Mathematical Introduction to Machine Learning

Lecture 6: Stochastic Gradient Descent

November 18, 2025

Lecturer: Lei Wu *Scribe: Lei Wu*

1 Problem Setup

In machine learning, the most common objective function is the empirical risk

$$\widehat{\mathcal{R}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(x_i;\theta), y_i), \tag{1}$$

where $\ell : \mathbb{R}\times\mathbb{R} \to \mathbb{R}_{\geq 0}$ denotes the loss function and $h(x_i;\theta)$ is the model prediction. For gradient descent (GD), computing the gradient at each step costs $O(n)$ operations, which can be prohibitively expensive when n is large (e.g., $n = 10^6$). To address this issue, stochastic gradient descent (SGD) is introduced as an efficient approximation to GD [Robbins and Monro, 1951].

To start, we consider a more general setting where the objective function admits an expectation representation:

$$f(x) = \mathbb{E}_{w\sim\pi}[f(x;w)]. \tag{2}$$

In the case of empirical risk minimization, this reduce to the special case $\pi = \mathrm{Unif}([n])$. The GD update for (2) takes the form

$$x_{t+1} = x_t - \eta_t\, \mathbb{E}_{w\sim\pi}[\nabla f(x_t;w)]; \tag{3}$$

SGD approximates the full expectation by sampling a small batch:

$$x_{t+1} = x_t - \eta_t \underbrace{\frac{1}{B}\sum_{j=1}^{B} \nabla f(x_t; w_{j,t})}_{\text{minibatch gradient}}, \tag{4}$$

where $\{w_{1,t},\ldots,w_{B,t}\}$ are i.i.d. samples drawn from π. Here, B is a crucial hyperparameter and commonly referred to as the **batch size**.

Then some natural questions are:

- What is the difference between GD and SGD?
- How the choice of B and η affect the convergence behavior of SGD
  - When B is large, the minibatch gradient is accurate; we can use a large learning rate?
  - When B is small, the minibatch gradient is far from being accurate, and a small learning rate should be used.
- Does SGD converge when B is a constant, e.g., $B = 1$?

To better understand the role of stochasticity in SGD, it is useful to rewrite (4) in the following equivalent form:

$$x_{t+1} = x_t - \eta_t\big(\nabla f(x_t) + \xi_t\big), \tag{5}$$

where ξ_t denotes the gradient noise introduced by the minibatch approximation. By construction,

$$\begin{aligned}\mathbb{E}[\xi_t] &= 0,\\ \mathbb{E}[\xi_t\xi_t^\top] &= \frac{1}{B}\,\mathbb{E}_{w\sim\pi}\Big[(\nabla f(x_t;w) - \nabla f(x_t))(\nabla f(x_t;w) - \nabla f(x_t))^\top\Big].\end{aligned}$$

Thus the noise level is on the order of $O(B^{-1/2})$, meaning that larger minibatches lead to smaller stochastic fluctuations.

Remark 1.1. The formulation (5) highlights the essential structure of stochastic gradient methods: an exact gradient corrupted by zero-mean noise. Whenever the update takes this form, we simply refer to it as *SGD*. When this noise arises from using a minibatch, the method is more precisely called *mini-batch SGD*.

A phenomenological comparison between SGD and GD. Figure 1 illustrates the trajectories of GD and SGD under different batch sizes. The SGD trajectories exhibit noticeable stochastic fluctuations, and the variance becomes larger when the batch size is small. Nevertheless, the iterates still tend to drift toward the minimizer over time.

It is worth emphasizing that, in minibatch optimization, the SGD path is not expected to track the GD path closely. What matters is that the stochastic dynamics remain stable and converge toward a solution in the long run, rather than matching the deterministic trajectory step by step.

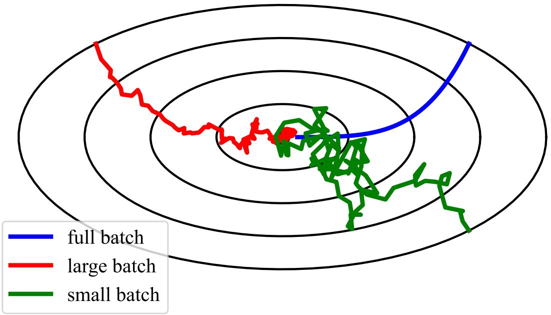

Figure 1: A visual illustration of how batch size affects SGD convergence.

