Lecture 18: Stochastic Optimization

Yudong Chen

1 Setup

The algorithms we've seen so far have access to a first order oracle, which returns the exact (sub)gradient at a given point, plus potentially the function value.

$$x \in \mathcal{X} \longrightarrow \boxed{\begin{array}{c}\text{1st order}\\ \text{oracle}\end{array}} \longrightarrow \begin{array}{c} g_x \in \partial f(x) \\ (\nabla f(x) \text{ if } f \text{ is differentiable}) \\ \text{maybe also } f(x) \end{array}$$

Stochastic optimization: We are given a *noisy* version of the (sub)gradient:

$$x \in \mathcal{X} \longrightarrow \boxed{\begin{array}{c}\text{1st order}\\ \text{stochastic oracle}\end{array}} \longrightarrow \widetilde{g}(x, \xi)$$

Here $\widetilde{g}(x, \xi)$ is a stochastic estimate of some $g_x \in \partial f(x)$, where ξ is a random variable (representing the randomness in the stochastic estimate).

Remark 1. Some models also assume access to stochastic estimates of the function value $f(x)$. We do not need that here.

1.1 Examples

Example 1. $\widetilde{g}(x, \xi) = g_x + \xi$, where ξ is additive noise from, e.g., inaccurate measurements in physical systems. Sometimes, the noise is added intentionally (for privacy).

Example 2. Finite sum minimization:

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

and n is large. We can take $\widetilde{g}(x, \xi) = \nabla f_{\bar{i}}(x)$, where $\bar{i}$ is an integer sampled uniformly at random from $\{1, 2, \ldots, n\}$.

Example 3. Empirical risk minimization (ERM): We want to minimize

$$f(x) = \mathbb{E}_{(x,y) \sim \Pi_{\text{data}}} \left[l(x; a, b) \right],$$

but we do not know how to compute the expectation exactly. Suppose we have collected n data points (a_i, b_i) that come from the distribution Π_{data}. We can consider minimizing the empirical loss

$$f_{\text{emp}}(x) = \frac{1}{n} \sum_{i=1}^{n} l(x; a_i; b_i).$$

When $n \to \infty$, $f_{\text{emp}} \to f$. Here we may view $\widetilde{g}(x, \xi) = \nabla f_{\text{emp}}(x)$ as a noisy estimate of $\nabla f(x)$.

1.2 Assumptions

Consider the problem

$$\min_{x \in \mathcal{X}} f(x). \tag{P}$$

We assume that

- f is convex and M-Lipschitz (w.r.t. $\|\cdot\|_2$).
- $\mathcal{X}$ is closed, convex and nonempty. The projection $P_{\mathcal{X}}$ can be efficiently computed.
- For all $x \in \mathcal{X}$, it holds that

$$\begin{aligned} &\text{(unbiased estimate)} \qquad \mathbb{E}_\xi \left[\widetilde{g}(x,\xi)\right] = g_x \in \partial f(x), \\ &\text{(bounded variance)} \qquad \mathbb{E}_\xi \left[\|\widetilde{g}(x,\xi) - g_x\|_2^2\right] \le \sigma^2 < \infty. \end{aligned}$$

2 Stochastic (projected sub)gradient descent

Consider the following S-PSubGD algorithm:

$$\begin{aligned} x_{k+1} &= \operatorname*{argmin}_{u \in \mathcal{X}} \left\{ a_k \left\langle \widetilde{g}(x_k,\xi_k), u - x_k \right\rangle + \frac{1}{2}\|u - x_k\|_2^2 \right\} \\ &= P_{\mathcal{X}}\left(x_k - a_k \widetilde{g}(x_k,\xi_k)\right), \end{aligned}$$

where $a_k > 0$ is the stepsize to be chosen later.

