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.

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

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

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

Here $\widetilde{g}(x, w)$ is a stochastic estimate of some $g_x \in \partial f(x)$, where w 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 it here.

1.1 Examples

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

Example 2. Finite sum minimization: Want to minimize

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

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

More generally, we can take $\widetilde{g}(x, w) = \frac{1}{n}\sum_{i \in S} \nabla f_i(x)$, where S is a random subset of $\{1, \ldots n\}$; here $w = S$ is sometimes called a mini-batch.

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

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

but we do not know how to exactly compute the expectation above. Suppose we have collected n data points (a_i, b_i) that come from the distribution Π_{data}. As an approximation we minimize 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 view $\widetilde{g}(x, w) = \nabla f_{\text{emp}}(x)$ 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. $\|\cdot\|_2$. (f may not be differentiable, but it is subdifferentiable)
- $\mathcal{X}$ is closed, convex and nonempty. The projection $P_{\mathcal{X}}$ can be efficiently computed.
- For all $x \in \mathcal{X}$, it holds that

$$\begin{aligned} &\text{(unbiased estimate)} && \mathbb{E}_w\left[\widetilde{g}(x,w)\right] = g_x \in \partial f(x), \\ &\text{(bounded variance)} && \mathbb{E}_w\left[\|\widetilde{g}(x,w) - g_x\|_2^2\right] \le \sigma^2 < \infty. \end{aligned} \tag{1}$$

2 Stochastic (projected sub)gradient descent

Consider the following S-PSubGD algorithm:

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

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

2.1 Convergence analysis

In the sequel, we assume that $w_0, w_1, \ldots, w_k, \ldots$ are *independent* and identically distributed (i.i.d.).

We introduce the shorthands $g_k \equiv g_{x_k}$ (true subgradient) and $\widetilde{g}_k \equiv \widetilde{g}(x_k, w_k)$ (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:

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

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

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

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

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

hence

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

It follows that

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

Taking expectation of both sides, we get

$$\mathbb{E}\left[A_k G_k - A_{k-1} G_{k-1}\right] \leq \frac{1}{2}\mathbb{E}\left[\left\|x_k - x^*\right\|_2^2 - \left\|x_{k+1} - x^*\right\|_2^2\right] + \frac{a_k^2 M^2}{2} + a_k \mathbb{E}\left[\left\langle g_k - \widetilde{g}_k, x_{k+1} - x^* \right\rangle\right].$$

To compute the last term on the right hand side, we need some basic facts from probability, including the linearity and independence property of expectation (reviewed at the end of this document). Moreover, by the Law of Total Expectation,[1] we can write

$$\mathbb{E}\left[\left\langle g_k - \widetilde{g}_k, x_{k+1} - x^* \right\rangle\right] = \mathbb{E}\left[\mathbb{E}\left[\left\langle g_k - \widetilde{g}_k, x_{k+1} - x^* \right\rangle \mid w_0^{k-1}\right]\right],$$

where $w_0^{k-1} := (w_0, \ldots, w_{k-1})$ denotes all the previous randomness in iterations 0 through $k-1$ (excluding w_k). Let us compute the inner expectation:

$\mathbb{E}\left[\left\langle g_k - \widetilde{g}_k, x_{k+1} - x^* \right\rangle \mid w_0^{k-1}\right]$	
$= \mathbb{E}\left[\left\langle g_k - \widetilde{g}_k, x_{k+1} \right\rangle \mid w_0^{k-1}\right]$	$\mathbb{E}\left[\left\langle g_k - \widetilde{g}_k, x^* \right\rangle \mid w_0^{k-1}\right] = 0$ as $\widetilde{g}_k$ is unbiased; see (6)
$= \mathbb{E}\left[\left\langle g_k - \widetilde{g}_k, P_{\mathcal{X}}\left(x_k - a_k \widetilde{g}_k\right) \right\rangle \mid w_0^{k-1}\right]$	
$= \mathbb{E}\left[\left\langle g_k - \widetilde{g}_k, P_{\mathcal{X}}\left(x_k - a_k \widetilde{g}_k\right) - P_{\mathcal{X}}\left(x_k - a_k g_k\right) \right\rangle \mid w_0^{k-1}\right]$	$\mathbb{E}\left[\left\langle g_k - \widetilde{g}_k, P_{\mathcal{X}}\left(x_k - a_k g_k\right) \right\rangle \mid w_0^{k-1}\right] = 0$ as $\widetilde{g}_k$ is unbiased and independent of x_k and w_0^{k-1} see (7)
$\leq \mathbb{E}\left[\left\|g_k - \widetilde{g}_k\right\|_2 \cdot \left\|P_{\mathcal{X}}\left(x_k - a_k \widetilde{g}_k\right) - P_{\mathcal{X}}\left(x_k - a_k g_k\right)\right\|_2 \mid w_0^{k-1}\right]$	Cauchy-Schwarz
$\leq \mathbb{E}\left[a_k \left\|g_k - \widetilde{g}_k\right\|_2^2 \mid w_0^{k-1}\right]$	$P_{\mathcal{X}}$ is non-expansive
$\leq a_k \sigma^2.$	bounded variance assumption

