Math 273A Notes: Chapter 2

Ernest K. Ryu

November 12, 2025

Stochastic optimization Consider the stochastic optimization problem

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \quad \mathop{\mathbb{E}}_{\omega}[f(x;\omega)] = F(x),$$

where ω is a random variable. In machine learning, such problems arise in the finite-sum form

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \quad \mathop{\mathbb{E}}_{I\sim\text{Uniform}\{1,\dots,N\}}[f_I(x)] = \frac{1}{N}\sum_{i=1}^N f_i(x),$$

or

$$\underset{\theta\in\mathbb{R}^d}{\text{minimize}} \quad \frac{1}{N}\sum_{i=1}^N \ell(f_\theta(X_i), Y_i).$$

Stochastic (sub)gradients Under mild conditions, we have

$$\nabla F(x) = \nabla \mathop{\mathbb{E}}_{\omega}[f(x;\omega)] = \mathop{\mathbb{E}}_{\omega}[\nabla_x f(x;\omega)].$$

Therefore, $\nabla_x f(x;\omega)$ is an unbiased estimate of $\nabla F(x)$, and we say $\nabla_x f(x;\omega)$ is a *stocahstic gradient* of F at $x\in\mathbb{R}^d$.

Let $g_\omega \in \partial f(x;\omega)$ be a random subgradient at $x\in\mathbb{R}^d$. Then,

$$\begin{aligned} F(y) = \mathop{\mathbb{E}}_{\omega}[f(y;\omega)] &\geq \mathop{\mathbb{E}}_{\omega}[f(x;\omega) + \langle g_\omega, y-x\rangle] \\ &= F(x) + \langle \mathop{\mathbb{E}}_{\omega}[g_\omega], y-x\rangle, \qquad \forall\, y\in\mathbb{R}^d \end{aligned}$$

and $\mathbb{E}_\omega[g_\omega]\in\partial F(x)$, provided that $\mathbb{E}_\omega[g_\omega]$ is well defined. In this case, we say $g_\omega\in\partial f(x;\omega)$ is a *stocahstic subgradient* of F at $x\in\mathbb{R}^d$.

Stochastic (sub)gradient descent (SGD) Consider the algorithm stochastic (sub)gradient descent (SGD)

$$x_{k+1} = x_k - \alpha_k g_k$$

for $k = 0, 1, \ldots$, where g_k is a stochastic (sub)gradient of F at x_k.

More specifically, we assume that

$$\mathbb{E}_k[g_k] \in \partial F(x_k),$$

where

$$\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot \mid x_0, x_1, \ldots, x_k]$$

is the conditional expectation, conditioned on the iterates up to x_k. We will also assume that the conditional variance is bounded:

$$\mathrm{Var}_k(g_k) = \mathbb{E}_k\big[\|g_k - \mathbb{E}_k[g_k]\|^2\big] \leq \sigma^2$$

Analysis of SGD

Theorem 1. *Let $F\colon \mathbb{R}^d \to \mathbb{R}$ be a G-Lipschitz continuous convex function. Assume F has a minimizer $x_\star$. Let $x_0 \in \mathbb{R}^d$ be a starting point. Let $K > 0$ be the total iteration count. Assume the stochastic subgradient g_k satisfies*

$$\mathbb{E}_k[g_k] \in \partial F(x_k), \qquad \mathrm{Var}_k(g_k) \leq \sigma^2$$

for $k = 0, 1, \ldots$. Then, SGD with the constant stepsize

$$\alpha_k = \alpha = \frac{\|x_0 - x_\star\|_2}{\sqrt{G^2 + \sigma^2}\sqrt{K+1}}$$

exhibits the rate

$$\mathbb{E}\big[f(\bar{x}^K) - f(x_\star)\big] \leq \frac{\sqrt{G^2 + \sigma^2}\|x_0 - x_\star\|_2}{\sqrt{K+1}},$$

where

$$\bar{x}^K = \frac{1}{K+1}\sum_{k=0}^{K} x_k.$$

Proof. First,

$$\begin{aligned}\mathbb{E}_k\big[\|x_{k+1} - x_\star\|_2^2\big] &= \|x_k - x_\star\|_2^2 - 2\alpha\langle \mathbb{E}_k[g_k], x_k - x_\star\rangle + \alpha^2 \mathbb{E}_k[\|g_k\|^2] \\ &\leq \|x_k - x_\star\|_2^2 - 2\alpha(F(x_k) - F(x_\star)) + \alpha^2(G^2 + \sigma^2).\end{aligned}$$

We take the total expectation on both sides to get

$$\mathbb{E}\big[\|x_{k+1} - x_\star\|_2^2\big] \leq \mathbb{E}\big[\|x_k - x_\star\|_2^2\big] - 2\alpha\mathbb{E}[F(x_k) - F(x_\star)] + \alpha^2(G^2 + \sigma^2).$$

