CIS3333: Mathematics of Machine Learning Fall 2025

Lecture: Convergence Analysis of SGD

Date: November 17th, 2025 *Author: Surbhi Goel*

1 Introduction

In the last lecture, we proved that for smooth, convex functions, Gradient Descent (GD) converges at a rate of $O(1/T)$. The proof relied on two key ingredients: **convexity**, which ensures that every step moves us in a globally meaningful direction, and **smoothness**, which guarantees that each gradient step decreases the objective function. In this lecture, we will see that for the basic convergence guarantee we prove for SGD, only convexity (plus simple assumptions on the stochastic gradient) is needed; smoothness will not play a direct role in the analysis.

However, in large-scale machine learning, we rarely use batch gradient descent. The objective is typically a sum over a large dataset:

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

Computing the full gradient $\nabla f(x)$ is too expensive. Instead, we use Stochastic Gradient Descent (SGD), which uses a single sample (or a mini-batch) to estimate the gradient:

$$x_{t+1} = x_t - \eta_t \nabla f_{i_t}(x_t), \quad i_t \sim \text{Uniform}(\{1, \dots, N\})$$

The key challenge with SGD is that the stochastic gradient $\nabla f_{i_t}(x_t)$ is a *noisy* estimate of the true gradient $\nabla f(x_t)$. A single SGD step is **not guaranteed** to decrease the objective function.

If an SGD step can make things worse, why does it converge at all?

Today, we will show that while a single step can be bad, the updates are correct *on average*. We will prove that SGD converges in expectation, but at a slower rate than batch GD.

2 Setting and Assumptions

We consider the same basic setting as batch GD: minimizing a function $f(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x)$ where each f_i is convex. This implies f is also convex.

The key difference is that at each step, instead of computing the full gradient $\nabla f(x_t) = \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(x_t)$, we sample a random index $i_t \sim \text{Uniform}(\{1, \dots, N\})$ and use the stochastic gradient $\nabla f_{i_t}(x_t)$.

Key Properties of the Stochastic Gradient. The stochastic gradient has two important properties:

1. **Unbiased:** By linearity of expectation,

$$\mathbb{E}_{i_t}[\nabla f_{i_t}(x_t) \mid x_t] = \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(x_t) = \nabla f(x_t)$$

   On average, the stochastic gradient points in the same direction as the true gradient.

2. **Bounded Second Moment:** We assume there exists a constant G^2 such that

$$\mathbb{E}_{i_t}[\|\nabla f_{i_t}(x_t)\|_2^2 \mid x_t] \leq G^2$$

   This bounds how noisy the stochastic gradients can be.

These are the only places where probability enters the analysis: we treat the stochastic gradient as a random variable and apply the tools you have already seen (unbiased estimators, variance bounds) to control its effect.

3 SGD Convergence Proof

The proof structure is parallel to the one for gradient descent from the last lecture. In the GD proof, we used two lemmas: the **Descent Lemma** (which guaranteed $f(x_{t+1}) \leq f(x_t) - \frac{\eta}{2}\|\nabla f(x_t)\|_2^2$) and the **Progress Lemma** (which related the decrease in distance to the optimum to the optimality gap). For SGD, the Descent Lemma *fails*—a single stochastic step can increase the function value due to noise. However, we can still prove an analogous Progress Lemma that holds *in expectation*. We present the lemma and main theorem first, then discuss why the Descent Lemma fails and what assumptions replace smoothness.

Notation. In what follows, $\mathbb{E}_{i_t}[\cdot \mid x_t]$ denotes expectation over the random choice of index i_t at iteration t, conditional on the current iterate x_t (which depends on all previous randomness $i_0, \ldots, i_{t-1}$). When we write $\mathbb{E}[\cdot]$ without subscript, it denotes expectation over all the randomness in SGD (all index choices $i_0, i_1, \ldots$). The initial point x_0, the optimum x^*, and the step size η are deterministic.

