Lecture #12: Gradient Descent Variants

Instructor: Aditya Bhaskara *Scribe: Yosuke Mizutani*

CS 5966/6966: Theory of Machine Learning

February 17th, 2022

Abstract

This lecture reviewed the basic theorem of the gradient descent for a convex, L-Lipschitz function and introduced its variants, online convex optimization and stochastic gradient descent. Further, we learned that more structure on the function, such as smoothness and strong convexity, provides a faster convergence rate.

1 Recap: Gradient Descent Analysis

Gradient descent is an iterative algorithm for finding a (local) minimum of a differentiable function. Particularly, we are interested in finding the minimizer (or the minimum value) for a convex function.

A function $f : \mathcal{D} \to \mathbb{R}$ is convex if for all $x, y \in \mathcal{D}$ and $0 \leq t \leq 1$, $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$. Any local minimum of a convex function is the global minimum.

The gradient of a scalar-valued multivariable function $f(x_1, x_2, \ldots, x_d)$, denoted by ∇f, is given by $\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_d}\right)$.

Gradient Descent for Convex Functions

Given a convex function $f : \mathcal{D} \to \mathbb{R}$ defined over a convex domain $\mathcal{D} \subseteq \mathbb{R}^d$, the algorithm starts with some feasible point $w_0 \in \mathcal{D}$. Then, for $t = 0, 1, \ldots, T-1$ iteratively set $w_{t+1} = \Pi\left(w_t - \eta_t \nabla f(w_t)\right)$, where η_t is the learning rate at step t and $\Pi : \mathbb{R}^d \to \mathcal{D}$ is a projection to the feasible set.

Projection to the feasible set can be "hard" if the domain is not "simple".

1.1 *Basic theorem*

For analysis, we assume that the convex function f is L-Lipschitz and its domain is all of $\mathbb{R}^d$. A function is called L-Lipschitz when:

$$|f(x) - f(y)| \leq L\|x - y\| \qquad \text{for every } x, y \tag{1}$$

This implies the following, although we left the proof as an exercise.

One may use the multi-variable Mean Value Theorem, which states that for every x, y, there exists $z \in [x, y]$ such that $f(x) - f(y) = \langle \nabla f(z),\ x - y \rangle$.

$$\|\nabla f(x)\| \leq L \qquad \text{for every } x \tag{2}$$

Let $w^* \in \mathbb{R}^d$ be the optimum minimizer, that is, $w^* = \arg\min_{x \in \mathbb{R}^d} f(x)$. Then, we introduce a parameter $B \in \mathbb{R}$, which bounds how bad the starting point is.

$$|w_0 - w^*| \leq B \tag{3}$$

For the given step t, how far is $f(w_t)$ from $f(w^*)$? By the convexity of f, we have:

$$f(w^*) \geq f(w_t) + \langle w^* - w_t,\ \nabla f(w_t) \rangle \tag{4}$$

For a differentiable function f, the tangent at x is the linear function $\ell(y) = f(x) + \langle \nabla f(x),\ y - x \rangle$. If f is convex, it must hold that $f(y) \geq \ell(y)$.

And we get:

$$\langle w_t - w^*,\ \nabla f(w_t) \rangle \geq f(w_t) - f(w^*) \tag{5}$$

Now, let us define the *potential function* $\Phi_t := \|w_t - w^*\|^2$ for the given t, and consider the *potential drop* $\Phi_t - \Phi_{t+1}$.

$$\begin{aligned}
\Phi_t - \Phi_{t+1} &= \|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2 \\
&= \|w_t - w^*\|^2 - \|w_t - \eta \nabla f(w_t) - w^*\|^2 && w_{t+1} = w_t - \eta \nabla f(w_t) \\
&= \|w_t - w^*\|^2 - \|(w_t - w^*) - \eta \nabla f(w_t)\|^2 \\
&= \|w_t - w^*\|^2 - && \text{(norm of a sum of vectors)} \\
&\qquad \Big[\|w_t - w^*\|^2 - 2 \langle w_t - w^*,\ \eta \nabla f(w_t) \rangle + \|\eta \nabla f(w_t)\|^2 \Big] \\
&= 2\eta \langle w_t - w^*,\ \nabla f(w_t) \rangle - \eta^2 \|\nabla f(w_t)\|^2 \\
&\geq 2\eta \left[f(w_t) - f(w^*) \right] - \eta^2 \|\nabla f(w_t)\|^2 && \text{from (5)} \\
&\geq 2\eta \left[f(w_t) - f(w^*) \right] - \eta^2 L^2 && \text{from (2)}
\end{aligned} \tag{6}$$

