COS 324, Precept #5: Convexity, Smoothness, and Properties

Sobhan Miryoosefi

March 6, 2020

1 Introduction

In this precept we cover different types of convexity (and how they are related), smoothness, and basic operations which preserve convexity and smoothness. Topics that are covered in this mini-course will be useful throughout the course, for example, it justifies why the objective that we are minimizing in the class (Training loss with respect to both squared loss and logistic loss) is convex. For the sake of conciseness, proofs are not included however you can find most of the proofs in "Understanding Machine Learning from Theory to Algorithms" Chapters 12-14. In this precept all the vectors are column vector and all the norms are Euclidean norm.

2 Convexity and Smoothness

Definition 1 (Convexity)**.** Consider function $f : \mathbb{R}^n \to \mathbb{R}$. These are equivalent definitions for f being **convex**

0. $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$ for all $x, y \in \mathbb{R}^n$ and $\lambda \in [0, 1]$
1. $f(y) \geq f(x) + \nabla f(x)^T (y - x)$ for all $x, y \in \mathbb{R}^n$
2. (case of $n = 1$) $f''(x) \geq 0$ for all $x \in \mathbb{R}^n$

First statement is saying that we want the function to be below the line-segment connecting any two points on the graph. Second statement is saying that we want the function to be above its first order Taylor approximation. The last statement is saying that we want the second derivative of the function to be non-negative meaning that its first derivative should be always non-decreasing.
Here we give two more definition of strict and strong convexity which are strictly stronger than being convex, meaning that every strictly or strongly convex function is convex but the inverse is not necessarily true.

Definition 2 (Strict Convexity)**.** Consider function $f : \mathbb{R}^n \to \mathbb{R}$. The equivalent definitions for f being **strictly convex**

0. $f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y)$ for all $x \neq y \in \mathbb{R}^n$ and $\lambda \in (0,1)$
1. $f(y) > f(x) + \nabla f(x)^T (y - x)$ for all $x \neq y \in \mathbb{R}^n$

and a sufficient condition for being strictly convex

2. (case of $n = 1$) $f''(x) > 0$ for all $x \in \mathbb{R}^n$ (it's only sufficient condition for being strictly convex)

First statement is saying that we want the function to be strictly below the line-segment connecting any two points on the graph. Second statement is saying that we want the function to be strictly above its first order Taylor approximation. The last statement (only sufficient condition not necessary) is saying that we want the second derivative of the function to be positive meaning that its first derivative should be always increasing.

Definition 3 (Strong Convexity)**.** Consider function $f : \mathbb{R}^n \to \mathbb{R}$. These are equivalent definitions for f being α**-strongly convex** where $\alpha > 0$

0. $f(x) - \frac{\alpha}{2} \|x\|^2$ is convex
1. $f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\alpha}{2} \|y - x\|^2$ for all $x, y \in \mathbb{R}^n$
2. (case of $n = 1$) $f''(x) \geq \alpha$ for all $x \in \mathbb{R}^n$

First statement is saying that we want the function to be convex even after subtracting $\frac{\alpha}{2} \|x\|^2$. Second statement is saying that we want the function to be at least $\frac{\alpha}{2} \|y - x\|^2$ above its first order Taylor approximation. The last statement is saying that we want the second derivative of the function to be at least $\alpha > 0$, meaning that not only its first derivative is increasing but also the rate of increase to be at least $\alpha > 0$. Now we state the claim we said previously in the following theorem:

Theorem 1 (Relationship Between Different Types of Convexity)**.**

$$f \text{ is } \alpha\text{-strongly convex} \implies f \text{ is strictly convex} \implies f \text{ is convex}$$

Convexity gives us a lower bound on the value of the function; Smoothness gives us a similar upper bound.

Definition 4 (Smoothness)**.** Consider function $f : \mathbb{R}^n \to \mathbb{R}$. These are equivalent definitions for f being β**-smooth** where $\beta > 0$

1. $f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{\beta}{2} \|y - x\|^2$ for all $x, y \in \mathbb{R}^n$
2. (case of $n = 1$) $f''(x) \leq \beta$ for all $x \in \mathbb{R}^n$

First statement is saying that we want the function to be at most $\frac{\beta}{2}\|y-x\|^2$ above its first order Taylor approximation. The last statement is saying that we want the second derivative of the function to be at most $\beta > 0$, meaning that the rate of increase of its first derivative is at most $\beta > 0$.

Example 1. Consider function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^2$. Its first and second derivatives are

$$f'(x) = 2x \quad f''(x) = 2$$

Therefore using (2.) in Definitions 3 and 4 we have that f is 2-strongly convex (consequently convex and strictly convex) and 2-smooth.

Remark 1. It's also true for the higher dimensions. $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|^2$ is 2-strongly convex (consequently convex and strictly convex) and 2-smooth.

Example 2. Consider function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \log(1+e^x)$. Its first and second derivatives are

$$f'(x) = \frac{e^x}{1+e^x} \quad f''(x) = \frac{e^x}{(1+e^x)^2}$$

its second derivatives is always positive therefore it's strictly convex (and consequently convex). However it's not α-strongly convex for any $\alpha > 0$ since for every $\alpha > 0$ there exists some $x \in \mathbb{R}$ such that $f''(x) < \alpha$ because it's easy to check that $\lim_{x\to\infty} f''(x) = 0$. In order to check smoothness we can right $f''(x) = \frac{1}{(1+e^{-x})(1+e^x)} = \frac{1}{(1+a)(1+\frac{1}{a})} \leq \frac{1}{4}$ where $a = e^x$ since $(1+a)(1+\frac{1}{a}) \geq 4$ for $a > 0$. Therefore f is $\frac{1}{4}$-smooth.

