Lecture 14

14.1 Stochastic Optimization Problem

14.1.1 Motivation

We discussed stochastic optimization previously in Lecture 4. We showed that for smooth (but possibly non-convex) problems, stochastic gradient method of the form

$$x_{k+1} \quad = \quad x_k - \tfrac{1}{M} g_k, \tag{14.1}$$

where $g_k = g(\xi_k, x_k)$ is a stochastic unbiased estimate of the gradient $\nabla f(x_k)$ of our objective, will converge to a stationary point as soon as the step size parameter $M > 0$ is sufficiently small. Namely, we set

$$M \quad := \quad L \cdot \max\Big\{1, \tfrac{2\sigma^2}{\varepsilon^2}\Big\}, \tag{14.2}$$

where L is the Lipschitz constant of the gradient, σ^2 is bound for the variance of g_k and $\varepsilon > 0$ is the target accuracy in terms of the gradient norm. Then, in order to reach a random point $E\big[\|\nabla f(\bar{x})\|\big] \leq \varepsilon$, we showed (see Theorem 4.2.6) that it is enough to perform

$$K \quad = \quad O\Big(L(f(x_0) - f^\star) \cdot \big[\tfrac{1}{\varepsilon^2} + \tfrac{\sigma^2}{\varepsilon^4}\big]\Big)$$

iterations (14.1). We have two goals now:

- **Study** stochastic methods for *convex problems*: having explored various methods and proof techniques in convex optimization, it is natural to expect that convexity will also benefit stochastic problems.
- **Develop** an *adaptive stepsize* rule: the constant rule in (14.2) depends on two parameters, L and σ, which are usually unknown in practice. Unlike determenistic methods, we cannot use an adaptive *line-search* to ensure progress of every step. Furthermore, a constant stepsize (14.2) seems too conservative, as it prevents the method from improving convergence when *local values* of L and σ are small.

It appears that ideas from non-smooth convex optimization work nicely in the stochastic case. Informally speaking, stochasticity can be treated as a form of "non-smoothness" in the problem. Thus, the subgradient methods developed initially for non-smooth convex problems are simple to generalize to stochastic settings.

14.1.2 Problem Formulation

We consider the following convex optimization problem,

$$\min_{x \in Q} f(x),$$

where $Q \subseteq \mathbb{R}^n$ is a bounded convex set. The boundedness will be a crucial assumption for the analysis of our method. However, if the initial problem is over unbounded set, we can always introduce an additional simple constrain of the large enough Euclidean ball around the origin.

We denote the diameter of Q in the Euclidean norm by:

$$D \;:=\; \max_{x,y \in Q} \|x - y\|.$$

We assume that $f : Q \to \mathbb{R}$ is convex, possibly non-differentiable, and denote by M its Lipschitz constant.

Now we only have access to the *stochastic first-order oracle*, for any $x \in Q$ we assume we can sample a random variable ξ and compute a vector:

$$g(x;\xi) \;\in\; \mathbb{R}^n,$$

that is a stochastic substitute for a subgradient. We assume that

- This is unbiased estimator of a subgradient:

$$\mathbb{E}_\xi[g(x,\xi)] \;=\; f'(x) \qquad \text{for some} \qquad f'(x) \;\in\; \partial f(x).$$

- It has bounded variance:

$$\mathbb{E}_\xi\big[\|g(x;\xi) - \nabla f(x)\|^2\big] \;\le\; \sigma^2,$$

for $\sigma > 0$, which is a parameter of our problem. As a consequence, we have (formula (4.11) in Lecture 4):

$$E_\xi\big[\|g(x,\xi)\|^2\big] \;\le\; \sigma^2 + \|f'(x)\|^2 \;\le\; \sigma^2 + M^2.$$

14.1.3 Stochastic Subgradient Method

For the deterministic problem, we have analyzed the following variant of the subgradient method with normalized stepsizes:

$$x_{k+1} \;=\; \pi_Q\Big(x_k - \tfrac{\gamma}{\|f'(x_k)\|} f'(x_k)\Big), \qquad \text{with} \qquad \gamma \;:=\; \tfrac{D}{\sqrt{K+1}}, \tag{14.3}$$

where $K \ge 1$ is a fixed number of iterations. This method gave us the optimal complexity for the non-smooth convex optimization.

