CS 7545: Machine Learning Theory **Fall 2019**

Lecture 13: Stochastic Gradient Descent and Mirror Descent

Lecturer: Jacob Abernethy *Scribes: Rui Feng, Jing Wu*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

Let $\mathcal{D}$ be a distribution on data and $f(x, \xi)$ be a convex loss function of parameter x given data ξ. Let $F(x) = \mathbb{E}_\xi[f(x, \xi)]$. Suppose we have (uniformly sampled) samples $\widehat{D} = \{\xi_{1:n}\}$ and empirical loss $\widehat{F}(x) = \frac{1}{n}\sum f(x, \xi_i)$.

The goal of learning is

$$\text{minimize } F(x) \tag{13.1}$$

which involves expectation and intractable. Instead, the goal of optimization is

$$\text{minimize } \widehat{F}(x) \tag{13.2}$$

13.1 Stochastic Gradient Descent

The following Algorithm 1 is called *stochastic gradient descent.*

Algorithm 1 Stochastic Gradient Descent

1: **for** $t = 1, \ldots, T$ **do**
2: Sample $\xi_t \overset{\text{i.i.d.}}{\sim} \mathcal{D}$ (or $\sim \widehat{D}$)
3: Update $x_{t+1} = x_t - \eta_t \nabla f(x_t, \xi_t)$
4: **end for**

The following claim shows the generalization bound.

Claim 13.1 *Let* $\overline{x}_T = \frac{1}{T}\sum x_t$. *Then*

$$\mathbb{E}_{\xi_{1:n}}\left[F(\overline{x}_T) - F(x^*)\right] = O\left(\frac{GD}{\sqrt{T}}\right)$$

where x^* *minimizes the optimization goal.*

Proof:

$$\begin{aligned}
\mathbb{E}_{\xi_{1:n}}\left[F(\overline{x}_T) - F(x^*)\right] &\leq \mathbb{E}_{\xi_{1:n}}\left[\frac{1}{T}\sum_{t=1}^{T}(F(\overline{x}_T) - F(x^*))\right] \\
&= \mathbb{E}_{\xi_{1:n}}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\xi\sim\mathcal{D}}[f(x_t, \xi) - f(x^*, \xi)]\right]
\end{aligned}$$

where the first inequality is because of the convexity of f. Observe that x_t is independent of ξ_t, and ξ_t and ξ have the same distribution,

$$\begin{aligned}
\mathbb{E}_{\xi_{1:n}}\left[F(\overline{x}_T) - F(x^*)\right] &= \mathbb{E}_{\xi_{1:n}}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\xi_t}[f(x_t, \xi_t) - f(x^*, \xi_t)]\right] \\
&= \mathbb{E}_{\xi_{1:n}}\left[\frac{1}{T}\sum_{t=1}^{T}(f(x_t, \xi_t) - f(x^*, \xi_t))\right] \\
&= \mathbb{E}\left[\frac{\text{Regret}_T}{T}\right] \leq \frac{3GD}{2\sqrt{T}}
\end{aligned}$$

the inequality is the result from the last lecture. ■

Based on a similar proof, the following claim is also true:

Claim 13.2

$$\mathbb{E}_{\xi_{1:n}}\left[\widehat{F}(\overline{x}_T) - \widehat{F}(x^*)\right] = O\left(\frac{GD}{\sqrt{T}}\right)$$

Remark 1 *The proof of Claim 13.2 is not algorithm-dependent. It applies to any low-regret algorithm.*

Based on the generalization bound, we have the following corollary

Corollary 13.3 $\mathbb{E}\left[F(\widehat{x}_T) - f(\widehat{x}_T)\right] \leq \sqrt{\frac{\log(\frac{1}{\delta})}{2T}}$ *holds with probability* $1-\delta$.

The proof of the corollary is not detailed in class, but it involves Azuma's inequality from the martingale theory.