Moreover, in Figure 2, we compare GD and SGD in terms of computational efficiency and illustrate how learning rate and batch size influence this efficiency. Specifically, we consider the problem of solving linear regression:

$$\min_{w\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}(w^\top x_i - y_i)^2,$$

where $d = 100$, $n = 200$, and the training data $x_i \overset{iid}{\sim} \mathcal{N}(0, A)$. Here $A = HH^\top$ with H is randomly sampled by $H_{i,j} \overset{iid}{\sim} \mathcal{N}(0,1)$. We examine different learning rates and batch sizes, and the results are shown in Figure 2. Note that the term “epoch” denotes a single pass through the entire dataset. For GD, each epoch corresponds to a single iteration. However, for SGD, one epoch is equivalent to n/B iterations, where n is the total number of samples in the dataset and B is the batch size. Thus, the number of epochs reflects the computational cost required. We observe the following phenomena:

- When measured in epochs (i.e., computational cost), SGD converges faster than GD.
- Due to the stochastic noise, it is hard for SGD to reach the high-precision regime.
- The convergence process generally consists of two phases:
  - Signal-dominated phase: when $|\nabla f(x_t)| \gg |\xi_t|$, the objective decreases rapidly;
  - Noise-dominated phase: once the noise becomes comparable to or larger than the gradient signal, progress stalls and the iterates oscillate around a neighborhood of the minimizer.

  To further reduce the loss in this regime, it is often necessary to decay the learning rate or increase the batch size.

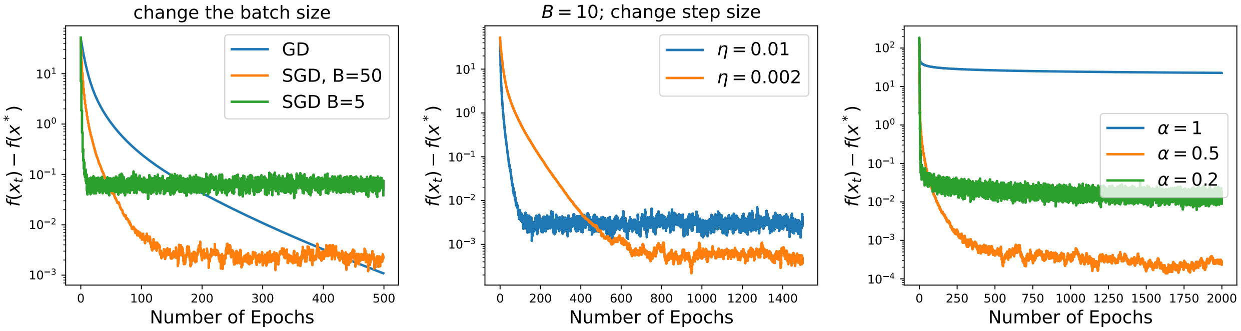

Figure 2: A convergence comparison between SGD and GD. **Left:** Varying the learning rate. **Middle:** Varying the batch size. **Right:** Learning rate decay of the form $\eta_t = \eta_0/(t+1)^\alpha$.

Continuous-time formulation. Consider SGD with a constant learning rate η:

$$x_{t+1} = x_t - \eta\big(\nabla f(x_t) + \xi_t\big),$$

where ξ_t denotes the stochastic gradient noise with $\mathbb{E}[\xi_t|x_t] = 0$ and covariance $\Sigma(x_t) = \mathbb{E}[\xi_t\xi_t^\top|x_t]$. We may rewrite the update as

$$x_{t+1} - x_t = -\eta\nabla f(x_t) + \underbrace{\sqrt{\eta}\big(-\sqrt{\eta}\,\xi_t\big)}_{\text{noise of size } O(\sqrt{\eta})}.$$

Assume that the noise is Gaussian ξ_t, then we have

$$x_{t+1} - x_t \approx -\eta\nabla f(x_t) + \sqrt{\eta\Sigma(x_t)}\mathcal{N}\left(0, \eta I\right)$$

When η is small, this recursion can be interpreted as the Euler–Maruyama discretization of the Itô SDE

$$\mathrm{d}X_t = -\nabla f(X_t)\,\mathrm{d}t + \sqrt{\eta\,\Sigma(X_t)}\,\mathrm{d}W_t, \tag{6}$$

where $(W_t)_{t\geq 0}$ is a standard Brownian motion. The stochastic term is of order $O(\sqrt{\eta})$, reflecting the fact that SGD becomes less noisy as the learning rate decreases.

