

# Course Notes for EE227C (Spring 2018): Convex Optimization and Approximation

Instructor: Moritz Hardt

Email: [hardt+ee227c@berkeley.edu](mailto:hardt+ee227c@berkeley.edu)

Graduate Instructor: Max Simchowitz

Email: [msimchow+ee227c@berkeley.edu](mailto:msimchow+ee227c@berkeley.edu)

March 8, 2018

## 2 Lecture 2: Gradient method

In this lecture we encounter the fundamentally important *gradient method* and a few ways to analyze its convergence behavior. The goal here is to solve a problem of the form

$$\min_{x \in \Omega} f(x)$$

where we'll make some additional assumptions on the function  $f: \Omega \rightarrow \mathbb{R}$ . The technical exposition closely follows the corresponding chapter in Bubeck's text [Bub15].

### 2.1 Gradient descent

For a differentiable function  $f$ , the basic gradient method starting from an initial point  $x_0$  is defined by the iterative description

$$x_{t+1} = x_t - \eta \nabla_t f(x_t) \quad (t \geq 0)$$

where  $\eta_t$  is the so-called *step size* that may vary with  $t$ .

The first assumption that leads to a convergence analysis is that the gradients of the function aren't too big over the domain. This turns out to follow from a natural Lipschitz continuity assumption.

**Definition 2.1** (*L-Lipschitz*). A function  $f: \Omega \rightarrow \mathbb{R}$  is *L-Lipschitz* if for every  $x, y \in \Omega$ , we have

$$|f(x) - f(y)| \leq L \|x - y\|$$

**Fact 2.2.** *If the function  $f$  is  $L$ -Lipschitz, differentiable, and convex, then*

$$\|\nabla f(x)\| \leq L.$$

How can we ensure that  $x_{t+1} \in \Omega$ ? One natural approach is to “project” each iterate back onto the domain  $\Omega$ .

**Definition 2.3** (Projection). The *projection* of a point  $x$  onto a set  $\Omega$  is defined as

$$\Pi_{\Omega}(x) = \arg \min_{y \in \Omega} \|x - y\|$$

**Example 2.4.** A projection onto the Euclidean ball  $B_2$  is just normalization:

$$\Pi_{B_2}(x) = \frac{x}{\|x\|}$$

A crucial property of projections is that when  $x \in \Omega$ , we have for any  $y$  (possibly outside  $\Omega$ ):

$$\|\Pi_{\Omega}(y) - x\|^2 \leq \|y - x\|^2$$

That is, the projection of  $y$  onto a convex set containing  $x$  is closer to  $x$ .

**Lemma 2.5.**

$$\|\Pi_{\Omega}(y) - x\|^2 \leq \|y - x\|^2 - \|y - \Pi_{\Omega}(y)\|^2$$

*Which follows from the Pythagorean theorem. Note that this lemma implies the above property.*

### 2.1.1 Modifying gradient descent with projections

So now we can modify our original procedure to use two steps.

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

$$x_{t+1} = \Pi_{\Omega}(y_{t+1})$$

And we are guaranteed that  $x_{t+1} \in \Omega$ . Note that computing the projection may be computationally the hardest part of the problem. However, there are convex sets for which we know explicitly how to compute the projection (see [Example 2.4](#)). We will see several other non-trivial examples in later lectures.

## 2.2 Convergence rate of gradient descent for Lipschitz functions

**Theorem 2.6** (Projected Gradient Descent for  $L$ -Lipschitz Functions). *Assume that function  $f$  is convex, differentiable, and closed with bounded gradients. Let  $L$  be the Lipschitz constant of  $f$  over the convex domain  $\Omega$ . Let  $R$  be the upper bound on the distance  $\|x_1 - x^*\|_2$  from the initial point  $x_1$  to the optimal point  $x^* = \arg \min_{x \in \Omega} f(x)$ . Let  $t$  be the number of iterations of project gradient descent. If the learning rate  $\eta$  is set to  $\eta = \frac{R}{L\sqrt{t}}$ , then*

$$f\left(\frac{1}{t} \sum_{s=1}^t x_s\right) - f(x^*) \leq \frac{RL}{\sqrt{t}}.$$

This means that the difference between the functional value of the average point during the optimization process from the optimal value is bounded above by a constant proportional to  $\frac{1}{\sqrt{t}}$ .

