Mathematical Introduction to Machine Learning

Lecture 6: Stochastic Gradient Descent

November 10, 2023

Lecturer: Lei Wu *Scribe: Lei Wu*

1 Problem Setup

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

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

To minimize this problem with GD, the computation cost of gradient in each step is $O(n)$, which is extremely expansive when n is large, e.g., $n = 10^6$. To speed up the training, one often apply a stochastic approximation to GD. In general, consider the objective function of expectation form:

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

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

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

where $\{w_{1,t},\dots,w_{B,t}\}$ are (nearly) i.i.d. samples drawn from π. We often refer B as the batch size. This optimizer is called (mini-batch) SGD. More generally, SGD iterates by

$$x_{t+1} = x_t - \eta_t g_t,$$

where g_t is an unbiased estimate of $\mathbb{E}_w[\nabla f(x_t;w)]$.

Then the natural questions are:

- What is the difference between GD and SGD?
- What is the trade-off between B and η?
  - 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 with $B = 1$?

It is also common to rewrite (4) in the following

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

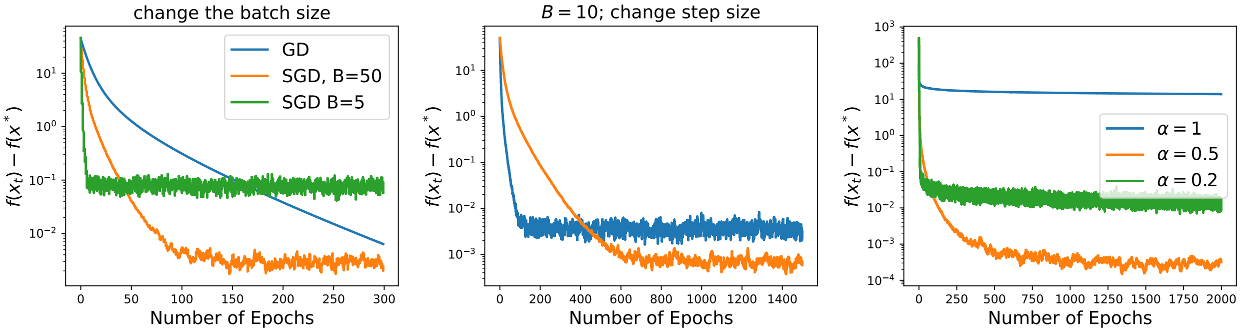
where ξ_t is the error of staochastic approximation, satisfying

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

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 Figure 1 shows the comparison between GD and SGD for training a logistic regression problem.

- In terms of number of steps, SGD converge slower than GD but for each step, the computation cost of SGD is much lower than GD.
- It is hard for SGD to reach high precision regime because of the noise.


change the batch size

$B = 10$; change step size

Figure 1: A visual comparison between SGD and GD. The objective function $f(x) = \frac{1}{2}x^T Ax$, where $x \in \mathbb{R}^{100}$ and $A = HH^T$ with H is randomly sampled by $H_{i,j} \overset{iid}{\sim} \mathcal{N}(0, 1)$. **Left:** Change the learning rate. **Middle:** Change the batch size. **Right:** Decay learning rate with $\eta_t = \eta_0/(t + 1)^\alpha$.

In Figure 1, the term 'epoch' denotes a single pass through the entire dataset. For Gradient Descent (GD), each epoch corresponds to a single iteration. However, for Stochastic Gradient Descent (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.

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

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

Lemma 2.2 (Energy dissipation)**.** *Under Assumption 2.1, if $\eta_t \leq 1/L$, then we have*

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

Proof. The smoothness implies

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

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

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

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

From this energy dissipation inequality, we have the following observation.

- If $\inf_t \sigma_t > 0$, we must set $\eta_t \to 0$ for convergence.
- Note that

$$\mathbb{E}[f(x_0)] - \mathbb{E}[f(x_\infty)] \geq \frac{1}{2}\sum_{t=0}^{\infty}\eta_t\,\mathbb{E}\,\|\nabla f(x_t)\|^2 - \frac{L\sigma^2}{2}\sum_{t=0}^{\infty}\eta_t^2. \tag{7}$$

  To ensure the decrease of energy of any finite amount, learning rates should satisfy the following Robbins-Monro condition [Robbins and Monro, 1951]:

$$\sum_t \eta_t = \infty, \qquad \sum_t \eta_t^2 < \infty. \tag{8}$$

  A typical example satisfying the above condition is $\eta_t = \alpha/t$.

Theorem 2.3. *Suppose the learning rates satisfies the Robbins-Monro condition, we have*

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

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. □

2.1 A convex analysis

Here we first provide the discrete-time analysis of GD.

