2 Deterministic Optimization

The discussion in Chapter 1 motivated how a significant number of inference and learning problems can be formalized as optimization problems:

$$\min_{w \in \mathbb{R}^M} J(w) \tag{2.1}$$

for some suitably chosen risk function $J(\cdot)$. The term "risk function" is common in the literature on statistical and machine learning. When describing more general optimization problems, it is also referred to as a "cost" or "objective" function, terms which we will use interchangebly. Strategies for optimal learning then boil down to finding optimal solutions to (2.1). The examples in Chapter **??** were based on least squares (LS) and mean squared error (MSE) metrics, which naturally arise from linear models. These kinds of learning objectives take the form of a quadratic, i.e.:

$$J(w) = \frac{1}{2} w^{\mathsf{T}} A w + b^{\mathsf{T}} w + c \tag{2.2}$$

for some suitably chosen constants $A \in \mathbb{R}^{M \times M}, b \in \mathbb{R}^M, c \in \mathbb{R}$. Quadratic optimization problems are appealing, because their optimal solutions can be pursued fairly efficiency and in closed form, provided that the dimension M is not too high. In particular, if we define

$$w^\star \triangleq \arg \min_{w \in \mathbb{R}^M} J(w) \tag{2.3}$$

it must hold that:

$$\nabla J(w^\star) = 0 \iff A w^\star + b = 0 \iff w^\star = A^{-1} b \tag{2.4}$$

In other words, the optimality conditions of a quadratic optimization problem translate to a linear system of equations, which can be solved uniquely provided that A is invertible. Unfortunately, many of the optimization problems that arise from more complex learning tasks are no longer quadratic, and their solutions can hence no longer be inferred directly from their optimality conditions in closed form. Instead, we need devise more elaborate schemes for finding $w^\star$.

Example 2.1 (**Least-Squares Estimation of a Linear Model**) The linear model $\boldsymbol{\gamma} = \boldsymbol{w}^{\mathsf{T}} \boldsymbol{h} + \boldsymbol{v}$ in Example 1.4 of Chapter (**??**) led to the quadratic optimization

© 2023. All rights reserved. This draft cannot be copied or distributed in print or electronically without the written consent of the authors S. Vlaski and A. H. Sayed.

problem:

$$w^\star = \arg\min_{w\in\mathbb{R}^M} \left\{ \frac{1}{2N}\sum_{n=1}^{N} \left(\gamma_n - h_n^\mathsf{T} w\right)^2 + \frac{1}{2N}\frac{\sigma_v^2}{\sigma_w^2}\|w\|^2 \right\} \tag{2.5}$$

which we can write equivalently as:

$$J(w) = \frac{1}{2N} w^\mathsf{T} \left(\sum_{n=1}^{N} h_n h_n^\mathsf{T} + \frac{\sigma_v^2}{\sigma_w^2} \right) w + \left(\frac{1}{N}\sum_{n=1}^{N} \gamma_n h_n \right)^\mathsf{T} w + \frac{1}{2N}\sum_{n=1}^{N}\gamma_n^2 \tag{2.6}$$

We can hence identify this least-squares problem as a special form of the quadratic problem (2.2) with:

$$\begin{aligned} A &= \frac{1}{N}\left(\sum_{n=1}^{N} h_n h_n^\mathsf{T} + \frac{\sigma_v^2}{\sigma_w^2} \right) \\ b &= \frac{1}{N}\sum_{n=1}^{N}\gamma_n h_n \\ c &= \frac{1}{2N}\sum_{n=1}^{N}\gamma_n^2 \end{aligned} \tag{2.7}$$

Hence:

$$w^\star = A^{-1}b = \left(\sum_{n=1}^{N} h_n h_n^\mathsf{T} + \frac{\sigma_v^2}{\sigma_w^2} \right)^{-1} \left(\sum_{n=1}^{N}\gamma_n h_n \right) \tag{2.8}$$

2.1 REGULARITY CONDITIONS

Our goal in this chapter will be to develop to develop optimization algorithms for a broad class of learning problems. Nevertheless, we will need to impose certain regularity conditions on the cost $J(w)$ in order to be able to make claims about the behavior of the resulting algorithms. We will introduce these conditons in this section and establish useful properties that will be useful as we develop and analyze optimization algorithms in subsequent sections and chapters.

2.1.1 Convexity and Strong Convexity

In general, $J(w)$ may not have a unique minimizer, such as in the case $J(w) = -\|w\|^2$, or may have inifinitely many minimizers, such as in the case $J(w) = \sin(\|w\|)$. To avoid ill-posed problems of this kind, in many cases, we will restrict ourselves to *convex* optimization problems. We say that $J(w)$ is convex, if for every $w_1, w_2 \in \mathbb{R}^M$, and every $0 \le \alpha \le 1$, it holds that:

$$J(\alpha w_1 + (1-\alpha)w_2) \le \alpha J(w_1) + (1-\alpha)J(w_2) \tag{2.9}$$

In words, this condition states that any line or hyperplane connecting $J(w_1)$ and $J(w_2)$ lies above the function $J(\cdot)$. While this definition is intuitive, when developing and analyzing algorithms for solving optimization problems, it is generally more useful to use different, but equivalent, formulations of convexity. If $J(w)$ is differentiable, we can instead write:

$$J(w_2) \geq J(w_1) + \nabla J(w_1)^{\mathsf{T}}(w_2 - w_1) \tag{2.10}$$

This condition again carries an intuitive interpretation, namely that at any point w_1, the function $J(\cdot)$ lies above its tangent hyperplane. If the objective function $J(w)$ happens to be twice differentiable, this translates an equivalent condition that:

$$\nabla^2 J(w) \succeq 0 \tag{2.11}$$

Here, $\nabla^2 J(w)$ denotes the Hessian matrix of $J(w)$, and the statement (2.11) indicates that the Hessian is positive semi-definite, i.e., all of its eigenvalues are greater than or equal to zero.

While the assumption of convexity excludes many ill-posed problems, it is not sufficient to ensure uniqueness of the optimal solution $w^\star$. Indeed, any constant function $J(w) = c$ is convex, yet it is minimized at any point in $\mathbb{R}^M$. One sufficient (though not necessary) condition to ensure uniqueness of the optimal solution $w^\star \triangleq \arg\min J(w)$ is to impose *strong* convexity. We will say that a function is ν-strongly convex, if:

$$J(w_2) \geq J(w_1) + \nabla J(w_1)^{\mathsf{T}}(w_2 - w_1) + \frac{\nu}{2}\|w_2 - w_1\|^2 \tag{2.12}$$

for some positive parameter $\nu > 0$. Note that strong convexity is a stronger condition than convexity, and in particular we would recover (2.10) from (2.12). Furthermore, any strongly convex function is also convex. Now suppose we have found, by whatever means, a point satisfying the first-order condition $\nabla J(w^\star) = 0$. Then, it follows from (2.12) that for any $w \in \mathbb{R}^M$:

$$\begin{aligned} J(w) &\geq J(w^\star) + \nabla J(w^\star)^{\mathsf{T}}(w - w^\star) + \frac{\nu}{2}\|w - w^\star\|^2 \\ &= J(w^\star) + \frac{\nu}{2}\|w - w^\star\|^2 \end{aligned} \tag{2.13}$$

Hence, $w^\star$ is the minimizing arguement of $J(w)$. If $J(w)$ happens to be twice-differentiable, (2.12) can be shown to be equivalent to:

$$\nabla^2 J(w) \succeq \nu I_M \tag{2.14}$$

where I_M defines a diagonal matrix of dimension $M \times M$ with ones on its diagonal. Relation (2.14) states that $\nabla^2 J(w) - \nu I_M$ is positive semidefinite, or equivalently, all eigenvalues of $\nabla^2 J(w)$ are greater than or equal to ν.

2.1.2 Smoothness

In Section 2.1.1 we examined a number of convexity conditions on the cost function $J(w)$, which impose a lower bound on the curvature of $J(w)$, and can be used to ensure uniqueness of the optimal solution $w^\star$. We now introduce a set of complimentary conditions, which ensure the *smoothness* of a function. In particular, we will say that $J(w)$ is smooth, if it satisfies:

$$J(w_2) \leq J(w_1) + \nabla J(w_1)^\mathsf{T}(w_2 - w_1) + \frac{\delta}{2}\|w_2 - w_1\|^2 \tag{2.15}$$

Condition (2.15) is essentially analogous to (2.12), except now we impose an *upper* bound on the growth above the tangent hyperplane, where δ controls the rate of growth. There are again a nuber of equivalent formulations of condition (2.15). In particular, (2.15) is equivalent to the statement:

$$\|\nabla J(w_2) - \nabla J(w_1)\| \leq \delta\|w_2 - w_1\| \tag{2.16}$$

In other words, the smoothness condition (2.15) translates into a Lipschitz condition on the gradients of $J(w)$. As we will see in later sections, most optimization algorithms are based on a number of local approximations, and the accuracy of these approximations will in large part be driven by the smoothness of the objective function measured through the parameter δ. If the objective function happens to be twice differentiable, relations (2.15) and (2.16) can further be shown to be equivalent to:

$$\nabla^2 J(w) \preceq \delta I_M \tag{2.17}$$

Example 2.2 (**Strong convexity and smoothness parameters of quadratic functions**) For a generic quadratic problem of the form (2.2), we have:

$$\nabla^2 J(w) = A \tag{2.18}$$

and hence:

$$\nu \overset{(2.14)}{=} \lambda_{\min}(A) \tag{2.19}$$

$$\delta \overset{(2.17)}{=} \lambda_{\max}(A) \tag{2.20}$$

Here, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the smallest and largest eigenvalue of A respectively. Furthermore, $J(w)$ is strongly convex if, and only if, $\lambda_{\min}(A) > 0$, which in turn ensures that A is invertible and $w^\star = A^{-1}b$ is well-defined.

2.2 GRADIENT DESCENT

The conditions established in Section 2.1 ensure the existence of a unique solution $w^\star$ characterized by $\nabla J(w^\star) = 0$, but so far we have not described a way for systematically finding $w^\star$. To motivate a classical algorithm for optimization, called gradient descent, we recall that the reason that we are able to very easily solve quadratic optimization problems such as (2.2) is that their first-order optimality condition $\nabla J(w^\star) = 0$ gives rise to a linear system of equations. If $J(w)$ is no longer quadratic, rather than minimize $J(w)$ directly we may then iteratively minimize a sequence of quadratic upper bounds on $J(w)$. Specifically, we construct iterates:

$$w_i = \arg\min_{w\in\mathbb{R}^M} \left\{ J(w_{i-1}) + \nabla J(w_{i-1})^{\mathsf{T}}(w - w_{i-1}) + \frac{1}{2\mu}\|w - w_{i-1}\|^2 \right\} \tag{2.21}$$

Whenever $\mu \leq \frac{2}{\delta}$, the objective of (2.21) is an upper bound on $J(w)$, as can be verified from (2.15). Since the objective of (2.21) is a quadratic, we can find its minimizing argument by differentiating and setting the derivative equal to zero. In this manner we find:

$$w_i = w_{i-1} - \mu\nabla J(w_{i-1}) \tag{2.22}$$

The name *gradient descent* arises from the fact that recursion (2.22) produces a sequence of iterates w_i, which descent along the objective function $J(w)$. Indeed, we can show that for any δ-smooth function $J(w)$ that:

$$\begin{aligned}
J(w_i) &\overset{(2.15)}{\leq} J(w_{i-1}) + \nabla J(w_{i-1})^{\mathsf{T}}(w_i - w_{i-1}) + \frac{\delta}{2}\|w_i - w_{i-1}\|^2 \\
&\overset{(2.22)}{=} J(w_{i-1}) + \nabla J(w_{i-1})^{\mathsf{T}}(-\mu\nabla J(w_{i-1})) + \frac{\delta}{2}\|\mu\nabla J(w_{i-1})\|^2 \\
&= J(w_{i-1}) - \mu\|\nabla J(w_{i-1})\|^2 + \mu^2\frac{\delta}{2}\|\nabla J(w_{i-1})\|^2 \\
&= J(w_{i-1}) - \mu\left(1 - \mu\frac{\delta}{2}\right)\|\nabla J(w_{i-1})\|^2 \\
&\overset{(a)}{\leq} J(w_{i-1})
\end{aligned} \tag{2.23}$$

where (a) holds whenever $1-\mu\frac{\delta}{2} \geq 0 \iff \mu \leq \frac{\delta}{2}$, which is the same condition that ensures that we are minimizing a quadratic *upper bound* in (2.21). In summary, we arrive at the following algorithm:

Gradient-descent algorithm for solving ((2.1)).

start from any initial condition, w_0**.**
repeat over $i \geq 0$ **until convergence** :

$$w_i = w_{i-1} - \mu\nabla J(w_{i-1})$$

end (2.24)

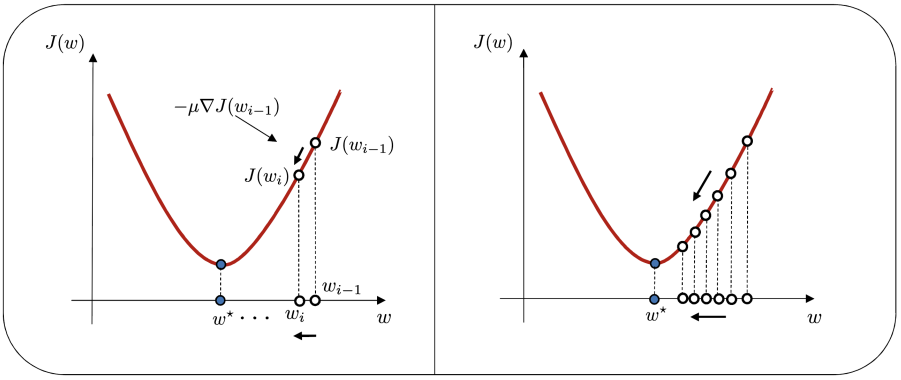

Figure 2.1 The panel on the left shows the mechanics of one update step where w_{i-1} is updated in the direction of the minimizer $w^\star$. The panel on the right shows the result of several successive steps with the iterates approaching $w^\star$.

