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

1. Matrix Calculus

2. Convex Set

Outline

1. Matrix Calculus
2. Convex Set

Topology in Euclidean space

- A subset $\mathcal{S}$ of $\mathbb{R}^n$ is called **open**, if for every $\mathbf{x} \in \mathcal{S}$ there exists $\delta > 0$ such that the ball $\mathcal{B}_\delta(\mathbf{x}) = \{\mathbf{y} : \|\mathbf{y} - \mathbf{x}\|_2 \leq \delta\}$ is included in $\mathcal{S}$.
  Example: $\{x|a < x < b\}$, $\{\mathbf{x}|\mathbf{x} > 0\}$, $\{\mathbf{x}|\,\|\mathbf{x} - \mathbf{a}\| < 1\}$.
- A subset $\mathcal{C}$ of $\mathbb{R}^n$ is called **closed**, if its complement $\mathcal{C}^c = \mathbb{R}^n \backslash \mathcal{C}$ is open.
  Example: $\{x|a \leq x \leq b\}$, $\{\mathbf{x}|\mathbf{x} \geq 0\}$, $\{\mathbf{x}|\,\|\mathbf{x} - \mathbf{a}\| \leq 1\}$.
- A subset $\mathcal{C}$ of $\mathbb{R}^n$ is called **bounded**, if there exists $r > 0$ such that $\|\mathbf{x}\|_2 < r$ for all $\mathbf{x} \in \mathcal{C}$.
  Example: $\{x|a \leq x < b\}$, $\{\mathbf{x}|1 > \mathbf{x} \geq 0\}$, $\{\mathbf{x}|\,\|\mathbf{x} - \mathbf{a}\| < 1\}$.
- A subset $\mathcal{C}$ of $\mathbb{R}^n$ is called **compact**, if it is both bounded and closed.
  Example: $\{x|a \leq x \leq b\}$, $\{\mathbf{x}|1 \geq \mathbf{x} \geq 0\}$, $\{\mathbf{x}|\,\|\mathbf{x} - \mathbf{a}\| \leq 1\}$.

Topology in Euclidean space

1. The **interior** of $\mathcal{C} \in \mathbb{R}^n$ is defined as

$$\mathcal{C}^\circ = \{\mathbf{y} : \text{there exist } \varepsilon > 0 \text{ such that } \mathcal{B}_\varepsilon(\mathbf{y}) \subset \mathcal{C}\}$$

2. The **closure** of $\mathcal{C} \in \mathbb{R}^n$ is defined as

$$\overline{\mathcal{C}} = \mathbb{R}^n \backslash (\mathbb{R}^n \backslash \mathcal{C})^\circ.$$

3. The **boundary** of $\mathcal{C} \in \mathbb{R}^n$ is defined as $\overline{\mathcal{C}} \backslash \mathcal{C}^\circ$.

Derivative (导数)

Suppose $f : \mathbb{R}^n \to \mathbb{R}^m$ and $\mathbf{x} \in (\operatorname{dom} f)^\circ$. The derivative at $\mathbf{x}$ is

$$\mathrm{D}f(\mathbf{x}) = \begin{bmatrix} \dfrac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \dfrac{\partial f_1(\mathbf{x})}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f_m(\mathbf{x})}{\partial x_1} & \cdots & \dfrac{\partial f_m(\mathbf{x})}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m\times n}.$$

This matrix is also called Jacobian matrix.

Gradient (梯度)

When f is real-valued, i.e., $f : \mathbb{R}^n \to \mathbb{R}$, the gradient of f is:

$$\nabla f(\mathbf{x}) = \mathrm{D} f(\mathbf{x})^\top = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{n \times 1}.$$

Gradient of matrix functions

Suppose that $f : \mathbb{R}^{m\times n} \to \mathbb{R}$. Then the gradient of f with respect to $\mathbf{X}$ is

$$\nabla f(\mathbf{X}) = \frac{\partial f}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial f(\mathbf{X})}{\partial x_{11}} & \cdots & \frac{\partial f(\mathbf{X})}{\partial x_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(\mathbf{X})}{\partial x_{m1}} & \cdots & \frac{\partial f(\mathbf{X})}{\partial x_{mn}} \end{bmatrix} \in \mathbb{R}^{m\times n}.$$

Example:

$$f(\mathbf{X}) = \|\mathbf{X}\|_F^2$$

Examples

1. For $\mathbf{a}, \mathbf{x} \in \mathbb{R}^n$, we have $\dfrac{\partial \mathbf{a}^\top \mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}$.
2. For $\mathbf{A}, \mathbf{X} \in \mathbb{R}^{m \times n}$, we have $\dfrac{\partial \operatorname{tr}(\mathbf{A}^\top \mathbf{X})}{\partial \mathbf{X}} = \mathbf{A}$.
3. For $\mathbf{A} \in \mathbb{R}^{n \times n}$ and $\mathbf{x} \in \mathbb{R}^n$, we have $\dfrac{\partial \mathbf{x}^\top \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}$.