By the same telescoping-sum argument as in the (non-stochastic) subgradient descent, we have

$$\mathbb{E}[F(\bar{x}^K) - F(x_\star)] \leq \frac{1}{K+1} \sum_{k=0}^{K} \mathbb{E}[F(x_k) - F(x_\star)] \leq \frac{\sqrt{G^2 + \sigma^2} R}{\sqrt{K+1}}.$$

□

Let $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots$ be a sequence of σ-algebras. Write $\mathbb{E}[X \mid \mathcal{F}_k]$ for the conditional expectation of a random variable X with respect to $\mathcal{F}_k$. In the context of this chapter, $\mathcal{F}_k$ represents the information before iteration k, and the quantity V_k is $\mathcal{F}_k$-measurable. Therefore, $\mathbb{E}[V_k \mid \mathcal{F}_k] = V_k$. To say this without using measure theoretic language, $\mathbb{E}[\cdot \mid \mathcal{F}_k]$ represents the expectation conditioned on the information before iteration k, and V_k is determined by the randomness of the iterations before k. For example, the Lyapunov function $V_k = \|x_k - x_\star\|^2$. Since x_k is determined by the starting point x_0 and the stochastic gradients $g_0, \ldots, g_{k-1}$, we have $\mathbb{E}[V_k \mid \mathcal{F}_k] = V_k$, since there is no randomness in V_k once we condition on the information before iteration k.

Theorem 2. Supermartingale convergence theorem. *Let V_k and S_k be $\mathcal{F}_k$-measurable random variables satisfying $V_k \geq 0$ and $S_k \geq 0$ almost surely for $k = 0, 1, \ldots$. Assume*

$$\mathbb{E}[V_{k+1} \mid \mathcal{F}_k] \leq V_k - S_k$$

holds for $k = 0, 1, \ldots$. Then

1. *$V_k \to V_\infty$*
2. *$\sum_{k=0}^{\infty} S_k < \infty$*

almost surely. (Note that the limit V_∞ is a random variable.)

We do not use the supermartingale convergence theorem itself, but we state it here for reference. The proof can be found in many standard textbooks on probability theory. (The standard supermartingale convergence theorem is slightly more general.)

Theorem 3 (Quasi-Martingale convergence theorem)**.** *Let V_k, S_k, and U_k be $\mathcal{F}_k$-measurable random variables satisfying $V_k \geq 0$, $S_k \geq 0$, and $U_k \geq 0$ almost surely for $k = 0, 1, \ldots$. Assume*

$$\mathbb{E}[V_{k+1} \mid \mathcal{F}_k] \leq V_k - S_k + U_k$$

and

$$\sum_{i=1}^{\infty} U_i < \infty$$

almost surely. Then

1. *$V_k \to V_\infty$*

2. $\sum_{k=0}^{\infty} S_k < \infty$

almost surely. (Note that the limit V_∞ is a random variable.)

This "almost supermartingale" convergence theorem is due to Robbins and Siegmund 1985.

Proof. Define

$$\tilde{V}_k = V_k - \sum_{j=0}^{k-1} U_j.$$

Then

$$\mathbb{E}[\tilde{V}_{k+1} \mid \mathcal{F}_k] \leq \tilde{V}_k - S_k$$

So we can apply the supermartingale convergence theorem. □

Theorem 4. *Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be convex and G-Lipschitz. Let α_k be a sequence of positive scalars such that*

$$\sum_k \alpha_k = \infty, \qquad \sum_k \alpha_k^2 < \infty$$

Then

$$g_k \in \partial f(x_k)$$
$$x_{k+1} = x_k - \alpha_k g_k$$

converges in the sense of $x_k \to x_\infty \in \operatorname{argmin} f$ almost surely.

Proof. Let $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot \mid x_0, \ldots, x_k]$. Then,

$$\begin{aligned}\mathbb{E}_k\left[\|x_{k+1} - \tilde{x}_\star\|_2^2\right] &= \|x_k - \tilde{x}_\star\|_2^2 - 2\alpha_k \langle \mathbb{E}_k[g_k], x_k - \tilde{x}_\star\rangle + \alpha_k^2 \mathbb{E}_k[\|g_k\|^2] \\ &\leq \|x_k - \tilde{x}_\star\|_2^2 - 2\alpha_k(F(x_k) - F(\tilde{x}_\star)) + \alpha_k^2(G^2 + \sigma^2).\end{aligned}$$

Now we apply the quasi-Martingale convergence theorem to conclude that

$$\|x_k - \tilde{x}_\star\| \to \text{limit}$$

converges to a limit almost surely. By Proposition 1, we can conclude $\|x_k - x_\star\| \to$ limit for all (uncountably many) $x_\star \in \operatorname{argmin} f$. The quasi-Martingale convergence theorem also allows us to conclude that

$$\sum_{k=0}^{\infty} \alpha_k (F(x_k) - F_\star) < \infty$$

almost surely, so

$$\liminf_{k\to\infty} F(x_k) - F_\star = 0$$

almost surely.

