

## Lecture 10: Convex Optimization (Draft)

Date: February 14, 2023

Author: Surbhi Goel

**Acknowledgements.** These notes are heavily inspired by material from Sham Kakade (Harvard) and Elad Hazan (Princeton).

**Disclaimer.** These notes have not been subjected to the usual scrutiny reserved for formal publications. If you notice any typos or errors, please reach out to the author.

## 1 Convex Optimization

In the course so far, we have often modelled our learning problem as a loss minimization over some constraint set. In Lecture 7, we looked at how to formulate these problems and when we can solve them using existing optimizers. In this lecture, we will look at perhaps the most intuitive and powerful tool to optimize our objective: *gradient descent*.

So here our goal will be the following:

$$\begin{array}{ll} \text{minimize over } w & \underbrace{F(w)}_{\text{objective}} \\ \text{such that} & \underbrace{w \in \mathcal{C}}_{\text{constraint}} \end{array}$$

Usually  $\mathcal{C} \subseteq \mathbb{R}^d$  and  $F : \mathbb{R}^d \rightarrow \mathbb{R}$ . The problem is unconstrained when  $\mathcal{C} = \mathbb{R}^d$ . In this lecture, we will focus on convex optimization problems where  $F$  is convex and  $\mathcal{C}$  is a convex set.

### 1.1 Convex functions and sets

Recall the definitions of convex function and sets.

**Definition 1** (Convex function). A function  $F : \mathbb{R}^d \rightarrow \mathbb{R}$  is convex if for all  $w, w' \in \mathbb{R}^d$  and  $\alpha \in [0, 1]$ ,

$$F(\alpha w + (1 - \alpha)w') \leq \alpha F(w) + (1 - \alpha)F(w').$$

**Definition 2** (Convex set). A set  $\mathcal{C} \subseteq \mathbb{R}^d$  is convex if for all  $w, w' \in \mathcal{C}$  and  $\alpha \in [0, 1]$ ,

$$\alpha w + (1 - \alpha)w' \in \mathcal{C}.$$

### 1.2 First and second order characterizations of convex functions

Suppose convex function  $F$  is twice differentiable, then the following statements are true:

- for all  $w, w' \in \mathbb{R}^d$ ,  $F(w') \geq F(w) + \nabla F(w)^\top (w' - w)$ .
- for all  $w \in \mathbb{R}^d$ ,  $\nabla^2 F(w) \succeq 0$ , that is,  $\nabla^2 F(w)$  is a PSD matrix.

The first statement says that the function always lies above the tangent at any point. The second says that the curvature of the function is always non-negative, that is, never downwards.

*Try to show that these statements are true from the definition of convex functions. For the latter one, try to show in 1-dimension first.*

**Theorem 3.** *For any convex differentiable function  $F$ , any  $w$  that satisfies  $\nabla F(w) = 0$  is a global minimum of  $F$ .*

*Proof.* From the first property above, we have for all  $w'$ ,

$$F(w') \geq F(w) + \nabla F(w)^\top (w' - w) \implies F(w') \geq F(w),$$

since  $\nabla F(w) = 0$ . □

This property highlights the local to global phenomenon, if the minimum is a local minimum then it is a global minimum. Also, this minimum is unique if  $F$  is strictly convex.

## 2 Gradient Descent

Let us first focus on the unconstrained setup, where  $\mathcal{C} = \mathbb{R}^d$ . Consider the following algorithm:

---

**Algorithm 1:** Gradient Descent (GD)

---

```

Initialize  $w_1 \in \mathbb{R}^d$ 
while  $t = 1, 2, \dots, T$  do
    Update  $w_{t+1} = w_t - \eta_t \nabla F(w_t)$ 
end
```

---

Here  $\eta_t$  are known as the learning rates. Usually we use a stopping condition to terminate the algorithm, for instance, when  $F(w_t) \leq \epsilon$  or when  $\|\nabla F(w_t)\|_2 \leq \epsilon$ .

**Choosing learning rate/ step size.** Choosing a good learning rate is paramount to the success of gradient descent. If the learning rate is too large, then we could diverge, that is, instead of making progress at each time step, we could instead make negative progress. If the learning rate is too small, then it could take us a very long time to converge to a good solution. One way of setting the learning rate is some factor (say 10) smaller than when the learning rate at which things tend to diverge.

