$$

so that we are interested in establishing the identity:

$$\sum_{\ell=1}^{L}q^{\mathcal{B}_\ell}(w) \;=\; \frac{C_{N-1}^{B-1}}{B}\sum_{m=0}^{N-1}q(w;x_m) \tag{18.119}$$

Consider first a few illustrative examples. Assume there are $N=3$ data samples x_0,x_1, and x_2 and that the mini-batch size is $B=2$. Then, there are $L=3$ candidate mini-batches $\{\mathcal{B}_1,\mathcal{B}_2,\mathcal{B}_3\}$ and, for this case,

$$q^{\mathcal{B}_1}(w) = \frac{1}{2}\Big(q(w;x_0)+q(w;x_1)\Big) \tag{18.120a}$$

$$q^{\mathcal{B}_2}(w) = \frac{1}{2}\Big(q(w;x_0)+q(w;x_2)\Big) \tag{18.120b}$$

$$q^{\mathcal{B}_3}(w) = \frac{1}{2}\Big(q(w;x_1)+q(w;x_2)\Big) \tag{18.120c}$$

As a result, it holds that

$$\sum_{\ell=1}^{L} q^{\mathcal{B}_\ell}(w) = q(w;x_0) + q(w;x_1) + q(w;x_2) = \sum_{m=0}^{N-1} q(w;x_m) \tag{18.121}$$

which satisfies (18.119). Assume next that there are $N = 4$ data samples x_0, x_1, x_2 and x_3 with the size of the mini-bath still at $B = 2$. Then, there are $L = C_4^2 = 6$ candidate mini-batches with:

$$q^{\mathcal{B}_1}(w) = \frac{1}{2}\Big(q(w;x_0) + q(w;x_1)\Big) \tag{18.122a}$$

$$q^{\mathcal{B}_2}(w) = \frac{1}{2}\Big(q(w;x_0) + q(w;x_2)\Big) \tag{18.122b}$$

$$q^{\mathcal{B}_3}(w) = \frac{1}{2}\Big(q(w;x_0) + q(w;x_3)\Big) \tag{18.122c}$$

$$q^{\mathcal{B}_4}(w) = \frac{1}{2}\Big(q(w;x_1) + q(w;x_2)\Big) \tag{18.122d}$$

$$q^{\mathcal{B}_5}(w) = \frac{1}{2}\Big(q(w;x_1) + q(w;x_3)\Big) \tag{18.122e}$$

$$q^{\mathcal{B}_6}(w) = \frac{1}{2}\Big(q(w;x_2) + q(w;x_3)\Big) \tag{18.122f}$$

As a result, it holds that

$$\sum_{\ell=1}^{L} q^{\mathcal{B}_\ell}(w) = \frac{1}{2}\Big(3q(w;x_0) + 3q(w;x_1) + 3q(w;x_2) + 3q(w;x_3)\Big) = \frac{3}{2}\sum_{m=0}^{N-1} q(w;x_m) \tag{18.123}$$

which again satisfies (18.119). In the third example, we assume there are $N = 4$ data samples x_0, x_1, x_2, and x_3 and increase the mini-batch size to $B = 3$. Then, there are $L = C_4^3 = 4$ candidate mini-batches with:

$$q^{\mathcal{B}_1}(w) = \frac{1}{3}\Big(q(w;x_0) + q(w;x_1) + q(w;x_2)\Big) \tag{18.124a}$$

$$q^{\mathcal{B}_2}(w) = \frac{1}{3}\Big(q(w;x_0) + q(w;x_1) + q(w;x_3)\Big) \tag{18.124b}$$

$$q^{\mathcal{B}_3}(w) = \frac{1}{3}\Big(q(w;x_0) + q(w;x_2) + q(w;x_3)\Big) \tag{18.124c}$$

$$q^{\mathcal{B}_4}(w) = \frac{1}{3}\Big(q(w;x_1) + q(w;x_2) + q(w;x_3)\Big) \tag{18.124d}$$

It follows that the following result holds, which satisfies (18.119):

$$\sum_{\ell=1}^{L} q^{\mathcal{B}_\ell}(w) = \frac{1}{3}\Big(3q(w;x_0) + 3q(w;x_1) + 3q(w;x_2) + 3q(w;x_3)\Big) = \sum_{m=0}^{N-1} q(w;x_m) \tag{18.125}$$