Remark 2 *Gradient descent converges at the rate* $\frac{GD}{\sqrt{T}}$. *Why is SGD better for optimizing? Both algorithms converge at* $O\left(\frac{1}{\sqrt{T}}\right)$ *but the time complexity differs. In fact, a step of GD takes* $O(n)$ *operations and SGD takes only* $O(1)$.

Remark 3 *In the proof of Claim 13.2, the fact that samples are i.i.d. is leveraged. In practice, people often go through the data sequentially for many runs. The first run is ok, but after that the i.i.d. assumption is violated and hence the generalization bound is no longer valid.*

However, if we re-sample from the samples with replacement, the i.i.d. assumption still holds.

13.2 Mirror Descent

The following describes a new algorithm called *mirror descent*.

Given a convex set $K \in \mathbb{R}^d$, a *regularizer* $R(\cdot)$ defined on K, $K \subseteq \text{int}(\text{dom}(R))$, and R is λ-strongly convex w.r.t. some norm $\|\cdot\|$.

At time t, let $\nabla_t = \nabla f_t(x_t)$. The mirror descent algorithm selects x_{t+1} as

$$x_{t+1} = \arg\min_{x \in K} \eta\langle \nabla_t, x\rangle + D_R(x, x_t)$$

where $D_R(x,y) = R(x) - R(y) - \langle \nabla R(y), x - y\rangle$ is the Bregman divergence of R, and R being λ-strongly convex is equivalent to $D_R(x,y) \geq \frac{\lambda}{2}\|x-y\|^2$.

Before proving the generalization bound for mirror descent, it is necessary to prepare the following tools:

1. **Lemma 13.4** *Let* $x, y, z \in K$, *then*

$$D_R(z,x) + D_R(x,y) - D_R(z,y) = \langle \nabla R(y) - \nabla R(x), z - x\rangle$$

2. The first-order optimality condition states that if $x^* = \arg\min_{x\in K}\Phi(x)$ with Φ convex, then $\forall u \in K, \langle \nabla\Phi(x^*), u - x^*\rangle \geq 0$.

   Let $\Phi(x) = \eta\langle \nabla_t, x\rangle + D_R(x, x_t)$. Then

$$\nabla\Phi(x) = \eta\nabla_t + \nabla(R(x) - R(x_t)) - \nabla(\langle \nabla R(x_t), x - x_t\rangle) = \nabla R(x) - \nabla R(x_t) + \eta\nabla_t$$

   The first-order optimality condition says that

$$\langle \nabla\Phi(x_{t+1}), u - x_{t+1}\rangle \geq 0$$

   Hence we've proved the following lemma:

Lemma 13.5

$$\langle \nabla R(x_{t+1}) - \nabla R(x_t) + \eta \nabla_t, u - x_{t+1} \rangle \geq 0$$

3. Recall that $ab \leq \frac{1}{p}a^p + \frac{1}{q}b^q$ if $\frac{1}{p} + \frac{1}{q} = 1$. Then

$$\langle v, r \rangle \leq \|v\| \|r\|_* = \left(\frac{1}{\lambda}\|v\|\right)(\lambda \|r\|_*) \leq \left(\frac{1}{2\lambda}\|v\|^2\right) + \left(\frac{\lambda}{2}\|r\|_*^2\right)$$

the second inequality is Hölder's.

4. And finally

Lemma 13.6 *Let η be changing and at each time it is η_t. By convexity of f_t,*

$$\eta_t \left[f_t(x_t) - f_t(u)\right] \leq \eta_t \langle \nabla_t, x_t - y \rangle \leq D_R(u, x_t) - D_R(u, x_{t+1}) + \frac{\eta_t^2}{2\lambda}\|\nabla_t\|_*^2$$

Theorem 13.7 *Let $u \in K$ be arbitrary. The regret mirror descent is given by*

$$\text{Regret}_T(\text{MD}) \leq D_R(u, x_1) + \frac{1}{2\lambda}\sum_{t=1}^{T} \eta_t^2 \|\nabla_t\|_*^2$$

In this lecture, there was not enough time for the complete proof of Theorem 13.7. Hence, it is reserved for the next lecture.