Statistical Optimization: Lecture 3

Optimization and Convexity

Zijian Guo

Zhejiang University
Center for data science

March 17, 2026

Outline

Introduction to Optimization

Convex Function

- Mathematical Preparation
- Convex sets
- Convex sets and functions
- Operations that preserve convexity

Characterization of Convexity

What is Optimization?

- **Optimization** is about making the *best choice* among many possibilities.
- We describe a choice by a variable x (a number or a vector).
- We measure how good a choice is by an **objective function** $f(x)$.
- In this course we mainly use **minimization**:
    - smaller $f(x)$ means a better choice.
- The goal: find a point x (among allowed choices) that makes $f(x)$ as small as possible.

Optimization Problem: Standard Form

An optimization problem has three basic ingredients:

1. **Decision variable:** x (what we choose);
2. **Objective:** $f(x)$ (what we want to minimize);
3. **Feasible set:** $\mathcal{X}$ (which choices are allowed).

Standard form:

$$\arg\min\{f(x) : x \in \mathcal{X}\} \qquad \text{or equivalently} \qquad \arg\min_{x \in \mathcal{X}} f(x).$$

If there are no restrictions, we can take $\mathcal{X} = \mathbb{R}^d$.

Global Minimum

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

(1) Global minimum value (best possible value):

$$f^\star := \inf_{x \in \mathcal{X}} f(x).$$

(2) Global minimizer (a best point): a point $x^\star \in \mathcal{X}$ such that

$$f(x^\star) \leq f(x) \quad \text{for all } x \in \mathcal{X}.$$

If such an $x^\star$ exists, then $f(x^\star) = f^\star$ and we write

$$x^\star \in \arg\min_{x \in \mathcal{X}} f(x).$$

The value $f^\star$ can be finite even if no minimizer exists (i.e., the infimum is not attained).

ϵ-Optimality

Consider

$$\min_{x \in \mathcal{X}} f(x), \qquad f^\star := \inf_{x \in \mathcal{X}} f(x).$$

In practice, finding an exact global minimizer may be difficult.

Definition: For a given $\epsilon \geq 0$, a point $x \in \mathcal{X}$ is called ϵ-optimal if

$$f(x) \leq f^\star + \epsilon.$$

This means the objective value at x is at most ϵ above the best possible value.

Special case:

- If $\epsilon = 0$, then x is an exact global minimizer.
- If $\epsilon > 0$, then x is an ϵ global minimizer.

Local Minimum

Definition: A point $x^\star \in \mathcal{X}$ is called a *local minimizer* if there exists some radius $r > 0$ such that

$$f(x^\star) \leq f(x) \quad \text{for all } x \in \mathcal{X} \text{ with } \|x - x^\star\| \leq r.$$

This means $x^\star$ is the best point only in a small neighborhood around it, not necessarily in the whole feasible set.

- A **global minimizer** is best over all $x \in \mathcal{X}$.
- A **local minimizer** is best only among nearby feasible points.

Every global minimizer is a local minimizer, but a local minimizer need not be global.

Outline

Introduction to Optimization

Convex Function

- Mathematical Preparation
- Convex sets
- Convex sets and functions
- Operations that preserve convexity

Characterization of Convexity

Outline

Introduction to Optimization

Convex Function

- Mathematical Preparation
- Convex sets
- Convex sets and functions
- Operations that preserve convexity

Characterization of Convexity

Derivatives and Partial Derivatives

Derivative. For $h : \mathbb{R} \to \mathbb{R}$, the derivative at x is

$$h'(x) = \lim_{t \to 0} \frac{h(x+t) - h(x)}{t},$$

if the limit exists.

Partial derivatives. For $f : \mathbb{R}^d \to \mathbb{R}$, the partial derivative with respect to x_i is

$$\frac{\partial f}{\partial x_i}(\mathbf{x}) = \lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{e}_i) - f(\mathbf{x})}{t}.$$

where e_i is the i-th coordinate vector.

Chain rule. Let $g : \mathbb{R}^d \to \mathbb{R}^m$ and $F : \mathbb{R}^m \to \mathbb{R}^r$ be differentiable. Then

$$(F \circ g)'(\mathbf{x}) = F'(g(\mathbf{x}))\, g'(\mathbf{x}).$$

Differentiability

Gradient: We collect the partial derivatives into the vector

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \frac{\partial f}{\partial x_2}(x), \dots, \frac{\partial f}{\partial x_d}(x) \right)^\top .$$

If all partial derivatives exist, we say f is **differentiable**.