Example 2.3 (**Quadratic problems**) Let us consider again the general quadratic objective (2.2). We have:

$$\nabla J(w) = Aw + b \tag{2.25}$$

Hence, the gradient descent recursion (2.22) specializes to:

$$w_i = w_{i-1} - \mu\,(Aw_{i-1} + b) = (I_M - \mu A)w_{i-1} - \mu b \tag{2.26}$$

While we could have pursued $w^\star$ for a quadratic cost more directly via $w^\star = A^{-1}b$, the recursive solution (2.26) can nevertheless be a useful alternative if computation of the inverse A^{-1} is infeasible, such as for example in large-scale settings.

Example 2.4 (**Logistic regression**) Quadradic loss functions are not appropriate when the label $\boldsymbol{\gamma}$ in an inference problem takes discrete values $\boldsymbol{\gamma} = \pm 1$. In this case, we may formulate the logistic regression problem:

$$J(w) = \frac{\rho}{2}\|w\|^2 \;+\; \frac{1}{N}\sum_{n=1}^{N} \ln\left(1 + e^{-\gamma_n h_n^{\mathsf{T}} w}\right) \tag{2.27a}$$

$$\nabla J(w) = \rho w \;-\; \frac{1}{N}\sum_{n=1}^{N} \frac{\gamma_n h_m}{1 + e^{\gamma_n h_n^{\mathsf{T}} w}} \tag{2.27b}$$

The gradient-descent recursion (2.22) in this case leads to

$$w_i \;=\; (1 - \mu\rho)\, w_{i-1} + \mu\left(\frac{1}{N}\sum_{n=1}^{N} \frac{\gamma_n h_n}{1 + e^{\gamma_n h_n^{\mathsf{T}} w_{i-1}}}\right), \quad i \geq 0 \tag{2.28}$$

Following the arguments of Chapter 1, we can interpret the logistic regression

loss as a Bayes' optimal estimate of $\boldsymbol{w}$ under statistical model:

$$f_{\boldsymbol{\gamma}|\boldsymbol{h},\boldsymbol{w}}(\gamma|h,w) = \frac{1}{1+e^{-\gamma h^{\mathsf{T}} w}} \tag{2.29}$$

with a Gaussian prior on $\boldsymbol{w}$, though the details of this derivation are not relevant to the discussion in this chapter, and left to the reader as an exercise. Once we have found $w^{\star}$, we can then perform inference on unlabeled data $\boldsymbol{h}$ by computing sign $\left(\boldsymbol{h}^{\mathsf{T}} w^{\star}\right)$.

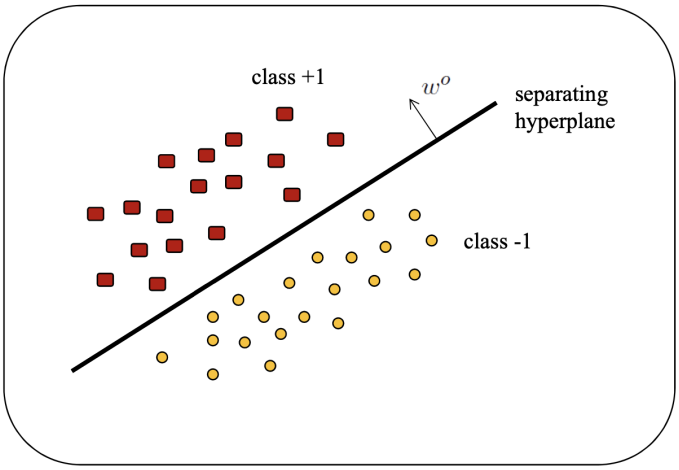

Figure 2.2 A linear classification problem.

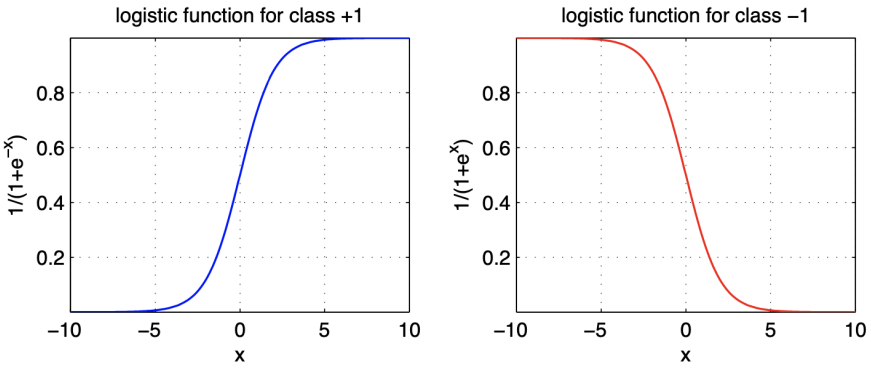

Figure 2.3 The left figure shows the logistic function (2.29) when $\gamma = +1$ as a function of $x = h^{\mathsf{T}} w$, the right firgure shows the same for $\gamma = -1$. This model imposes that it is likely that $\boldsymbol{\gamma}$ and $\boldsymbol{h}^{\mathsf{T}}\boldsymbol{w}$ have the same sign, with the likehood increasing with the absolute value of $\boldsymbol{h}^{\mathsf{T}}\boldsymbol{w}$.

2.3 CONVERGENCE ANALYSIS

While the descent relation (2.23) indicates that the gradient descent recursion (2.22) yields progressively improved iterates, it does not constitute a guarantee of convergence, and does not tell us how fast the iterates w_i approach the optimal solution $w^\star$. We now show how, using the regularity conditions of Section 2.1, we can derive such a convergence guarantee.

THEOREM 2.1 (**Convergence of gradient descent**). *Consider the gradient descent recursion (2.22) where $J(w)$ is a ν-strongly convex function with δ-Lipschitz gradients. Introduce the error vector $\widetilde{w}_i = w^\star - w_i$, which measures the difference between the $i-th$ iterate and the global minimizer of $J(w)$. If the step-size μ satisfies (i.e., is small enough):*

$$0 < \mu < 2\nu/\delta^2 \tag{2.30}$$

then w_i converges exponentially fast to $w^\star$ in the sense that

$$\|\widetilde{w}_i\|^2 \;\leq\; \lambda\,\|\widetilde{w}_{i-1}\|^2, \quad i \geq 0 \tag{2.31}$$

where

$$\lambda \;=\; 1 - 2\mu\nu + \mu^2\delta^2 \;\in [0,1) \tag{2.32}$$

It also holds that the risk value converges exponentially fast as follows

$$J(w_i) - J(w^\star) \;\leq\; \delta\lambda^i\|\widetilde{w}_0\|^2 \;=\; O(\lambda^i) \tag{2.33}$$

Proof: We subtract $w^\star$ from both sides of (2.22) to get

$$\widetilde{w}_i \;=\; \widetilde{w}_{i-1} \;+\; \mu\,\nabla J(w_{i-1}) \tag{2.34}$$

We next compute the squared Euclidean norms (or energies) of both sides of the above equality and use the fact that $\nabla J(w^\star) = 0$ to write

$$\begin{aligned}
\|\widetilde{w}_i\|^2 &= \|\widetilde{w}_{i-1}\|^2 + 2\mu\,(\nabla J(w_{i-1}))^{\mathsf{T}}\,\widetilde{w}_{i-1} + \mu^2\,\|\nabla J(w_{i-1})\|^2 \\
&= \|\widetilde{w}_{i-1}\|^2 + 2\mu\,(\nabla J(w_{i-1}))^{\mathsf{T}}\,\widetilde{w}_{i-1} + \mu^2\,\|\nabla J(w^\star) - \nabla J(w_{i-1})\|^2 \\
&\overset{(2.16)}{\leq} \|\widetilde{w}_{i-1}\|^2 + 2\mu\,(\nabla J(w_{i-1}))^{\mathsf{T}}\,\widetilde{w}_{i-1} + \mu^2\delta^2\|\widetilde{w}_{i-1}\|^2
\end{aligned} \tag{2.35}$$

We appeal to the strong-convexity property (2.12) and use $w_2 = w^\star$, $w_1 = w_{i-1}$ in step (a) below and $w_2 = w_{i-1}$, $w_1 = w^\star$ in step (b) to find that

$$\begin{aligned}
(\nabla J(w_{i-1}))^{\mathsf{T}}\,\widetilde{w}_{i-1} &\overset{(a)}{\leq} J(w^\star) - J(w_{i-1}) - \frac{\nu}{2}\|\widetilde{w}_{i-1}\|^2 \\
&\overset{(b)}{\leq} -\frac{\nu}{2}\|\widetilde{w}_{i-1}\|^2 \;-\; \frac{\nu}{2}\|\widetilde{w}_{i-1}\|^2 \\
&= -\nu\|\widetilde{w}_{i-1}\|^2
\end{aligned} \tag{2.36}$$

Substituting into (2.35) gives

$$\|\widetilde{w}_i\|^2 \;\leq\; (1 - 2\mu\nu + \mu^2\delta^2)\,\|\widetilde{w}_{i-1}\|^2 \tag{2.37}$$

which coincides with (2.31)–(2.32). Iterating we find that

$$\|\widetilde{w}_i\|^2 \;\leq\; \lambda^i \,\|\widetilde{w}_0\|^2 \tag{2.38}$$

which highlights the exponential convergence of $\|\widetilde{w}_i\|^2$ to zero at the rate λ^i. We next verify that condition (2.30) ensures $0 \leq \lambda < 1$. For this purpose, we refer to Figure 2.4, which plots the coefficient $\lambda(\mu)$ as a function of μ. The minimum value of $\lambda(\mu)$ occurs at location $\mu = \nu/\delta^2$ and is equal to $1 - \nu^2/\delta^2$. This value is nonnegative since $0 < \nu \leq \delta$. It is clear from the figure that $0 \leq \lambda < 1$ for $\mu \in (0, \frac{2\nu}{\delta^2})$.

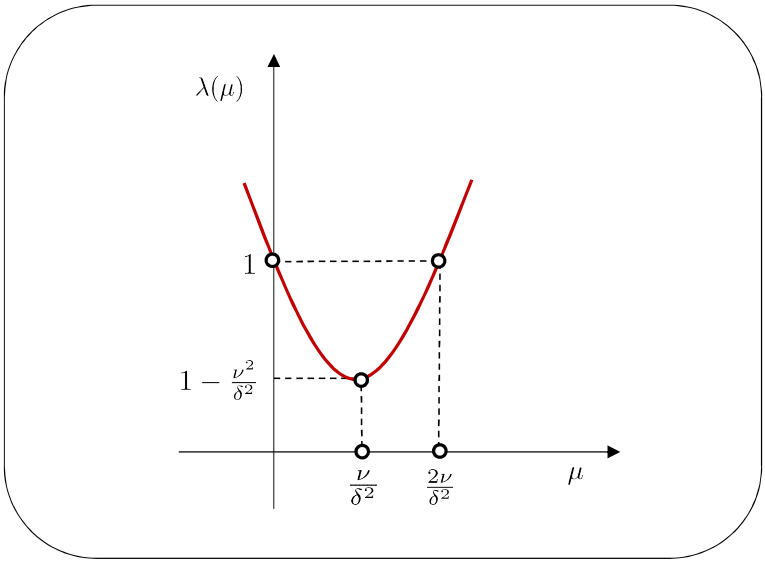

Figure 2.4 Plot of the function $\lambda(\mu) = 1 - 2\nu\mu + \mu^2\delta^2$ given by (2.32). It shows that the function $\lambda(\mu)$ assumes values below one in the range $0 < \mu < 2\nu/\delta^2$.

