Chapter 4

Subgradient Descent

Contents

4.1 Subgradients

Definition 4.1. *Let* $f : \mathbf{dom}(f) \to \mathbb{R}$. *Then* $\mathbf{g} \in \mathbb{R}^d$ *is a* subgradient *of* f *at* $\mathbf{x} \in \mathbf{dom}(f)$ *if*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^\top(\mathbf{y} - \mathbf{x}) \quad \forall \mathbf{y} \in \mathbf{dom}(f). \tag{4.1}$$

The set of subgradients of f *at* $\mathbf{x}$ *is denoted by* $\partial f(\mathbf{x})$.

The notion of a subgradient can be seen as a generalization of the gradient, for functions which are not necessarily differentiable. A prominent example is the ℓ_1-norm, which we have discussed in Exercise 7.

The above definition might look suspiciously familiar to the first-order characterization of convexity (1.2) we discussed earlier. Indeed, the only difference is that here we have replaced $\nabla f(\mathbf{x})$ by $\mathbf{g}$. It turns out that convexity is equivalent to the existence of subgradients everywhere. So we get a "first order characterization" of convexity that also covers the non-differentiable case.

Lemma 4.2 (Exercise 23)**.** *A function* $f : \mathbf{dom}(f) \to \mathbb{R}$ *is convex if and only if* $\mathbf{dom}(f)$ *is convex and* $\partial f(\mathbf{x}) \neq \emptyset$ *for all* $\mathbf{x} \in \mathbf{dom}(f)$.

It turns out that Lemma 2.2 also generalizes to subgradients.

Lemma 4.3 (Exercise 24)**.** *Let* $f : \mathbb{R}^d \to \mathbb{R}$ *be convex,* $B \in \mathbb{R}_+$. *Then the following two statements are equivalent.*

(i) $\|\mathbf{g}\| \leq B$ *for all* $\mathbf{x} \in \mathbb{R}^d$ *and all* $\mathbf{g} \in \partial f(\mathbf{x})$.

(ii) $|f(\mathbf{x}) - f(\mathbf{y})| \leq B\|\mathbf{x} - \mathbf{y}\|$ *for all* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

Subgradient optimality condition. Subgradients also allow us to describe cases of optimality for functions which are not necessarily differentiable (and not necessarily convex), generalizing Lemma 1.12:

Lemma 4.4. *Suppose that* f *is any function over* $\mathbf{dom}(f)$*, and* $\mathbf{x} \in \mathbf{dom}(f)$. *If* $\mathbf{0} \in \partial f(\mathbf{x})$*, then* $\mathbf{x}$ *is a global minimum.*

Proof. By (4.1), $\mathbf{g} = \mathbf{0} \in \partial f(\mathbf{x})$ gives

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^\top(\mathbf{y} - \mathbf{x}) = f(\mathbf{x})$$

for all $\mathbf{y} \in \mathbf{dom}(f)$, so $\mathbf{x}$ is a global minimum. □

4.2 The algorithm

An iteration of *subgradient descent* is defined as

$$\begin{aligned} &\text{Let } \mathbf{g}_t \in \partial f(\mathbf{x}_t) \\ &\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma \mathbf{g}_t. \end{aligned} \tag{4.2}$$

4.3 Bounded subgradients: $\mathcal{O}(1/\varepsilon^2)$ steps

The following result gives the convergence for Subgradient Descent. It is identical to Theorem 2.1, up to relaxing the requirement of differentiability.

Theorem 4.5. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and B-Lipschitz continuous on $\mathbb{R}^d$ with a global minimum $\mathbf{x}^\star$; furthermore, suppose that $\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq R$. Choosing the constant stepsize*

$$\gamma := \frac{R}{B\sqrt{T}},$$

subgradient descent (4.2) yields

$$\frac{1}{T}\sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{RB}{\sqrt{T}}.$$

Proof. The proof is identically to the vanilla analysis for gradient descent presented in Section 2.3. The only change is that the use of the first-order characterization of convexity as in the very first step (2.2) of the vanilla analysis is replaced by the subgradient property (4.1). □

Projected subgradient descent. Theorem 3.2 for constrained optimization in $\mathcal{O}(1/\varepsilon^2)$ steps directly extends to the case of subgradient descent as well.

4.4 Optimality of first-order methods

With all the convergence rates we have seen so far, a very natural question to ask is if these rates are best possible or not. Surprisingly, the rate can indeed not be improved in general.

