Optimization for Machine Learning 50.579

Instructor: Ioannis Panageas
Scribed by: Dan Jiang(1004712), Jacob Chen(1000308), Junhua Liu(1000378), Perry Lam(1000229)

Lecture 1. Convex Optimization and Gradient Descent.

1 Introduction

In machine learning tasks, especially for supervised learning, we always look for a function with some parameters $\theta \in \Theta$, that can minimize the distance between the real labels and prediction results, generated by the mentioned function. We call this "distance" as *Loss function*, or the *objective*.

Given n sample pairs of input data and labels (x_i, y_i), where x_i is the input (e.g. voice signal, pixels, ...) and y_i is the true label of each input, (e.g. gender of people, type of fruit ...), we want to minimize the (average) distance between the predicted label $f(x_i, \theta)$ and the true label:

$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n} l(f(x_i, \theta), y_i) \tag{1}$$

However, solving $min_{x\in\chi} L(\theta)$ in general is NP-hard(computational intractable). In this chapter, we restrict our objective to only *convex* functions for easier analysis, as they have strong theoretical *guarantees* and *efficient* optimization algorithms, and will be applying *Gradient Descent* to minimize the loss function.

2 Definitions

Let us first define some fundamental quantities to use later.

2.1 Convex Combination

$z \in \mathbb{R}^d$ is a *convex combination* of $x_1, x_2, \dots, x_n \in \mathbb{R}^d$ if:

$$z = \sum \lambda_i x_i,\ \lambda \geq \ for\ all\ i\ and\ \sum_i \lambda_i = 1 \tag{2}$$

2.2 Convex Set

$\mathcal{X}$ is a convex set if the *convex combination* of any two points in $\mathcal{X}$ also belongs in $\mathcal{X}$. The following figure depicts the relationship.

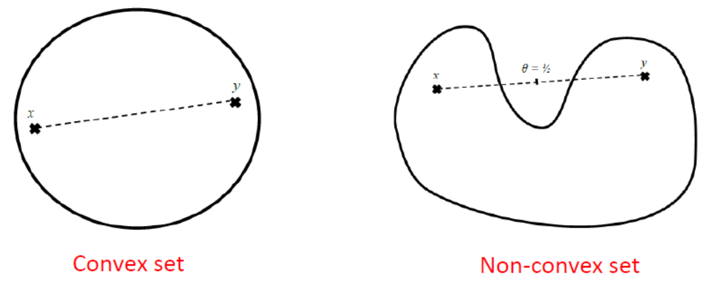

Figure 1: Schematic Diagram of Convex vs Non-Convex Set

2.3 Convex Function

A function $f(x)$ is convex iff the domain $dom(f)$ is a convex set and $\forall x, y \in dom(f), t \in [0, 1]$,

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) \tag{3}$$

This is known as Jensen's Inequality. Graphically, any line that intersects the function at two points should be above the function, as the following figure shows:

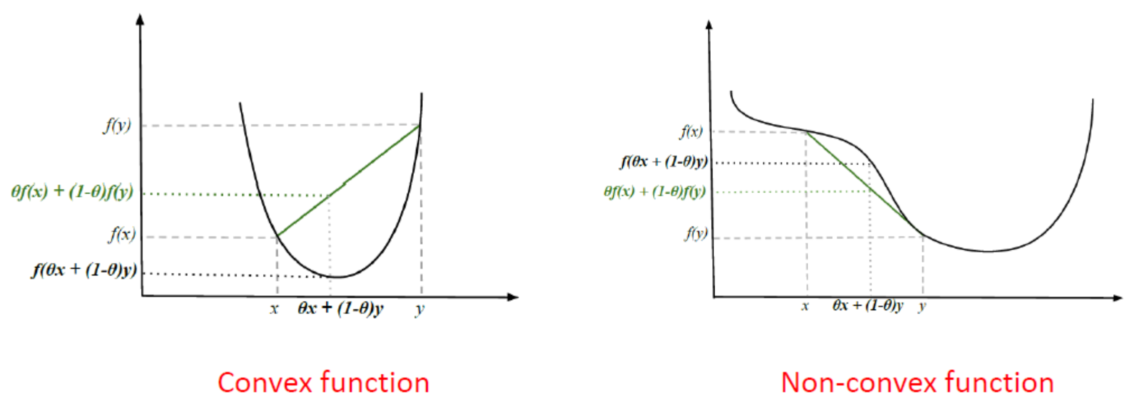

Figure 2: Convex vs. Non-Convex Function

Note: a *concave function* f will result in the reverse inequality. And f is called strictly convex when the inequality is $<$ instead of $\leq$.

