1 Outline

In this lecture, we study

- Variance-reduced (VR) stochastic methods,
- Proximal gradient descent.

2 Variance-reduced (VR) stochastic methods

Although the mini-batch SGD works with a reduced variance, it still cannot exploit the self-tuning property of smooth functions. In this section, we present another variant of SGD that achieves both variance reduction and improvement for smooth functions.

We focus on the following set up. We consider

$$\text{minimize}_{x\in\mathbb{R}^d} \quad f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

which is called the *finite-sum problem*. In stochastic optimization, we had the objective of

$$\mathbb{E}\left[f(x,\xi)\right].$$

Sampling n random vectors $\xi_1,\dots,\xi_n$, we obtain n sampled functions $f(x,\xi_1),\dots,f(x,\xi_n)$. Moreover,

$$\frac{1}{n}\sum_{i=1}^{n} f(x,\xi_i)$$

is an estimator of the original objective function. Taking $f_i(x) = f(x,\xi_i)$, we get the above optimization problem. Hence, in the context of stochastic optimization, the problem is often called the *empirical risk minimization (ERM)* and the *sample average approximation (SAA)*.

When it comes to defining the finite-sum problem for ERM (and SAA), we typically consider convex loss functions. On the other hand, the finite-sum model is also being used for *deep learning*, for which we use non-convex loss functions. Hereinafter, however, we focus on convex functions.

It is widely known that stochastic gradient descent works well for the finite-sum problem. In the previous section, we learned that taking a mini-batch of stochastic gradients can reduce the variance term. In fact, there are other ways of reducing the variance, and they are often called *variance reduced (VR) stochastic methods*. Among many of these methods, we mention a few below.

- Stochastic average gradient (SAG) [SLRB17].
- SAGA [DBLJ14].
- Stochastic variance reduced gradient (SVRG) [JZ13].

2.1 Stochastic variance reduced gradient (SVRG)

In particular, we introduce SVRG for this lecture. To elaborate, we select an index r from $\{1, \ldots, n\}$ uniformly at random. Then for any two points x and y, consider

$$\hat{g}_x = \nabla f_r(x) - (\nabla f_r(y) - \nabla f(y)).$$

By the random choice of r, it follows that

$$\begin{aligned}\mathbb{E}\left[\hat{g}_x\right] &= \mathbb{E}\left[\nabla f_r(x)\right] - (\mathbb{E}\left[\nabla f_r(y)\right] - \nabla f(y)) \\ &= \nabla f(x) - (\nabla f(y) - \nabla f(y)) \\ &= \nabla f(x).\end{aligned}$$

In particular, when $y = x^* \in \operatorname{argmin}_{x \in \mathbb{R}^d} f(x)$, we have

$$\hat{g}_x = \nabla f_r(x) - \nabla f_r(x^*).$$

Moreover, we can use

Lemma 16.1. *If $f_1, \ldots, f_n$ are convex and β-smooth in the ℓ_2 norm, then*

$$\mathbb{E}_{r \sim \mathbb{P}}\left[\|\nabla f_r(x) - \nabla f_r(x^*)\|_2^2\right] \leq 2\beta(f(x) - f(x^*))$$

where $\mathbb{P}$ is the uniform distribution over $\{1, \ldots, n\}$ and $x^* \in \operatorname{argmin}_{x \in \mathbb{R}^d} f(x)$.

Proof. Note that

$$g_r(x) = f_r(x) - \left(f_r(x^*) + \nabla f_r(x^*)^\top (x - x^*)\right) \geq 0$$

because f_r is convex. Moreover, f_r is β-smooth, and we have

$$\|\nabla g_r(x) - \nabla g_r(y)\|_2 = \|\nabla f_r(x) - \nabla f_r(x^*) - \nabla f_r(y) + \nabla f_r(x^*)\|_2 = \|\nabla f_r(x) - \nabla f_r(y)\|_2,$$

implying in turn that g_r is β-smooth. Then it follows that

$$g_r\left(x - \frac{1}{\beta}\nabla g_r(x)\right) \leq g_r(x) - \frac{1}{2\beta}\|\nabla g_r(x)\|_2^2.$$