This leads to two useful observations:

- **Vanishing noise as $\eta \to 0$.** When $\eta \to 0$, the diffusion term disappears and (6) reduces to the gradient flow ODE $\dot{X}_t = -\nabla f(X_t)$. Thus, infinitesimal steps eliminate stochasticity entirely.
- **Small but finite η.** For moderate stepsizes, the SDE (6) provides an effective continuous-time approximation to SGD dynamics. The accuracy of this modeling depends on the structure of the problem and the distribution of the gradient noise.

The SDE formulation of SGD in (6) was formalized in [Li et al., 2019] and has since become a standard tool for analyzing the dynamical behavior of SGD in modern machine learning.

2 Convergence Analysis

In this section, we analyze the discrete-time dynamics of SGD. For simplicity, we work with the general update rule (5) and denote the stochastic gradient by

$$g_t := \nabla f(x_t) + \xi_t.$$

In our analysis, we make the following assumptions about the objective function and the gradient noise.

Assumption 2.1. *The objective function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and it attains its minimum at some point x^*, so that $\inf_x f(x) = f(x^*)$. The noise $\{\xi_t\}$ are independent of x_t and satisfy*

$$\mathbb{E}[\xi_t] = 0, \qquad \sigma_t := \mathbb{E}[\|\xi_t\|^2] \leq \sigma^2 < \infty.$$

Remark 2.2. The above noise assumption is commonly used in theoretical analysis. However, In practice, two issues may arise: 1) the noise might be heavy-tailed, leading to $\mathbb{E}[\|\xi_t\|^2] = +\infty$, and 2) the noise may degenerate at the global minimum (see the homework for more details).

The following lemma provides the energy dissipation inequality for SGD, which is the starting point of our convergence analysis.

Lemma 2.3 (One-step energy dissipation)**.** *Under Assumption 2.1, if $\eta_t \leq 1/L$, then we have*

$$\mathbb{E}[f(x_{t+1})|x_t] \leq f(x_t) - \frac{\eta_t}{2}\|\nabla f(x_t)\|^2 + \frac{\eta_t^2 L\sigma^2}{2}. \tag{7}$$

Proof. Recall that $g_t = \nabla f(x_t) + \xi_t$ denotes the stochastic gradient. The smoothness implies

$$f(x_{t+1}) = f(x_t - \eta_t g_t) \leq f(x_t) + \eta_t \langle \nabla f(x_t), -\eta_t g_t \rangle + \frac{L\eta_t^2}{2}\|g_t\|^2.$$

Taking expectation and noticing $\mathbb{E}[\|g_t\|^2|x_t] = \mathbb{E}[\|\xi_t\|^2] + \|\nabla f(x_t)\|^2$, we have

$$\begin{aligned}\mathbb{E}[f(x_{t+1})|x_t] &\leq f(x_t) - \eta_t\|\nabla f(x_t)\|^2 + \frac{\eta_t^2 L}{2}\mathbb{E}[\|\xi_t\|^2] + \frac{\eta_t^2 L}{2}\|\nabla f(x_t)\|^2 \\ &\leq f(x_t) - \eta_t\left(1 - \frac{\eta_t L}{2}\right)\|\nabla f(x_t)\|^2 + \frac{\eta_t^2 L\sigma^2}{2},\end{aligned}$$

the last inequality follows from $\mathbb{E}[\|\xi_t\|^2] \leq \sigma^2$ and $\eta_t L \leq 1$. □

Theorem 2.4. *Under Assumption 2.1, we have*

$$\min_{t=0,\ldots,T} \mathbb{E}[\|\nabla f(x_t)\|^2] \le \frac{2(f(x_0) - f(x^*)) + L\sigma^2 \sum_{t=0}^{T} \eta_t^2}{\sum_{t=0}^{T} \eta_t} \tag{8}$$

