IFT 6085 - Lecture 2
Basics of convex analysis and gradient descent

This version of the notes has not yet been thoroughly checked. Please report any bugs to the scribes or instructor.

Scribes: Assya Trofimov, Mohammad Pezeshki, Reyhane Askari **Instructor:** Ioannis Mitliagkas

1 Summary

In this first lecture we cover some optimization basics with the following themes:

- Lipschitz continuity
- Some notions and definitions for convexity
- Smoothness and Strong Convexity
- Gradient Descent

Note: Most of this lecture has been adapted from [1].

2 Introduction

In this section we introduce the basic concepts of optimization.
The gradient descent algorithm is the workhorse of machine learning. It generally has two equivalent interpretations:

- downhill
- local minimization of a function

Definition 1 (Lipschitz continuity)**.** *A function $f(x)$ is L-Lipschitz if*

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

Intuitively, this is a measurement of how steep the function can get (Figure 1).

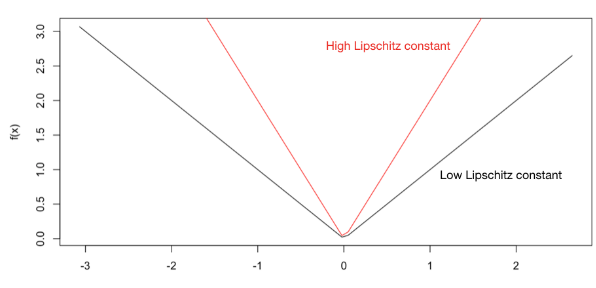

Figure 1: Lipschitz constant

This also implies that the derivative of the function cannot exceed L.

$$f'(x) = \lim_{\delta \to 0} \frac{f(x+\delta) - f(x)}{\delta}$$

and

$$f'(x) = \lim_{y \to x} \frac{f(x) - f(y)}{x - y} = \lim_{y \to x} \frac{|f(x) - f(y)|}{|x - y|} \leq L$$

As a consequence, L-Lipschitz implies that $f'(x)$ is bounded by L

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

Lipschitz continuity can be seen as a refinement of continuity. Example:

$$f(x) = \begin{cases} exp(-\lambda x), & \text{if } x > 0 \\ 1, & \text{otherwise} \end{cases}$$

Note that f(x) is L-lipschitz. As the λ value increases, the closer the function gets to discontinuity (Figure 2).

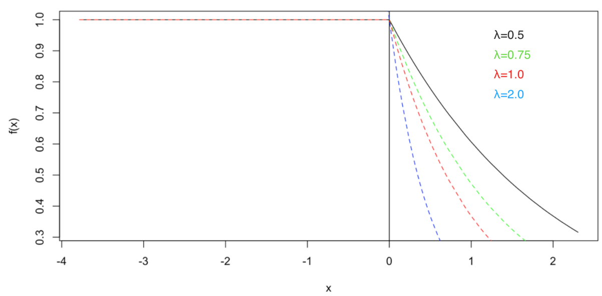

Figure 2: As λ value increases, the function is closer to being discontinuous

3 Convexity

Let us first look at the definition of convexity for a set.

Definition 2. *For a convex set* ***X****, for any two points x and y such that $x, y \in X$, the line between them lies within the set (Figure 3 A). That is:*

$$\forall \theta \in [0,1] \quad and \quad z = \theta x + (1-\theta)y, \quad then \quad z \in X$$

When parameter θ is equal to 1, we get x and when θ is 0, we get y. In contrast, a non-convex set is a set where z may lie outside of the set (Figure 3 B).

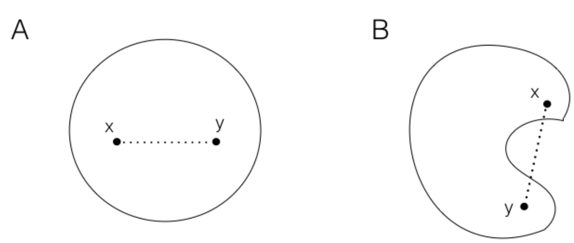

Figure 3: A) Convex set and B) Non-convex set

Definition 3 (Convex Linear Combination)**.** *The sum $\theta x + (1-\theta)y$ is termed as convex linear combination.*

We can apply the convex definition to functions.

Definition 4 (Convex function)**.** *A function $f(x)$ is convex if the following holds:*

- *The Domain(f) is convex.*
- *For any two members of the domain, the function value on a convex combination does not exceed the convex combination of values.*

$$\forall x, y \in Domain(f), \theta \in [0, 1]$$

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)$$

Another way to express this would be to check the line segment connecting x and y (the chord). If the chord lies above the function itself (Figure 4) the function is convex.

