Mathematical Introduction to Machine Learning

Lecture 6: Stochastic Gradient Descent

November 14, 2024

Lecturer: Lei Wu *Scribe: Lei Wu*

1 Problem Setup

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

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

For GD, the cost of computing gradient in each step is $O(n)$, which is extremely expansive when n is large, e.g., $n = 10^6$. Stochastic gradient descent is proposed to resolve this issue.

To start, we consider a more general form, where the objective function admits an expectation

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

For the empirical risk, $\pi = \text{Unif}([n])$. The GD of optimizing (2) is given by

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

Stochastic gradient descent (SGD) iterates as follows

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

where $\{w_{1,t}, \dots, w_{B,t}\}$ are i.i.d. samples drawn from π. Here, B is a crucial hyperparameter and often referred to as the batch size. The optimizer (4) is called (mini-batch) SGD.

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 stochastic gradient is accurate; we can use a large learning rate?
  - When B is small, the stochastic gradient is far from being accurate, and a small learning rate should be used.
- Does SGD converge when B is a constant, i.e.g., $B = 1$?

To understand these questions, it is more intuitive to rewrite (4) in the following form

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

where ξ_t is the noise induced by minibatch gradient, satisfying

$$\mathbb{E}[\xi_t] = 0$$
$$\mathbb{E}[\xi_t \xi_t^\top] = \frac{1}{B} \mathbb{E}_w[(\nabla f(x_t; w) - \nabla f(x_t))(\nabla f(x_t; w) - \nabla f(x_t))^\top] =: \frac{1}{B} \Sigma(x_t)$$

Remark 1.1. Comparing with (4), the formulation (5) is more general. One also generally refer (5) as SGD as long as the gradient noise ξ_t has a zero mean. One often refers (4) as mini-batch SGD for clarity.

A phenomenological comparison between SGD and GD In Figure 1, we provide a visual comparison between GD and SGD (with small and large batch sizes). It is not surprising that the trajectory of SGD exhibits more fluctuation, although it does converge to global minima.

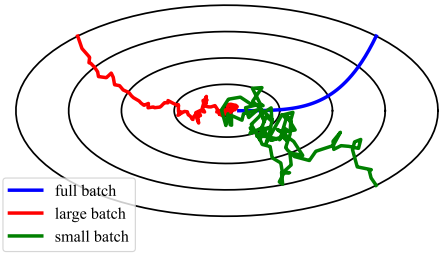

Figure 1: A visual illustration of how batch size affects SGD convergence.

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

$$\min_w \frac{1}{n} \sum_{i=1}^{n} (w^\top x_i - y_i)^2,$$

where $n = 200, x_i \overset{iid}{\sim} \mathcal{N}(0, A)$. 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:

- In terms of number of epochs, SGD converges faster than GD.
- It is hard for SGD to reach high precision regime because of the noise.
- The convergence process consists of two phases: In the first phase, where $\|\nabla f(x_t)\| \gg \|\xi_t\|$, the objective value decreases significantly. In the second phase, the noise dominates, and SGD no longer converges. To further reduce the objective value, it may be necessary to decay the learning rate.

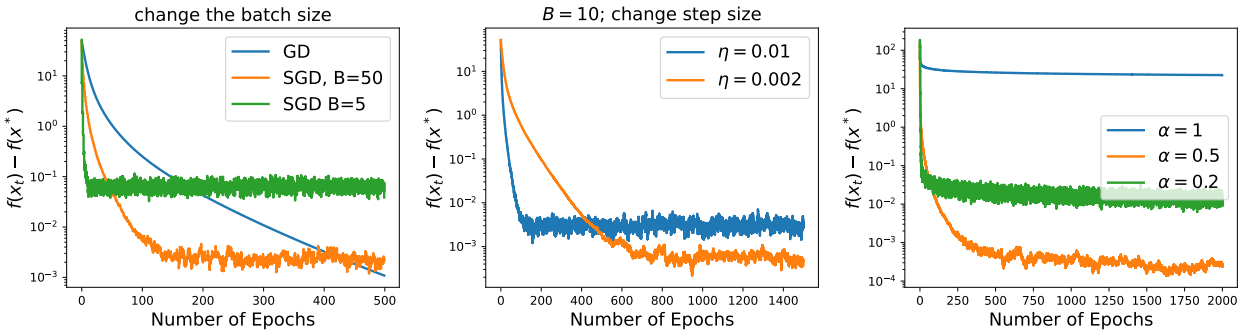

Figure 2: A visual comparison between SGD and GD. **Left:** Change the learning rate. **Middle:** Change the batch size. **Right:** Decay learning rate with $\eta_t = \eta_0/(t+1)^\alpha$.

Remark 1.2. The minibatch gradient can be viewed as a stochastic approximation of the full-batch gradient. It is similar to the Monte Carlo approximation, but with a key distinction: the convergence of SGD can be guaranteed even when the batch size B is a constant, e.g., $B = 1$. We will come back to this issue in Section 4.