Now choose a subsequence such that $F(x_{k_j}) \to F_\star$ and by passing to a further subsequence, we can have $x_{k_j} \to x_\infty$. Then, by continuity of F, we have that $x_\infty \in \operatorname{argmin} F$.

Since $\|x_k - x_\star\| \to$ limit, this holds for all $x_\star \in \operatorname{argmin} f$, including $x_\star = x_\infty$, we conclude that $\|x_k - x_\infty\| \to 0$ almost surely, i.e., x_k converges to a minimizer almost surely. □

The necessity of Proposition 1 is subtle. Since we choose $x_\star \in \operatorname{argmin} f$ first and then apply the supermartingale convergence theorem, the conclusion that $\lim_{k\to\infty} \|x_k - x_\star\|$ exists with probability 1 applies to one fixed point $x_\star$ at a time. Without a formal argument, this does not immediately imply that $\lim_{k\to\infty} \|x_k - x_\star\|$ for all $x_\star \in \operatorname{argmin} f$ with probability 1 in the case where Fix $\operatorname{argmin} f$ is not a singleton and therefore has uncountably many minimizers.

Proposition 1. *Let $Y \subseteq \mathbb{R}^n$ and let $x_0, x_1, \dots$ be a random sequence. Then statement 1 implies statement 2.*

1. *For all $y \in Y$* [with probability 1, $\lim_{k\to\infty} \|x_k - y\|$ exists].
2. *With probability 1* [for all $y \in Y$, $\lim_{k\to\infty} \|x_k - y\|$ exists].

Proof of Proposition 1. This proof uses the separability of $\mathbb{R}^n$, that is, $\mathbb{R}^n$ contains a countable, dense subset.

In particular, $Y \subseteq \mathbb{R}^n$ has a countable, dense subset $\{y^1, y^2, \dots\}$. By statement 1, given $i \in \{1, 2, \dots\}$, there is a probability 1 event $\Omega(y^i)$ such that $\lim_{k\to\infty} \|x_k(\omega) - y^i\|$ for all $\omega \in \Omega(y^i)$. Therefore $\lim_{k\to\infty} \|x_k(\omega) - y^i\|$ exists for all $i \in \{1, 2, \dots\}$ for $\omega \in \cap_{i=1,2,\dots}\Omega(y^i)$, and $\cap_{i=1,2,\dots}\Omega(y^i)$ is an event with probability 1 since it is a countable intersection of probability 1 events.

(In other words: with probability 1 [for all $i = 1, 2, \dots$, $\lim_{k\to\infty} \|x_k - y^i\|$ exists]. The subtlety is that an *uncountable* intersection of probability 1 events may not have probability 1.)

Now pick any $y \in Y$. Statement 2 is proved if we can show $\|x_k(\omega) - y\|$ converges for $\omega \in \cap_{i=1,2,\dots}\Omega(y^i)$. To this end, pick any $\varepsilon > 0$. Since $\{y^1, y^2, \dots\} \subseteq Y$ is dense, there exists $y^i \in Y$ such that $\|y^i - y\| \le \varepsilon$. We get the following lower and upper bounds with the triangle inequality:

$$\|x_k(\omega) - y\| \le \|x_k(\omega) - y^i\| + \|y^i - y\| \le \|x_k(\omega) - y^i\| + \varepsilon,$$
$$\|x_k(\omega) - y\| \ge \|x_k(\omega) - y^i\| - \|y^i - y\| \ge \|x_k(\omega) - y^i\| - \varepsilon.$$

Since $\omega \in \Omega \subset \Omega(y^i)$,

$$\limsup_{k\to\infty} \|x_k(\omega) - y\| \le \lim_{k\to\infty} \|x_k(\omega) - y^i\| + \varepsilon,$$
$$\liminf_{k} \|x_k(\omega) - y\| \ge \lim_{k\to\infty} \|x_k(\omega) - y^i\| - \varepsilon,$$

and together we have

$$0 \le \limsup_{k} \|x_k(\omega) - y\| - \liminf_{k} \|x_k(\omega) - y\| \le 2\varepsilon.$$

As $\varepsilon > 0$ is arbitrary, we conclude

$$\limsup_{k\to\infty} \|x_k(\omega) - y\| = \liminf_{k\to\infty} \|x_k(\omega) - y\| = \lim_{k\to\infty} \|x_k(\omega) - y\|.$$

□

In mathematical terms, the key idea of Proposition 1 is that (i) Y has a countable dense subset, (ii) the sequence of functions $\{\|x_k - \cdot\|\}_{k\in\mathbb{N}}$ has a limit on the countable dense subset of Y, and (iii) if an equicontinuous sequence of functions has a limit on the dense subset of a metric space, then the limit exists on the entire metric space.