2.4 Conditions for Convexity

Lemma 2.1 (First Order Condition (FOC)) *A differentiable function $f(x)$ is convex iff $dom(f)$ is a convex set and $\forall x, y \in dom(f)$,*

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) \tag{4}$$

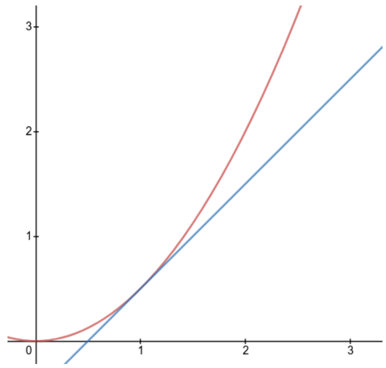

Figure 3: FOC for convexity: the tangent hyperplane at any point always gives values less than the function value of any other point.

Proof: If f is convex, then:

$$f(ty + (1-t)x) \leq tf(y) + (1-t)f(x) \tag{5}$$

Rearranging and dividing by t:

$$\begin{aligned} f(x + t(y-x)) &\leq t(f(y) - f(x)) + f(x) \\ f(y) - f(x) &\geq \frac{f(x + t(y-x)) - f(x)}{t} \end{aligned}$$

Hence

$$f(y) - f(x) \geq \lim_{t \to 0} \frac{f(x + t(y-x)) - f(x)}{t} = \nabla f(x)^\top (y - x) \tag{6}$$

Now we need to show the FOC implies convexity. Choose first $z = tx + (1-t)y$ for $t \in (0,1)$, then

$$f(x) \geq f(z) + \nabla f(z)^\top (x - z) \tag{7}$$

$$f(y) \geq f(z) + \nabla f(z)^\top (y - z) \tag{8}$$

Multiply (7) by t and (8) by $(1-t)$ and add them up, we have:

$$\begin{aligned} tf(x) + (1-t)f(y) &\geq f(z) + t\nabla f(z)^\top (x - z) + (1-t)\nabla f(z)^\top (y - z) \\ &= f(z) + \nabla f(z)^\top (tx - tz) + \nabla f(z)^\top ((1-t)(y-z)) \\ &= f(z) + \nabla f(z)^\top (tx - tz + y - ty - z + tz) \\ &= f(z) + \nabla f(z)^\top (tx + (1-t)y - z) \\ &= f(z) + \nabla f(z)^\top (0) \\ &= tf(y) + (1-t)f(x) \end{aligned}$$

■

Lemma 2.2 (Second Order Condition) *A twice-differentiable function $f(x)$ is convex iff $dom(f)$ is a convex set and $\forall x \in dom(f)$, the Hessian is positive semi-definite.*

$$\nabla^2 f(x) \succeq 0 \tag{9}$$

Proof: By convexity we have:

$$f(y) \geq f(x) + \nabla f(x)^{\top}(y - x) \tag{10}$$

$$f(x) \geq f(y) + \nabla f(y)^{\top}(x - y) \tag{11}$$

Rearrange the equations, we have

$$\nabla f(x)^{\top}(y - x) \leq f(y) - f(x) \leq \nabla f(y)^{\top}(y - x) \tag{12}$$

Dividing both side by $(y - x)^2$

$$\frac{\nabla f(y)^{\top} - \nabla f(x)^{\top}}{y - x} \geq 0 \tag{13}$$

■

2.5 Lipschitz Continuity

A function $f: \mathbb{R}^b \to \mathbb{R}^{d^{`}}$ *is L-Lipschitz continuous* $\iff$ *for* $L > 0$ *and* $\forall x, y \in dom(f)$ we have:

$$\|f(x) - f(y)\|_2 \leq L\|x - y\|_2 \tag{14}$$

This means the function must stay outside a double cone of steepness L. Like usual definitions of continuity (pointwise or uniform), it doesn't allow jumps, but it is stricter than just continuous.

