1 Outline

In this lecture, we study

- convex sets and functions,
- operations preserving convexity,
- first- and second-order characterizations of convex functiosn,
- convex optimization.

2 Convex Sets and Convex Functions

2.1 Convex Sets

Definition 2.1. A set $X \subseteq \mathbb{R}^d$ is *convex* if for any $u, v \in X$ and any $\lambda \in [0,1]$,

$$\lambda u + (1-\lambda)v \in X.$$

In words, the line segment joining any two points is entirely contained the set. In Figure 2.1, we have a convex set and a non-convex set.

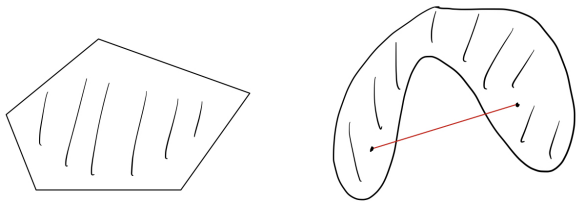

Figure 2.1: A convex set and a nonconvex set

There are many examples of convex sets.

1. Empty set, singletons (sets of the form $\{v\}$),
2. Norm ball: $\{x \in \mathbb{R}^d : \|x - c\| \leq r\}$ where c is the center.
3. Ellipsoid: $\{x \in \mathbb{R}^d : (x-c)^\top P(x-c) \leq 1\}$ where P is positive definitie and c is the center.
4. Hyperplane: $\{x \in \mathbb{R}^d : a^\top x = b\}$ where $a \in \mathbb{R}^d$ and $b \in \mathbb{R}$.
5. Half-space: $\{x \in \mathbb{R}^d : a^\top x \leq b\}$ where $a \in \mathbb{R}^d$ and $b \in \mathbb{R}$.
6. Polyhedron: A *polyhedron* is a finite intersection of half spaces, $\{x \in \mathbb{R}^d : Ax \leq b\}$ where $A \in \mathbb{R}^{m \times d}$ and $b \in \mathbb{R}^m$. Here, $Ax \leq b$ is a short-hand notation for system $a_k^\top x \leq b_k$ for $k \in [m]$.

7. Polytope: A *polytope* is a polyhedron that is bounded. Equivalently, a polytope is the convex hull of some finite set of vectors.

8. Simplex: A set of the form $\{x \in \mathbb{R}^d : 1^\top x = 1,\ x \geq 0\}$, which is equal to the convex hull of $e^1, \ldots, e^d$, the d-dimensional unit vectors.

9. Nonnegative orthant: $\mathbb{R}^d_+ = \{x \in \mathbb{R}^d : x \geq 0\}$.

10. Positive orthant: $\mathbb{R}^d_{++} = \{x \in \mathbb{R}^d : x > 0\}$.

11. Norm cone: $\{(x, t) \in \mathbb{R}^d \times \mathbb{R} : \|x\| \leq t\}$. When $\|\cdot\|$ is the Euclidean norm, the cone is called the *second-order cone*.

12. Positive semidefinite cone: The set of all positive semidefinite matrices of a fixed dimension.

2.2 Convex Functions

Definition 2.2. A function $f : \mathbb{R}^d \to \mathbb{R}$ is *convex* if the domain, denoted $\mathrm{dom}(f)$, is convex and for all $x, y \in \mathrm{dom}(f)$, we have

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \quad \text{for } 0 \leq \lambda \leq 1.$$

In words, function f evaluated at a point between x and y lies below the line segment joining $f(x)$ and $f(y)$.

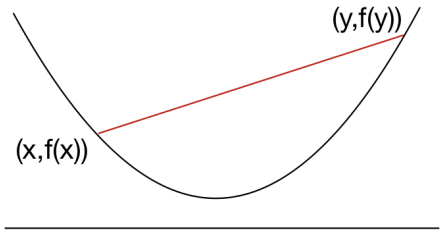

Figure 2.2: Illustration of a convex function in $\mathbb{R}^2$

Definition 2.3. We say that $f : \mathbb{R}^d \to \mathbb{R}$ is *concave* if $-f$ is convex.

Definition 2.4. A function $f : \mathbb{R}^d \to \mathbb{R}$ is

- *strictly convex* if $\mathrm{dom}(f)$ is convex and for any distinct $x, y \in \mathrm{dom}(f)$, we have

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda) f(y) \quad \text{for } 0 < \lambda < 1.$$

- *strongly convex* if $f(x) - \alpha\|x\|^2$ is convex for some $\alpha > 0$ and norm $\|\cdot\|$.

Note that strong convexity implies strict convexity, and strict convexity implies convexity.

There are many examples of convex functions.

- Exponential function: e^{ax} for any $a \in \mathbb{R}$.

