14

Stochastic Gradient Descent

Recall that the goal of learning is to minimize the risk function, $L_{\mathcal{D}}(h) = \mathbb{E}_{z\sim\mathcal{D}}[\ell(h,z)]$. We cannot directly minimize the risk function since it depends on the unknown distribution $\mathcal{D}$. So far in the book, we have discussed learning methods that depend on the empirical risk. That is, we first sample a training set S and define the empirical risk function $L_S(h)$. Then, the learner picks a hypothesis based on the value of $L_S(h)$. For example, the ERM rule tells us to pick the hypothesis that minimizes $L_S(h)$ over the hypothesis class, $\mathcal{H}$. Or, in the previous chapter, we discussed regularized risk minimization, in which we pick a hypothesis that jointly minimizes $L_S(h)$ and a regularization function over h.

In this chapter we describe and analyze a rather different learning approach, which is called *Stochastic Gradient Descent* (SGD). As in Chapter 12 we will focus on the important family of convex learning problems, and following the notation in that chapter, we will refer to hypotheses as vectors $\mathbf{w}$ that come from a convex hypothesis class, $\mathcal{H}$. In SGD, we try to minimize the risk function $L_{\mathcal{D}}(\mathbf{w})$ directly using a gradient descent procedure. Gradient descent is an iterative optimization procedure in which at each step we improve the solution by taking a step along the negative of the gradient of the function to be minimized at the current point. Of course, in our case, we are minimizing the risk function, and since we do not know $\mathcal{D}$ we also do not know the gradient of $L_{\mathcal{D}}(\mathbf{w})$. SGD circumvents this problem by allowing the optimization procedure to take a step along a random direction, as long as the expected value of the direction is the negative of the gradient. And, as we shall see, finding a random direction whose expected value corresponds to the gradient is rather simple even though we do not know the underlying distribution $\mathcal{D}$.

The advantage of SGD, in the context of convex learning problems, over the regularized risk minimization learning rule is that SGD is an efficient algorithm that can be implemented in a few lines of code, yet still enjoys the same sample complexity as the regularized risk minimization rule. The simplicity of SGD also allows us to use it in situations when it is not possible to apply methods that are based on the empirical risk, but this is beyond the scope of this book.

We start this chapter with the basic gradient descent algorithm and analyze its convergence rate for convex-Lipschitz functions. Next, we introduce the notion of

https://doi.org/10.1017/CBO9781107298019.015 Published online by Cambridge University Press

subgradient and show that gradient descent can be applied for nondifferentiable functions as well. The core of this chapter is Section 14.3, in which we describe the Stochastic Gradient Descent algorithm, along with several useful variants. We show that SGD enjoys an expected convergence rate similar to the rate of gradient descent. Finally, we turn to the applicability of SGD to learning problems.

14.1 GRADIENT DESCENT

Before we describe the stochastic gradient descent method, we would like to describe the standard gradient descent approach for minimizing a differentiable convex function $f(\mathbf{w})$.

The gradient of a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ at $\mathbf{w}$, denoted $\nabla f(\mathbf{w})$, is the vector of partial derivatives of f, namely, $\nabla f(\mathbf{w}) = \left(\frac{\partial f(\mathbf{w})}{\partial w[1]}, \ldots, \frac{\partial f(\mathbf{w})}{\partial w[d]} \right)$. Gradient descent is an iterative algorithm. We start with an initial value of $\mathbf{w}$ (say, $\mathbf{w}^{(1)} = \mathbf{0}$). Then, at each iteration, we take a step in the direction of the negative of the gradient at the current point. That is, the update step is

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \nabla f(\mathbf{w}^{(t)}), \tag{14.1}$$

where $\eta > 0$ is a parameter to be discussed later. Intuitively, since the gradient points in the direction of the greatest rate of increase of f around $\mathbf{w}^{(t)}$, the algorithm makes a small step in the opposite direction, thus decreasing the value of the function. Eventually, after T iterations, the algorithm outputs the averaged vector, $\bar{\mathbf{w}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{w}^{(t)}$. The output could also be the last vector, $\mathbf{w}^{(T)}$, or the best performing vector, $\operatorname{argmin}_{t \in [T]} f(\mathbf{w}^{(t)})$, but taking the average turns out to be rather useful, especially when we generalize gradient descent to nondifferentiable functions and to the stochastic case.

Another way to motivate gradient descent is by relying on Taylor approximation. The gradient of f at $\mathbf{w}$ yields the first order Taylor approximation of f around $\mathbf{w}$ by $f(\mathbf{u}) \approx f(\mathbf{w}) + \langle \mathbf{u} - \mathbf{w}, \nabla f(\mathbf{w}) \rangle$. When f is convex, this approximation lower bounds f, that is,

$$f(\mathbf{u}) \geq f(\mathbf{w}) + \langle \mathbf{u} - \mathbf{w}, \nabla f(\mathbf{w}) \rangle.$$

Therefore, for $\mathbf{w}$ close to $\mathbf{w}^{(t)}$ we have that $f(\mathbf{w}) \approx f(\mathbf{w}^{(t)}) + \langle \mathbf{w} - \mathbf{w}^{(t)}, \nabla f(\mathbf{w}^{(t)}) \rangle$. Hence we can minimize the approximation of $f(\mathbf{w})$. However, the approximation might become loose for $\mathbf{w}$, which is far away from $\mathbf{w}^{(t)}$. Therefore, we would like to minimize jointly the distance between $\mathbf{w}$ and $\mathbf{w}^{(t)}$ and the approximation of f around $\mathbf{w}^{(t)}$. If the parameter η controls the tradeoff between the two terms, we obtain the update rule

$$\mathbf{w}^{(t+1)} = \operatorname*{argmin}_{\mathbf{w}} \frac{1}{2} \|\mathbf{w} - \mathbf{w}^{(t)}\|^2 + \eta \left(f(\mathbf{w}^{(t)}) + \langle \mathbf{w} - \mathbf{w}^{(t)}, \nabla f(\mathbf{w}^{(t)}) \rangle \right).$$

Solving the preceding by taking the derivative with respect to $\mathbf{w}$ and comparing it to zero yields the same update rule as in Equation (14.1).

https://doi.org/10.1017/CBO9781107298019.015 Published online by Cambridge University Press

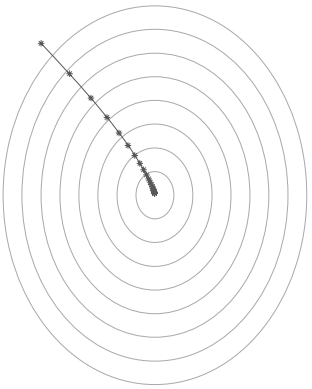

Figure 14.1. An illustration of the gradient descent algorithm. The function to be minimized is $1.25(x_1+6)^2+(x_2-8)^2$.

14.1.1 Analysis of GD for Convex-Lipschitz Functions

To analyze the convergence rate of the GD algorithm, we limit ourselves to the case of convex-Lipschitz functions (as we have seen, many problems lend themselves easily to this setting). Let $\mathbf{w}^\star$ be any vector and let B be an upper bound on $\|\mathbf{w}^\star\|$. It is convenient to think of $\mathbf{w}^\star$ as the minimizer of $f(\mathbf{w})$, but the analysis that follows holds for every $\mathbf{w}^\star$.

