10-725 Convex Optimization Course Notes

Yan Pan

February 13, 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*

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

Corollary 1.2.1 (Alternative Definition of Convexity) *A second order differentiable function over a convex set $\mathcal{D}$ is convex iff for every vector v and every $x \in \mathcal{D}$, we have*

$$v^\top \nabla^2 f(x) v \ge 0.$$

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. $\square$

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

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 $x, y \in \mathcal{D}$,

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

There are some nice properties about smoothness, for example in Corollary 2.2.1 and Corollary 2.2.2.

Corollary 2.2.1 (Smoothness Under Addition) *If f_1 and f_2 are L-smooth, then $f(x) = f_1(x) + f_2(x)$ is $2L$-smooth. (Trivial)*

Corollary 2.2.2 (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 $x \in \mathcal{D}$,*

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

The proof is omitted because it's highly non-trivial. There are also some facts related to this alternative definition, show in Corollary 1.2.1 and Corollary 2.2.3.

Corollary 2.2.3 *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 now derive a lemma for gradient descent about how much the function decreases.

Lemma 2.3 (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 2.4 (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 2.5 (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 2.6 (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

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

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

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

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

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

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

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

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

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

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

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

□

3 Momentum is All You Need

The intuition behind **accelerated gradient descent** is to use larger learning rate $\eta > \frac{1}{L}$ without entering zig-zag. Some functions are only non-smooth at some corners, and using a low learning rate $\frac{1}{L}$ might not be the most efficient option.

The key idea of momentum is to use a universal large learning rate and use the "weighted" sum of the gradients from the previous iterations to update the current point. When gradients point to the same direction, the sum will be large. Otherwise, when the gradients bump back and forth, the sum will be small. The weighted sum of the past gradients is called the **momentum**.

Definition 3.1 (Momentum) *Use a learning rate $\eta > \frac{1}{L}$, and update using*

$$x_{t+1} = x_t - \eta g_t.$$

Definition 3.2 (Nesterov's Accelerated Gradient Descent) *For a L-smooth function f, in each iteration, we compute*

$$z_{t+1} = x_t - \eta \nabla f(x_t)$$
$$x_{t+1} = (1 - \gamma_t) z_{t+1} + \gamma_t z_t$$
$$\lambda_0 = 0, \lambda_t = \frac{1 + \sqrt{1 + 4\lambda_{t-1}^2}}{2}, \qquad \gamma_t = \frac{1 - \lambda_t}{\lambda_{t+1}}.$$

Nesterov's accelerated gradient descent is guaranteed to work mathematically, but in practice it's not the best momentum. In contrast, people use the **heavy ball momentum** more often, for example in PyTorch `optim.SGD(momentum=0.9)`, but it doesn't have theoretical guarantee.

Definition 3.3 (Heavy Ball Momentum) *The update can be approximated by*

$$x_{t+1} \approx x_t - \eta g_t$$

$$g_t = \gamma \sum_{s \leq t} (1-\gamma)^{t-s} \nabla f(x_s)$$

where the last step can be updated easily using

$$g_{t+1} = g_t(1-\gamma) + \gamma \nabla f(x_{t+1}).$$

Therefore, we can choose $\eta = \frac{1}{\gamma L} > \frac{1}{L}$.

There is no proof for why heavy ball momentum works, but we can do a thought experiment to show why this works intuitively. The key observation is that when gradient is smaller than usual, we can use a larger learning rate. We fix a value $K > 0$, if $\|f(x_t)\|_2^2 \geq K$ holds for **every** t, using $\eta = \frac{1}{L}$ and the gradient descent lemma, we have

$$f(x_{t+1}) \leq f(x_t) - \frac{K}{2L}$$

so we need at most $\frac{Lf(x_0)}{K}$ iterations to find a point x_T with $f(x_T) \leq \frac{f(x_0)}{2}$. If $\|f(x_t)\|_2^2 < K$ holds for **every** t, using the telescoping sum of mirror descent lemma, we have

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

with the assumption that $f(x^*) = 0$. With $\eta = \frac{f(x_0)}{2K}$, we need at most $\frac{4K\|x_0 - x^*\|_2^2}{f(x_0)^2}$ iterations to find a point x_T with $f(x_T) \leq \frac{f(x_0)}{2}$.

Therefore, picking $K = \sqrt{\frac{Lf^3(x_0)}{4\|x_0 - x^*\|_2^2}}$, in both cases we need at most $\frac{2\|x_0 - x^*\|_2 \sqrt{L}}{\sqrt{f(x_0)}}$ iterations to find a point x_T with $f(x_T) \leq \frac{f(x_0)}{2}$. In the second case, when $f(x_0) \approx \|x_0 - x^*\|_2 \approx 1$, the learning rate is indeed much larger $\eta = \frac{f(x_0)}{K} \approx \frac{1}{\sqrt{L}} > \frac{1}{L}$. Therefore, the convergence of the thought experiment is, for $\varepsilon > 0$, we need at most $\frac{\sqrt{2L}}{\sqrt{\varepsilon}}$ iterations to find a point x_{T_ε} with $f(T_\varepsilon) \leq \varepsilon$.

However, it's likely a function is neither both cases for **every** t. To choose a reasonable K, we can combine gradient descent and momentum.

Definition 3.4 (Linear Coupling) *For a* $0 \leq \tau \leq 1$, *at every iteration, compute* $s_{t+1} = x_t - \frac{1}{L}\nabla f(x_t)$ *(gradient descent) and* $l_{t+1} = l_t - \eta \nabla f(x_t)$ *(momentum), and the updated value* $x_{t+1} = (1-\tau)s_{t+1} + \tau l_{t+1}$.