

Chapter 2 Convex Set

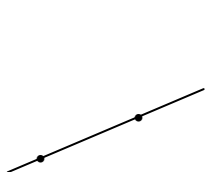
对于 $\mathbf{x}_1 \neq \mathbf{x}_2 \in \mathbf{R}^n$

连接两点的直线可表示为 $\{\mathbf{x} \mid \mathbf{x} = \theta \mathbf{x}_1 + (1 - \theta) \mathbf{x}_2, \theta \in \mathbf{R}\}$

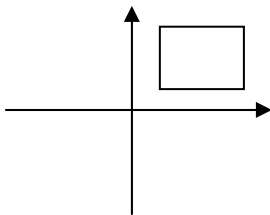
连接两点的线段可表示为 $\{\mathbf{x} \mid \mathbf{x} = \theta \mathbf{x}_1 + (1 - \theta) \mathbf{x}_2, \theta \in [0, 1]\}$

仿射集 (Affine Set)

定义: C 是仿射集 $\leftrightarrow \forall \mathbf{x}_1, \mathbf{x}_2 \in C$, 则连接两点的直线也在 C 内



一条直线是仿射集，一条线段则不是



平面是仿射集，一块矩形区域则不是

例: 线性方程组的解集是仿射集

证: 线性方程组的解集可表示为 $C = \{\mathbf{x} \mid A\mathbf{x} = \mathbf{b}, A \in \mathbf{R}^{m \times n}, \mathbf{b} \in \mathbf{R}^n\}$

设 $\mathbf{x}_1, \mathbf{x}_2 \in C$, 则 $A\mathbf{x}_1 = \mathbf{b}$, $A\mathbf{x}_2 = \mathbf{b}$

$A(\theta \mathbf{x}_1 + (1 - \theta) \mathbf{x}_2) = \theta A\mathbf{x}_1 + (1 - \theta) A\mathbf{x}_2 = \theta \mathbf{b} + (1 - \theta) \mathbf{b} = \mathbf{b}$, 得证

仿射组合 (Affine Combination)

设 k 个点 $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$, $\theta_1, \theta_2, \dots, \theta_k \in \mathbf{R}$
 $\theta_1 + \theta_2 + \dots + \theta_k = 1$

则称 $\theta_1 \mathbf{x}_1 + \theta_2 \mathbf{x}_2 + \dots + \theta_k \mathbf{x}_k$ 为 $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$ 的仿射组合

仿射包 (Affine Hull)

任意集合 $C \in \mathbf{R}^n$, C 中任意元素的仿射组合构成的集合称仿射包

$$\text{aff}C \triangleq \left\{ \theta_1 \mathbf{x}_1 + \theta_2 \mathbf{x}_2 + \dots + \theta_k \mathbf{x}_k \left| \begin{array}{l} \forall \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k \in C \\ \forall \theta_1, \theta_2, \dots, \theta_k \in \mathbf{R} \\ \theta_1 + \theta_2 + \dots + \theta_k = 1 \end{array} \right. \right\}$$

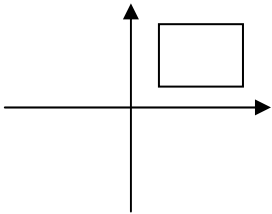
如果一个集合的仿射包就是它自己, 那么它就是一个仿射集

凸集 (Convex Set)

定义: C 是仿射集 $\leftrightarrow \forall \mathbf{x}_1, \mathbf{x}_2 \in C$, 则连接两点的**线段**也在 C 内



一条直线是凸集，一条线段也是凸集



平面是凸集，一块矩形区域也是凸集

凸组合 (Convex Combination)

设 k 个点 $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$, $\theta_1, \theta_2, \dots, \theta_k \in [0,1]$
 $\theta_1 + \theta_2 + \dots + \theta_k = 1$

则称 $\theta_1 \mathbf{x}_1 + \theta_2 \mathbf{x}_2 + \dots + \theta_k \mathbf{x}_k$ 为 $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$ 的凸组合

凸包 (Convex Hull)

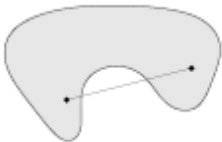
任意集合 $C \in \mathbf{R}^n$, C 中任意元素的凸组合构成的集合称凸包

$$\text{conv}C \triangleq \left\{ \theta_1 \mathbf{x}_1 + \theta_2 \mathbf{x}_2 + \dots + \theta_k \mathbf{x}_k \left| \begin{array}{l} \forall \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k \in C \\ \forall \theta_1, \theta_2, \dots, \theta_k \in [0,1] \\ \theta_1 + \theta_2 + \dots + \theta_k = 1 \end{array} \right. \right\}$$

一个凸集，它的凸包就是它本身



凸集



非凸集



非凸集



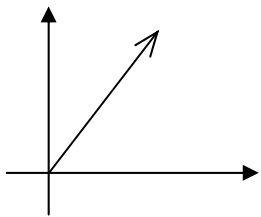
凸集的凸包 (本身)



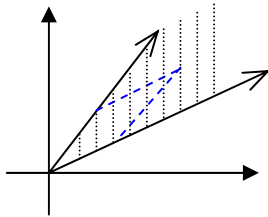
非凸集的凸包

凸锥 (Convex Cone)

定义: C 是仿射集 $\leftrightarrow \forall \mathbf{x}_1, \mathbf{x}_2 \in C$, 对 $\forall \theta_1, \theta_2 \geq 0$, 有 $\theta_1 \mathbf{x}_1 + \theta_2 \mathbf{x}_2 \in C$



过原点的射线是凸锥



过原点的两条射线之间的区域是凸锥

凸锥组合 (Convex Cone Combination)

设 k 个点 $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$, $\theta_1, \theta_2, \dots, \theta_k \geq 0$

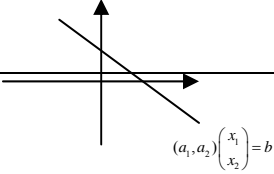
则称 $\theta_1 \mathbf{x}_1 + \theta_2 \mathbf{x}_2 + \dots + \theta_k \mathbf{x}_k$ 为 $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$ 的凸锥组合

凸锥包 (Convex Cone Hull)

任意集合 $C \in \mathbf{R}^n$, C 中任意元素的凸锥组合构成的集合称凸锥包

$$\left\{ \theta_1 \mathbf{x}_1 + \theta_2 \mathbf{x}_2 + \dots + \theta_k \mathbf{x}_k \left| \begin{array}{l} \forall \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k \in C \\ \forall \theta_1, \theta_2, \dots, \theta_k \geq 0 \end{array} \right. \right\}$$

一个凸锥集, 它的凸锥包就是它本身

集合名称	凸集	仿射集	凸锥
1. 空集	✓	✓	
2. 1 个点	✓	✓ $\theta x + (1-\theta)x = x$	
3. \mathbf{R}^n 空间	✓	✓	✓
4. \mathbf{R}^n 空间的子空间: $\{\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{0}\}$	✓	✓	✓
5. 任一直线	✓	✓	×
6. 任一线段	✓	×	×
7. 任一射线: $\{\mathbf{x}_0 + \theta \mathbf{v} \mid \theta \geq 0\}$	✓	×	×
8. 超平面: $\{\mathbf{x} \mid \mathbf{a}^T \mathbf{x} = b, \mathbf{x} \in \mathbf{R}^n, b \in \mathbf{R}\}$	✓	✓	×
9. 半空间: $\{\mathbf{x} \mid \mathbf{a}^T \mathbf{x} \geq b, \mathbf{x} \in \mathbf{R}^n, b \in \mathbf{R}\}$ $\{\mathbf{x} \mid \mathbf{a}^T \mathbf{x} \leq b, \mathbf{x} \in \mathbf{R}^n, b \in \mathbf{R}\}$		×	×
10. 欧几里得空间中的球: $\mathbf{B}(\mathbf{x}_c, r) = \{\mathbf{x} \mid \ \mathbf{x} - \mathbf{x}_c\ _2 \leq r, \mathbf{x}_c \in \mathbf{R}^n, r \in \mathbf{R}\}$	✓		
11. 欧几里得空间中的椭球: $\mathbf{E}(\mathbf{x}_c, r, \mathbf{P}) = \{\mathbf{x} \mid (\mathbf{x} - \mathbf{x}_c)^T \mathbf{P}^{-1} (\mathbf{x} - \mathbf{x}_c) \leq r, \mathbf{x}_c \in \mathbf{R}^n, r \in \mathbf{R}, \mathbf{P} \text{ 正定}\}$	✓		
12. 多面体 (Polyhedron): $\mathbf{P} = \left\{ \mathbf{x} \mid \begin{cases} \mathbf{a}_i^T \mathbf{x} \leq b_i, i=1, \dots, m \\ \mathbf{c}_j^T \mathbf{x} = d_j, j=1, \dots, n \end{cases} \right\}$ (很多超平面和半空间的交集)	✓		
13. 对称矩阵: $\mathbf{S}^n = \{\mathbf{A} \in \mathbf{R}^{n \times n} \mid \mathbf{A} = \mathbf{A}^T\}$	✓	✓	✓
14. 对称半正定矩阵: $\mathbf{S}_+^n = \{\mathbf{A} \in \mathbf{R}^{n \times n} \mid \mathbf{A} = \mathbf{A}^T, \mathbf{A} \succeq \mathbf{0}\}$	✓	×	✓
15. 对称正定矩阵: $\mathbf{S}_{++}^n = \{\mathbf{A} \in \mathbf{R}^{n \times n} \mid \mathbf{A} = \mathbf{A}^T, \mathbf{A} \succ \mathbf{0}\}$	✓	×	×

10) 证明: 设 $\mathbf{x}_1, \mathbf{x}_2 \in \mathbf{B}$, 则 $\begin{cases} \|\mathbf{x}_1 - \mathbf{x}_c\|_2 \leq r \\ \|\mathbf{x}_2 - \mathbf{x}_c\|_2 \leq r \end{cases}$

$$\begin{aligned}
 \|(\theta \mathbf{x}_1 + (1-\theta) \mathbf{x}_2) - \mathbf{x}_c\|_2 &= \|(\theta \mathbf{x}_1 + (1-\theta) \mathbf{x}_2) - (\theta \mathbf{x}_c + (1-\theta) \mathbf{x}_c)\|_2 \\
 &= \|\theta(\mathbf{x}_1 - \mathbf{x}_c) + (1-\theta)(\mathbf{x}_2 - \mathbf{x}_c)\|_2 \\
 &\leq \theta \|\mathbf{x}_1 - \mathbf{x}_c\|_2 + (1-\theta) \|\mathbf{x}_2 - \mathbf{x}_c\|_2 \\
 &\leq \theta r + (1-\theta)r = r
 \end{aligned}$$

13) 对称阵是仿射集，因而是凸集

证明：设 $A_1, A_2 \in \mathcal{S}^n$ ，则 $A_1^T = A_1$ ， $A_2^T = A_2$

对 $\forall \theta \in \mathbb{R}$ ， $(\theta A_1 + (1-\theta)A_2)^T = \theta A_1 + (1-\theta)A_2$ ，得证

14) 对称半正定阵是凸锥，因而是凸集

证明：设 $A_1, A_2 \in \mathcal{S}_+^n$ ，则 $A_1^T = A_1$ ， $A_2^T = A_2$ ，且对 $\forall \mathbf{x}$ ，有 $\mathbf{x}^T A_1 \mathbf{x} \geq 0$ ， $\mathbf{x}^T A_2 \mathbf{x} \geq 0$

对称已证明，现证明半正定，对 $\forall \mathbf{x}$ 和 $\forall \theta_1, \theta_2 \geq 0$ ， $\mathbf{x}^T (\theta_1 A_1 + \theta_2 A_2) \mathbf{x} = \theta_1 \mathbf{x}^T A_1 \mathbf{x} + \theta_2 \mathbf{x}^T A_2 \mathbf{x} \geq 0$ ，得证

14) 对称正定阵是凸集

证明：设 $A_1, A_2 \in \mathcal{S}_{++}^n$ ，则 $A_1^T = A_1$ ， $A_2^T = A_2$ ，且对 $\forall \mathbf{x} \neq 0$ ，有 $\mathbf{x}^T A_1 \mathbf{x} > 0$ ， $\mathbf{x}^T A_2 \mathbf{x} > 0$

对称已证明，现证明正定，对 $\forall \mathbf{x} \neq 0$ 和 $\forall \theta \in [0, 1]$ ， $\mathbf{x}^T (\theta A_1 + (1-\theta)A_2) \mathbf{x} = \theta \mathbf{x}^T A_1 \mathbf{x} + (1-\theta) \mathbf{x}^T A_2 \mathbf{x} > 0$ ，得证

保持凸集凸性的操作

➤ 交集 (Intersection)

若 S_1, S_2 为凸集, 则 $S_1 \cap S_2$ 为凸集。

若 S_α 为凸集, 对 $\forall \alpha \in A$, 则 $\bigcap_{\alpha \in A} S_\alpha$ 为凸集。

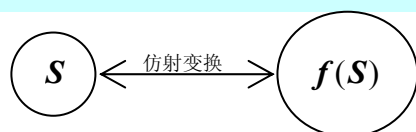
注: 并集不能保持凸集的凸性

➤ 仿射函数 (Affine Function)

定义: 对 $\text{dom } x \in \mathbf{R}^n$, $f(x) = Ax + b, A \in \mathbf{R}^{m \times n}, b \in \mathbf{R}^m$, 在称 $f: \mathbf{R}^n \rightarrow \mathbf{R}^m$ 是仿射的

定理: 若 $S \subseteq \mathbf{R}^n$ 是凸的, $f: \mathbf{R}^n \rightarrow \mathbf{R}^m$ 是仿射的, 则 S 在 f 下的映射 $f(S) = \{f(x) | x \in S\}$ 是凸的

同样, 若 $S \subseteq \mathbf{R}^n$ 是凸的, $g: \mathbf{R}^n \rightarrow \mathbf{R}^m$ 是仿射的, 则 S 在 g 下的反映射 $g^{-1}(S) = \{x | g(x) \in S\}$ 是凸的



例: 缩放 $\alpha S = \{\alpha x | x \in S\}$ 和位移 $S + a = \{x + a | x \in S\}$ 是保持凸性的

例: 两个凸集的**和** $S_1 + S_2 = \{x + y | x \in S_1, y \in S_2\}$ 是保持凸性的

解释如下, 集合 S_1 与集合 S_2 的**笛卡尔乘积** $S_1 \times S_2 \triangleq \{(x, y) | x \in S_1, y \in S_2\}$ 是保持凸性的

集合 $S_1 \times S_2$ 在仿射变换 $f(x, y) = x + y$ 下 $f(S_1 \times S_2) = S_1 + S_2$ 依然是保持凸性的

例: 线性矩阵不等式 (LMI: Linear Matrix Inequality): $A(x) = x_1 A_1 + \dots + x_n A_n \preceq B$, $B, A_i \in \mathbf{S}^n$ (对称阵)

类比一般的线性不等式 (Linear Inequality): $a^T x = x_1 a_1 + \dots + x_n a_n \leq b$, $b, x_i \in \mathbf{R}$

求证: 线性矩阵不等式 $A(X) \preceq B$ 的解集 $\{X | A(X) \preceq B\}$ 是凸集

证明: $f(X) \triangleq B - A(X) \in \mathbf{S}_+^n$ 是凸集

$f^{-1}(X) = \{X | B - A(X) \in \mathbf{S}_+^n\}$ 也是凸集

例: 椭圆是球的仿射映射, 因此是凸的

球 $\{u | \|u\|_2 \leq 1, u \in \mathbf{R}^n\}$ 是凸的, 它在仿射变换 $x(u) \triangleq P^{\frac{1}{2}} u + x_c$ 下 $\{x(u) | \|u\|_2 \leq 1, u \in \mathbf{R}^n\}$ 也是凸的

而 $\{x(u) | \|u\|_2 \leq 1, u \in \mathbf{R}^n\} \stackrel{u = P^{-\frac{1}{2}}(x - x_c)}{=} \left\{x(u) \left| \left\| P^{-\frac{1}{2}}(x - x_c) \right\|_2 \leq 1, u \in \mathbf{R}^n \right\} = \left\{x(u) \left| \left\| P^{-\frac{1}{2}}(x - x_c) \right\|_2^2 \leq 1, u \in \mathbf{R}^n \right\}$
 $\stackrel{\|a\|^2 = a^T a}{=} \left\{x(u) | (x - x_c)^T P^{-1} (x - x_c) \leq 1, u \in \mathbf{R}^n \right\}$ 就是椭圆