To establish (2.33), we first note that $J(w_i) \geq J(w^\star)$ since $w^\star$ is the minimizer of $J(w)$. Using (2.12) again with $w_2 = w^\star$ and $w_1 = w_i$ we get

$$\begin{aligned}
0 \leq J(w_i) - J(w^\star) \;&\leq\; -(\nabla J(w_i))^{\mathsf{T}}\,\widetilde{w}_i \\
&\leq\; \|\nabla J(w_i)\|\,\|\widetilde{w}_i\| \quad \text{(by Cauchy-Schwarz)} \\
&=\; \|\nabla J(w_i) - \nabla J(w^\star)\|\,\|\widetilde{w}_i\| \quad \text{(since } \nabla J(w^\star) = 0\text{)} \\
&\overset{(2.16)}{\leq}\; \delta\|\widetilde{w}_i\|^2 \\
&\overset{(2.37)}{\leq}\; \delta\lambda^i\|\widetilde{w}_0\|^2
\end{aligned} \tag{2.39}$$

□

Remark 2.1 (**Big-O and little-o notation**). The statement (2.33) uses the big-O notation. In other locations, we will also employ the little-o notation. We therefore explain their meaning in this remark. The big-O notation is used to compare the asymptotic growth rate of two sequences. Thus, writing $a_i = O(b_i)$, with a big O, means that there exists some constant c such that $|a_i| \leq c|b_i|$ for large enough $i > i_o$. This also means that the decay rates of both sequences a_i and b_i are comparable. For example, writing $a_i = O(1/i)$ means that the samples of the sequence a_i decay asymptotically at

a rate that is comparable to $1/i$. On the other hand, the little-o notation, $a_i = o(b_i)$, means that, asymptotically, the sequence a_i decays faster than the sequence b_i so that it should hold $|a_i|/|b_i| \to 0$ as $i \to \infty$. In this case, the notation $a_i = o(1/i)$ implies that the samples of a_i decay at a faster rate than $1/i$.

Remark 2.2 (**Exponential or linear convergence)**. Recursions evolving according to a dynamics of the form (2.31), such as $a_i \leq \lambda\, a_{i-1}$ for some $\lambda \in [0, 1)$, are said to converge exponentially fast since, by iterating, we get $a_i \leq \lambda^i a_0$. This expression shows that a_i decays to zero exponentially at the rate λ^i. This mode of convergence is also referred to as *linear* convergence because, when plotted on a semi-log scale, the curve $\ln a_i \times i$ will turn out to be linear in i with slope λ, namely,

$$\ln a_i \leq \lambda\, i + \ln a(0) \tag{2.40}$$

for sequences $a_i > 0$ for all i.

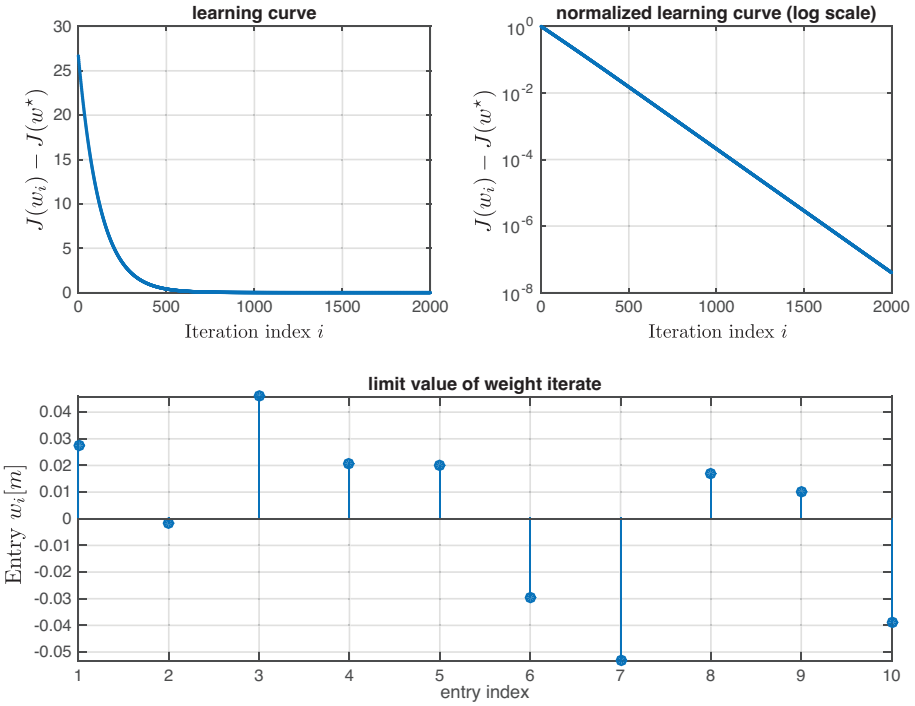

Figure 2.5 Learning curves $J(w_i)$ relative to the minimum risk value $J(w^\star)$ in linear scale (on the left) and in normalized logarithmic scale (on the right). This latter plot confirms the linear convergence of the risk value towards $J(w^\star)$. (*Bottom*) Limiting value of the weight iterate w_i, which tends to the minimizer $w^\star$ according to result (2.31).

2.4 PROBLEMS

2.1 Find the smoothness and strong-convexity parameters δ and ν of the logistic regression cost (2.27a). Use these and Theorem 2.1 to determine the rate of convergence of gradient descent when applied to problem (2.27a). Which choice of the step-size parameter μ gives the fastest rate of convergence?

3 Stochastic Gradient Approximations

In this chapter, we will continue to study optimization problems of the form:

$$\min_{w \in \mathbb{R}^M} J(w) \tag{3.1}$$

In the previous Chapter 2, we studied the performance of the gradient descent algorithm, which takes the form:

$$w_i = w_{i-1} - \mu \nabla J(w_{i-1}) \tag{3.2}$$

The gradient descent recurison (3.2) requires the evaluation of the gradient $\nabla J(w_{i-1})$ at every iteration. In many learning settings, depending on the structure of the cost $J(w)$, the evaluation of the exact gradient $\nabla J(w)$ may be infeasible or prohibitively expensive. The two most common such scenarios are online learning from streaming data, and empirical risk minimization, which we describe in more detail in the sequel.

Example 3.1 (**Risk minimization**) We will consider now the important setting, where the function value $J(w)$ as well as its gradient are no longer directly accessible by the learning algorithm, and instead, the loss $J(w)$ is defined as the expected value of some risk $Q(w; \boldsymbol{x})$ over the distribution of some data $\boldsymbol{x}$.

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

In the absence of knowledge about the distribution of the data $\boldsymbol{x}$, even evaluation of the function value $J(w)$ at any given point w is infeasible, since evaluation of the function value $J(w)$ in (3.3) requires evaluation of the expected value of the risk $Q(w; \boldsymbol{x})$ relative to the distribution of $\boldsymbol{x}$. Similarly, evaluation of the gradient of $J(w)$ would require:

$$\nabla_w J(w) = \nabla_w \left(\mathbb{E}_{\boldsymbol{x}} Q(w; \boldsymbol{x}) \right) \tag{3.4}$$

which again is not possible in the absence of knowledge about the distribution of $\boldsymbol{x}$. Suppose at time i, we are given access to a single sample $\boldsymbol{x}_i$ following the distribution of $\boldsymbol{x}$. We can then construct:

$$\widehat{\nabla J}(w; \boldsymbol{x}_i) \triangleq \nabla Q(w; \boldsymbol{x}_i) \tag{3.5}$$

© 2023. All rights reserved. This draft cannot be copied or distributed in print or electronically without the written consent of the authors S. Vlaski and A. H. Sayed.

and iterate as an approximation to (3.2) instead:

$$\boldsymbol{w}_i = \boldsymbol{w}_{i-1} - \mu\widehat{\nabla J}(\boldsymbol{w}_{i-1};\boldsymbol{x}_i) = \boldsymbol{w}_{i-1} - \mu\nabla Q(\boldsymbol{w}_{i-1};\boldsymbol{x}_i) \tag{3.6}$$

Observe that the stochastic gradient approximation $\nabla Q(\boldsymbol{w}_{i-1};\boldsymbol{x}_i)$ is a function of the random variable $\boldsymbol{x}_i$, and is hence random itself. This randomness propagates into the recursion describing the evolution of $\boldsymbol{w}_i$; a fact that we emphasize by employing bold font for the now random iterates $\boldsymbol{w}_i$. Nevertheless, under fairly weak conditions, which allow for the exchange of expectation and gradient operations, it then holds that:

$$\mathbb{E}\,\widehat{\nabla J}(w;\boldsymbol{x}_i) = \mathbb{E}\,\nabla Q(w;\boldsymbol{x}_i) = \nabla\mathbb{E}\,Q(w;\boldsymbol{x}_i) = \nabla J(w) \tag{3.7}$$

Hence, the instantaneous approximation $\nabla Q(w;\boldsymbol{x}_i)$ can be seen to be an *unbiased* estimate of the gradient $\nabla J(\boldsymbol{w}_{i-1})$. We then expect the stochastic recursion (3.6) to track the evolution of the deterministic gradient recursion (3.2) with reasonable accuracy (at least in the mean).

Example 3.2 (**Empirical risk minimization**) A second class of problems that appear in learning contexts are empirical risk minimization problems of the form:

$$J(w) \triangleq \frac{1}{N}\sum_{n=1}^{N} Q(w;x_n) \tag{3.8}$$

where the set of points $\{x_n\}_{n=1}^N$ corresponds to a batch of N independent realizations of the random variable $\boldsymbol{x}$. In contrast to the previous exmaple, where evaluation of the gradient $\nabla J(w)$ is infeasible, for empirical risk minimization problems, it is in principle possible to iterate:

$$w_i = w_{i-1} - \mu\nabla J(w_{i-1}) = w_{i-1} - \frac{\mu}{N}\sum_{n=0}^{N}\nabla Q(w_{i-1};x_n) \tag{3.9}$$

The drawback of the gradient descent recursion, applied to empirical risk minimization problems of the form (3.8), is that every iteration requires the evaluation of N gradients of the risk $\nabla Q(w_{i-1};x_n)$. This results in a costly procedure, particular in the large-scale setting where the sample size N is large. We may then construct a stochastic approximation to the gradient recursion (3.9) as follows. At any time i, we sample from the set of data points $\{x_n\}_{n=1}^N$ uniformly at random. We may model this by defining a random index $\boldsymbol{n}_i$, with uniform distribution:

$$\boldsymbol{n}_i = \begin{cases} 1, & \text{with probability } \frac{1}{N}, \\ 2, & \text{with probability } \frac{1}{N}, \\ \vdots & \\ N, & \text{with probability } \frac{1}{N}. \end{cases} \tag{3.10}$$

Then $x_{\boldsymbol{n}_i}$ denotes the sample picked randomly from $\{x_n\}_{n=1}^N$ at time i. We may then construct a stochastic approximation of the batch gradient as:

$$\widehat{\nabla J}(w; x_{\boldsymbol{n}_i}) \triangleq \nabla Q(w; x_{\boldsymbol{n}_i}) \tag{3.11}$$

and obtain a stochastic variant of the gradient recursion (3.9) as:

$$\boldsymbol{w}_i = \boldsymbol{w}_{i-1} - \mu \nabla Q(\boldsymbol{w}_{i-1}, x_{\boldsymbol{n}_i}) \tag{3.12}$$

Observe that again we have the property:

$$\mathbb{E}\, \widehat{\nabla J}(w; x_{\boldsymbol{n}_i}) = \mathbb{E}\, \nabla Q(w; x_{\boldsymbol{n}_i}) = \frac{1}{N} \sum_{n=1}^{N} Q(w; x_n) = \nabla J(w) \tag{3.13}$$

Note that if we define $\boldsymbol{x}_i^{\text{emp}} \triangleq x_{\boldsymbol{n}_i}$, it follows that:

$$J(w) \triangleq \frac{1}{N} \sum_{n=1}^{N} Q(w; x_n) = \mathbb{E}_{\boldsymbol{x}_i^{\text{emp}}} Q(w; \boldsymbol{x}_i^{\text{emp}}) = \mathbb{E}_{\boldsymbol{x}^{\text{emp}}} Q(w; \boldsymbol{x}^{\text{emp}}) \tag{3.14}$$

where the last equality holds since the indices $\boldsymbol{n}_i$ follow identical distributions over time, and we can hence replace $\boldsymbol{n}_i$ by $\boldsymbol{n}$. In other words, the empirical risk minimization problem (3.8) is a special case of the expected risk minimization problem (3.3), where we define a new random variable $\boldsymbol{x}^{\text{emp}} \triangleq x_{\boldsymbol{n}}$, which follows a uniform empirical distribution over the sample batch of data $\{x_n\}_{n=1}^N$. Note that for any finite N, there will be a difference between $\boldsymbol{x}$, the underlying random variable that led to the batch of samples $\{x_n\}_{n=1}^N$, and $\boldsymbol{x}^{\text{emp}}$, which is sampled from $\{x_n\}_{n=1}^N$. Hence, it will generally be the case that:

$$\mathbb{E}_{\boldsymbol{x}} Q(w; \boldsymbol{x}) \neq \mathbb{E}_{\boldsymbol{x}^{\text{emp}}} Q(w; \boldsymbol{x}^{\text{emp}}) \tag{3.15}$$

This observation means that while the empirical risk for a sample $\{x_n\}_n^N$ of a random variable $\boldsymbol{x}$ is in general distinct from the expected risk for $\boldsymbol{x}$, we can define a new random variable x^{emp} for which the empirical risk corresponds to an expected risk. In this sense, we may view empirical risk minimization as a *special case* of expected risk minimization. Most of our discussion in the sequel will be focused on the expected risk minimization problem, with the understanding that we can always recover results and algorithms for empirical risk minimization via (3.14).