3 Convexity/Smoothness Preserving Operations

Convexity is preserved under composition with affine mapping, which is very useful since most of the functions we are dealing with in machine learning and this course are in this form.

Theorem 2 (Composition with Affine Mapping)**.** *Let $g : \mathbb{R}^n \to \mathbb{R}$ be a convex function and let $A \in \mathbb{R}^{n\times m}$ and $b \in \mathbb{R}^n$. Then function $f : \mathbb{R}^m \to \mathbb{R}$ defined as*

$$f(x) = g(Ax + b)$$

is also convex.

Sum of convex functions is also convex, more generally non-negative weighted sum of convex functions is convex; It also preserves strong convexity and smoothness. Let's formalize it as the following theorem

Theorem 3 (Non-negative Weighted Sum)**.** *Let $g_1, \ldots, g_r : \mathbb{R}^n \to \mathbb{R}$ and let $c_1, \ldots, c_r \geq 0$ be non-negative scalars. Define $f : \mathbb{R}^n \to \mathbb{R}$ as*

$$f(x) = \sum_{i=1}^{r} c_i g_i(x)$$

Then we have

1. *If $g_1, \ldots, g_r$ are all convex then f is also convex.*

2. *If each g_i is α_i-strongly convex then f is $(\sum_{i=1}^r c_i\alpha_i)$-strongly convex*

3. *If each g_i is β_i-smooth then f is $(\sum_{i=1}^r c_i\beta_i)$-smooth*

Pointwise maximum also preserves convexity. For example function $|x| = \max\{x, -x\}$ is convex since both x and $-x$ are convex.

Theorem 4 (Pointwise Maximum)**.** *Let $g_1, \ldots, g_r : \mathbb{R}^n \to \mathbb{R}$. Define $f : \mathbb{R}^n \to \mathbb{R}$ as*

$$f(x) = \max_{i=1,\ldots,r} g_i(x)$$

then f is convex.

The last theorem of this section shows that smoothness is preserved under affine mapping with bounded norm. Let's formalize it as

Theorem 5 (Composition with Affine Mapping)**.** *Let $g : \mathbb{R} \to \mathbb{R}$ be a β-smooth function and let $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$ for which we have $\|a\|^2 \leq R^2$. Then function $f : \mathbb{R}^n \to \mathbb{R}$ defined as*

$$f(x) = g(a^T x + b)$$

is $(R^2\beta)$-smooth.

4 Example: Training Loss

Let's try to check whether the Training Loss for squared loss and logistic loss that you saw in the class are convex or smooth. Assume you have m examples (x_i, y_i) where each $x_i \in \mathbb{R}^d$ and each $y_i \in \{-1, +1\}$. In addition to that assume that $\|x_i\| \leq 1$. The training error is

$$L(w) = \frac{1}{m}\sum_{i=1}^m \ell_i(w)$$

where we have for

- **Squared Loss**: $\ell_i(w) = \ell_i^{\text{square}}(w) = (w^T x_i - y_i)^2$
- **Logistic Loss**: $\ell_i(w) = \ell_i^{\text{logistic}}(w) = \log(1 + e^{-y_i w^T x_i})$

Theorem 6. *$L^{\text{square}}(w) = \frac{1}{m}\sum_{i=1}^m \ell_i^{\text{square}}(w)$ is convex and 2-smooth*

Proof. Let $g(a) = a^2$ and note that $\ell_i^{\text{square}} = g(w^T x_i - y_i)$. g is convex and 2-smooth and $w \mapsto w^T x_i - y_i$ is an affine mapping (also $\|x_i\|^2 \leq 1$); therefore using Theorem 2 and Theorem 5 we conclude that ℓ_i^{square} is convex and 2-smooth. $L^{\text{square}}(w)$ is non-negative weighted sum of ℓ_i^{square}, therefore it's convex and smooth with $\beta = \sum_{i=1}^n (\frac{1}{n})(2) = 2$ using Theorem 3. □

Theorem 7. *$L^{\text{logistic}}(w) = \frac{1}{m}\sum_{i=1}^{m} \ell_i^{\text{logistic}}(w)$ is convex and $\frac{1}{4}$-smooth*

Proof. Let $g(a) = log(1 + e^a)$ and note that $\ell_i^{\text{logistic}} = g(-y_i w^T x_i)$. g is convex and $\frac{1}{4}$-smooth and $w \mapsto -y_i w^T x_i$ is an affine mapping (also $\|x_i\|^2 \leq 1$); therefore using Theorem 2 and Theorem 5 we conclude that ℓ_i^{square} is convex and $\frac{1}{4}$-smooth. $L^{\text{logistic}}(w)$ is non-negative weighted sum of ℓ_i^{logistic}, therefore it's convex and smooth with $\beta = \sum_{i=1}^{n}(\frac{1}{n})(1/4) = \frac{1}{4}$ using Theorem 3. □