- Power function: x^a for $a \geq 1$ over $\mathbb{R}_+$ and x^a for $a < 0$ over $\mathbb{R}_{++}$.
  x^a for $0 \leq a < 1$ over $\mathbb{R}_+$ is concave.
- Logarithm: $\log x$ is concave on $\mathbb{R}_{++}$.
- Negative entropy: $x \log x$ on $\mathbb{R}_{++}$.
- Linear function: $a^\top x + b$ where $a \in \mathbb{R}^d$ and $b \in \mathbb{R}$ are both convex and concave.
- Quadratic function: $\frac{1}{2}x^\top Ax + b^\top x + c$ where $A \succeq 0$, $b \in \mathbb{R}^d$, and $c \in \mathbb{R}$.
- Least squares loss: $\|b - Ax\|_2^2$ for any A.
- Norm: Any norm $\|\cdot\|$ is concex, because a norm is subadditive and homogeneous.
- Maximum eigenvalue of a symmetric matrix.
- Indicator function: When C is convex, its indicator function, given by,

$$I_C(x) = \begin{cases} 0, & x \in C \\ \infty, & x \notin C \end{cases}$$

  is convex.
- Support function: Given a convex set C, its support function is defined as

$$I_C^*(x) = \sup_{y \in C} \left\{ y^\top x \right\}.$$

- Conjugate function: Given an arbitrary function $f : \mathbb{R}^d \to \mathbb{R}$, the conjugate function f^* is defined as

$$f^*(x) = \sup_{y \in \mathbb{R}^d} \left\{ y^\top x - f(y) \right\}.$$

Example 2.5. Recall the linear regression problem with a given set of data $(x_1, y_1), \ldots, (x_n, y_n)$. We considered the mean squared error given by

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - w^\top x_i)^2.$$

Let y denote the vector whose components are $y_1, \ldots, y_n$, and let X denote the matrix whose rows are $x_1^\top, \ldots, x_n^\top$. Then it follows that

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - w^\top x_i)^2 = \frac{1}{n} \|y - Xw\|_2^2.$$

Here, we know that $\|y - Xw\|_2^2$ is convex in w, so the mean squared loss is a convex function of w.

2.3 Operations Preserving Convexity

For many problems, it is important to recognize underlying convex structures. We can determine whether certain sets and functions are convex by understanding basic rules. Moreover, based on these rules, we can build complex convex sets and functions from simpler ones.

Set Operations We first consider set operations that preserve convexity.

- Intersection: The intersection of any (possibly infinite) collection of covex sets is convex.
- Scaling: Given a convex set C and $\alpha \in \mathbb{R}$,

$$\alpha C = \{\alpha x : \ x \in C\}.$$

- Minkowski sum: Given convex sets $C_i \subseteq \mathbb{R}^d$ for $i = 1, \ldots, k$, the Minkowski sum of them, defined by

$$C_1 + \cdots + C_k = \{x^1 + \cdots + x^k : \ x^i \in C_i \text{ for } i = 1, \ldots, k\}$$

  is convex.

- Cartesian Product: Given convex sets $C_i \subseteq \mathbb{R}^{d_i}$ for $i = 1, \ldots, k$, the Cartesian product of them, defined by

$$C_1 \times \cdots \times C_k = \{(x^1, \ldots, x^k) \in \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_k} : \ x^i \in C_i \text{ for } i = 1, \ldots, k\}$$

  is convex.

- Affine image: Given a convex set C and matrices $A \in \mathbb{R}^{p \times d}$, $b \in \mathbb{R}^p$, we define an affine mapping $f(x) = Ax + b : \mathbb{R}^d \to \mathbb{R}^p$. Then

$$f(C) = \{Ax + b : \ x \in C\}.$$

- Inverse affine image: Given a convex set C and matrices $A \in \mathbb{R}^{p \times d}$, $b \in \mathbb{R}^p$, we define an affine mapping $f(x) = Ax + b : \mathbb{R}^d \to \mathbb{R}^p$. Then

$$f^{-1}(C) = \{x : \ Ax + b \in C\}.$$

Function Operations We next consider function operations preserving convexity.

- Nonnegative weighted sum: Let $f_1, \ldots, f_k : \mathbb{R}^d \to \mathbb{R}$ be convex functions. Then for any $\alpha_1, \ldots, \alpha_k \geq 0$,

$$\alpha_1 f_1 + \cdots + \alpha_k f_k$$

  is convex.