3.1 STOCHASTIC GRADIENT DESCENT

The previous examples illustrate how the construction of *stochastic gradient approximations* based on sampled data can facilitate the optimization of data-dependent risk functions when exact gradient are either inaccessible (as in the case of expected risk minimization, Example 3.1) or computationally expensive

(as in the case of empirical risk minimization, example 3.2). In both cases, we were able to verify that the constructions are unbiased, and hence we expect the behavior of the resulting stochastic gradient algorithms to approximate that of their deterministic counterparts in some sense that we are yet to make precise. At the same time, we expect some degradation in performance resulting from the fact that approximations still exhibit some variance around the true gradient $\nabla J(\cdot)$. In this chapter, we will quantify in great detail the effect of utilizing stochastic gradient approximations on learning performance, and show how these insight can be leveraged to make informed decisions when designing learning algorithms.

Definition 3.1 (**Stochastic Gradient Approximation**). Given a risk function $J(w)$ we denote by $\widehat{\nabla J}(w)$ a stochastic gradient approximation of the true gradient $\nabla J(w)$, if $\widehat{\nabla J}(w)$ can be evaluated solely from data available to the algorithm, and yet:

$$\widehat{\nabla J}(w) \approx \nabla J(w) \tag{3.16}$$

in some sense. On occasion, we may wish to make precise the exact data used to compute $\widehat{\nabla J}(w)$, in which case we may for example write:

$$\widehat{\nabla J}(w; \boldsymbol{x}_i) \approx \nabla J(w) \tag{3.17}$$

to emphasize that $\widehat{\nabla J}(w; \boldsymbol{x}_i)$ is constructed based on a realization of the random variable $\boldsymbol{x}_i$. As a general rule, we will employ $\widehat{\nabla J}(w)$ whenever the data used to construct the gradient approximation is either clear from context (to lighten the notation), or irrelevant to the discussion (to keep results general). □

Note that our definition of a stochastic gradient approximation is rather generic. Indeed, while the two constructions in Examples 3.1 and 3.2 were both unbiased, it is in possible to effectively employ biased approximations in certain settings. In this chapter, we will primarily focus on unbiased gradient approximations, and will explore biased constructions in later chapters.

3.1.1 Gradient Noise

Motivated by the discussion so far, we will now study generic expected risk minimization problems of the form:

$$\arg\min_{w} J(w) \triangleq \arg\min \mathbb{E}_{\boldsymbol{x}} Q(w; \boldsymbol{x}) \tag{3.18}$$

with the understanding that this formulation covers empirical risk minimization problems as laid out in Example 3.2. We will consider the stochastic gradient algorithm with generic gradient approximation:

$$\boldsymbol{w}_i = \boldsymbol{w}_{i-1} - \mu \widehat{\nabla J}(\boldsymbol{w}_{i-1}) \tag{3.19}$$

It is instructive to reformulate the recursion as:

$$\boldsymbol{w}_i = \boldsymbol{w}_{i-1} - \mu\nabla J(\boldsymbol{w}_{i-1}) - \mu\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \tag{3.20}$$

Here, we introduce the *gradient noise* term:

$$\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) = \widehat{\nabla J}(\boldsymbol{w}_{i-1}) - \nabla J(\boldsymbol{w}_{i-1}) \tag{3.21}$$

In light of this definition, we may view (3.20) as a perturbed version of the deterministic gradient recursion (3.2). The choice of indices in defining $\boldsymbol{s}_i(\boldsymbol{w}_{i-1})$ is deliberate. The gradient approximation $\widehat{\nabla J}(\boldsymbol{w}_{i-1})$ is constructed using data available at time i and evaluated at $\boldsymbol{w}_{i-1}$. Hence, $\boldsymbol{s}_i(\cdot)$ will be determined by data at time i, and will be a function of $\boldsymbol{w}_{i-1}$. How closely (3.20) tracks the original gradient descent recursion (3.2) will then depend on the statistical properties of the gradient noise $\boldsymbol{s}_i(\boldsymbol{w}_{i-1})$, which is in turn driven by the quality of the gradient approximation $\widehat{\nabla J}(\boldsymbol{w}_{i-1})$. In this chapter, we will study the performance of stochastic gradient algorithms under the following conditions on the first and second-order moment of the gradient noise:

Condition 3.1 (**Gradient Noise Conditions**). The gradient approximations $\widehat{\nabla J}(\boldsymbol{w}_{i-1})$ are unbiased conditioned on the iterate $\boldsymbol{w}_{i-1}$. Specifically:

$$\mathbb{E}\left\{\widehat{\nabla J}(\boldsymbol{w}_{i-1})|\boldsymbol{w}_{i-1}\right\} = \nabla J(\boldsymbol{w}_{i-1}) \Longleftrightarrow \mathbb{E}\left\{\boldsymbol{s}_i(\boldsymbol{w}_{i-1})|\boldsymbol{w}_{i-1}\right\} = 0 \tag{3.22}$$

Unbiased gradient estimates ensure that the gradient approximation $\widehat{\nabla J}(\boldsymbol{w}_{i-1})$ is, on average, a good estimate of the descent direction. Unbiased estimates are good estimates, but only if their variance is bounded. We hence impose an additional condition:

$$\begin{aligned}\mathbb{E}\left\{\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2|\boldsymbol{w}_{i-1}\right\} \leq \alpha^2\|\nabla J(\boldsymbol{w}_{i-1})\|^2 + \beta^2\|w^o - \boldsymbol{w}_{i-1}\|^2 \\ + \gamma^2\left(J(\boldsymbol{w}_{i-1}) - J^o\right) + \sigma^2\end{aligned} \tag{3.23}$$

This second condition imposes a *relative* bound on the variance of the gradient approximation because we allow the righthand-side of (3.23) to grow with various measures of suboptimality. □

We refer to Condition 3.1 as a *condition*, rather than an assumption, because for optimization problems that arise from most learning applications, we are able to establish relations (3.22) and (3.23) for suitably chosen constants $\alpha, \beta, \gamma, \sigma > 0$. We illustrate this in the following sections.

3.1.2 Ordinary Stochastic Gradient Descent

Perhaps the most commonly employed generic gradient approximation is obtained by from $\nabla J(w) \triangleq \nabla\left\{\mathbb{E}\, Q(w;\boldsymbol{x})\right\}$ by simply dropping the expected value and evaluating the loss using a single sample $\boldsymbol{x}_i$ available at time i. This results in:

$$\widehat{\nabla J}^{\text{ord}}(w) \triangleq \nabla Q(w;\boldsymbol{x}_i) \tag{3.24}$$

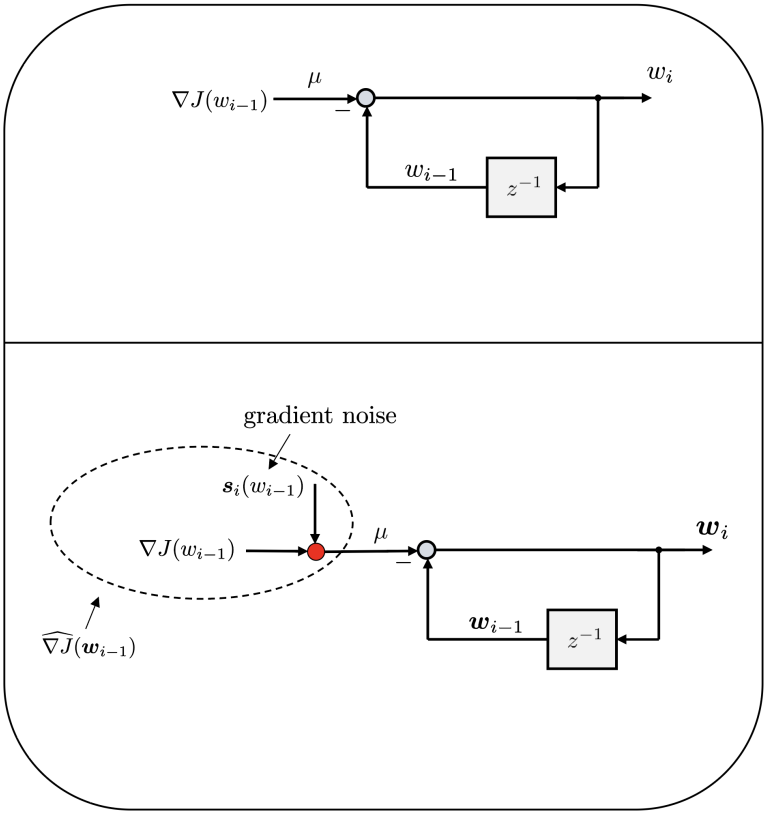

Figure 3.1 The panel on top shows the dynamics of the original gradient-descent recursion (3.2) for empirical risk minimization while the panel in the bottom shows the dynamics of the stochastic-gradient recursion (3.20). The true gradient vector is perturbed by the gradient noise process (3.21), which seeps into the operation of the algorithm. The block with z^{-1} represents a unit delay element.

We will be referring to such approximations as *ordinary* stochastic gradient approximations, as they are the most common and simplest constructions. When we wish to emphasize the fact that the construction is of the ordinary type, or to contrast it from other constructions, we will do so by attaching "ord" as a super- or subscript to the relevant quantities. For example, $\widehat{\nabla J}^{\,\mathrm{ord}}(w)$ refers to the ordinary gradient approximation constructed according to (3.24), while $\boldsymbol{s}_i^{\mathrm{ord}}(w)$ refers to the induced gradient noise and $\alpha^2_{\mathrm{ord}}, \beta^2_{\mathrm{ord}}, \gamma^2_{\mathrm{ord}}, \sigma^2_{\mathrm{ord}}$ the resulting gradient noise constants. When clear from context or irrelevant to the discussion, we will omit "ord" for notational convenience and generality.

Despite their simplicity, ordinary gradient approximations are commonly used and effective in practice, and will serve as fundamental building blocks of many variations of stochastic gradient descent we will encounter in Section 3.1.3 further

ahead. First, we illustrate the application of ordinary gradient approximations in the cases of least-mean square and logistic regression.

