Optimization for Machine Learning
机器学习中的优化方法

陈 程

华东师范大学 软件工程学院

chchen@sei.ecnu.edu.cn

Outline

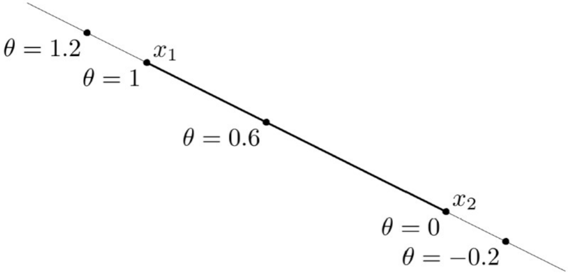
Outline

line through $\mathbf{x}_1$ and $\mathbf{x}_2$: all points

$$\mathbf{x} = \theta \mathbf{x}_1 + (1-\theta)\mathbf{x}_2, \quad \theta \in \mathbb{R}.$$

line segment between $\mathbf{x}_1$ and $\mathbf{x}_2$: all points

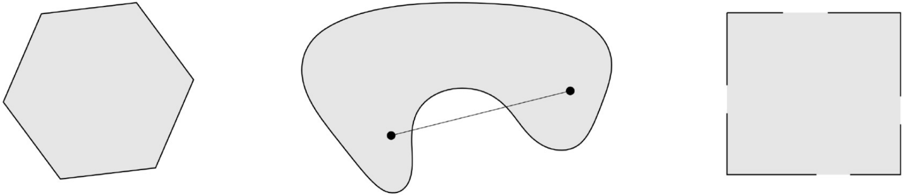
$$\mathbf{x} = \theta \mathbf{x}_1 + (1-\theta)\mathbf{x}_2, \quad 0 \leq \theta \leq 1.$$



Convex sets (凸集)

A set $\mathcal{S} \subseteq \mathbb{R}^n$ is **convex** if the line segment between any two points of $\mathcal{S}$ lies in $\mathcal{S}$, i.e., if for any $\mathbf{x}, \mathbf{y} \in \mathcal{S}$ and $\theta \in [0, 1]$, we have

$$\theta \mathbf{x} + (1 - \theta) \mathbf{y} \in \mathcal{S}.$$



Every two points can see each other.

Properties of convex sets

- If $\mathcal{S}$ is a convex set, then $k\mathcal{S} = \{k\mathbf{s} | k \in \mathbb{R}, \mathbf{s} \in \mathcal{S}\}$ is convex.
- If $\mathcal{S}$ and $\mathcal{T}$ are convex sets, then $\mathcal{S} + \mathcal{T} = \{\mathbf{s} + \mathbf{t} | \mathbf{s} \in \mathcal{S}, \mathbf{t} \in \mathcal{T}\}$ is convex.
- If $\mathcal{S}$ and $\mathcal{T}$ are convex sets, then $\mathcal{S} \times \mathcal{T} = \{(\mathbf{s}, \mathbf{t}) | \mathbf{s} \in \mathcal{S}, \mathbf{t} \in \mathcal{T}\}$ is convex.
- If $\mathcal{S}$ and $\mathcal{T}$ are convex sets, then $\mathcal{S} \cap \mathcal{T}$ is convex.

Convex combination (凸组合)

Convex combination of $\mathbf{x}_1, \ldots, \mathbf{x}_k$: any point $\mathbf{x}$ of the form

$$\mathbf{x} = \theta_1 \mathbf{x}_1 + \theta_2 \mathbf{x}_2 + \cdots + \theta_k \mathbf{x}_k$$

with $\theta_1 + \cdots + \theta_k = 1$, $\theta_i \geq 0$.

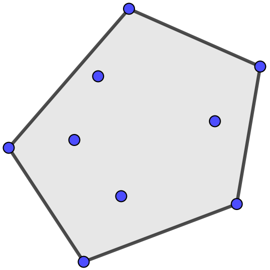
Theorem: If $\mathbf{x}_1, \ldots, \mathbf{x}_k$ belong to a convex set $\mathcal{S}$, then their convex combination $\mathbf{x}$ also belongs to $\mathcal{S}$.