Theorem 4.6 (Nesterov). *For any $T \leq d-1$ and starting point $\mathbf{x}_0$, there is a function f in the problem class of B-Lipschitz functions over $\mathbb{R}^d$, such that any (sub)gradient method has an objective error at least*

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \geq \frac{RB}{2(1+\sqrt{T+1})} .$$

The above theorem applies to all first-order methods which form iterates by linearly combining past iterates and (sub)gradients, and requires the dimension d to be sufficiently large.

4.5 Exercises

Exercise 23. *Prove the easy direction of Lemma 4.2, meaning that the existence of subgradients everywhere implies convexity!*

Solution: Let's assume that we have subgradients everywhere. With $\mathbf{g} \in \partial f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y})$, (4.1) yields

$$\begin{array}{rcl} f(\mathbf{x}) & \geq & f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y}) + \mathbf{g}^T((1-\lambda)(\mathbf{x}-\mathbf{y})), \\ f(\mathbf{y}) & \geq & f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y}) + \mathbf{g}^T(\lambda(\mathbf{y}-\mathbf{x})). \end{array}$$

Adding up these two inequalities with multiples λ and $1-\lambda$ cancels the subgradient terms and yields

$$\lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y}) \geq f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y}),$$

which is convexity.

Exercise 24. *Prove Lemma 4.3 (Lipschitz continuity and bounded subgradients).*

Solution:

Chapter 5

Stochastic Gradient Descent

Contents

5.1 The algorithm

Many objective functions occurring in machine learning are formulated as *sum structured objective functions*

$$f(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}). \tag{5.1}$$

Here f_i is typically the cost function of the i-th datapoint, taken from a training set of n elements in total.

We have already seen an example for this: the loss function (1.9) in the handwritten digit recognition (Section 1.6.1) has one term for each of the n training images $\mathbf{x} \in P$:

$$\ell(W) = -\sum_{\mathbf{x}\in P} \ln z_{d(\mathbf{x})}(W\mathbf{x}).$$

The normalizing factor $1/n$ that we assume in the general setting (5.1) will just simplify the following a bit.

An iteration of *stochastic gradient descent* (SGD) in its basic form is defined as

$$\begin{aligned} &\text{sample } i \in [n] \text{ uniformly at random} \\ &\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma_t \nabla f_i(\mathbf{x}_t). \end{aligned} \tag{5.2}$$

This update looks almost identical to the classical gradient method, the only difference being that we have computed the gradient not of the entire f but only of one particular (randomly chosen) function f_i. As we will need varying stepsizes a bit later, we allow for the stepsize to depend on t now.

In the above setting, the update vector $\mathbf{g}_t := \nabla f_i(\mathbf{x}_t)$ is called a *stochastic gradient*. Formally, $\mathbf{g}_t$ is a vector of d random variables, but we will also simply call this a random variable.

The vector $\mathbf{g}_t$ may be far from the true gradient, and of high variance, but in expectation over the random choice of i, it does coincide with the full gradient of f. We formalize this as

$$\mathbb{E}\big[\mathbf{g}_t \big| \mathbf{x}_t\big] = \nabla f(\mathbf{x}_t). \tag{5.3}$$

Here, $\mathbb{E}\big[\mathbf{g}_t \big| \mathbf{x}_t\big]$ is itself a random variable, the conditional expectation of $\mathbf{g}_t$, given the random variable $\mathbf{x}_t$. Similarly, the gradient $\nabla f(\mathbf{x}_t)$ is—as a function of the random variable $\mathbf{x}_t$—now also a random variable. Hence, (5.3) is an equality between two random variables. It says that for all $\mathbf{x}$,

$$\mathbb{E}\big[\mathbf{g}_t \big| \mathbf{x}_t\big](\mathbf{x}) = \mathbb{E}\big[\mathbf{g}_t \big| \mathbf{x}_t = \mathbf{x}\big] = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\mathbf{x}) = \nabla f(\mathbf{x}) = \nabla f(\mathbf{x}_t)(\mathbf{x}).$$

Exercise 25 lets you recall some basics around conditional expectations. Under (5.3) we say that the stochastic gradient $\mathbf{g}_t$ is an *unbiased* estimator of the gradient, for any time-step t.

