Optimization in Machine Learning: Lecture 3

Convex Functions

by Xiaolin Huang xiaolinhuang@sjtu.edu.cn SEIEE 2-429

Institute of Image Processing and Pattern Recognition

http://www.pami.sjtu.edu.cn/

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

目录 Contents

Convex Set

- convex optimization is to minimize a convex function over a **convex set**
    - convex combination
    - convex sets
    - operations that preserve convexity
    - separating hyperplane
    - supporting hyperplane

Convex Function

- convex optimization is to minimize a **convex function** over a **convex set**
  - convex combination
  - convex sets
  - operations that preserve convexity
  - separating hyperplane
  - supporting hyperplane

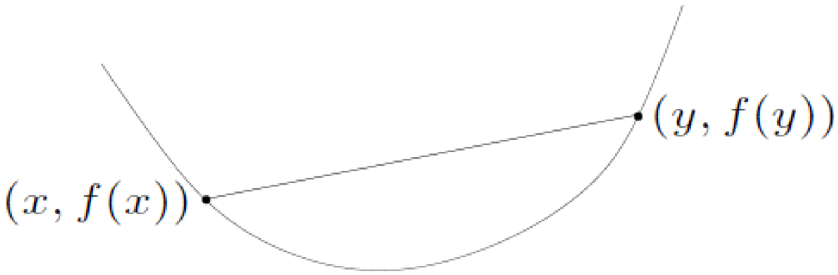
目录 Contents

Definition

- $f: R^n \to R$ is convex, if **dom** f is a convex set and

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y) \quad \forall x, y \in \textbf{dom}\, f, \theta \in [0,1]$$

- **"line is above the curve":**



- **strictly convex** if strict inequality holds for $x \neq y$ and $\theta \in (0,1)$
- concave function: f is concave iff $-f$ is convex.

Convex Function and Convex Sets

- $f: R^n \to R$ is convex, if **dom** f is a convex set and

$$f(\theta x + (1-\theta)y) \le \theta f(x) + (1-\theta)f(y) \quad \forall x, y \in \mathbf{dom}\, f, \theta \in [0,1]$$

- if $f(x)$ is convex, then $\{x: f(x) \le 0\}$ is a convex set

- if C is a convex set, then it could be represented as the solution of a system of (maybe infinite number of) convex inequalities

supporting hyperplanes

Equivalent Conditions

- $f: \mathbf{R}^n \to \mathbf{R}$ is convex, if $\mathbf{dom}\, f$ is a convex set and

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y) \quad \forall x, y \in \mathbf{dom}\, f, \theta \in [0,1]$$

Jensen's Inequality

- extension to multiple points

$$f\left(\sum_{i=1}^{m} \theta_i x_i\right) \leq \sum_{i=1}^{m} \theta_i f(x_i), \forall x_i \in \mathbf{dom}\, f, \theta_i \geq 0, \sum_{i=1}^{m} \theta_i = 1.$$

- **restriction on R:**

$g(t) = f(x+tv), \mathbf{dom}\, g = \{t: x+tv \in \mathbf{dom}\, f\}$ is convex for $\forall x \in \mathbf{dom}\, f, v \in R^n$

Equivalent Conditions

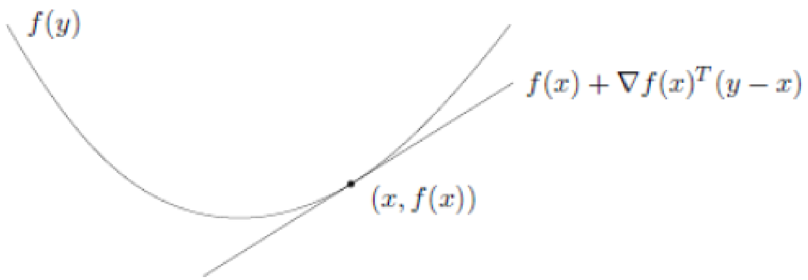
- **First-order condition**

$$f(y) \geq f(x) + \nabla f(x)^{\top}(y-x) \quad \forall x, y \in \textbf{dom}\, f$$

"tangent plane is below the surface"

proof. consider univariate functions

given by the gradient.