➤ 透视函数 (Perspective Function)

定义:

$$P: \mathbf{R}^{n+1} \rightarrow \mathbf{R}^n, \text{ dom } P = \mathbf{R}^n \times \mathbf{R}_{++} \quad (\mathbf{R}_{++} = \{t \in \mathbf{R} \mid t > 0\})$$

$$P(\mathbf{z}, t) = \frac{\mathbf{z}}{t}, \mathbf{z} \in \mathbf{R}^n, t \in \mathbf{R}_{++}$$

定理:

若 $C \subseteq \text{dom } P$ 是凸集, 则 $P(C) = \{P(\mathbf{x}) \mid \mathbf{x} \in C\}$ 也是凸集

例: 考虑 \mathbf{R}^{n+1} 内的一个线段

设 $\mathbf{x} = (\tilde{\mathbf{x}}, x_{n+1}), x_{n+1} > 0$
 $\mathbf{y} = (\tilde{\mathbf{y}}, y_{n+1}), y_{n+1} > 0$, 则 \mathbf{R}^{n+1} 内的一个线段可表示为 $\theta \mathbf{x} + (1-\theta) \mathbf{y}, \theta \in [0, 1]$

该线段经透视变换后仍为线段

$$\begin{aligned} P(\theta \mathbf{x} + (1-\theta) \mathbf{y}) &= \frac{\theta \tilde{\mathbf{x}} + (1-\theta) \tilde{\mathbf{y}}}{\theta x_{n+1} + (1-\theta) y_{n+1}} = \frac{\theta x_{n+1}}{\theta x_{n+1} + (1-\theta) y_{n+1}} \frac{\tilde{\mathbf{x}}}{x_{n+1}} + \frac{(1-\theta) y_{n+1}}{\theta x_{n+1} + (1-\theta) y_{n+1}} \frac{\tilde{\mathbf{y}}}{y_{n+1}} \\ &\stackrel{\mu = \frac{\theta x_{n+1}}{\theta x_{n+1} + (1-\theta) y_{n+1}} \in [0, 1]}{=} \mu P(\mathbf{x}) + (1-\mu) P(\mathbf{y}) \end{aligned}$$

➤ 线性分数函数 (Linear-fractional Function)

设仿射函数 $\mathbf{g}: \mathbf{R}^n \rightarrow \mathbf{R}^{m+1}$ 满足 $\mathbf{g}(\mathbf{x}) = \begin{pmatrix} \mathbf{A} \\ \mathbf{c}^T \end{pmatrix} \mathbf{x} + \begin{pmatrix} \mathbf{b} \\ d \end{pmatrix}, \mathbf{A} \in \mathbf{R}^{m \times n}, \mathbf{b} \in \mathbf{R}^m, \mathbf{c} \in \mathbf{R}^n, d \in \mathbf{R}$

设透视函数 $P: \mathbf{R}^{m+1} \rightarrow \mathbf{R}^m$ 满足 $P(\mathbf{z}, t) = \frac{\mathbf{z}}{t}, \mathbf{z} \in \mathbf{R}^m, t \in \mathbf{R}_{++}$

定义线性分数函数 $\mathbf{f}: \mathbf{R}^n \rightarrow \mathbf{R}^m$ 满足 $\mathbf{f} = P \circ \mathbf{g}$, 即

$$\mathbf{f}(\mathbf{x}) = P(\mathbf{g}(\mathbf{x})) = P\left(\frac{\mathbf{Ax} + \mathbf{b}}{\mathbf{c}^T \mathbf{x} + d}\right) = \frac{\mathbf{Ax} + \mathbf{b}}{\mathbf{c}^T \mathbf{x} + d} \quad \text{dom } \mathbf{f} = \{\mathbf{x} \mid \mathbf{c}^T \mathbf{x} + d > 0\}$$

例: 两个随机变量的联合概率与条件概率

设两个离散型随机变量 (Random Variable) $U: \{1, \dots, n\}, V: \{1, \dots, m\}$

联合概率: $p_{ij} = P\{U=i, V=j\}$; 条件概率: $q_{ij} = P\{U=i \mid V=j\} = \frac{P\{U=i, V=j\}}{P\{V=j\}} = \frac{p_{ij}}{\sum_{k=1}^n p_{kj}}$

$$\text{设 } \mathbf{p} \triangleq \begin{pmatrix} p_{11} \\ \vdots \\ p_{1n} \\ \vdots \\ p_{1m} \\ \vdots \\ p_{nm} \end{pmatrix} = \begin{pmatrix} p_{11} \\ \vdots \\ p_{n1} \\ \vdots \\ p_{1m} \\ \vdots \\ p_{nm} \end{pmatrix}, \quad \mathbf{q} \triangleq \begin{pmatrix} q_{11} \\ \vdots \\ q_{n1} \\ \vdots \\ q_{1m} \\ \vdots \\ q_{nm} \end{pmatrix} = \begin{pmatrix} q_{11} \\ \vdots \\ q_{n1} \\ \vdots \\ q_{1m} \\ \vdots \\ q_{nm} \end{pmatrix}, \quad \mathbf{I}_j \triangleq (\mathbf{0} \quad \mathbf{0} \quad \cdots \quad \underbrace{\mathbf{I}_{n \times n}}_{j\text{th block}} \quad \cdots \quad \mathbf{0})_{nm \times n}$$

$$\mathbf{e}_j \triangleq (\mathbf{0} \quad \mathbf{0} \quad \cdots \quad \underbrace{1 \cdots 1}_{j\text{th block, n 1s}} \quad \cdots \quad \mathbf{0})_{nm \times 1}, \quad \text{于是 } \mathbf{q}_j = \mathbf{f}(\mathbf{p}) = \frac{\mathbf{I}_j \mathbf{p}}{\mathbf{e}_j^T \mathbf{p}}$$

$\{\mathbf{q}_j\}$ 是凸集, $\{\mathbf{q}\}$ 是凸集

Chapter 3 Convex Function

凸函数:

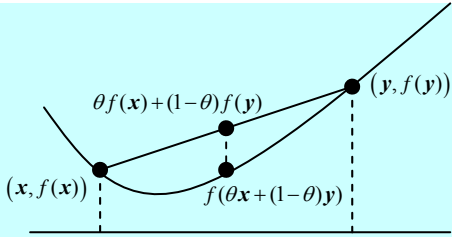
一个函数 $f: \mathbf{R}^n \rightarrow \mathbf{R}$ 为凸函数 \Leftrightarrow

若 $\text{dom}f$ 为凸集, 且对 $\forall \mathbf{x}, \mathbf{y} \in \text{dom}f$, $\forall \theta \in [0,1]$, 有

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

即凸组合的函数值 \leq 函数值的凸组合

$\text{dom}f$ 为凸集
是为了保证 $\theta \mathbf{x} + (1-\theta)\mathbf{y}$ 在定义域内



严格凸函数:

一个函数 $f: \mathbf{R}^n \rightarrow \mathbf{R}$ 为凸函数 \Leftrightarrow

若 $\text{dom}f$ 为凸集, 且对 $\forall \mathbf{x} \neq \mathbf{y} \in \text{dom}f$, $\forall \theta \in (0,1)$, 有

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

凹函数:

若 $-f$ 是凸函数, 则 f 是凹函数

严格凹函数:

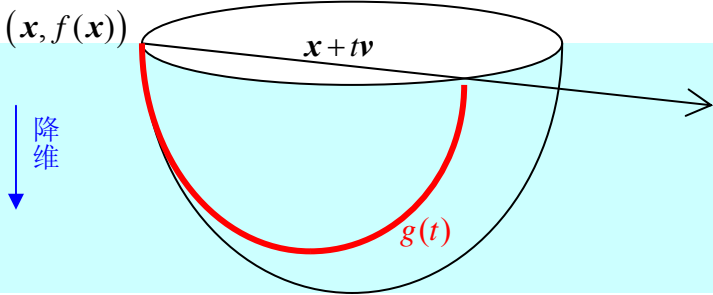
若 $-f$ 是严格凸函数, 则 f 是严格凹函数

凸函数的另一种定义:

一个函数 $f: \mathbf{R}^n \rightarrow \mathbf{R}$ 为凸函数 \Leftrightarrow

若 $\text{dom}f$ 为凸集, 且对 $\forall \mathbf{x} \in \text{dom}f$, $\forall \mathbf{v} \in \mathbf{R}^n$, 有

$g(t) = f(\mathbf{x} + t\mathbf{v})$ 在 $\text{dom}g = \{t \mid \mathbf{x} + t\mathbf{v} \in \text{dom}f\}$ 上是凸的



扩展的凸函数:

设函数 $f: \mathbf{R}^n \rightarrow \mathbf{R}$ 为凸函数, $\text{dom}f \in \mathbf{R}^n$, 则定义

$$\tilde{f}(\mathbf{x}) = \begin{cases} f(\mathbf{x}) & \mathbf{x} \in \text{dom}f \\ \infty & \mathbf{x} \notin \text{dom}f \end{cases}$$

例: 已知凸集 $C \in \mathbf{R}^n$

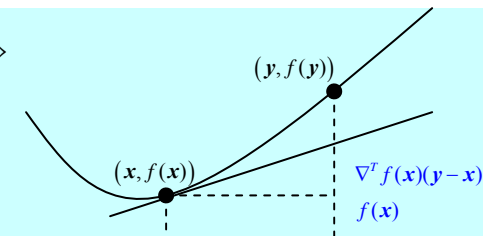
$f(\mathbf{x}) = \begin{cases} 0 & \mathbf{x} \in C \\ \text{no defined} & \mathbf{x} \notin C \end{cases}$	$I(\mathbf{x}) = \begin{cases} 0 & \mathbf{x} \in C \\ \infty & \mathbf{x} \notin C \end{cases}$	$JC(\mathbf{x}) = \begin{cases} 0 & \mathbf{x} \in C \\ 1 & \mathbf{x} \notin C \end{cases}$
是凸函数	是凸函数	不是凸函数

凸函数的一阶条件（可微情况下的等价定义）:

设函数 $f: \mathbf{R}^n \rightarrow \mathbf{R}$ 可微，即梯度 ∇f 在 $\text{dom}f$ 均存在，则 f 为凸函数 \Leftrightarrow

若 $\text{dom}f$ 为凸集，且对 $\forall \mathbf{x}, \mathbf{y} \in \text{dom}f$ ，有

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



性质:

若 f 为凸函数， $\exists \mathbf{x}_0 \in \text{dom}f$ ，使 $\nabla f(\mathbf{x}_0) = \mathbf{0}$ ，则

对 $\forall \mathbf{y} \in \text{dom}f$ ， $f(\mathbf{y}) \geq f(\mathbf{x}_0) + \nabla^T f(\mathbf{x}_0)(\mathbf{y} - \mathbf{x}_0) = f(\mathbf{x}_0)$

即 $f(\mathbf{x}_0)$ 是 f 的最小值

同理可定义严格凸函数的一阶条件

设函数 $f: \mathbf{R}^n \rightarrow \mathbf{R}$ 可微，即梯度 ∇f 在 $\text{dom}f$ 均存在，则 f 为凸函数 \Leftrightarrow

若 $\text{dom}f$ 为凸集，且对 $\forall \mathbf{x} \neq \mathbf{y} \in \text{dom}f$ ，有

$$f(\mathbf{y}) > f(\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x})$$

附录：求证凸函数的一阶条件

Step1: 考虑一维情况，即 f 为凸函数 \Leftrightarrow 若 $\text{dom}f$ 为凸集，且对 $\forall x, y \in \text{dom}f$ ，有 $f(y) \geq f(x) + f'(x)(y-x)$

充分性 (\Rightarrow)

若 f 为凸函数，则对 $\forall x, y \in \text{dom}f$ ($\text{dom}f$ 为凸集)，有

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

$$\theta f(y) \geq \theta f(x) + f(x + \theta(y-x)) - f(x)$$

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

$$\text{当 } \theta \rightarrow 0 \text{ 时, } f(y) \geq f(x) + f'(x)(y-x)$$

必要性 (\Leftarrow)

对 $\forall x, y \in \text{dom}f$ ($\text{dom}f$ 为凸集)，构造 $z = \theta x + (1-\theta)y$

$$f(x) \geq f(z) + f'(z)(x-z) \quad (1)$$

$$f(y) \geq f(z) + f'(z)(y-z) \quad (2)$$

把 $\theta \times (1) + (1-\theta) \times (2)$ ，得 $\theta f(x) + (1-\theta)f(y) \geq f(z) - f'(z)(\theta x + (1-\theta)y - z) = f(z)$ ，得证

Step2: 考虑多维情况，即 f 为凸函数 \Leftrightarrow 若 $\text{dom}f$ 为凸集，且对 $\forall \mathbf{x}, \mathbf{y} \in \text{dom}f$ ，有 $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y}-\mathbf{x})$

充分性 (\Rightarrow)

若 f 为凸函数，则对 $\forall \mathbf{x}, \mathbf{y} \in \text{dom}f$ ($\text{dom}f$ 为凸集)，有

$$g(\theta) = f(\mathbf{x} + \theta(\mathbf{y}-\mathbf{x})), \quad \forall \theta \in [0,1] \quad \longrightarrow \quad \text{构造一个低维凸函数}$$

$$g'(\theta) = \nabla^T f(\mathbf{x} + \theta(\mathbf{y}-\mathbf{x})) \cdot (\mathbf{y}-\mathbf{x})$$

$$g(\theta_1) \geq g(\theta_2) + g'(\theta_2)(\theta_1 - \theta_2) \quad \longrightarrow \quad \text{Step1 的结论}$$

$$g(1) \geq g(0) + \nabla^T f(\mathbf{x}) \cdot (\mathbf{y}-\mathbf{x}) \quad \longrightarrow \quad g(1) = f(\mathbf{y}), g(0) = f(\mathbf{x})$$

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

必要性 (\Leftarrow)

对 $\forall \mathbf{x}, \mathbf{y} \in \text{dom}f$ ($\text{dom}f$ 为凸集)，构造 $z = \theta x + (1-\theta)y$

$$\begin{cases} g(\theta_1) = f(\mathbf{x} + \theta_1(\mathbf{y}-\mathbf{x})) \\ g(\theta_2) = f(\mathbf{x} + \theta_2(\mathbf{y}-\mathbf{x})) \\ g'(\theta_2) = \nabla^T f(\mathbf{x} + \theta_2(\mathbf{y}-\mathbf{x})) \cdot (\mathbf{y}-\mathbf{x}) \end{cases}, \quad \forall \theta_1, \theta_2 \in [0,1]$$

$$f(\mathbf{x} + \theta_1(\mathbf{y}-\mathbf{x})) \geq f(\mathbf{x} + \theta_2(\mathbf{y}-\mathbf{x})) + \nabla^T f(\mathbf{x} + \theta_2(\mathbf{y}-\mathbf{x})) \cdot (\mathbf{y}-\mathbf{x})(\theta_1 - \theta_2)$$

$$g(\theta_1) \geq g(\theta_2) + g'(\theta_2)(\theta_1 - \theta_2)$$

凸函数的二阶条件（二阶可微情况下的等价定义）：

定义：

设向量 \mathbf{x} 的数量值函数 $f(\mathbf{x})$ 二阶可微，则有

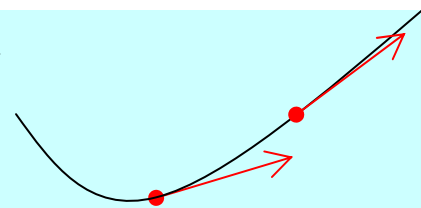
$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad \nabla f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}, \quad \nabla^2 f(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix} \quad (\text{Hessian 矩阵})$$

定义：

设函数 $f: \mathbf{R}^n \rightarrow \mathbf{R}$ 二阶可微，即梯度 $\nabla^2 f$ 在 $\text{dom}f$ 均存在，则 f 为凸函数 \Leftrightarrow

若 $\text{dom}f$ 为凸集，且对 $\forall \mathbf{x} \in \text{dom}f$ ，有 $\nabla^2 f(\mathbf{x})$ 半正定

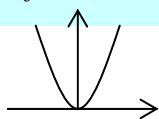
Note：对于一维情况，则要求该函数的二阶偏导 ≥ 0



定义：

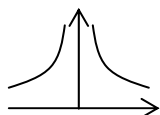
f 为严格凸函数 \Leftarrow 若 $\text{dom}f$ 为凸集，且对 $\forall \mathbf{x} \in \text{dom}f$ ，有 $\nabla^2 f(\mathbf{x})$ 正定