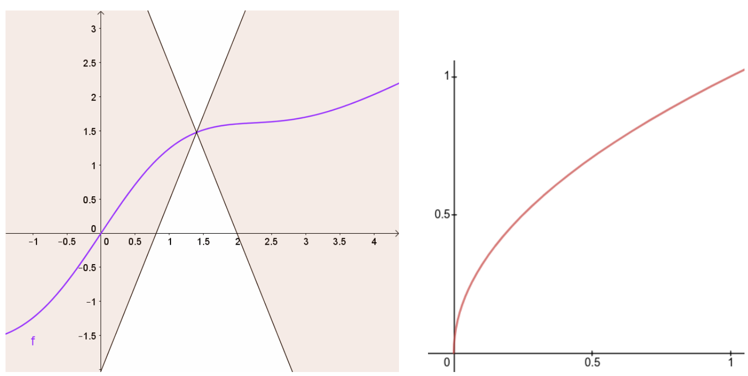

Figure 4: The function on the left is L-continuous but $\sqrt{x}$ on the right is not L-continuous, the gradient becomes infinitely steep at 0.

2.6 Smoothness

A continuously differentiable function $f(x)$ is *L-smooth* if its gradient is *L-Lipschitz*, i.e., there exists a $L > 0$ and $\forall x, y \in dom(f)$

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \tag{15}$$

The function $|x|$ is L-continuous but not L-smooth, since $\nabla f(x) - \nabla f(y) = 2$ when $x = 0^+$ and $y = 0^-$.

One important consequence of L-smoothness is this: there is a maximum bound on the difference between $f(y)$ and the predicted $f(y)$ if you drew a tangent line from x to y.

Claim 2.3 *Let f be a differentiable and L-smooth, then:*

$$f(y) - f(x) - \nabla f(x)^\top (y - x) \leq \frac{L}{2}\|y - x\|_2^2 \tag{16}$$

Proof: *For a differentiable function f(x), the difference in the f-value of x, y is simply the sum of the small differences from x to y:*

$$f(y) - f(x) = \int_x^y \nabla f(z) dz \tag{17}$$

$$set\ z = ty + (1 - t)x \tag{18}$$

$$\begin{aligned}
f(y) - f(x) - \nabla f(x)^\top (y - x) &= \int_x^y \nabla f(z) dz - \nabla f(x)^\top (y - x) && (19)\\
&= \int_0^1 \nabla f^\top (x + t(y - x))(y - x) dt - \nabla f^\top (y - x) && (20)\\
&= \int_0^1 \left[\nabla f^\top \left(x + t(y - x) - \nabla f^\top (x)\right)\right] (y - x) dt && (21)
\end{aligned}$$

For two vectors, we know that $a \cdot b \leq \|a\|\|b\|$

$$\leq \int_0^1 \left\|\nabla f^\top \left(x + t(y - x) - \nabla f^\top (x)\right)\right\| \|y - x\| dt \tag{22}$$

Now apply L-smooth definition

$$\begin{aligned}
&\leq \int_0^1 L\|x + t(y - x) - x\|\|y - x\| dt && (23)\\
&= \int_0^1 tL\|y - x\|\|y - x\| dt && (24)\\
&= \int_0^1 t dt L\|y - x\|_2^2 && (25)\\
&= \frac{L}{2}\|y - x\|_2^2 && (26)
\end{aligned}$$