As $g_r \geq 0$, we obtain

$$\|\nabla g_r(x)\|_2^2 \leq 2\beta g_r(x).$$

By the definition of g_r, this is equivalent to the following.

$$\|\nabla f_r(x) - \nabla f_r(x)\|_2 \leq 2\beta\left(f_r(x) - f_r(x^*) - \nabla f_r(x^*)^\top (x - x^*)\right).$$

Taking the expection of each side with respect to r,

$$\begin{aligned}\mathbb{E}\left[\|\nabla f_r(x) - \nabla f_r(x)\|_2\right] &\leq 2\beta\left(\mathbb{E}\left[f_r(x)\right] - \mathbb{E}\left[f_r(x^*)\right] - \mathbb{E}\left[\nabla f_r(x^*)^\top (x - x^*)\right]\right) \\ &= 2\beta\left(f(x) - f(x^*) - \nabla f(x^*)^\top (x - x^*)\right) \\ &= 2\beta\left(f(x) - f(x^*)\right),\end{aligned}$$

as required. □

Lemma 16.1 basically bounds the variance term $\mathbb{E}\left[\|\hat{g}_x\|_2^2\right]$ given by $\hat{g}_x = \nabla f_r(x) - \nabla f_r(x^*)$. Based on this result, we consider the following algorithm.

In the inner loop, we obtain a stochastic estimator of the gradient, $\nabla f_r(y_k)$, as in each iteration of SGD. On the other hand, the outer loop requires computing the exact gradient, $\nabla f(x_t)$.

Algorithm 1 Stochastic variance reduced gradient (SVRG) descent

Initialize $x_1 \in C$.
for $t = 1, \ldots, T$ **do**
  $y_1 = x_t$.
  for $k = 1, \ldots, B$ **do**
    Sample r from $\{1, \ldots, n\}$ uniformly at random.
    Update $y_{k+1} = y_k - \eta(\nabla f_r(y_k) - (\nabla f_r(x_t) - \nabla f(x_t)))$.
  end for
  Update $x_{t+1} = \frac{1}{B}\sum_{k=1}^{B} y_k$.
end for
Return x_{T+1}.

2.2 SVRG analysis

Theorem 16.2. *Assume that $f_1, \ldots, f_n$ are β-smooth and $f = (1/n)\sum_{i=1}^{n} f_i$ is α-strongly convex with respect to the ℓ_2 norm. Setting $\eta = 1/(6\beta)$ and $B = 36\beta/\alpha$, x_{T+1} returned by Algorithm 1 satisfies*

$$\mathbb{E}\left[f(x_{T+1})\right] - f(x^*) \leq \left(\frac{3}{4}\right)^T (f(x_1) - f(x^*))$$

where $x^* \in argmin_{x \in \mathbb{R}^d} f(x)$.

Proof. Let

$$g_k = \nabla f_r(y_k) - \nabla f_r(x_t) + \nabla f(x_t).$$

Note that

$$\begin{aligned} \|y_{k+1} - x^*\|_2^2 &= \|y_k - \eta g_k - x^*\|_2^2 \\ &= \|y_k - x^*\|_2^2 - 2\eta g_k^\top (y_k - x^*) + \eta^2 \|g_k\|_2^2. \end{aligned} \tag{16.1}$$

Let us consider the third term $\eta^2\|g_k\|_2^2$ in the right-hand side of (16.1). Note that

$$\begin{aligned} &\mathbb{E}\left[\|g_k\|_2^2 \mid y_k\right] \\ &= \mathbb{E}\left[\|\nabla f_r(y_k) - \nabla f_r(x_t) + \nabla f(x_t)\|_2^2 \mid y_k\right] \\ &= \mathbb{E}\left[\|\nabla f_r(y_k) - \nabla f_r(x^*) + \nabla f_r(x^*) - \nabla f_r(x_t) + \nabla f(x_t)\|_2^2 \mid y_k\right] \\ &\leq \mathbb{E}\left[2\|\nabla f_r(y_k) - \nabla f_r(x^*)\|_2^2 + 2\| - \nabla f_r(x^*) + \nabla f_r(x_t) - \nabla f(x_t)\|_2^2 \mid y_k\right] \\ &= 2\mathbb{E}\left[\|\nabla f_r(y_k) - \nabla f_r(x^*)\|_2^2 \mid y_k\right] + 2\mathbb{E}\left[\| - \nabla f_r(x^*) + \nabla f_r(x_t) - \nabla f(x_t)\|_2^2 \mid y_k\right] \end{aligned} \tag{16.2}$$