Hessian: If f is **twice differentiable**, its Hessian matrix is

$$\nabla^2 f(x) = \left(\frac{\partial^2 f}{\partial x_i \, \partial x_j}(x) \right)_{i,j=1}^{d} = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d}(x) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_2^2}(x) & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_d}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1}(x) & \frac{\partial^2 f}{\partial x_d \partial x_2}(x) & \cdots & \frac{\partial^2 f}{\partial x_d^2}(x) \end{pmatrix} .$$

Open and Closed Sets

Open set.
A set $C \subseteq \mathbb{R}^d$ is called *open* if for every $x \in C$ there exists $\varepsilon > 0$ such that

$$\{y : \|y - x\| < \varepsilon\} \subseteq C.$$

This means that every point in C has a small neighborhood fully contained in C.

Closed set.
A set $C \subseteq \mathbb{R}^d$ is called *closed* if it contains all its limit points.
Equivalently, whenever a sequence $x_k \in C$ satisfies $x_k \to x$, then $x \in C$.

Examples:

$$(0,1) \text{ is open}, \qquad [0,1] \text{ is closed.}$$

Taylor Expansion

Let $f : \mathbb{R}^d \to \mathbb{R}$.

First-order Taylor expansion. If f is differentiable at x, then

$$f(y) = f(x) + \nabla f(x)^\top (y - x) + o(\|y - x\|_2).$$

Second-order Taylor expansion. If f is twice differentiable near x, then

$$f(y) = f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2}(y - x)^\top \nabla^2 f(x)(y - x) + o(\|y - x\|_2^2).$$

Outline

Introduction to Optimization

Convex Function

- Mathematical Preparation
- Convex sets
- Convex sets and functions
- Operations that preserve convexity

Characterization of Convexity

Definition: Convex sets

Definition. A set $C \subseteq \mathbb{R}^d$ is *convex* if for any two points $\mathbf{x}, \mathbf{y} \in C$, the connecting line segment is contained in C. In formulas, if for all $\lambda \in [0, 1]$,

$$\lambda \mathbf{x} + (1 - \lambda)\mathbf{y} \in C.$$

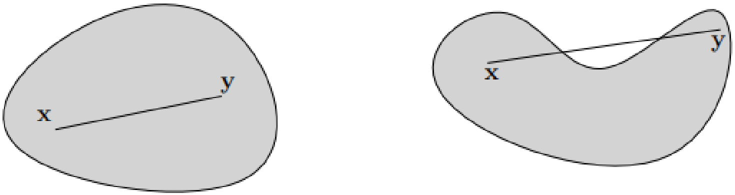


Figure 2.2: A convex set (left) and a non-convex set (right)

Examples of Convex Sets

Convex sets:

- **Interval on a line:** $[a, b] \subset \mathbb{R}$.
- **Line (or plane):** a straight line in $\mathbb{R}^2$, a plane in $\mathbb{R}^3$.
- **Ball:** $\{x : \|x - x_0\|_2 \leq r\}$.
- **Rectangle:** $\{x : \ell \leq x \leq u\}$.

Non-convex:

- **Donut:** $\{x : r_1 \leq \|x\|_2 \leq r_2\}$.
- **Two separated balls:** $B(x_1, r) \cup B(x_2, r)$.

Outline

Introduction to Optimization

Convex Function

- Mathematical Preparation
- Convex sets
- Convex sets and functions
- Operations that preserve convexity

Characterization of Convexity

Convex Function

We consider real-valued functions

$$f : \mathrm{dom}(f) \to \mathbb{R}, \qquad \mathrm{dom}(f) \subseteq \mathbb{R}^d,$$

where $\mathrm{dom}(f)$ is the *domain* of f, i.e. the set of all $\mathbf{x} \in \mathbb{R}^d$ for which $f(\mathbf{x})$ is well-defined.