例如：函数 $f(x) = x^4$



它是一个严格凸函数，但 $f(x)$ 在 $x=0$ 这一点不满足正定条件，故 $\frac{\partial^2 f(x)}{\partial x^2}$ 正定是严格凸函数的充分而不必要条件

例：函数 $f(x) = \frac{1}{x^2}$



虽然 $\frac{\partial^2 f(x)}{\partial x^2} = 6x^{-4} > 0$ ，但 $\text{dom}f$ 不是凸集，故不能通过凸函数的二阶条件来判定凸函数

事实上，该函数确实不是凸函数

例：二次函数 $f: \mathbf{R}^n \rightarrow \mathbf{R}$ ， $\text{dom}f = \mathbf{R}^n$ ， $f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{P} \mathbf{x} + \mathbf{q}^T \mathbf{x} + r$ ($\mathbf{P} \in \mathbf{S}^n$ ， $\mathbf{q} \in \mathbf{R}^n$ ， $r \in \mathbf{R}$)

$\nabla^2 f(\mathbf{x}) = \mathbf{P}$ ，
if $\mathbf{P} \succeq 0$, convex function
if $\mathbf{P} \succ 0$, strictly convex function
if $\mathbf{P} \preceq 0$, concave function

例：仿射函数 $f: \mathbf{R}^n \rightarrow \mathbf{R}^m$ ， $\text{dom}f = \mathbf{R}^n$ ， $f(\mathbf{x}) = \mathbf{A} \mathbf{x} + \mathbf{b}$ ($\mathbf{A} \in \mathbf{R}^{m \times n}$ ， $\mathbf{b} \in \mathbf{R}^m$)

$\nabla^2 f(\mathbf{x}) = \mathbf{0}$ ，故 $f(\mathbf{x})$ 即是凸函数又是凹函数

例：指数函数

$$f(x) = e^{ax}, x \in \mathbf{R}$$

$$f'(x) = e^{ax} a$$

$$f''(x) = e^{ax} a^2 \geq 0$$

故指数函数是凸函数

例：幂函数

$$f(x) = x^a, x \in \mathbf{R}_{++}$$

$$f'(x) = ax^{a-1}$$

$$f''(x) = a(a-1)x^{a-2} \begin{cases} \geq 0 & a \geq 1 \text{ or } a \leq 0 \\ \leq 0 & 0 \leq a \leq 1 \end{cases}$$

故幂函数 $f(x)$ is $\begin{cases} \text{convex} & a \geq 1 \text{ or } a \leq 0 \\ \text{concave} & 0 \leq a \leq 1 \end{cases}$

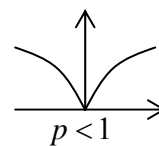
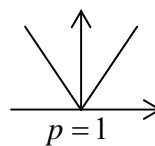
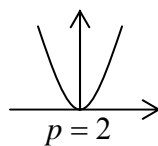
例：绝对值的幂函数

$$f(x) = |x|^p, x \in \mathbf{R}, p \geq 0 (\text{avoid singularity})$$

$$f'(x) = \begin{cases} px^{p-1} & x \geq 0 \\ -p(-x)^{p-1} & x < 0 \end{cases}$$

$$f''(x) = \begin{cases} p(p-1)x^{p-2} & x \geq 0 \\ p(p-1)(-x)^{p-2} & x < 0 \end{cases}$$

故 $f(x)$ is $\begin{cases} \text{convex} & p \geq 2 \\ \text{convex} & p \in [1, 2) \\ \text{not convex, not concave} & p \in (0, 1) \end{cases}$



例：对数函数

$$f(x) = \log x, x \in \mathbf{R}_{++}$$

$$f'(x) = \frac{1}{x}$$

$$f''(x) = -\frac{1}{x^2} < 0$$

故对数函数是凹函数

Note: 极大似然函数采用对数形式就是因为极大化凹函数是方便可行的

例：负熵

$$f(x) = x \log x, x \in \mathbf{R}_{++}$$

$$f'(x) = 1 + \log x$$

$$f''(x) = \frac{1}{x} > 0$$

故负熵是凸函数

Note: 极小化负熵也是因为极小化凸函数是方便可行的

例：范数

若 $p(\mathbf{x})$ 是范数，则

1) 正定性: $p(\mathbf{x}) \geq 0$ and $p(\mathbf{x}) = 0$, if $\mathbf{x} = \mathbf{0}$

2) 齐次性: $p(a\mathbf{x}) = |a| p(\mathbf{x})$

3) 三角不等式: $p(\mathbf{x} + \mathbf{y}) \leq p(\mathbf{x}) + p(\mathbf{y})$

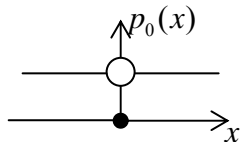
$\forall \mathbf{x}, \mathbf{y} \in \mathbf{R}^n, \forall \theta \in [0, 1]$

$$\begin{aligned} p(\theta \mathbf{x} + (1-\theta)\mathbf{y}) &\leq p(\theta \mathbf{x}) + p((1-\theta)\mathbf{y}) \\ &= |\theta| p(\mathbf{x}) + |1-\theta| p(\mathbf{y}) \\ &= \theta p(\mathbf{x}) + (1-\theta)p(\mathbf{y}) \end{aligned}$$

所以 $p(\mathbf{x})$ 是凸函数

例：零范数: $p_0(\mathbf{x}) = \|\mathbf{x}\|_0 \triangleq \text{the number of non-zero elements in } \mathbf{x}$

若 x 为标量，则



显然 $p_0(x)$ 不是凸函数

因为零范数不满足范数的条件 2) 和 3)，是一个伪范数。

例：极大值函数

$$f(\mathbf{x}) = \max\{x_1, x_2, \dots, x_n\} \quad \mathbf{x} \in \mathbf{R}^n$$

$$\forall \mathbf{x}, \mathbf{y} \in \mathbf{R}^n, \quad \forall \theta \in [0, 1]$$

$$\begin{aligned} f(\theta \mathbf{x} + (1-\theta)\mathbf{y}) &= \max\{\theta x_i + (1-\theta)y_i, i=1, \dots, n\} \\ &\leq \theta \max\{x_i, i=1, \dots, n\} + (1-\theta) \max\{y_i, i=1, \dots, n\} \\ &= \theta f(\mathbf{x}) + (1-\theta)f(\mathbf{y}) \end{aligned}$$

故极大值函数是一个凸函数

例：log-sum-exp 函数

$$f(\mathbf{x}) = \log(e^{x_1} + \dots + e^n) \quad \mathbf{x} \in \mathbf{R}^n$$

log-sum-exp 函数是极大值函数的解析近似，因为

$$\max\{x_1, x_2, \dots, x_n\} \leq \log(e^{x_1} + \dots + e^n) \leq \max\{x_1, x_2, \dots, x_n\} + \log n$$

下面判断此函数的凹凸性

首先求此函数的 Hessian 矩阵

$$\begin{aligned} \frac{\partial f}{\partial x_i} &= \frac{e^{x_i}}{e^{x_1} + \dots + e^{x_n}} \\ \frac{\partial^2 f}{\partial x_i \partial x_j} &= \frac{-e^{x_i} e^{x_j}}{(e^{x_1} + \dots + e^{x_n})^2}, i \neq j \\ \frac{\partial^2 f}{\partial x_i^2} &= \frac{e^{x_i}(e^{x_1} + \dots + e^{x_n}) - e^{2x_i}}{(e^{x_1} + \dots + e^{x_n})^2} \\ \mathbf{H} &= \frac{1}{(e^{x_1} + \dots + e^{x_n})^2} \left\{ \begin{bmatrix} e^{x_1}(e^{x_1} + \dots + e^{x_n}) & & \\ & \ddots & \\ & & e^{x_n}(e^{x_1} + \dots + e^{x_n}) \end{bmatrix} - \begin{pmatrix} e^{x_1} \\ \vdots \\ e^{x_n} \end{pmatrix} \begin{pmatrix} e^{x_1} & \dots & e^{x_n} \end{pmatrix} \right\} \\ \text{令 } \mathbf{z} &= \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix} = \begin{pmatrix} e^{x_1} \\ \vdots \\ e^{x_n} \end{pmatrix}, \quad \mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \quad \text{于是 } \mathbf{H} = \frac{1}{(\mathbf{1}^T \mathbf{z})^2} ((\mathbf{1}^T \mathbf{z}) \text{diag}(\mathbf{z}) - \mathbf{z} \mathbf{z}^T) \end{aligned}$$

然后判定此矩阵的正定性（若 $\forall \mathbf{v} \in \mathbf{R}^n$ ，有 $\mathbf{v}^T \mathbf{H} \mathbf{v} \geq 0$ ，则 $\mathbf{H} \succeq \mathbf{0}$ ）

$$\begin{aligned} &\mathbf{v}^T ((\mathbf{1}^T \mathbf{z}) \text{diag}(\mathbf{z}) - \mathbf{z} \mathbf{z}^T) \mathbf{v} \\ &= (\mathbf{1}^T \mathbf{z}) \mathbf{v}^T \text{diag}(\mathbf{z}) \mathbf{v} - \mathbf{v}^T \mathbf{z} \mathbf{z}^T \mathbf{v} \\ &= \sum_{i=0}^n z_i \sum_{i=0}^n v_i^2 z_i - \left(\sum_{i=0}^n v_i z_i \right)^2 \quad \begin{matrix} a_i = v_i \sqrt{z_i}, b_i = \sqrt{z_i}, a = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}, b = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \end{matrix} \\ &= \mathbf{b}^T \mathbf{b} \mathbf{a}^T \mathbf{a} - (\mathbf{a}^T \mathbf{b})^2 \quad \begin{matrix} \text{Cauch-Schwarz inequality} \\ \geq 0 \end{matrix} \end{aligned}$$

$$\text{因为 } \frac{1}{(\mathbf{1}^T \mathbf{z})^2} > 0, \text{ 故 } \mathbf{v}^T \mathbf{H} \mathbf{v} \geq 0, \text{ 故 } \mathbf{H} \succeq \mathbf{0}$$

综上，log-sum-exp 函数是凸函数

例：几何平均

$$f(\mathbf{x}) = (x_1 x_2 \cdots x_n)^{\frac{1}{n}} \quad \mathbf{x} \in \mathbf{R}_{++}^n$$

首先求此函数的 Hessian 矩阵

$$\frac{\partial f}{\partial x_i} = \frac{1}{n} \frac{\prod_{i=1}^n x_i}{x_i} \left(\prod_{i=1}^n x_i \right)^{\frac{1}{n}-1} = \frac{1}{n} \frac{\left(\prod_{i=1}^n x_i \right)^{\frac{1}{n}}}{x_i}$$

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{1}{n^2} \frac{\left(\prod_{i=1}^n x_i \right)^{\frac{1}{n}}}{x_i x_j}, i \neq j$$

$$\frac{\partial^2 f}{\partial x_i^2} = \left(\frac{1}{n^2} - \frac{1}{n} \right) \frac{\left(\prod_{i=1}^n x_i \right)^{\frac{1}{n}}}{x_i^2}$$

$$\mathbf{H} = -\frac{1}{n^2} \left(\prod_{i=1}^n x_i \right)^{\frac{1}{n}} \left\{ n \begin{bmatrix} \frac{1}{x_1^2} & & \\ & \ddots & \\ & & \frac{1}{x_n^2} \end{bmatrix} - \begin{pmatrix} \frac{1}{x_1} \\ \vdots \\ \frac{1}{x_n} \end{pmatrix} \begin{pmatrix} \frac{1}{x_1} & \cdots & \frac{1}{x_n} \end{pmatrix} \right\}$$

$$\text{令 } \mathbf{z} = \begin{pmatrix} \frac{1}{x_1} \\ \vdots \\ \frac{1}{x_n} \end{pmatrix}, \text{ 于是 } \mathbf{H} = -\frac{1}{n^2} \left(\prod_{i=1}^n x_i \right)^{\frac{1}{n}} \left(n \text{diag}(\mathbf{z}^2) - \mathbf{z} \mathbf{z}^T \right)$$

然后判定此矩阵的正定性（若 $\forall \mathbf{v} \in \mathbf{R}^n$ ，有 $\mathbf{v}^T \mathbf{H} \mathbf{v} \geq 0$ ，则 $\mathbf{H} \succeq \mathbf{0}$ ）

$$\begin{aligned} & \mathbf{v}^T \left(n \text{diag}(\mathbf{z}^2) - \mathbf{z} \mathbf{z}^T \right) \mathbf{v} \\ &= n \mathbf{v}^T \text{diag}(\mathbf{z}^2) \mathbf{v} - \mathbf{v}^T \mathbf{z} \mathbf{z}^T \mathbf{v} \\ &= n \sum_{i=0}^n v_i^2 z_i^2 - \left(\sum_{i=0}^n v_i z_i \right)^2 = n \mathbf{v}^T \mathbf{v} \mathbf{z}^T \mathbf{z} - (\mathbf{v}^T \mathbf{z})^2 \stackrel{\text{Cauch-Schwarz inequality}}{\geq} 0 \end{aligned}$$

因为 $-\frac{1}{n^2} \left(\prod_{i=1}^n x_i \right)^{\frac{1}{n}} < 0$ ，故 $\mathbf{v}^T \mathbf{H} \mathbf{v} \leq 0$ ，故 $\mathbf{H} \preceq \mathbf{0}$

综上，几何平均是凹函数

例：行列式的对数

$$f(X) = \log \det(X) \quad X \in \mathbf{S}_{++}^n$$

对 $\forall X \in \text{dom} f$, $\forall V \in \mathbf{R}^{n \times n}$

设 $g(t) = f(X + tV)$, $\text{dom} g = \{t \mid X + tV \in \mathbf{S}_{++}^n\}$

$$\begin{aligned} g(t) &= f(X + tV) \\ &= \log \det(X + tV) \\ &= \log \det \left(X^{\frac{1}{2}} (I + tX^{-\frac{1}{2}} V X^{-\frac{1}{2}}) X^{\frac{1}{2}} \right) \\ &\stackrel{\det(AB) = \det(A)\det(B)}{=} \log \left(\det(X^{\frac{1}{2}}) \det(I + tX^{-\frac{1}{2}} V X^{-\frac{1}{2}}) \det(X^{\frac{1}{2}}) \right) \\ &= \log \det(X) + \log \det \left(I + tX^{-\frac{1}{2}} V X^{-\frac{1}{2}} \right) \end{aligned}$$

因为 $X + tV \in \mathbf{S}_{++}^n$, $X \in \mathbf{S}_{++}^n$, 所以 $V \in \mathbf{S}^n$

于是 $X^{-\frac{1}{2}} V X^{-\frac{1}{2}} \in \mathbf{S}^n$, 设 $X^{-\frac{1}{2}} V X^{-\frac{1}{2}} \stackrel{SVD}{=} Q \Lambda Q^T$, $Q Q^T = I, \Lambda = \text{diag}(\lambda_i)$

$$\begin{aligned} \det \left(I + tX^{-\frac{1}{2}} V X^{-\frac{1}{2}} \right) &= \det(I + tQ \Lambda Q^T) \\ &= \det(Q(I + t\Lambda)Q^T) \\ &= \det(Q) \det(I + t\Lambda) \det(Q^T) \\ &= \det(Q Q^T) \det(I + t\Lambda) \\ &= \det(I + t\Lambda) \\ &= \prod_{i=1}^n (1 + t\lambda_i) \end{aligned}$$

$$g(t) = \log \det \left(X^{\frac{1}{2}} \right) + \log \det \left(I + tX^{-\frac{1}{2}} V X^{-\frac{1}{2}} \right) = \log \det \left(X^{\frac{1}{2}} \right) + \sum_{i=1}^n \log(1 + t\lambda_i)$$

$$g'(t) = \sum_{i=1}^n \frac{\lambda_i}{1 + t\lambda_i}$$

$$g''(t) = \sum_{i=1}^n \frac{-\lambda_i^2}{(1 + t\lambda_i)^2} \leq 0$$

所以 $g(t)$ 是凹函数

所以行列式的对数是凹函数

保持凸函数凸性的操作

➤ 非负加权求和 (Non-negative weighted sum)

先求值，再做线性变换
变换的是值域

1) 若 f_1, \dots, f_m 为凸函数，则 $f \triangleq \sum_{i=1}^m \omega_i f_i$, $\omega_i \geq 0$ 为凸函数。

2) 若 $f(x, y)$ 对 $\forall y \in A$ 均为 x 的凸函数，则 $g(x) \triangleq \int_{y \in A} \omega(y) f(x, y) dy$, $\omega(y) \geq 0$ 为凸函数。