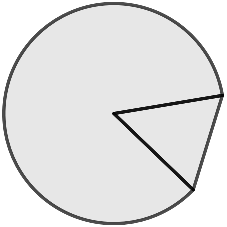
Convex hull (凸包)

Convex hull $\mathrm{conv}\mathcal{S}$: set of all convex combinations of points in $\mathcal{S}$.

$$\mathrm{conv}\mathcal{S} = \{\theta_1 \mathbf{x}_1 + \cdots + \theta_k \mathbf{x}_k | \mathbf{x}_i \in \mathcal{S}, \theta_i \geq 0, i = 1, \ldots, k, \theta_1 + \cdots + \theta_k = 1\}.$$

Example: convex hull of $\{0, 1\}$ is $[0, 1]$.





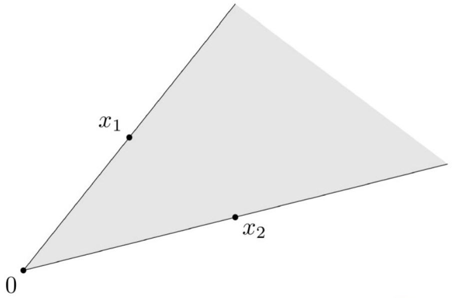
Affine sets (仿射集)

A set is called **affine set** if it contains the line through any two distinct points in the set.

Example: solution set of linear equations $\{\mathbf{x} | \mathbf{A}\mathbf{x} = \mathbf{b}\}$.

Cones (锥)

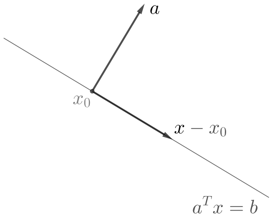
- A set $\mathcal{C}$ is called a **cone** if for every $\mathbf{x} \in \mathcal{C}$ and $\theta > 0$ we have $\theta\mathbf{x} \in \mathcal{C}$.
- A set $\mathcal{C}$ is called a **convex cone** if it is convex and a cone, which means that for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{C}$ and $\theta_1, \theta_2 > 0$, we have $\theta_1\mathbf{x}_1 + \theta_2\mathbf{x}_2 \in \mathcal{C}$.

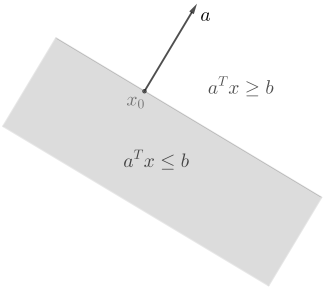


Hyperplanes and halfspaces (超平面与半平面)

Hyperplane: set of the form $\{\mathbf{x}|\mathbf{a}^\top\mathbf{x} = \mathbf{b}\}$ ($a \neq 0$).

Halfplane: set of the form $\{\mathbf{x}|\mathbf{a}^\top\mathbf{x} \leq \mathbf{b}\}$ ($a \neq 0$).

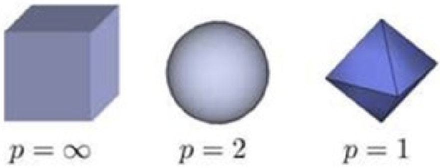




Hyperplane is affine set.

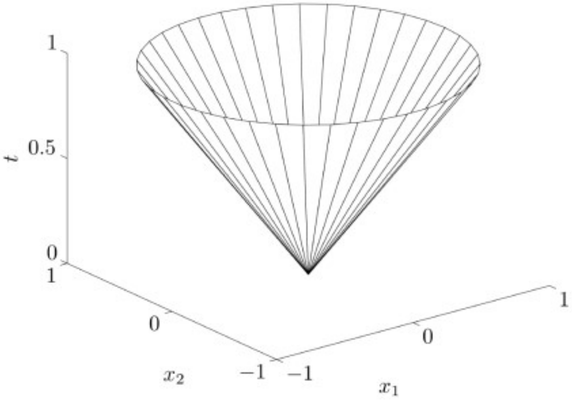
Norm balls (范数球)

Norm ball with center $\mathbf{x}_c$ and radius r: $\{\mathbf{x} \mid \|\mathbf{x} - \mathbf{x}_c\| \leq r\}$.



Norm cones (范数锥)

Norm cone: $\{(\mathbf{x}, t) | \|\mathbf{x}\| \leq t\}$.



Operations that preserve convexity (保凸运算)

Affine functions (仿射函数).

