1 Stochastic Optimization

In today's lecture we will study the stochastic gradient descent, a widely used algorithm which is the backbone of the success of machine learning.

A typical machine learning problem can be written as the following:

$$\min_w F(w) = \mathbb{E}_z[f(w; z)].$$

Here, we are again faced with an optimization problem. The goal is to minimize some function $F(x)$, which we assume to be convex. However, the difference compared to previous settings is that in typical machine learning scenarios, we do not actually know the function $F(\cdot)$, nor can we evaluate its gradients. However, we often can get some estimates of the function values and gradients, and we want to understand how we can use these estimates to solve the problem.

Why does this come up? For now, let's take a brief detour. Consider the standard machine learning setup, described below. Actually, many examples we have already talked about in the course are examples of machine learning tasks.

- We have some instance space $\mathcal{X}$ and some label space $\mathcal{Y}$. For example, $\mathcal{X}$ can be the space of all images, and $\mathcal{Y}$ can represent whether there is a dog in the picture or not. Another example we can use is $\mathcal{X}$ is some information about a patient, and $\mathcal{Y}$ can be some prediction about the patient, like their blood sugar level. We can even have more complicated settings. Let $\mathcal{X}$ be some sequence of words, and $\mathcal{Y}$ be a new word. Then our task here is to predict the next word in a sequence.
- There is some distribution in nature $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ pairs.
- As a machine learning practitioner, you get examples (x, y) which are drawn from the distribution. We assume that the examples are identically and independently distributed.
- You are also given a hypothesis, or predictor class $\mathcal{F} \subseteq (\mathcal{X} \to \mathcal{Y})$, which is a collection of mappings from instances to labels. For example, $\mathcal{F}$ can be a linear function class, or even a deep neural network.
- Lastly, there is also a loss function $\ell(f, (x, y))$ which measures the predictive power or performance of f. Some examples are $\ell(f, (x, y)) = (f(x) - y)^2$ (the square loss), or $\ell(f, (x, y)) = \max\{0, 1 - f(x)y\}$ (hinge loss). Or we can consider the cross entropy loss $-y \cdot \log f(x) - (1 - y) \cdot \log(1 - f(x))$.
- The goal of machine learning is to minimize the expected loss:

  $$\min_{f \in \mathcal{F}} L(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f, (x, y))].$$

  It is easy to see that this is exactly the previous stochastic optimization problem, but with different notation. So it is beneficial to understand how to solve this general stochastic optimization problem.
- A popular strategy for solving this is empirical risk minimization, which solves the following optimization problem.

  $$\min_{f \in \mathcal{F}} \widehat{L}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f, (x_i, y_i)).$$

The key idea in SGD is to use *estimates* of $F(\cdot)$ in order to approximately solve the optimization problem. In particular, we will adopt the following abstraction:

- In every iteration, we get *unbiased* estimates of the true gradient. That is, we will get a random estimate g_t such that $\mathbb{E}[g_t|w_t] = \nabla f(w_t)$.

In machine learning, this assumption is (roughly) satisfied. We usually get a dataset of i.i.d. pairs (x, y) and we can evaluate the loss or the gradient of the loss $\ell(f, (x, y)$. For empirical risk minimization, we can sample a random index $i \in [n]$ in every round and evaluate the gradient at that round.

However, if we wanted to do gradient descent on the full dataset, this would require us to iterate through all n datapoints before taking a step. This can be potentially wasteful, because we are waiting to evaluate n gradients before doing any optimization.

A subtle point is that if we iterate through the same dataset multiple times, as is common in practice called multi-pass SGD, it turns out that the estimates will not be independent of each other for the stochastic optimization problem! Usually, this is not that big of a problem, but it is something to keep in mind.

1.1 SGD for Convex Lipschitz Bounded

In stochastic gradient descent, we will instead make the update:

$$w_{t+1} = w_t - \eta_t g_t.$$

We provide some intuition on the benefits of SGD for ERM:

- Practical data usually has a lot of redundancy, so waiting to evaluate all the gradients before taking a step could be inefficient.
- In contrast, SGD will be extremely efficient at the beginning of optimization, as it gets fast initial performance with low cost-per-iteration.

