10-725 Convex Optimization Course Notes

Yan Pan

February 10, 2021

Course notes for CMU's 10-725 Convex Optimization in Spring 2021 by Yuanzhi Li.

1 Introduction

Definition 1.1 *A set $\mathcal{D}$ is **convex** if for every $x, y \in \mathcal{D}$ and $0 \le \lambda \le 1$, we have*

$$(1-\lambda)x + \lambda y \in \mathcal{D}.$$

Definition 1.2 *A function f is **convex** over a convex set $\mathcal{D}$ if for every $x, y \in \mathcal{D}$ and $0 \le \lambda \le 1$, we have*

$$f\left((1-\lambda)x + \lambda y\right) \le (1-\lambda)f(x) + \lambda f(y).$$

Lemma 1.3 (Lower Linear Bound) *For every (differentiable) convex function f over a convex set $\mathcal{D}$, for every $x, y \in \mathcal{D}$, we have*

$$f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle.$$

Proof. Assume for contradiction that exists $y \in \mathcal{D}$ such that $f(y) \le f(x) + \langle \nabla f(x), y - x \rangle - \varepsilon$ for $\varepsilon > 0$, then by convexity, for all $0 \le \lambda \le 1$, we have

$$f((1-\lambda)x + \lambda y) \le (1-\lambda)f(x) + \lambda f(y) \le f(x) + \lambda \langle \nabla f(x), y - x \rangle - \lambda\varepsilon,$$

$$f(x + \lambda(y-x)) \le f(x) + \lambda \langle \nabla f(x), y - x \rangle - \lambda\varepsilon$$

which is (for non-zero λ)

$$\frac{f(x + \lambda(y-x)) - f(x)}{\lambda} \le \langle \nabla f(x), y - x \rangle - \varepsilon.$$

By definition,

$$\lim_{\lambda \to 0} \frac{f(x + \lambda(y-x)) - f(x)}{\lambda} = \langle \nabla f(x), y - x \rangle.$$

So let $\lambda \to 0^+$ we have $\langle \nabla f(x), y - x \rangle \le \langle \nabla f(x), y - x \rangle - \varepsilon$, which is a contradiction. □

Definition 1.4 *An optimization problem is considered as **convex optimization** if f is a convex function and $\mathcal{D}$ is a convex set.*

- A convex function does not need to be differentiable, such as $f(x) = |x|$.
- Examples of convex optimization problems:
  - Linear regression: $\min_x \|y - Ax\|_2^2$.
  - Ridge regression: $\min_x \|y - Ax\|_2^2 + \lambda \|x\|_2^2$.
  - Logistic regression: $\min_x \sum_i -\log \frac{1}{1+e^{-y_i \langle A_i, x \rangle}}$.
- Convex function is considered easy because it has no local minima.

Theorem 1.5 *For a first order differentiable convex function f, $\nabla f(x^*) = 0$ iff $f(x^*) = \min_x f(x)$.*

Proof. ($\Rightarrow$) By the lower linear bound, for every $x \in \mathcal{D}$, $f(x) \geq f(x^*)$ since $\nabla f(x^*) = 0$.
($\Leftarrow$) Next lecture. □

Corollary 1.5.1 *For general Lipschitz function, $\exists y \in \partial f(x^*), y = 0$ iff $f(x^*) = \min_x f(x)$.*

For neural networks, even with one ReLU, the loss function is non-convex. However with millions of ReLUs, the function becomes close to convex[1].

2 Gradient Descent and Mirror Descent

Definition 2.1 (Gradient Descent Algorithm) *Given a starting point x_0 and a learning rate η, for $t = 0, 1, 2, \cdots$,*

$$x_{t+1} = x_t - \eta \nabla f(x_t).$$

The selection of the learning rate is important in gradient descent algorithm. Intuitively, we might want to choose large learning rates for "smooth" functions and small learning rates for "steep" functions. We need to define the smoothness of functions.

