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

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).

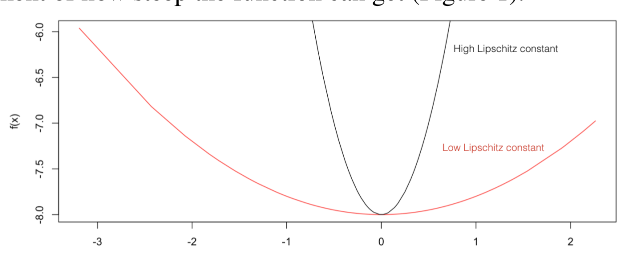

Figure 1: Lipschitz constant

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

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

and

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

By consequences, L-Lipschitz implies that $f'(x)$ is bounded by L

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

Example:

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

As the λ value increases, the closer the function gets to discontinuity (Figure 2).

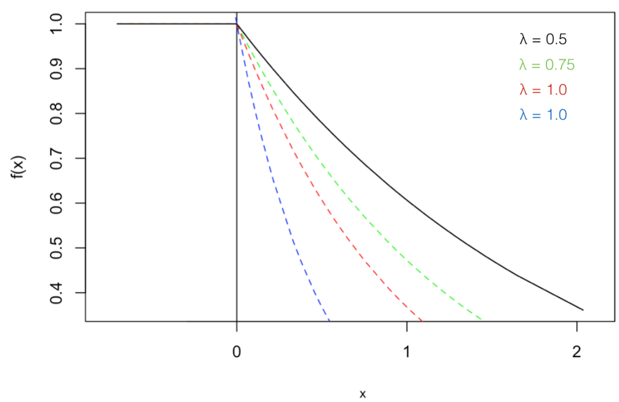

Figure 2: As λ value increases, the function is closer to being discontinuous

3 Convexity

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

Definition 2. *For a convex set, for any two points x and y picked, the line between them lies within the set (Figure 3 A).*

$$z = \theta x + (1 - \theta)y$$

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

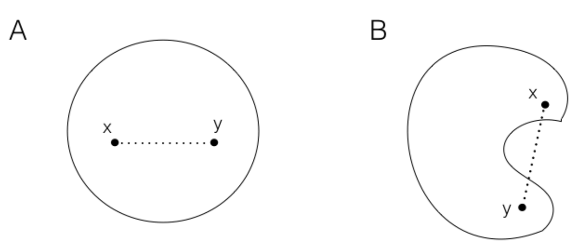

Figure 3: A) Convex set and B) Non-convex set

We can apply this definition to functions.

Definition 3 (Convex function)**.** *A function $f(x)$ is convex if the following holds:*

- *The domain of f (**dom f**) is convex*
- *For any two members of the domain, the objective value does not exceed individual combination.*

$$\forall x, y \in \textbf{\textit{dom f}}, \theta \in [0, 1]$$

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

Note: *The sum $\theta x + (1-\theta)y$ is termed as convex linear combination.*

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

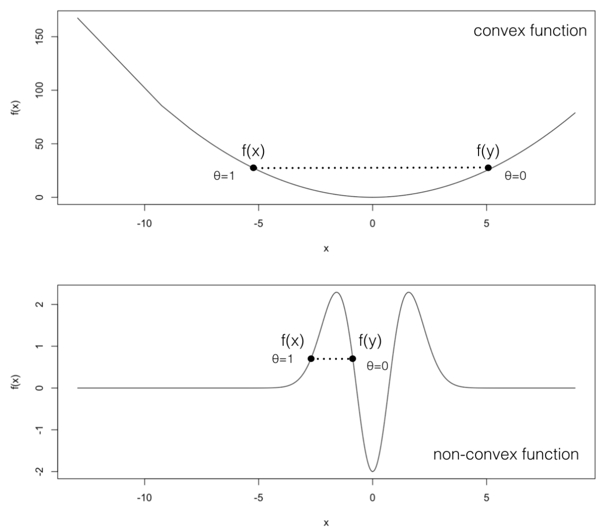

Figure 4: Example of convex and non-convex functions

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

Definition 4 (First order condition for convexity)**.** *f(x) is convex if and only if domain(f) is convex and the following holds for $\forall x, y \in domain(f)$*

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

In other words, the function should be lower bounded by all its tangents. Indeed, if we select x as a point for a Taylor approximation, we find that $f(y) \sim f(x) - f(x)(y-x)$.
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 curvature changes in the function. For a matrix, the Hessian is calculated with $f(x) = \frac{1}{2}x^T Hx$, where H is the Hessian.
Moreover, curvature may be described by the eigen values, with Λ is a diagonal matrix of eigen values.

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

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

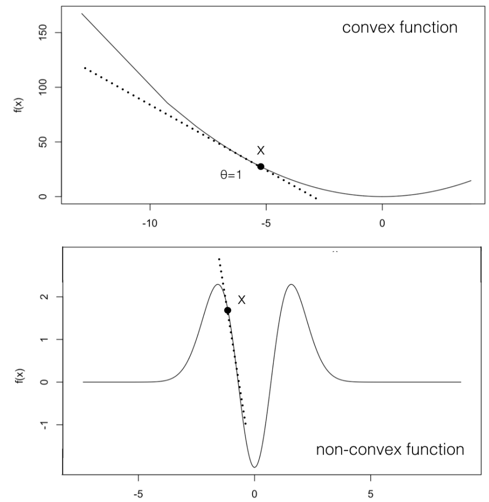

Figure 5: Example of convex and non-convex function relative to the tangent at point x

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.

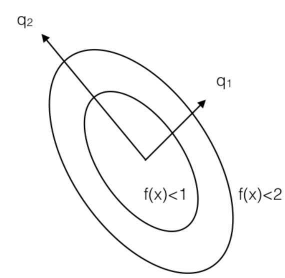

Figure 6: Looking along q_1 and q_2, it appears that q_1 has the higher slope, therefore the higher curvature

If the function is twice differentiable, another convexity definition applies.

Definition 5 (Second order condition for convexity)**.** *A function f is convex if:*

$$\nabla^2 f(x) \geq 0$$

In this definition it is assumed that f is twice differentiable. Also, the Hessian needs to be positive and semidefinite, in other words, eigenvalues need to be non-negative.

Note: All the definitions of convexity are equal.

4 Smoothness and Strong Convexity

Definition 6 (Smoothness)**.** *A function $f(x)$ is β-smooth if the following holds:*

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

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 7 (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:

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

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 which only one of them is strongly convex.

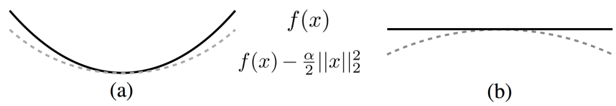

Figure 7: (a) A convex function which is also strongly convex. (b) A convex function which is not strongly convex.

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 is sufficiently small, it is guaranteed that as $t \to \infty$, $x_k \to x^*$. The following holds for the optimal value $x*$:

$$x^* = \underset{x \in \mathrm{Dom}(f(x))}{\operatorname{argmin}} f(x). \tag{4}$$

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

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

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

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

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

The the following holds:

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

where we can consider $\frac{||x_1 - x^*||L}{\sqrt{T}}$ as an ϵ.

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

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

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

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

By change of the variable $||x_k - x^*||_2^2$ to D_k, and applying telescoping sum, we have,

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

By summing all the equations, we have,

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

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

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

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

So from Equation 14 and 15 the following holds:

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

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

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

$\square$