The crucial advantage of SGD versus its classical gradient descent counterpart is the efficiency per iteration: While computing the full gradient for a sum structured problem (5.1) would require us to compute n individual gradients of the f_i functions, an iteration of SGD requires only a single one of those, and therefore is n times cheaper. SGD has therefore become the main workhorse for training machine learning models. Whether such cheaper iterations also give similar progress is another question, which we analyze next.

5.2 Stochastic vanilla analysis

It turns out that we can redo major parts of the vanilla analysis with $\nabla f(\mathbf{x}_t)$ replaced by $\mathbf{g}_t$, except that we cannot get started with

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star).$$

Indeed, this inequality only holds in expectation, a fact that we prove and exploit later. But we can continue rewriting the right-hand side exactly as we did in the vanilla analysis. For now, let's assume fixed stepsize $\gamma_t := \gamma$.

By definition of stochastic gradient descent (5.2), $\mathbf{g}_t = (\mathbf{x}_t - \mathbf{x}_{t+1})/\gamma$, hence

$$\mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star) = \frac{1}{\gamma}(\mathbf{x}_t - \mathbf{x}_{t+1})^\top (\mathbf{x}_t - \mathbf{x}^\star). \tag{5.4}$$

The basic vector equation $2\mathbf{v}^\top\mathbf{w} = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v}-\mathbf{w}\|^2$ yields

$$\begin{aligned}
\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star) &= \frac{1}{2\gamma}\left(\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\right) \\
&= \frac{1}{2\gamma}\left(\gamma^2\|\mathbf{g}_t\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\right), \qquad (5.5)
\end{aligned}$$

using the definition (5.2) of SGD again. Finally, the telescoping sum:

$$\begin{aligned}
\sum_{t=0}^{T-1}\left(\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star)\right) &\le \frac{\gamma}{2}\sum_{t=0}^{T-1}\|\mathbf{g}_t\|^2 + \frac{1}{2\gamma}\left(\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 - \|\mathbf{x}_T - \mathbf{x}^\star\|^2\right) \\
&\le \frac{\gamma}{2}\sum_{t=0}^{T-1}\|\mathbf{g}_t\|^2 + \frac{1}{2\gamma}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2. \qquad (5.6)
\end{aligned}$$

5.2.1 Bounded stochastic gradients: $\mathcal{O}(1/\varepsilon^2)$ steps

To get a first result out of the vanilla analysis, we assumed in Section 2.3 that $\|\nabla f(\mathbf{x})\|^2 \le L^2$ for all $\mathbf{x} \in \mathbb{R}^d$, where L was a constant. Here, we are assuming the same for the *expected* squared norms of our stochastic gradients, except that the constant is now called B^2. And we are getting the same result, expect that it now holds for the *expected* function values.

Theorem 5.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable, $\mathbf{x}^\star$ a global minimum; furthermore, suppose that $\|\mathbf{x}_0 - \mathbf{x}^\star\| \le R$, and that $\mathbb{E}\big[\|\mathbf{g}_t\|^2\big] \le B^2$ for all t. Choosing the constant stepsize*

$$\gamma := \frac{R}{B\sqrt{T}}$$

stochastic gradient descent (5.2) yields

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[f(\mathbf{x}_t)\big] - f(\mathbf{x}^\star) \le \frac{RB}{\sqrt{T}}.$$

Proof. Using convexity and unbiasedness of $\mathbf{g}_t$, we compute

$$\begin{aligned}
\mathbb{E}\big[f(\mathbf{x}_t)\big] - f(\mathbf{x}^\star) &= \mathbb{E}\big[f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big] \\
&\le \mathbb{E}\big[\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{x}^\star)\big] \\
&= \mathbb{E}\big[\mathbb{E}\big[\mathbf{g}_t \big| \mathbf{x}_t\big]^\top(\mathbf{x}_t - \mathbf{x}^\star)\big] \\
&= \mathbb{E}\big[\mathbb{E}\big[\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star)\big|\mathbf{x}_t\big]\big] \\
&= \mathbb{E}\big[\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star)\big],
\end{aligned}$$

where the second-to-last step uses linearity of (conditional) expectations, while the last step is known as the *tower rule*; see again Exercise 25. Now we can again use linearity of expectation and then (5.6). We get