In stochastic setting, it is a natural idea to replace $f'(x_k) \mapsto g_k$ in formula (14.3). We obtain a step in the following normalized stochastic direction:

$$x_{k+1} \;=\; \pi_Q\Big(x_k - \tfrac{\gamma}{\|g_k\|} g_k\Big),$$

which has the following drawbacks:

- It is more difficult to analyze the normalized random variable $\frac{g_k}{\|g_k\|}$ than the deterministic update in (14.3). We can also use instead the following update, which is easier and resembles our primary approach (14.1):

$$x_{k+1} \quad = \quad \pi_Q\Big(x_k - \eta g_k\Big), \tag{14.4}$$

- However, in both cases, the method becomes too chaotic, performing a lot of random fluctuations and we have to tune parameter $\eta > 0$ to be really small to ensure convergence. For update of (14.4) to work, we have to know variance σ to choose the step-size, as in (14.2).
- In the stepsize formula for (14.3), even in the deterministic case, we still need to fix the number of iterations $K > 0$ in advance and use it for the stepsize. If, after K iterations, we want to continue running our method, proposed γ will not work anymore.

14.2 Adaptive Stepsizes

We consider is a more advanced *adaptive* stepsize rule that solves all these problems at once. Namely, we perform the following algorithm.

Algorithm 14.1: *Stochastic Subgradient Method with Adaptive Stepsizes.*

Initialization: $x_0 \in Q$ and $S_0 = 0$.

For $k = 0 \dots K-1$ **iterate:**

1. Sample ξ_k and compute stochastic gradient $g_k := g(x_k, \xi_k)$
2. Update $S_{k+1} := S_k + \|g_k\|^2$ and set $\beta_k := \frac{\sqrt{S_{k+1}}}{D}$.
3. Perform the step:

$$x_{k+1} \quad = \quad \arg\min_{y \in Q}\Big[\langle g_k, y - x_k\rangle + \tfrac{\beta_k}{2}\|y - x_k\|^2\Big]$$

Return $\bar{x}_K = \frac{1}{K}\sum\limits_{i=1}^{K} x_i$.

Remark 14.2.1. One step of this method is given by:

$$x_{k+1} \quad = \quad \pi_Q\Big(x_k - \tfrac{1}{\beta_k} g_k\Big), \qquad \text{where} \qquad \beta_k \quad := \quad \tfrac{1}{D}\sqrt{\|g_0\|^2 + \ldots + \|g_k\|^2}, \tag{14.5}$$

while for deterministic normalized subgradient method (14.3) we had:

$$x_{k+1} \quad = \quad \pi_Q\Big(x_k - \tfrac{1}{\alpha_k} f'(x_k)\Big), \qquad \text{where} \qquad \alpha_k \quad := \quad \tfrac{\|f'(x_k)\|\sqrt{K+1}}{D}.$$

Therefore, new more advanced formula (14.5) replaces the subgradient norm at the current point by the average of all (stochastic) subgradient norms:

$$\|f'(x_k)\|\sqrt{k+1} \quad \approx \quad \sqrt{S_{k+1}} \quad = \quad \sqrt{\|g_0\|^2 + \ldots \|g_k\|^2}.$$

Remark 14.2.2. Steps of the method are *independent* of a fixed number of iterations $K \geq 1$. We use K only as a stopping condition and to form the output. If after K iterations we decide to continue running the method, we can easily do that without any restarts.

