

# CS182 Introduction to Machine Learning

## Recitation 3

2025.3.12

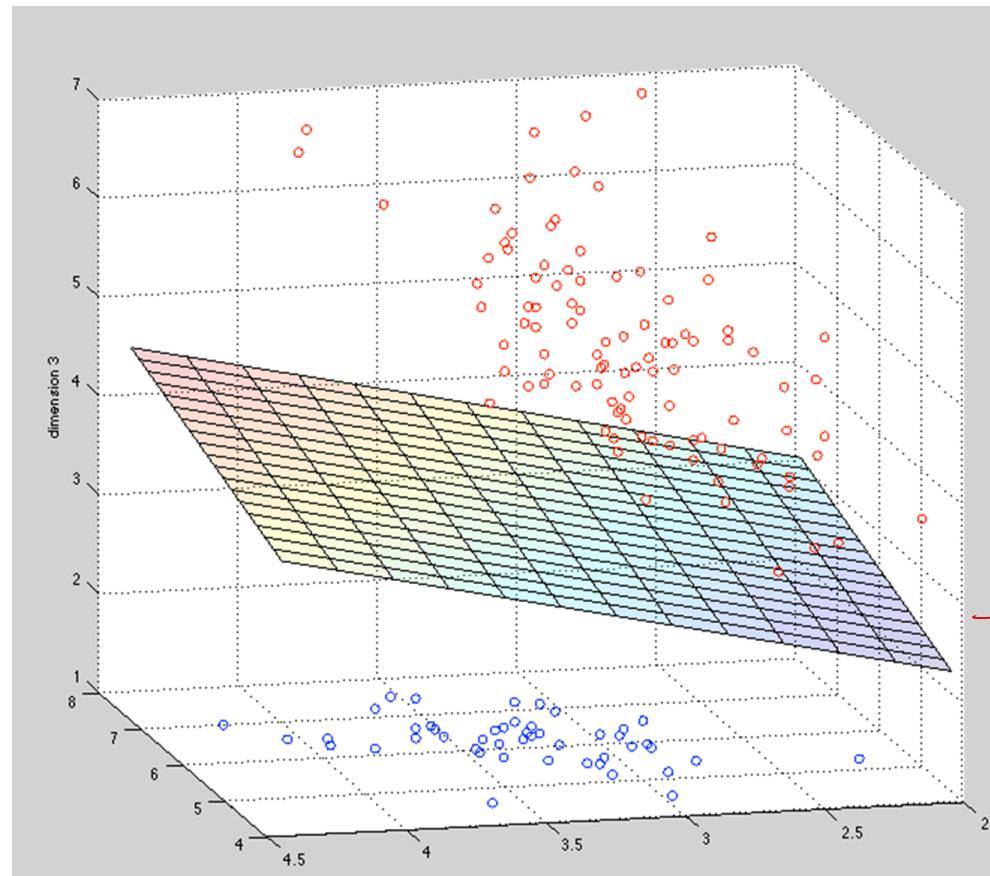
# Outline

- Perceptron
- Review(Preview): Optimization

# Linear Classification

$$\hat{y} = \text{sign}(\underline{\mathbf{w}}^\top \mathbf{x} + b) = \begin{cases} +1 & \text{if } \mathbf{w}^\top \mathbf{x} + b \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

$$\underline{\mathbf{w}}^\top (\underline{x}, \underline{x^2}, \underline{x^3}, \dots)$$



$$\underline{w} = (w, b)$$

$$\begin{array}{c} \underline{w}^\top \underline{x} + b \\ \downarrow \\ \underline{\mathbf{w}}^\top (\underline{x}, 1) \end{array}$$

# Perceptron

update rules

$$\mathcal{L} = - \sum_i y_i (\mathbf{w}^\top \mathbf{x}_i + b) \geq - \sum_i y_i \sum_j y_j [\mathbf{x}_j^\top \mathbf{x}_i] + b$$

$$\mathbf{w} = \sum_j y_j \mathbf{x}_j$$

$$k(x_i, x_j)$$

$$\mathbf{w} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

## Perceptron Algorithm: (without the intercept term)

- Set  $t=1$ , start with all-zeroes weight vector  $\mathbf{w}_1$ .
- Given example  $x$ , predict positive iff  $\mathbf{w}_t \cdot x \geq 0$ .
- On a mistake, update as follows:
  - Mistake on positive, update  $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + x$
  - Mistake on negative, update  $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - x$

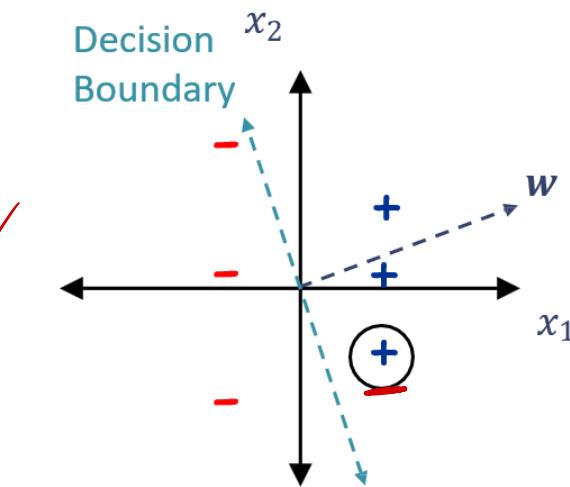
$x_1$	$x_2$	$\hat{y}$	$y$	Mistake?
-1	2	+	-	Yes
1	0	+	+	No
1	1	-	+	Yes
-1	0	-	-	No
-1	-2	+	-	Yes
1	-1	+	+	No

$$\mathbf{w} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

+ |

- |

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + y_i x$$



# Perceptron Convergence

Given dataset  $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^N$ , suppose:

1. Finite size inputs:  $\|x^{(i)}\| \leq R$
2. Linearly separable data:  $\exists \theta^*$  and  $\gamma > 0$  s.t.  $\|\theta^*\| = 1$  and  $y^{(i)}(\theta^* \cdot x^{(i)}) \geq \gamma, \forall i$

Then, the number of mistakes  $k$  made by the perceptron algorithm on  $\mathcal{D}$  is bounded by  $(R/\gamma)^2$ .