Note:

分别凸: $f(x, y)$ 对 $\forall y$ 均为 x 的凸函数, $f(x, y)$ 对 $\forall x$ 均为 y 的凸函数; **联合凸**: $f(x, y)$ 对 $\forall \begin{pmatrix} x \\ y \end{pmatrix}$ 为凸函数

分别凸 \nRightarrow 联合凸, 联合凸 \Rightarrow 分别凸

➤ 仿射映射的复合 (Composition with an affine mapping)

先做线性变换，再求值
变换的是定义域

若 $f(x): \mathbf{R}^n \rightarrow \mathbf{R}$ 为凸函数，则 $g(x) \triangleq f(Ax + b)$, $A \in \mathbf{R}^{n \times n}, b \in \mathbf{R}^n, Ax + b \in \text{dom} f$ 为凸函数。

证明:

$\forall x, y \in \text{dom} g, \forall \theta \in [0, 1]$

$$\begin{aligned} g(\theta x + (1-\theta)y) &= f(A(\theta x + (1-\theta)y) + b) \\ &= f(\theta(Ax + b) + (1-\theta)(Ay + b)) \\ &\stackrel{f \text{ is a convex function}}{\leq} \theta f(Ax + b) + (1-\theta)f(Ay + b) \\ &= \theta g(x) + (1-\theta)g(y) \end{aligned}$$

➤ 两个函数的极大值 (Point-wise Maximum)

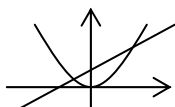
若 f_1, f_2 为凸函数，则 $f(x) \triangleq \max\{f_1(x), f_2(x)\}$, $\text{dom} f = \text{dom} f_1 \cap \text{dom} f_2$ 为凸函数。

证明:

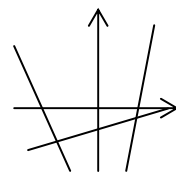
$\forall x, y \in \text{dom} f, \forall \theta \in [0, 1]$

$$\begin{aligned} f(\theta x + (1-\theta)y) &= \max\{f_1(\theta x + (1-\theta)y), f_2(\theta x + (1-\theta)y)\} \\ &\stackrel{f \text{ is a convex function}}{\leq} \max\{\theta f_1(x) + (1-\theta)f_1(y), \theta f_2(x) + (1-\theta)f_2(y)\} \\ &\stackrel{\max\{a+b, c+d\} \leq \max\{a, c\} + \max\{b, d\}}{\leq} \theta \max\{f_1(x), f_2(x)\} + (1-\theta) \max\{f_1(y), f_2(y)\} \\ &= \theta f(x) + (1-\theta)f(y) \end{aligned}$$

例: $f(x) = \max\{x^2, x\}$



例: 分段线性函数 (piece-wise linear function): $f(x) = \max\{a_1^T x + b_1, \dots, a_m^T x + b_m\}$



例: 向量中 r 个最大元素之和 (sum of r largest components)

设 $x \in \mathbf{R}^n$, $x_{[i]}$ 为第 i 大元素，则 $f(x) \triangleq \sum_{i=1}^r x_{[i]} = \max\{x_{i_1} + \dots + x_{i_r} \mid 1 \leq i_1 \leq \dots \leq i_r \leq n\}$

➤ 无限个凸函数的极大值函数 (Point-wise Supremum over an infinite set of convex functions)

若 $f(\mathbf{x}, y)$ 对 $\forall y \in A$ 均为 \mathbf{x} 的凸函数, 则 $g(\mathbf{x}) \triangleq \sup_{y \in A} f(\mathbf{x}, y)$ 为凸函数。

例: 从一点 \mathbf{x} 到集合 C 的最远距离 (distance to the farthest point in a set) : $f(\mathbf{x}) = \sup_{y \in C} \|\mathbf{x} - \mathbf{y}\|$

$\|\mathbf{x} - \mathbf{y}\|$ 是关于 \mathbf{x} 的凸函数

例: 实对称矩阵的最大特征值 (maximum eigenvalue of symmetric matrix) : $\lambda_{\max}(\mathbf{X}) = \sup_{\|\mathbf{y}\|_2=1} \mathbf{y}^T \mathbf{X} \mathbf{y}$

$\mathbf{y}^T \mathbf{X} \mathbf{y}$ 是关于 \mathbf{X} 的凸函数

$$\lambda \mathbf{y} = \mathbf{X} \mathbf{y} \quad \|\mathbf{y}\|_2 = 1 \text{ (normalized eigenvector)}$$

$$\Rightarrow \lambda \mathbf{y}^T \mathbf{y} = \mathbf{y}^T \mathbf{X} \mathbf{y}$$

$$\Rightarrow \lambda = \mathbf{y}^T \mathbf{X} \mathbf{y} \quad \text{eigenvalue} \quad \text{Note: 关于特征值的表达式}$$

➤ 函数的复合 (Composition of functions)

$$h: \mathbf{R}^k \rightarrow \mathbf{R} \\ g: \mathbf{R}^n \rightarrow \mathbf{R}^k, \quad f = h \circ g: \mathbf{R}^n \rightarrow \mathbf{R}, \quad \text{dom} f = \{\mathbf{x} \in \text{dom} g \mid g(\mathbf{x}) \in \text{dom} h\}$$

考虑一维情况, 假定 h, g 的定义域为全空间, 且它们均二阶可微

$$f(x) = h(g(x))$$

$$f'(x) = h'(g(x)) g'(x)$$

$$f''(x) = h''(g(x)) (g'(x))^2 + h'(g(x)) g''(x)$$

若 g 为凸, h 为凸且单增, 则 f 为凸

若 g 为凹, h 为凸且单减, 则 f 为凸

若 g 为凹, h 为凹且单增, 则 f 为凹

若 g 为凸, h 为凹且单减, 则 f 为凹

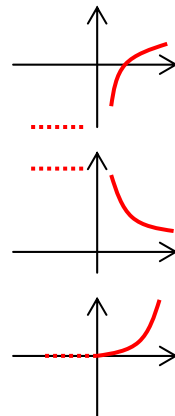
例: 若 g 为凸, 则 $\exp(g(x))$ 为凸

例: 若 g 为凹且 $g > 0$, 则 $\log(g(x))$ 为凹

例: 若 g 为凹且 $g > 0$, 则 $\frac{1}{g(x)}$ 为凸

例: 若 g 为凸且 $g \geq 0$, $p \geq 1$, 则 $g^p(x)$ 为凸

例: $g(x) = x^2$ $\text{dom} g = \mathbf{R}$ convex
 $h(y) = 0$ $\text{dom} h = [1, 2]$ convex, monotone, $f = h \circ g = 0$? convex



No! 因为 $\text{dom} f = [-\sqrt{2}, -1] \cup [1, \sqrt{2}]$ 不是凸集

➤ 函数的透视 (Perspective of a function)

若 $f(\mathbf{x}): \mathbf{R}^n \rightarrow \mathbf{R}$ 是凸函数 $(\mathbf{x}, t, s) \in \text{epi} f \Leftrightarrow tf(\frac{\mathbf{x}}{t}) \leq s$

则 $g(\mathbf{x}, t) = tf(\frac{\mathbf{x}}{t}): \mathbf{R}^n \times \mathbf{R}_{++} \rightarrow \mathbf{R}$, $\text{dom} g = \{(\mathbf{x}, t) \mid \frac{\mathbf{x}}{t} \in \text{dom} f, t > 0\}$ 是凸函数 $\Leftrightarrow f(\frac{\mathbf{x}}{t}) \leq \frac{s}{t}$

$\Leftrightarrow (\frac{\mathbf{x}}{t}, \frac{s}{t}) \in \text{epi} f$

例：欧氏范数的平方 (Euclidean norm squared)

$$f(\mathbf{x}) = \mathbf{x}^T \mathbf{x}, \quad \text{dom} f = \mathbf{R}^n \text{ is convex}$$

$$g(\mathbf{x}, t) = t(\frac{\mathbf{x}}{t})^T (\frac{\mathbf{x}}{t}) = \frac{\mathbf{x}^T \mathbf{x}}{t}, \quad \text{dom} g = \mathbf{R}^n \times \mathbf{R}_{++} \text{ is convex}$$

例：负对数 (Negative logarithm)

$$f(x) = -\log x, \quad \text{dom} f = \mathbf{R}_{++} \text{ is convex}$$

$$g(x, t) = t(-\log \frac{x}{t}) = t \log \frac{t}{x}, \quad \text{dom} g = \mathbf{R}_{++}^2 \text{ is convex}$$

$$g(\mathbf{u}, \mathbf{v}) \triangleq \sum_{i=1}^n u_i \log \frac{u_i}{v_i}, \quad \mathbf{u}, \mathbf{v} \in \mathbf{R}_{++}^n \text{ is convex}$$

KL 散度 / 相对熵 (Kullback-Leibler divergence / relative entropy)

$$D_{KL}(\mathbf{u}, \mathbf{v}) \triangleq \sum_{i=1}^n (u_i \log \frac{u_i}{v_i} - u_i + v_i)$$

Bregman 散度 / 相对熵 (Bregman divergence)

$$D_B(\mathbf{u}, \mathbf{v}) \triangleq f(\mathbf{u}) - f(\mathbf{v}) - \nabla^T f(\mathbf{v})(\mathbf{u} - \mathbf{v})$$

$$1) \text{ 当 } f(\mathbf{u}) = -\sum_{i=1}^n \log u_i$$

$$D_B(\mathbf{u}, \mathbf{v}) = -\sum_{i=1}^n \log u_i + \sum_{i=1}^n \log v_i + \sum_{i=1}^n \frac{1}{v_i} (u_i - v_i) = \sum_{i=1}^n \frac{1}{v_i} (v_i \log \frac{v_i}{u_i} - v_i + u_i)$$

$$2) \text{ 当 } f(\mathbf{u}) = \sum_{i=1}^n u_i \log u_i$$

$$D_B(\mathbf{u}, \mathbf{v}) \stackrel{\frac{\partial f}{\partial u_i} = \log u_i + 1}{=} \sum_{i=1}^n u_i \log u_i - \sum_{i=1}^n v_i \log v_i - \sum_{i=1}^n (u_i - v_i)(\log v_i + 1) = \sum_{i=1}^n (u_i \log \frac{u_i}{v_i} - u_i + v_i)$$

$$3) \text{ 当 } f(\mathbf{u}) = \sum_{i=1}^n u_i \log u_i - \sum_{i=1}^n u_i$$

$$D_B(\mathbf{u}, \mathbf{v}) \stackrel{\frac{\partial f}{\partial u_i} = \log u_i + 1 - 1 = \log u_i}{=} (\sum_{i=1}^n u_i \log u_i - \sum_{i=1}^n u_i) - (\sum_{i=1}^n v_i \log v_i - \sum_{i=1}^n v_i) - \sum_{i=1}^n (u_i - v_i) \log v_i = \sum_{i=1}^n (u_i \log \frac{u_i}{v_i} - u_i + v_i)$$

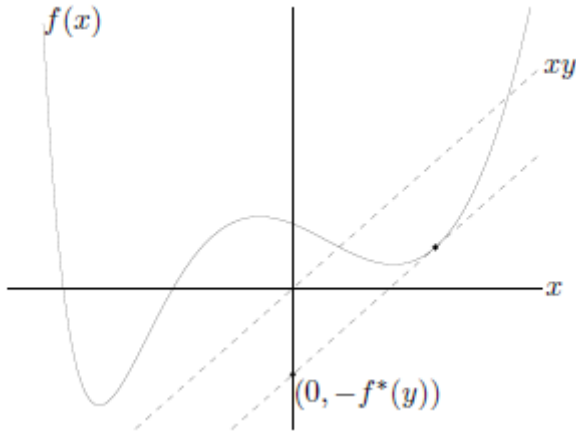
➤ 共轭函数 (The conjugate function)

设 $f(x): \mathbf{R}^n \rightarrow \mathbf{R}$ (不要求是凸函数)

则 $f^*(y) \triangleq \sup_{x \in \text{dom} f} (y^T x - f(x)): \mathbf{R}^n \rightarrow \mathbf{R}$ 是凸函数

Note: $f^*(y)$ 对 y 来说是分段线性函数, 所以 $f^*(y)$ 一定是凸函数

几何意义:



A function $f(x): \mathbf{R} \rightarrow \mathbf{R}$, and a value $y \in \mathbf{R}$.

The conjugate function $f^*(y)$ is the maximum gap between the linear function yx and $f(x)$.

If $f(x)$ is differentiable, this occurs at a point x where $f'(x) = y$.

例: $f(x) = ax + b$, $\text{dom} f = \mathbf{R}$

$$f^*(y) = \sup_{x \in \text{dom} f} (yx - (ax + b)) = \sup_{x \in \text{dom} f} ((y - a)x - b) = \begin{cases} -b & y = a \\ +\infty & y \neq a \end{cases} \text{ 是凸函数}$$

例: $f(x) = -\log x$, $\text{dom} f = \mathbf{R}_{++}$

$$f^*(y) = \sup_{x \in \text{dom} f} (yx + \log x) \stackrel{\because y + \frac{1}{x} = 0 (x > 0) \Rightarrow x = -\frac{1}{y}}{=} \begin{cases} +\infty & y \geq 0 \\ -1 + \log(-\frac{1}{y}) & y < 0 \end{cases}$$

例: $f(x) = \frac{1}{2} x^T Q x$, $Q = S_{++}^n$, $\text{dom} f = \mathbf{R}^n$

$$f^*(y) = \sup_{x \in \text{dom} f} \left(y^T x - \frac{1}{2} x^T Q x \right) \stackrel{y = Qx \Rightarrow x = Q^{-1}y}{=} y^T Q^{-1} y - \frac{1}{2} (Q^{-1} y)^T Q Q^{-1} y = \frac{1}{2} y^T Q^{-1} y$$

可以看出, 此函数共轭的共轭就是它自身 (如同复数一样)。

但并不是所有函数都有此性质。例如, 一个非凸函数经过两次共轭就不能复原, 因为任何函数的共轭一定是凸函数。

➤ **Log-concave 函数**

一个函数 $f: \mathbf{R}^n \rightarrow \mathbf{R}$ 为 Log-concave 函数 \Leftrightarrow

$f(\mathbf{x}) > 0$, 且对 $\forall \mathbf{x} \in \text{dom}f$, $\log f(\mathbf{x})$ 为凹函数

比凹弱

➤ **Log-convex 函数**

一个函数 $f: \mathbf{R}^n \rightarrow \mathbf{R}$ 为 Log-convex 函数 \Leftrightarrow

$f(\mathbf{x}) > 0$, 且对 $\forall \mathbf{x} \in \text{dom}f$, $\log f(\mathbf{x})$ 为凸函数

比凸强

凸集与凸函数的关系

定义:

α 次水平集 (α -sub-level set)

一个函数 $f: \mathbf{R}^n \rightarrow \mathbf{R}$ 的 α -sub-level set 为 $C_\alpha = \{\mathbf{x} \in \text{dom}f \mid f(\mathbf{x}) \leq \alpha\}$

定理:

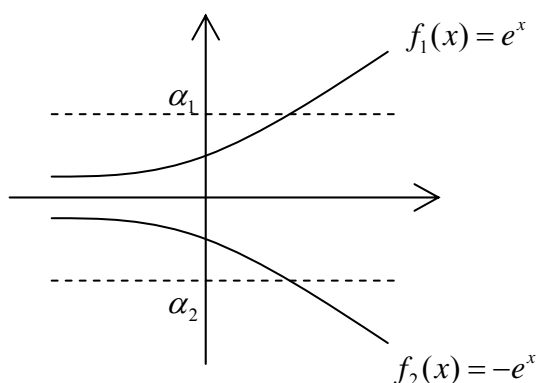
如果 $f: \mathbf{R}^n \rightarrow \mathbf{R}$ 是凸函数 $\Rightarrow f$ 的所有 α -sub-level set 为凸集

证明:

设 $\forall \mathbf{x}, \mathbf{y} \in C_\alpha$, 有 $f(\mathbf{x}) \leq \alpha, f(\mathbf{y}) \leq \alpha, \mathbf{x}, \mathbf{y} \in \text{dom}f$