We would like to obtain an upper bound on the suboptimality of our solution with respect to $\mathbf{w}^\star$, namely, $f(\bar{\mathbf{w}})-f(\mathbf{w}^\star)$, where $\bar{\mathbf{w}}=\frac{1}{T}\sum_{t=1}^{T}\mathbf{w}^{(t)}$. From the definition of $\bar{\mathbf{w}}$, and using Jensen's inequality, we have that

$$\begin{aligned} f(\bar{\mathbf{w}})-f(\mathbf{w}^\star) &= f\left(\tfrac{1}{T}\sum_{t=1}^{T}\mathbf{w}^{(t)}\right)-f(\mathbf{w}^\star) \\ &\leq \frac{1}{T}\sum_{t=1}^{T}\left(f(\mathbf{w}^{(t)})\right)-f(\mathbf{w}^\star) \\ &= \frac{1}{T}\sum_{t=1}^{T}\left(f(\mathbf{w}^{(t)})-f(\mathbf{w}^\star)\right). \end{aligned} \tag{14.2}$$

For every t, because of the convexity of f, we have that

$$f(\mathbf{w}^{(t)})-f(\mathbf{w}^\star) \leq \langle \mathbf{w}^{(t)}-\mathbf{w}^\star, \nabla f(\mathbf{w}^{(t)})\rangle. \tag{14.3}$$

Combining the preceding we obtain

$$f(\bar{\mathbf{w}})-f(\mathbf{w}^\star) \leq \frac{1}{T}\sum_{t=1}^{T}\langle \mathbf{w}^{(t)}-\mathbf{w}^\star, \nabla f(\mathbf{w}^{(t)})\rangle.$$

To bound the right-hand side we rely on the following lemma:

Lemma 14.1. *Let $\mathbf{v}_1,\ldots,\mathbf{v}_T$ be an arbitrary sequence of vectors. Any algorithm with an initialization $\mathbf{w}^{(1)}=0$ and an update rule of the form*

$$\mathbf{w}^{(t+1)}=\mathbf{w}^{(t)}-\eta\mathbf{v}_t \tag{14.4}$$

https://doi.org/10.1017/CBO9781107298019.015 Published online by Cambridge University Press

satisfies

$$\sum_{t=1}^{T} \langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t \rangle \leq \frac{\|\mathbf{w}^\star\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\mathbf{v}_t\|^2. \tag{14.5}$$

In particular, for every $B, \rho > 0$, if for all t we have that $\|\mathbf{v}_t\| \leq \rho$ and if we set $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$, then for every $\mathbf{w}^\star$ with $\|\mathbf{w}^\star\| \leq B$ we have

$$\frac{1}{T} \sum_{t=1}^{T} \langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t \rangle \leq \frac{B\rho}{\sqrt{T}}.$$

Proof. Using algebraic manipulations (completing the square), we obtain:

$$\begin{aligned}
\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t \rangle &= \frac{1}{\eta} \langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \eta \mathbf{v}_t \rangle \\
&= \frac{1}{2\eta} \left(- \|\mathbf{w}^{(t)} - \mathbf{w}^\star - \eta \mathbf{v}_t\|^2 + \|\mathbf{w}^{(t)} - \mathbf{w}^\star\|^2 + \eta^2 \|\mathbf{v}_t\|^2 \right) \\
&= \frac{1}{2\eta} \left(- \|\mathbf{w}^{(t+1)} - \mathbf{w}^\star\|^2 + \|\mathbf{w}^{(t)} - \mathbf{w}^\star\|^2 \right) + \frac{\eta}{2} \|\mathbf{v}_t\|^2,
\end{aligned}$$

where the last equality follows from the definition of the update rule. Summing the equality over t, we have

$$\sum_{t=1}^{T} \langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t \rangle = \frac{1}{2\eta} \sum_{t=1}^{T} \left(- \|\mathbf{w}^{(t+1)} - \mathbf{w}^\star\|^2 + \|\mathbf{w}^{(t)} - \mathbf{w}^\star\|^2 \right) + \frac{\eta}{2} \sum_{t=1}^{T} \|\mathbf{v}_t\|^2. \tag{14.6}$$

The first sum on the right-hand side is a telescopic sum that collapses to

$$\|\mathbf{w}^{(1)} - \mathbf{w}^\star\|^2 - \|\mathbf{w}^{(T+1)} - \mathbf{w}^\star\|^2.$$

Plugging this in Equation (14.6), we have

$$\begin{aligned}
\sum_{t=1}^{T} \langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t \rangle &= \frac{1}{2\eta} (\|\mathbf{w}^{(1)} - \mathbf{w}^\star\|^2 - \|\mathbf{w}^{(T+1)} - \mathbf{w}^\star\|^2) + \frac{\eta}{2} \sum_{t=1}^{T} \|\mathbf{v}_t\|^2 \\
&\leq \frac{1}{2\eta} \|\mathbf{w}^{(1)} - \mathbf{w}^\star\|^2 + \frac{\eta}{2} \sum_{t=1}^{T} \|\mathbf{v}_t\|^2 \\
&= \frac{1}{2\eta} \|\mathbf{w}^\star\|^2 + \frac{\eta}{2} \sum_{t=1}^{T} \|\mathbf{v}_t\|^2,
\end{aligned}$$

where the last equality is due to the definition $\mathbf{w}^{(1)} = 0$. This proves the first part of the lemma (Equation (14.5)). The second part follows by upper bounding $\|\mathbf{w}^\star\|$ by B, $\|\mathbf{v}_t\|$ by ρ, dividing by T, and plugging in the value of η. □

Lemma 14.1 applies to the GD algorithm with $\mathbf{v}_t = \nabla f(\mathbf{w}^{(t)})$. As we will show later in Lemma 14.7, if f is ρ-Lipschitz, then $\|\nabla f(\mathbf{w}^{(t)})\| \leq \rho$. We therefore satisfy

https://doi.org/10.1017/CBO9781107298019.015 Published online by Cambridge University Press

the lemma's conditions and achieve the following corollary:

Corollary 14.2. *Let f be a convex, ρ-Lipschitz function, and let $\mathbf{w}^\star \in \operatorname{argmin}_{\{\mathbf{w}:\|\mathbf{w}\|\leq B\}} f(\mathbf{w})$. If we run the GD algorithm on f for T steps with $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$, then the output vector $\bar{\mathbf{w}}$ satisfies*

$$f(\bar{\mathbf{w}}) - f(\mathbf{w}^\star) \leq \frac{B\rho}{\sqrt{T}}.$$

Furthermore, for every $\epsilon > 0$, to achieve $f(\bar{\mathbf{w}}) - f(\mathbf{w}^\star) \leq \epsilon$, it suffices to run the GD algorithm for a number of iterations that satisfies

$$T \geq \frac{B^2\rho^2}{\epsilon^2}.$$

14.2 SUBGRADIENTS

The GD algorithm requires that the function f be differentiable. We now generalize the discussion beyond differentiable functions. We will show that the GD algorithm can be applied to nondifferentiable functions by using a so-called subgradient of $f(\mathbf{w})$ at $\mathbf{w}^{(t)}$, instead of the gradient.

To motivate the definition of subgradients, recall that for a convex function f, the gradient at $\mathbf{w}$ defines the slope of a tangent that lies below f, that is,

$$\forall \mathbf{u}, \quad f(\mathbf{u}) \geq f(\mathbf{w}) + \langle \mathbf{u} - \mathbf{w}, \nabla f(\mathbf{w}) \rangle. \tag{14.7}$$

An illustration is given on the left-hand side of Figure 14.2.

The existence of a tangent that lies below f is an important property of convex functions, which is in fact an alternative characterization of convexity.