Therefore, we get:

$$f(w_t) - f(w^*) \leq \frac{\Phi_t - \Phi_{t+1}}{2\eta} + \frac{L^2 \eta}{2} \tag{7}$$

Finally, consider the sum over these values after T steps.

$$\begin{aligned}
\sum_{t=0}^{T-1} f(w_t) - f(w^*) &\leq \sum_{t=0}^{T-1} \frac{\Phi_t - \Phi_{t+1}}{2\eta} + \frac{L^2 \eta}{2} \\
&= \frac{1}{2\eta} \left(\sum_{t=0}^{T-1} \Phi_t - \Phi_{t+1} \right) + T \cdot \frac{L^2 \eta}{2} \\
&= \frac{\Phi_0 - \Phi_T}{2\eta} + T \cdot \frac{L^2 \eta}{2} && \text{(telescoping sum)} \\
&= \frac{\|w_0 - w^*\|^2 - \|w_T - w^*\|^2}{2\eta} + T \cdot \frac{L^2 \eta}{2} \\
&\leq \frac{\|w_0 - w^*\|^2}{2\eta} + T \cdot \frac{L^2 \eta}{2} \\
&\leq \frac{B^2}{2\eta} + T \cdot \frac{L^2 \eta}{2} && \text{from (3)}
\end{aligned} \tag{8}$$

This gives the following theorem.

1.1 theorem. *Consider running T steps of gradient descent with a fixed learning rate η. Then we have:*

$$\frac{1}{T} \sum_{t=1}^{T} f(w_t) - f(w^*) \leq \frac{B^2}{2\eta T} + \frac{L^2 \eta}{2}$$

Here we use indices $t \in [1, T]$, and (8) still holds.

When we set $\eta = \frac{B}{L\sqrt{T}}$, we get a nicer form:

$$\frac{1}{T}\sum_{t=1}^{T} f(w_t) - f(w^*) \leq \frac{BL}{\sqrt{T}} \tag{9}$$

We say this is $\frac{1}{\sqrt{T}}$-convergence.

1.2 *Constrained domain*

The same proof works if we had a constrained domain.

Let $w_{t+\frac{1}{2}} = w_t - \eta \nabla f(w_t)$, which may or may not be feasible. Then, we have the following due to the convex domain.

$$\|w^* - \Pi\left(w_{t+\frac{1}{2}}\right)\|^2 \leq \|w^* - w_{t+\frac{1}{2}}\|^2 \tag{10}$$

It turns out that the projection to the feasible set does not violate any inequalities in the main proof.

1.3 *Different functions at time steps*

The proof works even if functions at different time steps were different. Let $f_1, f_2, \ldots, f_T$ be all convex, L-Lipschitz functions.

We update our gradient descent to $w_{t+1} = w_t - \nabla f_t(w_t)$. Now the potential drop becomes:

$$\begin{aligned} \Phi_t - \Phi_{t+1} &= 2\eta \langle \nabla f_t(w_t),\ w_t - w^* \rangle - \eta^2 \|\nabla f_t(w_t)\|^2 \\ &\geq 2\eta \left[f_t(w_t) - f_t(w^*)\right] - \eta^2 L^2 \end{aligned} \tag{11}$$

This implies the following result.

$$\sum_{t=1}^{T} f_t(w_t) - f_t(w^*) \leq \frac{B^2}{2\eta} + \frac{L^2 \eta T}{2} \tag{12}$$

When $\eta = \frac{B}{L\sqrt{T}}$, we have:

$$\sum_{t=1}^{T} f_t(w_t) - f_t(w^*) \leq \sqrt{T} \tag{13}$$

2 Online Convex Optimization

Now, let us see some application of gradient descent. The previous result is directly applicable to the online convex optimization, where a learner makes a series of decisions to minimize the total loss, and loss functions (convex, L-Lipschitz) $f_1, \ldots, f_T$ over the same domain $\mathcal{D}$ are given sequentially.