Suppose $\mathcal{S}$ is convex and $f : \mathbb{R}^n \to \mathbb{R}^m$ is an affine function:

$$f(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}.$$

Then the image of $\mathcal{S}$ under f:

$$f(\mathcal{S}) = \{f(\mathbf{x}) | \mathbf{x} \in \mathcal{S}\}$$

is convex. The inverse image:

$$f^{-1}(\mathcal{S}) = \{\mathbf{x} \in \mathbb{R}^n | f(\mathbf{x}) \in \mathcal{S}\}$$

is convex.

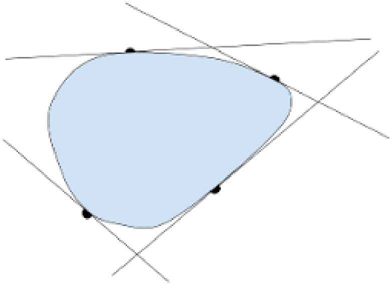
Operations that preserve convexity (保凸运算)

Intersection (取交集).

The intersection of (any number of) convex sets is convex, i.e., if $\mathcal{S}_\alpha$ is convex for any $\alpha \in \mathcal{A}$, then $\cap_{\alpha\in\mathcal{A}}\mathcal{S}_\alpha$ is convex.

Example: A closed convex set $\mathcal{S}$ is the intersection of all halfspaces contain it:

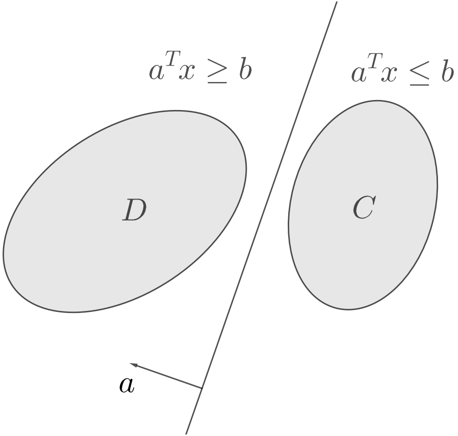
$$\mathcal{S} = \bigcap\{\mathcal{H} | \mathcal{H} \text{ is halfspace}, \mathcal{S} \subseteq \mathcal{H}\}$$



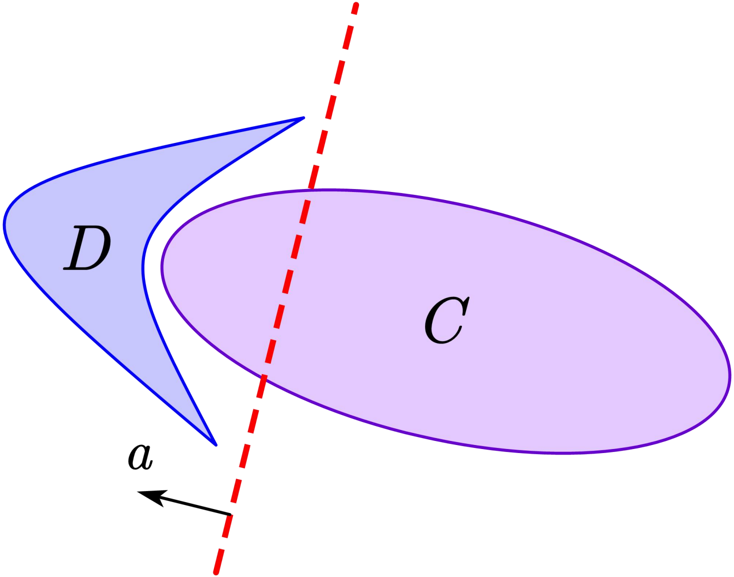
Hyperplane separation theorem

If $\mathcal{C}$ and $\mathcal{D}$ are nonempty disjoint convex sets, there exists $\mathbf{a} \neq 0$ and b s.t.

$$\mathbf{a}^\top \mathbf{x} \leq b \text{ for } \mathbf{x} \in \mathcal{C}, \quad \mathbf{a}^\top \mathbf{x} \geq b \text{ for } \mathbf{x} \in \mathcal{D}.$$



Hyperplane Separation Theorem



Strict separation theorem

Suppose $\mathcal{C}$ and $\mathcal{D}$ are nonempty disjoint convex sets. If $\mathcal{C}$ is closed and $\mathcal{D}$ is compact, there exists $\mathbf{a} \neq 0$ and b s.t.

$$\mathbf{a}^\top \mathbf{x} < b \text{ for } \mathbf{x} \in \mathcal{C}, \quad \mathbf{a}^\top \mathbf{x} > b \text{ for } \mathbf{x} \in \mathcal{D}.$$

Example: a point and a closed convex set.