Lemma 14.3. *Let S be an open convex set. A function $f : S \to \mathbb{R}$ is convex iff for every $\mathbf{w} \in S$ there exists $\mathbf{v}$ such that*

$$\forall \mathbf{u} \in S, \quad f(\mathbf{u}) \geq f(\mathbf{w}) + \langle \mathbf{u} - \mathbf{w}, \mathbf{v} \rangle. \tag{14.8}$$

The proof of this lemma can be found in many convex analysis textbooks (e.g., (Borwein & Lewis 2006)). The preceding inequality leads us to the definition of subgradients.

Definition 14.4. (Subgradients). A vector $\mathbf{v}$ that satisfies Equation (14.8) is called a *subgradient* of f at $\mathbf{w}$. The set of subgradients of f at $\mathbf{w}$ is called the *differential set* and denoted $\partial f(\mathbf{w})$.

An illustration of subgradients is given on the right-hand side of Figure 14.2. For scalar functions, a subgradient of a convex function f at w is a slope of a line that touches f at w and is not above f elsewhere.

14.2.1 Calculating Subgradients

How do we construct subgradients of a given convex function? If a function is differentiable at a point $\mathbf{w}$, then the differential set is trivial, as the following claim shows.

https://doi.org/10.1017/CBO9781107298019.015 Published online by Cambridge University Press

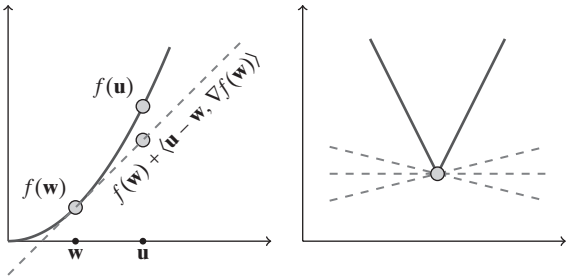

Figure 14.2. Left: The right-hand side of Equation (14.7) is the tangent of f at $\mathbf{w}$. For a convex function, the tangent lower bounds f. Right: Illustration of several subgradients of a nondifferentiable convex function.

Claim 14.5. *If f is differentiable at $\mathbf{w}$ then $\partial f(\mathbf{w})$ contains a single element – the gradient of f at $\mathbf{w}$, $\nabla f(\mathbf{w})$.*

Example 14.1 (The Differential Set of the Absolute Function). Consider the absolute value function $f(x) = |x|$. Using Claim 14.5, we can easily construct the differential set for the differentiable parts of f, and the only point that requires special attention is $x_0 = 0$. At that point, it is easy to verify that the subdifferential is the set of all numbers between -1 and 1. Hence:

$$\partial f(x) = \begin{cases} \{1\} & \text{if } x > 0 \\ \{-1\} & \text{if } x < 0 \\ [-1, 1] & \text{if } x = 0 \end{cases}$$

For many practical uses, we do not need to calculate the whole set of subgradients at a given point, as one member of this set would suffice. The following claim shows how to construct a sub-gradient for pointwise maximum functions.

Claim 14.6. *Let $g(\mathbf{w}) = \max_{i \in [r]} g_i(\mathbf{w})$ for r convex differentiable functions $g_1, \ldots, g_r$. Given some $\mathbf{w}$, let $j \in \operatorname{argmax}_i g_i(\mathbf{w})$. Then $\nabla g_j(\mathbf{w}) \in \partial g(\mathbf{w})$.*

Proof. Since g_j is convex we have that for all $\mathbf{u}$

$$g_j(\mathbf{u}) \geq g_j(\mathbf{w}) + \langle \mathbf{u} - \mathbf{w}, \nabla g_j(\mathbf{w}) \rangle.$$

Since $g(\mathbf{w}) = g_j(\mathbf{w})$ and $g(\mathbf{u}) \geq g_j(\mathbf{u})$ we obtain that

$$g(\mathbf{u}) \geq g(\mathbf{w}) + \langle \mathbf{u} - \mathbf{w}, \nabla g_j(\mathbf{w}) \rangle,$$

which concludes our proof. □

Example 14.2 (A Subgradient of the Hinge Loss). Recall the hinge loss function from Section 12.3, $f(\mathbf{w}) = \max\{0, 1 - y\langle \mathbf{w}, \mathbf{x} \rangle\}$ for some vector $\mathbf{x}$ and scalar y. To calculate a subgradient of the hinge loss at some $\mathbf{w}$ we rely on the preceding claim and obtain that the vector $\mathbf{v}$ defined in the following is a subgradient of the hinge loss at $\mathbf{w}$:

$$\mathbf{v} = \begin{cases} \mathbf{0} & \text{if } 1 - y\langle \mathbf{w}, \mathbf{x} \rangle \leq 0 \\ -y\mathbf{x} & \text{if } 1 - y\langle \mathbf{w}, \mathbf{x} \rangle > 0 \end{cases}$$

https://doi.org/10.1017/CBO9781107298019.015 Published online by Cambridge University Press

14.2.2 Subgradients of Lipschitz Functions

Recall that a function $f : A \to \mathbb{R}$ is ρ-Lipschitz if for all $\mathbf{u}, \mathbf{v} \in A$

$$|f(\mathbf{u}) - f(\mathbf{v})| \leq \rho \|\mathbf{u} - \mathbf{v}\|.$$

The following lemma gives an equivalent definition using norms of subgradients.

Lemma 14.7. *Let A be a convex open set and let $f : A \to \mathbb{R}$ be a convex function. Then, f is ρ-Lipschitz over A iff for all $\mathbf{w} \in A$ and $\mathbf{v} \in \partial f(\mathbf{w})$ we have that $\|\mathbf{v}\| \leq \rho$.*

Proof. Assume that for all $\mathbf{v} \in \partial f(\mathbf{w})$ we have that $\|\mathbf{v}\| \leq \rho$. Since $\mathbf{v} \in \partial f(\mathbf{w})$ we have

$$f(\mathbf{w}) - f(\mathbf{u}) \leq \langle \mathbf{v}, \mathbf{w} - \mathbf{u} \rangle.$$

Bounding the right-hand side using Cauchy-Schwartz inequality we obtain

$$f(\mathbf{w}) - f(\mathbf{u}) \ \leq \ \langle \mathbf{v}, \ \mathbf{w} - \mathbf{u} \rangle \leq \|\mathbf{v}\| \, \|\mathbf{w} - \mathbf{u}\| \leq \rho \, \|\mathbf{w} - \mathbf{u}\|.$$

An analogous argument can show that $f(\mathbf{u}) - f(\mathbf{w}) \leq \rho \|\mathbf{w} - \mathbf{u}\|$. Hence f is ρ-Lipschitz.

Now assume that f is ρ-Lipschitz. Choose some $\mathbf{w} \in A, \mathbf{v} \in \partial f(\mathbf{w})$. Since A is open, there exists $\epsilon > 0$ such that $\mathbf{u} = \mathbf{w} + \epsilon \mathbf{v}/\|\mathbf{v}\|$ belongs to A. Therefore, $\langle \mathbf{u} - \mathbf{w}, \ \mathbf{v} \rangle = \epsilon \|\mathbf{v}\|$ and $\|\mathbf{u} - \mathbf{w}\| = \epsilon$. From the definition of the subgradient,

$$f(\mathbf{u}) - f(\mathbf{w}) \geq \langle \mathbf{v}, \mathbf{u} - \mathbf{w} \rangle = \epsilon \|\mathbf{v}\|.$$

On the other hand, from the Lipschitzness of f we have

$$\rho \epsilon = \rho \, \|\mathbf{u} - \mathbf{w}\| \geq f(\mathbf{u}) - f(\mathbf{w}).$$

Combining the two inequalities we conclude that $\|\mathbf{v}\| \leq \rho$. □

14.2.3 Subgradient Descent

The gradient descent algorithm can be generalized to nondifferentiable functions by using a subgradient of $f(\mathbf{w})$ at $\mathbf{w}^{(t)}$, instead of the gradient. The analysis of the convergence rate remains unchanged: Simply note that Equation (14.3) is true for subgradients as well.