2 Convergence analysis

Different from GD, SGD does not have a clear continuous-time limit. The analysis in this section will focus on the discrete-time case. For brevity, we consider the general form (5) and let $g_t = \nabla f(x_t) + \xi_t$ be the stochastic gradient.

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

Assumption 2.1. *Suppose that $f \in C^1(\mathbb{R})$ is L-smooth. We also assume that the gradient noise ξ_t and x_t are independent, and $\sigma_t := \mathbb{E}[\|\xi_t\|^2] \le \sigma^2 < \infty$.*

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 $\eta_t \le 1/L$, then we have*

$$\mathbb{E}[f(x_{t+1})|x_t] \le f(x_t) - \frac{\eta_t}{2}\|\nabla f(x_t)\|^2 + \frac{\eta_t^2 L\sigma^2}{2}. \tag{6}$$

Proof. The smoothness implies

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

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

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

$$\leq f(x_t) - \eta_t(1 - \frac{\eta_t L}{2})\|\nabla f(x_t)\|^2 + \frac{\eta_t^2 L\sigma^2}{2},$$

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

From this energy dissipation inequality, we can observe that if $\inf_t \sigma_t > 0$, we must set $\eta_t \to 0$ for convergence.

Theorem 2.4. *Suppose learning rates satisfy the Robbins-Monro condition [Robbins and Monro, 1951]:*

$$\sum_t \eta_t = \infty, \qquad \sum_t \eta_t^2 < \infty. \tag{7}$$

Then, we have

$$\min_{t=0,1\ldots,T} \mathbb{E}\,\|\nabla f(x_t)\|^2 \to 0, \ \textit{as} \ \ T \to \infty.$$

Proof. Applying telescoping sum to (6) gives

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

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

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

2.1 A convex analysis

The convergence of SGD is similar as stated in the following theorem.

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

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

- 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.
- The Robins-Monro condition is very weak condition that is sufficient to ensure that $f(\bar{x}_T) \to f(x^*)$ as $T \to \infty$.

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

$$\mathbb{E}[f(\bar{x}_T) - f(x^*)] \leq \underbrace{\frac{\|x_0 - x^*\|^2}{2T\eta}}_{\text{GD decay}} + \underbrace{\eta\sigma^2}_{\text{noise effect}} . \tag{8}$$

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

- Taking the constant learning rate $\eta = 1/\sqrt{T}$ yields the overall rate $O(1/\sqrt{T})$. This, however, needs to know T a priori. Considering $\eta_t = 1/\sqrt{t}$, we obtain the rate $O(\log T/\sqrt{T})$ without needing to know T.

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

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

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,

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

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

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

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

Comparison with GD. The convergence rate of GD 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

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

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,

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

$$= (1 - \mu\eta)^t e_0 + \frac{L\sigma^2}{2\mu}\eta.$$

□

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)**.** *Choosing the learning rate* $\eta_t = \frac{1}{\mu(t+1)}$*, we have*

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

Proof. Still let $e_t = \mathbb{E}[f(x_t)] - f(x^*)$. Then, by Lemma (2.3) and the PL condition, we have

$$e_{t+1} \leq (1 - \mu\eta_t)\, e_t + \frac{\eta_t^2 L\sigma^2}{2}. \tag{9}$$

Plugging $\eta_t = 1/(\mu(t+1))$ yields,

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

Let $\tilde{e}_t = te_t$. Then,

$$\tilde{e}_{t+1} \leq \tilde{e}_t + \frac{L\sigma^2}{2\mu^2}\frac{1}{1+t}.$$

By telescoping sum, we have

$$\tilde{e}_T \leq \tilde{e}_0 + \frac{L\sigma^2}{2\mu^2}\sum_{t=0}^{T-1}\frac{1}{1+t} \leq \tilde{e}_0 + \frac{L\sigma^2}{2\mu^2}\log(1+T).$$

Noticing that $e_T = \tilde{e}_T/T$, we complete the proof. □

Remark 2.8. The $\log T$ factors in above theorem can be removed by a refined analysis.

Summary. We summarize the implications of the above analysis as follows.

- SGD can converge by reducing learning rates.
- SGD convergences slower than GD in terms of number of iterations: $O(1/T)$ vs $O(1/\sqrt{T})$ for convex problems; $O(e^{-T})$ vs. $O(1/T)$ for PL problems.
- Figure 2 shows SGD actually converges faster in terms of number of epochs. Can we establish theoretical foundations for this phenomenon?
- Typically, SGD slows down the training only in the late training phase.