**Proof:**

Part 1: For some  $A$ ,  $Ak \leq \|\theta^{(k+1)}\|$

$$\begin{aligned}\theta^{(k+1)} \cdot \theta^* &= (\theta^{(k)} + y^{(i)}x^{(i)}) \cdot \theta^*, \text{ Perceptron algorithm update} \\ &= \theta^{(k)} \cdot \theta^* + y^{(i)}(\theta^* \cdot x^{(i)}) \\ &\geq \theta^{(k)} \cdot \theta^* + \gamma, \text{ by assumption} \\ \implies \theta^{(k+1)} \cdot \theta^* &\geq k\gamma, \text{ by induction on } k \text{ since } \theta^{(1)} = 0 \\ \implies \|\theta^{(k+1)}\| &\geq k\gamma, \text{ since } \|w\| \times \|u\| \geq w \cdot u \text{ and } \|\theta^*\| = 1\end{aligned}$$

Part 2: For some  $B$ ,  $\|\theta^{(k+1)}\| \leq B\sqrt{k}$

$$\begin{aligned}\|\theta^{(k+1)}\|^2 &= \|\theta^{(k)} + y^{(i)}x^{(i)}\|^2, \text{ Perceptron algorithm update} \\ &= \|\theta^{(k)}\|^2 + (y^{(i)})^2\|x^{(i)}\|^2 + 2y^{(i)}(\theta^{(k)} \cdot x^{(i)}) \\ &\leq \|\theta^{(k)}\|^2 + (y^{(i)})^2\|x^{(i)}\|^2, \text{ since } k^{\text{th}} \text{ mistake } \implies y^{(i)}(\theta^{(k)} \cdot x^{(i)}) \leq 0 \\ &= \|\theta^{(k)}\|^2 + R^2, \text{ since } (y^{(i)})^2\|x^{(i)}\|^2 = \|x^{(i)}\|^2 \leq R^2, \text{ by assumption and } (y^{(i)})^2 = 1 \\ &\implies \|\theta^{(k+1)}\|^2 \leq kR^2, \text{ by induction on } k \text{ since } (\theta^{(i)})^2 = 0 \\ &\implies \|\theta^{(k+1)}\| \leq \sqrt{k}R\end{aligned}$$



$$R = \max_i \|x_i\|$$

Part 3: Combine the bounds

$$\begin{aligned}k\gamma &\leq \|\theta^{(k+1)}\| \leq \sqrt{k}R \\ \implies k &\leq (R/\gamma)^2\end{aligned}$$

- Perceptron will not converge.
- However, we can achieve a similar bound on the number of mistakes made in one pass (Freund, Schapire)

Main Takeaway: For linearly separable data, if the perceptron algorithm repeatedly cycles through the data, it will converge in a finite number of steps.

If data has margin  $\gamma$  and all points inside a ball of radius  $R$ , then Perceptron

$$\leq \left(\frac{R}{\gamma}\right)^2 \text{ mistakes}$$

# Review(Preview): Optimization

- 通常讨论凸优化的范围
  - 凸集
  - 凸函数
  - 凸优化问题
- 优化方法
  - Lagrange Duality
  - KKT method

# Review(Preview) Outline

- Matrix Derivative
- Convex Function
- Convex Problem
- Duality, KKT Condition

# Matrix Derivatives 矩阵求导

Jacobian

Types	Scalar	Vector	Matrix	
Scalar	$\frac{dy}{dx}$	$\frac{dy}{dx} = \left( \frac{\partial y}{\partial x_1}, \dots, \frac{\partial y}{\partial x_i} \right)$	$\frac{dy}{dx} = \left( \frac{\partial y_i}{\partial x} \right)$ $\frac{\partial y}{\partial x} = (x, y)$ $\frac{\partial y}{\partial x} = (r, \theta)$	$\frac{dY}{dx} = \left( \frac{\partial Y_{i,j}}{\partial x} \right)$
Vector	$\frac{\partial y}{\partial x} = \left( \frac{\partial y}{\partial x_1}, \dots, \frac{\partial y}{\partial x_i} \right)$			
Matrix	$\frac{\partial y}{\partial x}$			$x = r \cos \theta, y = r \sin \theta$

$\frac{\partial(x, y)}{\partial(r, \theta)} = \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{pmatrix}$

$$\underline{\nabla f^T(x)} \cdot \underline{x}$$

$$\frac{\nabla^2 f}{m \times n} \cdot \underline{\underline{x}}^{n \times 1}$$

# layout

- 分子布局

numerator layout:

求导结果的维度以分子为主

- 分母布局

denominator layout:

求导结果的维度以分母为主

- 机器学习通常使用混合布局:  
向量或者矩阵对标量求导,

则使用分子布局为准, 如果是标量对向量或者矩阵求导, 则以分母布局为准

具体总结如下:

自变量\因变量	标量 $y$	列向量 $\underline{y}$	矩阵 $\underline{\underline{Y}}^{p \times q}$
标量 $x$	/	$\frac{\partial y}{\partial x}$ 分子布局: $m$ 维列向量 (默认布局) 分母布局: $m$ 维行向量	$\frac{\partial \underline{\underline{Y}}}{\partial x}$ 分子布局: $p \times q$ 矩阵 (默认布局) 分母布局: $q \times p$ 矩阵
列向量 $\underline{x}$	$\frac{\partial y}{\partial \underline{x}}$ 分子布局: $n$ 维行向量 分母布局: $n$ 维列向量 (默认布局)	$\frac{\partial y}{\partial \underline{x}}$ 分子布局: $m \times n$ 雅克比矩阵 (默认布局) 分母布局: $n \times m$ 梯度矩阵	/
矩阵 $\underline{\underline{X}}^{n \times m}$	$\frac{\partial y}{\partial \underline{\underline{X}}}^{n \times m}$ 分子布局: $n \times m$ 矩阵 分母布局: $m \times n$ 矩阵 (默认布局)	/	/

<https://blog.csdn.net/keeppractice>

# Matrix Derivatives

常见求导:

- $\frac{\partial \mathbf{a}^\top \mathbf{x}}{\partial \mathbf{x}} = \frac{\partial \mathbf{x}^\top \mathbf{a}}{\partial \mathbf{x}} = \mathbf{a}$
- $\frac{\partial \mathbf{x}^\top \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = \underbrace{(\mathbf{A} + \mathbf{A}^\top) \mathbf{x}}$
- more details:

Matrix cookbook

- Chain Rule 矩阵求导链式法则  
注意讲矩阵的维度对上

<https://www.cnblogs.com/yifanrensheng/p/12639539.html>

$$Ax \approx b$$

## least square approximation

$$\min_{\mathbf{x} \in \mathbb{R}^n} \mathcal{L} = \|\mathbf{b} - Ax\|_2^2$$

$$\mathcal{L} = \|\mathbf{b} - Ax\|_2^2 = (\mathbf{b} - Ax)^\top (\mathbf{b} - Ax) = \mathbf{b}^\top \mathbf{b} - \mathbf{b}^\top A \mathbf{x} - \mathbf{x}^\top A^\top \mathbf{b} + \mathbf{x}^\top A^\top A \mathbf{x}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{x}} = -2A^\top \mathbf{b} + 2A^\top A \mathbf{x} = 0 \quad \leftarrow \quad -A^\top \mathbf{b} - A^\top \mathbf{b} + (A^\top A + (A^\top A)^\top) \mathbf{x}$$

$$\Rightarrow \boxed{A^\top A \mathbf{x} = A^\top \mathbf{b}}$$

$$\frac{\partial^2 \mathcal{L}}{\partial \mathbf{x}^2} = \underline{2A^\top A} \geq 0$$

$$\frac{\partial a^\top \mathbf{x}}{\partial \mathbf{x}} = a$$

$$\frac{\partial x^\top A \mathbf{x}}{\partial \mathbf{x}} = (A + A^\top) \mathbf{x}$$

## Convex set

**line segment** between  $x_1$  and  $x_2$ : all points

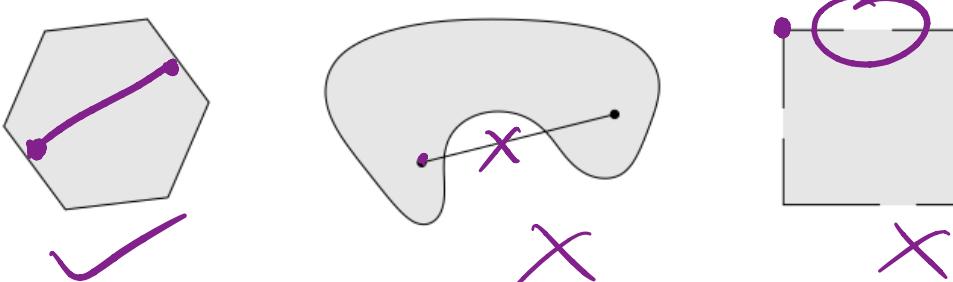
$$x = \theta x_1 + (1 - \theta)x_2$$

with  $0 \leq \theta \leq 1$

**convex set**: contains line segment between any two points in the set

$$\underbrace{x_1, x_2 \in C,}_{\text{ }} \underbrace{0 \leq \theta \leq 1}_{\text{ }} \implies \boxed{\theta x_1 + (1 - \theta)x_2} \in C$$

**examples** (one convex, two nonconvex sets)



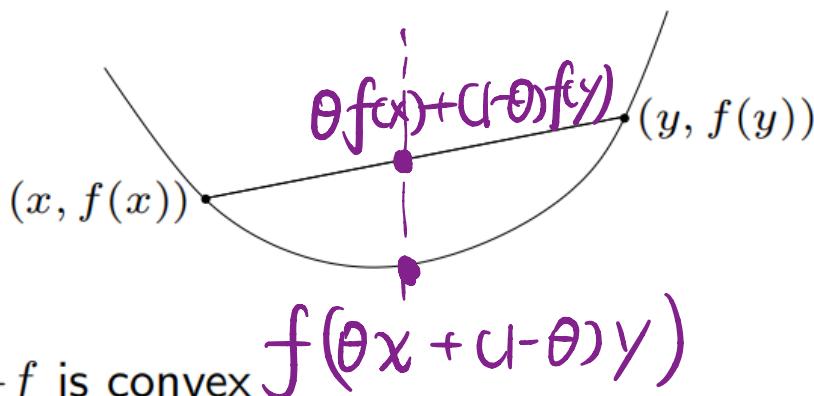
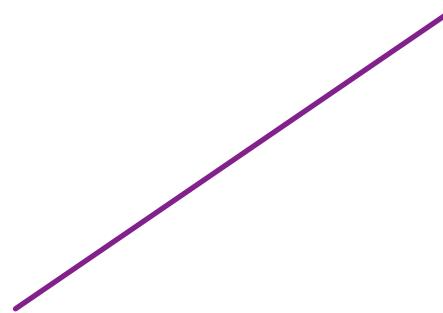
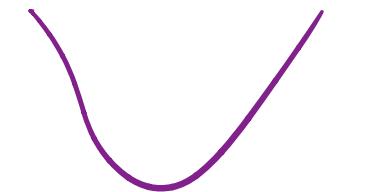
## Convex function

$f : \mathbf{R}^n \rightarrow \mathbf{R}$  is convex if  $\text{dom } f$  is a convex set and