*In modern machine learning, several very crazy learning rate schedules work well, counter-intuitive to what we can get guarantees for in convex problems. This is a very active area of research.*

### 3 GD for Smooth Functions

Let us look at convex functions that are smooth, that is, they don't change value drastically. More formally,

**Definition 4** (Smooth function). *A function  $F : \mathbb{R}^d \rightarrow \mathbb{R}$  is  $L$ -smooth if for all  $w, w' \in \mathbb{R}^d$ ,*

$$F(w') \leq F(w) + \nabla F(w)^\top (w' - w) + \frac{L}{2} \|w' - w\|_2^2.$$

This condition is equivalent to saying that the  $\nabla F$  is  $L$ -Lipschitz, that is, for all  $w, w' \in \mathbb{R}^d$

$$\|\nabla F(w) - \nabla F(w')\|_2 \leq L \|w - w'\|_2.$$

*We will prove this in the homework.*

**Interpretation of gradient descent.** For a smooth function, we have

$$F(w') \leq F(w) + \nabla F(w)^\top (w' - w) + \frac{L}{2} \|w' - w\|_2^2.$$

One way to locally minimize  $F$  at  $w$  would be to minimize the upper bound above,

$$\min_{w'} F(w) + \nabla F(w)^\top (w' - w) + \frac{L}{2} \|w' - w\|_2^2.$$

Differentiating, this gives us

$$w' = w - \frac{1}{L} \nabla F(w).$$

This is exactly the gradient step with learning rate  $\eta = 1/L$ .

#### 3.1 Proof of Convergence

**Theorem 5.** *Suppose we run GD on  $L$ -smooth function  $F$  with fixed constant learning rate  $\eta_t = 1/L$  for all  $t \in [T]$ . Then for all time  $\tau$ , we have*

$$F(w_{\tau+1}) - F(w_*) \leq \frac{L \|w_1 - w_*\|_2^2}{2\tau}$$

*where  $w_*$  is the global minimum.*

Let us interpret this result. Suppose we initialize at  $w_1 = 0$  and  $\|w_*\|_2 = 1$ , then this implies that after  $T$  iterations

$$F(w_{T+1}) - F(w_*) \leq \frac{L}{2T}.$$

This implies that after  $T = \frac{L}{2\epsilon} = O(1/\epsilon)$  steps, we get that  $F(w_{T+1}) - F(w_*) \leq \epsilon$ . Therefore we say that gradient descent has a convergence rate of  $O(1/T)$ .

*Proof.* The proof follows in three steps:

- **Step 1:** Upper bound the difference between function value at the next iterate and the current iterate for every time  $t$ .

$$F(w_{t+1}) - F(w_t) \leq -\frac{L}{2} \|w_{t+1} - w_t\|^2.$$

This step shows that  $F(w_t)$  is non-increasing.

- **Step 2:** Upper bound the difference between function value at the next iterate and the global minimum for every time  $t$ .

$$F(w_{t+1}) - F(w_*) \leq \frac{L}{2} (\|w_t - w_*\|_2^2 - \|w_{t+1} - w_*\|_2^2).$$

This step shows that  $\|w_t - w_*\|_2$  is non-increasing.

- **Step 3:** Use the above to upper bound  $F(w_{\tau+1}) - F(w_*)$ , that is, the difference after  $\tau$  iterates and the global minimum.

**Step :** By the definition of smoothness, we have at time  $t$ ,

$$F(w_{t+1}) - F(w_t) \leq \nabla F(w_t)^\top (w_{t+1} - w_t) + \frac{L}{2} \|w_{t+1} - w_t\|_2^2.$$

Substituting for  $L(w_t - w_{t+1}) = \nabla F(w_t)$  gives us,

$$\begin{aligned} F(w_{t+1}) - F(w_t) &\leq \nabla L(w_t - w_{t+1})^\top (w_{t+1} - w_t) + \frac{L}{2} \|w_{t+1} - w_t\|_2^2 \\ &= -L \|w_{t+1} - w_t\|_2^2 + \frac{L}{2} \|w_{t+1} - w_t\|_2^2 \\ &= -\frac{L}{2} \|w_{t+1} - w_t\|^2. \end{aligned} \tag{1}$$

The above shows that  $F(w_t)$  is decreasing with every step.