We will analyze the performance of this algorithm. First we will study convex Lipschitz bounded functions.

Theorem 1. *Let F be a convex function with minimizer $w^\star = \operatorname{argmin}_{w \in B_2(B)} F(w)$. Suppose we run SGD for T iterations with constant step size η. Also assume that $\|g_t\|_2 \leq G$ with probability 1. Then we have*

$$\mathbb{E}[F(\bar{w})] - F(w^\star) \leq \frac{BG}{\sqrt{T}}.$$

Proof. We will use $g_{1:t}$ to denote the sequence of vectors $g_1, \cdots, g_t$.

By Jensen's inequality we have

$$\mathbb{E}_{g_{1:T}}[F(\bar{w}) - F(w^\star)] \leq \mathbb{E}_{g_{1:T}}\left[\frac{1}{T}\sum_{t=1}^{T} F(w_t) - F(w^\star)\right].$$

Now we will prove a guarantee which holds for any $g_{1:T}$.

Lemma 1. *For any sequence of vectors $g_{1:T}$ and vector $w^\star$, the OGD algorithm with initialization $w_1 = 0$ satisfies*

$$\sum_{t=1}^{T} \langle w_t - w^\star, g_t \rangle \leq \frac{\|w^\star\|}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|g_t\|^2.$$

In particular if $\|g_t\| \leq G$ and $\|w^\star\| \leq B$ by setting $\eta = \sqrt{\frac{B^2}{G^2 T}}$ we have

$$\sum_{t=1}^{T} \langle w_t - w^\star, g_t \rangle \leq \frac{BG}{\sqrt{T}}.$$

Using the lemma, we get that

$$\mathbb{E}_{g_{1:T}} \left[\sum_{t=1}^{T} \langle w_t - w^\star, g_t \rangle \right] \leq \frac{BG}{\sqrt{T}}$$

So it is left to show that

$$\mathbb{E}_{g_{1:T}} \left[\frac{1}{T} \sum_{t=1}^{T} F(w_t) - F(w^\star) \right] \leq \mathbb{E}_{g_{1:T}} \left[\sum_{t=1}^{T} \langle w_t - w^\star, g_t \rangle \right].$$

Recall the law of total expectation: for every pair of random variables X, Y and function g, we have $\mathbb{E}_X[g(X)] = \mathbb{E}_Y \mathbb{E}_X[g(X)|Y]$. Setting $X = g_{1:t}$ and $Y = g_{1:t-1}$ we get that

$$\mathbb{E}_{g_{1:T}}[\langle w_t - w^\star, g_t \rangle] = \mathbb{E}_{g_{1:t}}[\langle w_t - w^\star, g_t \rangle] = \mathbb{E}_{g_{1:t-1}} \mathbb{E}_{g_{1:t}}[\langle w_t - w^\star, g_t \rangle \mid g_{1:t-1}]$$

Once $g_{1:t-1}$ are known, the value of w_t is not random so we have

$$\mathbb{E}_{g_{1:t-1}} \mathbb{E}_{g_{1:t}}[\langle w_t - w^\star, g_t \rangle \mid g_{1:t-1}] = \mathbb{E}_{g_{1:t-1}} \langle w_t - w^\star, \mathbb{E}_{g_t}[g_t \mid g_{1:t-1}] \rangle$$

We know that w_t only depends on $g_{1:t-1}$ and SGD requires that $\mathbb{E}_{g_t}[g_t \mid w_t] = \nabla f(w_t)$, so therefore we have $\mathbb{E}_{g_t}[g_t \mid g_{1:t-1}] = \nabla f(w_t)$ and by convexity.

$$\mathbb{E}_{g_{1:t-1}} \langle w_t - w^\star, \mathbb{E}_{g_t}[g_t \mid g_{1:t-1}] \rangle \geq \mathbb{E}_{g_{1:t-1}}[F(w_t) - F(w^\star)] = \mathbb{E}_{g_{1:T}}[F(w_t) - F(w^\star)].$$

Therefore averaging over t and using linearity of expectation we have shown the inequality. □

It is left to prove the lemma.