Definition 2.2 (L-smooth) *A first order differentiable function (not necessarily convex) f over a set (not necessarily convex) $\mathcal{D}$ is called **L-smooth** for some $L > 0$ if for every $x, y \in \mathcal{D}$,*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|_2^2.$$

The above formula is also referred to as the **upper quadratic bound**. Now for convex function f, we have both the lower linear bound and upper quadratic bound, giving us for every $x, y \in \mathcal{D}$,

$$f(x) + \langle \nabla f(x), y - x \rangle \leq f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|_2^2.$$

Lemma 2.3 (Smoothness Under Addition) *If f_1 and f_2 are L-smooth, then $f(x) = f_1(x) + f_2(x)$ is $2L$-smooth.*

Proof. Trivial. □

Theorem 2.4 (Alternative Definition of L-smoothness) *A second order differentiable function over a **convex** set $\mathcal{D}$ is L-smooth iff for every unit vector v, for every $x \in \mathcal{D}$,*

$$v^T \nabla^2 f(x) v \leq L.$$

The proof is omitted because it's highly non-trivial. With this alternative definition we have this following fact.

Lemma 2.5 *Every third order differentiable function over a closed, bounded convex set $\mathcal{D}$ is L-smooth for some finite L.*

With the definition of L-smoothness, we can derive a lemma for gradient descent about how much the function decreases.

Lemma 2.6 (Gradient Descent Lemma) *For any L-smooth function f, gradient descent decreases the function value by*

$$f(x_{t+1}) \leq f(x_t) - \frac{\eta}{2} \|\nabla f(x_t)\|_2^2$$

as long as $\eta \leq \frac{1}{L}$.

Proof. Notice that $x_{t+1} = x_t - \eta \nabla f(x_t)$ in gradient descent. By definition of L-smoothness, we have

$$f(x_{t+1}) \leq f(x_t) - \langle \nabla f(x_t), \eta \nabla f(x_t) \rangle + \eta^2 \frac{L}{2} \|\nabla f(x_t)\|_2^2.$$

Since $\eta^2 \frac{L}{2} \leq \eta$ for $\eta \leq \frac{1}{L}$, we have

$$f(x_{t+1}) \leq f(x_t) - \frac{\eta}{2} \|\nabla f(x_t)\|_2^2.$$

□

Theorem 2.7 (Convergence Rate of Gradient) *For every $\varepsilon > 0$, within*

$$T_\varepsilon = \frac{2(f(x_0) - \min_x f(x))}{\eta \varepsilon}$$

many iterations, there must be a $t \leq T_\varepsilon$ such that $\|\nabla f(x_t)\|_2^2 \leq \varepsilon$.

Proof. By contradiction. Suppose for every $t \leq T_\varepsilon$, $\|\nabla f(x_t)\|_2^2 > \varepsilon$. Then,

$$f(x_{t+1}) < f(x_t) - \frac{\eta}{2} \varepsilon.$$

Which implies that

$$f(x_{T_\varepsilon}) < f(x_0) - \frac{\eta}{2} T_\varepsilon \varepsilon = f(x_0) - (f(x_0) - \min_x f(x)) = \min_x f(x)$$

which is a contradiction. □

The lemma indicates that the learning rate does not depend on how large the gradient is, although in practice we do want to tune the learning rate for faster convergence. The lemma shows that we can decrease the objective, however if the gradient is too small, it might take forever for gradient to reach exactly zero. Take the function $f(x) = \varepsilon^2 x^2$ as an example, when $x = \frac{1}{\varepsilon}$ we have $|\nabla f(x)| = \varepsilon$ but $f(x) = 1$. We need to study the convergence rate of gradient descent.

