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

1. Convex Function

2. Subgradient

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

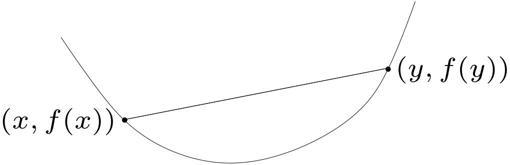
$$f(\theta \mathbf{x} + (1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y})$$

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

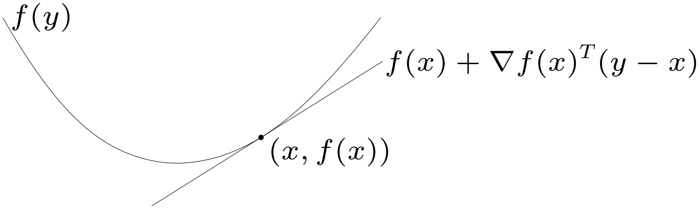
- exponential: e^{ax}.
- power: $x^{\alpha}\,(x > 0, \alpha \geq 1)$.
- logarithm: $\log_a x\,(0 < a < 1)$.
- negative entropy: $x \log x$
- affine: $\mathbf{a}^\top \mathbf{x} + b$.
- norms: $\|\mathbf{x}\|$.

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



First-order condition

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.

Strict convex:

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

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: $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$
- 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

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

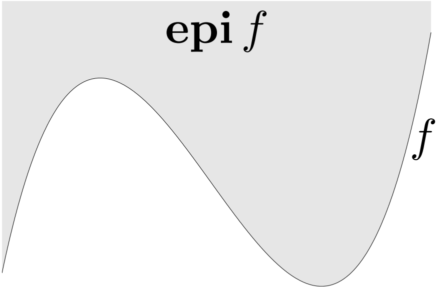
Sublevel sets of convex functions are convex for any value α.

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

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



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

Jensen inequality

Jensen Inequality:

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

can be proved by induction

Extensions:

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

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

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: $\text{dist}(\mathbf{x}, \mathcal{S}) = \inf_{\mathbf{y} \in \mathcal{S}} \|\mathbf{x} - \mathbf{y}\|$ is convex if $\mathcal{S}$ is convex.

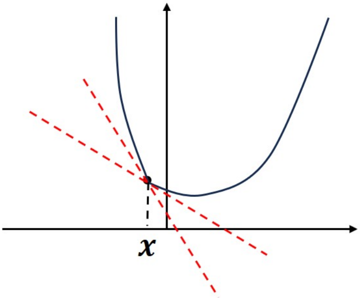
Convex optimization

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 $f(\mathbf{x}^*) \le f(\bar{\mathbf{x}})$. Then $\mathbf{x}^*$ is a global solution of

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

Outline

Subgradient (次梯度)

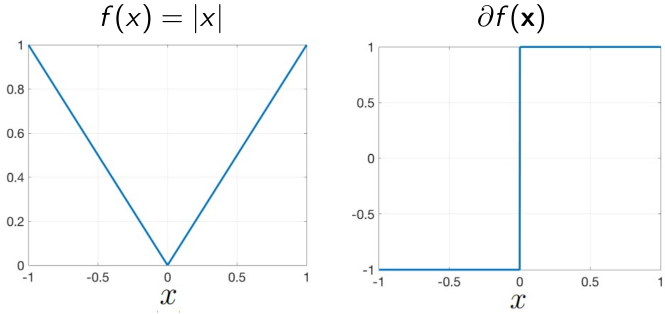


We say **g** is a subgradient of f at the point **x** if

$$f(\mathbf{y}) \geq \underbrace{f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle}_{\text{a linear under-estimate of } f}, \qquad \forall \mathbf{y} \in \operatorname{dom} f$$

The set of all subgradients of f at **x** is called the subdifferential of f at **x**, denoted by $\partial f(\mathbf{x})$.