Proof. Taking the expectation with respect in (7) yields

$$\mathbb{E}[f(x_{t+1})] \le \mathbb{E}[f(x_t)] - \frac{\eta_t}{2}\, \mathbb{E}\, \|\nabla f(x_t)\|^2 + \frac{\eta_t^2 L\sigma^2}{2}.$$

This is equivalent to

$$\eta_t\, \mathbb{E}\, \|\nabla f(x_t)\|^2 \le 2\left(\mathbb{E}[f(x_t)] - \mathbb{E}[f(x_{t+1})]\right) + \eta_t^2 L\sigma^2.$$

Applying telescoping sum gives

$$\begin{aligned} \frac{\sum_{t=0}^{T} \eta_t\, \mathbb{E}\, \|\nabla f(x_t)\|^2}{\sum_{t=0}^{T} \eta_t} &\le \frac{2\, \mathbb{E}[f(x_0) - f(x_{T+1})] + L\sigma^2 \sum_{t=0}^{T} \eta_t^2}{\sum_{t=0}^{T} \eta_t} \\ &\le \frac{2\big(f(x_0) - f(x^*)\big) + L\sigma^2 \sum_{t=0}^{T} \eta_t^2}{\sum_{t=0}^{T} \eta_t}. \end{aligned}$$

Noticing $\frac{\sum_{t=0}^{T} \eta_t\, \mathbb{E}\, \|\nabla f(x_t)\|^2}{\sum_{t=0}^{T} \eta_t} \ge \min_{t=0,\ldots,T} \mathbb{E}[\|\nabla f(x_t)\|^2]$, we complete the proof. □

The theorem above shows that SGD converges when an appropriately decaying learning rate is used. In particular, the bound in (8) indicates that SGD converges under the standard conditions

$$\sum_{t=1}^{\infty} \eta_t = +\infty, \qquad \lim_{T\to\infty} \frac{\sum_{t=1}^{T} \eta_t^2}{\sum_{t=1}^{T} \eta_t} = 0,$$

which balance exploration and variance reduction in the stochastic updates.

Question: Is the condition $\sum_{t=0}^{\infty} \eta_t = \infty$ also necessary?

2.1 A Convex Analysis

Theorem 2.5. *Suppose that Assumption (2.1) holds and f is convex. Let $\bar{x}_T$ be the average solution*

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

If $\eta_t \le 1/L$ for any $t \in \mathbb{N}$, then

$$\mathbb{E}[f(\bar{x}_T) - f(x^*)] \le \frac{\|x_0 - x^*\|^2 + 2\sigma^2 \sum_{t=0}^{T-1} \eta_t^2}{2\sum_{t=1}^{T} \eta_t}. \tag{9}$$

- Here we only consider the average solution $\bar{x}_T$ instead of the last-iterate solution x_T. Note that averaging has a variance-reduction effect, and as a result, the convergence of $\bar{x}_T$ is much more smooth and the corresponding analysis is also much easier. On the contrary, x_T oscillates much more significantly and the convergence analysis of x_T is more complicated.

- Considering the constant learning rate $\eta_t = \eta$, then the upper bound becomes

$$\mathbb{E}[f(\bar{x}_T) - f(x^*)] \leq \underbrace{\frac{\|x_0 - x^*\|^2}{2T\eta}}_{\text{GD decay}} + \underbrace{\eta\sigma^2}_{\text{noise effect}} . \tag{10}$$

  This upper bound indicates that the dynamics of SGD consist of two distinct phases: a **gradient-dominated phase**, in which the function value decreases at rate $O(1/(\eta T))$; and a **noise-dominated phase**, where the function value fluctuates around $O(\eta\sigma^2)$. In particular, the learning rate η determines the balance between these two phases.

- **Constant learning rate vs. learning rate decay.** If we already know the total number of iterations T, we may choose $\eta = \frac{1}{\sqrt{T}}$, and SGD achieves the optimal $O(1/\sqrt{T})$ convergence rate. This is the best rate achievable in the stochastic-gradient setting. However, this choice depends critically on knowing T in advance, which is often unrealistic in practice.