14.2.1 Analysis

We want to show the convergence for this algorithm. Let us start our analyses the same way as we did for the subgradient method. We denote our model by

$$m_k(y) \;:=\; \langle g_k, y - x_k\rangle + \tfrac{\beta_k}{2}\|y - x_k\|^2,$$

and x_{k+1} is defined as the minimizer of this model over Q. Due to strong convexity of the model, we have that

$$m_k(y) \;\geq\; m_k(x_{k+1}) + \tfrac{\beta_k}{2}\|y - x_{k+1}\|^2.$$

We also notice that

$$\begin{aligned} m_k(x_{k+1}) &= \langle g_k, x_{k+1} - x_k\rangle + \tfrac{\beta_k}{2}\|x_{k+1} - x_k\|^2 \\ &\geq \min_{h\in\mathbb{R}^n}\Big[\langle g_k, h\rangle + \tfrac{\beta_k}{2}\|h\|^2\Big] \;=\; -\tfrac{1}{2\beta_k}\|g_k\|^2. \end{aligned}$$

Hence, we get

$$\tfrac{1}{2\beta_k}\|g_k\|^2 + \tfrac{\beta_k}{2}\|y - x_k\|^2 \;\geq\; \tfrac{\beta_k}{2}\|y - x_{k+1}\|^2 + \langle g_k, x_k - y\rangle.$$

Further, we can substitute $y := x^\star$ (any minimizer for our problem). We get:

$$\tfrac{1}{2\beta_k}\|g_k\|^2 + \tfrac{\beta_k}{2}\|x^\star - x_k\|^2 \;\geq\; \tfrac{\beta_k}{2}\|x^\star - x_{k+1}\|^2 + \langle g_k, x_k - x^\star\rangle. \tag{14.6}$$

Notice that

$$\mathbb{E}_{\xi_k}\big[\langle g_k, x_k - x^\star\rangle\big] \;=\; \langle f'(x_k), x_k - x^\star\rangle \;\geq\; f(x_k) - f^\star.$$

Therefore, the right hand side of (14.6) gives us the progress of the method, and the left hand side contains an "error" of one step: $\frac{1}{2\beta_k}\|g_k\|^2$. However, to make (14.6) telescoping, we want to have β_{k+1} in the right hand side instead of β_k.

We can use the following simple but crucial observation, for $\beta_{k+1} \geq \beta_k$, using the boundedness of the feasible set:

$$\begin{aligned} \tfrac{\beta_{k+1}}{2}\|x^\star - x_{k+1}\|^2 &= \tfrac{\beta_k}{2}\|x^\star - x_{k+1}\|^2 + \tfrac{\beta_{k+1}-\beta_k}{2}\|x^\star - x_{k+1}\|^2 \\ &\leq \tfrac{\beta_k}{2}\|x^\star - x_{k+1}\|^2 + \tfrac{\beta_{k+1}-\beta_k}{2}D^2. \end{aligned}$$

Thus, we have established the following consequence for one step of the method.

Lemma 14.2.3. *Let $\beta_{k+1} \geq \beta_k$. Then,*

$$\tfrac{1}{2\beta_k}\|g_k\|^2 + \tfrac{\beta_{k+1}-\beta_k}{2}D^2 + \tfrac{\beta_k}{2}\|x^\star - x_k\|^2 \;\geq\; \tfrac{\beta_{k+1}}{2}\|x^\star - x_{k+1}\|^2 + \langle g_k, x_k - x^\star\rangle. \tag{14.7}$$

Now, we have two "error terms" in the left hand side of (14.7). For convenience, let us denote $\boxed{\beta_{-1} := 0}$. Then, we notice that, by the definition of β_k, for every $k \geq 0$:

$$\begin{aligned} (\beta_k - \beta_{k-1})D^2 &= (\sqrt{S_{k+1}} - \sqrt{S_k})D \;=\; \tfrac{S_{k+1}-S_k}{\sqrt{S_{k+1}}+\sqrt{S_k}}\cdot D \;=\; \tfrac{\|g_k\|^2}{\sqrt{S_{k+1}}+\sqrt{S_k}}\cdot D \\ &\geq \tfrac{\|g_k\|^2}{\sqrt{S_{k+1}}}\cdot\tfrac{D}{2} \;=\; \tfrac{\|g_k\|^2}{2\beta_k}. \end{aligned}$$

This important observations allows us to simplify the left hand side of (14.7).