Why we must restrict both sets $\mathcal{C}$ and $\mathcal{D}$ to be closed and one of them to be bounded?

- If both $\mathcal{C}$ and $\mathcal{D}$ are closed and unbounded:

$$\mathcal{C} = \left\{ (x, y) \;\middle|\; y \geq \frac{1}{x}, x > 0 \right\}, \quad \mathcal{D} = \{(x, y) \mid y \leq 0\}.$$

- If $\mathcal{C}$ is open and $\mathcal{D}$ is compact:

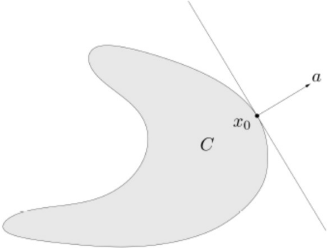
$$\mathcal{C} = \{(x, y) \mid x \in (0, 1)\}, \quad \mathcal{D} = \{x, y) \mid y \in [1, 2]\}.$$

Supporting hyperplane theorem

Supporting hyperplane to set $\mathcal{C}$ at boundary point $\mathbf{x}_0$:

$$\{\mathbf{a}^\top \mathbf{x} = \mathbf{a}^\top \mathbf{x}_0\}$$

where $\mathbf{a} \neq 0$ and $\mathbf{a}^\top \mathbf{x} \leq \mathbf{a}^\top \mathbf{x}_0$ for all $\mathbf{x} \in \mathcal{C}$.



Supporting hyperplane theorem: if $\mathcal{C}$ is convex, then there exists a supporting hyperplane at every boundary point of $\mathcal{C}$.

Outline

1. Convex Set
2. Convex Function

Convex Function (凸函数)

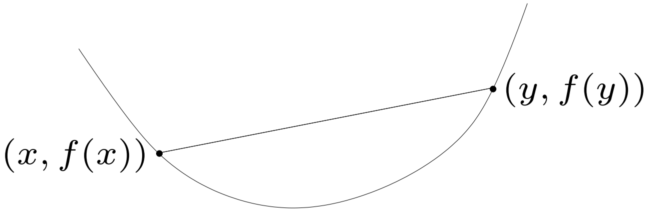
- A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if $\mathrm{dom}\ f$ is a convex set and

$$f(\theta \mathbf{x} + (1-\theta)\mathbf{y}) \le \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y})$$

  for all $\mathbf{x}, \mathbf{y} \in \mathrm{dom}\ f$, $\theta \in [0, 1]$.
- A function f is concave if $-f$ is convex.

Strict convex function:

$$f(\theta \mathbf{x} + (1-\theta)\mathbf{y}) < \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y}),\ \ t \in (0,1),\ \ \mathbf{x} \neq \mathbf{y}$$



Examples

- exponential: e^{ax}.
- power: $x^\alpha \, (x > 0, \alpha \geq 1)$.
- logarithm: $\log_a x \, (0 < a < 1)$.
- negative entropy: $x \log x$
- affine: $\mathbf{a}^\top \mathbf{x} + b$.
- norms: $\|\mathbf{x}\|$.

First-order condition

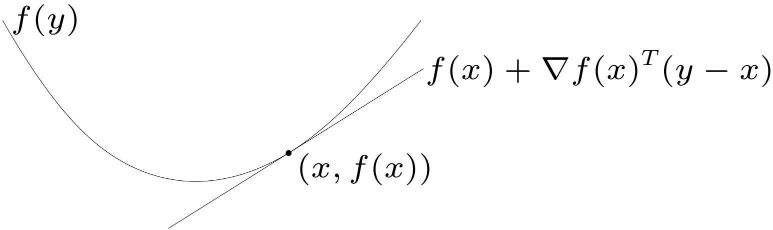
Suppose f is differentiable and has convex domain, then f is convex if and only if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$$

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

Strict convex:

$$f(\mathbf{y}) \ > \ f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle, \ \text{if } \mathbf{y} \neq \mathbf{x}.$$



First-order condition

Suppose f is differentiable and has convex domain, then f is convex if and only if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$$

