Columbia University in the City of New York
Optimization Methods for Machine Learning
Instructors: Satyen Kale
Authors: Gurpreet Singh
Email: gurpreet.s@columbia.edu

SCRIBE

4

PROPERTIES OF CONVEX FUNCTIONS

1. Recap

This section covers a short recap of the previous lecture where we defined convexity of sets and functions.

Definition 4.1 (Convex Set) A set $\mathcal{S} \subseteq \mathbb{R}^d$ is *convex* if $\forall x, y \in \mathcal{S}$, $\forall \lambda \in [0, 1]$,

$$\lambda x + (1 - \lambda)y \in \mathcal{S}$$

Definition 4.2 (Convex Function) A function $f : \mathbb{R}^d \to \mathbb{R}$ $\mathcal{S} \subseteq \mathbb{R}^d$ is convex if and only if the epigraph of the function f defined as the set $\{(x, y) \in \mathbb{R}^{d+1} \mid y \geq f(x)\}$ is convex.

An equivalent definition is the following:

Definition 4.3 (Convex Function) A function $f : \mathbb{R}^d \to \mathbb{R}$ $\mathcal{S} \subseteq \mathbb{R}^d$ is convex if and only if $\forall x, y \in \mathbb{R}^d$ and $\forall \lambda \in [0, 1]$, we have

$$f(\lambda x + (1 - \lambda)y) \;\leq\; \lambda f(x) + (1 - \lambda)f(y)$$

In the following sections, we define some terms that will be used over the course of the lectures and then we move on to understand what convexity means and some direct results and requirements of convexity.

2. Definitions

An important inequality satisfied by convex functions is *Jensen's Inequality* below which is used extensively for proving results that follow from convexity of functions.

Theorem 4.1 (Jensen's Inequality) For a convex function f, suppose we have $\{x_n\}_{n=1}^N$ where $\forall n \in [N]$, $x_n \in \mathbb{R}^d$, we have $\{\lambda_n\}_{n=1}^N$ such that $\lambda_n \geq 0$ for all $n \in [N]$ and $\sum_{n=1}^N \lambda_n = 1$, then

$$f\left(\sum_{n=1}^N \lambda_n x_n\right) \leq \sum_{n=1}^N \lambda_n f(x_n)$$

Definition 4.4 (Convex Hull) For a convex set $\mathrm{S} \subseteq \mathbb{R}^d$, the *convex hull* of S is the smallest convex set $C \subseteq \mathbb{R}^d$ such that $S \subseteq C$. For a set $\mathrm{S} = \{x_n\}_{n=1}^{N}$, the convex hull of S can be written as

$$\text{Convex Hull}\,(\mathrm{S}) \;=\; \left\{\sum_{n=1}^{N} \lambda_n x_n \,|\, \sum_{n=1}^{N} \lambda_n = 1; \forall n \in [\,N\,]\,, \lambda_n \geq 0\right\}$$

3. Meaning of Convexity

In this section, we discuss some results that follow from the convexity of functions that are differentiable everywhere in the real space. Before moving to that, we shall discuss some basic definitions required to analyse these results.

Definition 4.5 (Gradient) A gradient of a function $f : \mathbb{R}^d \to \mathbb{R}$ at point $x \in \mathbb{R}^d$, represented by $\nabla f(x)$ is given as

$$\nabla f(x) \;=\; \left[\frac{\partial f(x)}{\partial x_1} \ldots \frac{\partial f(x)}{\partial x_i} \ldots \frac{\partial f(x)}{\partial x_d}\right]^{\mathrm{T}}.$$

The second derivate of the function f is known as the *hessian* matrix of the function f. This is defined formally as below.

Definition 4.6 (Hessian) The *hessian* of a function $f : \mathbb{R}^d \to \mathbb{R}$ at a point $x \in \mathbb{R}^d$, represented as $\mathrm{H}_f(x)$ is defined as the second derivative of f at the point x, given as

$$\mathrm{H}_f(x) \;=\; \nabla^2 f(x) \;=\; \left[\frac{\partial^2 f(x)}{\partial x_i \partial x_j}\right]_{ij}$$

We are now equipped with enough definitions to look at some results of convexity of functions.