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[f(\mathbf{x}_t)\big]-f(\mathbf{x}^\star) &\leq \frac{1}{T}\mathbb{E}\big[\sum_{t=0}^{T-1}\mathbf{g}_t^\top(\mathbf{x}_t-\mathbf{x}^\star)\big] \\
&= \frac{1}{T}\mathbb{E}\big[\frac{\gamma}{2}\sum_{t=0}^{T-1}\|\mathbf{g}_t\|^2+\frac{1}{2\gamma}\|\mathbf{x}_0-\mathbf{x}^\star\|^2\big] \\
&= \frac{1}{T}\left(\frac{\gamma}{2}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\mathbf{g}_t\|^2\big]+\frac{1}{2\gamma}\|\mathbf{x}_0-\mathbf{x}^\star\|^2\right) \\
&\leq \frac{RB}{\sqrt{T}},
\end{aligned}$$

after plugging in our value of γ and the assumption on $\mathbb{E}\big[\|\mathbf{g}_t\|^2\big]$ and $\|\mathbf{x}_0-\mathbf{x}^\star\|$. □

Stochastic Subgradient Descent. For problems which are not necessarily differentiable, we modify SGD to use a subgradient of f_i in each iteration. The update of stochastic subgradient descent is given by

$$\begin{aligned}
&\text{sample } i\in[n] \text{ uniformly at random} \\
&\text{let } \mathbf{g}_t\in\partial f_i(\mathbf{x}_t) \qquad (5.7)\\
&\mathbf{x}_{t+1}:=\mathbf{x}_t-\gamma_t\mathbf{g}_t.
\end{aligned}$$

In other words, we are using an unbiased estimate of a subgradient at each step, $\mathbb{E}\big[\mathbf{g}_t\big|\mathbf{x}_t\big]\in\partial f(\mathbf{x}_t)$.

The above analysis of convergence in $\mathcal{O}(1/\varepsilon^2)$ steps directly extends to the case of subgradient descent here as well, by using the subgradient property (4.1) at the beginning of the proof, where convexity was applied.

Constrained optimization. For constrained optimization, Theorem 5.1 for the convergence in $\mathcal{O}(1/\varepsilon^2)$ steps directly extends to constrained problems as well. After every step of SGD, projection back to X is applied as usual. The resulting algorithm is called *projected SGD*.

5.2.2 Strong convexity: $\mathcal{O}(1/\varepsilon)$ steps

It is possible to strengthen our above SGD analysis. One way to do so is under the additional assumption of strong convexity of the objective function f (as in Definition 2.8). For this case, we will now for the first time depart from algorithm variants with a constant stepsize γ, but instead use a time-varying stepsize γ_t decreasing over the time t.

Theorem 5.2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and strongly convex with parameter $\mu > 0$; let $\mathbf{x}^\star$ be the unique global minimum of f, and $\mathbb{E}\big[\|\mathbf{g}_t\|^2\big] \leq B^2$ for all $\mathbf{x}$. Choosing the decreasing stepsize*

$$\gamma_t := \frac{2}{\mu(t+1)}$$

stochastic gradient descent (5.2) yields

$$\mathbb{E}\Big[f\Big(\frac{2}{T(T+1)}\sum_{t=1}^{T} t\cdot\mathbf{x}_t\Big) - f(\mathbf{x}^\star)\Big] \leq \frac{2B^2}{\mu(T+1)}.$$

Proof. We use the definition of the SGD step, and the basic vector equation $2\mathbf{v}^\top\mathbf{w} = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v}-\mathbf{w}\|^2$ which we have also used in the vanilla analysis, we have

$$\begin{aligned}\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 &= \|\mathbf{x}_t - \gamma_t\mathbf{g}_t - \mathbf{x}^\star\|^2 \\ &= \|\mathbf{x}_t - \mathbf{x}^\star\|^2 + \gamma_t^2\,\|\mathbf{g}_t\|^2 - 2\gamma_t\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star)\end{aligned}$$

Taking conditional expectation on both sides, and using unbiasedness (5.3), we get

$$\begin{aligned}&\mathbb{E}\big[\,\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \;\big|\, \mathbf{x}_t\big] \\ &= \|\mathbf{x}_t - \mathbf{x}^\star\|^2 + \gamma_t^2\mathbb{E}\big[\,\|\mathbf{g}_t\|^2 \;\big|\, \mathbf{x}_t\big] - 2\gamma_t\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{x}^\star)\end{aligned} \tag{5.8}$$