■

2.7 Strongly Convex

A function $f(x)$ is μ-strongly convex if for $\alpha > 0$ and $\forall x \in dom(f)$:

$$f(x) - \frac{\mu}{2}\|x\|_2^2 \ is\ convex \tag{27}$$

If a μ-strongly convex function $g(x)$ is differentiable, then $\forall x, y \in dom(g)$, by applying the definition for convexity we have:

$$
\begin{aligned}
g(y) - g(x) &\geq \nabla g(x)(y - x) \\
f(y) - \frac{\mu}{2}\|y\|^2 &\geq f(x) - \frac{\mu}{2}\|x\|^2 + \nabla(f(x) - \frac{\mu}{2}\|x\|^2)^\top (y - x) \\
f(y) - f(x) &\geq \frac{\mu}{2}(\|y\|^2 - \|x\|^2) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}(-2x^\top y + 2\|x\|^2) \\
&= \nabla f(x)^\top (y - x) + \frac{\mu}{2}(\|y\|^2 - 2x^\top y + \|x\|^2) \\
&= \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|^2
\end{aligned}
$$

Note that this is similar to the L-smooth claim before, but with the inequality reversed (i.e. there is a *lower* bound on the difference between $f(y)$ and the predicted $f(y)$. Hence strongly-convex functions are generally $O(x^2)$.

2.8 Minimizing Convex Functions

Lemma 2.4 (Gradient Zero) *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and convex. x^* is a minimizer iff $\nabla f(x^*) = 0$. Hence all minimizers give same f-value, i.e $f(x_1^*) = f(x_2^*)$.*

If $\nabla f(x^*) = 0$, then from convexity:

$$
\begin{aligned}
f(y) &\geq f(x^*) + \nabla f(x^*)^\top (y - x^*) \\
&= f(x^*)
\end{aligned}
\tag{28}
$$

For some small $t > 0$, let $y = x^* - t\nabla f(x^*) \in f$. By Taylor expansion of $f(y)$,

$$
\begin{aligned}
f(y) &= f(x^*) + \nabla f(x^*)^\top (y - x^*) + o(\|y - x^*\|^2) \\
&= f(x^*) - t\|\nabla f(x^*)\|^2 + o(\|t\nabla f(x^*)\|^2)
\end{aligned}
\tag{29}
$$

Small t means $-t\|\nabla f(x^*)\|^2$ dominates, and if $\nabla f(x^*) \neq 0$, $f(y) < f(x^*)$ (x^* isn't a minimizer anymore). Hence $\nabla f(x^*)$ must be 0.

3 Gradient Descent Algorithm

3.1 Gradient Descent (GD)

Now that we have defined the classes of objective functions to minimize, we use *Gradient Descent*[1] to optimize the function.

Definition 3.1 (Gradient Descent) *Let $f : \mathbb{R}^d \to \mathbb{R}$ be our objection function and differentiable.*

$$x_{t+1} = x_t - \alpha \nabla f(x_t) \tag{30}$$

where α is the step-size or learning rate. Smaller α makes convergence slower, but larger α may make the algorithm oscillate.

We will show, given appropriate choices of α, that the GD estimate converges for the above classes of functions:

Class	α	Type of Convergence	Rate of Convergence
L-continuous	$\dfrac{\epsilon}{L^2}$	Average: $f\left(\dfrac{1}{T}\sum x_T\right) \to f(x^*)$	$O\left(\dfrac{L^2}{\epsilon^2}\right)$
L-smooth	$\dfrac{1}{L}$	Value: $f(x_T) \to f(x^*)$	$O\left(\dfrac{L}{\epsilon}\right)$
μ-strongly convex	$\dfrac{1}{L}$	Point: $x_T \to x^*$	$\dfrac{L}{\mu} \ln \dfrac{1}{\epsilon}$

Figure 5: Covergence for different classes of objective function

where $R = \|x_0 - x^*\|_2$ is the distance between starting point and minimizer, L and μ are as defined previously, and $\epsilon = \|f(x_T) - f(x^*)\|_2$ is the max allowed error.