Example 3.3 (**Least-Mean Square**) Let us consider a linear model:

$$\boldsymbol{\gamma} = \boldsymbol{h}^{\mathsf{T}} w^o + \boldsymbol{v} \tag{3.25}$$

where $\boldsymbol{h} \in \mathbb{R}^M$ denotes the regressor and $\boldsymbol{\gamma}$ denotes the observation. The measurement noise is denoted by $\boldsymbol{v}$ and assumed to be zero-mean. Given observations $\boldsymbol{x} \triangleq \mathrm{col}\{\boldsymbol{h}, \boldsymbol{\gamma}\}$, we may pursue w^o via:

$$w^o = \arg\min_{w} \mathbb{E}\,\|\boldsymbol{\gamma} - \boldsymbol{h}^{\mathsf{T}} w\|^2 \tag{3.26}$$

For notational convenience and consistency, we define $\boldsymbol{x} \triangleq \mathrm{col}\{\boldsymbol{h}, \boldsymbol{\gamma}\}$ and hence $Q(w; \boldsymbol{x}) = Q(w; \boldsymbol{h}, \boldsymbol{\gamma}) = \|\boldsymbol{\gamma} - \boldsymbol{h}^{\mathsf{T}} w\|^2$. By differentiating, we find:

$$\nabla J(w) = \mathbb{E}\,\boldsymbol{h}\left(\boldsymbol{\gamma} - \boldsymbol{h}^{\mathsf{T}} w\right) = r_{\boldsymbol{h}\boldsymbol{\gamma}} - R_{\boldsymbol{h}} w \tag{3.27}$$

where we defined $r_{\boldsymbol{h}\boldsymbol{\gamma}} = \mathbb{E}\,\boldsymbol{h}\boldsymbol{\gamma}$ and $R_{\boldsymbol{h}} \triangleq \mathbb{E}\,\boldsymbol{h}\boldsymbol{h}^{\mathsf{T}}$. Following Example 3.1, we can then construct:

$$\widehat{\nabla J}(w) = \nabla Q(w; \boldsymbol{x}_i) = \boldsymbol{h}_i\left(\boldsymbol{\gamma}_i - \boldsymbol{h}_i^{\mathsf{T}} w\right) = \boldsymbol{h}_i\boldsymbol{\gamma}_i - \boldsymbol{h}_i\boldsymbol{h}_i^{\mathsf{T}} w \tag{3.28}$$

We can then compute the gradient noise $\boldsymbol{s}_i(w)$ as:

$$\boldsymbol{s}_i(w) = \boldsymbol{h}_i\boldsymbol{\gamma}_i - \boldsymbol{h}_i\boldsymbol{h}_i^{\mathsf{T}} w - r_{\boldsymbol{h}\boldsymbol{\gamma}} + R_{\boldsymbol{h}} w \tag{3.29}$$

To verify the zero-mean condition (3.22), compute:

$$\begin{aligned}
&\mathbb{E}\,\{\boldsymbol{s}_i(\boldsymbol{w}_{i-1})|\boldsymbol{w}_{i-1}\} \\
&= \mathbb{E}\left\{\boldsymbol{h}_i\boldsymbol{\gamma}_i - \boldsymbol{h}_i\boldsymbol{h}_i^{\mathsf{T}}\boldsymbol{w}_{i-1} - r_{\boldsymbol{h}\boldsymbol{\gamma}} + R_{\boldsymbol{h}}\boldsymbol{w}_{i-1}|\boldsymbol{w}_{i-1}\right\} \\
&= \mathbb{E}\,\{\boldsymbol{h}_i\boldsymbol{\gamma}_i|\boldsymbol{w}_{i-1}\} - \mathbb{E}\left\{\boldsymbol{h}_i\boldsymbol{h}_i^{\mathsf{T}}\boldsymbol{w}_{i-1}|\boldsymbol{w}_{i-1}\right\} - \mathbb{E}\,\{r_{\boldsymbol{h}\boldsymbol{\gamma}}|\boldsymbol{w}_{i-1}\} + \mathbb{E}\,\{R_{\boldsymbol{h}}\boldsymbol{w}_{i-1}|\boldsymbol{w}_{i-1}\} \\
&\overset{(a)}{=} \mathbb{E}\,\{\boldsymbol{h}_i\boldsymbol{\gamma}_i|\boldsymbol{w}_{i-1}\} - \mathbb{E}\left\{\boldsymbol{h}_i\boldsymbol{h}_i^{\mathsf{T}}|\boldsymbol{w}_{i-1}\right\}\boldsymbol{w}_{i-1} - r_{\boldsymbol{h}\boldsymbol{\gamma}} + R_{\boldsymbol{h}}\boldsymbol{w}_{i-1}
\end{aligned} \tag{3.30}$$

where (a) follows after accounting for deterministic terms when conditioning on $\boldsymbol{w}_{i-1}$. To simplify the expression further, we need to argue that $\boldsymbol{h}_i$ and $\boldsymbol{\gamma}_i$ are independent of $\boldsymbol{w}_{i-1}$. A sufficient condition for independence is to assume that $\{\boldsymbol{h}_i, \boldsymbol{\gamma}_i\}$ are sampled *independently* over time. Since $\boldsymbol{w}_{i-1}$ is constructed with data up to and including time-instant $i-1$, it then follows that $\{\boldsymbol{h}_i, \boldsymbol{\gamma}_i\}$ are independent of $\boldsymbol{w}_{i-1}$. Then the conditioning can be removed from (3.30) and we obtain:

$$\mathbb{E}\,\{\boldsymbol{s}_i(\boldsymbol{w}_{i-1})|\boldsymbol{w}_{i-1}\} = \mathbb{E}\,\{\boldsymbol{h}_i\boldsymbol{\gamma}_i\} - \mathbb{E}\left\{\boldsymbol{h}_i\boldsymbol{h}_i^{\mathsf{T}}\right\}\boldsymbol{w}_{i-1} - r_{\boldsymbol{h}\boldsymbol{\gamma}} + R_{\boldsymbol{h}}\boldsymbol{w}_{i-1} = 0 \tag{3.31}$$

To compute the variance, it will be useful to reformulate expressions in terms of the regressor $\boldsymbol{h}$ and noise $\boldsymbol{v}$, rather than the regressor $\boldsymbol{h}$ and the observation

$\boldsymbol{\gamma}$. This is because regressor and noise are assumed to be independent, while regressor and observation are not. To this end:

$$\begin{aligned}
r_{\boldsymbol{h}\boldsymbol{\gamma}} &= \mathbb{E}\boldsymbol{h}\boldsymbol{\gamma} \\
&= \mathbb{E}\left\{\boldsymbol{h}\left(\boldsymbol{h}^{\mathsf{T}}w^o + \boldsymbol{v}\right)\right\} \\
&= \left(\mathbb{E}\boldsymbol{h}\boldsymbol{h}^{\mathsf{T}}\right)w^o + \mathbb{E}\boldsymbol{h}\boldsymbol{v} \\
&\overset{(a)}{=} \left(\mathbb{E}\boldsymbol{h}\boldsymbol{h}^{\mathsf{T}}\right)w^o + (\mathbb{E}\boldsymbol{h})\cdot(\mathbb{E}\boldsymbol{v}) \\
&\overset{(b)}{=} R_h w^o
\end{aligned} \tag{3.32}$$

Here, (a) follows since $\boldsymbol{h}$ and $\boldsymbol{v}$ are independent, and (b) follows since the noise is zero-mean. For the gradient noise, we then find:

$$\begin{aligned}
\boldsymbol{s}_i(w) &= \boldsymbol{h}_i\boldsymbol{\gamma}_i - \boldsymbol{h}_i\boldsymbol{h}_i^{\mathsf{T}}w - r_{\boldsymbol{h}\boldsymbol{\gamma}} + R_{\boldsymbol{h}}w \\
&= \boldsymbol{h}_i\left(\boldsymbol{h}_i^{\mathsf{T}}w^o + \boldsymbol{v}_i\right) - \boldsymbol{h}_i\boldsymbol{h}_i^{\mathsf{T}}w - R_{\boldsymbol{h}}w^o + R_{\boldsymbol{h}}w \\
&= \left(\boldsymbol{h}_i\boldsymbol{h}_i^{\mathsf{T}} - R_{\boldsymbol{h}}\right)(w^o - w) + \boldsymbol{v}_i\boldsymbol{h}_i
\end{aligned} \tag{3.33}$$

We recognize a relative component $\left(\boldsymbol{h}_i\boldsymbol{h}_i^{\mathsf{T}} - R_{\boldsymbol{h}}\right)(w^o - w)$, which vanishes for $w = w^o$, and an absolute component $\boldsymbol{v}_i\boldsymbol{h}_i$, which remains irrespective of the quality of the estimate w. For the variance, we then have:

$$\begin{aligned}
&\mathbb{E}\left\{\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2|\boldsymbol{w}_{i-1}\right\} \\
&= \mathbb{E}\left\{\left\|\left(\boldsymbol{h}_i\boldsymbol{h}_i^{\mathsf{T}} - R_{\boldsymbol{h}}\right)(w^o - \boldsymbol{w}_{i-1}) + \boldsymbol{v}_i\boldsymbol{h}_i\right\|^2|\boldsymbol{w}_{i-1}\right\} \\
&= \mathbb{E}\left\{\left\|\boldsymbol{h}_i\boldsymbol{h}_i^{\mathsf{T}} - R_{\boldsymbol{h}}\right\|^2\right\}\|w^o - \boldsymbol{w}_{i-1}\|^2 + \mathbb{E}\|\boldsymbol{v}_i\boldsymbol{h}_i\|^2 \\
&\quad + 2\mathbb{E}\left\{\boldsymbol{v}_i\boldsymbol{h}_i^{\mathsf{T}}\left(\boldsymbol{h}_i\boldsymbol{h}_i^{\mathsf{T}} - R_{\boldsymbol{h}}\right)\right\}(w^o - \boldsymbol{w}_{i-1}) \\
&\overset{(a)}{=} \mathbb{E}\left\{\left\|\boldsymbol{h}_i\boldsymbol{h}_i^{\mathsf{T}} - R_{\boldsymbol{h}}\right\|^2\right\}\|w^o - \boldsymbol{w}_{i-1}\|^2 + \left(\mathbb{E}\|\boldsymbol{v}_i\|^2\right)\cdot\left(\mathbb{E}\|\boldsymbol{h}_i\|^2\right) \\
&\quad + 2\mathbb{E}\{\boldsymbol{v}_i\}\,\mathbb{E}\left\{\boldsymbol{h}_i^{\mathsf{T}}\left(\boldsymbol{h}_i\boldsymbol{h}_i^{\mathsf{T}} - R_{\boldsymbol{h}}\right)\right\}(w^o - \boldsymbol{w}_{i-1}) \\
&\overset{(b)}{=} \mathbb{E}\left\{\left\|\boldsymbol{h}_i\boldsymbol{h}_i^{\mathsf{T}} - R_{\boldsymbol{h}}\right\|^2\right\}\|w^o - \boldsymbol{w}_{i-1}\|^2 + \left(\mathbb{E}\|\boldsymbol{v}_i\|^2\right)\cdot\left(\mathbb{E}\|\boldsymbol{h}_i\|^2\right) \\
&= \mathbb{E}\left\{\left\|\boldsymbol{h}_i\boldsymbol{h}_i^{\mathsf{T}} - R_{\boldsymbol{h}}\right\|^2\right\}\|w^o - \boldsymbol{w}_{i-1}\|^2 + \sigma_{\boldsymbol{v}}^2\mathrm{Tr}\left(R_{\boldsymbol{h}}\right) \\
&\overset{(c)}{=} \beta^2\|w^o - \boldsymbol{w}_{i-1}\|^2 + \sigma^2
\end{aligned} \tag{3.34}$$

where (a) follows by independence of $\boldsymbol{h}_i$, $\boldsymbol{v}_i$ and $\boldsymbol{w}_{i-1}$, and (b) follows since $\mathbb{E}\boldsymbol{v}_i = 0$. In (c) we defined:

$$\beta^2 \triangleq \mathbb{E}\left\|\boldsymbol{h}_i\boldsymbol{h}_i^{\mathsf{T}} - R_{\boldsymbol{h}}\right\|^2 \tag{3.35}$$

$$\sigma^2 \triangleq \sigma_{\boldsymbol{v}}^2 \mathrm{Tr}\left(R_{\boldsymbol{h}}\right) \tag{3.36}$$

We conclude that the stochastic approximation (3.28) of the mean square error cost (3.26) satisfies the gradient noise conditions (3.22)–(3.23) with $\alpha^2 = \gamma^2 = 0$ and β^2, σ^2 as defined in (3.35)–(3.36). A notable fact is that in the case of the mean square error cost, the variance of the gradient noise (3.34) can be computed in closed form, and includes a relative component of the form $\beta^2\|w^o - \boldsymbol{w}_{i-1}\|^2$. This illustrates that allowing and accounting for relative gradient noise components is important when developing performance analysis of stochastic gradient algorithms.