14.3 STOCHASTIC GRADIENT DESCENT (SGD)

In stochastic gradient descent we do not require the update direction to be based exactly on the gradient. Instead, we allow the direction to be a random vector and only require that its *expected value* at each iteration will equal the gradient direction. Or, more generally, we require that the expected value of the random vector will be a subgradient of the function at the current vector.

https://doi.org/10.1017/CBO9781107298019.015 Published online by Cambridge University Press

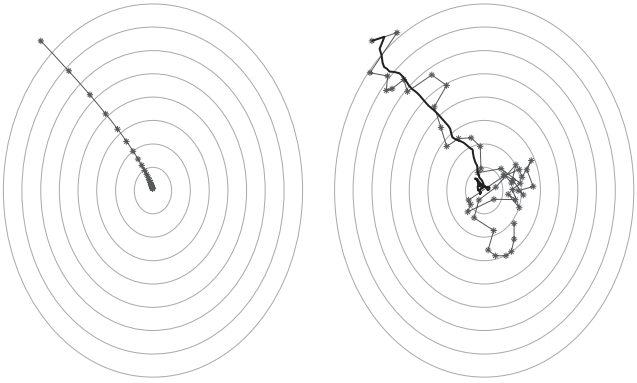

Figure 14.3. An illustration of the gradient descent algorithm (left) and the stochastic gradient descent algorithm (right). The function to be minimized is $1.25(x+6)^2+(y-8)^2$. For the stochastic case, the solid line depicts the averaged value of $\mathbf{w}$.

Stochastic Gradient Descent (SGD) for minimizing $f(\mathbf{w})$

parameters: Scalar $\eta > 0$, integer $T > 0$
initialize: $\mathbf{w}^{(1)} = \mathbf{0}$
for $t = 1, 2, \ldots, T$
  choose $\mathbf{v}_t$ at random from a distribution such that $\mathbb{E}[\mathbf{v}_t \mid \mathbf{w}^{(t)}] \in \partial f(\mathbf{w}^{(t)})$
  update $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \mathbf{v}_t$
output $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}^{(t)}$

An illustration of stochastic gradient descent versus gradient descent is given in Figure 14.3. As we will see in Section 14.5, in the context of learning problems, it is easy to find a random vector whose expectation is a subgradient of the risk function.

14.3.1 Analysis of SGD for Convex-Lipschitz-Bounded Functions

Recall the bound we achieved for the GD algorithm in Corollary 14.2. For the stochastic case, in which only the expectation of $\mathbf{v}_t$ is in $\partial f(\mathbf{w}^{(t)})$, we cannot directly apply Equation (14.3). However, since the expected value of $\mathbf{v}_t$ is a subgradient of f at $\mathbf{w}^{(t)}$, we can still derive a similar bound on the *expected* output of stochastic gradient descent. This is formalized in the following theorem.

Theorem 14.8. *Let $B, \rho > 0$. Let f be a convex function and let $\mathbf{w}^\star \in \operatorname{argmin}_{\mathbf{w}:\|\mathbf{w}\|\le B} f(\mathbf{w})$. Assume that SGD is run for T iterations with $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$. Assume also that for all t, $\|\mathbf{v}_t\| \le \rho$ with probability 1. Then,*

$$\mathbb{E}[f(\bar{\mathbf{w}})] - f(\mathbf{w}^\star) \le \frac{B\rho}{\sqrt{T}}.$$

https://doi.org/10.1017/CBO9781107298019.015 Published online by Cambridge University Press

Therefore, for any $\epsilon > 0$, to achieve $\mathbb{E}[f(\bar{\mathbf{w}})] - f(\mathbf{w}^\star) \leq \epsilon$, it suffices to run the SGD algorithm for a number of iterations that satisfies

$$T \geq \frac{B^2 \rho^2}{\epsilon^2}.$$

Proof. Let us introduce the notation $\mathbf{v}_{1:t}$ to denote the sequence $\mathbf{v}_1, \ldots, \mathbf{v}_t$. Taking expectation of Equation (14.2), we obtain

$$\mathop{\mathbb{E}}_{\mathbf{v}_{1:T}}[f(\bar{\mathbf{w}}) - f(\mathbf{w}^\star)] \leq \mathop{\mathbb{E}}_{\mathbf{v}_{1:T}}\left[\frac{1}{T}\sum_{t=1}^{T}(f(\mathbf{w}^{(t)}) - f(\mathbf{w}^\star))\right].$$

Since Lemma 14.1 holds for any sequence $\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_T$, it applies to SGD as well. By taking expectation of the bound in the lemma we have

$$\mathop{\mathbb{E}}_{\mathbf{v}_{1:T}}\left[\frac{1}{T}\sum_{t=1}^{T}\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t\rangle\right] \leq \frac{B\rho}{\sqrt{T}}. \tag{14.9}$$

It is left to show that

$$\mathop{\mathbb{E}}_{\mathbf{v}_{1:T}}\left[\frac{1}{T}\sum_{t=1}^{T}(f(\mathbf{w}^{(t)}) - f(\mathbf{w}^\star))\right] \leq \mathop{\mathbb{E}}_{\mathbf{v}_{1:T}}\left[\frac{1}{T}\sum_{t=1}^{T}\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t\rangle\right], \tag{14.10}$$

which we will hereby prove.

Using the linearity of the expectation we have

$$\mathop{\mathbb{E}}_{\mathbf{v}_{1:T}}\left[\frac{1}{T}\sum_{t=1}^{T}\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t\rangle\right] = \frac{1}{T}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\mathbf{v}_{1:T}}[\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t\rangle].$$

Next, we recall the *law of total expectation*: For every two random variables α, β, and a function g, $\mathbb{E}_\alpha[g(\alpha)] = \mathbb{E}_\beta \mathbb{E}_\alpha[g(\alpha)|\beta]$. Setting $\alpha = \mathbf{v}_{1:t}$ and $\beta = \mathbf{v}_{1:t-1}$ we get that

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathbf{v}_{1:T}}[\langle \mathbf{w}^{(t)} - \mathbf{w}^\star,\ \mathbf{v}_t\rangle] &= \mathop{\mathbb{E}}_{\mathbf{v}_{1:t}}[\langle \mathbf{w}^{(t)} - \mathbf{w}^\star,\ \mathbf{v}_t\rangle] \\ &= \mathop{\mathbb{E}}_{\mathbf{v}_{1:t-1}}\mathop{\mathbb{E}}_{\mathbf{v}_{1:t}}[\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t\rangle \,|\, \mathbf{v}_{1:t-1}].\end{aligned}$$

Once we know $\mathbf{v}_{1:t-1}$, the value of $\mathbf{w}^{(t)}$ is not random any more and therefore

$$\mathop{\mathbb{E}}_{\mathbf{v}_{1:t-1}}\mathop{\mathbb{E}}_{\mathbf{v}_{1:t}}[\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t\rangle \,|\, \mathbf{v}_{1:t-1}] = \mathop{\mathbb{E}}_{\mathbf{v}_{1:t-1}}\langle \mathbf{w}^{(t)} - \mathbf{w}^\star,\ \mathop{\mathbb{E}}_{\mathbf{v}_t}[\mathbf{v}_t \,|\, \mathbf{v}_{1:t-1}]\rangle.$$