Let us consider next the general scenario with N data samples and mini-batches of size B. Then, there are

$$L = C_N^B = \frac{N!}{B!(N-B)!} \tag{18.126}$$

candidate batches. We denote these batches by $\mathcal{B}_1, \mathcal{B}_2, \cdots, \mathcal{B}_L$. It then holds that

$$\begin{aligned}\sum_{\ell=1}^{L} q^{\mathcal{B}_\ell}(w) &= \sum_{\ell=1}^{L}\left(\frac{1}{B}\sum_{b\in\mathcal{B}_\ell} q(w;x_b)\right) \qquad (18.127a)\\ &= \frac{1}{B}\sum_{\ell=1}^{L}\sum_{\ell=1}^{B} q(w;x_b)\\ &= \frac{1}{B}\Big(\alpha_0 q(w;x_0)+\alpha_1 q(w;x_1)+\cdots+\alpha_{N-1}q(w;x_{N-1})\Big)\end{aligned}$$

where the $\{\alpha_m\}$ are integers; each α_m counts how many times the term $q(w;x_m)$ appears in (18.127a). A critical observation here is that α_m is equal to the number of mini-batches that involve the data sample x_m, as is evident from the previous examples. Thus, suppose x_m is already selected. Then, the number of mini-batches that will contain x_m can be determined by counting in how many ways $B-1$ data samples (that exclude x_m) can be selected from the remaining $N-1$ data samples. This number is given by C_{N-1}^{B-1}. That is,

$$\alpha_0 = \cdots = \alpha_{N-1} = C_{N-1}^{B-1} = \frac{(N-1)!}{(B-1)!(N-B)!} \qquad (18.128)$$

from which we conclude that (18.117) holds.

18.B AUXILIARY VARIANCE RESULT

In this appendix we establish the following result.

Lemma 18.1. (Variance expression) *Consider N vectors $\{x_0, x_1, \ldots, x_{N-1}\}$ satisfying*

$$\frac{1}{N}\sum_{n=0}^{N-1} x_n = 0 \qquad (18.129)$$

Assume we sample B of the vectors without replacement and obtain the random sequence $\{\boldsymbol{x}_{\boldsymbol{\sigma}(1)}, \boldsymbol{x}_{\boldsymbol{\sigma}(2)}, \ldots, \boldsymbol{x}_{\boldsymbol{\sigma}(B)}\}$. Then, it holds that:

$$\mathbb{E}\,\Big\|\sum_{j=1}^{B}\boldsymbol{x}_{\boldsymbol{\sigma}(j)}\Big\|^2 \;=\; \frac{B(N-B)}{N(N-1)}\sum_{n=0}^{N-1}\|x_n\|^2 \qquad (18.130)$$

Proof: The proof employs mathematical induction and follows the derivation from Ying *et al.* (2018). We introduce the notation:

$$f(B) \;\triangleq\; \mathbb{E}\,\Big\|\sum_{j=1}^{B}\boldsymbol{x}_{\boldsymbol{\sigma}(j)}\Big\|^2 \qquad (18.131)$$

and note that for any single sample selected at random from the collection of N samples:

$$f(1) = \mathbb{E}\,\|\boldsymbol{x}_{\boldsymbol{\sigma}(j)}\|^2 \;=\; \frac{1}{N}\sum_{n=0}^{N-1}\|x_n\|^2 \;\triangleq\; \mathrm{var}(x) \qquad (18.132)$$

where we are using the notation $\mathrm{var}(x)$ to refer to the average squared value of the

samples. It follows that (18.130) holds for $B = 1$. Next we assume result (18.130) holds up to B and establish that it also holds at $B + 1$. Indeed, note that, by definition,

$$\begin{aligned}
f(B+1) &= \mathbb{E}\left\|\sum_{j=1}^{B+1} \boldsymbol{x}_{\boldsymbol{\sigma}(j)}\right\|^2 \\
&= \mathbb{E}\left\|\sum_{j=1}^{B} x_{\boldsymbol{\sigma}(j)} + x_{\boldsymbol{\sigma}(B+1)}\right\|^2 \\
&= \mathbb{E}\left\|\sum_{j=1}^{B} x_{\boldsymbol{\sigma}(j)}\right\|^2 + \mathbb{E}\,\|x_{\boldsymbol{\sigma}(B+1)}\|^2 + 2\,\mathbb{E}\left(\sum_{j=1}^{B} x_{\boldsymbol{\sigma}(j)}\right)^{\mathsf{T}} x_{\boldsymbol{\sigma}(B+1)} \\
&= f(B) + \mathrm{var}(x) + 2\,\mathbb{E}\left(\sum_{j=1}^{B} x_{\boldsymbol{\sigma}(j)}\right)^{\mathsf{T}} x_{\boldsymbol{\sigma}(B+1)}
\end{aligned} \tag{18.133}$$

where we used

$$\mathbb{E}\,\|x_{\boldsymbol{\sigma}(B+1)}\|^2 \overset{(18.132)}{=} \mathrm{var}(x) \tag{18.134}$$