Proof. We can calculate that

$$\begin{aligned}
\langle w_t - w^\star, g_t \rangle &= \frac{1}{\eta} \langle w_t - w^\star, \eta g_t \rangle \\
&= \frac{1}{2\eta} \big(-\|w_t - w^\star - \eta g_t\|^2 + \|w_t - w^\star\|^2 + \eta^2 \|g_t\|^2 \big) \\
&= \frac{1}{2\eta} \big(-\|w_{t+1} - w^\star\|^2 + \|w_t - w^\star\|^2 \big) + \frac{\eta}{2} \|g_t\|^2.
\end{aligned}$$

Now we will sum over t for both sides to get

$$\sum_{t=1}^{T}\langle w_t - w^\star, g_t\rangle = \frac{1}{2\eta}\left(\|w_1 - w^\star\|^2 - \|w_{T+1} - w^\star\|^2\right) + \frac{\eta}{2}\sum_{t=1}^{T}\|g_t\|^2$$
$$\leq \frac{1}{2\eta}\|w^\star\|^2 + \frac{\eta}{2}\sum_{t=1}^{T}\|g_t\|^2.$$

For the second part of the lemma, we just need to apply the bounds on $\|w^\star\|$, $\|g_t\|$, and the value of η. □

This rate matches what we got in the deterministic case with exact gradients in each round. This means that SGD is an extremely robust algorithm; the convergence rate is barely affected by the amount of noise.

What happens if we also assume smoothness?Unfortunately, in stochastic optimization, smoothness does not really help at all. The intuitive reason is because the noise of the gradients can still be large near the optimum. This is in contrast for the deterministic case, where we can get $1/T$ rate for smooth optimization (or a $1/T^2$ rate with acceleration).

Theorem 2. *Suppose F is μ-smooth, and suppose the stochastic oracle is such that $\mathbb{E}\|\nabla f(x) - g\|_2^2 \leq \sigma^2$. Also let B denote the bound on the initial distance to the optimum, i.e., $\|w^\star\| \leq B$. Then stochastic gradient descent with $\eta = 1/\left(\beta + 1/\frac{B}{\sigma}\sqrt{\frac{2}{T}}\right)$ achieves a guarantee*

$$\mathbb{E}[F(\bar{w})] - F(w^\star) \leq B\sigma\sqrt{\frac{2}{T}} + \frac{\beta B^2}{T}.$$

This result allows us to interpolate between the deterministic setting and the noisy setting.

1.2 SGD for Smooth and Strongly Convex

Interestingly, once we use SGD, we cannot get the linear convergence anymore, even for smooth and strongly convex optimization. See e.g., https://arxiv.org/pdf/1109.5647.pdf. But strong convexity does help a little bit.

Theorem 3. *Suppose F is λ-strongly convex and μ-smooth, and that $\mathbb{E}\|g_t\|^2 \leq G^2$. If we pick $\eta_t = 1/(\lambda t)$ then it holds for any T that*

$$\mathbb{E}[F(w_T) - F(w^\star)] \leq \frac{2\mu G^2}{\lambda^2 T}.$$

To prove this theorem, we will prove this important lemma.

Lemma 2. *Suppose F is λ-strongly convex and $\mathbb{E}\|g_t\|^2 \leq G^2$. Then picking $\eta_t = 1/(\lambda t)$ then it holds for any T that*

$$\mathbb{E}\left[\|w_T - w^\star\|^2\right] \leq \frac{4G^2}{\lambda^2 T}.$$

The theorem is a straightforward corollary of this lemma by using μ-smoothness.