  A more practical and widely used approach is to employ a *decaying* learning rate, for example $\eta_t = \frac{1}{\sqrt{t}}$. This schedule does *not* require prior knowledge of T, and guarantees $O\big(\log t/\sqrt{t}\big)$ for every t. Thus, we pay only a mild $\log t$ factor in exchange for a method that works uniformly over time and avoids dependence on the training horizon.

  Table 1 summarizes the differences between the two learning rate schedules.

Table 1: Comparison of constant learning rate and learning rate decay in SGD.

Method	**Rate**	**Requires knowing T?**	**Uniform in t?**
Constant LR $\eta = 1/\sqrt{T}$	$O(1/\sqrt{T})$	Yes	No
LR decay $\eta_t = 1/\sqrt{t}$	$O(\log t/\sqrt{t})$	No	Yes

Proof. By the energy dissipation inequality (Lemma 2.3), we have

$$\begin{aligned}
\mathbb{E}[f(x_{t+1}) - f(x^*)] &\leq \mathbb{E}[f(x_t)] - f(x^*) - \frac{\eta_t}{2}\,\mathbb{E}\,\|\nabla f(x_t)\|^2 + \frac{\eta_t^2 L\sigma^2}{2} \\
&\leq \mathbb{E}[\langle \nabla f(x_t), x_t - x^*\rangle] - \frac{\eta_t}{2}\,\mathbb{E}\,\|\nabla f(x_t)\|^2 + \frac{\eta_t^2 L\sigma^2}{2} \\
&= -\frac{1}{2\eta_t}\big(\mathbb{E}[\|x_t - \eta_t\nabla f(x_t) - x^*\|^2 - \|x_t - x^*\|^2]\big) + \frac{\eta_t^2 L\sigma^2}{2},
\end{aligned}$$

where the second step follows from the convexity of f. Note that

$$\begin{aligned}
\mathbb{E}[\|x_{t+1} - x^*\|^2] &= \mathbb{E}[\|x_t - \eta_t\nabla f(x_t) - \eta_t\xi_t - x^*\|^2] \\
&= \mathbb{E}[\|x_t - \eta_t\nabla f(x_t) - x^*\|^2] + \eta_t^2\,\mathbb{E}[\|\xi_t\|^2] \\
&\leq \mathbb{E}[\|x_t - \eta_t\nabla f(x_t) - x^*\|^2] + \eta_t^2\sigma^2.
\end{aligned}$$

Then,

$$\mathbb{E}[f(x_{t+1}) - f(x^*)] \leq -\frac{1}{2\eta_t}\big(\mathbb{E}[\|x_{t+1} - x^*\|^2 - \|x_t - x^*\|^2]\big) + \frac{\eta_t}{2}\sigma^2 + \frac{L\eta_t^2\sigma^2}{2}$$

$$\leq -\frac{1}{2\eta_t}\left(\mathbb{E}[\|x_{t+1}-x^*\|^2 - \|x_t - x^*\|^2]\right) + \eta_t\sigma^2,$$

where we use $\eta_t L \leq 1$. Therefore,

$$\begin{aligned}
\mathbb{E}[f(\bar{x}_T) - f(x^*)] &\leq \frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\,\mathbb{E}[f(x_t) - f(x^*)] \\
&\leq \frac{1}{2\sum_{0=1}^{T-1}\eta_t}\sum_{t=0}^{T-1}\left(\mathbb{E}\,\|x_{t-1}-x^*\|^2 - \mathbb{E}\,\|x_t - x^*\|^2 + 2\eta_t^2\sigma^2\right) \\
&\leq \frac{\|x_0 - x^*\|^2 + 2\sigma^2\sum_{t=0}^{T-1}\eta_t^2}{2\sum_{t=0}^{T-1}\eta_t},
\end{aligned}$$

where the first step follows from the convexity of f. □

Comparison with GD. The convergence rate of GD for convex problem is $O(1/T)$. Therefore, SGD is slower than GD in terms of number of iteration. However, in terms of computational efficiency, SGD can outperform GD. Consider the batch size $B = 1$ and learning rate $\eta = 1/\sqrt{T}$; under this condition, the converge rate of SGD becomes $O(1/\sqrt{T})$. Consequently, to achieve an error of ϵ, SGD requires $\Omega(1/\epsilon^2)$ iterations; while GD needs only $\Omega(1/\epsilon)$ iterations. However, in terms of computation cost, SGD and GD require $\Omega(1/\epsilon^2)$ and $\Omega(n/\epsilon)$, respectively. As long as, $\epsilon \geq 1/n$, SGD is more efficient.