3.1.1 Analysis of GD for L-continuous

Theorem 3.1 (Gradient Descent for L-continuous) *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable, convex and* L-Lipschitz*. Let $R = \|x_0 - x^*\|_2$ be the distance between the initial point x_0 and minimizer x^*.It holds for $T = \frac{R^2L^2}{\epsilon^2}$ that*

$$f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - f(x^*) \leq \epsilon \tag{31}$$

with appropriate choosing $\alpha = \frac{\epsilon}{L^2}$. This L-Lipschitz *GD gives a rate of $O\left(\frac{1}{\epsilon^2}\right)$ but we can optimize it further in the next part.*

Proof:[for Theorem 3.1] It holds that from FOC for convexity functions we have

$$f(x_t) - f(x^*) \leq \nabla f^\top(x_t)(x_t - x^*) \tag{32}$$

then, by substituting $\nabla f^\top(x_t)$ with definition of GD, we get

$$f(x_t) - f(x^*) \leq \frac{1}{\alpha}(x_t - x_{t+1})^\top (x_t - x^*) \tag{33}$$

From the law of Cosines, i.e. For a triangular with sides **a**, **b** and **c**, we have

$$c^2 = a^2 + b^2 - 2a^\top b \tag{34}$$

with $a = (x_t - x_{t+1}), b = (x_t - x^*), c = (x_{t+1} - x^*)$,then

$$f(x_t) - f(x^*) \leq \frac{1}{2\alpha}\left(\|x_t - x^*\|_2^2 + \|x_t - x_{t+1}\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) \tag{35}$$

$$= \frac{1}{2\alpha}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) + \frac{\alpha}{2}\|\nabla f(x_t)\|_2^2 \tag{36}$$

Supposing that f *is* $L - $*Lipschitz continuous* , then $\forall x \in dom(f)$ exists

$$\|\nabla f(x)\|_2 \leq L \tag{37}$$

therefore,

$$f(x_t) - f(x^*) = \frac{1}{2\alpha}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) + \frac{\alpha L^2}{2} \tag{38}$$

Summing from 1 to T:

$$\frac{1}{T}\sum_{t=1}^{T} f(x_t) - f(x^*) \leq \frac{1}{2\alpha T}\sum_{t=1}^{T}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) + \frac{\alpha L^2}{2} \tag{39}$$

Taking the telescopic sum, the terms cancel, leaving the first and last:

$$\leq \frac{1}{2\alpha T}\left(\|x_1 - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) + \frac{\alpha L^2}{2} \tag{40}$$

$$\leq \frac{R^2}{2\alpha T} + \frac{\alpha L^2}{2} = \epsilon \tag{41}$$

Finally, from Jensen's inequality it induces the[2]

$$f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - f(x^*) \leq \epsilon \tag{42}$$

■

Note that this theorem does not imply the point-wise convergence like $f(x_T) \to f(x^*)$.

3.1.2 Analysis of GD for *L-smooth*

Theorem 3.2 (Gradient Descent for L-Smooth) *Let $f : \mathbb{R}^d \to \mathbb{R}$ differentiable, convex and* L-Smooth. *Let $R = \|x_1 - x^*\|_2$ be the distance between the initial point x_1 and minimizer x^*. It holds for $T = \dfrac{LR^2}{\epsilon}$ that*

$$f(x_{t+1}) - f(x^*) \leq \epsilon \tag{43}$$

with appropriate choosing $\alpha = \frac{1}{L}$. This L-Smooth *GD gives a rate of $O\left(\dfrac{1}{\epsilon}\right)$ but we can optimize it further in the next part.*

Proof:[for Theorem 3.2] Let's first try to find an expression for on the f-value improvement with each iteration, $f(x_{t+1}) - f(x_t)$. We can sum it up later to obtain the total improvement we need. From L-smoothness (see Claim 2.3) we have:

$$\begin{aligned}
f(x_{t+1}) - f(x_t) &\leq \nabla f(x_t)^\top (x_{t+1} - x_t) + \frac{L}{2}\|x_{t+1} - x_t\|_2^2 \\
&= -\frac{1}{L}\|\nabla f(x_t)\|_2^2 + \frac{L}{2} * \frac{1}{L^2}\|\nabla f(x_t)\|_2^2 \\
&= -\frac{1}{2L}\|\nabla f(x_t)\|_2^2 \\
\|\nabla f(x_t)\|^2 &\leq 2L(f(x_t) - f(x_{t+1}))
\end{aligned}$$