Lemma 1 (Progress towards Optimum (SGD))**.** *Let f be convex, and let the SGD assumptions hold. For a single SGD step with step size η, we have:*

$$f(x_t) - f(x^*) \leq \frac{1}{2\eta}\Big(\|x_t - x^*\|_2^2 - \mathbb{E}_{i_t}[\|x_{t+1} - x^*\|_2^2 \mid x_t]\Big) + \frac{\eta G^2}{2}.$$

Proof. We analyze the change in distance to the optimum x^* after one SGD step. The expectation below is over the random choice of i_t, treating x_t as given:

$$\begin{aligned}\mathbb{E}_{i_t}[\|x_{t+1} - x^*\|_2^2 \mid x_t] &= \mathbb{E}_{i_t}[\|x_t - \eta\nabla f_{i_t}(x_t) - x^*\|_2^2 \mid x_t] \quad \text{(substitute SGD update)}\\ &= \mathbb{E}_{i_t}[\|x_t - x^*\|_2^2 - 2\eta(x_t - x^*)^\top \nabla f_{i_t}(x_t) + \eta^2\|\nabla f_{i_t}(x_t)\|_2^2 \mid x_t] \quad \text{(expand squared norm)}\end{aligned}$$

Taking expectation inside and using unbiasedness ($\mathbb{E}_{i_t}[\nabla f_{i_t}(x_t) \mid x_t] = \nabla f(x_t)$) and the bounded second moment:

$$\begin{aligned} \mathbb{E}_{i_t}[\|x_{t+1} - x^*\|_2^2 \mid x_t] &= \|x_t - x^*\|_2^2 - 2\eta(x_t - x^*)^\top \nabla f(x_t) + \eta^2 \mathbb{E}_{i_t}[\|\nabla f_{i_t}(x_t)\|_2^2 \mid x_t] \quad \text{(linearity of expectation)} \\ &\le \|x_t - x^*\|_2^2 - 2\eta(x_t - x^*)^\top \nabla f(x_t) + \eta^2 G^2 \quad \text{(bounded second moment)} \end{aligned}$$

Using the first-order condition for convexity, $\nabla f(x_t)^\top (x_t - x^*) \ge f(x_t) - f(x^*)$, we can bound the middle term:

$$\mathbb{E}_{i_t}[\|x_{t+1} - x^*\|_2^2 \mid x_t] \le \|x_t - x^*\|_2^2 - 2\eta(f(x_t) - f(x^*)) + \eta^2 G^2 \quad \text{(apply convexity)}$$

Rearranging the terms and dividing by 2η gives the desired result:

$$f(x_t) - f(x^*) \le \frac{1}{2\eta}\big(\|x_t - x^*\|_2^2 - \mathbb{E}_{i_t}[\|x_{t+1} - x^*\|_2^2 \mid x_t]\big) + \frac{\eta G^2}{2} \quad \text{(rearrange and divide by } 2\eta\text{)}$$

$\square$

With this lemma in hand, the main theorem follows by summing over iterations and applying Jensen's inequality.

Theorem 2 (SGD Convergence with Constant Step Size)**.** *Let f be convex, and let the SGD assumptions hold. Run SGD for T steps with constant step size $\eta > 0$, and define the averaged iterate*

$$\bar{x}_T := \frac{1}{T}\sum_{t=0}^{T-1} x_t.$$

Then

$$\mathbb{E}[f(\bar{x}_T)] - f(x^*) \quad \le \quad \frac{\|x_0 - x^*\|_2^2}{2\eta T} + \frac{\eta G^2}{2}.$$

Proof. Apply the one-step lemma at each iteration t:

$$f(x_t) - f(x^*) \quad \le \quad \frac{1}{2\eta}\Big(\|x_t - x^*\|_2^2 - \mathbb{E}_{i_t}[\|x_{t+1} - x^*\|_2^2 \mid x_t]\Big) + \frac{\eta G^2}{2} \quad \text{(Progress Lemma)}$$

Summing over $t = 0, \dots, T-1$ gives

$$\sum_{t=0}^{T-1}(f(x_t) - f(x^*)) \quad \le \quad \frac{1}{2\eta}\sum_{t=0}^{T-1}\Big(\|x_t - x^*\|_2^2 - \mathbb{E}_{i_t}[\|x_{t+1} - x^*\|_2^2 \mid x_t]\Big) + \frac{T\eta G^2}{2} \quad \text{(sum over } t\text{)}$$