$\textbf{dom}\, f$ is open

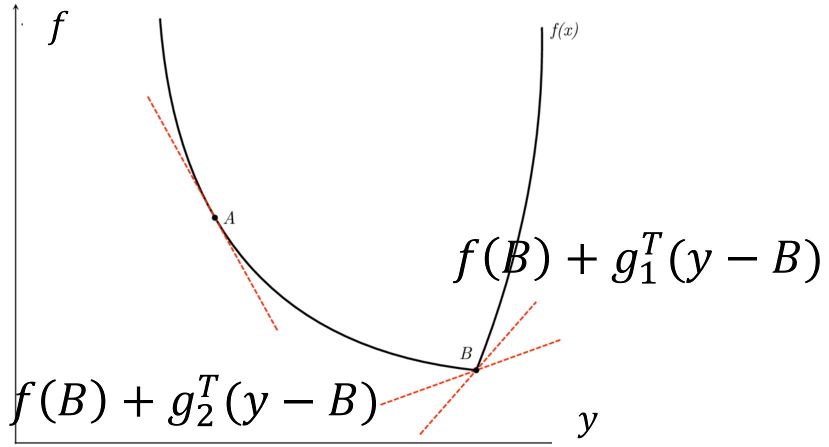
Gradient and Sub-Gradient

- **First-order condition**

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

- consider a convex set $\{x: f(x) \leq 0\}$
- $\nabla f(x)$ gives a supporting hyperplane
- how about non-smooth function?
  - there are multiple hyperplanes
  - the set containing the vectors of all supporting hyperplanes is called sub-gradient $\partial f(x)$

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

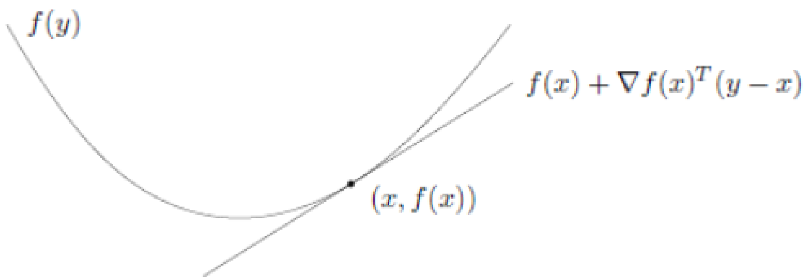


Equivalent Conditions

- **First-order condition**

$$f(y) \geq f(x) + \nabla f(x)^{\mathsf{T}}(y-x) \quad \forall x, y \in \mathbf{dom}\, f$$

"tangent plane is below the surface"



proof. consider univariate functions given by the gradient.

- **Second-order condition**

$\nabla^2 f(x) \geq 0, \forall x \in \mathbf{dom}\, f \iff f$ is convex

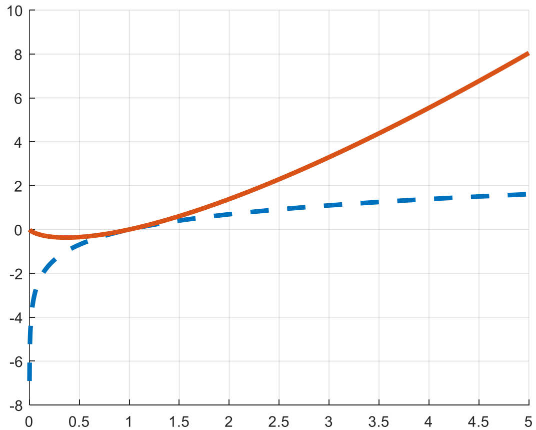
$\nabla^2 f(x) > 0, \forall x \in \mathbf{dom}\, f \implies f$ is strictly convex

Examples

- affine functions $f(x) = a^{\top}x + b$
- exponential $f(x) = e^{ax}$
- powers $f(x) = x^{\alpha}$, with $x > 0, \alpha > 1$ or $\alpha \leq 0$
- negative entropy $f(x) = x\log x$ with $x > 0$
- (minus powers) $f(x) = x^{\alpha}$, $0 \leq \alpha \leq 1$
- (minus logarithm) $f(x) = \log x$ with $x > 0$