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

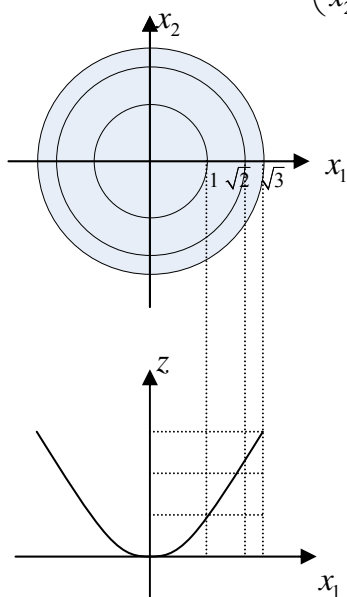
该定理反之则不一定成立



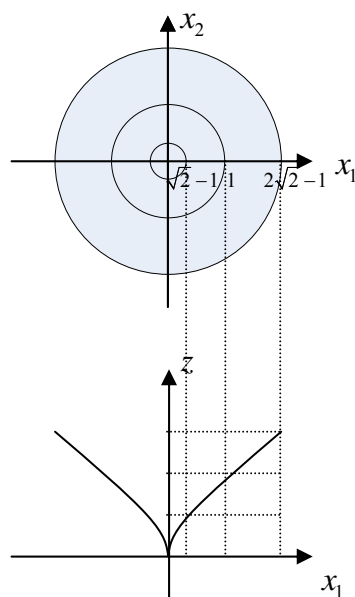
$f_1(x) = e^x$ 的 α_1 -sub-level set 为 $(-\infty, \ln \alpha_1]$ 是凸集, $f_1(x)$ 也是凸函数

$f_2(x) = -e^x$ 的 α_2 -sub-level set 为 $[\ln(-\alpha_2), +\infty)$ 是凸集, $f_2(x)$ 不是凸函数

例: $f(\mathbf{x}) = \|\mathbf{x}\|_2 = \mathbf{x}^T \mathbf{x}, \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$



例: $f(\mathbf{x}) = \log((\mathbf{x}+1)^T(\mathbf{x}+1)), \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$



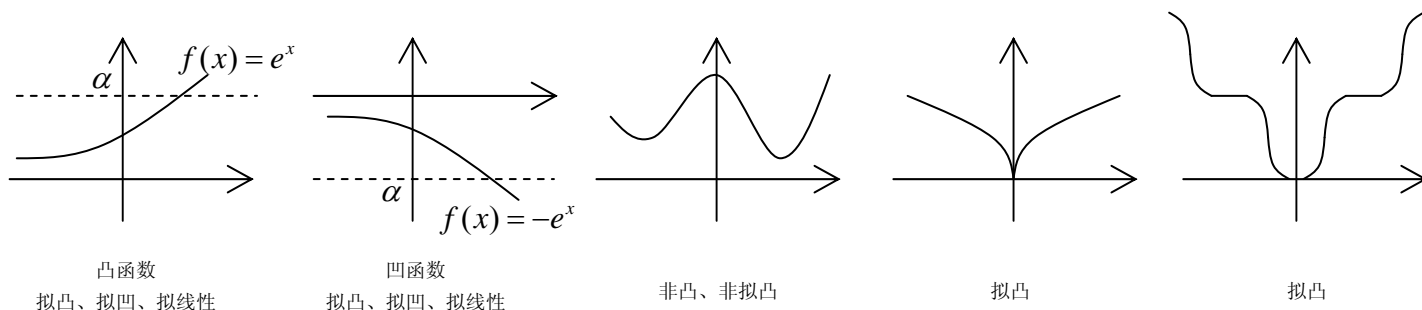
以上两例说明根据某一函数 α -sub-level set (等高线) 的分布可判定该函数的凹凸性

定义 1:

拟凸函数 (Quasi-convex): 一个函数 $f: \mathbf{R}^n \rightarrow \mathbf{R}$ 所有的 α -sub-level set 为 $C_\alpha \triangleq \{\mathbf{x} \in \text{dom}f \mid f(\mathbf{x}) \leq \alpha\}$ 是凸集

拟凹函数 (Quasi-concave): 一个函数 $f: \mathbf{R}^n \rightarrow \mathbf{R}$ 所有的 α -super-level set 为 $C'_\alpha \triangleq \{\mathbf{x} \in \text{dom}f \mid f(\mathbf{x}) \geq \alpha\}$ 是凸集

拟线性函数 (Quasi-linear): 一个函数 $f: \mathbf{R}^n \rightarrow \mathbf{R}$ 所有的 α -level set 为 $C''_\alpha \triangleq \{\mathbf{x} \in \text{dom}f \mid f(\mathbf{x}) = \alpha\}$ 是凸集



例: 向量 $\mathbf{x} \in \mathbf{R}^n$ 的长度 (即最后一个非零元素的位置)

$C_\alpha \triangleq \{\mathbf{x} \mid f(\mathbf{x}) \leq \alpha\} = \{\mathbf{x} \mid f(\mathbf{x}) \leq \lfloor \alpha \rfloor\} = \{\mathbf{x} \mid x_i = 0, i = \lfloor \alpha \rfloor + 1, \dots, n\}$ 是 \mathbf{R}^n 的子空间 (凸集)

所以该函数拟凸

例: 线性分数函数: $f(\mathbf{x}) = \frac{\mathbf{a}^T \mathbf{x} + b}{\mathbf{c}^T \mathbf{x} + d}$ $\text{dom}f = \{\mathbf{x} \mid \mathbf{c}^T \mathbf{x} + d > 0\}$

$C_\alpha \triangleq \left\{ \mathbf{x} \mid \frac{\mathbf{a}^T \mathbf{x} + b}{\mathbf{c}^T \mathbf{x} + d} \leq \alpha, \mathbf{c}^T \mathbf{x} + d > 0 \right\} = \left\{ \mathbf{x} \mid (\mathbf{a} - \alpha \mathbf{c})^T \mathbf{x} \leq \alpha d - b, \mathbf{c}^T \mathbf{x} + d > 0 \right\}$ 是两个半空间的交集 (凸集)

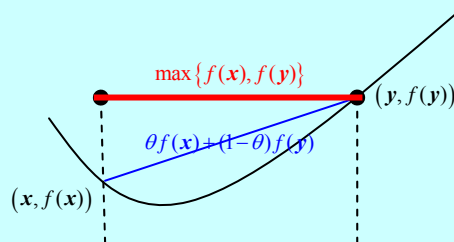
所以该函数拟凸

定义 2:

一个函数 $f: \mathbf{R}^n \rightarrow \mathbf{R}$ 为拟凸函数 \Leftrightarrow

若 $\text{dom}f$ 为凸集, 且对 $\forall \mathbf{x}, \mathbf{y} \in \text{dom}f$, 有

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



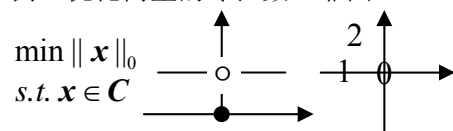
可微拟凸函数的一阶条件:

一个函数 $f: \mathbf{R}^n \rightarrow \mathbf{R}$ 为拟凸函数 $\Leftrightarrow \forall \mathbf{x}, \mathbf{y} \in \text{dom}f$, $f(\mathbf{y}) \leq f(\mathbf{x}) \Rightarrow \nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x}) \leq 0$ (一个逻辑关系)

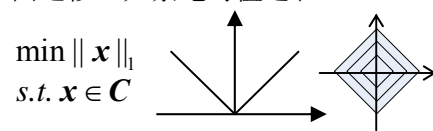
可微拟凸函数的二阶条件:

一个函数 $f: \mathbf{R}^n \rightarrow \mathbf{R}$ 为拟凸函数 $\Leftrightarrow \forall \mathbf{x}, \mathbf{y} \in \text{dom}f$, $\nabla^T f(\mathbf{x})\mathbf{y} = 0 \Rightarrow \mathbf{y}^T \nabla^2 f(\mathbf{x})\mathbf{y} \geq 0$ (一个逻辑关系)

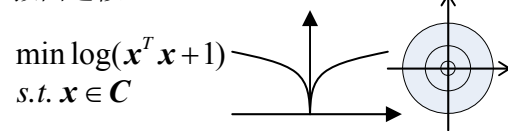
例: 优化向量的零范数 (非凸)



凸近似 (元素绝对值之和)



拟凸近似



Chapter 4 Convex Optimization Problems

优化问题的标准形式 (Optimization Problems in Standard Form)

$$\min f_o(\mathbf{x})$$

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

$$h_j(\mathbf{x}) = 0 \quad j = 1, \dots, p$$

\mathbf{x} : 优化变量 (optimization variable)

f_o : 目标函数/损失函数 (objective function / cost function)

f_i : 不等式约束 (inequality constraint)

h_j : 等式约束 (equality constraint)

转化不等式约束

$$\max f_o(\text{cost})$$

$$\text{s.t.} \quad \text{cost} < \$100$$

\Downarrow

$$\max f_o(\text{cost})$$

$$\text{s.t.} \quad \text{cost} \leq \$99.99$$

域 (Domain)

$$\mathbf{D} \triangleq \bigcap_{i=1}^m \text{dom} f_i \cap \bigcap_{j=1}^p \text{dom} h_j$$

可行解集 (feasible set)

$$\mathbf{X} = \left\{ \mathbf{x} \in \mathbf{D} \left| \begin{array}{ll} f_i(\mathbf{x}) \leq 0 & i = 1, \dots, m \\ h_j(\mathbf{x}) = 0 & j = 1, \dots, p \end{array} \right. \right\}$$

最优值 (optimal value)

$$p^* = \inf \{ f_o(\mathbf{x}) \mid \mathbf{x} \in \mathbf{X} \}$$

最优解集 (optimal set)

$$\mathbf{X}^* = \{ \mathbf{x}^* \in \mathbf{X} \mid f_o(\mathbf{x}^*) = p^* \}$$

ε 次优解 (ε -suboptimal set)

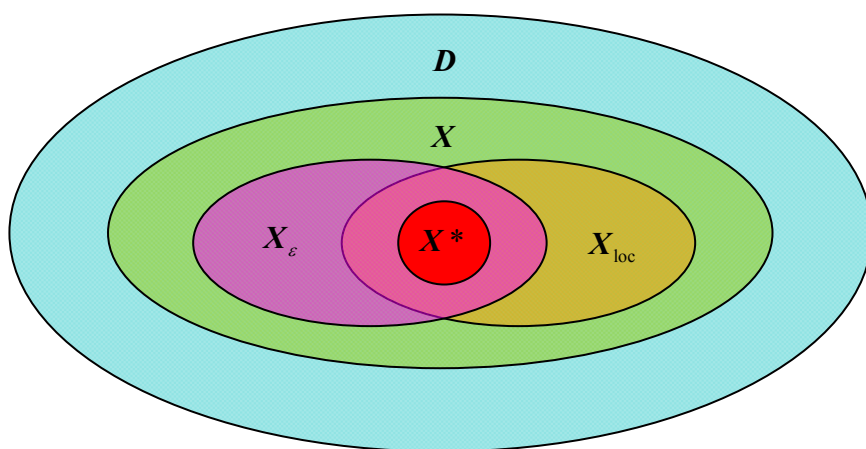
$$\mathbf{X}_\varepsilon = \{ \mathbf{x} \in \mathbf{X} \mid f_o(\mathbf{x}) = p^* + \varepsilon \}$$

局部最优值 (local optimal value)

$$\exists R > 0, \quad p_{\text{loc}} = \inf \{ f_o(\mathbf{z}) \mid \|\mathbf{z} - \mathbf{x}\|_2 \leq R, \mathbf{x} \in \mathbf{X}, \mathbf{z} \in \mathbf{X} \}$$

局部最优解集 (local optimal set)

$$\mathbf{X}_{\text{loc}} = \{ \mathbf{x}_{\text{loc}} \in \mathbf{X} \mid f_o(\mathbf{x}_{\text{loc}}) = p_{\text{loc}} \}$$



凸优化问题 (Convex Optimization Problems)

$$\begin{aligned} \min \quad & f_o(\mathbf{x}) \\ \text{s.t.} \quad & f_i(\mathbf{x}) \leq 0 \quad i = 1, \dots, m \\ & \mathbf{A}_j^T \mathbf{x} = \mathbf{b}_j \quad j = 1, \dots, p \end{aligned}$$

1) 目标函数 $f_o(\mathbf{x})$ 为凸函数

2) 不等式约束 $f_1(\mathbf{x}), \dots, f_m(\mathbf{x})$ 为凸函数 (Note: 凸函数 $f_i(\mathbf{x}) \leq 0$ 的解集一定是凸集)

3) 等式约束 $\mathbf{A}_1^T \mathbf{x} = \mathbf{b}_1, \dots, \mathbf{A}_p^T \mathbf{x} = \mathbf{b}_p$ 为仿射函数 (Note: 仿射函数 $\mathbf{A}_j^T \mathbf{x} = \mathbf{b}_j$ 的解集一定是凸集)

非凸优化问题转化为凸优化问题

例:

$$\begin{aligned} \min \quad & f_o(\mathbf{x}) = x_1^2 + x_2^2 \quad \text{convex} \\ \text{s.t.} \quad & f_1(\mathbf{x}) = \frac{x_1}{1+x_2^2} \leq 0 \quad \text{nonconvex} \\ & h_1(\mathbf{x}) = (x_1 + x_2)^2 = 0 \quad \text{nonlinear} \end{aligned}$$

转化为

$$\begin{aligned} \min \quad & f_o(\mathbf{x}) = x_1^2 + x_2^2 \quad \text{convex} \\ \text{s.t.} \quad & x_1 \leq 0 \quad \text{convex} \\ & x_1 + x_2 = 0 \quad \text{linear} \end{aligned}$$

凸优化问题的一个重要性质

局部最优=全局最优

证明: 反证法

令 \mathbf{x}_{loc} 是局部最优解, 则 $\exists R > 0$, 当 $\|\mathbf{z} - \mathbf{x}_{\text{loc}}\|_2 \leq R, \mathbf{z} \in X$, 有 $f_o(\mathbf{y}) \geq f_o(\mathbf{x})$

假设 \mathbf{x}_{loc} 不是全局最优解, 则 $\exists \mathbf{y} \in X$, 当 $\|\mathbf{y} - \mathbf{x}_{\text{loc}}\|_2 > R, \mathbf{y} \in X$ 时, $f_o(\mathbf{y}) < f_o(\mathbf{x})$

考虑 $\mathbf{z} = (1-\theta)\mathbf{x}_{\text{loc}} + \theta\mathbf{y}$, 其中 $\theta = \frac{R}{2\|\mathbf{y} - \mathbf{x}_{\text{loc}}\|_2} \in (0,1)$

因为 $\|\mathbf{z} - \mathbf{x}_{\text{loc}}\|_2 = \|\theta\mathbf{y} - \theta\mathbf{x}_{\text{loc}}\|_2 = \theta\|\mathbf{y} - \mathbf{x}_{\text{loc}}\|_2 = \frac{R}{2}$

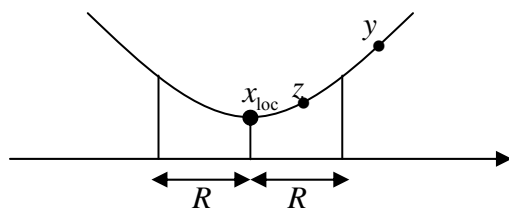
所以 $f_o(\mathbf{z}) \geq f_o(\mathbf{x}_{\text{loc}})$

再由凸函数的性质

$$f_o(\mathbf{z}) \leq (1-\theta)f_o(\mathbf{x}_{\text{loc}}) + \theta f_o(\mathbf{y}) < f_o(\mathbf{x}_{\text{loc}})$$

矛盾, 假设不成立

于是局部最优解 \mathbf{x}_{loc} 必定是全局最优解 \mathbf{x}^*



可微目标函数下最优解的性质 (optimality criterion for differentiable f_o)