Since $\mathbf{w}^{(t)}$ only depends on $\mathbf{v}_{1:t-1}$ and SGD requires that $\mathbb{E}_{\mathbf{v}_t}[\mathbf{v}_t \,|\, \mathbf{w}^{(t)}] \in \partial f(\mathbf{w}^{(t)})$ we obtain that $\mathbb{E}_{\mathbf{v}_t}[\mathbf{v}_t \,|\, \mathbf{v}_{1:t-1}] \in \partial f(\mathbf{w}^{(t)})$. Thus,

$$\mathop{\mathbb{E}}_{\mathbf{v}_{1:t-1}}\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathop{\mathbb{E}}_{\mathbf{v}_t}[\mathbf{v}_t \,|\, \mathbf{v}_{1:t-1}]\rangle \ \geq\ \mathop{\mathbb{E}}_{\mathbf{v}_{1:t-1}}[f(\mathbf{w}^{(t)}) - f(\mathbf{w}^\star)].$$

Overall, we have shown that

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathbf{v}_{1:T}}[\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t\rangle] &\geq \mathop{\mathbb{E}}_{\mathbf{v}_{1:t-1}}[f(\mathbf{w}^{(t)}) - f(\mathbf{w}^\star)] \\ &= \mathop{\mathbb{E}}_{\mathbf{v}_{1:T}}[f(\mathbf{w}^{(t)}) - f(\mathbf{w}^\star)].\end{aligned}$$

https://doi.org/10.1017/CBO9781107298019.015 Published online by Cambridge University Press

Summing over t, dividing by T, and using the linearity of expectation, we get that Equation (14.10) holds, which concludes our proof. □

14.4 VARIANTS

In this section we describe several variants of Stochastic Gradient Descent.

14.4.1 Adding a Projection Step

In the previous analyses of the GD and SGD algorithms, we required that the norm of $\mathbf{w}^\star$ will be at most B, which is equivalent to requiring that $\mathbf{w}^\star$ is in the set $\mathcal{H} = \{\mathbf{w} : \|\mathbf{w}\| \leq B\}$. In terms of learning, this means restricting ourselves to a B-bounded hypothesis class. Yet any step we take in the opposite direction of the gradient (or its expected direction) might result in stepping out of this bound, and there is even no guarantee that $\bar{\mathbf{w}}$ satisfies it. We show in the following how to overcome this problem while maintaining the same convergence rate.

The basic idea is to add a *projection step*; namely, we will now have a two-step update rule, where we first subtract a subgradient from the current value of $\mathbf{w}$ and then project the resulting vector onto $\mathcal{H}$. Formally,

1. $\mathbf{w}^{(t+\frac{1}{2})} = \mathbf{w}^{(t)} - \eta \mathbf{v}_t$
2. $\mathbf{w}^{(t+1)} = \operatorname{argmin}_{\mathbf{w} \in \mathcal{H}} \|\mathbf{w} - \mathbf{w}^{(t+\frac{1}{2})}\|$

The projection step replaces the current value of $\mathbf{w}$ by the vector in $\mathcal{H}$ closest to it.

Clearly, the projection step guarantees that $\mathbf{w}^{(t)} \in \mathcal{H}$ for all t. Since $\mathcal{H}$ is convex this also implies that $\bar{\mathbf{w}} \in \mathcal{H}$ as required. We next show that the analysis of SGD with projections remains the same. This is based on the following lemma.

Lemma 14.9 (Projection Lemma). *Let $\mathcal{H}$ be a closed convex set and let $\mathbf{v}$ be the projection of $\mathbf{w}$ onto $\mathcal{H}$, namely,*

$$\mathbf{v} = \operatorname*{argmin}_{\mathbf{x} \in \mathcal{H}} \|\mathbf{x} - \mathbf{w}\|^2.$$

Then, for every $\mathbf{u} \in \mathcal{H}$,

$$\|\mathbf{w} - \mathbf{u}\|^2 - \|\mathbf{v} - \mathbf{u}\|^2 \geq 0.$$

Proof. By the convexity of $\mathcal{H}$, for every $\alpha \in (0, 1)$ we have that $\mathbf{v} + \alpha(\mathbf{u} - \mathbf{v}) \in \mathcal{H}$. Therefore, from the optimality of $\mathbf{v}$ we obtain

$$\begin{aligned} \|\mathbf{v} - \mathbf{w}\|^2 &\leq \|\mathbf{v} + \alpha(\mathbf{u} - \mathbf{v}) - \mathbf{w}\|^2 \\ &= \|\mathbf{v} - \mathbf{w}\|^2 + 2\alpha \langle \mathbf{v} - \mathbf{w}, \mathbf{u} - \mathbf{v} \rangle + \alpha^2 \|\mathbf{u} - \mathbf{v}\|^2. \end{aligned}$$

Rearranging, we obtain

$$2\langle \mathbf{v} - \mathbf{w}, \mathbf{u} - \mathbf{v} \rangle \geq -\alpha \|\mathbf{u} - \mathbf{v}\|^2.$$

Taking the limit $\alpha \to 0$ we get that

$$\langle \mathbf{v} - \mathbf{w}, \mathbf{u} - \mathbf{v} \rangle \geq 0.$$

https://doi.org/10.1017/CBO9781107298019.015 Published online by Cambridge University Press

Therefore,

$$\begin{aligned}\|\mathbf{w}-\mathbf{u}\|^2 &= \|\mathbf{w}-\mathbf{v}+\mathbf{v}-\mathbf{u}\|^2 \\ &= \|\mathbf{w}-\mathbf{v}\|^2+\|\mathbf{v}-\mathbf{u}\|^2+2\langle \mathbf{v}-\mathbf{w},\ \mathbf{u}-\mathbf{v}\rangle \\ &\geq \|\mathbf{v}-\mathbf{u}\|^2.\end{aligned}$$

□

Equipped with the preceding lemma, we can easily adapt the analysis of SGD to the case in which we add projection steps on a closed and convex set. Simply note that for every t,

$$\begin{aligned}&\|\mathbf{w}^{(t+1)}-\mathbf{w}^\star\|^2-\|\mathbf{w}^{(t)}-\mathbf{w}^\star\|^2 \\ &= \|\mathbf{w}^{(t+1)}-\mathbf{w}^\star\|^2-\|\mathbf{w}^{(t+\frac{1}{2})}-\mathbf{w}^\star\|^2+\|\mathbf{w}^{(t+\frac{1}{2})}-\mathbf{w}^\star\|^2-\|\mathbf{w}^{(t)}-\mathbf{w}^\star\|^2 \\ &\leq \|\mathbf{w}^{(t+\frac{1}{2})}-\mathbf{w}^\star\|^2-\|\mathbf{w}^{(t)}-\mathbf{w}^\star\|^2.\end{aligned}$$

Therefore, Lemma 14.1 holds when we add projection steps and hence the rest of the analysis follows directly.

14.4.2 Variable Step Size

Another variant of SGD is decreasing the step size as a function of t. That is, rather than updating with a constant η, we use η_t. For instance, we can set $\eta_t = \frac{B}{\rho\sqrt{t}}$ and achieve a bound similar to Theorem 14.8. The idea is that when we are closer to the minimum of the function, we take our steps more carefully, so as not to "overshoot" the minimum.

14.4.3 Other Averaging Techniques

We have set the output vector to be $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}^{(t)}$. There are alternative approaches such as outputting $\mathbf{w}^{(t)}$ for some random $t \in [t]$, or outputting the average of $\mathbf{w}^{(t)}$ over the last αT iterations, for some $\alpha \in (0,1)$. One can also take a weighted average of the last few iterates. These more sophisticated averaging schemes can improve the convergence speed in some situations, such as in the case of strongly convex functions defined in the following.