Examples

- affine functions $f(x) = a^{\top}x + b$
- exponential $f(x) = e^{ax}$
- powers $f(x) = x^{\alpha}$, with $x > 0, \alpha > 1$ or $\alpha \leq 0$
- negative entropy $f(x) = x\log x$ with $x > 0$
- (minus powers) $f(x) = x^{\alpha}, 0 \leq \alpha \leq 1$
- (minus logarithm) $f(x) = \log x$ with $x > 0$



Examples

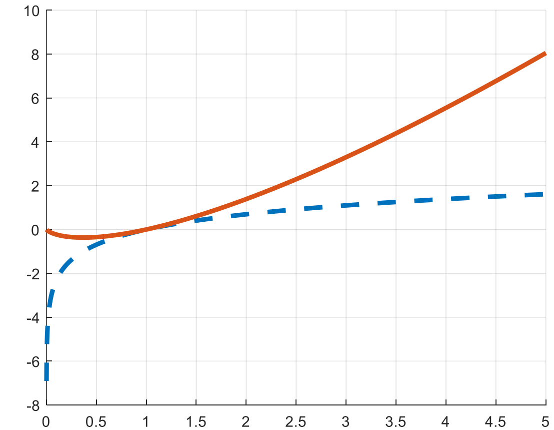
- affine functions $f(x) = a^\top x + b$
- exponential $f(x) = e^{ax}$
- powers $f(x) = x^\alpha$, with $x > 0, \alpha > 1$ or $\alpha \leq 0$
- negative entropy $f(x) = x\log x$ with $x > 0$
- (minus powers) $f(x) = x^\alpha$, $0 \leq \alpha \leq 1$
- (minus logarithm) $f(x) = \log x$ with $x > 0$

$(x\log x)' = \log x + 1$ $\qquad$ $(x\log x)'' = 1/x > 0$

$(\log x)' = 1/x$ $\qquad$ $(\log x)'' = -1/x^2 < 0$

logarithm functions are widely used in machine learning

- for independent data, the joint probability will be in the form of multiplication, which could be transformed to sum by log
- entropy, a good measure for information/uncertainty, is in the form of logarithm



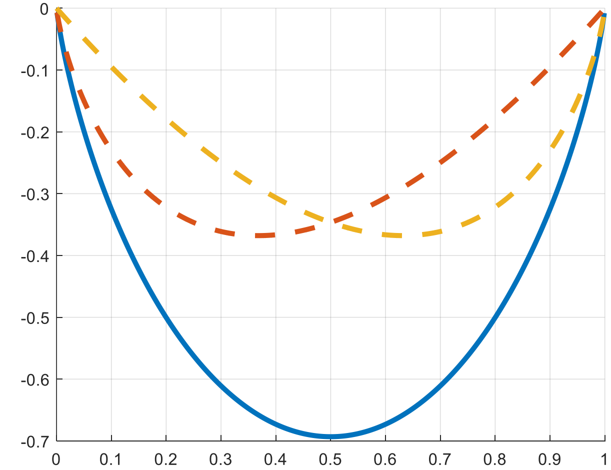
Measure of Uncertainty

- the cancel of uncertainty is **information**
  - case 1: $x \in \{0,1,2,3\}$, $x_1 = 1$
  - case 2: $x \in \{0,1\}$, $x_1 = 1$
  - the information of case 1 is larger
- quantitative measurement
  - no uncertainty, equal zero
  - proportional to uncertainty
  - consider case 2: $p \triangleq p(x=0), 1-p = p(x=1)$

$$\text{negative entropy} = p\log p + (1-p)\log(1-p)$$

logarithm functions are widely used in machine learning

- for independent data, the joint probability will be in the form of multiplication, which could be transformed to sum by log
- entropy, a good measure for information/uncertainty, is in the form of logarithm



Sublevel Set and Epigraph

- **sublevel set** of $f: \mathrm{R}^n \to \mathrm{R}$ $\quad C_\alpha = \{x \in \textbf{dom}\, f : f(x) \leq \alpha\}$
- sublevel sets of convex functions are convex