2.1 Convergence analysis

In the sequel, we assume that $\xi_0, \xi_1, \ldots, \xi_k, \ldots$ are independent and identically distributed (i.i.d.). To avoid cluttered notation, we introduce the shorthands $g_k \equiv g_{x_k}$ (true subgradient) and $\widetilde{g}_k \equiv \widetilde{g}(x_k,\xi_k)$ (noisy subgradient).

As in the previous lecture, we analyze the averaged iterate $x_k^{\text{out}} := \frac{1}{A_k}\sum_{i=0}^k a_i x_i$, where $A_k := \sum_{i=0}^k a_i$, and we use the same U_k, L_k and G_k:

$$\begin{aligned} \text{upper bound:} \qquad & U_k := \frac{1}{A_k}\sum_{i=0}^k a_i f(x_i) \ge f(x_k^{\text{out}}), \\ \text{lower bound:} \qquad & L_k := \frac{1}{A_k}\sum_{i=0}^k a_i f(x_i) + \frac{1}{A_k}\sum_{i=0}^k a_i \left\langle g_i, x^* - x_i \right\rangle \le f(x^*), \\ \text{optimality gap bound:} \qquad & G_k := U_k - L_k = -\frac{1}{A_k}\sum_{i=0}^k a_i \left\langle g_i, x^* - x_i \right\rangle \ge f(x_k^{\text{out}}) - f(x^*). \end{aligned}$$

The analysis is similar to last lecture, except that we need to keep track of the stochastic error $g_k - \widetilde{g}_k$. We have

$$A_0 G_0 = -a_0 \left\langle g_0, x^* - x_0 \right\rangle,$$

and

$$\begin{aligned} A_k G_k - A_{k-1} G_{k-1} &= -a_k \langle g_k, x^* - x_k \rangle \\ &= a_k \langle g_k, x_k - x_{k+1} \rangle + a_k \langle g_k, x_{k+1} - x^* \rangle \\ &= \underbrace{a_k \langle g_k, x_k - x_{k+1} \rangle + a_k \langle \widetilde{g}_k, x_{k+1} - x^* \rangle}_{\text{similar to last lecture}} + \underbrace{a_k \langle g_k - \widetilde{g}_k, x_{k+1} - x^* \rangle}_{\text{additional error term}}. \end{aligned}$$

Note that $x_{k+1} = P_{\mathcal{X}}(x_k - a_k \widetilde{g}_k)$ satisfies the minimum principle:

$$\langle a_k \widetilde{g}_k + x_{k+1} - x_k, x^* - x_{k+1} \rangle \geq 0,$$

hence

$$\begin{aligned} a_k \langle \widetilde{g}_k, x_{k+1} - x^* \rangle &\leq \langle x_{k+1} - x_k, x^* - x_{k+1} \rangle \\ &= \frac{1}{2} \|x_k - x^*\|_2^2 - \frac{1}{2} \|x_{k+1} - x^*\|_2^2 - \frac{1}{2} \|x_k - x_{k+1}\|_2^2. \end{aligned}$$

It follows that

$$\begin{aligned} & A_k G_k - A_{k-1} G_{k-1} \\ \leq & \underbrace{a_k \langle g_k, x_k - x_{k+1} \rangle + \left(\frac{1}{2} \|x_k - x^*\|_2^2 - \frac{1}{2} \|x_{k+1} - x^*\|_2^2 - \frac{1}{2} \|x_k - x_{k+1}\|_2^2 \right)}_{\text{same as last lecture}} + a_k \langle g_k - \widetilde{g}_k, x_{k+1} - x^* \rangle \\ \leq & \underbrace{\frac{a_k^2 M^2}{2} + \frac{1}{2} \|x_k - x^*\|_2^2 - \frac{1}{2} \|x_{k+1} - x^*\|_2^2}_{\text{same as last lecture}} + a_k \langle g_k - \widetilde{g}_k, x_{k+1} - x^* \rangle. \end{aligned}$$