We introduce the notation $\boldsymbol{\sigma}(1\colon B)$ to denote the collection of sample indexes selected during steps 1 through B. To evaluate the last cross term in (18.133), we exploit the conditional mean property $\mathbb{E}\,\boldsymbol{a} = \mathbb{E}\,(\mathbb{E}\,(\boldsymbol{a}|\boldsymbol{b}))$, for any two random variables $\boldsymbol{a}$ and $\boldsymbol{b}$, to write:

$$\begin{aligned}
&\mathbb{E}\left(\sum_{j=1}^{B} x_{\boldsymbol{\sigma}(j)}\right)^{\mathsf{T}} x_{\boldsymbol{\sigma}(B+1)} \\
&= \mathbb{E}_{\boldsymbol{\sigma}(1:B)}\left[\mathbb{E}_{\boldsymbol{\sigma}(B+1)}\left\{\left(\sum_{j=1}^{B} x_{\boldsymbol{\sigma}(j)}\right)^{\mathsf{T}} x_{\boldsymbol{\sigma}(B+1)}\,\middle|\,\boldsymbol{\sigma}(1:B)\right\}\right] \\
&\overset{(a)}{=} \mathbb{E}_{\boldsymbol{\sigma}(1:B)}\left[\left(\sum_{j=1}^{B} x_{\boldsymbol{\sigma}(j)}\right)^{\mathsf{T}}\left(\frac{1}{N-B}\sum_{j'\notin\boldsymbol{\sigma}(1:B)} x_{j'}\right)\right] \\
&= \frac{1}{N-B}\mathbb{E}_{\boldsymbol{\sigma}(1:B)}\left[\left(\sum_{j=1}^{B} x_{\boldsymbol{\sigma}(j)}\right)^{\mathsf{T}}\left(\sum_{j'=0}^{N-1} x_{j'} - \sum_{j'=1}^{B} x_{\boldsymbol{\sigma}(j')}\right)\right] \\
&\overset{(b)}{=} -\frac{1}{N-B}\mathbb{E}_{\boldsymbol{\sigma}(1:B)}\left[\left(\sum_{j=1}^{B} x_{\boldsymbol{\sigma}(j)}\right)^{\mathsf{T}}\sum_{j'=1}^{B} x_{\boldsymbol{\sigma}(j')}\right] \\
&= -\frac{1}{N-B}\mathbb{E}_{\boldsymbol{\sigma}(1:B)}\left(\sum_{j=1}^{B}\|x_{\boldsymbol{\sigma}(j)}\|^2\right) - \\
&\quad \frac{1}{N-B}\mathbb{E}_{\boldsymbol{\sigma}(1:B)}\left[\sum_{j=1}^{B}\left(\sum_{j'=1,j'\neq j}^{B} x_{\boldsymbol{\sigma}(j)}^{\mathsf{T}} x_{\boldsymbol{\sigma}(j')}\right)\right] \\
&\overset{(18.132)}{=} -\frac{B}{N-B}\mathrm{var}(x) - \frac{1}{N-B}\mathbb{E}_{\boldsymbol{\sigma}(1:B)}\left[\sum_{j=1}^{B}\left(\sum_{j'=1,j'\neq j}^{B} x_{\boldsymbol{\sigma}(j)}^{\mathsf{T}} x_{\boldsymbol{\sigma}(j')}\right)\right]
\end{aligned} \tag{18.135}$$

where in step (a) we used the fact that

$$\mathbb{E}\left(x_{\boldsymbol{\sigma}(B+1)} \,|\, \boldsymbol{\sigma}(1:B)\right) \;=\; \frac{1}{N-B}\sum_{j'\notin\boldsymbol{\sigma}(1:B)} x_{j'} \tag{18.136}$$

since the expectation is over the distribution of $x_{\boldsymbol{\sigma}(B+1)}$, and in step (b) we used the condition