14.4.4 Strongly Convex Functions*

In this section we show a variant of SGD that enjoys a faster convergence rate for problems in which the objective function is strongly convex (see Definition 13.4 of strong convexity in the previous chapter). We rely on the following claim, which generalizes Lemma 13.5.

Claim 14.10. *If f is λ-strongly convex then for every $\mathbf{w}, \mathbf{u}$ and $\mathbf{v} \in \partial f(\mathbf{w})$ we have*

$$\langle \mathbf{w}-\mathbf{u}, \mathbf{v}\rangle \ \geq\ f(\mathbf{w})-f(\mathbf{u})+\tfrac{\lambda}{2}\|\mathbf{w}-\mathbf{u}\|^2.$$

The proof is similar to the proof of Lemma 13.5 and is left as an exercise.

https://doi.org/10.1017/CBO9781107298019.015 Published online by Cambridge University Press

SGD for minimizing a λ-strongly convex function

Goal: Solve $\min_{\mathbf{w}\in\mathcal{H}} f(\mathbf{w})$
parameter: T
initialize: $\mathbf{w}^{(1)} = \mathbf{0}$
for $t = 1, \ldots, T$
  Choose a random vector $\mathbf{v}_t$ s.t. $\mathbb{E}[\mathbf{v}_t | \mathbf{w}^{(t)}] \in \partial f(\mathbf{w}^{(t)})$
  Set $\eta_t = 1/(\lambda t)$
  Set $\mathbf{w}^{(t+\frac{1}{2})} = \mathbf{w}^{(t)} - \eta_t \mathbf{v}_t$
  Set $\mathbf{w}^{(t+1)} = \arg\min_{\mathbf{w}\in\mathcal{H}} \|\mathbf{w} - \mathbf{w}^{(t+\frac{1}{2})}\|^2$
output: $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}^{(t)}$

Theorem 14.11. *Assume that f is λ-strongly convex and that $\mathbb{E}[\|\mathbf{v}_t\|^2] \le \rho^2$. Let $\mathbf{w}^\star \in \operatorname{argmin}_{\mathbf{w}\in\mathcal{H}} f(\mathbf{w})$ be an optimal solution. Then,*

$$\mathbb{E}[f(\bar{\mathbf{w}})] - f(\mathbf{w}^\star) \le \frac{\rho^2}{2\lambda T}(1 + \log(T)).$$

Proof. Let $\nabla^{(t)} = \mathbb{E}[\mathbf{v}_t | \mathbf{w}^{(t)}]$. Since f is strongly convex and $\nabla^{(t)}$ is in the subgradient set of f at $\mathbf{w}^{(t)}$ we have that

$$\langle \mathbf{w}^{(t)} - \mathbf{w}^\star,\ \nabla^{(t)} \rangle \ \ge\ f(\mathbf{w}^{(t)}) - f(\mathbf{w}^\star) + \tfrac{\lambda}{2}\|\mathbf{w}^{(t)} - \mathbf{w}^\star\|^2. \tag{14.11}$$

Next, we show that

$$\langle \mathbf{w}^{(t)} - \mathbf{w}^\star,\ \nabla^{(t)} \rangle \ \le\ \frac{\mathbb{E}[\|\mathbf{w}^{(t)} - \mathbf{w}^\star\|^2 - \|\mathbf{w}^{(t+1)} - \mathbf{w}^\star\|^2]}{2\eta_t} + \frac{\eta_t}{2}\rho^2. \tag{14.12}$$

Since $\mathbf{w}^{(t+1)}$ is the projection of $\mathbf{w}^{(t+\frac{1}{2})}$ onto $\mathcal{H}$, and $\mathbf{w}^\star \in \mathcal{H}$ we have that $\|\mathbf{w}^{(t+\frac{1}{2})} - \mathbf{w}^\star\|^2 \ge \|\mathbf{w}^{(t+1)} - \mathbf{w}^\star\|^2$. Therefore,

$$\begin{aligned}\|\mathbf{w}^{(t)} - \mathbf{w}^\star\|^2 - \|\mathbf{w}^{(t+1)} - \mathbf{w}^\star\|^2 &\ge \|\mathbf{w}^{(t)} - \mathbf{w}^\star\|^2 - \|\mathbf{w}^{(t+\frac{1}{2})} - \mathbf{w}^\star\|^2 \\ &= 2\eta_t \langle \mathbf{w}^{(t)} - \mathbf{w}^\star,\ \mathbf{v}_t \rangle - \eta_t^2 \|\mathbf{v}_t\|^2.\end{aligned}$$

Taking expectation of both sides, rearranging, and using the assumption $\mathbb{E}[\|\mathbf{v}_t\|^2] \le \rho^2$ yield Equation (14.12). Comparing Equation (14.11) and Equation (14.12) and summing over t we obtain

$$\begin{aligned}&\sum_{t=1}^{T} (\mathbb{E}[f(\mathbf{w}^{(t)})] - f(\mathbf{w}^\star)) \\ &\quad\le \mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{\|\mathbf{w}^{(t)} - \mathbf{w}^\star\|^2 - \|\mathbf{w}^{(t+1)} - \mathbf{w}^\star\|^2}{2\eta_t} - \tfrac{\lambda}{2}\|\mathbf{w}^{(t)} - \mathbf{w}^\star\|^2\right)\right] + \frac{\rho^2}{2}\sum_{t=1}^{T}\eta_t.\end{aligned}$$

Next, we use the definition $\eta_t = 1/(\lambda t)$ and note that the first sum on the right-hand side of the equation collapses to $-\lambda T \|\mathbf{w}^{(T+1)} - \mathbf{w}^\star\|^2 \le 0$. Thus,

$$\sum_{t=1}^{T} (\mathbb{E}[f(\mathbf{w}^{(t)})] - f(\mathbf{w}^\star)) \le \frac{\rho^2}{2\lambda}\sum_{t=1}^{T}\frac{1}{t} \le \frac{\rho^2}{2\lambda}(1 + \log(T)).$$

https://doi.org/10.1017/CBO9781107298019.015 Published online by Cambridge University Press

The theorem follows from the preceding by dividing by T and using Jensen's inequality. □

Remark 14.3. Rakhlin, Shamir, and Sridharan ((2012)) derived a convergence rate in which the $\log(T)$ term is eliminated for a variant of the algorithm in which we output the average of the last $T/2$ iterates, $\bar{\mathbf{w}} = \frac{2}{T}\sum_{t=T/2+1}^{T} \mathbf{w}^{(t)}$. Shamir and Zhang (2013) have shown that Theorem 14.11 holds even if we output $\bar{\mathbf{w}} = \mathbf{w}^{(T)}$.

14.5 LEARNING WITH SGD

We have so far introduced and analyzed the SGD algorithm for general convex functions. Now we shall consider its applicability to learning tasks.

14.5.1 SGD for Risk Minimization

Recall that in learning we face the problem of minimizing the risk function

$$L_{\mathcal{D}}(\mathbf{w}) = \mathop{\mathbb{E}}_{z\sim\mathcal{D}}[\ell(\mathbf{w}, z)].$$

We have seen the method of empirical risk minimization, where we minimize the empirical risk, $L_S(\mathbf{w})$, as an estimate to minimizing $L_{\mathcal{D}}(\mathbf{w})$. SGD allows us to take a different approach and minimize $L_{\mathcal{D}}(\mathbf{w})$ directly. Since we do not know $\mathcal{D}$, we cannot simply calculate $\nabla L_{\mathcal{D}}(\mathbf{w}^{(t)})$ and minimize it with the GD method. With SGD, however, all we need is to find an unbiased estimate of the gradient of $L_{\mathcal{D}}(\mathbf{w})$, that is, a random vector whose conditional expected value is $\nabla L_{\mathcal{D}}(\mathbf{w}^{(t)})$. We shall now see how such an estimate can be easily constructed.