We take expectation of both sides. By the Law of Iterated Expectation,[1] we can write

$$\mathbb{E}[\text{RHS}] = \mathbb{E}\left[\mathbb{E}\left[\text{RHS} \mid \xi_0^{k-1}\right]\right],$$

where $\xi_0^{k-1} := (\xi_0, \ldots, \xi_{k-1})$ denotes all the previous randomness in iterations 0 through $k-1$ (not including ξ_k). Observe that

$$\begin{aligned} & \mathbb{E}\left[a_k \langle g_k - \widetilde{g}_k, x_{k+1} - x^* \rangle \mid \xi_0^{k-1}\right] \\ = & a_k \mathbb{E}\left[\langle g_k - \widetilde{g}_k, x_{k+1} \rangle \mid \xi_0^{k-1}\right] && \mathbb{E}\left[\langle g_k - \widetilde{g}_k, x^* \rangle \mid \xi_0^{k-1}\right] = 0 \\ & && \text{as } \widetilde{g}_k \text{ is unbiased and independent of } x^* \\ = & a_k \mathbb{E}\left[\langle g_k - \widetilde{g}_k, P_{\mathcal{X}}(x_k - a_k \widetilde{g}_k) \rangle \mid \xi_0^{k-1}\right] \\ = & a_k \mathbb{E}\left[\langle g_k - \widetilde{g}_k, P_{\mathcal{X}}(x_k - a_k \widetilde{g}_k) - P_{\mathcal{X}}(x_k - a_k g_k) \rangle \mid \xi_0^{k-1}\right] && \mathbb{E}\left[\langle g_k - \widetilde{g}_k, P_{\mathcal{X}}(x_k - a_k g_k) \rangle \mid \xi_0^{k-1}\right] = 0 \\ & && \text{as } \widetilde{g}_k \text{ is unbiased and independent of } x_k \\ \leq & a_k \mathbb{E}\left[\|g_k - \widetilde{g}_k\|_2 \cdot \|P_{\mathcal{X}}(x_k - a_k \widetilde{g}_k) - P_{\mathcal{X}}(x_k - a_k g_k)\|_2 \mid \xi_0^{k-1}\right] && \text{Cauchy-Schwarz} \\ \leq & a_k \mathbb{E}\left[a_k \|g_k - \widetilde{g}_k\|_2^2 \mid \xi_0^{k-1}\right] && P_{\mathcal{X}} \text{ is nonexpansive} \\ \leq & a_k^2 \sigma^2. && \text{bounded variance assumption} \end{aligned}$$

[1] Also known as Law of Total Expectation, or Tower Rule

It follows that

$$\mathbb{E}\left[A_k G_k - A_{k-1} G_{k-1}\right] \le \mathbb{E}\left[\frac{1}{2}\left\|x_k - x^*\right\|_2^2 - \frac{1}{2}\left\|x_{k+1} - x^*\right\|_2^2\right] + \frac{a_k^2\left(M^2 + 2\sigma^2\right)}{2}.$$

Summing both sides over k and telescoping, we get

$$\begin{aligned}\mathbb{E}\left[f(x_K^{\text{out}}) - f(x^*)\right] &\le \mathbb{E}\left[G_K\right] \\ &\le \frac{\left\|x_0 - x^*\right\|_2^2 + \left(M^2 + 2\sigma^2\right)\sum_{k=0}^K a_k^2}{2A_K}.\end{aligned}$$

The expression on the right-hand side is the same as what we got the last time for projected subgradient descent (PSubGD), except for having $M^2 + 2\sigma^2$ in place of M^2. The rest of the analysis is similar to that for PSubGD:

- Using constant stepsize $a_k = \frac{\|x_0 - x^*\|_2}{\sqrt{M^2+2\sigma^2}\sqrt{K+1}}, \forall k$, we get an $O\left(\frac{1}{\sqrt{K}}\right)$ convergence rate.
- Same discussion about anytime algorithm, unknown/unbounded diameter of $\mathcal{X}$, unknown M, etc.