Sublevel Set and Epigraph

- **sublevel set** of $f: \mathrm{R}^n \to \mathrm{R}$ $\quad C_\alpha = \{x \in \textbf{dom}\, f : f(x) \le \alpha\}$
- sublevel sets of convex functions are convex
- **epigraph** of $f: \mathrm{R}^n \to \mathrm{R}$

$$\textbf{epi}\, f = \{(x, t) \in \mathbf{R}^{n+1}\ x \in \textbf{dom}\, f : f(x) \le t\}$$

- f is convex iff **epi** f is a convex set
  - from first order condition $f(y) \ge f(x) + \nabla f(x)^\top (y - x)$
  - for $(y, t) \in \textbf{epi}\, f$ $\quad t \ge f(y) \ge f(x) + \nabla f(x)^\top (y - x)$

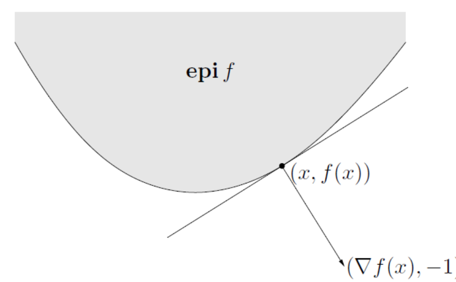
Sublevel Set and Epigraph

- **sublevel set** of $f\colon \mathrm{R}^n \to \mathrm{R}$ $\quad C_\alpha = \{x \in \mathbf{dom}\, f : f(x) \le \alpha\}$
- sublevel sets of convex functions are convex
- **epigraph** of $f\colon \mathrm{R}^n \to \mathrm{R}$

$$\mathbf{epi}\, f = \{(x,t) \in \mathbf{R}^{n+1}\ x \in \mathbf{dom}\, f : f(x) \le t\}$$

- f is convex iff **epi** f is a convex set
  - from first order condition $f(y) \ge f(x) + \nabla f(x)^\top (y - x)$
  - for $(y,t) \in \mathbf{epi}\, f$ $\qquad t \ge f(y) \ge f(x) + \nabla f(x)^\top (y - x)$

$$\begin{bmatrix} \nabla f(x) \\ -1 \end{bmatrix}^{-1} \left(\begin{bmatrix} y \\ t \end{bmatrix} - \begin{bmatrix} x \\ f(x) \end{bmatrix} \right) \le 0$$ supporting hyperplane



目录 Contents

Nonnegative Weighted Sum and Composition

- nonnegative weighted sum $f = w_1 f_1 + \cdots + w_m f_m, w_m \geq 0$
  - nonnegative multiple: αf is convex with $\alpha \geq 0$
  - integral $\int w(y) f(x, y) dy$
  - the set of convex function is a conic
- composition with affine function $f(Ax + b)$
  - residual $r_i = b_i - a_i^\top x$
  - measuring residual by convex loss is convex
    - minus log (MLE) $-\Sigma_{i=1}^{m} \log(b_i - a_i^\top x)$
    - norm $\|Ax - b\|_2^2$

Nonnegative Weighted Sum and Composition

- nonnegative weighted sum $f = w_1 f_1 + \cdots + w_m f_m, w_m \geq 0$
  - nonnegative multiple: αf is convex with $\alpha \geq 0$
  - integral $\int w(y) f(x, y) dy$
  - the set of convex function is a conic
- composition with affine function $f(Ax + b)$
  - residual $r_i = b_i - a_i^\top x$
  - measuring residual by convex loss is convex
    - minus log (MLE) $-\Sigma_{i=1}^{m} \log(b_i - a_i^\top x)$
    - norm $||Ax - b||_2^2$

$r_i = b_i - a_i^\top \phi(x)$

composition of $g: \mathrm{R}^n \to \mathrm{R}$ and : $h: \mathrm{R} \to \mathrm{R}$