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

Proof. (Part 1) If f is convex, then for all $\mathbf{x}, \mathbf{y} \in \mathrm{dom}\ f$ we have

$$\theta f(\mathbf{y}) + (1-\theta) f(\mathbf{x}) \geq f(\theta \mathbf{y} + (1-\theta)\mathbf{x})$$

Let $\theta \to 0^+$, we obtain

$$f(\mathbf{y}) - f(\mathbf{x}) \geq \lim_{\theta \to 0^+} \frac{f(\mathbf{x} + \theta(\mathbf{y} - \mathbf{x})) - f(\mathbf{x})}{\theta} = \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle,$$

where the last inequality uses the following lemma:

Lemma. For any vector $\mathbf{h}$, we have $\langle \nabla f(\mathbf{x}), \mathbf{h} \rangle = \lim_{t \to 0^+} \frac{f(\mathbf{x}+t\mathbf{h}) - f(\mathbf{x})}{t}$.

First-order condition

Suppose f is differentiable and has convex domain, then f is convex if and only if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$$

holds for all $\mathbf{x}, \mathbf{y} \in \mathrm{dom}\ f$.
Proof. (Part 2) Assume $f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$ all $\mathbf{x}, \mathbf{y} \in \mathrm{dom}\ f$.
Let $\mathbf{z} = \theta \mathbf{x} + (1-\theta)\mathbf{y}$. We have

$$\begin{aligned} f(\mathbf{x}) &\geq f(\mathbf{z}) + \langle \nabla f(\mathbf{z}), \mathbf{x} - \mathbf{z} \rangle, \\ f(\mathbf{y}) &\geq f(\mathbf{z}) + \langle \nabla f(\mathbf{z}), \mathbf{y} - \mathbf{z} \rangle. \end{aligned}$$

Combine them together, we get

$$\theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y}) \geq f(\mathbf{z}) + \langle \nabla f(\mathbf{z}), \theta \mathbf{x} + (1-\theta)\mathbf{y} - \mathbf{z} \rangle = f(\mathbf{z}).$$

First-order condition

Theorem. Assume $f(\mathbf{x})$ is convex. If $\nabla f(\mathbf{x}) = 0$, then for all $\mathbf{y} \in \mathrm{dom}\ f$, $f(\mathbf{y}) \geq f(\mathbf{x})$, i.e., $\mathbf{x}$ is a global minimizer of f.

Second-order condition

Suppose f is twice differentiable and has convex domain, then f is convex if and only if

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

Strict convex:

$$\nabla^2 f(\mathbf{x}) \succ \mathbf{0}.$$

Examples

- least-square: $f(\mathbf{x}) = \|\mathbf{Ax} - \mathbf{b}\|_2^2$
- quadratic-over-linear: $f(x, y) = x^2/y$, $y > 0$
- log-sum-exp: $f(\mathbf{x}) = \log \sum_{i=1}^{n} \exp(x_i)$

Sublevel set（水平子集）

The α-sublevel set of a function f is defined as

$$\mathcal{C}_\alpha = \{\mathbf{x} \in \mathrm{dom}\ f | f(\mathbf{x}) \leq \alpha\}$$

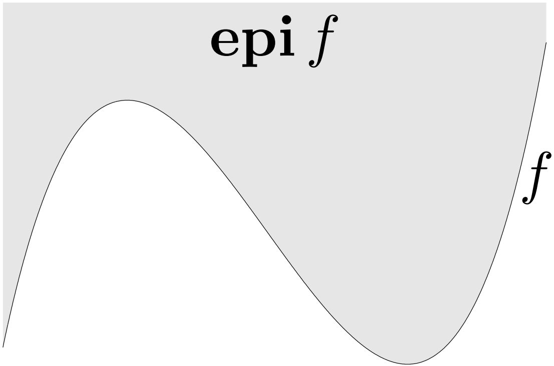
Sublevel sets of convex functions are convex for any value α.

The converse is not true: a function can have all its sublevel sets convex, but not be a convex function.

Epigraph (上方图)

The epigraph of a function $f : \mathcal{S} \to \mathbb{R}$ is defined as the set

$$\operatorname{epi} f \triangleq \{(\mathbf{x}, u) \in \mathcal{S} \times \mathbb{R} : f(\mathbf{x}) \leq u\}.$$



Epigraph (上方图)

Theorem. A function f is convex if and only if its epigraph is a convex set.

