**Lecture 15**



## 15.1 Smooth Stochastic Optimization

### 15.1.1 Smooth Problems

We consider unconstrained minimization problem,

$$\min_{x \in \mathbb{R}^n} f(x),$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex, and is gradient is Lipschitz coninuous:

$$\|\nabla f(y) - \nabla f(x)\| \quad \leq \quad L\|y - x\|, \qquad x, y \in \mathbb{R}^n.$$

Let us study the performance of the stochastic gradient method on this class of problems.

*Which rate we can expect?*

For simplicity, we consider unconstrained optimization, so $\nabla f(x^\star) = 0$.

**Direct Approach.** Notice that

$$\|\nabla f(y)\| \quad = \quad \|\nabla f(y) - \nabla f(x^\star)\| \quad \leq \quad L\|y - x^\star\| \quad \leq \quad LD \quad =: \quad M,$$

where $D$ is a distance $D \geq \|x_0 - x^\star\|$. If we can estimate $D$ somehow, we can introduce a ball $Q := \{x : \|x - x_0\| \leq D\}$ and apply the method from the previous lecture, which will give us the rate

$$\mathbb{E}[f(\bar{x}_K) - f^\star] \quad \leq \quad \tfrac{LD^2}{\sqrt{k}} + \tfrac{\sigma D}{\sqrt{k}}, \tag{15.1}$$

where $\sigma > 0$ is the uniform bound on the variance of stochastic gradients.

It appears that the we cannot improve the last "variance term" in (15.1) for the general stochastic optimization problems. However, the first term can be improved. For the basic stochastic gradient method we can ensure the rate of:

$$\mathbb{E}[f(\bar{x}_K) - f^\star] \quad = \quad O\Big(\tfrac{LD^2}{k} + \tfrac{\sigma D}{\sqrt{k}}\Big), \tag{15.2}$$

and, for the accelerated stochastic gradient method, we can have the optimal rate:

$$\mathbb{E}[f(\bar{x}_K) - f^\star] \quad = \quad O\Big(\tfrac{LD^2}{k^2} + \tfrac{\sigma D}{\sqrt{k}}\Big). \tag{15.3}$$

In this note, we establish (15.2).

**Useful Inequality** Let us recall the following useful inequality, which is a consequence of convexity and smoothness (see Theorem 5.2.1 in Lecture 5):

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle \quad \geq \quad \tfrac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2, \qquad \forall x, y \in \mathbb{R}^n. \tag{15.4}$$

**Main Lemma.** For simplicity, we analyze the method with a constant stepsize $\eta > 0$, while an employment of adaptive stepsizes, like in the previous lecture, is also possible.

Thus, we perform iterations, starting from some $x_0 \in \mathbb{R}^n$:

$$x_{k+1} \;=\; x_k - \eta g_k, \qquad k \geq 0,$$

where $g_k \in \mathbb{R}^n$ is a stochastic gradient.

Note that

$$\begin{aligned} \tfrac{1}{2}\|x_{k+1} - x^\star\|^2 &= \tfrac{1}{2}\|x_k - x^\star - \eta g_k\|^2 \\ &= \tfrac{1}{2}\|x_k - x^\star\|^2 + \tfrac{\eta^2}{2}\|g_k\|^2 - \gamma\langle g_k, x_k - x^\star\rangle. \end{aligned}$$

Rearranging the terms we get a familiar expression:

$$\tfrac{\gamma^2}{2}\|g_k\|^2 + \tfrac{1}{2}\|x_k - x^\star\|^2 \;=\; \tfrac{1}{2}\|x_{k+1} - x^\star\|^2 + \gamma\langle g_k, x_k - x^\star\rangle.$$

Now, we notice that

$$\mathbb{E}_{\xi_k}\|g_k - \nabla f(x_k)\|^2 \;=\; \mathbb{E}_{\xi_k}\|g_k\|^2 - \|\nabla f(x_k)\|^2,$$

thus

$$\mathbb{E}\|g_k - \nabla f(x_k)\|^2 \;=\; \mathbb{E}\|g_k\|^2 - \mathbb{E}\|\nabla f(x_k)\|^2.$$

At the same time, we have

$$\mathbb{E}_\xi\langle g_k, x_k - x^\star\rangle \;=\; \langle \nabla f(x_k), x_k - x^\star\rangle \;\overset{(15.4)}{\geq}\; f(x_k) - f^\star + \tfrac{1}{2L}\|\nabla f(x_k)\|^2.$$

Thus,

$$\mathbb{E}\langle g_k, x_k - x^\star\rangle \;\geq\; \mathbb{E}[f(x_k) - f^\star] + \tfrac{1}{2L}\mathbb{E}\|\nabla f(x_k)\|^2$$

Combining these observations together, we get the following progress of one step. Denote

$$R_k^2 \;:=\; \mathbb{E}\|x_k - x^\star\|^2.$$

**Lemma 15.1.1.** *We have:*

$$R_k^2 - R_{k+1}^2 + \tfrac{\gamma^2}{2}E\|\nabla f(x_k)\|^2 + \tfrac{\gamma^2}{2}\mathbb{E}\|\nabla f(x_k) - g_k\|^2 \;\geq\; \gamma\mathbb{E}[f(x_k) - f^\star] + \tfrac{\gamma}{2L}\|\nabla f(x_k)\|^2.$$