**Step 2:** By convexity, we have

$$\begin{aligned} F(w_*) &\geq F(w_t) + \nabla F(w_t)^\top (w_* - w_t) \\ \implies F(w_t) - F(w_*) &\leq L(w_t - w_{t+1})^\top (w_t - w_*). \end{aligned} \tag{2}$$

Observe that

$$\begin{aligned} \|w_{t+1} - w_*\|_2^2 &= \|w_{t+1} - w_t + w_t - w_*\|_2^2 \\ &= \|w_{t+1} - w_t\|_2^2 + \|w_t - w_*\|_2^2 + 2(w_{t+1} - w_t)^\top (w_t - w_*). \end{aligned}$$

This implies

$$(w_t - w_{t+1})^\top (w_t - w_*) = \frac{1}{2} (\|w_{t+1} - w_t\|_2^2 + \|w_t - w_*\|_2^2 - \|w_{t+1} - w_*\|_2^2). \tag{3}$$

Substituting (3) in (2), we get

$$F(w_t) - F(w_*) \leq \frac{L}{2} (\|w_{t+1} - w_t\|_2^2 + \|w_t - w_*\|_2^2 - \|w_{t+1} - w_*\|_2^2). \tag{4}$$

Adding (1) and (2) gives us,

$$\begin{aligned} F(w_{t+1}) - F(w_*) &\leq \frac{L}{2} (\|w_{t+1} - w_t\|_2^2 + \|w_t - w_*\|_2^2 - \|w_{t+1} - w_*\|_2^2) - \frac{L}{2} \|w_{t+1} - w_t\|^2 \\ &= \frac{L}{2} (\|w_t - w_*\|_2^2 - \|w_{t+1} - w_*\|_2^2) \end{aligned} \quad (5)$$

Since  $F(w_*)$  is the optimal, we have  $0 \geq F(w_{t+1}) - F(w_*)$ . Therefore, (5) gives us  $\|w_{t+1} - w_*\|_2 \leq \|w_t - w_*\|_2$ . this implies we are getting closer to  $w_*$  with each iteration.

**Step 3:** Now summing (5) over all  $t \leq \tau$  and using the fact that  $F(w_t)$  is non-increasing, we get

$$\tau(F(w_{\tau+1}) - F(w_*)) = \sum_{t=1}^{\tau} (F(w_{t+1}) - F(w_*)) \leq \frac{L}{2} (\|w_1 - w_*\|_2^2 - \|w_{\tau+1} - w_*\|_2^2) \leq \frac{L}{2} \|w_1 - w_*\|_2^2.$$

Rearranging the above gives us the desired result.  $\square$

Let us interpret this result. Suppose we initialize at  $w_1 = 0$  and  $\|w_*\|_2 = 1$ , then this implies that after  $T$  iterations

$$F(w_{T+1}) - F(w_*) \leq \frac{L}{2T}.$$

This implies that after  $T = \frac{L}{2\epsilon} = O(1/\epsilon)$  steps, we get that  $F(w_{T+1}) - F(w_*) \leq \epsilon$ . Therefore we say that gradient descent has a convergence rate of  $O(1/T)$ .

## 4 GD on Smooth and Strongly Convex Functions

Let us look at convex functions that are strongly convex, that is, they have high curvature at all points. More formally,

**Definition 6** (Strongly Convex function). *A function  $F : \mathbb{R}^d \rightarrow \mathbb{R}$  is  $\mu$ -strongly convex if for all  $w, w' \in \mathbb{R}^d$ ,*

$$F(w') \geq F(w) + \nabla F(w)^\top (w' - w) + \frac{\mu}{2} \|w' - w\|_2^2.$$

### 4.1 Proof of Convergence

**Theorem 7.** *Suppose we run GD on  $L$ -smooth and  $\mu$ -strongly convex function  $F$  with fixed constant learning rate  $\eta_t = 1/L$  for all  $t \in [T]$ . Then for all time  $\tau$ , we have*

$$\|w_{\tau+1} - w_*\|_2^2 \leq \left(1 - \frac{\mu}{L}\right)^\tau \|w_1 - w_*\|_2^2.$$

where  $w_*$  is the global minimum.

Let us interpret this result. Suppose we initialize at  $w_1 = 0$  and  $\|w_*\|_2 = 1$ , then this implies that after  $T$  iterations

$$\|w_{T+1} - w_*\|_2^2 \leq \left(1 - \frac{\mu}{L}\right)^T.$$

This implies that after  $T = \frac{L}{\mu} \log(1/\epsilon) = O(\log(1/\epsilon))$  steps, we get that  $\|w_{T+1} - w_*\|_2 \leq \epsilon$ . Therefore we say that gradient descent has a convergence rate of  $O(\exp(-T))$ . Note that this is exponentially faster than the previous case.