where the inequality is because $\|a - b\|_2^2 \leq 2\|a\|_2^2 + 2\|b\|_2^2$. Moreover, the second term in the

right-hand side of (16.2) can be bounded as follows.

$$
\begin{aligned}
&\mathbb{E}\left[\| - \nabla f_r(x^*) + \nabla f_r(x_t) - \nabla f(x_t)\|_2^2 \mid \ y_k\right] \\
&= \mathbb{E}\left[\| - \nabla f_r(x^*) + \nabla f_r(x_t)\|_2^2 - 2\nabla f(x_t)^\top(\nabla f_r(x_t) - \nabla f_r(x^*)) + \|\nabla f(x_t)\|_2^2 \mid \ y_k\right] \\
&= \mathbb{E}\left[\| - \nabla f_r(x^*) + \nabla f_r(x_t)\|_2^2 \mid \ y_k\right] - 2\nabla f(x_t)^\top \mathbb{E}\left[\nabla f_r(x_t) - \nabla f_r(x^*) \mid \ y_k\right] \\
&\quad + \mathbb{E}\left[\|\nabla f(x_t)\|_2^2 \mid \ y_k\right] \\
&= \mathbb{E}\left[\| - \nabla f_r(x^*) + \nabla f_r(x_t)\|_2^2 \mid \ y_k\right] - 2\nabla f(x_t)^\top(\nabla f(x_t) - \nabla f(x^*)) \\
&\quad + \mathbb{E}\left[\|\nabla f(x_t)\|_2^2 \mid \ y_k\right] \\
&= \mathbb{E}\left[\| - \nabla f_r(x^*) + \nabla f_r(x_t)\|_2^2 \mid \ y_k\right] - 2\nabla f(x_t)^\top \nabla f(x_t) + \mathbb{E}\left[\|\nabla f(x_t)\|_2^2 \mid \ y_k\right] \\
&= \mathbb{E}\left[\| - \nabla f_r(x^*) + \nabla f_r(x_t)\|_2^2 \mid \ y_k\right] - \mathbb{E}\left[\|\nabla f(x_t)\|_2^2 \mid \ y_k\right] \\
&\leq \mathbb{E}\left[\| - \nabla f_r(x^*) + \nabla f_r(x_t)\|_2^2 \mid \ y_k\right].
\end{aligned}
\tag{16.3}
$$

Combining (16.2) and (16.3), it follows that

$$
\begin{aligned}
&\mathbb{E}\left[\|g_k\|_2^2 \mid \ y_k\right] \\
&\leq 2\mathbb{E}\left[\|\nabla f_r(y_k) - \nabla f_r(x^*)\|_2^2 \mid \ y_k\right] + 2\mathbb{E}\left[\| - \nabla f_r(x^*) + \nabla f_r(x_t)\|_2^2 \mid \ y_k\right] \\
&\leq 4\beta(f(y_k) - f(x^*)) + 4\beta(f(x_t) - f(x^*)) \\
&= 4\beta(f(y_k) - f(x^*) + f(x_t) - f(x^*)).
\end{aligned}
\tag{16.4}
$$

Applying the tower rule to (16.4),

$$
\begin{aligned}
\mathbb{E}\left[\|g_k\|_2^2 \mid \ x_t\right] &= \mathbb{E}\left[\mathbb{E}\left[\|g_k\|_2^2 \mid \ y_k\right] \mid \ x_t\right] \\
&\leq \mathbb{E}\left[4\beta(f(y_k) - f(x^*) + f(x_t) - f(x^*)) \mid \ x_t\right] \\
&= 4\beta(\mathbb{E}\left[f(y_k) \mid \ x_t\right] - f(x^*) + f(x_t) - f(x^*)).
\end{aligned}
\tag{16.5}
$$