3 Continuous-time Limit?

When η is small, the continuous-time courterpart of SGD is the following Ito-type stochastic differential equation (SDE):

$$\mathrm{d}X_t = -\nabla f(X_t) + \sqrt{2\eta\Sigma(X_t)}\,\mathrm{d}W_t, \tag{10}$$

where $(W_t)_{t\geq 0}$ is the Brownian motion and $\Sigma(X_t) = \mathbb{E}[\xi_t\xi_t^T]$ is the covariance of gradient noise. For mini-batch SGD,

$$\Sigma(x) = \mathbb{E}_w\left[(\nabla f(x;w) - \nabla f(x))(\nabla f(x;w) - \nabla f(x))^T\right].$$

Note that the stochastic term is $O(\sqrt{\eta})$. We have the following observation

- When $\eta \to 0$, SGD converges to gradient flow. No stochasticity!!!
- When η is finite but small, the stochasticity can be modeled with Brownian motion. However, whether this modeling is accurate or not highly depends on the problem.
- The closeness between SGD and SDE (10) only holds for a finite time. Their long-time behaviors can be very different. We refer interested readers to [Li et al., 2019].

Remark 3.1. Currently, many works analyze the dynamical property of SDE (10) for training machine learning models. However, whether the results can be generalized to SGD (in particular with large LR) or not is still questionable.

4 Stochastic Approximation

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

- Stochastic approximation enables us to reduce the computational cost per iteration.
- The impact of approximation noise can be mitigated by adjusting the learning rate at an appropriate decay rate (following the Robins-Monro condition) 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

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

The stochastic approximation is given by

$$\begin{aligned} &w_{1,t}, w_{2,t}, \ldots, w_{B,t} \overset{iid}{\sim} \pi_t \\ &x_{t+1} = (1-\eta_t)x_t + \eta_t \frac{1}{B}\sum_{j=1}^{B} G(x_t; w_{j,t}), \end{aligned} \tag{12}$$

The key idea in stochastic approximation is the introduction of a *convex combination* to mitigate the impact of noise in the stochastic estimate. During the iteration, η_t is gradually reduced to zero, helping to diminish the influence of noise and ensuring convergence to the desired solution.

In particular, when $B = 1$,

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

Compared with SGD, the iteration (11) is more general and SGD is a special case of (13). Let $G(x) = x - \alpha\nabla f(x)$ with $f(x) = \mathbb{E}_{w\sim\pi}[f(x; w)]$. Then,

$$\begin{aligned} x_{t+1} &= (1 - \eta_t)x_t + \eta_t G(x_t; w_t) = (1 - \eta_t)x_t + \eta_t(x_t - \alpha\nabla f(x_t; w_t)) \\ &= x_t - \alpha\eta_t\nabla f(x_t; w_t), \end{aligned}$$

which recovers SGD. Moreover, π_t is not necessary fixed for different t's.

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 4.1. *Consider the stochastic approximation* (13) *and let* $\eta_t = \frac{1}{(1-\alpha)(t+1)}$. *If there exists a* $\alpha \in (0, 1)$ *such that* $\|G(x) - G(x')\| \le \alpha\|x - x'\|$, *then. Then, we have*

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

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

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

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

Stochastic EM. Consider the problem of learning a latent variable model:

$$\max L(\theta) := \log \int p(x, z; \theta)\,\mathrm{d}z. \tag{14}$$

The EM iteration is

$$\theta_{t+1} = \operatorname{argmax} Q(\theta|\theta_t) = \operatorname{argmax} \mathbb{E}_{z|x,\theta_t}[\log p(x, z|\theta)],$$

where the right-hand side is an expectation. The stochastic approximation is given by

$$\begin{aligned} z_t &\sim p(\cdot|x, \theta_t) \\ \theta_{t+1} &= (1 - \eta_t)\theta_t + \eta_t \operatorname*{argmax}_\theta \log p(x, z_t|\theta). \end{aligned}$$

For each step, the output is a convex combination between the last-step solution and the current estimate.

The Log Derivative Trick. Still consider the optimization problem (14). But this time, we consider SGD to solve it. First,

$$\begin{aligned}\nabla L(\theta) &= \frac{\int \nabla p(x,z;\theta)\,\mathrm{d}z}{\int p(x,z;\theta)\,\mathrm{d}z} \\ &= \frac{\int p(x,z;\theta)\nabla \log p(x,z;\theta)\,\mathrm{d}z}{\int p(x,z;\theta)\,\mathrm{d}z} \\ &= \frac{\int p(z|x;\theta)p(x;\theta)\nabla \log p(x,z;\theta)\,\mathrm{d}z}{p(x;\theta)} \\ &= \mathbb{E}_{z|x,\theta}[\nabla \log p(x,z|\theta)], \end{aligned} \tag{15}$$

where the second step is called the log derivative trick. *The trick formulates the derivative of marginal likelihood in an expectation form*, facilitating the gradient's estimation. Then, we can solve (14) using the following SGD:

$$\begin{aligned} z_t &\sim p(z|x,\theta_t) \\ \theta_{t+1} &= \theta_t - \eta_t \nabla \log p(x,z_t|\theta_t). \end{aligned}$$

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.