For simplicity, let us denote $\delta t = x_t - x*$ and $\nabla_t = \nabla f(x_t)$. We already have some bound on $\|\nabla_t\|^2$ above, and we want to force it out. ∇_t appears in gradient descent: $x_{t+1} - x^* = x_t - x^* - \frac{1}{L}\nabla_t$. So we square both sides to get positive values and the $\|\nabla_t\|^2$:

$$\|x_{t+1} - x^*\|^2 = \|x_t - x^*\|^2 + \frac{1}{L^2}\|\nabla_t\|^2 - \frac{2}{L}(x_t - x^*)^\top \nabla_t$$

Because f is convex, $(x_t - x^*)^\top \nabla_t \geq \delta_t$

$$\begin{aligned}
\|x_{t+1} - x^*\|^2 &\leq \|x_t - x^*\|^2 + \frac{1}{L^2}\|\nabla_t\|^2 - \frac{2}{L}\delta_t \\
\|x_{t+1} - x^*\|^2 - \|x_t - x^*\|^2 &\leq \frac{1}{L^2}\|\nabla_t\|^2 - \frac{2}{L}\delta_t
\end{aligned}$$

Summing from 1 to t, we can see that the left side reduces to the first and last terms.

$$\|x_{t+1} - x^*\|^2 - \|x_1 - x^*\|^2 \leq \frac{1}{L^2}\sum \|\nabla_t\|^2 - \frac{2}{L}\sum \delta_t$$

Now $\|x_{t+1} - x^*\|^2$ must be ≥ 0 and $\|x_1 - x^*\|^2 = R^2$

$$-R^2 \leq \frac{1}{L^2}\sum \|\nabla_t\|^2 - \frac{2}{L}\sum \delta_t$$

δ_t is decreasing. Thus $\sum \delta_t \geq t\delta_t$

$$-R^2 \leq \frac{1}{L^2}\sum \|\nabla_t\|^2 - \frac{2t}{L}\delta_t$$

Now we need to find a bound for $\sum \|\nabla_t\|^2$. Fortunately we did it earlier!

$$\begin{aligned}
\sum \|\nabla t\|^2 &\leq \sum 2L(f(x_t) - f(x_{t+1}) \\
&= 2L(f(x_1) - f(x_{t+1}) \\
&\leq 2L(f(x_1) - f(x_*)) \qquad \text{since } f(x^*) \leq f(x_{t+1})
\end{aligned}$$

From L-smoothness (Claim 2.3) we know

$$
\begin{aligned}
f(x_1) - f(x^*) - \nabla f(x^*)^\top (x_1 - x^*) &\leq \frac{L}{2}\|x_1 - x^*\|^2 \\
f(x_1) - f(x^*) - 0 \cdot (x_1 - x^*) &\leq \frac{L}{2}R^2
\end{aligned}
$$

Hence,

$$
\sum \|\nabla t\|^2 \leq 2L\left(\frac{LR^2}{2}\right) = L^2R^2
$$

Now we sub this back into where we left off:

$$
\begin{aligned}
-R^2 &\leq \frac{1}{L^2}\sum \|\nabla_t\|^2 - \frac{2t}{L}\delta_t \\
&\leq \frac{1}{L^2}(L^2R^2) - \frac{2t}{L}\delta_t \\
-2R^2 &\leq -\frac{2t}{L}\delta_t \\
\delta_t &\leq \frac{LR^2}{t}
\end{aligned}
$$

Thus the minimum number of steps T to reach $\delta_t \leq \epsilon$ is LR^2/ϵ. ■

Theorem 3.3 (Gradient Descent for L-smooth and μ-convex) *Let $f : \mathbb{R}^d \to \mathbb{R}$ differentiable, μ-*convex *and L-Smooth. Let $R = \|x_0 - x^*\|_2$ be the distance between the initial point x_0 and minimizer x^*.It holds for $T = \frac{2L}{\mu}\ln(\frac{R}{\epsilon})$ that*

$$
\|x_T - x^*\|_2^2 \leq \epsilon \tag{44}
$$

*with appropriate choosing $\alpha = \frac{1}{L}$. This μ-*convex *and L-smooth function GD gives a rate of $O\left(\ln\frac{1}{\epsilon}\right)$*