   If $\mathbf{A}$ is symmetric, we have $\dfrac{\partial \mathbf{x}^\top \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$.

We can find more results in the matrix cookbook:
https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

Chain rules

Suppose $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $\mathbf{x} \in \mathrm{dom}\ f$ and $g : \mathbb{R}^m \to \mathbb{R}^p$ is differentiable at $f(\mathbf{x}) \in (\mathrm{dom}\ g)^\circ$. Define the composition $h : \mathbb{R}^n \to \mathbb{R}^p$ by $h(\mathbf{z}) = g(f(\mathbf{z}))$. Then h is is differentiable at $\mathbf{x}$ and

$$\mathrm{D}h(\mathbf{x}) = \mathrm{D}(g(f(\mathbf{x})))\mathrm{D}(f(\mathbf{x})).$$

Examples:

- Suppose $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$ and $h(\mathbf{x}) = g(f(\mathbf{x}))$. Then

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

- Suppose $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{A} \in \mathbb{R}^{n \times p}$ and $b \in \mathbb{R}^n$. Define $h : \mathbb{R}^p \to \mathbb{R}$ as $h(\mathbf{x}) = f(\mathbf{A}\mathbf{x} + \mathbf{b})$. Then,

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

Gradient of logistic regression

What is the gradient of the following loss function?

$$f(\mathbf{x}) = \log \sum_{i=1}^{m} \exp(\mathbf{a}_i^\top \mathbf{x} + b_i) \tag{1}$$

The Hessian matrix

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth function that takes as input a matrix $\mathbf{x} \in \mathbb{R}^n$ and returns a real value. Then the Hessian matrix with respect to $\mathbf{x}$, written as $\nabla^2 f(\mathbf{x})$, which is defined as

$$\nabla^2 f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_n} \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

Taylor's expansion for multivariable function $f : \mathbb{R}^n \to \mathbb{R}$

$$f(\mathbf{x}) \approx f(\mathbf{a}) + \nabla f(\mathbf{a})^\top (\mathbf{x} - \mathbf{a}) + \frac{1}{2}(\mathbf{x} - \mathbf{a})^\top \nabla^2 f(\mathbf{a})(\mathbf{x} - \mathbf{a})$$

Chain rules for second derivative

- Suppose $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$ and $h(\mathbf{x}) = g(f(\mathbf{x}))$. Then

$$\nabla^2 h(\mathbf{x}) = g'(f(\mathbf{x}))\nabla^2 f(\mathbf{x}) + g''(f(\mathbf{x}))\nabla f(\mathbf{x})\nabla f(\mathbf{x})^\top.$$

- Suppose $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{A} \in \mathbb{R}^{n\times p}$ and $b \in \mathbb{R}^n$. Define $h : \mathbb{R}^p \to \mathbb{R}$ as $h(\mathbf{x}) = f(\mathbf{Ax} + \mathbf{b})$. Then,

$$\nabla^2 h(\mathbf{x}) = \mathbf{A}^\top \nabla^2 f(\mathbf{Ax} + \mathbf{b})\mathbf{A}.$$

Bonus homework: Compute the Hessian matrix of loss function (1).

Outline

1 Matrix Calculus

2 Convex Set

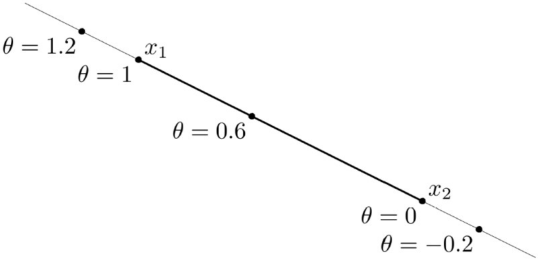
Lines and Line Segments (直线与线段)

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

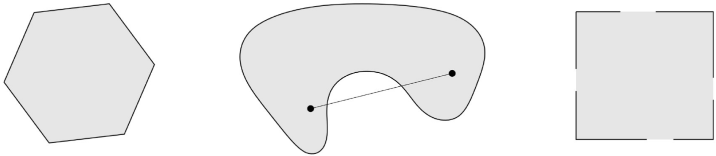
$$\mathbf{x} = \theta \mathbf{x}_1 + (1 - \theta)\mathbf{x}_2, \quad 0 \le \theta \le 1.$$



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

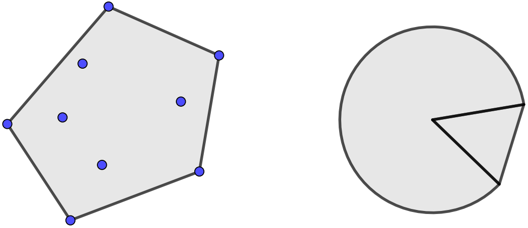
If $\mathbf{x}_1, \ldots, \mathbf{x}_k$ belong to a convex set $\mathcal{S}$, then their convex combination $\mathbf{x}$ also belongs to $\mathcal{S}$.