Lemma 2.8 (Mirror Descent Lemma) *For any point y, we have*

$$f(x_t) \leq f(y) + \frac{1}{2\eta} \left(\|y - x_t\|_2^2 - \|y - x_{t+1}\|_2^2 + \|x_{t+1} - x_t\|_2^2 \right).$$

Proof. By lower linear bound we have

$$f(y) \geq f(x_t) + \langle \nabla f(x_t), y - x_t \rangle = f(x_t) + \frac{1}{\eta} \langle x_t - x_{t+1}, y - x_t \rangle.$$

We observe that

$$\begin{aligned} \langle x_t - x_{t+1}, y - x_t \rangle &= \langle x_{t+1}, x_t \rangle + \langle y, x_t \rangle - \langle x_{t+1}, y \rangle - \langle x_t, x_t \rangle \\ &= -\frac{1}{2} \left(\|y - x_t\|_2^2 - \|y - x_{t+1}\|_2^2 + \|x_{t+1} - x_t\|_2^2 \right). \end{aligned}$$

Hence we have the desired result. □

The mirror descent lemma implies that if the current function value $f(x_t)$ is much larger than $f(x^*)$, then since $\|x_{t+1} - x_t\|_2^2$ is small, the values $\|x^* - x_{t+1}\|_2^2$ must be much smaller than $\|x^* - x_t\|_2^2$, so x_{t+1} will be much closer to x^* compared to x_t. So the mirror descent lemma looks at decreasing the distance between x_t to x^*.

Theorem 2.9 (Convergence Rate) *For every $\eta \leq \frac{1}{L}$, we have the rate of convergence*

$$f(x_T) \leq f(x^*) + \frac{\|x^* - x_0\|_2^2}{\eta T}.$$

Proof. Let $y = x^*$, we sum up the mirror descent lemma for $t = 0, \cdots, t-1$, we have the telescope sum inequality

$$\sum_{t=0}^{T-1} f(x_t) \leq Tf(x^*) + \frac{1}{2\eta}\left(\|x^* - x_0\|_2^2 - \|x^* - x_T\|_2^2 + \sum_{t=0}^{T-1}\|x_{t+1} - x_t\|_2^2\right)$$

$$\frac{1}{T}\sum_{t=0}^{T-1} f(x_t) \leq f(x^*) + \frac{1}{2\eta T}\left(\|x^* - x_0\|_2^2 - \|x^* - x_T\|_2^2 + \sum_{t=0}^{T-1}\|x_{t+1} - x_t\|_2^2\right)$$

$$\frac{1}{T}\sum_{t=0}^{T-1} f(x_t) \leq f(x^*) + \frac{1}{2\eta T}\left(\|x^* - x_0\|_2^2 + \sum_{t=0}^{T-1}\|x_{t+1} - x_t\|_2^2\right)$$

which implies using $x_{t+1} = x_t - \eta\nabla f(x_t)$

$$\frac{1}{T}\sum_{t=0}^{T-1} f(x_t) \leq f(x^*) + \frac{1}{2\eta T}\|x^* - x_0\|_2^2 + \frac{\eta}{2T}\sum_{t=0}^{T-1}\|\nabla f(x_t)\|_2^2.$$

When $\eta \leq \frac{1}{L}$, by Gradient Descent Lemma we have

$$\frac{\eta}{2}\sum_{t=0}^{T-1}\|\nabla f(x_t)\|_2^2 \leq f(x_0) - f(x_T) \leq f(x_0) - f(x^*)$$

since x^* is a minimizer of f. Then,

$$\frac{1}{T}\sum_{t=0}^{T-1} f(x_t) \leq f(x^*) + \frac{1}{T}\left(\frac{1}{2\eta}\|x^* - x_0\|_2^2 + f(x_0) - f(x^*)\right).$$

Using the L-smoothness of f that shows $f(x_0) - f(x^*) \leq \frac{L}{2}\|x^* - x_0\|_2^2$ and the gradient descent lemma that shows $f(x_T) \leq f(x_t)$, we have

$$f(x_T) \leq f(x^*) + \frac{\|x^* - x_0\|_2^2}{T}\left(\frac{1}{2\eta} + \frac{L}{2}\right)$$

which implies when $\eta \leq \frac{1}{L}$

$$f(x_T) \leq f(x^*) + \frac{\|x^* - x_0\|_2^2}{\eta T}.$$

□

3 Momentum

The intuition behind **accelerated gradient descent** is to use larger learning rate $\eta > \frac{1}{L}$ without entering zig-zag.

References

[1] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019.