Theorem 2.4. *Assume that $f \in C^1(\mathbb{R})$ is L-smooth and convex. If the learning rate $\eta \leq 1/L$, then the GD solution satisfies*

$$f(x_T) - \inf f \leq \frac{d^2(x_0, S_f)}{2T\eta}.$$

Proof. By the convexity, we have for any $x^* \in S_f$ that

$$f(x_t) \leq f(x^*) + \langle \nabla f(x_t), x_t - x^* \rangle . \tag{9}$$

When $\eta \leq 1/L$, the energy dissipation satisfies

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

Combining with (9), we have

$$\begin{aligned} f(x_{t+1}) - f(x^*) &\leq \langle \nabla f(x_t), x_t - x^* \rangle - \frac{\eta_t}{2} \|\nabla f(x_t)\|^2 \\ &\leq -\frac{1}{2\eta} \left(\|x_t - \eta \nabla f(x_t) - x^*\|^2 - \|x_t - x^*\|^2 \right) \\ &= -\frac{1}{2\eta} \left(\|x_{t+1} - x^*\|^2 - \|x_t - x^*\|^2 \right) . \end{aligned} \tag{10}$$

Hence,

$$\begin{aligned} f(x_T) - f(x^*) \leq \frac{1}{T} \sum_{t=0}^{T} (f(x_t) - f(x^*)) &\leq \frac{1}{T} \sum_{t=1}^{T} -\frac{1}{2\eta} \left(\|x_t - x^*\|^2 - \|x_{t-1} - x^*\|^2 \right) \\ &\leq \frac{1}{2\eta T} (\|x_0 - x^*\|^2 - \|x_T - x^*\|^2) \leq \frac{\|x_0 - x^*\|^2}{2\eta T}. \end{aligned}$$

□

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 bound becomes

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

  This bound shows a clear trade-off when tuning the learning rate η.
- 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.
- The convergence rate of GD is $O(1/T)$. Therefore, SGD is slower than GD.

Proof. By the energy dissipation inequality (Lemma 2.2), we have

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

Analogous to (10), we have

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

Note that

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

Then,

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

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

$$\mathbb{E}[f(\bar{x}_T) - f(x^*)] \le \frac{1}{\sum_{t=1}^T \eta_t}\sum_{t=1}^{T}\eta_t\,\mathbb{E}[f(x_t) - f(x^*)]$$

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

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

Question: The above analysis measure the convergence with $\bar{x}_T$. A natural question is can prove the convergence of the last iterate:

- Does $f(x_T) \to f(x^*)$ as $T \to \infty$?
- Does the convergence of $f(x_T)$ share the same rate as $f(\bar{x}_T)$?

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.,*

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

Then

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

We have the following observations.

- 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 2.2) leads to

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

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

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

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

□

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)**.** *Choosing the learning rate* $\eta_t = \frac{1}{2\mu(t+1)}$*, we have*

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

Proof. Let $\mathbb{E}[f(x_t)] - f(x^*)$. Then, by Lemma (2.2) and the PL condition, we have

$$
e_{t+1} \leq \left(1-\frac{\mu\eta_t}{2}\right)e_t + \frac{\eta_t^2 L\sigma^2}{2}. \tag{11}
$$

Plugging $\eta_t = 2/(\mu(t+1))$ yields,

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

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

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

By telescoping sum, we have

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

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 1 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):

$$\mathrm{d}X_t = -\nabla f(X_t) + \sqrt{2\eta\Sigma(X_t)}\,\mathrm{d}W_t, \tag{12}$$

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 Browninion motion. However, whether this modeling is accurate or not highly depends on the problem.
- The closeness between SGD and SDE (12) 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 (12) 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

Consider a general iteration

$$x_{t+1} = G(x_t) = \mathbb{E}_{w\sim\pi_t}[G(x_t;w)]. \tag{13}$$

The stochastic approximation is given by

$$\begin{aligned} &w_{1,t}, w_{2,t}, \dots, w_{B,t} \overset{iid}{\sim} \pi_t \\ &x_{t+1} = (1-\eta_t)x_t + \eta_t\frac{1}{B}\sum_{j=1}^{B} G(x_t;w_{j,t}), \end{aligned} \tag{14}$$

Here, we introuce a convex combination to increase the contribution of current estimate, since the stochastic estimate is not fully trustable. In particular, when $B=1$,

$$x_{t+1} = (1-\eta_t)x_t + \eta_t G(x_t;w_t), \tag{15}$$

with $w_t \sim \pi_t$.

Note that the iteration (13) is not necessary a gradient system. Compared with SGD,

- The iteration (13) is more general and SGD is a special case of (15). 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 recover the SGD.

- π_t is not necessary fixed for different t's.

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

Theorem 4.1. *Consider the stochastic approximation* (15) *and let* $\eta_t = \frac{1}{(1-\alpha)t}$*. 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:

$$\max L(\theta) := \log \int p(x, z; \theta)\, \mathrm{d}z. \tag{16}$$

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

$$\theta_{t+1} = (1-\eta_t)\theta_t + \eta_t \operatorname*{argmax}_\theta \log p(x, z_t|\theta),$$

where $z_t \sim p(\cdot|x, \theta_t)$. For each step, the output is a convex combination between the last-step solution and the current estimate. The convergence of this method is left to homework.

The Log Derivative Trick. Still consider the optimization problem (16). But this time, we consider SGD to solve it. First,

$$\begin{aligned} \nabla L(\theta) &= \frac{\int \nabla p(x, z; \theta)\, \mathrm{d}z}{\int p(x, z; \theta)\, \mathrm{d}z} \\ &= \frac{\int p(x, z; \theta)\nabla \log p(x, z; \theta)\, \mathrm{d}z}{\int p(x, z; \theta)\, \mathrm{d}z} \\ &= \mathbb{E}_{z|x,\theta}[\nabla \log p(x, z|\theta)], \end{aligned} \tag{17}$$

where the second step is called the log derivative trick. The trick formulates the derivative in an expectation form, facilitating the gradient's estimation. Then, we can solve (16) with 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.