Definition. A function $f : \mathrm{dom}(f) \to \mathbb{R}$ is *convex* if
(i) $\mathrm{dom}(f)$ is convex, and
(ii) for all $\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$ and all $\lambda \in [0, 1]$,

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}). \tag{2.2}$$

An important special case arises when $f : \mathbb{R}^d \to \mathbb{R}$ is an *affine function*, i.e. $f(\mathbf{x}) = \mathbf{c}^\top \mathbf{x} + c_0$ for some vector $\mathbf{c} \in \mathbb{R}^d$ and scalar $c_0 \in \mathbb{R}$.

Examples of Convex Functions

Convex examples:

- $f(x) = x^2$
- $f(x) = ax + b$
- $f(x) = |x|$

Non-convex examples:

- $f(x) = -x^2$
- $f(x) = \sin x \quad$ (on $\mathbb{R}$)
- $f(x) = x^3$

Jensen's inequality

Theorem (Jensen's inequality). Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function, $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathrm{dom}(f)$, and $\lambda_1, \ldots, \lambda_m \in \mathbb{R}_+$ such that $\sum_{i=1}^m \lambda_i = 1$. Then

$$f\left(\sum_{i=1}^m \lambda_i \mathbf{x}_i\right) \leq \sum_{i=1}^m \lambda_i f(\mathbf{x}_i).$$

For $m = 2$, this is (2.2).

Convexity: Local Minimum is Global Minimum

The main feature that makes convex functions attractive in optimization is that every local minimum is a global one, so we cannot “get stuck” in local optima.

Definition. A local minimum of $f : \mathrm{dom}(f) \to \mathbb{R}$ is a point $\mathbf{x}$ such that there exists $\varepsilon > 0$ with

$$f(\mathbf{x}) \leq f(\mathbf{y}), \quad \forall \mathbf{y} \in \mathrm{dom}(f) \text{ satisfying } \|\mathbf{y} - \mathbf{x}\| < \varepsilon.$$

Theorem. Let $\mathbf{x}^\star$ be a local minimum of a convex function $f : \mathrm{dom}(f) \to \mathbb{R}$. Then $\mathbf{x}^\star$ is a global minimum, meaning that

$$f(\mathbf{x}^\star) \leq f(\mathbf{y}), \quad \forall \mathbf{y} \in \mathrm{dom}(f).$$

Convexity: Local Minimum is Global Minimum

Proof. Suppose there exists $\mathbf{y} \in \mathrm{dom}(f)$ such that $f(\mathbf{y}) < f(\mathbf{x}^\star)$ and define $\mathbf{y}' := \lambda\mathbf{x}^\star + (1-\lambda)\mathbf{y}$ for $\lambda \in (0,1)$. From convexity (2.2), we get that $f(\mathbf{y}') < f(\mathbf{x}^\star)$. Choosing λ so close to 1 that $\|\mathbf{y}' - \mathbf{x}^\star\| < \varepsilon$ yields a contradiction to $\mathbf{x}^\star$ being a local minimum. $\square$

This does not mean that a convex function always has a global minimum. Think of $f(x) = x$ as a trivial example. But also if f is bounded from below over $\mathrm{dom}(f)$, it may fail to have a global minimum ($f(x) = e^x$).

Outline

Introduction to Optimization

Convex Function

- Mathematical Preparation
- Convex sets
- Convex sets and functions
- Operations that preserve convexity

Characterization of Convexity

Operations that preserve convexity

(i) Let $f_1, f_2, \ldots, f_m$ be convex functions, and let $\lambda_1, \lambda_2, \ldots, \lambda_m \in \mathbb{R}_+$. Then the functions

$$f := \max_{i=1,\ldots,m} f_i \quad \text{as well as} \quad f := \sum_{i=1}^{m} \lambda_i f_i$$

are convex on

$$\mathrm{dom}(f) := \bigcap_{i=1}^{m} \mathrm{dom}(f_i).$$

(ii) Let f be a convex function with $\mathrm{dom}(f) \subseteq \mathbb{R}^d$, and let $g : \mathbb{R}^m \to \mathbb{R}^d$ be an affine function, meaning that

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

for some matrix $A \in \mathbb{R}^{d \times m}$ and some vector $b \in \mathbb{R}^d$. Then the composition $f \circ g$ is convex on

$$\mathrm{dom}(f \circ g) := \{x \in \mathbb{R}^m : g(x) \in \mathrm{dom}(f)\}.$$