2.2 A PL Analysis

Theorem 2.6 (Constant learning rate). *Under Assumption (2.1), we further assume that f is μ-PL, i.e.,*

$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f(x^*)).$$

Then

$$\mathbb{E}[f(x_T)] - f(x^*) \leq \underbrace{(1-\mu\eta)^T\left(f(x_0) - f(x^*)\right)}_{\text{exponential decay}} + \underbrace{\frac{L\sigma^2}{2\mu}\eta}_{\text{noise effect}}\ .$$

We have the following observations.

- Still the SGD dynamics consists of two phases. When $f(x_t)$ is large with respect to η, the decay is exponential, and this exponential decay comes from the GD step. When $f(x_t)$ is in the same order as η, the decay induced by GD is dominated by the gradient noise. Consequently, we must reduce the learning rate if we would like to further reduce $f(x_t)$.
- Taking $\eta = \frac{2\log(T)}{\mu T}$, we obtain

$$\mathbb{E}[f(x_T)] - f(x^*) \leq O\left(\frac{1+\log T}{T}\right).$$

  This rate is faster than $O(1/\sqrt{T})$, the rate of the general convex case, but is significantly slower than the rate of GD, which is exponential.

Proof. Plugging the PL condition into the energy dissipation inequality (Lemma 2.3) leads to

$$\begin{aligned}\mathbb{E}[f(x_t)] - f(x^*) &\leq \mathbb{E}[f(x_t)] - f(x^*) - \frac{\eta}{2}\|\nabla f(x_t)\|^2 + \frac{L\sigma^2\eta^2}{2} \\ &\leq \mathbb{E}[f(x_{t-1})] - f(x^*) - \mu\eta(\mathbb{E}[f(x_{t-1})] - f(x^*)) + \frac{L\eta^2\sigma^2}{2}\end{aligned}$$

Let $e_t = \mathbb{E}[f(x_t)] - f(x^*)$. Then,

$$\begin{aligned}e_{t+1} &\leq (1-\mu\eta)e_t + \frac{L\eta^2\sigma^2}{2} \\ &\leq (1-\mu\eta)^t e_0 + \frac{L\eta^2\sigma^2}{2}\sum_{k=0}^{t}(1-\mu\eta)^{t-k} \\ &\leq (1-\mu\eta)^t e_0 + \frac{L\eta^2\sigma^2}{2}\frac{1}{1-(1-\mu\eta)} \\ &= (1-\mu\eta)^t e_0 + \frac{L\sigma^2}{2\mu}\eta.\end{aligned}$$

□

Note that setting $\eta = 1/T$ means that we need to know the number of iterations a priori. The following theorem shows that a similar convergence rate can be achieved with decaying learning rates.

Theorem 2.7 (Decay learning rate)**.** *Let f be L-smooth and satisfy the PL condition with parameter $\mu > 0$. Run SGD with the learning rate schedule:*

$$\eta_t = \min\left\{\frac{1}{L}, \frac{2}{\mu(t+1)}\right\}.$$

Then for all sufficiently large T,

$$\mathbb{E}[f(x_T)] - f(x^*) \;\leq\; \frac{2L\sigma^2}{\mu^2}\frac{1}{T}.$$

Proof. Let $e_t = \mathbb{E}[f(x_t)] - f(x^*)$. By Lemma 2.3 (one-step descent) and the PL inequality $\|\nabla f(x_t)\|^2 \geq 2\mu(f(x_t) - f(x^*))$, we have

$$e_{t+1} \;\leq\; (1-\mu\eta_t)\, e_t + \frac{L\sigma^2}{2}\eta_t^2 \leq \left(1 - \frac{\mu\eta_t}{2}\right)^2 e_t + \frac{L\sigma^2}{2}\eta_t^2.$$