Next, we consider the term $-2\eta g_k^\top(y_k - x^*)$ in (16.1).

$$
\begin{aligned}
\mathbb{E}\left[-2\eta g_k^\top(y_k - x^*) \mid \ y_k\right] &= -2\eta\mathbb{E}\left[g_k \mid \ y_k\right]^\top(y_k - x^*) \\
&= -2\eta\mathbb{E}\left[\nabla f_r(y_k) - \nabla f_r(x_t) + \nabla f(x_t) \mid \ y_k\right]^\top(y_k - x^*) \\
&= -2\eta\nabla f(y_k)^\top(y_k - x^*) \\
&\leq -2\eta(f(y_k) - f(x^*)).
\end{aligned}
\tag{16.6}
$$

Again, applying the tower rule to (16.6),

$$
\begin{aligned}
\mathbb{E}\left[-2\eta g_k^\top(y_k - x^*) \mid \ x_t\right] &= \mathbb{E}\left[\mathbb{E}\left[-2\eta g_k^\top(y_k - x^*) \mid \ y_k\right] \mid \ x_t\right] \\
&\leq \mathbb{E}\left[-2\eta(f(y_k) - f(x^*)) \mid \ x_t\right] \\
&= -2\eta(\mathbb{E}\left[f(y_k) \mid \ x_t\right] - f(x^*))
\end{aligned}
\tag{16.7}
$$

Combining (16.1), (16.5), and (16.7), we obtain

$$
\begin{aligned}
\mathbb{E}\left[\|y_{k+1} - x^*\|_2^2 \mid \ x_t\right] &\leq \mathbb{E}\left[\|y_k - x^*\|_2^2 \mid \ x_t\right] - 2\eta(\mathbb{E}\left[f(y_k) \mid \ x_t\right] - f(x^*)) \\
&\quad + 4\eta^2\beta(\mathbb{E}\left[f(y_k) \mid \ x_t\right] - f(x^*) + f(x_t) - f(x^*)) \\
&= \mathbb{E}\left[\|y_k - x^*\|_2^2 \mid \ x_t\right] - 2\eta(1 - 2\eta\beta)(\mathbb{E}\left[f(y_k) \mid \ x_t\right] - f(x^*)) \\
&\quad + 4\eta^2\beta(f(x_t) - f(x^*))
\end{aligned}
\tag{16.8}
$$

Summing (16.8) over $k = 1, \ldots, B$, we obtain

$$\begin{aligned} 2\eta(1-2\eta\beta)\sum_{k=1}^{B}(\mathbb{E}\left[f(y_k) \mid \ x_t\right] - f(x^*)) &\leq \mathbb{E}\left[\|y_1 - x^*\|_2^2 \mid \ x_t\right] - \mathbb{E}\left[\|y_{B+1} - x^*\|_2^2 \mid \ x_t\right] \\ &\quad + 4\eta^2\beta B(f(x_t) - f(x^*)) \\ &\leq \|x_t - x^*\|_2^2 + 4\eta^2\beta B(f(x_t) - f(x^*)) \\ &\leq \left(\frac{2}{\alpha} + 4\eta^2\beta B\right)(f(x_t) - f(x^*)). \end{aligned} \tag{16.9}$$

Dividing each side of (16.9) by B,

$$\begin{aligned} 2\eta(1-2\eta\beta)(\mathbb{E}\left[f\left(x_{t+1}\right) \mid \ x_t\right] - f(x^*)) &= 2\eta(1-2\eta\beta)(\mathbb{E}\left[f\left(\frac{1}{B}\sum_{k=1}^{B} y_k\right) \mid \ x_t\right] - f(x^*)) \\ &\leq 2\eta(1-2\eta\beta)\frac{1}{B}\sum_{k=1}^{B}(\mathbb{E}\left[f\left(y_k\right) \mid \ x_t\right] - f(x^*)) \\ &\leq \left(\frac{2}{\alpha B} + 4\eta^2\beta\right)(f(x_t) - f(x^*)). \end{aligned} \tag{16.10}$$