$f(x) = h(g(x))$ is convex, if

- g convex, h convex and non-decreasing
- g concave, h convex and non-increasing

Pointwise Maximum/Supremum

- $f(x) = \max\{f_1(x), \dots, f_m(x)\}$ is convex
  - piecewise-linear function: $f(x) = \max_i\{a_i^\top x + b_i\}$
  - largest error: $f(x) = \max_i\left\{\left\|a_i^\top x - b_i\right\|_2^2\right\}$
  - sum of r largest errors
  - sum of r largest components: $x_{[r_1]} \geq x_{[r_2]}$ if $r_1 \leq r_2$

$$f(x) = x_{[1]} + x_{[2]} + \cdots x_{[r]}$$

- robust learning
- risk control
- robust control

- $f(x, y)$ is convex in x for each $y \in Y$, then $\sup_y f(x, y)$ is convex
  - distance to farthest point in a set $f(x) = \sup_{y \in C} \|x - y\|$
  - maximum eigenvalue $\lambda_{\max}(X) = \sup_{\|y\|_2 = 1} y^\top X y$

Minimization and Schur Complement

- if $f(x, y)$ is convex w.r.t. (x, y) and C is a convex set, then
  the minimization $g(x) = \inf_{y \in C} f(x, y)$ is convex
- distance to a convex set S

$$\mathrm{dist}(x, S) \triangleq g(x) = \inf_{y \in S} \|x - y\|$$

difference to

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

- Shur complement
  - $f(x, y) = x^\top A x + 2x^\top B y + y^\top C y$ with $\begin{bmatrix} A & B \\ B^\top & C \end{bmatrix} \geq 0, C > 0$

$C^\uparrow = C^\dagger$

  - minimizing over y, i.e., $g(x) = \inf_{y} f(x, y) = x^\top (A - BC^\uparrow B^\top) x$, is convex
  - Shur complement $A - BC^\uparrow B^\top \geq 0$

widely used in matrix inverse, and related tasks, e.g., control, SLAM, etc.

Perspective Function and KL Divergence

- for a function $f(x): R^n \to R$, its perspective function is defined as

$$g(x,t) = t\, f(x/t)$$

- if $f(x)$ is convex (concave), so is $g(x,t)$

proof:
- epigraph
- perspective function preserves convexity

- relative entropy

$$\sum_{i=1}^{n} (u_i \log(u_i/v_i))$$

  - $f(x) = -\log x$ is convex and its perspective function is

$$g(x,t) = tf(x/t) = -t \log\frac{x}{t} = t \log t - t \log x$$

- Kullback-Leibler divergence

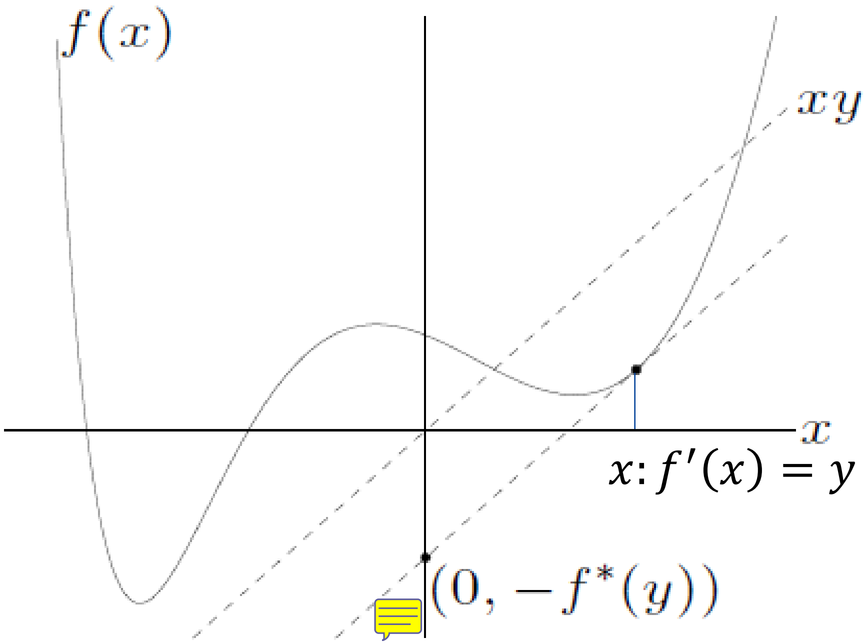
$$D_{kl}(u,v) = \sum_{i}^{n} (u_i \log(u_i/v_i) - u_i + v_i)$$