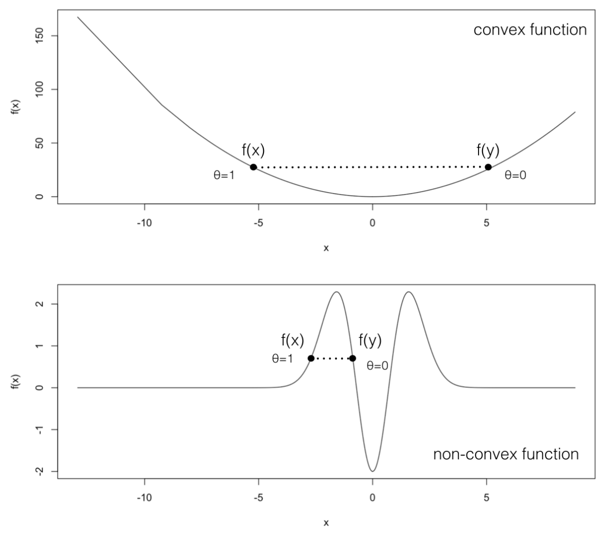

Figure 4: Example of convex and non-convex functions

Moreover, for differentiable or twice differentiable functions, it is possible to define convexity with the following first and second order conditions for convexity.

Definition 5 (First order condition for convexity)**.** *f(x) is convex if and only if domain(f) is convex and the following holds for $\forall x, y \in domain(f)$*

$$f(y) \geq f(x) + \nabla f(x)^T(y-x)$$

In other words, the function should be lower bounded by all its tangents.
In Figure 5, part of the non-convex function is below the tangent at point x. This is not the case for the convex function. The convex function should therefore be *lower-bounded* by all the tangents at any point.
As a reminder, the Hessian is a measure of curvature. It is the multivariate generalization for second derivative. Indeed, for function $f(x) = \frac{h}{2}x^2$, the second derivative $f''(x) = h$, which corresponds to a measure of how quickly the shape changes in the function. A multivariate quadratic can be written as $f(x) = \frac{1}{2}x^T Hx$, where H is the Hessian.
Curvature along the eigenvectors of the Hessian is given by the corresponding eigenvalues.

$$H = Q\Lambda Q^T$$

$$\Lambda = \begin{bmatrix} h_1 & & & \\ & h_2 & & \\ & & ... & \\ & & & h_d \end{bmatrix}$$

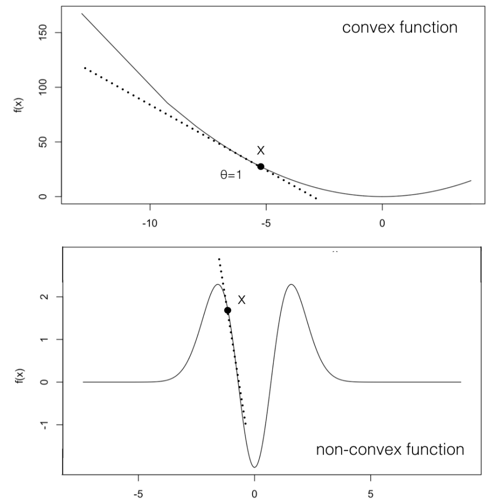

Figure 5: Example of convex and non-convex function relative to the tangent at point x

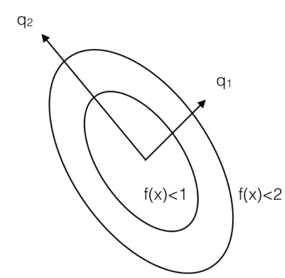

Figure 6: Looking along the principle directions of the quadratic, it appears that along q_1 we reach higher values more quickly. This means the curvature is higher along q_1.

Changing the basis with Q, we decompose the matrix and focus on the direction described by $Q = [q_1, q_2, ..., q_d]$. Along the direction of q_i, we see the curvature for h_i (Figure 6). Note that $[h_1, h_2, ..., h_d]$ are sorted in order of magnitude.
If the function is twice differentiable, another convexity definition applies.

Definition 6 (Second order condition for convexity)**.** *A twice differentiable function f is convex if and only if:*

$$\nabla^2 f(x) \succeq 0 \quad where \quad x \in domain(f(x))$$

This also implies that the Hessian needs to be positive semi-definite, in other words, eigenvalues need to be non-negative.

Note: All the definitions of convexity are equivalent when the right level of differentiability holds.

4 Smoothness and Strong Convexity

Definition 7 (Smoothness)**.** *A function $f(x)$ is β-smooth if the following holds:*

$$||\nabla f(x) - \nabla f(y)|| \leq \beta||x - y|| \quad where \quad x, y \in domain(f(x)). \tag{1}$$

It is noted that β-smoothness of $f(x)$ is equivalent to β-Lipschitz of $\nabla f(x)$. Smoothness constraint requires the gradient of $f(x)$ to not change rapidly.

Definition 8 (Strong Convexity). *A function $f(x)$ is α-strongly convex if $f(x) - \frac{\alpha}{2}||x||^2$ is convex.*

If $f(x)$ is α-*strongly convex* then the following hold:

$$\nabla^2 f(x) \succeq \alpha I \Leftrightarrow \nabla^2 f(x) - \alpha I \succeq 0. \tag{2}$$

It informally means that the curvature of f(x) is not very close to zero. For instance, in 1-D case, $f(x) = \frac{h}{2}x^2$ is h-strongly convex but not $(h + \epsilon)$-strongly convex. Figure 7 illustrates examples of two convex functions of which only one is strongly convex.