Before proving the theorem, recall the “Fundamental Theorem of Optimization”, which is that an inner product can be written as a sum of norms:  $u^\top v = \frac{1}{2}(\|u\|^2 + \|v\|^2 - \|u - v\|^2)$ . This property can be seen from  $\|u - v\|^2 = \|u\|^2 + \|v\|^2 - 2u^\top v$ .

*Proof of Theorem 2.6.* The proof begins by first bounding the difference in function values  $f(x_s) - f(x^*)$ .

$$f(x_s) - f(x^*) \leq \nabla f(x_s)^\top (x_s - x^*) \quad (1)$$

$$= \frac{1}{\eta} (x_s - y_{s+1})^\top (x_s - x^*) \quad (2)$$

$$= \frac{1}{2\eta} (\|x_s - x^*\|^2 + \|x_s - y_{s+1}\|^2 - \|y_{s+1} - x^*\|^2) \quad (3)$$

$$= \frac{1}{2\eta} (\|x_s - x^*\|^2 - \|y_{s+1} - x^*\|^2) + \frac{\eta}{2} \|\nabla f(x_s)\|^2 \quad (4)$$

$$\leq \frac{1}{2\eta} (\|x_s - x^*\|^2 - \|y_{s+1} - x^*\|^2) + \frac{\eta L^2}{2} \quad (5)$$

$$\leq \frac{1}{2\eta} (\|x_s - x^*\|^2 - \|x_{s+1} - x^*\|^2) + \frac{\eta L^2}{2} \quad (6)$$

Equation 1 comes from the definition of convexity. Equation 2 comes from the update rule for projected gradient descent. Equation 3 comes from the “Fundamental Theorem of Optimization.” Equation 4 comes from the update rule for projected gradient descent. Equation 5 is because  $f$  is  $L$ -Lipschitz. Equation 6 comes from Lemma 2.5.

Now, sum these differences from  $s = 1$  to  $s = t$ :

$$\sum_{s=1}^t f(x_s) - f(x^*) \leq \frac{1}{2\eta} \sum_{s=1}^t (\|x_s - x^*\|^2 - \|x_{s+1} - x^*\|^2) + \frac{\eta L^2 t}{2} \quad (7)$$

$$= \frac{1}{2\eta} (\|x_1 - x^*\|^2 - \|x_t - x^*\|^2) + \frac{\eta L^2 t}{2} \quad (8)$$

$$\leq \frac{1}{2\eta} \|x_1 - x^*\|^2 + \frac{\eta L^2 t}{2} \quad (9)$$

$$\leq \frac{R^2}{2\eta} + \frac{\eta L^2 t}{2} \quad (10)$$

Equation 8 is because Equation 7 is a telescoping sum. Equation 9 is because  $\|x_t - x^*\|^2 \geq 0$ . Equation 10 is by the assumption that  $\|x_1 - x^*\|^2 \leq R^2$ .

Finally,

$$\begin{aligned}
f\left(\frac{1}{t} \sum_{s=1}^t x_s\right) - f(x^*) &\leq \frac{1}{t} \sum_{s=1}^t f(x_s) - f(x^*) && \text{(by convexity)} \\
&\leq \frac{R^2}{2\eta t} + \frac{\eta L^2}{2} && \text{(by Equation 10)} \\
&\leq \frac{RL}{\sqrt{t}} && \text{(for } \eta = R/L\sqrt{t}.)
\end{aligned}$$

■

## 2.3 Convergence rate for smooth functions

The next property we'll encounter is called *smoothness* and it often leads to stronger convergence guarantees.

**Definition 2.7** ( $\beta$ -smoothness). A continuously differentiable function  $f$  is  $\beta$  smooth if the gradient  $\nabla f$  is  $\beta$ -Lipschitz, i.e

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|$$

**Lemma 2.8.** Let  $f$  be a  $\beta$ -smooth function on  $\mathbb{R}^n$ . Then for any  $x, y \in \mathbb{R}^n$ , one has

$$|f(x) - f(y) - \nabla f(y)^\top (x - y)| \leq \frac{\beta}{2} \|x - y\|^2$$

*Proof.* Express  $f(x) - f(y)$  as an integral, then apply Cauchy-Schwarz and  $\beta$ -smoothness as follows:

$$\begin{aligned}
|f(x) - f(y) - \nabla f(y)^\top (x - y)| &= \left| \int_0^1 \nabla f(y + t(x - y))^\top (x - y) dt - \nabla f(y)^\top (x - y) \right| \\
&\leq \int_0^1 \|\nabla f(y + t(x - y)) - \nabla f(y)\| \cdot \|x - y\| dt \\
&\leq \int_0^1 \beta t \|x - y\|^2 dt \\
&= \frac{\beta}{2} \|x - y\|^2
\end{aligned}$$

■

We also need the following lemma.