零阶

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

for all  $x, y \in \text{dom } f, 0 \leq \theta \leq 1$

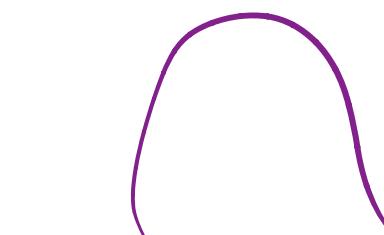


- $f$  is concave if  $-f$  is convex
- $f$  is strictly convex if  $\text{dom } f$  is convex and



$$f(\theta x + (1 - \theta)y) < \theta f(x) + (1 - \theta)f(y)$$

for  $x, y \in \text{dom } f, x \neq y, 0 < \theta < 1$



$$\nabla^2 f \succ 0$$

# 多元函数微分

nabla

- $\nabla$  算子:  $\nabla_x f$ : 函数  $f(\mathbf{x})$  对  $\mathbf{x}$  的梯度

$$\nabla f = \frac{\partial f}{\partial \mathbf{x}}$$

- $\nabla f$ : 一阶导 (Jacobian matrix):  $\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \dots & \frac{\partial f}{\partial x_n} \end{bmatrix}^\top$
- $\nabla^2 f$ : 二阶导 (Hessian matrix)

$$\nabla^2 f = \nabla(\nabla f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \dots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \dots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

$$(\nabla^2 f)^\top = \nabla^2 f$$

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$$

$\nabla f$ ,  $\nabla^2 f$

## Convex Function 凸函数

判据:  $f(\mathbf{x})$  是凸函数当且仅当

- $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \forall \theta \in [0, 1], f(\theta\mathbf{x} + (1 - \theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1 - \theta)f(\mathbf{y})$  0阶
- $\forall \mathbf{x} \in \mathbb{R}^n, \nabla^2 f(\mathbf{x}) \succeq 0$  2阶
- $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$  1阶

| 这三条本质是等价的, 可以互相推导



## Second-order conditions

$f$  is **twice differentiable** if  $\text{dom } f$  is open and the Hessian  $\nabla^2 f(x) \in \mathbf{S}^n$ ,

$$\nabla^2 f(x)_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}, \quad i, j = 1, \dots, n,$$

exists at each  $x \in \text{dom } f$

**2nd-order conditions:** for twice differentiable  $f$  with convex domain

- $f$  is convex if and only if

$$\nabla^2 f(x) \succeq 0 \quad \text{for all } x \in \text{dom } f$$

- if  $\nabla^2 f(x) \succ 0$  for all  $x \in \text{dom } f$ , then  $f$  is strictly convex

$$\begin{aligned}f(x) &= w^T x + b \\Df &= w \\D^2 f &= 0\end{aligned}$$

$$f(x+d) = f(x) + f'(x)d + \frac{1}{2}f''(x)d^2$$

# Taylor Expansion 泰勒展开

$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ :

$$f(\mathbf{x}+\mathbf{d}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top \mathbf{d} + \frac{1}{2} \mathbf{d}^\top \nabla^2 f(\mathbf{x}) \mathbf{d}$$

- 泰勒展开

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{1}{2} (\mathbf{y} - \mathbf{x})^\top \nabla^2 f(\mathbf{x}) (\mathbf{y} - \mathbf{x}) + o(\|\mathbf{y} - \mathbf{x}\|^2)$$

- 中值定理:

$$\exists \theta \in [0, 1], s.t. \mathbf{z} = \theta \mathbf{x} + (1 - \theta) \mathbf{y}:$$

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{z})^\top (\mathbf{y} - \mathbf{x})$$

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \underbrace{\frac{1}{2} (\mathbf{y} - \mathbf{x})^\top \nabla^2 f(\mathbf{z}) (\mathbf{y} - \mathbf{x})}_{\geq 0}$$

- 凸函数  $\nabla^2 f(\mathbf{x}) \succeq 0$

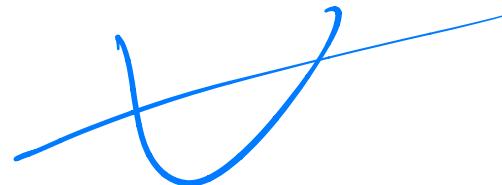
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$$

# Jensen's Inequality

对于一个凸函数 $f(x)$ , 有

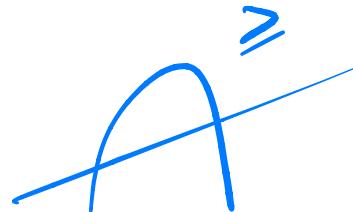
- 概率论角度:

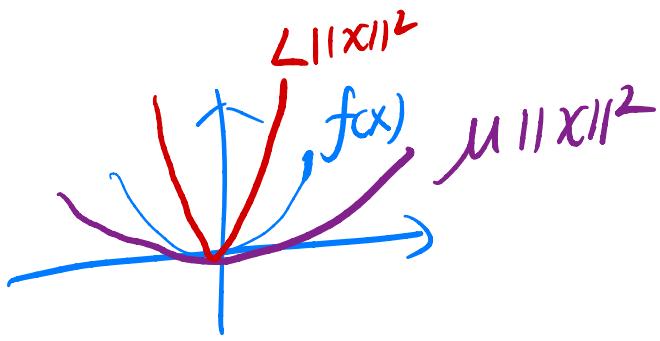
$$\underline{f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]}$$



- 优化角度:

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y), \forall x, y \in \mathbb{R}^n, \theta \in [0, 1]$$





$$A = \nabla^2 f(x)$$

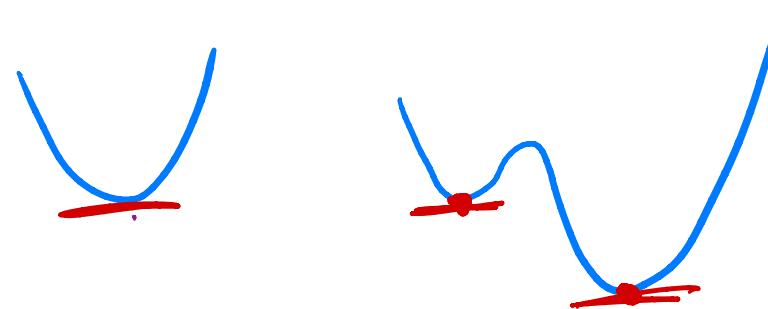
$$\mu \leq \lambda_1, \dots, \lambda_n \leq L$$

$\mu$ -strongly convex &  $L$ -smooth

$$\underline{\mu} I \preceq A \preceq L I$$

- 如果一个函数  $f(\mathbf{x})$  满足:  $\mu \|\mathbf{x}\|_2^2 \leq \mathbf{x}^\top \mathbf{A} \mathbf{x}$ , 或写作  $\mu I \preceq \mathbf{A}$ , 则称  $f(\mathbf{x})$  是  $\mu$ -strongly convex 的
- 如果一个函数  $f(\mathbf{x})$  满足:  $\mathbf{x}^\top \mathbf{A} \mathbf{x} \leq L \|\mathbf{x}\|_2^2$ , 或写作  $\mathbf{A} \preceq L I$ , 则称  $f(\mathbf{x})$  是  $L$ -smooth 的
- 一个函数的条件数(condition number)为  $\kappa = \frac{L}{\mu}$ , 这决定了 Gradient Descent 的收敛速度