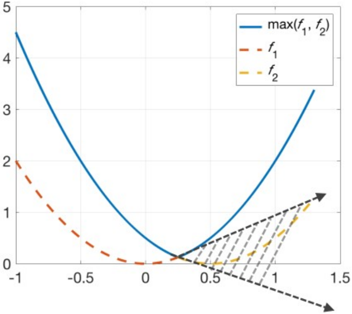
Example: $f(x) = |x|$



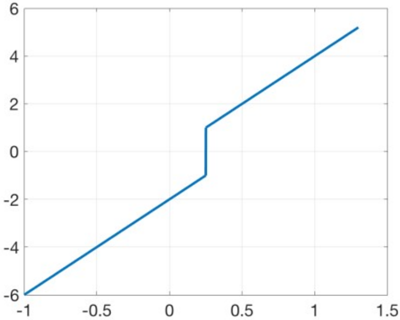
$$f(x) = |x| \qquad \partial f(\mathbf{x}) = \begin{cases} \{-1\}, & \text{if } x < 0 \\ [-1, 1], & \text{if } x = 0 \\ \{1\}, & \text{if } x > 0 \end{cases}$$

Example: $\max\{f_1(x), f_2(x)\}$

$\max\{f_1(x), f_2(x)\}$



$\partial f(\mathbf{x})$



$f(x) = \max\{f_1(x), f_2(x)\}$ where $f_1(x)$ and $f_2(x)$ are differentiable.

$$\partial f(\mathbf{x}) = \begin{cases} \{f_1'(x)\}, & \text{if } f_1(x) > f_2(x) \\ [f_1'(x), f_2'(x)], & \text{if } f_1(x) = f_2(x) \\ \{f_2'(x)\}, & \text{if } f_1(x) < f_2(x) \end{cases}$$

Subgradient of differentiable functions

If a function is differentiable, the only subgradient at each point is the gradient, i.e.,

$$\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}.$$

Basic rules of subgradient

- **scaling:** $\partial(\alpha f) = \alpha \partial f$, for $\alpha > 0$
- **summation:** $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$

Example: Compute the subdifferential of ℓ_1 norm

$$f(\mathbf{x}) = \|\mathbf{x}\|_1 = \sum_{i=1}^{d} |x_i|.$$

Basic rules of subgradient (cont.)

- **chain rule:** suppose f is convex, and g is differentiable, nondecreasing, and convex. Let $h(\mathbf{x}) = g(f(\mathbf{x}))$, then

$$\partial h(\mathbf{x}) = g'(f(\mathbf{x}))\partial f(\mathbf{x})$$

- Suppose f is convex, and let $h(\mathbf{x}) = f(\mathbf{A}\mathbf{x} + \mathbf{b})$. Then

$$\partial h(\mathbf{x}) = \mathbf{A}^\top \partial f(\mathbf{A}\mathbf{x} + \mathbf{b})$$

Example: Find a subgradient of $\|\mathbf{A}\mathbf{x} + \mathbf{b}\|_1$.

Basic rules of subgradient (cont.)

- **pointwise maximum:** if $f(\mathbf{x}) = \max_{1\leq i\leq k} f_i(\mathbf{x})$, then

$$\partial f(\mathbf{x}) = \mathrm{conv}\left\{\bigcup\{\partial f_i(\mathbf{x}) | f_i(\mathbf{x}) = f(\mathbf{x})\}\right\}$$

- **pointwise supremum:** if $f(\mathbf{x}) = \sup_{\alpha\in\mathcal{F}} f_\alpha(\mathbf{x})$, then

$$\partial f(\mathbf{x}) = \mathrm{closure}\left(\mathrm{conv}\left\{\bigcup\{\partial f_\alpha(\mathbf{x}) | f_\alpha(\mathbf{x}) = f(\mathbf{x})\}\right\}\right)$$

Example: Find subgradients of following functions:

$$f(\mathbf{x}) = \max_{1\leq i\leq k}\{\mathbf{a}_i^\top \mathbf{x} + b_i\}$$

$$f(\mathbf{x}) = \|\mathbf{x}\|_\infty = \max_{1\leq i\leq d} |x_i|$$

Subgradient characterization of convexity

A function f is convex if and only if $\mathrm{dom}\ f$ is convex and $\partial f(\mathbf{x}) \neq \emptyset$ for all $\mathbf{x} \in (\mathrm{dom}\ f)^\circ$.

Summary

- **convex function**
  - definition
  - first-order condition, second-order condition
  - sublevel set, epigraph
  - Jensen inequality
  - operations that preserve convexity
- **subgradient**
  - definition
  - basic properties