$\Longrightarrow$: Suppose $f : \mathcal{C} \to \mathbb{R}$ is convex. Let $(\mathbf{x}_1, u_1)$ and $(\mathbf{x}_2, u_2)$ be two points in the epigraph. For any $\alpha \in [0, 1]$, the point $\alpha(\mathbf{x}_1, u_1) + (1-\alpha)(\mathbf{x}_2, u_2)$ satisfies

$$f(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2) \leq \alpha u_1 + (1-\alpha) u_2,$$

Hence, the point $\alpha(\mathbf{x}_1, u_1) + (1-\alpha)(\mathbf{x}_2, u_2)$ is in the epigraph, which means the epigraph is convex.

$\Longleftarrow$: Suppose the epigraph is convex. It is easy to see $\mathcal{C}$ is convex by fixing some u. Let $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{C}$, $u_1 = f(\mathbf{x}_1)$ and $u_2 = f(\mathbf{x}_2)$. The convexity of epigraph means

$$\alpha(\mathbf{x}_1, u_1) + (1-\alpha)(\mathbf{x}_2, u_2) = (\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2, \alpha u_1 + (1-\alpha) u_2) \in \operatorname{epi} f,$$

which leads to $f(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha u_1 + (1-\alpha) u_2 = \alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2)$.

Jensen inequality

Jensen Inequality:

$$f(\theta_1 \mathbf{x}_1 + \cdots + \theta_k \mathbf{x}_k) \leq \theta_1 f(\mathbf{x}_1) + \cdots + \theta_k f(\mathbf{x}_k), \ \theta_1 + \cdots + \theta_k = 1, \theta_i \geq 0$$

can be proved by induction

Extensions:

$$f\left(\int_S p(\mathbf{x}) \mathrm{d}\,\mathbf{x}\right) \leq \int_S f(\mathbf{x}) p(\mathbf{x}) \mathrm{d}\,\mathbf{x}$$

$$f(\mathbb{E}[\mathbf{x}]) \leq \mathbb{E}[f(\mathbf{x})], \ \text{for any random variable } \mathbf{x}$$

Operations that preserve convexity

Nonnegative weighted sums:
A nonnegative weighted sum of convex functions

$$f = w_1 f_1 + \cdots + w_m f_m$$

is convex.

Composition with affine function:
If f is convex, then $f(\mathbf{Ax} + \mathbf{b})$ is convex.

Operations that preserve convexity

Pointwise maximum:
If $f_1, \ldots, f_m$ are convex, then $f(x) = \max\{f_1(x), \ldots, f_m(x)\}$ is convex.

Example:

- piecewise-linear function: $f(x) = \max_{i=1,\ldots,m}(\mathbf{a}_i^\top \mathbf{x} + \mathbf{b}_i)$ is convex
- sum of r largest components of $\mathbf{x} \in \mathbb{R}^n$:

$$f(\mathbf{x}) = x_{[1]} + \cdots + x_{[r]}$$

  is convex. ($\mathbf{x}_{[i]}$ is i-th largest component of $\mathbf{x}$)

Operations that preserve convexity

Pointwise supremum:
If $f(x, y)$ is convex in x for each $y \in \mathcal{A}$, then

$$g(x) = \sup_{y \in \mathcal{A}} f(x, y)$$

is convex.

Example:

- distance to farthest point in a set $\mathcal{C}$:

$$f(\mathbf{x}) = \sup_{\mathbf{y} \in \mathcal{C}} \|\mathbf{x} - \mathbf{y}\|$$

Operations that preserve convexity

Minimization:
If $f(x, y)$ is convex in (x, y) and $\mathcal{C}$ is a convex set, then

$$g(x) = \inf_{y \in \mathcal{C}} f(x, y)$$

is convex.

Example: distance to a set: $\mathrm{dist}(\mathbf{x}, \mathcal{S}) = \inf_{\mathbf{y} \in \mathcal{S}} \|\mathbf{x} - \mathbf{y}\|$ is convex if $\mathcal{S}$ is convex.

Convex optimization

Theorem. Let f be a convex function on a convex set $\mathcal{C}$. Suppose $\mathbf{x}^*$ is a local minima of f, i.e., there exist some $\delta > 0$ such that any $\bar{\mathbf{x}} \in \mathcal{B}_\delta \cap \mathcal{C}$ holds $f(\mathbf{x}^*) \leq f(\bar{\mathbf{x}})$. Then $\mathbf{x}^*$ is a global solution of

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