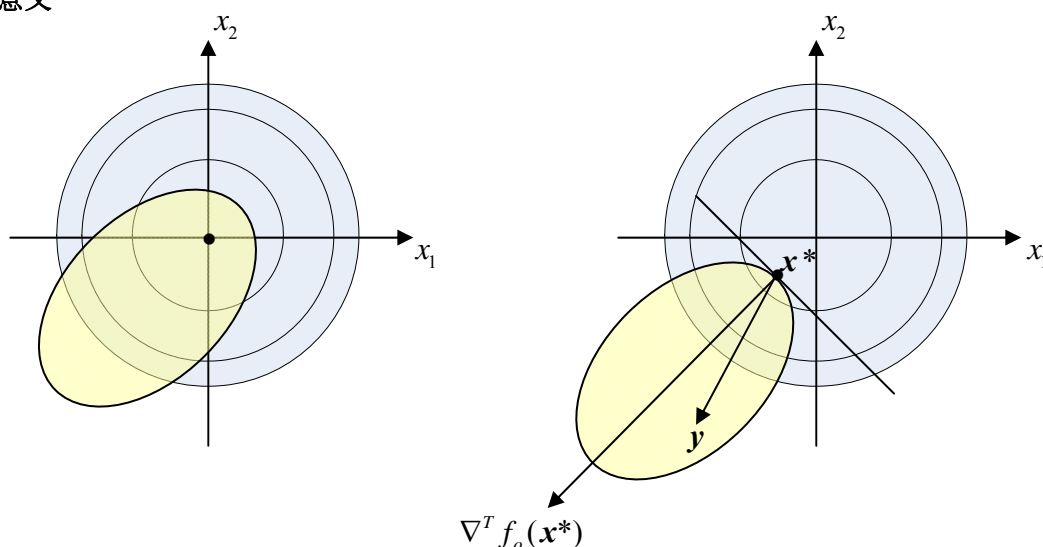
若 f_o 为可微且为凸函数, 则 $f_o(y) \geq f_o(x) + \nabla^T f_o(x)(y-x)$, $\forall x, y \in \text{dom} f$

对无约束问题 $\begin{cases} \min & f_o(x) \\ \text{s.t.} & x \in \text{dom} f_o \end{cases}$ 其最优解 x^* 满足 $\nabla^T f_o(x^*)(y-x^*) \geq 0, \forall y \in \text{dom} f_o$

同理

对有约束问题 $\begin{cases} \min & f_o(x) \\ \text{s.t.} & x \in \text{dom} f_o \cap X \end{cases}$ 其最优解 x^* 满足 $\nabla^T f_o(x^*)(y-x^*) \geq 0, \forall y \in \text{dom} f_o \cap X$
 \downarrow
 可行解集

几何意义



$$\nabla^T f_o(x^*) = 0$$

$$\nabla^T f_o(x^*)(y-x^*) = 0, \forall y \in \text{dom} f_o$$

The angle between $\nabla^T f_o(x^*)$ and $(y-x^*)$ is acute

$$\nabla^T f_o(x^*)(y-x^*) > 0, \forall y \in \text{dom} f_o$$

例: 等式约束优化问题 (Problems with equality constraints only)

$$\begin{cases} \min & f_o(x) \\ \text{s.t.} & Ax = b \end{cases}$$

设 x^* 是最优解, 有 $Ax^* = b$

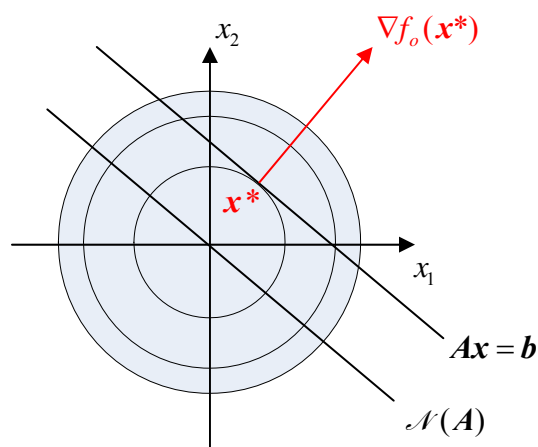
由最优解条件, 对 $\forall y \in \{y \mid Ay = b\}$, 有 $\nabla^T f_o(x^*)(y-x^*) \geq 0$

$$\text{因 } \begin{cases} Ax^* = b \\ Ay = b \end{cases} \Rightarrow y = x^* + v, \quad v \in \mathcal{N}(A) \longrightarrow \text{即满足 } Av = 0$$

故最优解条件可转换为, 对 $\forall v \in \mathcal{N}(A)$, 有 $\nabla^T f_o(x^*)v \geq 0$

由 v 的任意性, 知 $\nabla^T f_o(x^*)v = 0$

亦即 $\nabla f_o(x^*) \perp \mathcal{N}(A)$



例：非负象限约束优化问题（Minimization over the nonnegative orthant）

$$\begin{cases} \min & f_o(\mathbf{x}) \\ \text{s.t.} & \mathbf{x} \succeq \mathbf{0} \end{cases}$$

设 \mathbf{x}^* 是最优解，有 $\mathbf{x}^* \succeq \mathbf{0}$ (1)

由最优解条件，有 $\forall \mathbf{y} \succeq \mathbf{0}$ ， $\nabla^T f_o(\mathbf{x}^*)(\mathbf{y} - \mathbf{x}^*) \geq 0 \longrightarrow \nabla^T f_o(\mathbf{x}^*)\mathbf{y} - \nabla^T f_o(\mathbf{x}^*)\mathbf{x}^* \geq 0$

由 \mathbf{y} 的任意性，上述条件可转换为 $\begin{cases} \nabla f_o(\mathbf{x}^*) \succeq 0 & (2) \\ \nabla^T f_o(\mathbf{x}^*)\mathbf{x}^* \leq 0 & (3) \end{cases}$

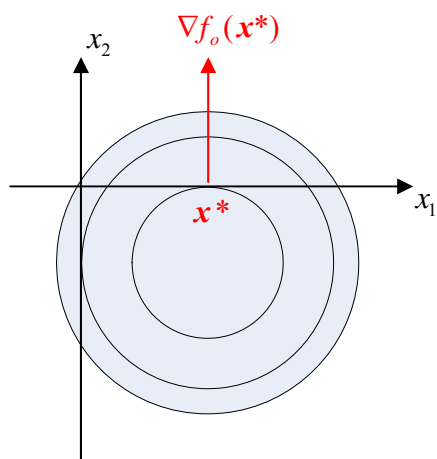
再由条件 (1) $\mathbf{x}^* \succeq \mathbf{0}$ 和 (2) $\nabla f_o(\mathbf{x}^*) \succeq 0$ ，条件 (3) 可进一步化为 $\nabla^T f_o(\mathbf{x}^*)\mathbf{x}^* = 0$ (3-b)

条件 (3-b) 称作互补性 (complementarity)，因为

$$\nabla^T f_o(\mathbf{x}^*)\mathbf{x}^* = 0 \Rightarrow \left(\nabla f_o(\mathbf{x}^*) \right)_i x_i^* = 0 \Rightarrow \begin{cases} \left(\nabla f_o(\mathbf{x}^*) \right)_i = 0 & x_i^* \neq 0 \\ \left(\nabla f_o(\mathbf{x}^*) \right)_i \neq 0 & x_i^* = 0 \end{cases} \quad (3-c)$$

再由条件 (1) $\mathbf{x}^* \succeq \mathbf{0}$ 和 (2) $\nabla f_o(\mathbf{x}^*) \succeq 0$ ，条件 (3-c) 可写为

$$\begin{cases} \left(\nabla f_o(\mathbf{x}^*) \right)_i = 0 & x_i^* \geq 0 \\ \left(\nabla f_o(\mathbf{x}^*) \right)_i \geq 0 & x_i^* = 0 \end{cases}$$



$$\begin{cases} \left(\nabla f_o(\mathbf{x}^*) \right)_1 = 0 & x_1^* \geq 0 \\ \left(\nabla f_o(\mathbf{x}^*) \right)_2 \geq 0 & x_2^* = 0 \end{cases}$$

线性规划 (Linear Programming / LP)

Linear Program

Linear Programming Problem

$$\left\{ \begin{array}{ll} \min & \mathbf{c}^T \mathbf{x} + d \\ \text{s.t.} & \mathbf{G}\mathbf{x} \preceq \mathbf{h} \\ & \mathbf{A}\mathbf{x} = \mathbf{b} \end{array} \right\} \longrightarrow P: \text{多面体}$$

Kantorovich(康托洛维奇)

Dantzig (丹齐格)

Tobias Dantzig, a Baltic German mathematician and linguist

Henry Poincare Dantzig

George Bernard Dantzig

Simplex algorithm (单纯形法)

John von Neumann

线性规划的等价问题 (引入松弛变量——Slack Variable)

$$\left\{ \begin{array}{ll} \min_{\mathbf{x}, \mathbf{s}} & \mathbf{c}^T \mathbf{x} + d \\ \text{s.t.} & \mathbf{G}\mathbf{x} + \mathbf{s} = \mathbf{h} \\ & \mathbf{A}\mathbf{x} = \mathbf{b} \\ & \mathbf{s} \succeq \mathbf{0} \end{array} \right\} \quad \left\{ \begin{array}{ll} \min_{\mathbf{x}^+, \mathbf{x}^-, \mathbf{s}} & \mathbf{c}^T \mathbf{x}^+ - \mathbf{c}^T \mathbf{x}^- + d \\ \text{s.t.} & \mathbf{G}\mathbf{x}^+ - \mathbf{G}\mathbf{x}^- + \mathbf{s} = \mathbf{h} \\ & \mathbf{A}\mathbf{x}^+ - \mathbf{A}\mathbf{x}^- = \mathbf{b} \\ & \mathbf{x}^+ \succeq \mathbf{0}, \mathbf{x}^- \succeq \mathbf{0}, \mathbf{s} \succeq \mathbf{0} \end{array} \right\}$$

松弛变量 $(\mathbf{x}^*, \mathbf{s}^*)$

松弛变量 $((\mathbf{x}^+)^*, (\mathbf{x}^-)^*, \mathbf{s}^*)$

★ 证明两个问题等价的方法:

优化变量之间存在映射, 使解可行 (feasible), 且使目标函数值一样 (same objective value)

★ Matlab 里线性规划的命令: **linprog**

例: 食谱问题

一个健康食谱所含 m 种营养元素的最小量分别为 b_1, \dots, b_m

现有 n 种食物, 单位数量的第 j 种食物所含的营养元素分别为 a_{j1}, \dots, a_{jm} ,

第 j 种食物价格为 c_j , 数量为 x_j

如何合理地确定这 n 种食物的数量, 使价格最小且营养充分

$$\left\{ \begin{array}{ll} \min & \sum_{j=1}^n c_j x_j \\ \text{s.t.} & \sum_{j=1}^n a_{ij} x_j \geq b_i, \quad i = 1, \dots, m \\ & x_j \geq 0, \quad j = 1, \dots, n \end{array} \right.$$

例：线性分数规划

$$\begin{cases} \min & f_o(\mathbf{x}) \\ \text{s.t.} & \mathbf{G}\mathbf{x} \preceq \mathbf{h} \\ & \mathbf{A}\mathbf{x} = \mathbf{b} \end{cases}$$

其中 $f_o(\mathbf{x}) = \frac{\mathbf{c}^T \mathbf{x} + d}{\mathbf{e}^T \mathbf{x} + f}$, $\text{dom} f = \{\mathbf{x} \mid \mathbf{e}^T \mathbf{x} + f > 0\}$ Note: 目标函数是拟凸函数

将拟凸优化 P0 转化为线性优化 P1

$$P0: \begin{cases} \min & \frac{\mathbf{c}^T \mathbf{x} + d}{\mathbf{e}^T \mathbf{x} + f} \\ \text{s.t.} & \mathbf{G}\mathbf{x} \preceq \mathbf{h} \\ & \mathbf{A}\mathbf{x} = \mathbf{b} \\ & \mathbf{e}^T \mathbf{x} + f > 0 \end{cases} \Rightarrow P1: \begin{cases} \min_{\mathbf{y}, z} & \mathbf{c}^T \mathbf{y} + dz \\ \text{s.t.} & \mathbf{G}\mathbf{y} - \mathbf{h}z \preceq \mathbf{0} \\ & \mathbf{A}\mathbf{y} - \mathbf{b}z = \mathbf{0} \\ & \mathbf{e}^T \mathbf{y} + fz = 1 \\ & z \geq 0 \end{cases}$$

证明：

(1) 若 \mathbf{x} 在 P0 中可行，则 $\mathbf{y} = \frac{\mathbf{x}}{\mathbf{e}^T \mathbf{x} + f}, z = \frac{1}{\mathbf{e}^T \mathbf{x} + f}$ 在 P1 中可行，且目标函数值相同

(2) 若 (\mathbf{y}, z) 在 P1 中可行

(2-a) 且 $z \neq 0$ ，则 $\mathbf{x} = \frac{\mathbf{y}}{z}$ 在 P0 中可行，且目标函数值相同

(2-b) 且 $z = 0$ ，此时 P1 退化为

$$P1: \begin{cases} \min_{\mathbf{y}, z} & \mathbf{c}^T \mathbf{y} \\ \text{s.t.} & \mathbf{G}\mathbf{y} \preceq \mathbf{0} \\ & \mathbf{A}\mathbf{y} = \mathbf{0} \\ & \mathbf{e}^T \mathbf{y} = 1 \end{cases}$$

设 \mathbf{x}_0 为 P0 的可行解，必有 $\forall t \geq 0$ ， $\mathbf{x} = \mathbf{x}_0 + t\mathbf{y}$ 在 P0 中可行

$$\text{因为} \begin{cases} \mathbf{G}\mathbf{y} \preceq \mathbf{0} \\ \mathbf{A}\mathbf{y} = \mathbf{0} \\ \mathbf{e}^T \mathbf{y} = 1 \end{cases} + \begin{cases} \mathbf{G}\mathbf{x}_0 \preceq \mathbf{h} \\ \mathbf{A}\mathbf{x}_0 = \mathbf{b} \\ \mathbf{e}^T \mathbf{x}_0 + f > 0 \end{cases} \Rightarrow \begin{cases} \mathbf{G}\mathbf{x}_0 + t\mathbf{G}\mathbf{y} \preceq \mathbf{h} \\ \mathbf{A}\mathbf{x}_0 + t\mathbf{A}\mathbf{y} = \mathbf{b} \\ \mathbf{e}^T \mathbf{x}_0 + t\mathbf{e}^T \mathbf{y} + f > 0 \end{cases} \Rightarrow \begin{cases} \mathbf{G}\mathbf{x} \preceq \mathbf{h} \\ \mathbf{A}\mathbf{x} = \mathbf{b} \\ \mathbf{e}^T \mathbf{x} + f > 0 \end{cases}$$

$$\text{而} \frac{\mathbf{c}^T \mathbf{x} + d}{\mathbf{e}^T \mathbf{x} + f} = \frac{\mathbf{c}^T \mathbf{x}_0 + t\mathbf{c}^T \mathbf{y} + d}{\mathbf{e}^T \mathbf{x}_0 + t\mathbf{e}^T \mathbf{y} + f} \xrightarrow{t \rightarrow \infty} \mathbf{c}^T \mathbf{y}, \text{ 即目标函数相同}$$

综上，P0 与 P1 等价

二次规划 (Quadratic Programming / QP)

$$\begin{cases} \min & \frac{1}{2} \mathbf{x}^T \mathbf{P} \mathbf{x} + \mathbf{q}^T \mathbf{x} + r \\ \text{s.t.} & \mathbf{G} \mathbf{x} \preceq \mathbf{h} \\ & \mathbf{A} \mathbf{x} = \mathbf{b} \end{cases}$$

这里只考虑凸优化问题，故要求 $\mathbf{P} \succ \mathbf{0}$ ($\mathbf{P} = \mathbf{0}$ 则退化为线性规划问题)

★ Matlab 里二次规划的命令: **quadprog**

二次约束二次规划 (Quadratically Constrained Quadratic Programming / QCQP)

$$\begin{cases} \min & \frac{1}{2} \mathbf{x}^T \mathbf{P} \mathbf{x} + \mathbf{q}^T \mathbf{x} + r \\ \text{s.t.} & \frac{1}{2} \mathbf{x}^T \mathbf{P}_i \mathbf{x} + \mathbf{q}_i^T \mathbf{x} + r_i \leq 0, i = 1, \dots, m \\ & \mathbf{A} \mathbf{x} = \mathbf{b} \end{cases}$$

这里要求 $\mathbf{P} \succ \mathbf{0}$, $\mathbf{P}_i \succ \mathbf{0}, i = 1, \dots, m$

例: 线性测量方程—— $\mathbf{b} = \mathbf{A} \mathbf{x} + \boldsymbol{\rho}$ ($\boldsymbol{\rho}$ 为误差项)

1) 选择合适的 \mathbf{x}^* 使误差项最小, 即

$$\begin{aligned} \min \quad & \|\mathbf{b} - \mathbf{A} \mathbf{x}\|_2^2 \\ \min \quad & \mathbf{x}^T \mathbf{A}^T \mathbf{A} \mathbf{x} - 2 \mathbf{b}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{b} \end{aligned} \quad \text{——QP}$$