Proof:[for Theorem 3.3] Consider the left side of the inequality, we have

$$
\begin{aligned}
\|x_T - x^*\|_2^2 &= \|x_{T-1} - \frac{1}{L}\nabla f(x_{T-1}) - x^*\|_2^2 &(45)\\
&= \|x_{T-1} - x^*\|_2^2 + \frac{1}{L^2}\|\nabla f(x_{T-1})\|_2^2 - \frac{2}{L}\nabla f(x_{T-1})^\top(x_{T-1} - x^*)) &(46)
\end{aligned}
$$

since f is *μ-strong convex* and L-smooth, from 2.7 and 3.1.2

$$
\begin{aligned}
\frac{2}{L}\nabla f(x_{T-1})^\top(x^* - x_{T-1}) &\leq \frac{2}{L}(f(x^*) - f(x_{T-1})) - \frac{\mu}{L}\|x^* - x_{T-1}\|_2^2 &(47)\\
&\leq -\frac{1}{L^2}\|\nabla f(x_{T-1})\|_2^2 - \frac{\mu}{L}\|x^* - x_{T-1}\|_2^2 &(48)\\
\textit{then } \|x_T - x^*\|_2^2 &\leq (1 - \frac{\mu}{L})\|x_{T-1} - x^*\|_2^2 &(49)\\
&\leq (1 - \frac{\mu}{L})^\top R^2 \leq e^{-\frac{\mu T}{L}}R^2 &(50)
\end{aligned}
$$

substituting $T = \frac{2L}{\mu}\ln\frac{R}{\epsilon}$ we get

$$
\|x_T - x^*\|_2^2 \leq \epsilon \tag{51}
$$

Note that in this theorem, it will convergence at last iteration. ■

3.2 Projected Gradient Descent

In previous settings, we focus on how to find solutions of the *unconstrained optimization problem.* However, in general machine learning problems we are likely to encounter some constrained problems. In this subsection, we discuss how to solve constrained optimization problem:

$$min_{x \in \mathcal{X}} f(x)$$

where f is a convex function and $\mathcal{X}$ is a convex set. Consider that when we use gradient descent to update the x_t by step-size α, or $x_{t+1} = x_t - \alpha \nabla f(x)$, it is possible that x_{t+1} may not belong to the constraint, i.e. convex set $\mathcal{X}$. In this part, we introduce *Projected Gradient Descent* to deal with the issue.

Definition 3.2 (Projected Gradient Descent) *The projection of a point y, onto a set $\mathcal{X}$ is defined as the nearest point in the set to y.*

$$\Pi_{\mathcal{X}}(y) = argmin_{x \in \mathcal{X}} \frac{1}{2} \|x - y\|_2^2$$

Let f: $\mathbb{R}^d$ be differentiable function in some convex *set $\mathcal{X}$. The algorithm below is called* Projected *Gradient Descent:*

$$x_{t+1} = \Pi_{\mathcal{X}}(x_k - \alpha \nabla f(x_k)) \tag{52}$$

where the projection GD may not be that efficient and the minimizer x^* does not necessarily satisfy $\nabla f(x^*) = 0$.

In this part, we mainly analysis the *Projected Gradient Descent* for *L-lipschitz*

Theorem 3.4 (Projected Gradient Descent) *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable, convex and* L-Lipschitz. *Let $R = \|x_0 - x^*\|_2$ be the distance between the initial point x_0 and minimizer x^*.It holds for $T = \frac{R^2 L^2}{\epsilon^2}$ that*

$$f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - f(x^*) \leq \epsilon \tag{53}$$

with appropriate choosing $\alpha = \frac{\epsilon}{L^2}$. Same guarantees as in the unconstrained case.