For sufficiently large t, the stepsize is $\eta_t = \frac{2}{\mu(t+1)}$, hence

$$e_{t+1} \;\leq\; \frac{t^2}{(t+1)^2}e_t + \frac{2L\sigma^2}{\mu^2(t+1)^2}.$$

Multiply both sides by $(t+1)^2$ and let $a_t := t^2 e_t$. We obtain

$$a_{t+1} \;\leq\; a_t + \frac{2L\sigma^2}{\mu^2}.$$

Iterating from 1 to t yields

$$a_t \;\le\; a_1 + \frac{2L\sigma^2}{\mu^2}(t-1) \;\le\; \frac{2L\sigma^2}{\mu^2}t \qquad \text{(absorbing } a_1 \text{ into the constant)}.$$

Since $e_t = a_t/t^2$, we obtain

$$e_t \;\le\; \frac{2L\sigma^2}{\mu^2}\cdot\frac{1}{t},$$

which completes the proof. □

3 Stochastic Approximation

First, let us briefly summarize key insights we gain from Analyzing the convergence of SGD:

- Stochastic approximation enables us to reduce the computational cost per iteration.
- The impact of approximation noise can be mitigated by decaying the learning rate appropriately to ensure convergence.

These two insights form the foundation of stochastic approximation [Robbins and Monro, 1951], a concept that extends beyond SGD and can be applied in various contexts.

Stochastic Approximation (SA). Consider a general iteration

$$x_{t+1} = \bar{G}(x_t) = \mathbb{E}_{w\sim\pi_t}[G(x_t; w)]. \tag{11}$$

This iteration shall converges to the fixed point satisfying $x^* = \bar{G}(x^*)$. The stochastic approximation to (11) is obtained by replacing the exact expectation with a minibatch of B independent samples:

$$\begin{aligned} &w_{1,t}, \ldots, w_{B,t} \overset{iid}{\sim} \pi_t, \\ &x_{t+1} = (1-\alpha_t)x_t + \alpha_t \frac{1}{B}\sum_{j=1}^{B} G(x_t; w_{j,t}), \end{aligned} \tag{12}$$

where $\alpha_t \in (0,1)$ is a stepsize parameter that typically decreases to zero. A particularly important special case is the single-sample update ($B=1$):

$$x_{t+1} = (1-\alpha_t)x_t + \alpha_t\, G(x_t; w_t), \qquad w_t \sim \pi_t. \tag{13}$$

The key idea in stochastic approximation is to use a convex combination in the update so as to smooth out the noise in the stochastic estimate. By gradually decreasing the stepsize $\alpha_t \to 0$, the influence of the noise is reduced over time, allowing the iterates to stabilize and converge to the desired solution.

Remark 3.1. In the stochastic update (11), the sampling distribution π_t may itself evolve over the course of the iterations. This allows the approximation to adapt dynamically as the algorithm progresses.

Stochastic Gradient Descent. SGD can be viewed as a stochastic-approximation version of GD. Consider the GD iteration

$$x_{t+1} = x_t - \eta \nabla f(x_t), \qquad f(x) = \mathbb{E}_{w\sim\pi}[\, f(x; w)\,]. \tag{14}$$

Applying the SA update (12) to (14), we replace the true gradient $\nabla f(x_t)$ by a minibatch estimate: $\nabla f(x_t) \approx \frac{1}{B}\sum_{j=1}^{B} \nabla f(x_t; w_{j,t})$. Plugging this estimator into the SA form,

$$\begin{aligned} x_{t+1} &= (1-\alpha_t)x_t + \alpha_t \left(x_t - \eta\, \frac{1}{B}\sum_{j=1}^{B} \nabla f(x_t; w_{j,t}) \right) \\ &= x_t - \alpha_t \eta \cdot \frac{1}{B}\sum_{j=1}^{B} \nabla f(x_t; w_{j,t}). \end{aligned}$$

Thus we recover exactly the SGD update with learning rate

$$\eta_t = \alpha_t\, \eta.$$

The following theorems shows that when G is contractive, we have that x_t converges to the fixed point in a rate of $O(1/t)$.