Example 3.4 (**Logistic Regression**) If the labels $\boldsymbol{\gamma}$ are sampled from a discrete distribution, say $\boldsymbol{\gamma} = \pm 1$, the least squares loss may not be appropriate. We can then resort to the logistic loss:

$$Q(w; \boldsymbol{h}_i, \boldsymbol{\gamma}_i) \triangleq \ln\left(1 + e^{-\boldsymbol{\gamma}_i \boldsymbol{h}_i^{\mathsf{T}} w}\right) + \frac{\rho}{2}\|w\|^2 \tag{3.37}$$

$$J(w) \triangleq \mathbb{E}Q(w; \boldsymbol{h}_i, \boldsymbol{\gamma}_i) = \mathbb{E}\ln\left(1 + e^{-\boldsymbol{\gamma}_i \boldsymbol{h}_i^{\mathsf{T}} w}\right) + \frac{\rho}{2}\|w\|^2 \tag{3.38}$$

Then, it follows that:

$$\begin{aligned} \nabla Q(w; \boldsymbol{h}_i, \boldsymbol{\gamma}_i) &= -\frac{\boldsymbol{\gamma}_i \boldsymbol{h}_i}{1 + e^{\boldsymbol{\gamma}_i \boldsymbol{h}_i^{\mathsf{T}} w}} + \rho w \\ \nabla J(w) &= -\mathbb{E}\left\{\frac{\boldsymbol{\gamma}_i \boldsymbol{h}_i}{1 + e^{\boldsymbol{\gamma}_i \boldsymbol{h}_i^{\mathsf{T}} w}}\right\} + \rho w \end{aligned} \tag{3.39}$$

Hence, we can construct again $\widehat{\nabla J}(w) = \nabla Q(w; \boldsymbol{h}_i, \boldsymbol{\gamma}_i)$ and immediately verify the zero-mean condition (3.22). For simplicity, in the following, we assume an equal prior $\Pr\{\boldsymbol{\gamma}_i = +1\} = \Pr\{\boldsymbol{\gamma}_i = -1\} = \frac{1}{2}$. For the variance of the gradient

noise, we have:

$$
\begin{aligned}
&\mathbb{E}\left\{\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2|\boldsymbol{w}_{i-1}\right\} \\
&=\mathbb{E}\left\{\|\widehat{\nabla J}(\boldsymbol{w}_{i-1})-\nabla J(\boldsymbol{w}_{i-1})\|^2|\boldsymbol{w}_{i-1}\right\} \\
&=\mathbb{E}\left\{\|\nabla Q(\boldsymbol{w}_{i-1};\boldsymbol{h}_i,\boldsymbol{\gamma}_i)-\nabla J(\boldsymbol{w}_{i-1})\|^2|\boldsymbol{w}_{i-1}\right\} \\
&=\mathbb{E}\left\{\left\|-\frac{\boldsymbol{\gamma}_i\boldsymbol{h}_i}{1+e^{\boldsymbol{\gamma}_i\boldsymbol{h}_i^{\mathsf{T}}\boldsymbol{w}_{i-1}}}+\rho\boldsymbol{w}_{i-1}+\mathbb{E}\left\{\frac{\boldsymbol{\gamma}_i\boldsymbol{h}_i}{1+e^{\boldsymbol{\gamma}_i\boldsymbol{h}_i^{\mathsf{T}}\boldsymbol{w}_{i-1}}}|\boldsymbol{w}_{i-1}\right\}-\rho\boldsymbol{w}_{i-1}\right\|^2|\boldsymbol{w}_{i-1}\right\} \\
&=\mathbb{E}\left\{\left\|\frac{\boldsymbol{\gamma}_i\boldsymbol{h}_i}{1+e^{\boldsymbol{\gamma}_i\boldsymbol{h}_i^{\mathsf{T}}\boldsymbol{w}_{i-1}}}-\mathbb{E}\left\{\frac{\boldsymbol{\gamma}_i\boldsymbol{h}_i}{1+e^{\boldsymbol{\gamma}_i\boldsymbol{h}_i^{\mathsf{T}}\boldsymbol{w}_{i-1}}}|\boldsymbol{w}_{i-1}\right\}\right\|^2|\boldsymbol{w}_{i-1}\right\} \\
&\overset{(a)}{=}\mathbb{E}\left\{\left\|\frac{\boldsymbol{\gamma}_i\boldsymbol{h}_i}{1+e^{\boldsymbol{\gamma}_i\boldsymbol{h}_i^{\mathsf{T}}\boldsymbol{w}_{i-1}}}\right\|^2|\boldsymbol{w}_{i-1}\right\}-\left\|\mathbb{E}\left\{\frac{\boldsymbol{\gamma}_i\boldsymbol{h}_i}{1+e^{\boldsymbol{\gamma}_i\boldsymbol{h}_i^{\mathsf{T}}\boldsymbol{w}_{i-1}}}|\boldsymbol{w}_{i-1}\right\}\right\|^2 \\
&\leq\mathbb{E}\left\{\left\|\frac{\boldsymbol{\gamma}_i\boldsymbol{h}_i}{1+e^{\boldsymbol{\gamma}_i\boldsymbol{h}_i^{\mathsf{T}}\boldsymbol{w}_{i-1}}}\right\|^2|\boldsymbol{w}_{i-1}\right\} \\
&=\mathbb{E}\left\{\frac{\|\boldsymbol{\gamma}_i\boldsymbol{h}_i\|^2}{\left(1+e^{\boldsymbol{\gamma}_i\boldsymbol{h}_i^{\mathsf{T}}\boldsymbol{w}_{i-1}}\right)^2}|\boldsymbol{w}_{i-1}\right\} \\
&\overset{(b)}{\leq}\mathbb{E}\left\{\|\boldsymbol{\gamma}_i\|^2\|\boldsymbol{h}_i\|^2|\boldsymbol{w}_{i-1}\right\} \\
&\overset{(c)}{=}\mathbb{E}\|\boldsymbol{h}_i\|^2=\mathrm{Tr}\left(R_{\boldsymbol{h}}\right)
\end{aligned}
\tag{3.40}
$$

Here, (a) follows after cross-multiplying, (b) follows since $\left(1+e^{\boldsymbol{\gamma}_i\boldsymbol{h}_i^{\mathsf{T}}\boldsymbol{w}_{i-1}}\right)^2\geq 1$ almost surely, and (c) is due to the fact that $\boldsymbol{\gamma}_i\in\pm 1$, and hence $\boldsymbol{\gamma}_i^2=1$ almost surely. We conclude that the gradient approximation $\nabla J(w)=\nabla Q(w;\boldsymbol{h}_i,\boldsymbol{\gamma}_i)$ satisfies the gradient noise conditions (3.22)–(3.23) with $\alpha^2=\beta^2=\gamma^2=0$ and $\sigma^2=\mathrm{Tr}(R_{\boldsymbol{h}})$.

General Relations for Ordinary Gradient Approximations

The examples of linear and logistic regression demonstrated that the ordinary gradient approximation satisfies the gradient noise conditions (3.22)–(3.23) in applications of significant practical importance. It turns out that the ordinary gradient approximation satisfies (3.22)–(3.23) for a broad family of loss functions $Q(\cdot;\cdot)$ and distributions of the data $\boldsymbol{x}$.

3.1.3 Variants of Stochastic Gradient Descent

In the previous section, we saw that the gradient approximation $\widehat{\nabla J}^{\,\mathrm{ord}}(w)\triangleq\nabla Q(w;\boldsymbol{x}_i)$ satisfies the gradient noise conditions (3.22)–(3.23) for many risk

functions of interest. The resulting constants quantify the quality of the gradient approximation and are a function of the statistical properties of the data $\boldsymbol{x}_i$, along with properties of the loss $Q(\cdot;\cdot)$. Importantly, these are immutable and generally not under the control of the algorithm designer. As we will see in the analysis further ahead, the stability and performance of stochastic gradient algorithms will depend critically on the noise constants $\alpha^2, \beta^2, \gamma^2, \sigma^2$. In this section, we will show how these constants can be modified by constructing alternative gradient approximations.

Mini-Batch Stochastic Gradient Descent

Suppose instead of being provided with a single data point $\boldsymbol{x}_i$ at any given time i, we have access to B independent samples $\{\boldsymbol{x}_{b,i}\}_{b=1}^B$. Clearly, the fact that multiple samples are available conveys additional information about the distribution of $\boldsymbol{x}$, and should be exploited in construction the gradient approximation $\widehat{\nabla J}(\boldsymbol{w}_{i-1})$ of $\nabla J(\boldsymbol{w}_{i-1})$. One such approach is to compute the average:

$$\widehat{\nabla J}(\boldsymbol{w}_{i-1};\{\boldsymbol{x}_{b,i}\}_{b=1}^B) \triangleq \frac{1}{B}\sum_{b=1}^{B}\nabla Q(\boldsymbol{w}_{i-1};\boldsymbol{x}_{b,i}) \tag{3.41}$$

It is straightforward to verify that:

$$\begin{aligned}
&\mathbb{E}\left\{\widehat{\nabla J}(\boldsymbol{w}_{i-1};\{\boldsymbol{x}_{b,i}\}_{b=1}^B)|\boldsymbol{w}_{i-1}\right\}\\
&=\mathbb{E}\left\{\frac{1}{B}\sum_{b=1}^{B}\nabla Q(\boldsymbol{w}_{i-1};\boldsymbol{x}_{b,i})\right\}\\
&=\frac{1}{B}\sum_{b=1}^{B}\mathbb{E}\left\{\nabla Q(\boldsymbol{w}_{i-1};\boldsymbol{x}_{b,i})\right\}\\
&\overset{(a)}{=}\frac{1}{B}\sum_{b=1}^{B}\nabla\mathbb{E}\left\{Q(\boldsymbol{w}_{i-1};\boldsymbol{x}_{b,i})\right\}\\
&=\frac{1}{B}\sum_{b=1}^{B}\nabla J(\boldsymbol{w}_{i-1})\\
&=J(\boldsymbol{w}_{i-1})
\end{aligned} \tag{3.42}$$

Step (a) follows whenever the expectation and gradient operations can be exchanged, which is the case under fairly loose regularity conditions on the loss $Q(\cdot;\cdot)$ and data distributions. For the variance, we expect a variance reduction, since the estimate (3.41) is constructed by averaging multiple realizations of

$\nabla Q(\boldsymbol{w}_{i-1};\boldsymbol{x}_{b,i})$. Indeed:

$$\begin{aligned}
&\mathbb{E}\left\{\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2|\boldsymbol{w}_{i-1}\right\}\\
&=\mathbb{E}\left\{\left\|\widehat{\nabla J}\left(\boldsymbol{w}_{i-1};\{\boldsymbol{x}_{b,i}\}_{b=1}^{B}\right)-\nabla J(\boldsymbol{w}_{i-1})\right\|^2|\boldsymbol{w}_{i-1}\right\}\\
&=\mathbb{E}\left\{\left\|\frac{1}{B}\sum_{b=1}^{B}\nabla Q(\boldsymbol{w}_{i-1};\boldsymbol{x}_{b,i})-\nabla J(\boldsymbol{w}_{i-1})\right\|^2|\boldsymbol{w}_{i-1}\right\}\\
&=\mathbb{E}\left\{\left\|\frac{1}{B}\sum_{b=1}^{B}(\nabla Q(\boldsymbol{w}_{i-1};\boldsymbol{x}_{b,i})-\nabla J(\boldsymbol{w}_{i-1}))\right\|^2|\boldsymbol{w}_{i-1}\right\}\\
&\overset{(a)}{=}\frac{1}{B^2}\sum_{b=1}^{B}\mathbb{E}\left\{\|(\nabla Q(\boldsymbol{w}_{i-1};\boldsymbol{x}_{b,i})-\nabla J(\boldsymbol{w}_{i-1}))\|^2|\boldsymbol{w}_{i-1}\right\}\\
&\overset{(b)}{=}\frac{1}{B^2}\sum_{b=1}^{B}\mathbb{E}\left\{\left\|\boldsymbol{s}_{b,i}^{\mathrm{ord}}(\boldsymbol{w}_{i-1})\right\|^2|\boldsymbol{w}_{i-1}\right\}
\end{aligned}\tag{3.43}$$

In (a), we used the fact that samples $\boldsymbol{x}_{b,i}$ and $\boldsymbol{x}_{b',i}$ are independent, and hence cross-terms can be eliminated. In (b) we defined $\boldsymbol{s}_{b,i}^{\mathrm{ord}}(\boldsymbol{w}_{i-1})$ as the gradient noise arising from the *ordinary* stochastic gradient approximation $\widehat{\nabla J}^{\mathrm{ord}}(\boldsymbol{w}_{i-1};\boldsymbol{x}_{i,b}) \triangleq \nabla Q(\boldsymbol{w}_{i-1};\boldsymbol{x}_{b,i})$. In other words:

$$\begin{aligned}
\boldsymbol{s}_{b,i}^{\mathrm{ord}}(\boldsymbol{w}_{i-1}) &\triangleq \widehat{\nabla J}^{\mathrm{ord}}(\boldsymbol{w}_{i-1};\boldsymbol{x}_{i,b})-\nabla J(\boldsymbol{w}_{i-1})\\
&=\nabla Q(\boldsymbol{w}_{i-1},\boldsymbol{x}_{i,b})-\nabla J(\boldsymbol{w}_{i-1})
\end{aligned}\tag{3.44}$$