$$\sum_{j'=0}^{N-1} x_{j'} = 0 \tag{18.137}$$

We continue with (18.135). Without loss of generality, we assume $j < j'$ in the following argument. If $j > j'$, exchanging the places of $x_{\boldsymbol{\sigma}(j)}$ and $x_{\boldsymbol{\sigma}(j')}$ leads to the same conclusion:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\sigma}(1:B)}\left(x_{\boldsymbol{\sigma}(j)}^{\mathsf{T}} x_{\boldsymbol{\sigma}(j')}\right) &= \mathbb{E}_{\boldsymbol{\sigma}(j),\boldsymbol{\sigma}(j')}\left(x_{\boldsymbol{\sigma}(j)}^{\mathsf{T}} x_{\boldsymbol{\sigma}(j')}\right) \\
&= \mathbb{E}_{\boldsymbol{\sigma}(j)}\left\{x_{\boldsymbol{\sigma}(j)}^{\mathsf{T}}\left(\mathbb{E}_{\boldsymbol{\sigma}(j')}[x_{\boldsymbol{\sigma}(j')} \,|\, \boldsymbol{\sigma}(j)]\right)\right\} \\
&= \mathbb{E}_{\boldsymbol{\sigma}(j)}\left\{x_{\boldsymbol{\sigma}(j)}^{\mathsf{T}}\left(\frac{1}{N-1}\sum_{j'\neq j}^{N-1} x_n\right)\right\} \\
&= \mathbb{E}_{\boldsymbol{\sigma}(j)}\left\{x_{\boldsymbol{\sigma}(j)}^{\mathsf{T}}\left(\frac{1}{N-1}\left[\sum_{j'=1}^{N-1} x_{j'} - x_{\boldsymbol{\sigma}(j)}\right]\right)\right\} \\
&\overset{(18.137)}{=} -\frac{1}{N-1}\mathbb{E}_{\boldsymbol{\sigma}(j)}\|x_{\boldsymbol{\sigma}(j)}\|^2 \\
&= -\frac{1}{N-1}\mathrm{var}(x)
\end{aligned} \tag{18.138}$$

Substituting (18.138) into (18.135), we obtain:

$$\mathbb{E}\left(\sum_{j=1}^{B} x_{\boldsymbol{\sigma}(j)}\right)^{\mathsf{T}} x_{\boldsymbol{\sigma}(B+1)} = -\frac{B}{N-1}\mathrm{var}(x) \tag{18.139}$$

Combining (18.133), (18.134), and (18.139), we get:

$$\begin{aligned}
f(B+1) &= f(B) + \mathrm{var}(x) - \frac{2B}{N-1}\mathrm{var}(x) \\
&\overset{(a)}{=} \frac{B(N-B)}{N-1}\mathrm{var}(x) + \mathrm{var}(x) - \frac{2B}{N-1}\mathrm{var}(x) \\
&= \left(\frac{(B+1)(N-B-1)}{N-1}\right)\mathrm{var}(x)
\end{aligned} \tag{18.140}$$

where in step (a) we used the induction assumption on $f(B)$ and form (18.130). The same form turns out to be valid for $f(B+1)$ and we conclude that (18.130) is valid. ■

REFERENCES

Agarwal, A., P. L. Bartlett, P. Ravikumar, and M. J. Wainwright (2012), "Information-theoretic lower bounds on the oracle complexity of convex optimization," *IEEE Trans. Information Theory*, vol. 58, no. 5, pp. 3235–3249.

Bertsekas, D. P. (1999), *Nonlinear Programming*, Athena Scientific, Belmont, MA, 2nd edition.

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.

Chen, J. and A. H. Sayed (2012a), "Diffusion adaptation strategies for distributed optimization and learning over networks," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4289–4305.

Nedic, A. and A. Ozdaglar (2009), "Distributed subgradient methods for multiagent optimization,"*IEEE Trans. Automatic Control*, vol. 54, no. 1, pp. 48–61.

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.

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

Polyak, B. T. and Y. Z. Tsypkin (1973), "Pseudogradient adaptation and training algorithms," *Automat. Remote Control*, vol. 12, pp. 83–94.

Ram, S. S., A. Nedic, and V. V. Veeravalli (2010), "Distributed stochastic subgradient projection algorithms for convex optimization," *J. Optim. Theory Appl.*, vol. 147, no. 3, pp. 516–545.

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.

Srivastava, K. and A. Nedic (2011), "Distributed asynchronous constrained stochastic optimization," *IEEE J. Sel. Topics. Signal Process.*, vol. 5, no. 4, pp. 772–790.

Ying, B. and A. H. Sayed (2018), "Performance limits of stochastic sub-gradient learning, Part I: Single-agent case," *Signal Processing*, vol. 144, pp. 271–282.

Ying, B., K. Yuan, S. Vlaski, and A. H. Sayed (2018), "Stochastic learning under random reshuffling with constant step-sizes," *IEEE Trans. Signal Process.*, vol. 67, no. 2, pp. 474–489.