目录 Contents

Definition and Understanding

- the **conjugate** of a function $f: \mathrm{R}^n \to \mathrm{R}$ is

$$f^*(y) = \sup_{x \in \mathbf{dom}\, f} (y^\top x - f(x))$$



for a given y, $f^*(y)$ is the largest gap between the affine function xy and the function $f(x)$

when f is differentiable, the optimality condition tells

$$\nabla(y^\top x - f(x)) = y - f'(x) = 0$$

Definition and Understanding

- the **conjugate** of a function $f: \mathrm{R}^n \to \mathrm{R}$ is

$$f^*(y) = \sup_{x \in \mathbf{dom}\, f} (y^\top x - f(x))$$

- examples
  - affine functions:

$$f(x) = ax + b \to f^*(y) = \sup_{x \in \mathbf{dom}\, f} (x^\top (y - a) - b) = \begin{cases} -b, & y = a \\ +\infty, & y \neq a \end{cases}$$

  - exponential functions

$$f(x) = e^x \to f^*(y) = \sup_{x \in \mathbf{dom}\, f} (xy - e^x) = \begin{cases} y\log y - y, & y > 0 \\ \infty, & y < 0 \end{cases}$$

  - indicator function

$$f(x) = \begin{cases} \infty, & x \notin S \\ 0, & x \in S \end{cases} \to f^*(y) = \sup_{x \in \boldsymbol{S}} y^\top x \qquad \text{supporting of } S$$

Properties

- the **conjugate** of a function $f: \mathrm{R}^n \to \mathrm{R}$ is

$$f^*(y) = \sup_{x \in \mathbf{dom}\, f} (y^\top x - f(x))$$

- a conjugate function is convex, even when f is not.
- Fenchel inequality: $f(x) + f^*(y) \geq x^\top y$
  - example $x^\top Q x + y^\top Q^{-1} y \geq 2x^\top y$, when $Q \in S^n_{++}$
- conjugate of conjugate: when f is convex, proper, and closed, then $f^{**} = f$

Conjugate and sub-Gradient

- the **conjugate** of a function $f: \mathrm{R}^n \to \mathrm{R}$ is

$$f^*(y) = \sup_{x \in \mathbf{dom}\, f} (y^\top x - f(x))$$

- for a convex, proper, and closed function f, the following three are equal
  - (1) $f(x) + f^*(y) = x^\top y$
  - (2) $y \in \partial f(x)$
  - (3) $x \in \partial f^*(y)$

Proof: (1) → (2)

- $y^\top x - f(x) = f^*(y) = \sup_{x \in \mathbf{dom}\, f} (y^\top x - f(x))$
- $y^\top x - f(x) \geq y^\top z - f(z), \forall z \in \mathbf{dom}\, f$
- $f(z) \geq f(x) + y^\top (z - x), \forall z \in \mathbf{dom}\, f$
- $y \in \partial f(x)$

- sub-gradient by conjugation

$$\partial f(x) = \mathrm{argmax}_y \{x^\top y - f^*(y)\}$$

$$\partial f^*(y) = \mathrm{argmax}_x \{x^\top y - f(x)\}$$

Proof: (2) → (1)

- definition of sub-Gradient + Fenchel inequality

Proof: (1) ↔ (3)

- $f^{**} = f$, denote f^* as g, then $g^*(x) + g(y) = x^\top y$
- by (2), $x \in \partial g(y) = \partial f^*(y)$

目录 Contents

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Conclusion and Home Work

- convex optimization is to minimize a **convex function** over a convex set
  - **definition of convex functions**
  - **first-, second condition**
  - **operations that preserve convexity**
  - **conjugate functions**
  - **subgradient and the use of conjugate function**

- **Excise 3.1:** convexity proof

  Excise 3.20: convexity preservation practice

  Excise 3.38: understating conjugate

THANKS