Proofs

(i) For the pointwise maximum, let $f(x) = \max_{i=1,\ldots,m} f_i(x)$. Then for any $x, y \in \mathrm{dom}(f)$ and $\lambda \in [0,1]$, we have

$$\begin{aligned} f(\lambda x + (1-\lambda)y) &= \max_{i=1,\ldots,m} f_i(\lambda x + (1-\lambda)y) \\ &\leq \max_{i=1,\ldots,m} \big(\lambda f_i(x) + (1-\lambda) f_i(y)\big) \\ &\leq \lambda \max_{i=1,\ldots,m} f_i(x) + (1-\lambda) \max_{i=1,\ldots,m} f_i(y) \\ &= \lambda f(x) + (1-\lambda) f(y), \end{aligned}$$

where the first inequality follows from the convexity of each f_i, and the second from the monotonicity of $\max$. Hence f is convex.
The convexity of $f := \sum_{i=1}^m \lambda_i f_i$ are convex is easy to derive.

Proofs

(ii) Now let f be convex and let $g(x) = Ax + b$ be affine. For any $x, y \in \mathrm{dom}(f \circ g)$ and $\lambda \in [0, 1]$, we have

$$\begin{aligned}
(f \circ g)(\lambda x + (1-\lambda)y) &= f(g(\lambda x + (1-\lambda)y)) \\
&= f(A(\lambda x + (1-\lambda)y) + b) \\
&= f(\lambda(Ax + b) + (1-\lambda)(Ay + b)) \\
&\leq \lambda f(Ax + b) + (1-\lambda)f(Ay + b) \\
&= \lambda(f \circ g)(x) + (1-\lambda)(f \circ g)(y),
\end{aligned}$$

where the inequality follows from the convexity of f. Thus, $f \circ g$ is convex. $\square$

Outline

Introduction to Optimization

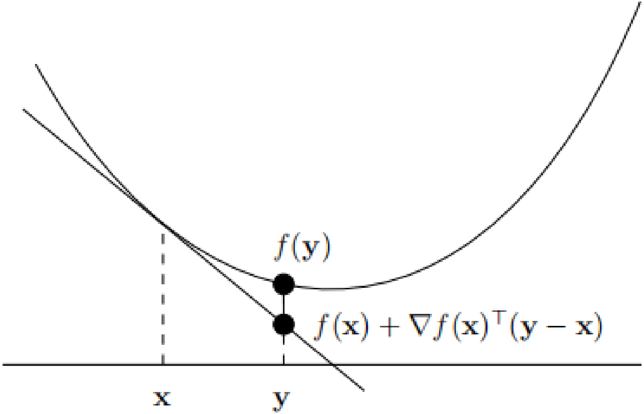
Convex Function

- Mathematical Preparation
- Convex sets
- Convex sets and functions
- Operations that preserve convexity

Characterization of Convexity

First-order characterization of convexity

Geometrically, this means that for all $x \in \mathrm{dom}(f)$, the graph of f lies above its tangent hyperplane at the point $(x, f(x))$.



First-order characterization of convexity

Theorem. Suppose that $\mathrm{dom}(f)$ is open and that f is differentiable; in particular, the gradient (vector of partial derivatives)

$$\nabla f(x) := \left(\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_d}(x) \right)^\top$$

exists at every point $x \in \mathrm{dom}(f)$. Then f is convex if and only if $\mathrm{dom}(f)$ is convex and

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) \tag{2.3}$$

holds for all $x, y \in \mathrm{dom}(f)$.

Proof of "(2.2) $\Rightarrow$ (2.3)"

Suppose that f is convex and differentiable. For any $\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$ and any $\lambda \in [0, 1]$,

$$f\big(\lambda \mathbf{y} + (1-\lambda)\mathbf{x}\big) \leq \lambda f(\mathbf{y}) + (1-\lambda) f(\mathbf{x}).$$

Equivalently,

$$f\big(\mathbf{x} + \lambda(\mathbf{y} - \mathbf{x})\big) - f(\mathbf{x}) \leq \lambda\big(f(\mathbf{y}) - f(\mathbf{x})\big).$$

For $\lambda > 0$, divide by λ:

$$\frac{f\big(\mathbf{x} + \lambda(\mathbf{y} - \mathbf{x})\big) - f(\mathbf{x})}{\lambda} \leq f(\mathbf{y}) - f(\mathbf{x}).$$