Strong convexity (2.12) with $\mathbf{y} = \mathbf{x}^*, \mathbf{x} = \mathbf{x}_t$ yields

$$\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{x}^\star) \geq f(\mathbf{x}_t) - f(\mathbf{x}^\star) + \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}^\star\|^2,$$

hence (5.8) further yields

$$\begin{aligned}&\mathbb{E}\Big[\,\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \;\Big|\, \mathbf{x}_t\Big] \\ &\leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2 + \gamma_t^2\mathbb{E}\Big[\,\|\mathbf{g}_t\|^2 \;\Big|\, \mathbf{x}_t\Big] - 2\gamma_t\Big[f(\mathbf{x}_t) - f(\mathbf{x}^\star) + \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}^\star\|^2\Big]\end{aligned}$$

Rearranging and again taking expectation $\mathbb{E}$ over the randomness of the entire sequence of steps $0, 1, \ldots, t$, as well as using $\mathbb{E}\big[\|\mathbf{g}_t\|^2\big] \leq B^2$, we have

$$\begin{aligned}
2\gamma_t \mathbb{E}[f(\mathbf{x}_t) - f(\mathbf{x}^\star)] &\leq \gamma_t^2 B^2 + (1 - \mu\gamma_t)\mathbb{E}\big[\,\|\mathbf{x}_t - \mathbf{x}^\star\|^2\,\big] - \mathbb{E}\big[\,\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\,\big] \\
\mathbb{E}[f(\mathbf{x}_t) - f(\mathbf{x}^\star)] &\leq \frac{B^2\gamma_t}{2} + \frac{(\gamma_t^{-1} - \mu)}{2}\mathbb{E}\big[\,\|\mathbf{x}_t - \mathbf{x}^\star\|^2\,\big] - \frac{\gamma_t^{-1}}{2}\mathbb{E}\big[\,\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\,\big]
\end{aligned}$$

Now using the stepsize $\gamma_t := \frac{2}{\mu(t+1)}$, and multiplying the above inequality by t both the sides,

$$\begin{aligned}
t\mathbb{E}\Big[f(\mathbf{x}_t) - f(\mathbf{x}^\star)\Big] &\leq \frac{B^2 t}{\mu(t+1)} + \frac{\mu}{4}\Big[t(t-1)\mathbb{E}\Big[\,\|\mathbf{x}_t - \mathbf{x}^\star\|^2\,\Big] - t(t+1)\mathbb{E}\Big[\,\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\,\Big]\Big] \\
&\leq \frac{B^2}{\mu} + \frac{\mu}{4}\Big[t(t-1)\mathbb{E}\Big[\,\|\mathbf{x}_t - \mathbf{x}^\star\|^2\,\Big] - t(t+1)\mathbb{E}\Big[\,\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\,\Big]\Big]
\end{aligned}$$

Summing from $t = 1, \ldots, T$, we obtain the following telescoping sum,

$$\sum_{t=1}^{T} t \cdot \mathbb{E}\big[f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big] \leq \frac{TB^2}{\mu} + \frac{\mu}{4}\Big[0 - T(T+1)\mathbb{E}\big[\,\|\mathbf{x}_T - \mathbf{x}^\star\|^2\,\big]\Big] \leq \frac{TB^2}{\mu}.$$

Since

$$\frac{2}{T(T+1)}\sum_{t=1}^{T} t = 1,$$

Jensen's inequality (Lemma 1.5) yields

$$f\left(\frac{2}{T(T+1)}\sum_{t=1}^{T} t \cdot \mathbf{x}_t\right) - f(\mathbf{x}^\star) \leq \frac{2}{T(T+1)}\sum_{t=1}^{T} t(f(\mathbf{x}_t) - f(\mathbf{x}^*)).$$

This in turn implies

$$\mathbb{E}\Big[f\left(\frac{2}{T(T+1)}\sum_{t=1}^{T} t \cdot \mathbf{x}_t\right) - f(\mathbf{x}^\star)\Big] \leq \frac{2B^2}{\mu(T+1)}\,.$$

□

Stochastic Subgradient Descent. Again as a corollary, we have the same convergence rate for the case of stochastic subgradient descent (5.7) here as well, by using the subgradient property (4.1) at the beginning of the proof in (5.8), where convexity was applied.