- Maximum of arbitrary collection of convex functions: Let $\{f_\gamma\}_{\gamma \in \Gamma}$ be a collection of convex functions. Then $\max_{\gamma \in \Gamma} f_\gamma$ is also convex. Here, Γ may be infinite.
- Minimizing out variables: Let $g(x, y)$ be convex function in (x, y). Define f by $f(x) = \inf_{y \in C} g(x, y)$ for some convex set C. Then f is convex.
- Perspective function: Let $g(x)$ be a convex function. Then $f(x, t) = tg(x/t)$ is a convex function in $(x, t) \in \mathbb{R}^d \times \mathbb{R}_{++}$. Here, f is called the perspective of g.
- Affine composition: Let $g : \mathbb{R}^p \to \mathbb{R}$ be a convex function, and take matrices $A \in \mathbb{R}^{p \times d}$, $b \in \mathbb{R}^p$. Then $f : \mathbb{R}^d \to \mathbb{R}$ defined by $f(x) = g(Ax + b)$ is convex.
- Compositions: Let $h : \mathbb{R} \to \mathbb{R}$ be a univariate non-decreasing convex function, and let $g : \mathbb{R}^d \to \mathbb{R}$ be convex. Then $f = h \circ g$ is convex.

Example 2.6. For linear regression, we considered the mean squared error with an ℓ_1-regularization term given by

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - w^\top x_i)^2 + \lambda\|w\|_1. \tag{2.1}$$

We saw that the mean squared error is convex in w. Moreover, the ℓ_1 norm is also convex. As (2.1) is the sum of two convex functions, it is a convex function.

3 First- and Second-Order Characterizations of Convex Functions

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function. Let e^i denote the ith unit vector. For example, $e^1 = (1, 0, \ldots, 0)^\top$ and $e^d = (0, \ldots, 0, 1)^\top$. Then the ith *partial derivative* of f is defined as

$$\frac{\partial f}{\partial x_i}(x) = \lim_{t\to 0}\frac{f(x + te^i) - f(x)}{t}.$$

Thus, the ith partial derivative is the directional derivative of f along the ith unit direction e^i. If all the partial derivatives of f exist at $x \in \mathbb{R}^d$, then we may define the *gradient* of f at x, given by

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

The following results provides a first-order characterization of convex functions.

Theorem 2.7. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. Then f is convex if and only if $dom(f)$ is convex and*

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

for all $x, y \in dom(f)$.

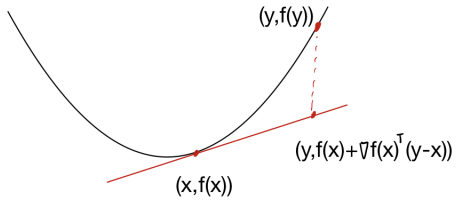

Figure 2.3: Illustration of the first-order characterization

What follows is another first-order characterization.

Theorem 2.8. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. Then f is convex if and only if $dom(f)$ is convex and*

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq 0$$

for all $x, y \in dom(f)$.

Next, we consider

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(x) \quad \text{for } i, j \in [d]$$

are the second partial derivatives of f. If all the second partial derivatives exist, then we may define the *Hessian* of f as follows

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

Moreover, if the second partial derivatives are continuous, then Schwarz's theorem implies that the Hessian is symmetric, i.e.,

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(x) = \frac{\partial^2 f}{\partial x_j \partial x_i}(x) \quad \text{for every } i, j \in [d].$$

Theorem 2.9. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a twice differentiable function[1]. Then f is convex if and only if $dom(f)$ is convex and*

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

for all $x \in dom(f)$.

4 Convex Optimization

A convex optimization problem is to minimize a convex function f over a convex domain C as follows.

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & x \in C. \end{aligned} \tag{P}$$

It is often difficult to obtain an exact minimizer of the problem. Instead, we seek for an approximately optimal solution. Namely, for a given error tolerance $\epsilon > 0$, the goal is to find a solution x such that

$$f(x) \leq \min_{x \in C} f(x) + \epsilon.$$

We say that a solution x satisfies this condition is ϵ-optimal. One can predict that as ϵ gets smaller, it is harder to find an ϵ-optimal solution. Then, how do we measure how hard it is to find an ϵ-optimal solution? Here, we need some notion of computational complexity.

A computational abstraction for convex optimization is the notion of *oracles*. A *first-order oracle* returns the gradient $\nabla f(x)$ of a given solution x. A *second-order oracle* returns the Hessian $\nabla^2 f(x)$ of a given solution x. A *zeroth-order oracle* returns the objective value $f(x)$ of a given solution x. Here, we say that an algorithm is a *first-order method* if it relies on a first-order oracle. How do we measure the complexity of an algorithm for convex optimization? Basically, we count the number of oracle calls to find an ϵ-optimal solution. For example, the famous gradient descent method finds an ϵ-optimal solution with $O(1/\epsilon^2)$ first-order oracle calls. Here, we often simply say "iterations" instead of "oracle calls".

[1] $\nabla^2 f$ exists at any point in $\mathrm{dom}(f)$, and $\mathrm{dom}(f)$ is open.