**Corollary 15.1.2.** *Consequently, for $\gamma \leq \frac{1}{L}$ we have:*

$$R_k^2 - R_{k+1}^2 + \tfrac{\gamma^2}{2}\mathbb{E}\|\nabla f(x_k) - g_k\|^2 \;\geq\; \gamma\mathbb{E}[f(x_k) - f^\star].$$

### 15.1.2 Convergence Rate

We can bound the variance of the gradients at iteration by $\sigma^2$. We obtain the following inequality,

$$\mathbb{E}[f(x_k) - f^\star] \;\leq\; \tfrac{1}{2\gamma}R_k^2 - \tfrac{1}{2\gamma}R_{k+1}^2 + \tfrac{\gamma}{2}\sigma^2.$$

Telescoping and using Jensen's inequality, we have

$$\begin{aligned} \gamma k\mathbb{E}[f(\bar{x}_k) - f^\star] &\leq \gamma\mathbb{E}[\textstyle\sum_{i=0}^{k-1}(f(x_i) - f^\star)] \;\leq\; \tfrac{1}{2\gamma}(R_0^2 - R_k^2) + \tfrac{\gamma}{2}\sigma^2 k \\ &\leq \tfrac{1}{2\gamma}R_0^2 + \tfrac{\gamma}{2}\sigma^2 k. \end{aligned}$$

Let us minimize the right hand side with respect to $\gamma$. We get the optimal choice

$$\gamma^\star \;=\; \tfrac{R_0}{\sigma\sqrt{k}}.$$

Taking into account the other condition, $\gamma \le \frac{1}{L}$, we set $\boxed{\gamma := \min\{\frac{1}{L}, \frac{R_0}{\sigma\sqrt{k}}\}}$

For that choice we get

$$\mathbb{E}[f(\bar{x}_k) - f^\star] \;\le\; \max\{\tfrac{LR_0^2}{2k}, \tfrac{R_0\sigma}{2\sqrt{k}}\} + \tfrac{R_0\sigma}{\sqrt{k}} \;=\; O\Big(\tfrac{LR_0^2}{k} + \tfrac{\sigma R}{\sqrt{k}}\Big).$$

The oracle complexity is

$$O\Big(\tfrac{LR_0^2}{\varepsilon} + \big[\tfrac{\sigma R_0}{\varepsilon}\big]^2\Big).$$

## 15.2 Variance Reduction

### 15.2.1 Minibatching

Instead of one sample, we use $m \ge 1$ samples:

$$g_k \;:=\; \tfrac{1}{m}\sum_{i=1}^m g(x_k, \xi_{k,i})$$

It is clear that $\mathbb{E}g_k = \nabla f(x_k)$. What will be the variance of $g_k$?

Denote $\delta_i := g(x_k, \xi_{k,i}) - \nabla f(x_k)$. Hence, $\mathbb{E}\delta_i = 0$. Observe that

$$\begin{aligned}
\mathbb{E}\Big[\big\|g_k - \nabla f(x_k)\big\|^2\Big] \;&=\; \mathbb{E}\Big[\big\|\tfrac{1}{m}\sum_{i=1}^m \delta_i\big\|^2\Big] \;=\; \mathbb{E}\Big[\langle \tfrac{1}{m}\sum_{i=1}^m \delta_i, \tfrac{1}{m}\sum_{j=1}^m \delta_j\rangle\Big] \\
&=\; \tfrac{1}{m^2}\mathbb{E}\sum_{i=1}^m \|\delta_i\|^2 + \tfrac{2}{m^2}\sum_{i<j}\mathbb{E}\langle\delta_i, \delta_j\rangle \\
&=\; \tfrac{1}{m^2}\mathbb{E}\sum_{i=1}^m \|\delta_i\|^2 + \;\le\; \tfrac{1}{m}\sigma^2.
\end{aligned}$$

Therefore, we can reduce the value $\sigma$ by $\sqrt{m}$, when using minibatch of size $m \ge 1$.

What is the total complexity of the method in terms of the *total sampled gradients*, where at each iteration sample $m$ gradients. A natural choice would be to make the two complexity terms equal:

$$\tfrac{LR_0^2}{\varepsilon} \;\stackrel{?}{=}\; \tfrac{1}{m}\big[\tfrac{\sigma R_0}{\varepsilon}\big]^2,$$

which leads to the choice:

$$m \;:=\; 1 + \tfrac{\varepsilon}{LR_0^2}\cdot\big[\tfrac{\sigma R_0}{\varepsilon}\big]^2 \;=\; 1 + \tfrac{\sigma^2}{\varepsilon L}.$$

Performing this amount of samples each iteration, we ensure the same rate as for the deterministic method. At the same time, the total number of samples over all iterations is:

$$m\cdot O(\tfrac{LR_0^2}{\varepsilon}) \;=\; \Big(1 + \tfrac{\sigma^2}{\varepsilon L}\Big)\cdot O(\tfrac{LR_0^2}{\varepsilon}) \;=\; O\Big(\tfrac{LR_0^2}{\varepsilon} + \tfrac{\sigma^2 R_0^2}{\varepsilon^2}\Big).$$

So, the total number of samples remains the same. In practice, it is always useful use a small mini-batch ($m \approx 100$ or more).