Now take expectation over all randomness $i_0, \dots, i_{T-1}$ on both sides. By the tower property, $\mathbb{E}_{i_0,\dots,i_{T-1}}[\mathbb{E}_{i_t}[\|x_{t+1} - x^*\|_2^2 \mid x_t]] = \mathbb{E}_{i_0,\dots,i_{T-1}}[\|x_{t+1} - x^*\|_2^2]$, so:

$$\mathbb{E}\left[\sum_{t=0}^{T-1}(f(x_t) - f(x^*))\right] \quad \le \quad \frac{1}{2\eta}\sum_{t=0}^{T-1}\Big(\mathbb{E}[\|x_t - x^*\|_2^2] - \mathbb{E}[\|x_{t+1} - x^*\|_2^2]\Big) + \frac{T\eta G^2}{2} \quad \text{(take expectation; tower proper}$$

The first sum on the right is a telescoping sum:

$$\begin{aligned} \sum_{t=0}^{T-1}\Big(\mathbb{E}[\|x_t - x^*\|_2^2] - \mathbb{E}[\|x_{t+1} - x^*\|_2^2]\Big) &= \|x_0 - x^*\|_2^2 - \mathbb{E}[\|x_T - x^*\|_2^2] \quad \text{(telescope; } x_0 \text{ is deterministic)} \\ &\le \|x_0 - x^*\|_2^2, \quad \text{(norms are non-negative)} \end{aligned}$$

Therefore

$$\mathbb{E}\left[\sum_{t=0}^{T-1}(f(x_t) - f(x^*))\right] \quad \leq \quad \frac{\|x_0 - x^*\|_2^2}{2\eta} + \frac{T\eta G^2}{2} \qquad \text{(substitute telescoping sum)}$$

Dividing both sides by T yields

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}(f(x_t) - f(x^*))\right] \quad \leq \quad \frac{\|x_0 - x^*\|_2^2}{2\eta T} + \frac{\eta G^2}{2} \qquad \text{(divide by } T\text{)}$$

By convexity of f and the definition of $\bar{x}_T$, we have

$$f(\bar{x}_T) = f\left(\frac{1}{T}\sum_{t=0}^{T-1} x_t\right) \leq \frac{1}{T}\sum_{t=0}^{T-1} f(x_t), \qquad \text{(convexity of } f\text{)}$$

so

$$f(\bar{x}_T) - f(x^*) \leq \frac{1}{T}\sum_{t=0}^{T-1}(f(x_t) - f(x^*)). \qquad \text{(subtract } f(x^*)\text{)}$$

Taking expectations on both sides and combining with the previous inequality gives the desired bound:

$$\mathbb{E}[f(\bar{x}_T)] - f(x^*) \leq \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}(f(x_t) - f(x^*))\right] \leq \frac{\|x_0 - x^*\|_2^2}{2\eta T} + \frac{\eta G^2}{2}.$$

□

3.1 Why the Descent Lemma Fails for SGD

Recall from the previous lecture that for an L-smooth function f, the Descent Lemma for GD follows from the quadratic upper bound

$$f(x_{t+1}) \leq f(x_t) + \nabla f(x_t)^\top (x_{t+1} - x_t) + \frac{L}{2}\|x_{t+1} - x_t\|_2^2.$$

Plugging in the *GD* update $x_{t+1} = x_t - \eta \nabla f(x_t)$ gives

$$f(x_{t+1}) \leq f(x_t) - \eta\|\nabla f(x_t)\|_2^2 + \frac{L\eta^2}{2}\|\nabla f(x_t)\|_2^2,$$

so for $\eta \leq 1/L$ we get a guaranteed decrease in $f(x_t)$ at every step.

For *SGD*, the update is $x_{t+1} = x_t - \eta \nabla f_{i_t}(x_t)$, where the stochastic gradient $\nabla f_{i_t}(x_t)$ may point in a different direction from the true gradient $\nabla f(x_t)$. Applying the same smoothness inequality to f now yields

$$f(x_{t+1}) \leq f(x_t) - \eta\, \nabla f(x_t)^\top \nabla f_{i_t}(x_t) + \frac{L\eta^2}{2}\|\nabla f_{i_t}(x_t)\|_2^2.$$

The middle term involves the inner product $\nabla f(x_t)^\top \nabla f_{i_t}(x_t)$, whose sign can be positive or negative on any given step. Smoothness controls *how much* f can increase or decrease, but because the stochastic gradient can occasionally point in the "wrong" direction, the right-hand side is not guaranteed to be smaller than $f(x_t)$. In other words, we can no longer prove that $f(x_{t+1}) \leq f(x_t)$ deterministically at each step; we can only control the behavior of SGD *on average* using unbiasedness and variance bounds.