# Convex Problem 凸优化问题



- 对于一个优化问题:

$$\begin{cases} \min_{\mathbf{x}} & \underline{f_0(\mathbf{x})} \\ \text{s.t.} & \underline{f_i(\mathbf{x})} \leq 0, i = 1, 2, \dots, m \\ & \underline{h_i(\mathbf{x})} = 0, i = 1, 2, \dots, n \end{cases}$$

$$f_i(\mathbf{x}) \geq 0$$

$$\begin{cases} & \underline{h_i(\mathbf{x})} \leq 0 \\ \xrightarrow{-} & \underline{h_i(\mathbf{x})} \leq 0 \end{cases}$$

- 其拉格朗日函数为:

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^n \nu_i h_i(\mathbf{x})$$

其中 $\boldsymbol{\lambda}$ 和 $\boldsymbol{\nu}$ 是拉格朗日乘子,  $\lambda_i \geq 0$ ,  $\nu_i$ 无约束

- 若 $f_0(\mathbf{x})$ 和 $f_i(\mathbf{x})$ 是凸函数,  $h_i(\mathbf{x})$ 是仿射/线性函数, 则原问题是凸优化问题

## Example

$$\begin{aligned} \min \quad & x_1^2 + x_2^2, \\ \text{s.t.} \quad & \underline{x_2 \leq \alpha}, \\ & \underline{x_1 + x_2 = 1} \end{aligned}$$

$$\begin{aligned} f_0(x) &= x_1^2 + x_2^2 \\ f_1(x) &= x_2 - \alpha \\ h_1(x) &= x_1 + x_2 - 1 \end{aligned}$$

其中  $(x_1, x_2) \in \mathbb{R}^2$ ,  $\alpha$  为实数

step1: 写出Lagrangian函数

$$\mathcal{L}(x_1, x_2, \mu, \lambda) = x_1^2 + x_2^2 + \lambda(x_2 - \alpha) + \mu(1 - x_1 - x_2), \text{ where } \underline{\lambda \geq 0}$$

$$\overline{f_0(x)} + \sum \lambda_i f_i(x) + \sum \mu_i h_i(x)$$

# Duality 对偶性

原问题(primal problem):

$$\min_{\mathbf{x}} f_0(\mathbf{x})$$

$$s.t. \quad f_i(\mathbf{x}) \leq 0, i = 1, 2, \dots, m$$
$$h_i(\mathbf{x}) = 0, i = 1, 2, \dots, n$$

对应的对偶问题(dual problem):

$$\max_{\boldsymbol{\lambda}, \boldsymbol{\nu}} g(\boldsymbol{\lambda}, \boldsymbol{\nu})$$

$$s.t. \quad \boldsymbol{\lambda} \succeq 0$$

$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$  取到  $g(\boldsymbol{\lambda}, \boldsymbol{\nu})$

$$f_0(\mathbf{x}) = \max_{\boldsymbol{\lambda} \succeq 0, \boldsymbol{\nu}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}), \quad g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \min_{\mathbf{x} \in \mathcal{D}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$$

无论原问题是否为凸优化问题, 对偶目标函数  $g(\boldsymbol{\lambda}, \boldsymbol{\nu})$  永远是凹函数!

$$x^* = \arg \min f_0(x)$$

## Duality 对偶性

$$\lambda^*, \mu^* = \arg \max g(\lambda, \mu)$$

$$f_0(\mathbf{x}) = \max_{\lambda \geq 0, \nu} \mathcal{L}(\mathbf{x}, \lambda, \nu), \quad g(\lambda, \nu) = \min_{\mathbf{x} \in \mathcal{D}} \mathcal{L}(\mathbf{x}, \lambda, \nu)$$

$$x^* \rightarrow p^* = \min_{\mathbf{x}} f_0(\mathbf{x})$$

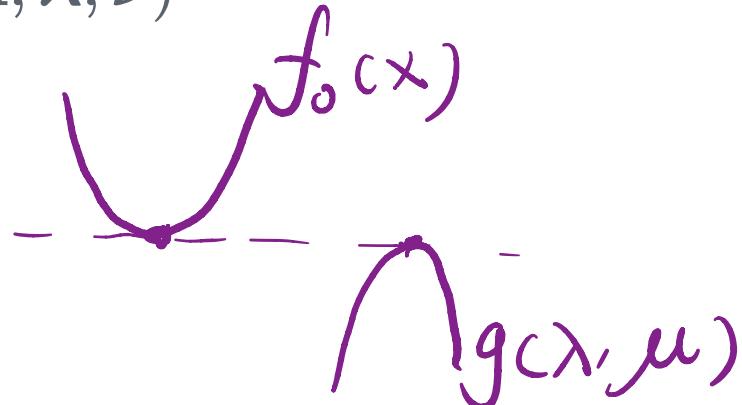
$$d^* = \max_{\lambda \geq 0, \nu} g(\lambda, \nu)$$

- Weak Duality  $p^* \geq d^*$

$$g(\lambda, \nu) = \min_{\mathbf{x} \in \mathcal{D}} \mathcal{L}(\mathbf{x}, \lambda, \nu) \leq \mathcal{L}(\underline{\mathbf{x}^*}, \lambda, \nu)$$

$$= f_0(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}^*) + \sum_{i=1}^n \nu_i h_i(\mathbf{x}^*) \leq f_0(\mathbf{x}^*) = p^*$$