Think about learning game where adversary chooses functions.

We want our total loss to be comparable with the best *fixed* minimizer x in hindsight. That is, $x^* = \arg\min_{x \in \mathcal{D}} \sum_{t=1}^{T} f_t(x)$.

From (13), we obtain:

$$\left(\sum_{t=1}^{T} f_t(x_t)\right) - \min_{x\in\mathcal{D}} \sum_{t=1}^{T} f_t(x) \leq \sqrt{T} \tag{14}$$

The left-hand side of (14) is called *regret*. If we want to compare our decisions to a dynamic minimizer, it is called *switching regret* or *dynamic regret*.

When there are k switches, the right-hand side of (14) becomes $\sqrt{kT}$.

3 Stochastic Gradient Descent

In stochastic gradient descent (SGD), we do not require to use the same function f for every iteration, but we randomly choose a function g such that the expected value of g over the possible functions is equal to f (i.e. $f = \mathbb{E}[g]$).

For example, ERM (empirical risk minimization) for training data $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$ can be viewed as loss:

$$f : \frac{1}{m}\sum_{i=1}^{m} \ell(h_w(x_i), y_i) \tag{15}$$

$[m]$ denotes $[1, m] = \{1, \ldots, m\}$.

Imagine we sample an index $i \sim [m]$ at random (with replacement). If we define $g_i(w) = \ell(h_w(x_i), y_i)$, then we have:

$$f(w) = \frac{1}{m}\sum_{i=1}^{m} g_i(w) \tag{16}$$

From the linearity of the gradient, $\nabla f = \frac{1}{m}\sum_{i=1}^{m} \nabla g_i$.

Now, consider T iterations where at each step t, we pick $i_t \sim [m]$. We write $h_t(w) = g_{i_t}(w)$ for the function chosen at step t as a random variable. Observe that at every step t, $\mathbb{E}[h_t] = f$.

$f = \mathbb{E}[h_t]$ holds even if we had larger batch sizes.

The stochastic gradient descent works as follows.

Stochastic Gradient Descent

Given a convex function $f : \mathcal{D} \to \mathbb{R}$ defined over a convex domain $\mathcal{D} \subseteq \mathbb{R}^d$ and a distribution $\mathcal{X}$ of functions ($\mathcal{D} \to \mathbb{R}$) such that $f = \mathbb{E}_{h\sim\mathcal{X}}[h]$, the algorithm starts with some feasible point $w_0 \in \mathcal{D}$.
Then, for $t = 0, 1, \ldots, T-1$ iteratively pick a random function $h_t \sim \mathcal{X}$ and set $w_{t+1} = \Pi\left(w_t - \eta \nabla h_t(w_t)\right)$, where η is the learning rate and $\Pi : \mathbb{R}^d \to \mathcal{D}$ is a projection to the feasible set.

Intuitively, we can decompose a function f into a number of simple functions $g_1, \ldots, g_m$ so that (16) holds. And for each step we randomly pick one of the g_i's and perform gradient descent as usual.

3.1 *Analysis of Stochastic Gradient Descent*

Here we want to prove the following theorem.

