

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

January 24, 2018

2 Lecture 2: Gradient Descent

2.1 Gradient Descent

The procedure of gradient descent is defined by the recursion:

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

where η is the step size. This works to solve the problem

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

for f convex, differentiable, and L -Lipschitz

Definition 2.1 (L -Lipschitz). A function is said to be L -Lipschitz if its gradient is bounded,

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

Fact 2.2. $f(x)$ is L -Lipschitz implies that the difference between two points in the range is bounded,

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

Question 2.1. How do we ensure that $x_{t+1} \in \Omega$?

Solution: Project onto Ω

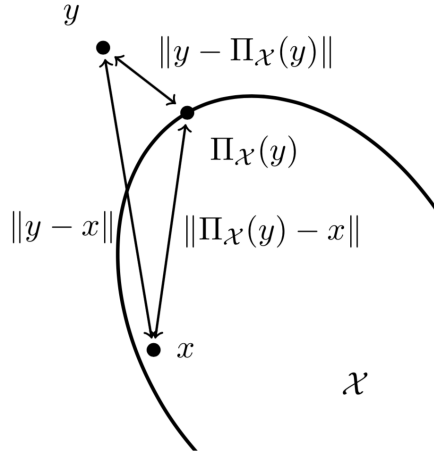


Figure 1: Projection of y onto set \mathcal{X} . (Figure taken from Bubeck)

Definition 2.3 (Projection). The *projection* of a point y onto a set Ω is defined as

$$\Pi_{\Omega}(x) = \underset{y \in \Omega}{\operatorname{argmin}} \|x - y\|$$

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

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

The crucial property of projections is that they satisfy the following condition:

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

i.e. the projection of y onto a convex set containing x is closer to x . See [Figure 1](#) for a geometric picture.

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 the hardest part of your problem, as you are computing an argmin. However, there are convex sets for which we know explicitly how to compute the projection (see [Example 2.4](#)).

2.2 Convergence Rate of Gradient Descent for L -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 Ω . 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 η 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 by writing $\|u - v\|^2$ as $\|u\|^2 + \|v\|^2 - 2u^\top v$.

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

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

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

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

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

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

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

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 -Lipchitz. Equation 6 comes from Lemma 2.5.

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

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

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

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

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

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 $\|x_1 - x^*\|^2 \leq R^2$.

Then bound $f\left(\frac{1}{t} \sum_{s=1}^t x_s\right) - f(x^*)$ by the above sum:

$$f\left(\frac{1}{t} \sum_{s=1}^t x_s\right) \leq \frac{1}{t} \sum_{s=1}^t f(x_s) \quad (11)$$

$$\left(\frac{1}{t} \sum_{s=1}^t x_s\right) - f(x^*) \leq \frac{1}{t} \sum_{s=1}^t f(x_s) - f(x^*) \quad (12)$$

$$\leq \frac{R^2}{2\eta} + \frac{\eta L^2 t}{2} \quad (13)$$

Equation 11 is by convexity. $\frac{R^2}{2\eta} + \frac{\eta L^2 t}{2}$, the upper bound of the difference between $f\left(\frac{1}{t} \sum_{s=1}^t x_s\right)$ and $f(x^*)$ is minimized when η is set to be $\frac{RL}{\sqrt{t}}$. ■

2.3 Convergence Rate of Gradient Descent for β -Smooth Functions

Definition 2.7 (β -smoothness). A continuously differentiable function f is β smooth if the gradient ∇f is β -Lipschitz, i.e

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

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 2.8. Let f be convex and β -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.

Lemma 2.9. Let f be a β -smooth function on \mathbb{R}^n . Then for any $x, y \in \mathbb{R}^n$, one has

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

Proof. First represent $f(x) - f(y)$ as an integral, apply Cauchy-Schwarz and then β -smoothness:

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

■

We also need

Lemma 2.10. *Let f be a β -smooth function, then for any $x, y \in \mathbb{R}^n$, one has*

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

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

$$f(x) - f(y)$$

$$= f(x) - f(z) + f(z) - f(y) \tag{14}$$

$$\leq \nabla f(x)^T(x - z) + \nabla f(y)^T(z - y) + \frac{\beta}{2} \|z - y\|^2 \tag{15}$$

$$= \nabla f(x)^T(x - y) + (\nabla f(x) - \nabla f(y))^T(y - z) + \frac{1}{2\beta} \|\nabla f(x) - \nabla f(y)\|^2 \tag{16}$$

$$= \nabla f(x)^T(x - y) - \frac{1}{2\beta} \|\nabla f(x) - \nabla f(y)\|^2 \tag{17}$$

■

Now we prove the theorem

Proof. By the update rule and lemma 2.9 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

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

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 2.10 one immediately gets

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

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

$$\|x_{s+1} - x^*\|^2 = \|x_s - \frac{1}{\beta} \nabla f(x_s) - x^*\|^2 \tag{18}$$

$$= \|x_s - x^*\|^2 - \frac{2}{\beta} \nabla f(x_s)^T(x_s - x^*) + \frac{1}{\beta^2} \|\nabla f(x_s)\|^2 \tag{19}$$

$$\leq \|x_s - x^*\|^2 - \frac{1}{\beta^2} \|\nabla f(x_s)\|^2 \tag{20}$$

$$\leq \|x_s - x^*\|^2 \tag{21}$$

which concludes the proof. ■