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

It follows that

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

Summing both sides over k and telescoping, we get the bound

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

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

$$\mathbb{E}\left[f(x_K^{\text{out}}) - f(x^*)\right] \leq \frac{\|x_0 - x^*\|_2\sqrt{M^2 + 2\sigma^2}}{\sqrt{K+1}}. \tag{2}$$

  Setting $\sigma = 0$ recovers the rate for PSubSGD from last lecture. Note that for first-order method, the $O(1/\sqrt{K})$ rate is optimal even when we have access to exact gradients (see last lecture).

- Same discussion about anytime algorithm, unknown/unbounded diameter of $\mathcal{X}$, unknown M, unknown σ^2, etc.

3 Analysis of SGD in other settings (Optional)

In this section, we state without proof several additional convergence results for (projected) stochastic (sub)gradient descent.[2] As before, we assume that f is convex and the stochastic gradient $g(x, w)$ is unbiased, but we will consider other additional properties of f and $g(x, w)$.

3.1 Role of smoothness

Still assume that stochastic gradient has variance bounded by σ^2; see equation (1). We make the additional assumption that f is L-smooth (w.r.t. $\|\cdot\|_2$). Let $D := \max_{x,y\in\mathcal{X}} \|x - y\|_2$ be the diameter of $\mathcal{X}$. With a constant stepsize $a_k = \frac{1}{L+(\sigma/D)\sqrt{(K+1)/2}}, \forall k$, one can show that

$$\mathbb{E}f\left(x_K^{\text{out}}\right) - f(x^*) \leq D\sigma\sqrt{\frac{2}{K+1}} + \frac{LD^2}{K+1}. \tag{3}$$

When K is large, the first term on the RHS dominates and thus we have an $O(1/\sqrt{K})$ rate. This rate is essentially the same as the bound (2) for nonsmooth f. Therefore, smoothness does not offer much benefit in the stochastic setting. In contrast, in the deterministic setting, smoothness leads to the faster rates of $O(1/K)$ (for GD) and $O(1/K^2)$ (for AGD).

[2] For details please see

- Section 6 of Sebastien Bubeck's monograph.
- Chapter 5 of Wright and Recht, Optimization for Data Analysis.

3.2 Role of strong convexity

Going back to the setting with M-Lipschitz f. Still assume that stochastic gradient has variance bounded by σ^2; see equation (1). We make the additional assumption that f is m-strongly convex (w.r.t. $\|\cdot\|_2$). Note that strong convexity and Lipschitzness can hold simultaneously only when $\mathcal{X}$ is bounded. [3]

For the diminishing stepsize $a_k = \frac{2}{m(k+2)}$, we have

$$\mathbb{E} f\left(\sum_{k=0}^{K} \frac{2(k+1)}{(K+1)(K+2)} x_k\right) - f(x^*) \le \frac{2(M^2+\sigma^2)}{m(K+2)}. \tag{4}$$

This $O(1/K)$ rate is better than the $O(1/\sqrt{K})$ rate for non-strongly convex f.

3.3 More general noise

We now consider a more general form of noise assumption: there exist some $L_g \ge 0$ and $B \ge 0$ such that for all $x \in \mathcal{X}$:

$$\mathbb{E}\left[\|g(x,w)\|_2^2\right] \le L_g^2 \|x - x^*\|_2^2 + B^2. \tag{5}$$

We consider three cases.

3.3.1 $L_g = 0$, $B > 0$, convex f

This setting is a slight generalization of the previous assumption (1) of M-Lipschitz f and σ^2-bounded variance. In particular, the assumption (1) implies that

$$\begin{aligned}\mathbb{E}\left[\|g(x,w)\|_2^2\right] &= \|\mathbb{E}[g(x,w)]\|_2^2 + \mathbb{E}_w\left[\|\widetilde{g}(x,w) - g_x\|_2^2\right] \\ &= \|g_x\|_2^2 + \mathbb{E}_w\left[\|\widetilde{g}(x,w) - g_x\|_2^2\right] \le M^2 + \sigma^2.\end{aligned}$$

Therefore, the more general assumption (5) is satisfied with $L_g = 0$ and $B^2 = M^2 + \sigma^2$. In this case, using the constant stepsize $a_k = \frac{\|x_0 - x^*\|_w}{B\sqrt{K+1}}$, $\forall k$, we have

$$\mathbb{E}\left[f(x_K^{\text{out}}) - f(x^*)\right] \le \frac{\|x_0 - x^*\|_2 B}{\sqrt{K+1}}.$$

This bound is essentially the same as the bound (2) proved earlier.