Theorem 4.2 If a function $f : \mathbb{R}^d \to \mathbb{R}$ is convex and differentiable at $x \in \mathbb{R}^d$, then $\forall y \in \mathbb{R}^d$ we have

$$f(y) \;\geq\; f(x) + \nabla f(x)^{\mathrm{T}}(y - x)$$

Proof. From convexity of f, we have

$$f\left((1-\lambda)x + \lambda y\right) \;\leq\; (1-\lambda)f(x) + \lambda f(y)$$
$$\implies f(x) + \frac{f(x) + \lambda(y - x)}{\lambda} \;\leq\; f(y)$$

As we limit λ to 0, we have

$$\lim_{\lambda \to 0} f(x) + \frac{f(x) + \lambda(y - x)}{\lambda} \;=\; f(x) + \nabla f(x)^{\mathrm{T}}(y - x)$$

The above follows by considering the directional derivative of f at x in the direction $(y-x)$. Hence, we have

$$f(y) \;\geq\; f(x) + \nabla f(x)^{\mathrm{T}}(y-x)$$

$\square$

Theorem 4.3 Suppose a function $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable everywhere and $\forall x, y \in \mathbb{R}^d$, we have

$$f(y) \;\geq\; f(x) + \nabla f(x)^{\mathrm{T}}(y-x)$$

then f is convex.

Proof. Suppose $z = \lambda x + (1-\lambda)y$. Then we can write

$$\begin{aligned} f(y) &\;\geq\; f(z) + \nabla f(z)(y-z) \\ f(x) &\;\geq\; f(z) + \nabla f(z)(x-z) \end{aligned}$$

Therefore, we have

$$\begin{aligned} \lambda f(x) + (1-\lambda) f(y) &\;\geq\; f(z) + \nabla f(z)(\lambda x + (1-\lambda)y - (\lambda + 1 - \lambda))z \\ \implies f(x) + (1-\lambda) f(y) &\;\geq\; f\left(\lambda x + (1-\lambda)y\right) \end{aligned}$$

Therefore, f is convex $\square$

4. Convexity and Gradients

Not all convex functions are differentiable everywhere. To overcome this, we define subgradients that do not require the function to be differentiable and are helpful in defining and analysing the convexity of functions.

Definition 4.7 (Subgradient) For a function $f : \mathbb{R}^d \to \mathbb{R}$ and a point $x \in \mathbb{R}^d$, a vector $\mathbf{g} \in \mathbb{R}^d$ is called a *subgradient* of the function f at point x iff

$$f(y) \;\geq\; f(x) + \mathbf{g}^{\mathrm{T}}(y-x)$$

If f is convex and

(a) if f is differentiable at x, then there is the only one subgradient and that is equal to the gradient $\nabla f(x)$

(b) if f is non-differentiable at x, there there exist infinite subgradients.

The notion of subgradients is only defined rigorously for convex functions even though it is possible for subgradients to exist for non-convex functions.

Example: Suppose we have a function $f : \mathbb{R} \to \mathbb{R}$ defined as $f(x) = |\,x\,|$ (note that this function is convex), we can find the subgradients at different points such as

1. for $x = 1$, the function is differentiable and hence there is only one subgradient which is equal to $\nabla f(1) \implies g = 1$
2. similarly, for $x = -1$, we have $g = -1$
3. in this case, we have infintely many subgradients since this is a point of non-differentiability. $\therefore x = 0, g \in [\,-1, 1\,]$ (refer to figure 1)