Lemma 14.2.4. *Let β_k be chosen as in Algorithm 14.1. Then,*

$$D^2 \cdot \left(\tfrac{\beta_{k+1}-\beta_k}{2} + \beta_k - \beta_{k-1}\right) + \tfrac{\beta_k}{2}\|x^\star - x_k\|^2 \;\; \geq \;\; \tfrac{\beta_{k+1}}{2}\|x^\star - x_k\|^2 + \langle g_k, x_k - x^\star\rangle.$$

Telescoping this inequality for the first $K \geq 0$ iterations, we get

$$\begin{aligned}\sum_{i=0}^{K-1}\langle g_i, x_i - x^\star\rangle &\leq \tfrac{\beta_0}{2}\|x^\star - x_0\|^2 + D^2 \cdot \left(\tfrac{\beta_K-\beta_0}{2} + \beta_{K-1} - \beta_{-1}\right)\\ &\leq \tfrac{3}{2}D^2\beta_K.\end{aligned}$$

Let us take the expectations for the left and the right hand sides. For the left hand side, we have:

$$\begin{aligned}\mathbb{E}\Big[\sum_{i=0}^{K-1}\langle g_i, x_i - x^\star\rangle\Big] &= \mathbb{E}\Big[\sum_{i=0}^{K-1}\langle f'(x_i), x_i - x^\star\rangle\Big] \geq K \cdot \mathbb{E}\Big[\tfrac{1}{K}\sum_{i=0}^{K-1}\big[f(x_i) - f^\star\big]\Big]\\ &\geq K \cdot \mathbb{E}\big[f(\bar{x}_K) - f^\star\big].\end{aligned}$$

For the right hand side, we obtain, using Jensen's inequality for concave function $\sqrt{\cdot}$, that

$$\mathbb{E}\Big[\beta_K\Big] = \tfrac{1}{D}\mathbb{E}\Big[\sqrt{\sum_{i=0}^{K-1}\|g_i\|^2}\Big] \leq \tfrac{1}{D}\sqrt{\sum_{i=0}^{K-1}\mathbb{E}[\|g_i\|^2]} \leq \tfrac{\sqrt{K}\cdot\sqrt{\sigma^2+M^2}}{D} \leq \tfrac{\sqrt{K}(\sigma+M)}{D}.$$

Combining these two bounds together, we have proved the following theorem.

Theorem 14.2.5. *It holds,*

$$\mathbb{E}\Big[f(\bar{x}_K) - f^\star\Big] \leq \tfrac{3D^2\mathbb{E}[\beta_K]}{2K} \leq \tfrac{3(\sigma+M)D}{2\sqrt{K}}.$$

This is the same rate as for the deterministic subgradient method, where we replaced M by $\sigma + M$ in the complexity estimate. Note that this adaptive strategy work for the deterministic method as well ($\sigma = 0$), eliminating the need to fix the number of iterations $K \geq 0$ within the stepsize.

Literature

Our analysis is based on recent work [RJS24], which demonstrates that adaptive stepsizes work universally well for stochastic problems with varying degrees of smoothness, biased oracles, composite, and acceleration methods, while automatically supporting variance reduction. See also the references therein for additional reading.

The adaptive stepsizes presented here are also referred to as *AdaGrad stepsizes*, as they resemble the popular AdaGrad algorithm (Adaptive Subgradient Method) [DHS11] from machine learning.

[DHS11] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[RJS24] Anton Rodomanov, Xiaowen Jiang, and Sebastian Stich. Universality of adagrad stepsizes for stochastic optimization: Inexact oracle, acceleration and variance reduction. *Advances in Neural Information Processing Systems*, 37:26770–26813, 2024.