上式  $\forall \lambda \geq 0, \nu$  成立,  $d^*$  符合该条件, 所以  $d^* \leq p^*$



# Duality 对偶性

- Strong Duality

$$p^* = d^*$$

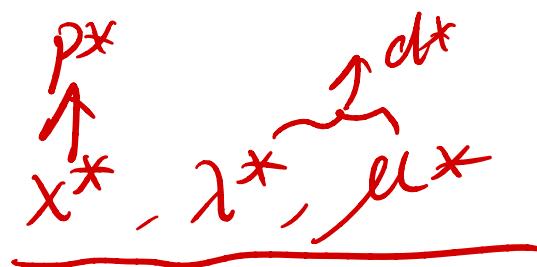
假设 strong duality 成立, 且  $\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*$  是原问题和对偶问题的最优解, 则有

$$\underline{f_0(\mathbf{x}^*) = g(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) = \min_{\mathbf{x} \in \mathcal{D}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)}$$

$$\begin{aligned} &= \min_{\mathbf{x} \in \mathcal{D}} \left( f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}) + \sum_{i=1}^n \nu_i^* h_i(\mathbf{x}) \right) \\ &\stackrel{x \in x^*}{\leq} f_0(\mathbf{x}^*) + \sum_{i=1}^m \underbrace{\lambda_i^* f_i(\mathbf{x}^*)}_{\geq 0} + \sum_{i=1}^n \underbrace{\nu_i^* h_i(\mathbf{x}^*)}_{\leq 0} \\ &\leq f_0(\mathbf{x}^*) \end{aligned}$$

- { 第一个不等号取等条件:  $\mathbf{x}^*$  minimizes  $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ , i.e.  $\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) = 0$   
第二个不等号取等条件:  $\lambda_i^* f_i(\mathbf{x}^*) = 0, i = 1, 2, \dots, m$

## KKT Condition



- primal feasibility:

$$\begin{cases} f_i(\mathbf{x}) \leq 0, i = 1, 2, \dots, m \\ h_i(\mathbf{x}) = 0, i = 1, 2, \dots, n \end{cases}$$

- dual feasibility:

$L(x, \mu, \lambda)$  取到  $g(\lambda, \mu)$   
 $\lambda \succeq 0$

- complementary slackness:

$$\lambda_i f_i(\mathbf{x}) = 0, i = 1, 2, \dots, m$$

- gradient of Lagrangian:

$$\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = 0$$

可以注意到我们推导出KKT条件的过程中完全没有要求原问题是一个凸优化问题

## Example

$$\begin{aligned} \min \quad & x_1^2 + x_2^2, \\ \text{s.t.} \quad & x_2 \leq \alpha, \\ & x_1 + x_2 = 1 \end{aligned}$$

其中  $(x_1, x_2) \in \mathbb{R}^2$ ,  $\alpha$  为实数

step1: 写出Lagrangian函数

$$\mathcal{L}(x_1, x_2, \mu, \lambda) = x_1^2 + x_2^2 + \lambda(x_2 - \alpha) + \mu(1 - x_1 - x_2), \text{ where } \lambda \geq 0$$

## step2: KKT condition

- primal feasibility:

$$\begin{cases} x_2 - \alpha \leq 0 \\ x_1 + x_2 = 1 \end{cases}$$

- dual feasibility:

$$\lambda \succeq 0$$

- complementary slackness:

$$\lambda(x_2 - \alpha) = 0$$

- gradient of Lagrangian:

$$\frac{\partial \mathcal{L}}{\partial x_i} = 0, \quad i = 1, 2$$

$$\frac{\partial \mathcal{L}}{\partial x_1} = 2x_1 - \mu = 0, \quad \frac{\partial \mathcal{L}}{\partial x_2} = 2x_2 - \mu + \lambda = 0$$

step3: Solve KKT or construct the dual problem

由 gradient of Lagrangian: 分别解出  $x_1 = \frac{\mu}{2}$  且  $x_2 = \frac{\mu}{2} - \frac{\lambda}{2}$ 。代入约束等式  $x_1 + x_2 = \mu - \frac{\lambda}{2} = 1$  或  $\mu = \frac{\lambda}{2} + 1$ 。合并上面结果,

$$x_1 = \frac{\lambda}{4} + \frac{1}{2}, \quad x_2 = -\frac{\lambda}{4} + \frac{1}{2}$$

最后再加入约束不等式  $-\frac{\lambda}{4} + \frac{1}{2} \leq \alpha$  或  $\lambda \geq 2 - 4\alpha$ 。底下分开三种情况讨论。

(1)  $\alpha > \frac{1}{2}$  : 不难验证  $\lambda = 0 > 2 - 4\alpha$  满足所有的 KKT 条件, 约束不等式是无效的,  $x_1^* = x_2^* = \frac{1}{2}$  是内部解, 目标函数的极小值是  $\frac{1}{2}$ 。

(2)  $\alpha = \frac{1}{2}$  : 如同 1,  $\lambda = 0 = 2 - 4\alpha$  满足所有的KKT条件,  $x_1^* = x_2^* = \frac{1}{2}$  是边界解, 因为  $x_2^* = \alpha$ 。

(3)  $\alpha < \frac{1}{2}$  : 这时约束不等式是有效的,  $\lambda = 2 - 4\alpha > 0$ , 则  $x_1^* = 1 - \alpha$  且  $x_2^* = \alpha$ , 目标函数的极小值是  $(1 - \alpha)^2 + \alpha^2$ 。

step3: Solve KKT or construct the dual problem

$$\begin{aligned}g(\lambda, \mu) &= \min_{x_1, x_2} \mathcal{L}(x_1, x_2, \lambda, \mu) = \min_{x_1, x_2} x_1^2 + x_2^2 + \lambda(x_2 - \alpha) + \mu(1 - x_1 - x_2) \\&= \min_{x_1, x_2} (x_1^2 - \mu x_1) + (x_2^2 + (\lambda - \mu)x_2) + \mu - \lambda\alpha \\&= -\frac{1}{2}\mu^2 + \frac{1}{2}\mu\lambda - \frac{1}{4}\lambda^2 + \mu - \lambda\alpha\end{aligned}$$

Dual problem:  $x_1 = \frac{\mu}{2}$      $x_2 = \frac{\mu - \lambda}{2}$

$$\begin{aligned}\max_{\mu, \lambda} \quad & -\frac{1}{2}\mu^2 + \frac{1}{2}\mu\lambda - \frac{1}{4}\lambda^2 + \mu - \lambda\alpha \\ \text{s.t.} \quad & \lambda \geq 0\end{aligned}$$