This set of subgradients is called *sub-differential set.*

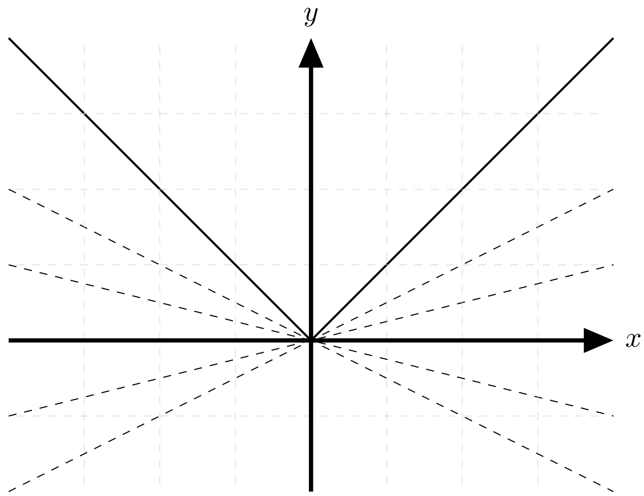

Figure 1: Subgradients for $f = |\,x\,|$

Theorem 4.4 If a function $f : \mathbb{R}^d \to \mathbb{R}$ is convex and twice differentiable at x, then the Hessian of f at x is a positive semi-definitive matrix, *i.e.* $\nabla^2 f(x) \succcurlyeq 0$.

Proof. From the properties of a convex function, we know $\forall x, y \in \mathbb{R}^d$ and $\forall \lambda \in \mathbb{R}$ we have

$$f(x + \lambda y) \;\geq\; f(x) + \lambda \nabla f(x)^{\mathrm{T}} y$$

Also, using Taylor's Theorem, we can write

$$f(x + \lambda y) \;=\; f(x) + \lambda \nabla f(x)^{\mathrm{T}} y + \frac{\lambda^2}{2} y^{\mathrm{T}} \nabla^2 f(\boldsymbol{\gamma}) y$$

for some $\boldsymbol{\gamma}$ on the line segment joining x and $x + \lambda y$. Combining the two equations, we have

$$\frac{\lambda^2}{2} y^{\mathrm{T}} \nabla^2 f(\boldsymbol{\gamma}) y \;\geq\; 0$$

$$\implies y^{\mathrm{T}} \nabla^2 f(\boldsymbol{\gamma}) y \;\geq\; 0$$

Allowing λ to tend to zero, we will have $\boldsymbol{\gamma} \to x$. Hence, we have

$$\implies y^{\mathrm{T}} \nabla^2 f(x) y \;\geq\; 0$$

Since this is true for all values of $x, y \in \mathbb{R}^d$, we have satisfied the sufficient condition for a matrix to be positive semi-definite. Therefore, $\nabla^2 f(x)$ is positive semi-definite. □

Theorem 4.5 If a function $f : \mathbb{R}^d \to \mathbb{R}$ is twice differentiable at all points $x \in \mathbb{R}^d$ and $\nabla^2 f(x) \succcurlyeq 0$, then f is convex.

Proof. Similar to last proof discussed, we can write the following $\forall x, y \in \mathbb{R}^d$ using Taylor's Theorem

$$f(y) \;=\; f(x) + \nabla f(y)^{\mathrm{T}}(y-x) + \frac{1}{2}(y-x)^{\mathrm{T}}\nabla^2 f(\boldsymbol{\gamma})(y-x)$$

for some $\boldsymbol{\gamma} \in [\,x, y\,]$. Since the hessian of f at all points is positive semi-definite, the term $(y-x)^{\mathrm{T}}\nabla^2 f(\boldsymbol{\gamma})(y-x)$ is positive. Hence, we can write

$$f(y) \;\geq\; f(x) + \nabla f(y)^{\mathrm{T}}(y-x)$$

From the definition of convexity, we know that the above condition is sufficient to show that f is convex. □

5. Strong Convexity

In this section we define one more concept of convexity, strong convexity. Strong convexity helps us provide strict conditions which can be used to do convergence analysis of algorithms which we will see later.

Definition 4.8 (Strong Convexity) A function $f : \mathbb{R}^d \to \mathbb{R}$ is α-strongly convex for some $\alpha > 0$ if and only if for all $x \in \mathbb{R}^d$, there is a subgradient $g \in \mathbb{R}^d$ of f at x such that for all $y \in \mathbb{R}^d$, we have

$$f(y) \;\geq\; f(x) + g^{\mathrm{T}}(y-x) + \frac{\alpha}{2}\,\|\, y - x\,\|^2$$