3.2 What Replaced Smoothness in the SGD Proof?

For GD, smoothness played a crucial role: it gave a quadratic upper bound on f and led to a deterministic decrease in $f(x_t)$ at every step (the Descent Lemma). In the SGD setting, we do *not* get such a per-step decrease because the stochastic gradient can point in the wrong direction, even if f is smooth. Instead, in the SGD proof above we replaced the smoothness assumption with a *bounded second moment* assumption on the stochastic gradients:

$$\mathbb{E}_{i_t}[\|\nabla f_{i_t}(x_t)\|_2^2 | x_t] \leq G^2.$$

This assumption is exactly what allows us to control the extra $\eta^2\|\nabla f_{i_t}(x_t)\|_2^2$ term that appears in the distance recursion (see the proof of the Progress Lemma). Because we track the squared distance $\|x_t - x^*\|_2^2$ instead of the function value $f(x_t)$, and use only

- convexity (to relate $\nabla f(x_t)^\top(x_t - x^*)$ to $f(x_t) - f(x^*)$), and
- the unbiasedness and bounded second moment of the stochastic gradient,

we never need the quadratic upper bound on f that smoothness provides. This is why smoothness does not explicitly appear in the SGD convergence theorem, even though it is still essential for the GD result from the previous lecture.

4 Impact of Learning Rate

The theorem gives the bound

$$\mathbb{E}[f(\bar{x}_T)] - f(x^*) \leq \frac{R^2}{2\eta T} + \frac{\eta G^2}{2}, \quad \text{where } R = \|x_0 - x^*\|_2.$$

This bound has two competing terms:

- **Optimization error:** $\frac{R^2}{2\eta T}$ goes to zero as $T \to \infty$, and is smaller when η is larger.
- **Noise term:** $\frac{\eta G^2}{2}$ is constant in T, and is smaller when η is smaller.

This captures the fundamental trade-off in SGD: a large step size makes fast progress initially but leaves us in a "noise ball" around the optimum; a small step size reduces the noise but slows down convergence.

Optimal Fixed Step Size. To minimize the bound for a given T, we can balance the two terms by setting $\eta \propto 1/\sqrt{T}$, which gives an overall rate of $O(1/\sqrt{T})$. This is slower than the $O(1/T)$ rate for batch gradient descent from the previous lecture, which is the price we pay for using noisy but cheap gradients. *You will prove this in the homework.*

Decreasing Step Size. To make the error go to zero as $T \to \infty$, we need a step size that decreases over time so that the noise term eventually vanishes. The most common choice is $\eta_t = \frac{\eta}{\sqrt{t+1}}$, which balances progress and noise accumulation.

To see why this works, consider a more general version of our theorem with varying step sizes η_t. The proof technique extends to give:

$$\mathbb{E}[f(\tilde{x}_T)] - f(x^*) \quad \leq \quad \frac{R^2}{2\sum_{t=0}^{T-1} \eta_t} + \frac{G^2}{2} \cdot \frac{\sum_{t=0}^{T-1} \eta_t^2}{\sum_{t=0}^{T-1} \eta_t},$$

where

$$\tilde{x}_T = \frac{1}{\sum_{t=0}^{T-1} \eta_t} \sum_{t=0}^{T-1} \eta_t x_t$$

is a weighted average of the iterates (with weights proportional to η_t). For the choice $\eta_t = \frac{\eta}{\sqrt{t+1}}$, we have $\sum_{t=0}^{T-1} \eta_t \approx \Theta(\sqrt{T})$ and $\sum_{t=0}^{T-1} \eta_t^2 \approx \Theta(\log T)$, which gives an overall rate of $O(1/\sqrt{T})$ (hiding logarithmic factors). This is slower than the $O(1/T)$ rate for batch GD, but the noise term now vanishes as $T \to \infty$, unlike the fixed step size case.