## Example: LP

- Primal problem:

$$\begin{aligned} \min_{\mathbf{x}} \quad & \mathbf{c}^T \mathbf{x} = f_0 \\ \text{s.t.} \quad & \mathbf{A}\mathbf{x} = \mathbf{b} + h(\mathbf{x}) \\ -\mathbf{x} \leq 0 \quad & \mathbf{x} \geq 0 \quad f_i(\mathbf{x}) \end{aligned}$$

- Dual problem:

$$\begin{aligned} \max_{\boldsymbol{\lambda}} \quad & \mathbf{b}^T \boldsymbol{\lambda} \\ \text{s.t.} \quad & \mathbf{A}^T \boldsymbol{\lambda} \leq \mathbf{c} \\ & \boldsymbol{\lambda} \leq 0 \end{aligned}$$

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = \mathbf{c}^T \mathbf{x} + \boldsymbol{\lambda}^T (\mathbf{A}\mathbf{x} - \mathbf{b}) - \boldsymbol{\nu}^T \mathbf{x} = (\mathbf{c} - \mathbf{A}^T \boldsymbol{\lambda} - \boldsymbol{\nu})^T \mathbf{x} - \mathbf{b}^T \boldsymbol{\lambda}$$

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = \begin{cases} -\mathbf{b}^T \boldsymbol{\lambda} & \text{if } \mathbf{c} - \mathbf{A}^T \boldsymbol{\lambda} - \boldsymbol{\nu} = 0 \\ -\infty & \text{otherwise} \end{cases}$$

$$\begin{aligned} \max & \quad -\mathbf{b}^T \boldsymbol{\lambda} \\ \text{s.t.} \quad & \boldsymbol{\nu} \geq 0 \end{aligned}$$

$$\mathbf{c} - \mathbf{A}^T \boldsymbol{\lambda} - \boldsymbol{\nu} = 0$$

## Example: LP

- Primal problem:

$$\min_{\mathbf{x}} \mathbf{c}^T \mathbf{x} = f_0$$

$$s.t. \quad \boxed{\mathbf{Ax} = \mathbf{b} + h_i \mathbf{x}}$$

$$-x \leq 0 \quad \underline{\mathbf{x} \geq 0} \quad f_i(x)$$

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = \underline{\mathbf{c}^T \mathbf{x}} + \underline{\boldsymbol{\lambda}^T (\mathbf{Ax} - \mathbf{b})} - \underline{\boldsymbol{\nu}^T \mathbf{x}} = (\mathbf{c} + \mathbf{A}^T \boldsymbol{\lambda} - \boldsymbol{\nu})^T \mathbf{x} - \mathbf{b}^T \boldsymbol{\lambda}$$

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = \begin{cases} -\mathbf{b}^T \boldsymbol{\lambda} & \text{if } \mathbf{c} + \mathbf{A}^T \boldsymbol{\lambda} - \boldsymbol{\nu} = 0 \\ -\infty & \text{otherwise} \end{cases}$$

- Dual problem:

$$\max_{\boldsymbol{\lambda}} \mathbf{b}^T \boldsymbol{\lambda}$$

$$s.t. \quad \mathbf{A}^T \boldsymbol{\lambda} \leq \mathbf{c}$$

$$\max -\mathbf{b}^T \boldsymbol{\lambda}$$

$$s.t. \quad \boldsymbol{\nu} \geq 0$$

$$\mathbf{c} - \mathbf{A}^T \boldsymbol{\lambda} - \boldsymbol{\nu} = 0$$

## Why dual problem?

- 减少变量的数量

e.g.  $A \in \mathbb{R}^{m \times n}$ ,  $m \ll n$

原问题  $n$  个变量, 对偶问题只有  $m$  个变量

- 拥有更好的形式

e.g. SVM 对偶问题引入了内积 -> 核函数

- 对偶目标函数是凹函数, 原问题未必是凸/凹函数

$$\theta g(\lambda_1) + (1-\theta)g(\lambda_2)$$

Primal	$x \in \mathbb{R}^n$	$n$
dual	$\lambda \in \mathbb{R}^m$	$m$

$$\lambda < (\lambda, \mu)$$

$$g(\lambda) = \min_x L(x, \lambda)$$

$$\begin{aligned} & \underline{g(\theta\lambda_1 + (1-\theta)\lambda_2)} \\ &= \min_x L(x, \theta\lambda_1 + (1-\theta)\lambda_2) \end{aligned}$$

$$= \min_x \theta L(x, \lambda_1) + (1-\theta) L(x, \lambda_2)$$

$$\leq \theta \min_x L(x, \lambda_1) + (1-\theta) \min_x L(x, \lambda_2)$$

# KKT condition 几何意义

对于不等式约束的互补条件

$$\lambda_i f_i(x) = 0, i = 1, 2, \dots, m$$

用只有一个不等式约束的情况来理解:

- 最优解在  $f_i(x) < 0$  处:  $f_i(x) \leq 0$   
不起作用, 起作用的为  $\nabla f_0(x) = 0$ .