**Lemma 2.9.** Let  $f$  be a  $\beta$ -smooth convex function, then for any  $x, y \in \mathbb{R}^n$ , one has

$$f(x) - f(y) \leq \nabla f(x)^\top (x - y) - \frac{1}{2\beta} \|\nabla f(x) - \nabla f(y)\|^2$$

*Proof.* Let  $z = y - \frac{1}{\beta}(\nabla f(y) - \nabla f(x))$ . Then,

$$\begin{aligned}
f(x) - f(y) &= f(x) - f(z) + f(z) - f(y) \\
&\leq \nabla f(x)^\top (x - z) + \nabla f(y)^\top (z - y) + \frac{\beta}{2} \|z - y\|^2 \\
&= \nabla f(x)^\top (x - y) + (\nabla f(x) - \nabla f(y))^\top (y - z) + \frac{1}{2\beta} \|\nabla f(x) - \nabla f(y)\|^2 \\
&= \nabla f(x)^\top (x - y) - \frac{1}{2\beta} \|\nabla f(x) - \nabla f(y)\|^2
\end{aligned}$$

Here, the inequality follows from convexity and smoothness. ■

We will show that gradient descent with the update rule

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

attains a faster rate of convergence under the smoothness condition.

**Theorem 2.10.** *Let  $f$  be convex and  $\beta$ -smooth on  $\mathbb{R}^n$  then gradient descent with  $\eta = \frac{1}{\beta}$  satisfies*

$$f(x_t) - f(x^*) \leq \frac{2\beta \|x_1 - x^*\|^2}{t - 1}$$

To prove this we will need the following two lemmas.

*Proof.* By the update rule and lemma 2.8 we have

$$f(x_{s+1}) - f(x_s) \leq -\frac{1}{2\beta} \|\nabla f(x_s)\|^2$$

In particular, denoting  $\delta_s = f(x_s) - f(x^*)$  this shows

$$\delta_{s+1} \leq \delta_s - \frac{1}{2\beta} \|\nabla f(x_s)\|^2$$

One also has by convexity

$$\delta_s \leq \nabla f(x_s)^\top (x_s - x^*) \leq \|x_s - x^*\| \cdot \|\nabla f(x_s)\|$$

We will prove that  $\|x_s - x^*\|$  is decreasing with  $s$ , which with the two above displays will imply

$$\delta_{s+1} \leq \delta_s - \frac{1}{2\beta \|x_1 - x^*\|^2} \delta_s^2$$

We solve the recurrence as follows. Let  $w = \frac{1}{2\beta \|x_1 - x^*\|^2}$ , then

$$w\delta_s^2 + \delta_{s+1} \leq \delta_s \iff w\frac{\delta_s}{\delta_{s+1}} + \frac{1}{\delta_s} \leq \frac{1}{\delta_{s+1}} \implies \frac{1}{\delta_{s+1}} - \frac{1}{\delta_s} \geq w \implies \frac{1}{\delta_t} \geq w(t - 1)$$

To finish the proof it remains to show that  $\|x_s - x^*\|$  is decreasing with  $s$ . Using lemma 2.9 one immediately gets

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geq \frac{1}{\beta} \|\nabla f(x) - \nabla f(y)\|^2$$

We use this and the fact that  $\nabla f(x^*) = 0$

$$\|x_{s+1} - x^*\|^2 = \|x_s - \frac{1}{\beta} \nabla f(x_s) - x^*\|^2 \tag{11}$$

$$= \|x_s - x^*\|^2 - \frac{2}{\beta} \nabla f(x_s)^\top (x_s - x^*) + \frac{1}{\beta^2} \|\nabla f(x_s)\|^2 \tag{12}$$

$$\leq \|x_s - x^*\|^2 - \frac{1}{\beta^2} \|\nabla f(x_s)\|^2 \tag{13}$$

$$\leq \|x_s - x^*\|^2 \tag{14}$$

which concludes the proof. ■

## References

[Bub15] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.