3.3.2 $L_g > 0$, $B = 0$, m-strongly convex f

In this setting, we have $\mathbb{E}\left[\|g(x,w)\|_2^2\right] \to 0 = \nabla f(x^*)$ as $x \to x^*$. That is, the stochastic gradient becomes more and more accurate near x^*. Moreover, we have

$$\begin{aligned}L_g^2 \|x - x^*\|_2^2 &\ge \mathbb{E}\left[\|g(x,w)\|_2^2\right] \\ &\ge \|\mathbb{E}[g(x,w)]\|_2^2 && \text{Jensen's} \\ &= \|\nabla f(x)\|_2^2 = \|\nabla f(x) - \nabla f(x^*)\|_2^2, && \text{unbiased, } \nabla f(x^*) = 0\end{aligned}$$

[3] For a strongly convex function, its subgradient grows linearly away from x^*: $\|\nabla f(x)\|_2 \ge \frac{m}{2}\|x - x^*\|_2$, hence $\|\nabla f(x)\| \le M$ cannot be over the entire $\mathbb{R}^d$.

so the gradient of f satisfies a "Lipschitz-like" assumption.

With a constant stepsize $a_k = \frac{m}{L_g^2}, \forall k$, we have

$$\mathbb{E}\left\|x_K - x^*\right\|_2^2 \leq \left(1 - \frac{m^2}{L_g^2}\right)^K \left\|x_0 - x^*\right\|^2.$$

We have geometric convergence thanks to strong convexity and the Lipschitz-like property. The contraction factor is $1 - \frac{m^2}{L_g^2}$, which is worse than the $1 - \frac{m}{L}$ (for GD) and $1 - \sqrt{\frac{m}{L}}$ (for AGD) factors we saw in the deterministic setting with m-strong convexity and L-Lipschitz gradient.

3.3.3 $L_g > 0, B > 0$, m-strongly convex f

With a diminishing stepsize $a_k = \frac{1}{2m(L_g^2/2m^2+k)}$, we have

$$\mathbb{E}\left\|x_K - x^*\right\|_2^2 \leq \frac{c_0 B^2}{2m(L_g^2/2m^2 + K)}.$$

For large K, this is an $O(1/K)$ rate.

Appendix: Basic properties of expectation

Linearity Let X and Y be two random variables, and a, b be two (deterministic) numbers. Then

$$\mathbb{E}\left[aX + bY\right] = a\mathbb{E}\left[X\right] + b\mathbb{E}\left[Y\right].$$

Multiplicity under independence Let X and Y be two *independent* random variables. Then

$$\mathbb{E}\left[XY\right] = \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right] \qquad \text{and} \qquad \mathbb{E}\left[X \mid Y\right] = \mathbb{E}\left[X\right].$$

These property hold for conditional expectation and random vectors as well. For example, we have

$$\begin{aligned}
\mathbb{E}\left[\left\langle g_k - \widetilde{g}_k, x^*\right\rangle \mid w_0^{k-1}\right] &= \left\langle g_k - \mathbb{E}\left[\widetilde{g}_k \mid w_0^{k-1}\right], x^*\right\rangle. && \text{linearity} \\
&= \left\langle g_k - \mathbb{E}\left[\widetilde{g}_k\right], x^*\right\rangle && \text{independence} \\
&= \left\langle \underbrace{g_k - g_k}_{=0}, x^*\right\rangle && \text{unbiased} \\
&= 0. && (6)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\left[\left\langle g_k - \widetilde{g}_k, P_{\mathcal{X}}\left(x_k - a_k g_k\right)\right\rangle \mid w_0^{k-1}\right] \\
=&\left\langle \mathbb{E}\left[g_k - \widetilde{g}_k \mid w_0^{k-1}\right], \mathbb{E}\left[P_{\mathcal{X}}\left(x_k - a_k g_k\right) \mid w_0^{k-1}\right]\right\rangle && \text{independence} \\
=&\left\langle g_k - \mathbb{E}\left[\widetilde{g}_k \mid w_0^{k-1}\right], \mathbb{E}\left[P_{\mathcal{X}}\left(x_k - a_k g_k\right) \mid w_0^{k-1}\right]\right\rangle && \text{linearity} \\
=&\left\langle g_k - \mathbb{E}\left[\widetilde{g}_k\right], \mathbb{E}\left[P_{\mathcal{X}}\left(x_k - a_k g_k\right) \mid w_0^{k-1}\right]\right\rangle && \text{independence} \\
=&\left\langle \underbrace{g_k - g_k}_{=0}, \mathbb{E}\left[P_{\mathcal{X}}\left(x_k - a_k g_k\right) \mid w_0^{k-1}\right]\right\rangle && \text{unbiased} \\
=&0. && (7)
\end{aligned}$$