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)

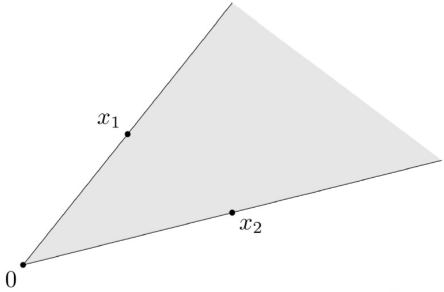
A set is called **affine set** if it contains the line through any two distinct points in the set.

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

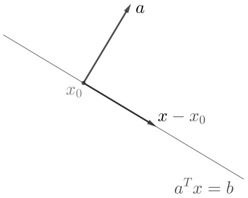
$$\theta_1\mathbf{x}_1 + \theta_2\mathbf{x}_2 \in \mathcal{C}.$$

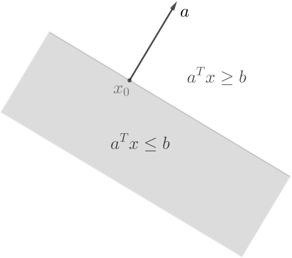


Hyperplanes and Halfspaces (超平面与半平面)

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

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

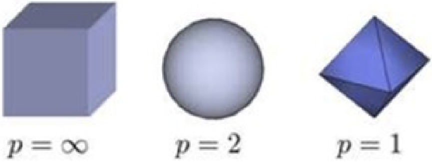




Hyperplane is affine set.

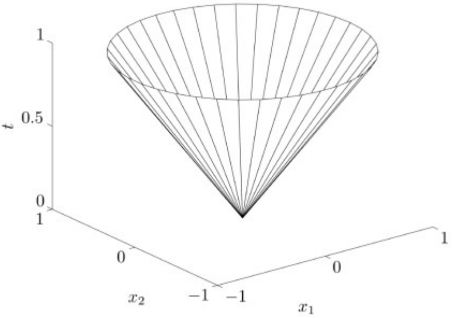
Norm Balls (范数球)

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



Norm Cones (范数锥)

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



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.

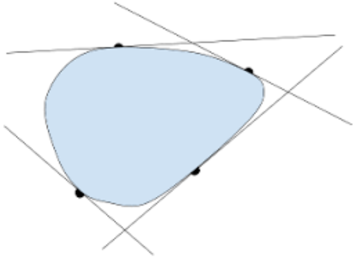
Operations that preserve convexity (保凸运算)

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.

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

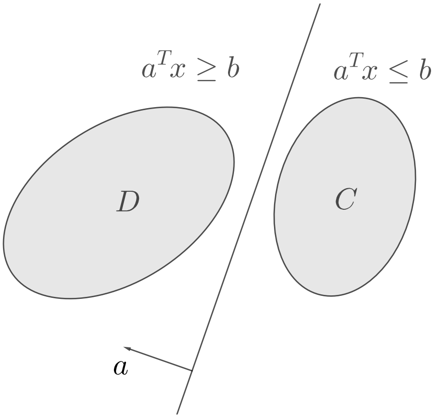
$$\mathcal{S} = \bigcap \{\mathcal{H} | \mathcal{H} \text{ is halfspace}, \mathcal{S} \subseteq \mathcal{H}\}$$



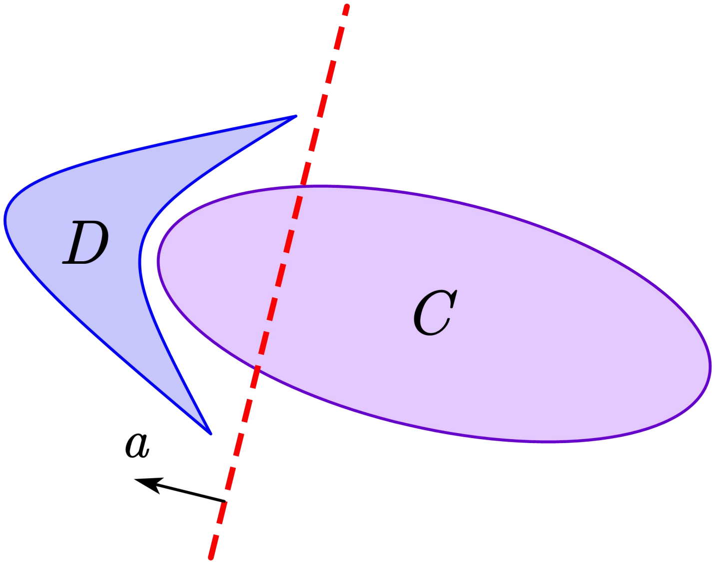
Hyperplane Separation Theorem

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.

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

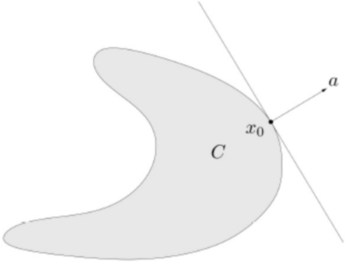
Example: a point and a closed convex set.

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 $\mathcal{C}$.