For simplicity, let us first consider the case of differentiable loss functions. Hence the risk function $L_{\mathcal{D}}$ is also differentiable. The construction of the random vector $\mathbf{v}_t$ will be as follows: First, sample $z \sim \mathcal{D}$. Then, define $\mathbf{v}_t$ to be the gradient of the function $\ell(\mathbf{w}, z)$ with respect to $\mathbf{w}$, at the point $\mathbf{w}^{(t)}$. Then, by the linearity of the gradient we have

$$\mathbb{E}[\mathbf{v}_t|\mathbf{w}^{(t)}] = \mathop{\mathbb{E}}_{z\sim\mathcal{D}}[\nabla\ell(\mathbf{w}^{(t)}, z)] = \nabla \mathop{\mathbb{E}}_{z\sim\mathcal{D}}[\ell(\mathbf{w}^{(t)}, z)] = \nabla L_{\mathcal{D}}(\mathbf{w}^{(t)}). \tag{14.13}$$

The gradient of the loss function $\ell(\mathbf{w}, z)$ at $\mathbf{w}^{(t)}$ is therefore an unbiased estimate of the gradient of the risk function $L_{\mathcal{D}}(\mathbf{w}^{(t)})$ and is easily constructed by sampling a single fresh example $z \sim \mathcal{D}$ at each iteration t.

The same argument holds for nondifferentiable loss functions. We simply let $\mathbf{v}_t$ be a subgradient of $\ell(\mathbf{w}, z)$ at $\mathbf{w}^{(t)}$. Then, for every $\mathbf{u}$ we have

$$\ell(\mathbf{u}, z) - \ell(\mathbf{w}^{(t)}, z) \geq \langle \mathbf{u} - \mathbf{w}^{(t)}, \mathbf{v}_t \rangle.$$

Taking expectation on both sides with respect to $z \sim \mathcal{D}$ and conditioned on the value of $\mathbf{w}^{(t)}$ we obtain

$$\begin{aligned} L_{\mathcal{D}}(\mathbf{u}) - L_{\mathcal{D}}(\mathbf{w}^{(t)}) &= \mathbb{E}[\ell(\mathbf{u}, z) - \ell(\mathbf{w}^{(t)}, z)|\mathbf{w}^{(t)}] \\ &\geq \mathbb{E}[\langle \mathbf{u} - \mathbf{w}^{(t)}, \mathbf{v}_t \rangle|\mathbf{w}^{(t)}] \\ &= \langle \mathbf{u} - \mathbf{w}^{(t)}, \mathbb{E}[\mathbf{v}_t|\mathbf{w}^{(t)}] \rangle. \end{aligned}$$

https://doi.org/10.1017/CBO9781107298019.015 Published online by Cambridge University Press

It follows that $\mathbb{E}[\mathbf{v}_t|\mathbf{w}^{(t)}]$ is a subgradient of $L_{\mathcal{D}}(\mathbf{w})$ at $\mathbf{w}^{(t)}$.

To summarize, the stochastic gradient descent framework for minimizing the risk is as follows.

Stochastic Gradient Descent (SGD) for minimizing $L_{\mathcal{D}}(\mathbf{w})$

parameters: Scalar $\eta > 0$, integer $T > 0$
initialize: $\mathbf{w}^{(1)} = \mathbf{0}$
for $t = 1, 2, \ldots, T$
  sample $z \sim \mathcal{D}$
  pick $\mathbf{v}_t \in \partial \ell(\mathbf{w}^{(t)}, z)$
  update $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \mathbf{v}_t$
output $\bar{\mathbf{w}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{w}^{(t)}$

We shall now use our analysis of SGD to obtain a sample complexity analysis for learning convex-Lipschitz-bounded problems. Theorem 14.8 yields the following:

Corollary 14.12. *Consider a convex-Lipschitz-bounded learning problem with parameters ρ, B. Then, for every $\epsilon > 0$, if we run the SGD method for minimizing $L_{\mathcal{D}}(\mathbf{w})$ with a number of iterations (i.e., number of examples)*

$$T \geq \frac{B^2 \rho^2}{\epsilon^2}$$

and with $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$, then the output of SGD satisfies

$$\mathbb{E}[L_{\mathcal{D}}(\bar{\mathbf{w}})] \leq \min_{\mathbf{w} \in \mathcal{H}} L_{\mathcal{D}}(\mathbf{w}) + \epsilon.$$

It is interesting to note that the required sample complexity is of the same order of magnitude as the sample complexity guarantee we derived for regularized loss minimization. In fact, the sample complexity of SGD is even better than what we have derived for regularized loss minimization by a factor of 8.

14.5.2 Analyzing SGD for Convex-Smooth Learning Problems

In the previous chapter we saw that the regularized loss minimization rule also learns the class of convex-smooth-bounded learning problems. We now show that the SGD algorithm can be also used for such problems.

Theorem 14.13. *Assume that for all z, the loss function $\ell(\cdot, z)$ is convex, β-smooth, and nonnegative. Then, if we run the SGD algorithm for minimizing $L_{\mathcal{D}}(\mathbf{w})$ we have that for every $\mathbf{w}^\star$,*

$$\mathbb{E}[L_{\mathcal{D}}(\bar{\mathbf{w}})] \leq \frac{1}{1 - \eta\beta} \left(L_{\mathcal{D}}(\mathbf{w}^\star) + \frac{\|\mathbf{w}^\star\|^2}{2\eta T} \right).$$

Proof. Recall that if a function is β-smooth and nonnegative then it is self-bounded:

$$\|\nabla f(\mathbf{w})\|^2 \leq 2\beta f(\mathbf{w}).$$

https://doi.org/10.1017/CBO9781107298019.015 Published online by Cambridge University Press

To analyze SGD for convex-smooth problems, let us define $z_1, \ldots, z_T$ the random samples of the SGD algorithm, let $f_t(\cdot) = \ell(\cdot, z_t)$, and note that $\mathbf{v}_t = \nabla f_t(\mathbf{w}^{(t)})$. For all t, f_t is a convex function and therefore $f_t(\mathbf{w}^{(t)}) - f_t(\mathbf{w}^\star) \leq \langle \mathbf{v}_t, \mathbf{w}^{(t)} - \mathbf{w}^\star \rangle$. Summing over t and using Lemma 14.1 we obtain

$$\sum_{t=1}^{T} (f_t(\mathbf{w}^{(t)}) - f_t(\mathbf{w}^\star)) \leq \sum_{t=1}^{T} \langle \mathbf{v}_t, \mathbf{w}^{(t)} - \mathbf{w}^\star \rangle \leq \frac{\|\mathbf{w}^\star\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\mathbf{v}_t\|^2.$$

Combining the preceding with the self-boundedness of f_t yields

$$\sum_{t=1}^{T} (f_t(\mathbf{w}^{(t)}) - f_t(\mathbf{w}^\star)) \leq \frac{\|\mathbf{w}^\star\|^2}{2\eta} + \eta\beta \sum_{t=1}^{T} f_t(\mathbf{w}^{(t)}).$$

Dividing by T and rearranging, we obtain

$$\frac{1}{T} \sum_{t=1}^{T} f_t(\mathbf{w}^{(t)}) \leq \frac{1}{1 - \eta\beta} \left(\frac{1}{T} \sum_{t=1}^{T} f_t(\mathbf{w}^\star) + \frac{\|\mathbf{w}^\star\|^2}{2\eta T} \right).$$