5.2.3 Mini-batch variants

Instead of using a single element f_i of our sum objective (5.1) to form a stochastic gradient $\mathbf{g}_t = \nabla f_i(\mathbf{x}_t)$, another variant is to use an average of several of them:

$$\tilde{\mathbf{g}}_t := \frac{1}{m}\sum_{j=1}^{m} \mathbf{g}_t^j. \tag{5.9}$$

where $\mathbf{g}_t^j = \nabla f_{i_j}(\mathbf{x}_t)$ for an index i_j. The set of the (distinct) i_j indices is called a mini-batch, and m is the mini batch size.

Using the step direction $\tilde{\mathbf{g}}_t$ defines mini-batch SGD. For $m = 1$, we recover SGD as originally defined, while for $m = n$ we recover full gradient descent.

Mini-batch SGD can be advantageous in several applications. For example, parallelization over up to m processors will easily give a speed-up for the gradient computation, which is typically the main cost of running SGD. Here, parallelization exploits the fact that all $\mathbf{g}_t^j$ are defined at the same iterate $\mathbf{x}_t$ and can therefore be computed independently.

Taking an average of many independent random variables reduces the variance. In the context of mini-batch SGD, we obtain that for larger size of the mini-batch m our estimate $\tilde{\mathbf{g}}_t$ will be closer to the true gradient, in expectation:

$$\begin{aligned}\mathbb{E}\left[\left\|\tilde{\mathbf{g}}_t - \nabla f(\mathbf{x}_t)\right\|^2\right] =& \mathbb{E}\left[\left\|\frac{1}{m}\sum_{j=1}^{m}\mathbf{g}_t^j - \nabla f(\mathbf{x}_t)\right\|^2\right] \\ =& \frac{1}{m}\mathbb{E}\left[\|\mathbf{g}_t^1 - \nabla f(\mathbf{x}_t)\|^2\right] \\ =& \frac{1}{m}\mathbb{E}\left[\|\mathbf{g}_t^1\|^2\right] - \frac{1}{m}\|\nabla f(\mathbf{x}_t)\|^2 \le \frac{B^2}{m}\ .\end{aligned}$$

Using a modification of the above analysis, it is possible to use this property to relate the above convergence rate of SGD to the rate of full gradient descent.

Exercise 25. *Let X, Y be two random variables over a finite probability space $(\Omega, \mathbb{P})$; this avoids subtleties in defining conditional probabilities and expectations; and it covers the random variables occurring in SGD, since in each step, we are randomly choosing among a finite set of n indices.*

The conditional expectation of Y given X *is the random variable* $\mathbb{E}\big[Y\big|X\big]$*, defined by*

$$\mathbb{E}\big[Y\big|X\big](x) := \mathbb{E}\big[Y\big|X = x\big],$$

where $X = x$ *is shorthand for the event* $\{\omega \in \Omega : X(\omega) = x\}$*.*

Hence, the domain of $\mathbb{E}\big[Y\big|X\big]$ *is* $X(\Omega)$ *(the image of* X*), and the probability of* $x \in X(\Omega)$ *is the probability of the event* $X = x$*, i.e.* $\mathbb{P}\big[X = x\big]$*.*

Also recall that

$$\mathbb{E}\big[Y\big|X = x\big] := \sum_{y \in Y(\Omega)} y \cdot \mathbb{P}\big[Y = y|X = x\big].$$

Finally, for two events A *and* B*, the conditional probability* $\mathbb{P}\big[A\big|B\big]$ *is defined as*

$$\mathbb{P}\big[A\big|B\big] := \frac{\mathbb{P}\big[A \cap B\big]}{\mathbb{P}\big[B\big]},$$

if $\mathbb{P}(B) \neq 0$*, and* 0 *otherwise. The equation*

$$\mathbb{P}\big[A\big|B\big]\mathbb{P}\big[B\big] = \mathbb{P}\big[A \cap B\big]$$

always holds.

Prove the following statements.