2-1) 一范数规范化 (Regularization with L1 norm) 优化问题

$$\min_{\mathbf{x}^+, \mathbf{x}^-} \|\mathbf{b} - \mathbf{A}(\mathbf{x}^+ - \mathbf{x}^-)\|_2^2 + \lambda_1 (\mathbf{1}^T \mathbf{x}^+ + \mathbf{1}^T \mathbf{x}^-) \quad \xrightarrow{\text{convert to QP}} \begin{cases} \min_{\mathbf{x}^+, \mathbf{x}^-} & \|\mathbf{b} - \mathbf{A}(\mathbf{x}^+ - \mathbf{x}^-)\|_2^2 + \lambda_1 (\mathbf{1}^T \mathbf{x}^+ + \mathbf{1}^T \mathbf{x}^-) \\ \text{s.t.} & \mathbf{x}^+ \succeq \mathbf{0}, \mathbf{x}^- \succeq \mathbf{0} \end{cases}$$

2-2) 一范数约束优化问题

$$\begin{cases} \min & \|\mathbf{b} - \mathbf{A} \mathbf{x}\|_2^2 \\ \text{s.t.} & \|\mathbf{x}\|_1 \leq \varepsilon_1 \end{cases} \quad \xrightarrow{\text{convert to QP}} \begin{cases} \min_{\mathbf{x}^+, \mathbf{x}^-} & \|\mathbf{b} - \mathbf{A}(\mathbf{x}^+ - \mathbf{x}^-)\|_2^2 \\ \text{s.t.} & \mathbf{1}^T \mathbf{x}^+ + \mathbf{1}^T \mathbf{x}^- \leq \varepsilon_1 \\ & \mathbf{x}^+ \succeq \mathbf{0}, \mathbf{x}^- \succeq \mathbf{0} \end{cases}$$

3-1) 二范数规范化 (Regularization with Euclidean norm) 优化问题

$$\min \|\mathbf{b} - \mathbf{A} \mathbf{x}\|_2^2 + \lambda_2 \|\mathbf{x}\|_2^2 \quad \text{——QP (岭回归: Ridge Regression)}$$

3-2) 二范数约束优化问题

$$\begin{cases} \min & \|\mathbf{b} - \mathbf{A} \mathbf{x}\|_2^2 \\ \text{s.t.} & \|\mathbf{x}\|_2^2 \leq \varepsilon_2 \end{cases} \quad \text{——QCQP}$$

例：投资组合问题（Risk/Return Trade-off in Portfolio Optimization）

asset initial investment overall return

#1	x_1	p_1x_1
\vdots	\vdots	\vdots
#n	x_n	p_nx_n

$$\begin{cases} \max & p_1x_1 + \cdots + p_nx_n \\ \text{s.t.} & x_1 + \cdots + x_n = B \\ & x_1, \cdots, x_n \geq 0 \end{cases}$$

$$\begin{cases} \max & \mathbf{p}^T \mathbf{x} \\ \text{s.t.} & \mathbf{1}^T \mathbf{x} = B, \text{ 其中 } \mathbf{p} = \begin{pmatrix} p_1 \\ \vdots \\ p_n \end{pmatrix}, \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \\ & \mathbf{x} \succeq 0 \end{cases} \quad \text{---LP}$$

Markowitz (马科维茨) portfolio optimization

We take a **stochastic model** (随机模型) for price changes: $\mathbf{p} \in \mathbf{R}^n$ is a **random vector** (随机变量),
with known mean $\bar{\mathbf{p}}$ and covariance Σ .

Therefore with portfolio $\mathbf{x} \in \mathbf{R}^n$, the return r is a (scalar) random variable with **mean** $\bar{\mathbf{p}}^T \mathbf{x}$ and **variance** $\mathbf{x}^T \Sigma \mathbf{x}$.

The choice of portfolio \mathbf{x} involves a trade-off between the **mean** of the return, and its **variance**.

$$\begin{cases} \min & risk \\ \text{s.t.} & Income \geq r_{\min} \\ & Resources \end{cases} \quad \begin{cases} \min & \mathbf{x}^T \Sigma \mathbf{x} \\ \text{s.t.} & \bar{\mathbf{p}}^T \mathbf{x} \geq r_{\min} \\ & \mathbf{1}^T \mathbf{x} = B \\ & \mathbf{x} \succeq 0 \end{cases} \quad \text{---QP}$$

半定规划 (Semi-Definite Programming / SDP)

$$\begin{cases} \max & \mathbf{c}^T \mathbf{x} \\ \text{s.t.} & x_1 \mathbf{F}_1 + x_2 \mathbf{F}_2 + \cdots + x_n \mathbf{F}_n + \mathbf{G} \preceq \mathbf{0}, \text{ with } \mathbf{F}_i, \mathbf{G} \in \mathcal{S}^k \\ & \mathbf{A}\mathbf{x} = \mathbf{b} \end{cases}$$

不等式约束为 **LMI** (线性矩阵不等式)

如果 \mathbf{F}_i, \mathbf{G} 均是对角阵, 则线性矩阵不等式约束 (LMI Constraint) 退化为线性不等式约束 (LI Constraint)

例:

$$\begin{cases} \min & \text{tr}(\mathbf{C}\mathbf{X}) \\ \text{s.t.} & \text{tr}(\mathbf{A}_i \mathbf{X}) = \mathbf{b}_i, \quad i = 1, \dots, p, \text{ with } \mathbf{X} \in \mathcal{S}^n \\ & \mathbf{X} \succeq \mathbf{0} \end{cases} \quad \text{---SDP}$$

若 $\mathbf{C}, \mathbf{X}, \mathbf{A}_i$ 均为对角阵

记 $\mathbf{c} = (\text{diag}\{\mathbf{C}\})$, $\mathbf{x} = (\text{diag}\{\mathbf{X}\})$, $\mathbf{a}_i = (\text{diag}\{\mathbf{A}_i\})$ 是相应对角阵的对角元素构成的向量

而对角阵乘积的迹=对角阵元素乘积之和, 即

$$\text{tr}(\mathbf{C}\mathbf{X}) = \text{tr}(\text{diag}\{\mathbf{C}\} \text{diag}\{\mathbf{X}\}) = (\text{diag}\{\mathbf{C}\})^T (\text{diag}\{\mathbf{X}\}) = \mathbf{c}^T \mathbf{x}$$

$$\text{tr}(\mathbf{A}_i \mathbf{X}) = \text{tr}(\text{diag}\{\mathbf{A}_i\} \text{diag}\{\mathbf{X}\}) = (\text{diag}\{\mathbf{A}_i\})^T (\text{diag}\{\mathbf{X}\}) = \mathbf{a}_i^T \mathbf{x}$$

原 SDP 问题就转化为 LP 问题, 即

$$\begin{cases} \min & \mathbf{c}^T \mathbf{x} \\ \text{s.t.} & \mathbf{a}_i^T \mathbf{x} = \mathbf{b}_i \\ & \mathbf{x} \succeq \mathbf{0} \end{cases} \quad \text{---LP}$$

例: 矩阵的谱范数定义为 $\|\mathbf{A}\|_2 = \sqrt{\max_i \lambda_i(\mathbf{A}^H \mathbf{A})}$

要求 优化 矩阵函数: $\mathbf{A}(\mathbf{x}) = \mathbf{A}_0 + x_1 \mathbf{A}_1 + \cdots + x_n \mathbf{A}_n$ 的谱范数²

$$P0: \min_{\mathbf{x}} \quad \|\mathbf{A}(\mathbf{x})\|_2^2$$

$$P1: \begin{cases} \min_{\mathbf{x}, s} & s \\ \text{s.t.} & \mathbf{A}^T(\mathbf{x}) \mathbf{A}(\mathbf{x}) \preceq s \mathbf{I} \end{cases}, \text{ with } s > 0$$

$$P2: \begin{cases} \min_{\mathbf{x}, t} & t \\ \text{s.t.} & \begin{pmatrix} t \mathbf{I} & \mathbf{A}(\mathbf{x}) \\ \mathbf{A}^T(\mathbf{x}) & t \mathbf{I} \end{pmatrix} \succeq \mathbf{0} \end{cases}, \text{ with } t > 0$$

例：Fastest mixing Markov chain problem

定义一个无向图 (undirected graph)

结点 (nodes): $1, \dots, n$; 边 (edges): $\mathcal{E} \subseteq \{1, \dots, n\} \times \{1, \dots, n\}$

由于是无向图, 故 \mathcal{E} 应是对称的

定义一个 Markov 链

t 时刻的状态 (state): $X(t) \in \{1, \dots, n\}$, $t \in \mathbb{Z}_+$

状态 i 与状态 j 之间的转移概率 (transition probability): $P_{ij} = \text{prob}\{X(t+1) = j \mid X(t) = i\}$

记转移概率矩阵 (transition probability matrix) 为 $\mathbf{P} = \begin{pmatrix} P_{11} & \cdots & P_{1n} \\ \vdots & \ddots & \vdots \\ P_{m1} & \cdots & P_{mn} \end{pmatrix}$

该矩阵满足

1) $P_{ij} = 0$, for $(i, j) \notin \mathcal{E}$

2) $P_{ij} \geq 0$, for $(i, j) \in \mathcal{E}$

3) $\sum_j P_{ij} = 1$ (或 $\mathbf{P} \cdot \mathbf{1} = \mathbf{1}$, 即行和为 1)

4) $\mathbf{P} = \mathbf{P}^T$ (无向的)

Convergence of the distribution of $X(t)$ to $\frac{1}{n}\mathbf{1}$ is determined by the second largest (in magnitude) eigenvalue of \mathbf{P} ,

i.e., by $r = \max\{\lambda_2, -\lambda_n\}$, where $1 = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq -1$ are the eigenvalues of \mathbf{P} .

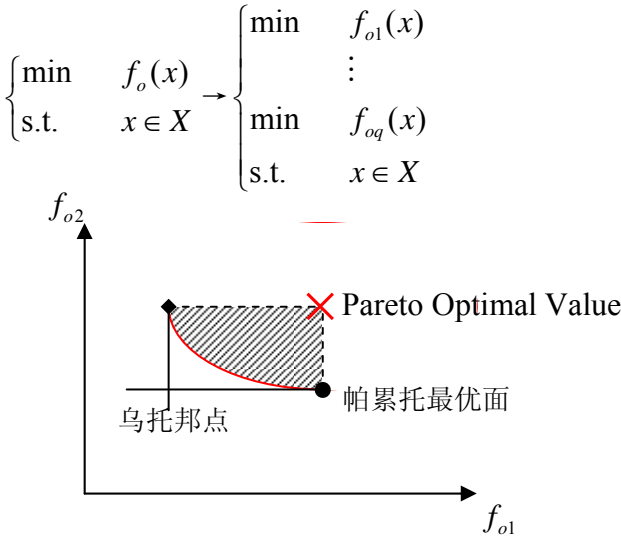
$$\min \quad \left\| \mathbf{P} - \frac{1}{n} \mathbf{1} \mathbf{1}^T \right\|_2$$

$$\text{s.t.} \quad \mathbf{P} \mathbf{1} = \mathbf{1}$$

$$P_{ij} \geq 0, \text{ for } (i, j) \in \mathcal{E}$$

$$P_{ij} = 0, \text{ for } (i, j) \notin \mathcal{E}$$

多目标优化



求解多目标优化问题（对各目标进行非负加权）

$$\begin{cases} \min & \lambda_1 f_{o1}(x) + \cdots + \lambda_q f_{oq}(x) \\ \text{s.t.} & x \in X \end{cases}$$

Chapter 5 Duality

标准最优化问题

$$\begin{array}{ll}\min & f_o(\mathbf{x}) \\ \text{s.t.} & f_i(\mathbf{x}) \leq 0 \quad i=1, \dots, m \\ & h_j(\mathbf{x}) = 0 \quad j=1, \dots, p\end{array}$$

域 (Domain): $D \triangleq \bigcap_{i=1}^m \text{dom} f_i \cap \bigcap_{j=1}^p \text{dom} h_j$

可行解集 (feasible set): $X = \left\{ \mathbf{x} \in D \left| \begin{array}{l} f_i(\mathbf{x}) \leq 0 \quad i=1, \dots, m \\ h_j(\mathbf{x}) = 0 \quad j=1, \dots, p \end{array} \right. \right\}$

最优值 (optimal value): $p^* = \inf \{ f_o(\mathbf{x}) \mid \mathbf{x} \in X \}$

最优解 (optimal solution) \mathbf{x}^* : $f_o(\mathbf{x}^*) = p^*$

拉格朗日函数 (Lagrangian Function)

$$L(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{v}) = f_o(\mathbf{x}) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}) + \sum_{j=1}^p \nu_j h_j(\mathbf{x})$$

$$\text{dom} L = D \times \mathbf{R}^m \times \mathbf{R}^p$$

\mathbf{x} : 原变量 (primal variable)

$\boldsymbol{\lambda}, \mathbf{v}$: 对偶变量 (dual variable)

λ_i, ν_i : 拉格朗日乘子 (Lagrange Multiplier)

对偶函数 (Lagrange Dual Function)

$$g(\boldsymbol{\lambda}, \mathbf{v}) = \inf_{\mathbf{x} \in D} L(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{v}) = \inf_{\mathbf{x} \in D} \left(f_o(\mathbf{x}) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}) + \sum_{j=1}^p \nu_j h_j(\mathbf{x}) \right)$$

(1) 对偶函数一定为凹函数

(2) $\forall \boldsymbol{\lambda} \succeq \mathbf{0}, \forall \mathbf{v}, \quad g(\boldsymbol{\lambda}, \mathbf{v}) \leq p^*$

证明: 设 \mathbf{x}^* 是最优解, 则 $f_i(\mathbf{x}^*) \leq 0$, $h_j(\mathbf{x}^*) = 0$, 且 $f_o(\mathbf{x}^*) = p^*$

$$L(\mathbf{x}^*, \boldsymbol{\lambda}, \mathbf{v}) = f_o(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}^*) + \sum_{j=1}^p \nu_j h_j(\mathbf{x}^*) \leq p^*$$

$$g(\boldsymbol{\lambda}, \mathbf{v}) = \inf_{\mathbf{x} \in D} L(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{v}) \leq L(\mathbf{x}^*, \boldsymbol{\lambda}, \mathbf{v}) \leq p^*$$

$$\text{例: } \begin{cases} \min & \mathbf{c}^T \mathbf{x} \\ \text{s.t.} & \mathbf{Ax} = \mathbf{b} \\ & \mathbf{x} \succeq \mathbf{0} \end{cases}$$

$$\begin{aligned} L(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{v}) &= \mathbf{c}^T \mathbf{x} - \boldsymbol{\lambda}^T \mathbf{x} + \mathbf{v}^T (\mathbf{Ax} - \mathbf{b}) \\ &= (\mathbf{c} - \boldsymbol{\lambda} + \mathbf{A}^T \mathbf{v})^T \mathbf{x} - \mathbf{v}^T \mathbf{b} \end{aligned}$$

$$\begin{aligned} g(\boldsymbol{\lambda}, \mathbf{v}) &= \inf_{\mathbf{x} \in D} L(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{v}) \\ &= \begin{cases} -\mathbf{b}^T \mathbf{v} & \mathbf{c} - \boldsymbol{\lambda} + \mathbf{A}^T \mathbf{v} = \mathbf{0} \\ -\infty & \text{others} \end{cases} \quad (\text{concave}) \end{aligned}$$

$$\text{例: } \begin{cases} \min & \mathbf{c}^T \mathbf{x} \\ \text{s.t.} & \mathbf{Ax} - \mathbf{b} \preceq \mathbf{0} \end{cases}$$

$$\begin{aligned} L(\mathbf{x}, \boldsymbol{\lambda}) &= \mathbf{c}^T \mathbf{x} + \boldsymbol{\lambda}^T (\mathbf{Ax} - \mathbf{b}) \\ &= (\mathbf{c} + \mathbf{A}^T \boldsymbol{\lambda})^T \mathbf{x} - \boldsymbol{\lambda}^T \mathbf{b} \end{aligned}$$

$$\begin{aligned} g(\boldsymbol{\lambda}) &= \inf_{\mathbf{x} \in D} L(\mathbf{x}, \boldsymbol{\lambda}) \\ &= \begin{cases} -\mathbf{b}^T \boldsymbol{\lambda} & \mathbf{c} + \mathbf{A}^T \boldsymbol{\lambda} = \mathbf{0} \\ -\infty & \text{others} \end{cases} \quad (\text{concave}) \end{aligned}$$

$$\text{例: } \begin{cases} \min & \mathbf{x}^T \mathbf{x} \\ \text{s.t.} & \mathbf{Ax} = \mathbf{b} \end{cases}$$

$$\begin{aligned} L(\mathbf{x}, \mathbf{v}) &= \mathbf{x}^T \mathbf{x} + \mathbf{v}^T (\mathbf{Ax} - \mathbf{b}) \\ &= \mathbf{x}^T \mathbf{x} + \mathbf{v}^T \mathbf{Ax} - \mathbf{v}^T \mathbf{b} \end{aligned} \quad \frac{\partial L}{\partial \mathbf{x}} = 2\mathbf{x} + \mathbf{A}^T \mathbf{v} = \mathbf{0} \Rightarrow \mathbf{x} = -\frac{1}{2} \mathbf{A}^T \mathbf{v}$$

$$\begin{aligned} g(\mathbf{v}) &= \inf_{\mathbf{x} \in D} L(\mathbf{x}, \mathbf{v}) \\ &= \frac{1}{4} \mathbf{v}^T \mathbf{A} \mathbf{A}^T \mathbf{v} - \frac{1}{2} \mathbf{v}^T \mathbf{A} \mathbf{A}^T \mathbf{v} - \mathbf{v}^T \mathbf{b} \\ &= -\frac{1}{4} \mathbf{v}^T \mathbf{A} \mathbf{A}^T \mathbf{v} - \mathbf{v}^T \mathbf{b} \quad (\text{concave}) \end{aligned}$$

$$\text{例: } \begin{cases} \min & \mathbf{x}^T \mathbf{W} \mathbf{x} \\ \text{s.t.} & x_i^2 = 1, \quad i = 1, \dots, n \end{cases} \quad (\text{非凸问题})$$

$$\begin{aligned} L(\mathbf{x}, \mathbf{v}) &= \mathbf{x}^T \mathbf{W} \mathbf{x} + \sum_{i=1}^n v_i (x_i^2 - 1) \\ &= \mathbf{x}^T (\mathbf{W} + \text{diag}\{\mathbf{v}\}) \mathbf{x} - \mathbf{1}^T \mathbf{v} \end{aligned}$$

$$\begin{aligned} g(\mathbf{v}) &= \inf_{\mathbf{x} \in D} L(\mathbf{x}, \mathbf{v}) \\ &= \begin{cases} -\mathbf{1}^T \mathbf{v} & \mathbf{W} + \text{diag}\{\mathbf{v}\} \succeq \mathbf{0} \\ -\infty & \text{others} \end{cases} \quad (\text{concave}) \end{aligned}$$