Proof:[for Theorem 3.4] Set $y := x_t - \alpha \nabla f(x_t)$. It holds that from FOC for convexity functions we have

$$f(x_t) - f(x^*) \leq \nabla f(x_t)^\top (x_t - x^*) \tag{54}$$

then, by substituting $\nabla f(x_t)^\top$ with definition of GD, we get

$$f(x_t) - f(x^*) \leq \frac{1}{\alpha}(x_t - y_t)^\top (x_t - x^*) \tag{55}$$

From the law of Cosines, i.e. For a triangular with sides **a**, **b** and **c**, we have

$$c^2 = a^2 + b^2 - 2a^\top b \tag{56}$$

with $a = (x_t - x_{t+1}), b = (x_t - x^*), c = (x_{t+1} - x^*)$,then

$$f(x_t) - f(x^*) \leq \frac{1}{2\alpha}\left(\|x_t - x^*\|_2^2 + \|x_t - y_t\|_2^2 - \|y_t - x^*\|_2^2\right) \tag{57}$$

$$= \frac{1}{2\alpha}\left(\|x_t - x^*\|_2^2 - \|y_t - x^*\|_2^2\right) + \frac{\alpha}{2}\|\nabla f(x_t)\|_2^2 \tag{58}$$

Recall that f *is* $L - Lipschitz$ *continuous* , then $\forall x \in dom(f)$ exists

$$\|\nabla f(x)\|_2 \leq L$$

Therefore,

$$f(x_t) - f(x^*) \leq \frac{1}{2\alpha}\left(\|x_t - x^*\|_2^2 - \|y_t - x^*\|_2^2\right) + \frac{\alpha L^2}{2} \tag{59}$$

Stop here and introduce the Claim.

Claim 3.5 *For projection of a point* y, *it holds that:*

$$(\Pi_{\mathcal{X}}(y) - x)^\top(\Pi_{\mathcal{X}}(y) - y) \leq 0$$

Proof:*[for Claim 3] From the following figure*

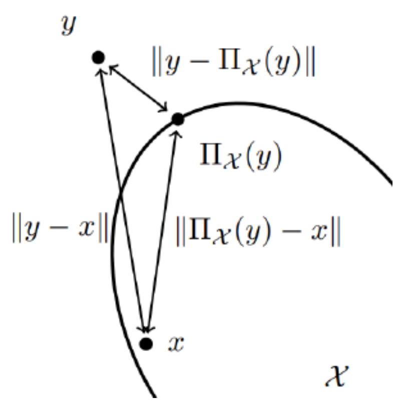

Figure 6: Projection of a point y on Convex Set

Since the projection property on the convex set, it is true that $\|y - x\|$ *must be the longest side of the triangular. Hence, from* law of Cosines *we have*

$$\|y - x\|_2^2 \geq \|\Pi_{\mathcal{X}}(y) - y\|_2^2 + \|\Pi_{\mathcal{X}}(y) - x\|_2^2$$

Therefore $\cos\langle \Pi_{\mathcal{X}}(y) - y, \Pi_{\mathcal{X}}(y) - x\rangle < 0$. *It is proved.* ■

Then, continue to prove **Theorem 3.4**. From Claim 3 we have:

$$\|y_t - x^*\|_2^2 \geq \|x_{t+1} - y_t\|_2^2 + \|x_{t+1} - x^*\|_2^2 \tag{60}$$

$$\geq \|x_{t+1} - x^*\|_2^2 \tag{61}$$

Note that x_{t+1} is in the *Convex Set* $\mathcal{X}$, since

$$f(x_t) - f(x^*) \leq \frac{1}{2\alpha}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) + \frac{\alpha L^2}{2} \tag{62}$$

Taking the telescopic sum we have

$$\frac{1}{T}\sum_{t=1}^{T} f(x_t) - f(x^*) \leq \frac{1}{2\alpha T}\left(\|x_1 - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) + \frac{\alpha L^2}{2} \tag{63}$$

$$\leq \frac{R^2}{2\alpha T} + \frac{\alpha L^2}{2} = \epsilon \tag{64}$$

Finally, from Jensen's inequality it induces the

$$f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - f(x^*) \leq \epsilon \tag{65}$$

It is the same as classic Gradient Descent. ■

References

[1] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[2] Daniel Carlton Smith. BFunctional Gradient Descent. `http://www.cs.cmu.edu/~16831-f12/notes/F12/16831_lecture21_danielsm.pdf`, 2013.