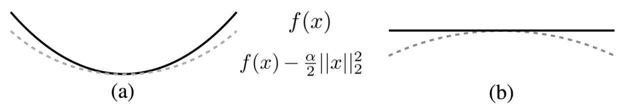

Figure 7: (a) A convex function which is also strongly convex. (b) A convex function which is not strongly convex.

5 Gradient Descent

Gradient descent is an optimization algorithm based on the fact that a function $f(x)$ decreases fastest in the direction of the negative gradient of $f(x)$ at a current point. Consequently, starting from a guess x_0 for a local minimum of $f(x)$ the sequence $x_0, x_1, ..., x_t \in \mathbb{R}^d$ is generated using the following rule:

$$x_{k+1} = x_k - \gamma \nabla f(x_k), \tag{3}$$

in which γ is called the *step size* or the *learning rate*. If $f(x)$ is convex and γ decays at the right rate, it is guaranteed that as $t \to \infty$, $x_k \to x^*$. The following holds for the optimal value $x*$:

$$x^* = \underset{x \in \text{Domain}(f(x))}{\text{argmin}} f(x). \tag{4}$$

Lemma 1. *From L-Lipschitz constraint the following holds:*

$$||\nabla f(x_k)||_2^2 \leq L^2. \tag{5}$$

This lemma is used in the proof on the following theorem.

Theorem 1 (Gradient Descent Theory). *Let $f(x)$ be convex and L-lipschitz, if T is the total number of steps taken and the learning rate is chosen as:*

$$\gamma = \frac{||x_1 - x^*||_2}{L\sqrt{T}} \tag{6}$$

Then the following holds:

$$f\Big(\frac{1}{T}\sum_{k=1}^{T} X_k\Big) - f(x^*) \leq \frac{||x_1 - x^*||L}{\sqrt{T}}, \tag{7}$$

Proof. By applying the Taylor expansion on $f(x)$ at the point x_k, we have,

$$\begin{aligned} f(x_k) - f(x^*) &\leq \Big\langle \nabla f(x_k), x_k - x^* \Big\rangle & (8) \\ &= \Big\langle \frac{1}{\gamma}(x_k - x_{k+1}), x_k - x^* \Big\rangle & (9) \\ &= \frac{1}{2\gamma}\Big(||x_k - x^*||_2^2 - ||x_{k+1} - x^*||_2^2\Big) + \gamma^2||\nabla f(x_k)||_2^2 & (10) \end{aligned}$$

From Equation (10) and Lemma 1, the following holds:

$$f(x_k) - f(x^*) \leq \frac{1}{2\gamma}\Big(||x_k - x*||_2^2 - ||x_{k+1} - x^*||_2^2\Big) + \frac{\gamma}{2}L^2 \tag{11}$$

By change of the variable $||x_k - x^*||_2^2$ to D_k:

$$\begin{aligned} &f(x_1) - f(x^*) \leq \frac{1}{2\gamma}[D_1^2 - D_2^2] + \frac{\gamma}{2}L^2 \\ &f(x_2) - f(x^*) \leq \frac{1}{2\gamma}[D_2^2 - D_3^2] + \frac{\gamma}{2}L^2 \\ &... \\ &f(x_T) - f(x^*) \leq \frac{1}{2\gamma}[D_T^2 - D_{T+1}^2] + \frac{\gamma}{2}L^2 \\ &\qquad\qquad\qquad \leq \frac{1}{2\gamma}[D_T^2] + \frac{\gamma}{2}L^2. \end{aligned} \tag{12}$$

Summing all the equations, most terms cancel. This is known as the telescoping sum. We get:

$$\sum_{k=1}^{T}(f(x_k) - f(x^*) \leq \frac{1}{2\gamma}D_1^2 + \frac{T\gamma L^2}{2} \tag{13}$$

$$\Rightarrow \frac{1}{T}\sum_{k=1}^{T} f(x_k) - f(x^*) \leq \frac{1}{2\gamma T}D_1^2 + \frac{\gamma L^2}{2} \tag{14}$$

From convexity of $f(x)$ we know:

$$f\big(\theta x + (1-\theta)y\big) \leq \theta f(x) + (1 - \theta f(y) \tag{15}$$

So from Equation 14 and 15 the following holds:

$$f\Big(\frac{1}{T}\sum_{k=1}^{T} x_k\Big) - f(x^*) \leq \frac{1}{2\gamma T}D_1^2 + \frac{\gamma L^2}{2} \tag{16}$$

Thus, if we set $\gamma = \frac{||x_1 - x^*||}{L\sqrt{T}}$, we get:

$$f\Big(\frac{1}{T}\sum_{k=1}^{T} X_k\Big) - f(x^*) \leq \frac{||x_1 - x^*||L}{\sqrt{T}}. \tag{17}$$

□

References

[1] Bubeck, Sbastien. "Convex optimization: Algorithms and complexity." Foundations and Trends in Machine Learning 8.3-4 (2015): 231-357.