Next, we take expectation of the two sides of the preceding equation with respect to $z_1, \ldots, z_T$. Clearly, $\mathbb{E}[f_t(\mathbf{w}^\star)] = L_{\mathcal{D}}(\mathbf{w}^\star)$. In addition, using the same argument as in the proof of Theorem 14.8 we have that

$$\mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} f_t(\mathbf{w}^{(t)}) \right] = \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} L_{\mathcal{D}}(\mathbf{w}^{(t)}) \right] \geq \mathbb{E}[L_{\mathcal{D}}(\bar{\mathbf{w}})].$$

Combining all we conclude our proof. □

As a direct corollary we obtain:

Corollary 14.14. *Consider a convex-smooth-bounded learning problem with parameters β, B. Assume in addition that $\ell(\mathbf{0}, z) \leq 1$ for all $z \in Z$. For every $\epsilon > 0$, set $\eta = \frac{1}{\beta(1+3/\epsilon)}$. Then, running SGD with $T \geq 12B^2\beta/\epsilon^2$ yields*

$$\mathbb{E}[L_{\mathcal{D}}(\bar{\mathbf{w}})] \leq \min_{\mathbf{w} \in \mathcal{H}} L_{\mathcal{D}}(\mathbf{w}) + \epsilon.$$

14.5.3 SGD for Regularized Loss Minimization

We have shown that SGD enjoys the same worst-case sample complexity bound as regularized loss minimization. However, on some distributions, regularized loss minimization may yield a better solution. Therefore, in some cases we may want to solve the optimization problem associated with regularized loss minimization, namely,[1]

$$\min_{\mathbf{w}} \left(\frac{\lambda}{2} \|\mathbf{w}\|^2 + L_S(\mathbf{w}) \right). \tag{14.14}$$

Since we are dealing with convex learning problems in which the loss function is convex, the preceding problem is also a convex optimization problem that can be solved using SGD as well, as we shall see in this section.

[1] We divided λ by 2 for convenience.

https://doi.org/10.1017/CBO9781107298019.015 Published online by Cambridge University Press

Define $f(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|^2 + L_S(\mathbf{w})$. Note that f is a λ-strongly convex function; therefore, we can apply the SGD variant given in Section 14.4.4 (with $\mathcal{H} = \mathbb{R}^d$). To apply this algorithm, we only need to find a way to construct an unbiased estimate of a subgradient of f at $\mathbf{w}^{(t)}$. This is easily done by noting that if we pick z uniformly at random from S, and choose $\mathbf{v}_t \in \partial\ell(\mathbf{w}^{(t)}, z)$ then the expected value of $\lambda\mathbf{w}^{(t)} + \mathbf{v}_t$ is a subgradient of f at $\mathbf{w}^{(t)}$.

To analyze the resulting algorithm, we first rewrite the update rule (assuming that $\mathcal{H} = \mathbb{R}^d$ and therefore the projection step does not matter) as follows

$$\begin{aligned}
\mathbf{w}^{(t+1)} &= \mathbf{w}^{(t)} - \frac{1}{\lambda t}\left(\lambda\mathbf{w}^{(t)} + \mathbf{v}_t\right) \\
&= \left(1 - \frac{1}{t}\right)\mathbf{w}^{(t)} - \frac{1}{\lambda t}\mathbf{v}_t \\
&= \frac{t-1}{t}\mathbf{w}^{(t)} - \frac{1}{\lambda t}\mathbf{v}_t \\
&= \frac{t-1}{t}\left(\frac{t-2}{t-1}\mathbf{w}^{(t-1)} - \frac{1}{\lambda(t-1)}\mathbf{v}_{t-1}\right) - \frac{1}{\lambda t}\mathbf{v}_t \\
&= -\frac{1}{\lambda t}\sum_{i=1}^{t}\mathbf{v}_i. \qquad (14.15)
\end{aligned}$$

If we assume that the loss function is ρ-Lipschitz, it follows that for all t we have $\|\mathbf{v}_t\| \leq \rho$ and therefore $\|\lambda\mathbf{w}^{(t)}\| \leq \rho$, which yields

$$\|\lambda\mathbf{w}^{(t)} + \mathbf{v}_t\| \leq 2\rho.$$

Theorem 14.11 therefore tells us that after performing T iterations we have that

$$\mathbb{E}[f(\bar{\mathbf{w}})] - f(\mathbf{w}^\star) \leq \frac{4\rho^2}{\lambda T}(1 + \log(T)).$$

14.6 SUMMARY

We have introduced the Gradient Descent and Stochastic Gradient Descent algorithms, along with several of their variants. We have analyzed their convergence rate and calculated the number of iterations that would guarantee an expected objective of at most ϵ plus the optimal objective. Most importantly, we have shown that by using SGD we can directly minimize the risk function. We do so by sampling a point i.i.d from $\mathcal{D}$ and using a subgradient of the loss of the current hypothesis $\mathbf{w}^{(t)}$ at this point as an unbiased estimate of the gradient (or a subgradient) of the risk function. This implies that a bound on the number of iterations also yields a sample complexity bound. Finally, we have also shown how to apply the SGD method to the problem of regularized risk minimization. In future chapters we show how this yields extremely simple solvers to some optimization problems associated with regularized risk minimization.

https://doi.org/10.1017/CBO9781107298019.015 Published online by Cambridge University Press

14.7 BIBLIOGRAPHIC REMARKS

SGD dates back to Robbins and Monro (1951). It is especially effective in large scale machine learning problems. See, for example, (Murata 1998, Le Cun 2004, Zhang 2004, Bottou & Bousquet 2008, Shalev-Shwartz, Singer & Srebro 2007, Shalev-Shwartz & Srebro 2008). In the optimization community it was studied in the context of *stochastic optimization*. See, for example, (Nemirovski & Yudin 1978, Nesterov & Nesterov 2004, Nesterov 2005, Nemirovski, Juditsky, Lan & Shapiro 2009, Shapiro, Dentcheva & Ruszczyński 2009).

The bound we have derived for strongly convex function is due to Hazan, Agarwal, and Kale (2007). As mentioned previously, improved bounds have been obtained in Rakhlin, Shamir & Sridharan (2012).

14.8 EXERCISES

14.1 Prove Claim 14.10. *Hint:* Extend the proof of Lemma 13.5.

14.2 Prove Corollary 14.14.

14.3 **Perceptron as a subgradient descent algorithm:** Let $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)) \in (\mathbb{R}^d \times \{\pm 1\})^m$. Assume that there exists $\mathbf{w} \in \mathbb{R}^d$ such that for every $i \in [m]$ we have $y_i \langle \mathbf{w}, \mathbf{x}_i \rangle \geq 1$, and let $\mathbf{w}^\star$ be a vector that has the minimal norm among all vectors that satisfy the preceding requirement. Let $R = \max_i \|\mathbf{x}_i\|$. Define a function

$$f(\mathbf{w}) = \max_{i \in [m]} (1 - y_i \langle \mathbf{w}, \mathbf{x}_i \rangle).$$

- Show that $\min_{\mathbf{w}:\|\mathbf{w}\| \leq \|\mathbf{w}^\star\|} f(\mathbf{w}) = 0$ and show that any $\mathbf{w}$ for which $f(\mathbf{w}) < 1$ separates the examples in S.
- Show how to calculate a subgradient of f.
- Describe and analyze the subgradient descent algorithm for this case. Compare the algorithm and the analysis to the Batch Perceptron algorithm given in Section 9.1.2.

14.4 **Variable step size (*):** Prove an analog of Theorem 14.8 for SGD with a variable step size, $\eta_t = \frac{B}{\rho\sqrt{t}}$.

https://doi.org/10.1017/CBO9781107298019.015 Published online by Cambridge University Press