Proof. By strong convexity, we know that

$$\langle \nabla f(w_t), w_t, w^\star \rangle \geq F(w_t) - F(w^\star) + \frac{\lambda}{2}\|w_t - w^\star\|^2,$$

and also that

$$F(w_t) - F(w^\star) \geq \frac{\lambda}{2}\|w_t - w^\star\|^2.$$

Therefore we can compute that

$$\begin{aligned}
\mathbb{E}\big[\|w_{t+1} - w^\star\|^2\big] &= \mathbb{E}\big[\|w_t - \eta_t g_t - w^\star\|^2\big] \\
&= \mathbb{E}\big[\|w_t - w^\star\|^2\big] - 2\eta_t\,\mathbb{E}[\langle g_t, w_t - w^\star\rangle] + \eta_t^2\,\mathbb{E}\|g_t\|^2 \\
&= \mathbb{E}\big[\|w_t - w^\star\|^2\big] - 2\eta_t\,\mathbb{E}[\langle \nabla f(x_t), w_t - w^\star\rangle] + \eta_t^2\,\mathbb{E}\|g_t\|^2 \\
&\leq \mathbb{E}\big[\|w_t - w^\star\|^2\big] - 2\eta_t\,\mathbb{E}\left[F(w_t) - F(w^\star) + \frac{\lambda}{2}\|w_t - w^\star\|^2\right] + \eta_t^2 G^2 \\
&\leq (1 - 2\eta_t\lambda)\,\mathbb{E}\big[\|w_t - w^\star\|^2\big] + \eta_t^2 G^2.
\end{aligned}$$

By choosing $\eta_t = 1/(\lambda t)$ we see that

$$\mathbb{E}\big[\|w_{t+1} - w^\star\|^2\big] \leq \left(1 - \frac{2}{t}\right)\mathbb{E}\big[\|w_t - w^\star\|^2\big] + \frac{G^2}{\lambda^2 t^2}.$$

Now we apply an inductive argument. Observe that at $t = 1$ we have

$$\frac{\lambda}{2}\|w_1 - w^\star\|^2 \leq \langle \nabla f(w_1), w_1 - w^\star\rangle,$$

which means that

$$\|\nabla f(w_1)\| \geq \frac{\lambda}{2}\|w_1 - w^\star\|.$$

In addition we know that

$$\begin{aligned}
\mathbb{E}\|g_1\|^2 &= \mathbb{E}\|\nabla f(w_1) + (g_1 - \nabla f(w_1))\|^2 \\
&= \mathbb{E}\|\nabla f(w)\|^2 + \mathbb{E}\|g_1 - \nabla f(w_1)\|^2 + 2\,\mathbb{E}[\langle g_1 - \nabla f(w_1), \nabla f(w_1)\rangle] \\
&\geq \mathbb{E}\|\nabla f(w)\|^2.
\end{aligned}$$

Therefore we get that

$$\mathbb{E}\big[\|w_1 - w^\star\|^2\big] \leq \frac{4}{\lambda^2}\,\mathbb{E}\big[\|\nabla f(w_1)\|^2\big] \leq \frac{4}{\lambda^2}\,\mathbb{E}\big[\|g_1\|^2\big] \leq \frac{4G^2}{\lambda^2}.$$

At $t = 2$, we can also get that

$$\mathbb{E}\big[\|w_2 - w^\star\|^2\big] \leq \frac{G^2}{4\lambda^2} \leq \frac{4G^2}{\lambda^2 \cdot 2}. \tag{1}$$

Now we apply induction. Suppose that at iterate t we have $\mathbb{E}\big[\|w_2 - w^\star\|^2\big] \leq \frac{4G^2}{\lambda^2 \cdot t}$. Then at iterate $t + 1$ we have

$$\mathbb{E}\big[\|w_{t+1} - w^\star\|^2\big] \leq \left(1 - \frac{2}{t}\right)\mathbb{E}\big[\|w_t - w^\star\|^2\big] + \frac{G^2}{\lambda^2 t^2}$$

$$\leq \left(1 - \frac{2}{t}\right) \cdot \frac{4G^2}{\lambda^2 t} + \frac{G^2}{\lambda^2 t^2}$$
$$= \frac{4G^2}{\lambda^2 t} - \frac{7G^2}{\lambda^2 t^2} \leq \frac{4G^2}{\lambda^2 (t+1)}.$$

So the induction step holds. □

Remark: a similar conclusion holds for the averaged iterate $\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w_t$ (but with slightly worse constants).

References