Proof of "(2.2) $\Rightarrow$ (2.3)"

Let $\lambda \downarrow 0$. Since f is differentiable at $\mathbf{x}$,

$$\lim_{\lambda \downarrow 0} \frac{f\big(\mathbf{x} + \lambda(\mathbf{y} - \mathbf{x})\big) - f(\mathbf{x})}{\lambda} = \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}).$$

Hence,

$$\nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \leq f(\mathbf{y}) - f(\mathbf{x}),$$

that is,

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}).$$

This proves (2.3). $\square$

Proof of "(2.3) $\Rightarrow$ (2.2)"

Assume that $\mathrm{dom}(f)$ is convex and that

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \qquad \forall \mathbf{x}, \mathbf{y} \in \mathrm{dom}(f).$$

Let $\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$ and $\lambda \in [0, 1]$. Define

$$\mathbf{z} = \lambda \mathbf{x} + (1 - \lambda)\mathbf{y}.$$

Apply (2.3) with $(\mathbf{z}, \mathbf{x})$ and $(\mathbf{z}, \mathbf{y})$:

$$f(\mathbf{x}) \geq f(\mathbf{z}) + \nabla f(\mathbf{z})^\top (\mathbf{x} - \mathbf{z}),$$

$$f(\mathbf{y}) \geq f(\mathbf{z}) + \nabla f(\mathbf{z})^\top (\mathbf{y} - \mathbf{z}).$$

Proof of "(2.3) $\Rightarrow$ (2.2)"

Multiply the first inequality by λ, the second by $1-\lambda$, and add:

$$\lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \geq f(\mathbf{z}) + \nabla f(\mathbf{z})^\top \big(\lambda(\mathbf{x}-\mathbf{z}) + (1-\lambda)(\mathbf{y}-\mathbf{z})\big).$$

Since

$$\mathbf{z} = \lambda \mathbf{x} + (1-\lambda)\mathbf{y},$$

we have

$$\lambda(\mathbf{x}-\mathbf{z}) + (1-\lambda)(\mathbf{y}-\mathbf{z}) = 0.$$

Therefore,

$$f(\mathbf{z}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}).$$

That is,

$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}),$$

so f is convex. $\square$

Monotonicity of the gradient

Theorem. Suppose that $\mathrm{dom}(f)$ is open and that f is differentiable. Then f is convex if and only if $\mathrm{dom}(f)$ is convex and

$$(\nabla f(y) - \nabla f(x))^\top (y - x) \geq 0 \tag{2.4}$$

holds for all $x, y \in \mathrm{dom}(f)$.
The inequality (2.4) is known as *monotonicity of the gradient*.

Proof of “(2.3) $\Rightarrow$ (2.4)”

Assume that f is convex. By the first-order characterization of convexity, for all $x, y \in \mathrm{dom}(f)$,

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

and, exchanging the roles of x and y,

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

Adding these two inequalities gives

$$0 \geq \nabla f(x)^\top (y - x) + \nabla f(y)^\top (x - y).$$

Since $x - y = -(y - x)$, this becomes

$$(\nabla f(y) - \nabla f(x))^\top (y - x) \geq 0.$$

This proves the monotonicity of the gradient. $\square$

Proof of "(2.4) $\Rightarrow$ (2.3)"

For the other direction, suppose that monotonicity of the gradient (2.4) holds. Then we in particular have

$$(\nabla f(x + t(y - x)) - \nabla f(x))^\top t(y - x) \geq 0$$

for all $x, y \in \mathrm{dom}(f)$ and $t \in (0, 1]$. Dividing by t, this yields

$$(\nabla f(x + t(y - x)) - \nabla f(x))^\top (y - x) \geq 0. \tag{2.5}$$

Fix $x, y \in \mathrm{dom}(f)$. For $t \in [0, 1]$, let $h(t) := f(x + t(y - x))$. In our case where f is real-valued, (2.1) yields $h'(t) = \nabla f(x + t(y - x))^\top (y - x)$, $t \in (0, 1)$.