Theorem 3.2. *Consider the stochastic approximation* (13) *and let* $\eta_t = \frac{1}{(1-\alpha)(t+1)}$ *and* $B = 1$. *If there exists a* $\alpha \in (0,1)$ *such that* $\|G(x) - G(x')\| \le \alpha \|x - x'\|$ *and* $\mathbb{E}_w[\|G(x; w) - \bar{G}(x)\|^2] \le \sigma^2$. *Then, we have*

$$\mathbb{E}[\|x_T - x^*\|^2] \le \frac{\sigma^2}{(1-\alpha)^2 T}.$$

Proof. By definition,

$$\begin{aligned} x_{t+1} - x^* &= (1-\eta_t)(x_t - x^*) + \eta_t (G(x_t; w_t) - x^*) \\ &= (1-\eta_t)(x_t - x^*) + \eta_t (G(x_t) - G(x^*) + \xi_t). \end{aligned}$$

Let $\Delta_t = \|x_t - x^*\|$, we have

$$\begin{aligned} \mathbb{E}[\Delta_{t+1}^2] &\le \mathbb{E}[(1-\eta_t)^2 \Delta_t^2 + 2(1-\eta_t)\eta_t \alpha \Delta_t^2 + \eta_t^2 \alpha^2 \Delta_t^2] + \eta_t^2 \sigma^2 \\ &= (1 - (1-\alpha)\eta_t)^2\, \mathbb{E}[\Delta_t^2] + \eta_t^2 \sigma^2. \end{aligned}$$

Then, we can complete the proof by following the same argument as the proof of Theorem 2.7. □

Latent-Variable Maximum Likelihood. Consider the latent-variable model

$$\max_{\theta}\ L(\theta) := \log \int p(x, z; \theta)\ \mathrm{d}z, \tag{15}$$

where, for simplicity, we assume that only a single observation x is available. There are two classical stochastic approaches to solving (15): the EM algorithm and stochastic gradient descent.

Stochastic EM. The EM algorithm iteratively maximizes the surrogate function

$$Q(\theta \mid \theta_t) := \mathbb{E}_{z|x,\theta_t}[\log p(x,z;\theta)]\,,$$

which leads to the update

$$\theta_{t+1} = \operatorname*{argmax}_{\theta} Q(\theta \mid \theta_t).$$

A stochastic approximation of EM is obtained by replacing the expectation with a minibatch version:

$$\begin{aligned} z_{1,t}, z_{2,t}, \ldots, z_{B,t} &\overset{iid}{\sim} p(\cdot \mid x, \theta_t), \\ \theta_{t+1} &= (1-\alpha_t)\theta_t + \alpha_t \ \operatorname*{argmax}_{\theta} \frac{1}{B}\sum_{j=1}^{B} \log p(x, z_{j,t}; \theta), \end{aligned} \tag{16}$$

where $\alpha_t \in (0,1)$ decreases to zero. This update can be viewed as a noisy version of the exact EM step.

Stochastic Gradient Descent via the Log-Derivative Trick. An alternative approach is to compute (or approximate) the gradient of $L(\theta)$ directly. Using the log-derivative trick,

$$\nabla L(\theta) = \frac{\int \nabla p(x,z;\theta)\ \mathrm{d}z}{\int p(x,z;\theta)\ \mathrm{d}z} = \frac{\int p(x,z;\theta)\nabla \log p(x,z;\theta)\ \mathrm{d}z}{p(x;\theta)} \tag{17}$$

$$= \mathbb{E}_{z|x,\theta}\left[\nabla \log p(x,z;\theta)\right]. \tag{18}$$

This expresses the gradient as an expectation over the posterior of z, allowing unbiased stochastic estimates. Drawing a single latent sample gives the SGD update

$$\begin{aligned} z_{1,t}, z_{2,t}, \ldots, z_{B,t} &\overset{iid}{\sim} p(\cdot \mid x, \theta_t), \\ \theta_{t+1} &= \theta_t + \eta_t \frac{1}{B}\sum_{j=1}^{B} \nabla \log p(x, z_{j,t} \mid \theta_t). \end{aligned}$$

Both stochastic EM and SGD therefore solve the same latent-variable maximum-likelihood problem using different stochastic approximations: EM performs a noisy maximization of a surrogate expectation, while SGD performs noisy ascent on the true marginal likelihood.

References

[Li et al., 2019] Li, Q., Tai, C., and Weinan, E. (2019). Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *The Journal of Machine Learning Research*, 20(1):1474–1520.

[Robbins and Monro, 1951] Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.