Remember that

$$\eta = \frac{1}{6\beta}, \quad B = \frac{36\beta}{\alpha}.$$

Then it follows from (16.10) that

$$\begin{aligned} \mathbb{E}\left[f\left(x_{t+1}\right) \mid \ x_t\right] - f(x^*) &\leq \frac{1}{2\eta(1-2\eta\beta)}\left(\frac{2}{\alpha B} + 4\eta^2\beta\right)(f(x_t) - f(x^*)) \\ &= \frac{3\beta}{1-1/3}\left(\frac{1}{18\beta} + \frac{1}{9\beta}\right)(f(x_t) - f(x^*)) \\ &= \frac{3}{4}(f(x_t) - f(x^*)). \end{aligned} \tag{16.11}$$

Applying the tower rule to (16.11),

$$\begin{aligned} \mathbb{E}\left[f\left(x_{t+1}\right)\right] - f(x^*) &\leq \frac{3}{4}(\mathbb{E}\left[f(x_t)\right] - f(x^*)) \\ &\leq \left(\frac{3}{4}\right)^t (f(x_1) - f(x^*)), \end{aligned} \tag{16.12}$$

as required. □

3 Proximal gradient descent

Recall the formulation of LASSO, given by

$$\min_{\beta} \quad \frac{1}{n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1.$$

Here, the objective function is non-differentiable because of the ℓ_1-regularization term $\lambda\|\beta\|_1$, and therefore, it is non-smooth. On the other hand, the objective is convex, and we have a characterization of the subdifferential of $\|\beta\|_1$, so we can simply apply the subgradient method. To bound the additive error by ϵ, the subgradient method requires $O(1/\epsilon^2)$ iterations.

If you take a closer look at the objective, it consists of two part. One part is smooth, and the other part is something whose subdifferential is well understood. Can we use this structure to obtain a better algorithm? The main subject of this section is developing an algorithm that converges to an ϵ-approximate solution after $O(1/\epsilon)$ iterations.

3.1 Projection and proximal operator

We studied the projected gradient descent method, where at each step, we take a projection to the constraint set. When the constraint set is given by C, the projection operator is given by

$$\mathrm{Proj}_C(x) = \operatorname*{argmin}_{u \in C} \frac{1}{2}\|u - x\|_2^2 = \operatorname*{argmin}_{u \in \mathbb{R}^d} \left\{ I_C(u) + \frac{1}{2}\|u - x\|_2^2 \right\}$$

where $I_C(u)$ is the indicator function of C. This definition is proper as there is a unique minimizer for the optimization problem. Hence, the projection operator is defined by the indicator function and the proximity term $(1/2)\|u - x\|_2^2$. The proximal operator is a generalization of the projection operator replacing the indicator function by other general functions.

The proximal operator with respect to a convex function h is defined as follows.

$$\mathrm{Prox}_h(x) = \operatorname*{argmin}_{u \in \mathbb{R}^d} \left\{ h(u) + \frac{1}{2}\|u - x\|_2^2 \right\}.$$

Again the definition is proper because the objective of the optimization problem is strongly convex. Hence, for any $\eta > 0$,

$$\mathrm{Prox}_{\eta h}(x) = \operatorname*{argmin}_{u \in \mathbb{R}^d} \left\{ h(u) + \frac{1}{2\eta}\|u - x\|_2^2 \right\}.$$

As projected gradient descent proceeds with the update rule

$$x_{t+1} = \mathrm{Proj}_C \left\{ x_t - \eta \nabla f(x_t) \right\},$$

we can defined the proximal gradient method with the update rule

$$x_{t+1} = \mathrm{Prox}_{\eta h}(x_t - \eta \nabla f(x_t)).$$

In particular, when we take the indicator function I_C for h, the proximal gradient method reduces to the projeced gradient descent method.

References

[DBLJ14] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. 2

[JZ13] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. 2

[SLRB17] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1):83–112, 2017. 2