Suppose the ordinary gradient stochastic gradient approximation $\widehat{\nabla J}^{\mathrm{ord}}(\boldsymbol{w}_{i-1};\boldsymbol{x}_{i,b})$ induces a gradient noise process satisfying the variance conditions (3.22)–(3.23) with constants $\alpha_{\mathrm{ord}}^2,\beta_{\mathrm{ord}}^2,\gamma_{\mathrm{ord}}^2,\sigma_{\mathrm{ord}}^2$. Then:

$$\begin{aligned}
&\mathbb{E}\left\{\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2|\boldsymbol{w}_{i-1}\right\}\\
&=\frac{1}{B^2}\sum_{b=1}^{B}\mathbb{E}\left\{\left\|\boldsymbol{s}_{b,i}^{\mathrm{ord}}(\boldsymbol{w}_{i-1})\right\|^2|\boldsymbol{w}_{i-1}\right\}\\
&\le\frac{1}{B^2}\sum_{b=1}^{B}\left(\alpha_{\mathrm{ord}}^2\|\nabla J(\boldsymbol{w}_{i-1})\|^2+\beta_{\mathrm{ord}}^2\|w^o-\boldsymbol{w}_{i-1}\|^2+\gamma_{\mathrm{ord}}^2\left(J(\boldsymbol{w}_{i-1})-J^o\right)+\sigma_{\mathrm{ord}}^2\right)\\
&=\frac{\alpha_{\mathrm{ord}}^2}{B}\|\nabla J(\boldsymbol{w}_{i-1})\|^2+\frac{\beta_{\mathrm{ord}}^2}{B}\|w^o-\boldsymbol{w}_{i-1}\|^2+\frac{\gamma_{\mathrm{ord}}^2}{B}\left(J(\boldsymbol{w}_{i-1})-J^o\right)+\frac{\sigma_{\mathrm{ord}}^2}{B}\\
&=\alpha_B^2\|\nabla J(\boldsymbol{w}_{i-1})\|^2+\beta_B^2\|w^o-\boldsymbol{w}_{i-1}\|^2+\gamma_B^2\left(J(\boldsymbol{w}_{i-1})-J^o\right)+\sigma_B^2
\end{aligned}\tag{3.45}$$

where in the last line we defined:

$$\alpha_B^2\triangleq\frac{\alpha_{\mathrm{ord}}^2}{B};\ \beta_B^2\triangleq\frac{\beta_{\mathrm{ord}}^2}{B};\ \gamma_B^2\triangleq\frac{\gamma_{\mathrm{ord}}^2}{B};\ \sigma_B^2\triangleq\frac{\sigma_{\mathrm{ord}}^2}{B}\tag{3.46}$$

We conclude that using B independent samples in constructing the gradient approximation (3.41) results in a B-fold reduction in the variance. This indicates that the gradient approximation based on a mini-batch is of higher quality, and we would expect the resulting stochastic gradient algorithm to exhibit improved performance. Of course, computing $\nabla Q(\boldsymbol{w}_{i-1}, \boldsymbol{x}_{i,b})$ multiple times results in a B-fold increase in computational complexity. In order to make on informed decision on the optimal mini-batch size, we will hence need to more carefully link the variance of the gradient approximation with the resulting learning performance.

Asynchronous Stochastic Gradient Descent

In the previous section we saw how the availability of multiple data samples can be leveraged to construct an improved gradient approximation. At the other end of the spectrum, we may be in a situation where data is more scarce, and we are forced to construct lower-quality gradient approximations. To illustrate this case, consider a setting where at each iteration i, we are only provided with a sample $\boldsymbol{x}_i$ with probability π. With probablility $1-\pi$ we are provided no data, and are hence unable to compute a stochastic gradient estimate. We define an i.i.d. Bernoulli random indicator variable:

$$\mathbb{1}_i = \begin{cases} 1 & \text{with probability } \pi, \\ 0 & \text{otherwise.} \end{cases} \tag{3.47}$$

where $\mathbb{1}_i$ indicates whether we receive a sample $\boldsymbol{x}_i$ at time i. We can then then construct:

$$\widehat{\nabla J}(w) \triangleq \begin{cases} \frac{1}{\pi}\nabla Q(w;\boldsymbol{x}_i) & \text{if } \mathbb{1}_i = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{3.48}$$

Note that in the case when data is available, the gradient approximation $\frac{1}{p}\nabla Q(w;\boldsymbol{x}_i)$ is scaled by a factor of $\frac{1}{\pi}$. Since $0 < \pi \leq 1$, this results in larger steps being taken when data is available to compensate for the fact that with probability $1-\pi$, no update occurs. This scaling ensures that we continue to have an unbiased approximation:

$$\begin{aligned} &\mathbb{E}\,\widehat{\nabla J}(w) \\ &= \mathbb{E}\left\{\widehat{\nabla J}(w)|\mathbb{1}_i = 1\right\} \cdot \text{Prob}\left\{\mathbb{1}_i = 1\right\} + \left\{\widehat{\nabla J}(w)|\mathbb{1}_i = 0\right\} \cdot \text{Prob}\left\{\mathbb{1}_i = 0\right\} \\ &= \mathbb{E}\left\{\frac{1}{\pi}\nabla Q(w;\boldsymbol{x}_i)|\mathbb{1}_i = 1\right\} \cdot \pi + 0 \cdot (1-\pi) \\ &= \mathbb{E}\left\{\nabla Q(w;\boldsymbol{x}_i)\right\} = \nabla J(w) \end{aligned} \tag{3.49}$$

To derive an expression for the gradient noise variance, we proceed:

$$
\begin{aligned}
&\mathbb{E}\left\{\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2|\boldsymbol{w}_{i-1}\right\} \\
&=\mathbb{E}\left\{\left\|\widehat{\nabla J}\left(\boldsymbol{w}_{i-1}\right)-\nabla J(\boldsymbol{w}_{i-1})\right\|^2|\boldsymbol{w}_{i-1}\right\} \\
&=\mathbb{E}\left\{\left\|\widehat{\nabla J}\left(\boldsymbol{w}_{i-1}\right)-\nabla J(\boldsymbol{w}_{i-1})\right\|^2|\boldsymbol{w}_{i-1},\mathbb{1}_i=1\right\}\cdot\Pr\{\mathbb{1}_i=1\} \\
&\quad+\mathbb{E}\left\{\left\|\widehat{\nabla J}\left(\boldsymbol{w}_{i-1}\right)-\nabla J(\boldsymbol{w}_{i-1})\right\|^2|\boldsymbol{w}_{i-1},\mathbb{1}_i=0\right\}\cdot\Pr\{\mathbb{1}_i=0\} \\
&\overset{(3.48)}{=}\mathbb{E}\left\{\left\|\frac{1}{\pi}\nabla Q(\boldsymbol{w}_{i-1};\boldsymbol{x}_i)-\nabla J(\boldsymbol{w}_{i-1})\right\|^2|\boldsymbol{w}_{i-1}\right\}\cdot\pi+\|\nabla J(\boldsymbol{w}_{i-1})\|^2\cdot(1-\pi) \\
&=\mathbb{E}\left\{\left\|\frac{1}{\pi}\nabla Q(\boldsymbol{w}_{i-1};\boldsymbol{x}_i)-\frac{1}{\pi}\nabla J(\boldsymbol{w}_{i-1})-\left(1-\frac{1}{\pi}\right)\nabla J(\boldsymbol{w}_{i-1})\right\|^2|\boldsymbol{w}_{i-1}\right\}\cdot\pi \\
&\quad+\|\nabla J(\boldsymbol{w}_{i-1})\|^2\cdot(1-\pi) \\
&=\mathbb{E}\left\{\left\|\frac{1}{\pi}\nabla Q(\boldsymbol{w}_{i-1};\boldsymbol{x}_i)-\frac{1}{\pi}\nabla J(\boldsymbol{w}_{i-1})\right\|^2|\boldsymbol{w}_{i-1}\right\}\cdot\pi \\
&\quad+\left(1-\frac{1}{\pi}\right)^2\|\nabla J(\boldsymbol{w}_{i-1})\|^2\cdot\pi+\|\nabla J(\boldsymbol{w}_{i-1})\|^2\cdot(1-\pi) \\
&=\frac{1}{\pi}\mathbb{E}\left\{\left\|\boldsymbol{s}_i^{\mathrm{ord}}(\boldsymbol{w}_i)\right\|^2|\boldsymbol{w}_{i-1}\right\}+\left(\left(1-\frac{1}{\pi}\right)^2\cdot\pi+1-\pi\right)\|\nabla J(\boldsymbol{w}_{i-1})\|^2 \\
&=\frac{1}{\pi}\mathbb{E}\left\{\left\|\boldsymbol{s}_i^{\mathrm{ord}}(\boldsymbol{w}_i)\right\|^2|\boldsymbol{w}_{i-1}\right\}+\frac{1-\pi}{\pi}\|\nabla J(\boldsymbol{w}_{i-1})\|^2 \\
&=\frac{\alpha_{\mathrm{ord}}^2+1-\pi}{\pi}\|\nabla J(\boldsymbol{w}_{i-1})\|^2+\frac{\beta_{\mathrm{ord}}^2}{\pi}\|w^o-\boldsymbol{w}_{i-1}\|^2+\frac{\gamma_{\mathrm{ord}}^2}{\pi}\left(J(\boldsymbol{w}_{i-1})-J^o\right)+\frac{\sigma_{\mathrm{ord}}^2}{\pi} \\
&=\alpha_{\mathrm{asy}}^2\|\nabla J(\boldsymbol{w}_{i-1})\|^2+\beta_{\mathrm{asy}}^2\|w^o-\boldsymbol{w}_{i-1}\|^2+\gamma_{\mathrm{asy}}^2\left(J(\boldsymbol{w}_{i-1})-J^o\right)+\sigma_{\mathrm{asy}}^2
\end{aligned}
\tag{3.50}
$$

We introduced:

$$
\alpha_{\mathrm{asy}}^2\triangleq\frac{\alpha_{\mathrm{ord}}^2+1-\pi}{\pi};\ \beta_{\mathrm{asy}}^2\triangleq\frac{\beta_{\mathrm{ord}}^2}{\pi};\ \gamma_{\mathrm{asy}}^2\triangleq\frac{\gamma_{\mathrm{ord}}^2}{\pi};\ \sigma_{\mathrm{asy}}^2\triangleq\frac{\sigma_{\mathrm{ord}}^2}{\pi} \tag{3.51}
$$

3.2 PERFORMANCE

As we have seen in our discussion so far, there are a plethora of potential constructions for gradient approximations, all of which are of varying quality, and hence introduce varying levels of gradient noise, as captured in the gradient noise constants $\alpha^2, \beta^2, \gamma^2, \sigma^2$. We will now present convergence analysis for the stochastic gradient recursion (3.19).