(i) Let X *be a random variable,* x *in the image of* X*. For random variables* $Y_1, \ldots, Y_m$ *and real numbers* $\lambda_1, \ldots, \lambda_m$*,*

$$\sum_{i=1}^{m} \lambda_i \mathbb{E}\big[Y_i\big|X\big] = \mathbb{E}\big[\sum_{i=1}^{m} \lambda_i Y_i\big|X\big]$$

(ii) Tower rule:

$$\mathbb{E}\big[\mathbb{E}\big[Y\big|X\big]\big] = \mathbb{E}\big[Y\big].$$

Solution: (i) For $x \in X(\Omega)$ we use

$$\begin{aligned}
\mathbb{P}(X = x)\mathbb{E}\big[Y\big|X = x\big] &= \sum_{y \in Y(\Omega)} y \cdot \mathbb{P}\big[Y = y, X = x\big] \\
&= \sum_{y \in Y(\Omega)} y \sum_{\omega \in \Omega : X(\omega) = x, Y(\omega) = y} \mathbb{P}(\omega) \\
&= \sum_{y \in Y(\Omega)} \sum_{\omega \in \Omega : X(\omega) = x, Y(\omega) = y} Y(\omega)\mathbb{P}(\omega) \\
&= \sum_{\omega \in \Omega : X(\omega) = x} Y(\omega)\mathbb{P}(\omega).
\end{aligned}$$

Now we compute

$$\begin{aligned}
\mathbb{P}(X=x)\sum_{i=1}^m \lambda_i \mathbb{E}\big[Y_i\big|X\big](x) &= \sum_{i=1}^m \lambda_i \mathbb{P}(X=x)\mathbb{E}\big[Y_i\big|X=x\big] \\
&= \sum_{i=1}^m \lambda_i \sum_{\omega\in\Omega:X(\omega)=x} Y_i(\omega)\mathbb{P}(\omega) \\
&= \sum_{\omega\in\Omega:X(\omega)=x}\sum_{i=1}^m \lambda_i Y_i(\omega)\mathbb{P}(\omega) \\
&= \mathbb{P}(X=x)\mathbb{E}\big[\sum_{i=1}^m \lambda_i Y_i\big|X\big](x).
\end{aligned}$$

If $\mathbb{P}(X=x)=0$, we have $\sum_{i=1}^m \lambda_i \mathbb{E}\big[Y_i\big|X\big](x) = \mathbb{E}\big[\sum_{i=1}^m \lambda_i Y_i\big|X\big](x)$ by definition, otherwise, the desired statement follows after dividing by $\mathbb{P}(X=x)=0$.

For (ii), we compute

$$\begin{aligned}
\mathbb{E}\big[\mathbb{E}\big[Y\big|X\big]\big] &= \sum_{x\in X(\Omega)} \mathbb{E}\big[Y\big|X\big](x)\mathbb{P}(X=x) \\
&= \sum_{x\in X(\Omega)} \mathbb{E}\big[Y\big|X=x\big]\mathbb{P}(X=x) \\
&= \sum_{x\in X(\Omega)}\sum_{y\in Y(\Omega)} y\cdot\mathbb{P}\big[Y=y|X=x\big]\mathbb{P}(X=x) \\
&= \sum_{y\in Y(\Omega)} y \sum_{x\in X(\Omega)} \mathbb{P}\big[Y=y|X=x\big]\mathbb{P}(X=x) \\
&= \sum_{y\in Y(\Omega)} y\cdot\mathbb{P}\big[Y=y\big] \\
&= E\big[Y\big].
\end{aligned}$$

Bibliography

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004. https://web.stanford.edu/~boyd/cvxbook/.

[Dav59] William C. Davidon. Variable metric method for minimization. Technical Report ANL-5990, AEC Research and Development, 1959.

[Dav91] William C. Davidon. Variable metric method for minimization. *SIAM J. Optimization*, 1(1):1–17, 1991.

[DSSSC08] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the 1-ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*, pages 272–279, 07 2008.

[Gol70] D. Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24(109):23–26, 1970.

[Gre70] J. Greenstadt. Variations on variable-metric methods. *Mathematics of Computation*, 24(109):1–22, 1970.

[Noc80] J. Nocedal. Updating quasi-newton matrices with limited storage. *Mathematics of Computation*, 35(151):773–782, 1980.

[NP06] Yurii Nesterov and B.T. Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, Aug 2006.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the LASSO. *J. R. Statist. Soc. B*, 58(1):267–288, 1996.

[Vis14] Nisheeth Vishnoi. Lecture notes on fundamentals of convex optimization, 2014. https://tcs.epfl.ch/files/content/sites/tcs/files/Lec3-Fall14-Web.pdf.

[Zim16] Judith Zimmermann. *Information Processing for Effective and Stable Admission*. PhD thesis, ETH Zurich, 2016. .