3.1 theorem. *Let $\overline{w}$ be the average point over $w_1, \ldots, w_T$ chosen by stochastic gradient descent, that is, $\overline{w} = \frac{1}{T}\sum_{i=1}^{T} w_i$, and w^* be the optimum minimizer $w^* =$*

$\arg\min_{x \in \mathcal{D}} f(x)$. *Then,*

$$\mathbb{E}[f(\overline{w})] - f(w^*) \leq \frac{B^2}{2\eta T} + \frac{L^2 \eta}{2},$$

where $|w_0 - w^*| \leq B$ *and* L *is a bound on the Lipschitz constant of all the* g_i*'s.*

Notice that in SGD w_i's and $\overline{w}$ are random variables. From Jensen's inequality, we have:

See Section 14.1.1 (page 186) of Shalev-Shwartz and Ben-David's book.

$$\begin{aligned} f(\overline{w}) - f(w^*) &= f\left(\frac{1}{T}\sum_{t=1}^{T} w_t\right) - f(w^*) \\ &\leq \frac{1}{T}\left(\sum_{t=1}^{T} f(w_t)\right) - f(w^*) \\ &= \frac{1}{T}\sum_{t=1}^{T} f(w_t) - f(w^*) \end{aligned} \tag{17}$$

A key idea to proceed the proof is to keep track of the *expected* potential drop.

$$\begin{aligned} \mathbb{E}_{h_t}[\Phi_t - \Phi_{t+1} \mid w_t] &= \mathbb{E}_{h_t}[2\eta \langle \nabla h_t(w_t),\ w_t - w^* \rangle] - \eta^2 \|\nabla h_t(w_t)\|^2] \\ &\geq 2\eta \cdot \mathbb{E}_{h_t}[\langle \nabla h_t(w_t),\ w_t - w^* \rangle] - \eta^2 L^2 && \text{from (2)} \\ &= 2\eta \cdot \langle \mathbb{E}_{h_t}[\nabla h_t(w_t)],\ w_t - w^* \rangle - \eta^2 L^2 \\ &= 2\eta \cdot \langle \nabla f(w_t),\ w_t - w^* \rangle - \eta^2 L^2 \\ &\geq 2\eta(f(w_t) - f(w^*)) - \eta^2 L^2 && \text{from (5)} \end{aligned} \tag{18}$$

The expectations are over the choice of h_t.

Now, the remaining task is to accumulate the left-hand side of (18). Notice that it still telescopes; what we condition on doesn't matter!

Convince yourself with $T = 2$.

$$\sum_{t=1}^{T} \mathbb{E}_{h_t}[\Phi_t - \Phi_{t+1} \mid w_t] = \mathbb{E}_{h_1, h_2, \ldots, h_T}[\Phi_1 - \Phi_{T+1}] \leq \Phi_0 \leq B^2 \tag{19}$$

The accumulated right-hand side would be:

$$\begin{aligned} \sum_{t=1}^{T} 2\eta(f(w_t) - f(w^*)) - \eta^2 L^2 &= 2\eta \left[\sum_{t=1}^{T} f(w_t) - f(w^*)\right] - T\eta^2 L^2 \\ &\geq 2\eta T\left(f(\overline{w}) - f(w^*)\right) - T\eta^2 L^2 && \text{from (17)} \\ &= \eta T\left[2\left(\mathbb{E}[f(\overline{w})] - f(w^*)\right) - \eta L^2\right] \end{aligned} \tag{20}$$

By combining (19) and (20), we obtain the following, which completes the proof.

$$B^2 \geq \eta T\left[2\left(\mathbb{E}[f(\overline{w})] - f(w^*)\right) - \eta L^2\right] \tag{21}$$

$$\mathbb{E}[f(\overline{w})] - f(w^*) \leq \frac{B^2}{2\eta T} + \frac{L^2 \eta}{2}$$

4 More Structure On Function

4.1 *Smooth functions*

A function is *smooth* if its gradient is also Lipschitz (i.e. the gradient does not change rapidly). The gradient descent for smooth functions achieves $\frac{1}{T}$-convergence. This topic was covered in Lecture #13.

4.2 *Strongly convex functions*

A function f is *strongly convex* if $f(y) \geq f(x) + \langle \nabla f(x),\ y - x \rangle + \mu \|y - x\|^2$ for some $\mu \geq 0$ and for all x, y. In this case, an error is bounded by e^{-T} after T steps.

A function is convex when $\mu = 0$.

Refer to Section 14.4.4 (page 195) of Shalev-Shwartz and Ben-David's book.