Proof of "(2.4) $\Rightarrow$ (2.3)"

Hence, (2.5) can be rewritten as

$$h'(t) \geq \nabla f(x)^\top (y - x), \qquad t \in (0, 1).$$

Lemma(Mean Value Theorem). If $h : [0, 1] \to \mathbb{R}$ is continuous on $[0, 1]$ and differentiable on $(0, 1)$, then there exists $c \in (0, 1)$ such that

$$h(1) - h(0) = h'(c).$$

By the mean value theorem, there is $c \in (0, 1)$ such that $h'(c) = h(1) - h(0)$. Then

$$\begin{aligned} f(y) = h(1) &= h(0) + h'(c) \\ &= f(x) + h'(c) \\ &\geq f(x) + \nabla f(x)^\top (y - x). \end{aligned}$$

This is the first-order characterization of convexity. $\square$

Positive Semidefinite and Positive definite Matrices

Positive semidefinite. A symmetric matrix $A \in \mathbb{R}^{d \times d}$ is *Positive Semidefinite* (PSD), written $A \succeq 0$, if

$$v^\top A v \geq 0 \qquad \text{for all } v \in \mathbb{R}^d.$$

Positive definite. It is *Positive Definite* (PD), written $A \succ 0$, if

$$v^\top A v > 0 \qquad \text{for all } v \neq 0.$$

Second-order characterization of convexity

Theorem. Suppose that $\mathrm{dom}(f)$ is open and that f is twice differentiable; in particular, the Hessian (matrix of second partial derivatives)

$$\nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d}(x) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_2^2}(x) & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_d}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1}(x) & \frac{\partial^2 f}{\partial x_d \partial x_2}(x) & \cdots & \frac{\partial^2 f}{\partial x_d^2}(x) \end{pmatrix}$$

exists at every point $x \in \mathrm{dom}(f)$ and is symmetric. Then f is convex if and only if $\mathrm{dom}(f)$ is convex, and for all $x \in \mathrm{dom}(f)$, we have

$$\nabla^2 f(x) \succeq 0 \qquad \text{(i.e. } \nabla^2 f(x) \text{ is positive semidefinite).} \tag{2.6}$$

Proof of "f convex $\Rightarrow \nabla^2 f(x) \succeq 0$"

Assume that f is convex. Fix $x \in \mathrm{dom}(f)$ and $v \in \mathbb{R}^d$. Since $\mathrm{dom}(f)$ is open, there exists an open interval $I \ni 0$ such that

$$x + tv \in \mathrm{dom}(f), \qquad \forall t \in I.$$

Define

$$h(t) := f(x + tv), \qquad t \in I.$$

Because f is convex, the one-dimensional function h is also convex.

Since f is twice differentiable,

$$h'(t) = \nabla f(x + tv)^\top v, \qquad h''(t) = v^\top \nabla^2 f(x + tv)\, v.$$

Proof of "f convex $\Rightarrow \nabla^2 f(x) \succeq 0$"

A twice differentiable convex function on an interval satisfies

$$h''(t) \geq 0 \qquad \forall t \in I.$$

In particular, at $t = 0$,

$$h''(0) \geq 0.$$

Therefore,

$$v^\top \nabla^2 f(x)\, v = h''(0) \geq 0.$$

Since this holds for every $v \in \mathbb{R}^d$, we conclude that

$$\nabla^2 f(x) \succeq 0.$$

$\square$

Proof of "$\nabla^2 f(x) \succeq 0 \;\Rightarrow f$ convex"

Now assume that $\mathrm{dom}(f)$ is convex and that

$$\nabla^2 f(x) \succeq 0 \qquad \forall x \in \mathrm{dom}(f).$$

Fix $x, y \in \mathrm{dom}(f)$, and let

$$v := y - x.$$

Define

$$h(t) := f(x + t(y - x)) = f(x + tv), \qquad t \in [0, 1].$$

Then

$$h'(t) = \nabla f(x + tv)^\top v, \qquad h''(t) = v^\top \nabla^2 f(x + tv)\, v \geq 0.$$

Hence h' is nondecreasing on $[0, 1]$, so

$$h'(1) - h'(0) \geq 0.$$

Proof of "$\nabla^2 f(x) \succeq 0 \;\Rightarrow f$ convex"

Spelling out the inequality $h'(1) - h'(0) \geq 0$, we obtain

$$\nabla f(y)^\top (y - x) - \nabla f(x)^\top (y - x) \geq 0,$$

that is,

$$(\nabla f(y) - \nabla f(x))^\top (y - x) \geq 0.$$

So the gradient is monotone.

$\square$