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

Instructor: Moritz Hardt

Email: hardt+ee227c@berkeley.edu

Graduate Instructor: Max Simchowitz

Email: msimchow+ee227c@berkeley.edu

April 27, 2018

## 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 $f : \Omega \to \mathbb{R}$. 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 $x_1$ is defined by the iterative update rule

$$x_{t+1} = x_t - \eta_t \nabla f(x_t)\,, \qquad t = 1, 2, \ldots$$

where the scalar $\eta_t$ is the so-called *step size*, sometimes called *learning rate*, that may vary with $t$. There are numerous ways of choosing step sizes that have a significant effect on the performance of gradient descent. What we will see in this lecture are several choices of step sizes that ensure the convergence of gradient descent by virtue of a theorem. These step sizes are not necessarily ideal for practical applications.

### 2.1.1 Projections

In cases where the constraint set $\Omega$ is not all of $\mathbb{R}^n$, the gradient update can take us outside the domain $\Omega$. How can we ensure that $x_{t+1} \in \Omega$? One natural approach is to "project" each iterate back onto the domain $\Omega$. As it turns out, this won't really make our analysis more difficult and so we include from the get-go.

**Definition 2.1** (Projection)**.** The *projection* of a point $x$ onto a set $\Omega$ is defined as

$$\Pi_\Omega(x) = \arg\min_{y \in \Omega} \|x - y\|_2 .$$

**Example 2.2.** 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 $\Omega$):

$$\|\Pi_\Omega(y) - x\|^2 \leqslant \|y - x\|^2$$

That is, the projection of $y$ onto a convex set containing $x$ is closer to $x$. In fact, a stronger claim is true that follows from the Pythagorean theorem.

**Lemma 2.3.**

$$\|\Pi_\Omega(y) - x\|^2 \leqslant \|y - x\|^2 - \|y - \Pi_\Omega(y)\|^2$$

So, now we can modify our original procedure as displayed in Figure 1.

| Starting from $x_1 \in \Omega$, repeat: | |
|---|---|
| $y_{t+1} = x_t - \eta \nabla f(x_t)$ | (gradient step) |
| $x_{t+1} = \Pi_\Omega(y_{t+1})$ | (projection) |

Figure 1: Projected gradient descent

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 2.2). We will see several other non-trivial examples in later lectures.

## 2.2 Lipschitz functions

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

**Definition 2.4** ($L$-Lipschitz)**.** A function $f : \Omega \to \mathbb{R}$ is *$L$-Lipschitz* if for every $x, y \in \Omega$, we have

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

**Fact 2.5.** *If the function $f$ is $L$-Lipschitz, differentiable, and convex, then*

$$\|\nabla f(x)\| \leqslant L\,.$$

We can now prove our first convergence rate for gradient descent.

**Theorem 2.6.** *Assume that function $f$ is convex, differentiable, and $L$-Lipschitz 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 an optimal point $x^* \in \arg\min_{x\in\Omega} f(x)$. Let $x_1, \ldots, x_t$ be the sequence of iterates computed by $t$ steps of projected gradient descent with constant step size $\eta = \frac{R}{L\sqrt{t}}$. Then,*

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

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:

$$u^\top v = \frac{1}{2}(\|u\|^2 + \|v\|^2 - \|u - v\|^2) \tag{1}$$

This property follows from the more familiar identity $\|u - v\|^2 = \|u\|^2 + \|v\|^2 - 2u^\top v$.

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

$$\begin{aligned}
f(x_s) - f(x^*) &\leqslant \nabla f(x_s)^\top (x_s - x^*) && \text{(by convexity)}\\
&= \frac{1}{\eta}(x_s - y_{s+1})^\top (x_s - x^*) && \text{(by the update rule)}\\
&= \frac{1}{2\eta}\left(\|x_s - x^*\|^2 + \|x_s - y_{s+1}\|^2 - \|y_{s+1} - x^*\|^2\right) && \text{(by Equation 1)}\\
&= \frac{1}{2\eta}\left(\|x_s - x^*\|^2 - \|y_{s+1} - x^*\|^2\right) + \frac{\eta}{2}\|\nabla f(x_s)\|^2 && \\
& && \text{(by the update rule)}\\
&\leqslant \frac{1}{2\eta}\left(\|x_s - x^*\|^2 - \|y_{s+1} - x^*\|^2\right) + \frac{\eta L^2}{2} && \text{(Lipschitz condition)}\\
&\leqslant \frac{1}{2\eta}\left(\|x_s - x^*\|^2 - \|x_{s+1} - x^*\|^2\right) + \frac{\eta L^2}{2} && \text{(Lemma 2.3)}
\end{aligned}$$

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

$$\begin{aligned}
\sum_{s=1}^{t} f(x_s) - f(x^*) &\leqslant \frac{1}{2\eta} \sum_{s=1}^{t} \left( \|x_s - x^*\|^2 - \|x_{s+1} - x^*\|^2 \right) + \frac{\eta L^2 t}{2} \\
&= \frac{1}{2\eta} \left( \|x_1 - x^*\|^2 - \|x_t - x^*\|^2 \right) + \frac{\eta L^2 t}{2} && \text{(telescoping sum)} \\
&\leqslant \frac{1}{2\eta} \|x_1 - x^*\|^2 + \frac{\eta L^2 t}{2} && \text{(since } \|x_t - x^*\| \geqslant 0\text{)} \\
&\leqslant \frac{R^2}{2\eta} + \frac{\eta L^2 t}{2} && \text{(since } \|x_1 - x^*\| \leqslant R\text{)}
\end{aligned}$$

Finally,

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

■

## 2.3 Smooth functions

The next property we'll encounter is called *smoothness*. The main point about smoothness is that it allows us to control the second-order term in the Taylor approximation. This often leads to stronger convergence guarantees at the expense of a relatively strong assumption.

**Definition 2.7** (Smoothness)**.** A continuously differentiable function f is $\beta$ smooth if the gradient gradient map $\nabla f \colon \mathbb{R}^n \to \mathbb{R}^n$ is $\beta$-Lipschitz, i.e,

$$\|\nabla f(x) - \nabla f(y)\| \leqslant \beta \|x - y\| .$$

We will need a couple of technical lemmas before we can analyze gradient descent for smooth functions. It's safe to skip the proof of these technical lemmas on a first read.

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

$$\left| f(x) - f(y) - \nabla f(y)^\top (x - y) \right| \leqslant \frac{\beta}{2} \|x - y\|^2 .$$

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

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

■

We also need the following lemma.

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

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

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

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

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*

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

To prove this we will need the following two lemmas.

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

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

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

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

One also has by convexity

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

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

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

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

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

To finish the proof it remains to show that $\|x_s - x^*\|$ is decreasing with $s$. Using Lemma 2.9, we get

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

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

$$\begin{aligned}
\|x_{s+1} - x^*\|^2 &= \|x_s - \frac{1}{\beta}\nabla f(x_s) - x^*\|^2 \\
&= \|x_s - x^*\|^2 - \frac{2}{\beta}\nabla f(x_s)^\top (x_s - x^*) + \frac{1}{\beta^2}\|\nabla f(x_s)\|^2 \\
&\leqslant \|x_s - x^*\|^2 - \frac{1}{\beta^2}\|\nabla f(x_s)\|^2 \\
&\leqslant \|x_s - x^*\|^2 .
\end{aligned}$$

■

# References


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