THEOREM 3.1 (Mean-square-behavior of stochastic gradient algorithms). *Let*

the objective function $J(w)$ be ν-strongly convex with δ-Lipschitz gradients (see Section 2.1). Suppose further that we employ a gradient approximation $\widehat{\nabla J}(\cdot)$ satisfying the conditions (3.22) and (3.23) with constants $\alpha^2, \beta^2, \gamma^2, \sigma^2$. Then, the error $\widetilde{\boldsymbol{w}}_i \triangleq w^o - \boldsymbol{w}_i$ of the iterates generated by the stochastic gradient descent algorithm (3.19) satisfies:

$$\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2 \leq \lambda \mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2 + \mu^2\sigma^2 \tag{3.52}$$

Then, for sufficiently small step-sizes

$$\mu \leq \frac{\nu}{(1+\alpha^2)\delta^2 + \beta^2 + \gamma^2\frac{\delta}{2}} \tag{3.53}$$

it holds that $\lambda < 1$ and we can iterate this relation to find:

$$\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \leq \gamma^i \|\widetilde{w}_0\|^2 + \frac{\mu\sigma^2}{\nu} \tag{3.54}$$

For the excess risk, we have:

$$\mathbb{E}\,J(\boldsymbol{w}_i) - J(w^o) \leq \gamma^i \left(\frac{\delta\|\widetilde{w}_0\|^2}{2}\right) + \frac{\mu\delta\sigma^2}{2\nu} \tag{3.55}$$

Proof: We subtract recursion (3.19) from w^o and square both sides to obtain for the error $\widetilde{\boldsymbol{w}}_i \triangleq w^o - \boldsymbol{w}_i$:

$$\|\widetilde{\boldsymbol{w}}_i\|^2 = \left\|\widetilde{\boldsymbol{w}}_{i-1} + \mu\widehat{\nabla J}(\boldsymbol{w}_{i-1})\right\|^2 \tag{3.56}$$

Next, we take expectations conditioned on $\boldsymbol{w}_{i-1}$ to obtain:

$$\begin{aligned}
&\mathbb{E}\left\{\|\widetilde{\boldsymbol{w}}_i\|^2 \big| \boldsymbol{w}_{i-1}\right\} \\
&= \mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_{i-1} + \mu\widehat{\nabla J}(\boldsymbol{w}_{i-1})\right\|^2 \Big| \boldsymbol{w}_{i-1}\right\} \\
&= \|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + \mu^2\mathbb{E}\left\{\left\|\widehat{\nabla J}(\boldsymbol{w}_{i-1})\right\|^2 \Big| \boldsymbol{w}_{i-1}\right\} + 2\mu\widetilde{\boldsymbol{w}}_{i-1}^{\mathsf{T}}\mathbb{E}\left\{\widehat{\nabla J}(\boldsymbol{w}_{i-1})\big|\boldsymbol{w}_{i-1}\right\}
\end{aligned} \tag{3.57}$$

We first simplify the terms involving expectations. First, for the linear term, we have:

$$\begin{aligned}
&2\mu\widetilde{\boldsymbol{w}}_{i-1}^{\mathsf{T}}\mathbb{E}\left\{\widehat{\nabla J}(\boldsymbol{w}_{i-1})\big|\boldsymbol{w}_{i-1}\right\} \\
&\overset{(3.22)}{=} 2\mu\widetilde{\boldsymbol{w}}_{i-1}^{\mathsf{T}}\nabla J(\boldsymbol{w}_{i-1}) \\
&\leq -2\mu\nu\|\widetilde{\boldsymbol{w}}_{i-1}\|^2
\end{aligned} \tag{3.58}$$

To verify the last step, we appeal to the strong-convexity property (2.12) and use $w_2 = w^\star$, $w_1 = \boldsymbol{w}_{i-1}$ in step (a) below and $w_2 = \boldsymbol{w}_{i-1}$, $w_1 = w^\star$ in step (b) to find that

$$\begin{aligned}
\widetilde{\boldsymbol{w}}_{i-1}^{\mathsf{T}}\nabla J(\boldsymbol{w}_{i-1}) &\overset{(a)}{\leq} J(w^\star) - J(\boldsymbol{w}_{i-1}) - \frac{\nu}{2}\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 \\
&\overset{(b)}{\leq} -\frac{\nu}{2}\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 - \frac{\nu}{2}\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 \\
&= -\nu\|\widetilde{\boldsymbol{w}}_{i-1}\|^2
\end{aligned} \tag{3.59}$$

For the quadratic term we have:

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\widehat{\nabla J}(\boldsymbol{w}_{i-1})\right\|^2\Big|\boldsymbol{w}_{i-1}\right\} \\
&= \mathbb{E}\left\{\|\nabla J(\boldsymbol{w}_{i-1}) + \boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2\big|\boldsymbol{w}_{i-1}\right\} \\
&= \|\nabla J(\boldsymbol{w}_{i-1})\|^2 + \mathbb{E}\left\{\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2\big|\boldsymbol{w}_{i-1}\right\} + 2\nabla J(\boldsymbol{w}_{i-1})^{\mathsf{T}}\mathbb{E}\left\{\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\big|\boldsymbol{w}_{i-1}\right\} \\
&\overset{(3.22)}{=} \|\nabla J(\boldsymbol{w}_{i-1})\|^2 + \mathbb{E}\left\{\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2\big|\boldsymbol{w}_{i-1}\right\} \\
&\overset{(3.23)}{\leq} \|\nabla J(\boldsymbol{w}_{i-1})\|^2 + \alpha^2\|\nabla J(\boldsymbol{w}_{i-1})\|^2 + \beta^2\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + \gamma^2\left(J(\boldsymbol{w}_{i-1} - J(w^o))\right) + \sigma^2 \\
&= (1+\alpha^2)\|\nabla J(\boldsymbol{w}_{i-1})\|^2 + \beta^2\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + \gamma^2\left(J(\boldsymbol{w}_{i-1} - J(w^o))\right) + \sigma^2 \\
&\overset{(a)}{=} (1+\alpha^2)\delta^2\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + \beta^2\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + \gamma^2\frac{\delta}{2}\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + \sigma^2 \\
&= \left((1+\alpha^2)\delta^2 + \beta^2 + \gamma^2\frac{\delta}{2}\right)\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + \sigma^2
\end{aligned} \tag{3.60}$$

where in (a) we made use of the smoothness conditions (2.15) and (2.16) with $w_2 = \boldsymbol{w}_{i-1}$ and $w_1 = w^o$. We then obtain for (3.57):

$$\begin{aligned}
&\mathbb{E}\left\{\|\widetilde{\boldsymbol{w}}_i\|^2\big|\boldsymbol{w}_{i-1}\right\} \\
&\leq \|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + \mu^2\left((1+\alpha^2)\delta^2 + \beta^2 + \gamma^2\frac{\delta}{2}\right)\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + \mu^2\sigma^2 - 2\mu\nu\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 \\
&= \left(1 - 2\mu\nu + \mu^2\left((1+\alpha^2)\delta^2 + \beta^2 + \gamma^2\frac{\delta}{2}\right)\right)\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + \mu^2\sigma^2 \\
&= \lambda\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + \mu^2\sigma^2
\end{aligned} \tag{3.61}$$

where we defined:

$$\lambda \triangleq 1 - 2\mu\nu + \mu^2\left((1+\alpha^2)\delta^2 + \beta^2 + \gamma^2\frac{\delta}{2}\right) \tag{3.62}$$

Upon taking expectations to remove the conditioning on the lefthand-side, we have:

$$\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \leq \lambda\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + \mu^2\sigma^2 \tag{3.63}$$

which contracts whenever $\lambda < 1$. This is equivalent to:

$$1 - 2\mu\nu + \mu^2\left((1+\alpha^2)\delta^2 + \beta^2 + \gamma^2\frac{\delta}{2}\right) < 1 \Longleftrightarrow \mu < \frac{2\nu}{(1+\alpha^2)\delta^2 + \beta^2 + \gamma^2\frac{\delta}{2}} \tag{3.64}$$

Iterating, we find:

$$\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \leq \gamma^i\|\widetilde{w}_0\|^2 + \frac{\mu\sigma^2}{2\nu - \mu\left((1+\alpha^2)\delta^2 + \beta^2 + \gamma^2\frac{\delta}{2}\right)} \tag{3.65}$$

If we assume that $\mu \leq \frac{\nu}{(1+\alpha^2)\delta^2+\beta^2+\gamma^2\frac{\delta}{2}}$, we can simplify the constant term to:

$$\frac{\mu\sigma^2}{2\nu - \mu\left((1+\alpha^2)\delta^2 + \beta^2 + \gamma^2\frac{\delta}{2}\right)} \leq \frac{\mu\sigma^2}{2\nu - \nu} = \frac{\mu\sigma^2}{\nu} \tag{3.66}$$

and obtain:

$$\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \leq \gamma^i\|\widetilde{w}_0\|^2 + \frac{\mu\sigma^2}{\nu} \tag{3.67}$$

For the excess risk, we have:

$$\mathbb{E}\, J(\boldsymbol{w}_i) - J(w^o) \overset{(a)}{\leq} \frac{\delta}{2}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \leq \gamma^i \left(\frac{\delta\|\widetilde{w}_0\|^2}{2}\right) + \frac{\mu\delta\sigma^2}{2\nu} \tag{3.68}$$

where (a) follows again from the smoothness condition (2.15).

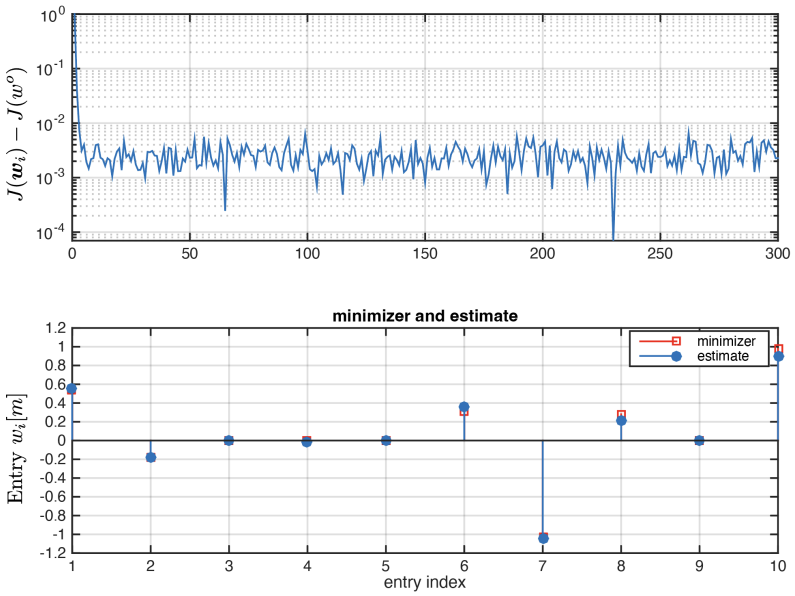

Figure 3.2 Evolution of the excess risk $J(\boldsymbol{w}_i) - J(w^o)$, averaged over several experiments. The top panel shows how the excess risk decays very quickly until it reaches a steady-state regime, as predicted by Theorem 3.1. The bottom panel shows a small error remaining even after a large number of iterations.

3.3 PROBLEMS

3.1 We are interested in solving the logistic regression problem (3.38) by employing the ordinary gradient approximation $\widehat{\nabla J}(w) \triangleq \nabla Q(w; \boldsymbol{h}_i, \boldsymbol{\gamma}_i)$, giving rise to the recursion:

$$\boldsymbol{w}_i = (1 - \mu\rho)\boldsymbol{w}_{i-1} + \mu\frac{\boldsymbol{\gamma}_i \boldsymbol{h}_i}{1 + e^{\boldsymbol{\gamma}_i \boldsymbol{h}_i^{\mathsf{T}} \boldsymbol{w}_{i-1}}} \tag{3.69}$$

Determine the strong convexity parameter ν of the logistic cost (3.38). Use it along with the gradient noise parameters derived in Example 3.4 and Theorem 3.1 to find a bound for the steady-state performance $\lim_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2$. How do you need to choose the step-size in terms of the other parameters to ensure that

$$\lim_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2 \leq \epsilon? \tag{3.70}$$

How will a very small ϵ affect the rate of convergence λ?

3.2 We consider again the logistic regression problem (3.38), but now employ the mini-batch stochastic gradient approximation (3.41), giving rise to:

$$\boldsymbol{w}_i = (1-\mu\rho)\boldsymbol{w}_{i-1} + \frac{\mu}{B}\sum_{b=1}^{B}\frac{\boldsymbol{\gamma}_{b,i}\boldsymbol{h}_{b,i}}{1+e^{\boldsymbol{\gamma}_{b,i}\boldsymbol{h}_{b,i}^{\mathsf{T}}\boldsymbol{w}_{i-1}}} \tag{3.71}$$

Determine an expression for the gradient noise parameters of this improved approximation, and use them with Theorem 3.1 to find a bound for the steady-state performance $\lim_{i\to\infty}\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2$. Keeping the step-size constant, how do you need to choose the mini-batch size B as a function of the remaining parameters to ensure:

$$\lim_{i\to\infty}\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2 \leq \epsilon? \tag{3.72}$$

How will a very small ϵ affect the rate of convergence λ?

3.3 Consider a gradient approximation similar to the mini-batch construction (3.41), where the number of available samples at time i is now a random variable $\boldsymbol{B}_i$, distributed uniformly between 1 and B. We can then construct:

$$\widehat{\nabla J}(w) \triangleq \frac{1}{\boldsymbol{B}_i}\sum_{b=1}^{\boldsymbol{B}_i}\nabla Q(w;\boldsymbol{x}_{b,i}) \tag{3.73}$$

Verify that this construction satisfies the unbiasedness condition (3.22) and determine an expression for the gradient noise constants of (3.23) in terms of those of the ordinary gradient approximation.