$$\lambda = 0 \Leftrightarrow \nabla \mathcal{L}(x) = \nabla f_0(x)$$

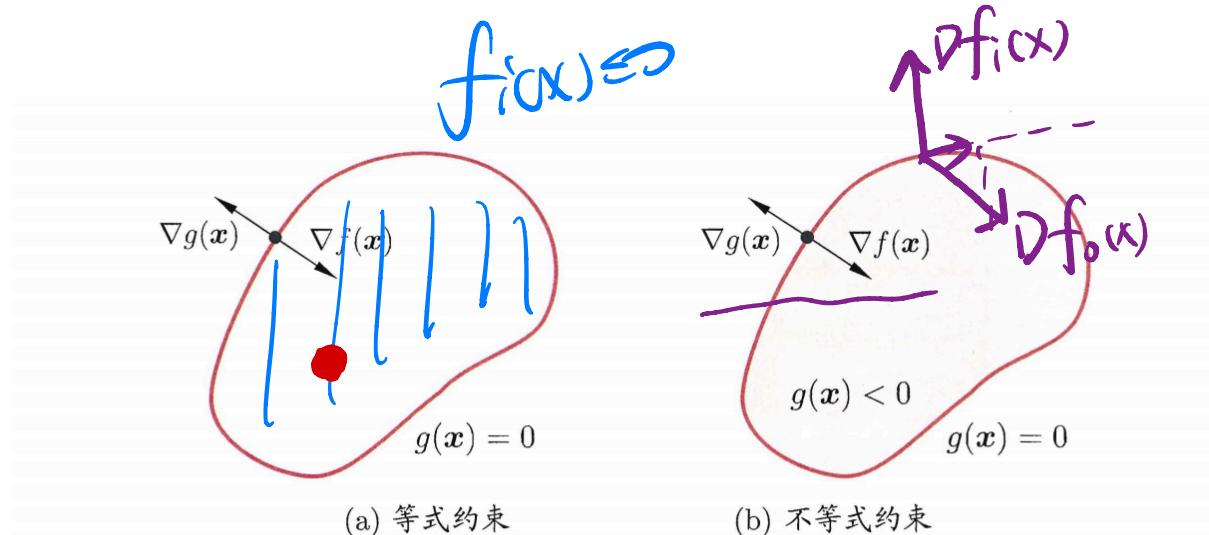
- 最优解在  $f_i(x) = 0$  处

一定有  $\nabla f_i(x)$  与  $\nabla f_0(x)$  反向, i.e.

$$\exists \lambda_i > 0, s.t. \nabla \mathcal{L}(x) = \nabla f_0(x) + \lambda_i \nabla f_i(x) = 0$$

结合两种情况可得  $\boxed{\lambda_i f_i(x) = 0}$

$$\lambda_i = -\frac{\nabla f_0(x)}{\nabla f_i(x)} \neq 0$$



附图B.1 拉格朗日乘子法的几何含义: 在 (a) 等式约束  $g(x) = 0$  或 (b) 不等式约束  $g(x) \leq 0$  下, 最小化目标函数  $f(x)$ . 红色曲线表示  $g(x) = 0$  构成的曲面, 而其围成的阴影区域表示  $g(x) < 0$ .

$$\begin{aligned} L &= f_0(x) + \lambda f_i(x) \\ \nabla L &= \nabla f_0(x) + \lambda \nabla f_i(x) \\ &\parallel \\ &= 0 \quad \lambda = 0 \end{aligned}$$

# KKT解一定是最优解吗？

- 必要性:

strong duality成立,  $\underline{\mathbf{x}, \lambda, \nu}$ 是原问题和对偶问题的最优解, 则他们满足KKT条件

- 充分性:

Theorem: 若原问题是一个凸优化问题, 且Slater's condition成立, 则KKT解一定是最优解

Slater's condition:  $\exists \mathbf{x} \in \text{int } \mathcal{D}, s.t. \underline{f_i(\mathbf{x}) < 0, i = 1, 2, \dots, m, \mathbf{A}\mathbf{x} = \mathbf{b}}$

或者有其他的Constrain Qualification(CQ)保证KKT解是最优解

# KKT 条件未必充分

$$\min_{\mathbf{x}} \quad f(\mathbf{x}) = x_1 + x_2$$

$$s.t. \quad c(\mathbf{x}) = x_1^2 + x_2^2 - 2 = 0$$

- Lagrange function:

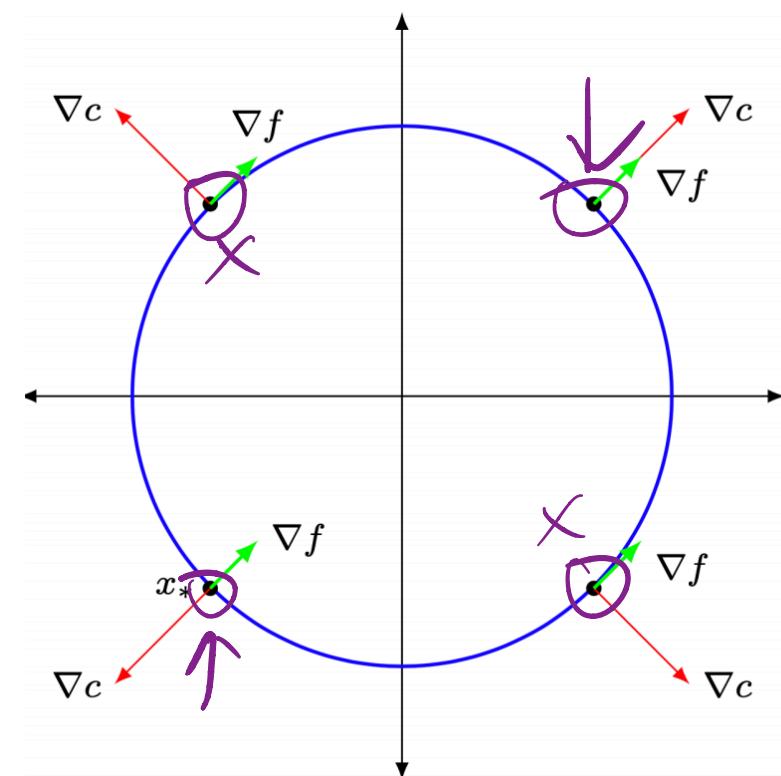
$$\mathcal{L}(\mathbf{x}, \lambda) = x_1 + x_2 + \lambda(x_1^2 + x_2^2 - 2)$$

- KKT condition:  $\begin{cases} 1 + 2\lambda x_1 = 0 \\ 1 + 2\lambda x_2 = 0 \\ x_1^2 + x_2^2 - 2 = 0 \end{cases}$

解KKT可获得两组解:

$$x = (1, 1), \lambda = -\frac{1}{2} \text{ 和 } x = (-1, -1), \lambda = \frac{1}{2}$$

但是 $(-1, -1)$ 是minimizer,  $(1, 1)$ 是maximizer



## 有最优解, 但KKT条件无解

$$\min_x \quad x$$

$$x = 0$$

$$s.t. \quad \underline{x^2 = 0}$$

- Lagrange function:  $\mathcal{L}(x, \lambda) = x + \lambda x^2$
- KKT condition:  $\begin{cases} \underline{1 + 2\lambda x = 0} \\ \underline{x^2 = 0} \end{cases} \quad \lambda \times$