对偶函数与共轭函数之间的联系

$f(\mathbf{x})$ 的共轭函数为 $f^*(\mathbf{y}) \triangleq \sup_{\mathbf{x} \in \text{dom} f} (\mathbf{y}^T \mathbf{x} - f(\mathbf{x}))$

$f_o(\mathbf{x})$ (在 $f_i(\mathbf{x}) \leq 0, h_j(\mathbf{x}) = 0$ 的约束下) 的对偶函数为 $g(\boldsymbol{\lambda}, \mathbf{v}) = \inf_{\mathbf{x} \in \mathcal{D}} \left(f_o(\mathbf{x}) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}) + \sum_{j=1}^p \nu_j h_j(\mathbf{x}) \right)$

$$= -\sup_{\mathbf{x} \in \mathcal{D}} \left(-\sum_{i=1}^m \lambda_i f_i(\mathbf{x}) - \sum_{j=1}^p \nu_j h_j(\mathbf{x}) - f_o(\mathbf{x}) \right)$$

如果 $f_i(\mathbf{x}), h_j(\mathbf{x})$ 是关于 \mathbf{x} 的线性函数, 则两者刚好对应

推导如下:

$$\begin{aligned} \min \quad & f_o(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{Ax} \leq \mathbf{b} \\ & \mathbf{Cx} = \mathbf{d} \end{aligned}$$

$$\begin{aligned} L(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{v}) &= f_o(\mathbf{x}) - \boldsymbol{\lambda}^T (\mathbf{Ax} - \mathbf{b}) + \mathbf{v}^T (\mathbf{Cx} - \mathbf{d}) \\ &= f_o(\mathbf{x}) + (\mathbf{A}^T \boldsymbol{\lambda} + \mathbf{C}^T \mathbf{v})^T \mathbf{x} - \boldsymbol{\lambda}^T \mathbf{b} - \mathbf{v}^T \mathbf{d} \end{aligned}$$

$$\begin{aligned} g(\boldsymbol{\lambda}, \mathbf{v}) &= \inf_{\mathbf{x} \in \mathcal{D}} L(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{v}) \\ &= \inf_{\mathbf{x} \in \mathcal{D}} \left(f_o(\mathbf{x}) + (\mathbf{A}^T \boldsymbol{\lambda} + \mathbf{C}^T \mathbf{v})^T \mathbf{x} - \boldsymbol{\lambda}^T \mathbf{b} - \mathbf{v}^T \mathbf{d} \right) \\ &= -f_o^*(-\mathbf{A}^T \boldsymbol{\lambda} - \mathbf{C}^T \mathbf{v}) - \boldsymbol{\lambda}^T \mathbf{b} - \mathbf{v}^T \mathbf{d} \end{aligned}$$

对偶问题 (Lagrange Dual Problem)

$$\begin{cases} \max & g(\lambda, v) \\ \text{s.t.} & \lambda \succeq \mathbf{0} \end{cases}$$

最优值 (optimal value): $d^* = \inf \{g(\lambda, v) \mid \lambda, v \in \text{dom } g \text{ \& } \lambda \succeq \mathbf{0}\}$

原问题 (Primal Problem)

$$\begin{aligned} \min & f_o(x) \\ \text{s.t.} & f_i(x) \leq 0 \quad i = 1, \dots, m \\ & h_j(x) = 0 \quad j = 1, \dots, p \end{aligned}$$

最优值 (optimal value): $p^* = \inf \{f_o(x) \mid x \in X\}$

由对偶函数的性质, $d^* \leq p^*$

对偶问题与原问题的关系

$$\begin{cases} \min & \mathbf{c}^T \mathbf{x} \\ \text{s.t.} & \mathbf{Ax} = \mathbf{b} \\ & \mathbf{x} \succeq \mathbf{0} \end{cases} \Rightarrow g(\lambda, v) = \begin{cases} -\mathbf{b}^T \mathbf{v} & \mathbf{c} - \lambda + \mathbf{A}^T \mathbf{v} = \mathbf{0} \\ -\infty & \text{others} \end{cases}$$

$$(D): \begin{cases} \max & g(\lambda, v) \\ \text{s.t.} & \lambda \succeq \mathbf{0} \end{cases} \Rightarrow \begin{cases} \max & -\mathbf{b}^T \mathbf{v} \\ \text{s.t.} & \mathbf{c} - \lambda + \mathbf{A}^T \mathbf{v} = \mathbf{0} \\ & \lambda \succeq \mathbf{0} \end{cases} \Rightarrow \begin{cases} \max & -\mathbf{b}^T \mathbf{v} \\ \text{s.t.} & \mathbf{c} + \mathbf{A}^T \mathbf{v} \succeq \mathbf{0} \end{cases} \Rightarrow \begin{cases} \min & \mathbf{b}^T \mathbf{v} \\ \text{s.t.} & \mathbf{c} + \mathbf{A}^T \mathbf{v} \succeq \mathbf{0} \end{cases}$$

$$\begin{cases} \min & \mathbf{c}^T \mathbf{x} \\ \text{s.t.} & \mathbf{Ax} - \mathbf{b} \preceq \mathbf{0} \end{cases} \Rightarrow g(\lambda) = \begin{cases} -\mathbf{b}^T \lambda & \mathbf{c} + \mathbf{A}^T \lambda = \mathbf{0} \\ -\infty & \text{others} \end{cases}$$

$$(D): \begin{cases} \max & g(\lambda, v) \\ \text{s.t.} & \lambda \succeq \mathbf{0} \end{cases} \Rightarrow \begin{cases} \max & -\mathbf{b}^T \mathbf{v} \\ \text{s.t.} & \mathbf{c} + \mathbf{A}^T \lambda = \mathbf{0} \\ & \lambda \succeq \mathbf{0} \end{cases} \Rightarrow \begin{cases} \min & \mathbf{b}^T \mathbf{v} \\ \text{s.t.} & \mathbf{c} + \mathbf{A}^T \lambda = \mathbf{0} \\ & \lambda \succeq \mathbf{0} \end{cases}$$

就线性优化问题而言, 其对偶的对偶就是它自身

但此结论不适用于其它优化问题

$$\begin{cases} \min & \mathbf{x}^T \mathbf{x} \\ \text{s.t.} & \mathbf{Ax} = \mathbf{b} \end{cases} \Rightarrow g(v) = -\frac{1}{4} \mathbf{v}^T \mathbf{AA}^T \mathbf{v} - \mathbf{v}^T \mathbf{b}$$

$$(D): \max \quad -\frac{1}{4} \mathbf{v}^T \mathbf{AA}^T \mathbf{v} - \mathbf{v}^T \mathbf{b} \quad \Rightarrow \quad \min \quad \frac{1}{4} \mathbf{v}^T \mathbf{AA}^T \mathbf{v} + \mathbf{v}^T \mathbf{b}$$

无约束问题无法写出其对偶问题

弱对偶性、强对偶性、对偶间隙

前面已经提到，对偶问题的最优值 d^* 与原问题的最优值 p^* 满足： $d^* \leq p^*$

$d^* \leq p^*$ ——弱对偶性 (Weak Duality)

$d^* = p^*$ ——强对偶性 (Strong Duality)

$p^* - d^*$ ——对偶间隙 (Duality Gap)

满足强对偶性的 Slater's Condition

定义： D 的 Relative Interior

$\text{relint} D = \{ \mathbf{x} \in D \mid \exists r > 0, (B(\mathbf{x}, r) \cap \text{aff} D) \in D \}$ (即 D 去掉边界后的集合)

定理：

①对于非凸问题，其对偶问题通常没有强对偶性

②对于凸问题，有充分条件 (Slater's Condition)，使得该问题的对偶问题满足强对偶性

Primal Problem (Convex)

$$\begin{aligned} \min \quad & f_o(\mathbf{x}) \\ \text{s.t.} \quad & f_i(\mathbf{x}) \leq 0 \quad i = 1, \dots, m \\ & A_j^T \mathbf{x} = \mathbf{b}_j \quad j = 1, \dots, p \end{aligned}$$

Slater's Condition

$\exists \mathbf{x} \in \text{relint} D$

$$\begin{cases} f_i(\mathbf{x}) < 0 & i = 1, \dots, m \\ A_j^T \mathbf{x} = \mathbf{b}_j & j = 1, \dots, p \end{cases}$$

Refined Slater's Condition

$\exists \mathbf{x} \in \text{relint} D$

$$\begin{cases} f_i(\mathbf{x}) \leq 0 & i = 1, \dots, k & \text{when } f_i \text{ is affine} \\ f_i(\mathbf{x}) < 0 & i = k+1, \dots, m \\ A_j^T \mathbf{x} = \mathbf{b}_j & j = 1, \dots, p \end{cases}$$

例：

$$\begin{cases} \min & \mathbf{x}^T \mathbf{x} \\ \text{s.t.} & A\mathbf{x} = \mathbf{b} \end{cases} \quad (\text{QP 问题}) \quad \Rightarrow \quad g(\mathbf{v}) = -\frac{1}{4} \mathbf{v}^T A A^T \mathbf{v} - \mathbf{v}^T \mathbf{b}$$

$$(D) : \max \quad -\frac{1}{4} \mathbf{v}^T A A^T \mathbf{v} - \mathbf{v}^T \mathbf{b}$$

原问题满足 Slater's Condition，故对偶问题满足强对偶性

例:

$$\begin{cases} \min & \frac{1}{2} \mathbf{x}^T \mathbf{P}_0 \mathbf{x} + \mathbf{q}_0^T \mathbf{x} + r_0 \\ \text{s.t.} & \frac{1}{2} \mathbf{x}^T \mathbf{P}_i \mathbf{x} + \mathbf{q}_i^T \mathbf{x} + r_i \leq 0, i=1, \dots, m \end{cases} \quad \text{with } \mathbf{P}_0 \in \mathbf{S}_{++}^n, \mathbf{P}_i \in \mathbf{S}_+^n \text{ (QCQP 问题)}$$

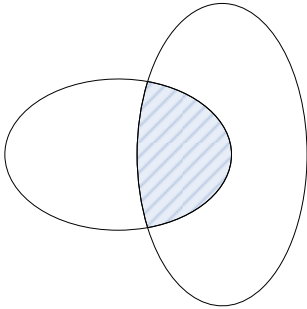
$$\begin{aligned} L(\mathbf{x}, \boldsymbol{\lambda}) &= \frac{1}{2} \mathbf{x}^T \mathbf{P}_0 \mathbf{x} + \mathbf{q}_0^T \mathbf{x} + r_0 + \sum_{i=1}^m \lambda_i \left(\frac{1}{2} \mathbf{x}^T \mathbf{P}_i \mathbf{x} + \mathbf{q}_i^T \mathbf{x} + r_i \right) \\ &= \frac{1}{2} \mathbf{x}^T \left(\mathbf{P}_0 + \sum_{i=1}^m \lambda_i \mathbf{P}_i \right) \mathbf{x} + \left(\mathbf{q}_0^T + \sum_{i=1}^m \lambda_i \mathbf{q}_i^T \right) \mathbf{x} + \left(r_0 + \sum_{i=1}^m \lambda_i r_i \right) \\ &= \frac{1}{2} \mathbf{x}^T \mathbf{P}(\boldsymbol{\lambda}) \mathbf{x} + \mathbf{q}(\boldsymbol{\lambda})^T \mathbf{x} + r(\boldsymbol{\lambda}) \end{aligned}$$

$$\frac{\partial L}{\partial \mathbf{x}} = \mathbf{P}(\boldsymbol{\lambda}) \mathbf{x} + \mathbf{q}(\boldsymbol{\lambda}) = \mathbf{0} \Rightarrow \mathbf{x} = -\mathbf{P}(\boldsymbol{\lambda})^{-1} \mathbf{q}(\boldsymbol{\lambda})$$

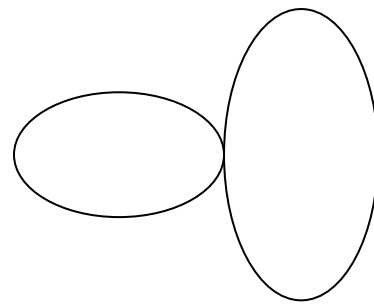
$$\begin{aligned} g(\boldsymbol{\lambda}) &= \inf_{\mathbf{x} \in \mathbf{R}^n} L(\mathbf{x}, \boldsymbol{\lambda}) \\ &= \frac{1}{2} \left(-\mathbf{P}(\boldsymbol{\lambda})^{-1} \mathbf{q}(\boldsymbol{\lambda}) \right)^T \mathbf{P}(\boldsymbol{\lambda}) \left(-\mathbf{P}(\boldsymbol{\lambda})^{-1} \mathbf{q}(\boldsymbol{\lambda}) \right) + \mathbf{q}(\boldsymbol{\lambda})^T \left(-\mathbf{P}(\boldsymbol{\lambda})^{-1} \mathbf{q}(\boldsymbol{\lambda}) \right) + r(\boldsymbol{\lambda}) \\ &= -\frac{1}{2} \mathbf{q}(\boldsymbol{\lambda})^T \mathbf{P}(\boldsymbol{\lambda})^{-1} \mathbf{q}(\boldsymbol{\lambda}) + r(\boldsymbol{\lambda}) \quad \text{when } \boldsymbol{\lambda} \succeq \mathbf{0} \end{aligned}$$

$$(D): \begin{cases} \max & -\frac{1}{2} \mathbf{q}(\boldsymbol{\lambda})^T \mathbf{P}(\boldsymbol{\lambda})^{-1} \mathbf{q}(\boldsymbol{\lambda}) + r(\boldsymbol{\lambda}) \\ \text{s.t.} & \boldsymbol{\lambda} \succeq \mathbf{0} \end{cases}$$

假设原问题的可行解集存在（即约束条件中各椭圆有交集），分两种情况说明对偶问题的对偶性



1) 原问题满足 Slater's Condition
故对偶问题满足强对偶性



2) 原问题不满足 Slater's Condition
但因为可行解集只有一个解（必为最优解）
故对偶问题也满足强对偶性

此例的第二种情况也说明 Slater's Condition 是强对偶性的充分而非必要条件

再举一个不满足 Slater's Condition，而对偶间隙依然是零的例子

$$\begin{cases} \min & \mathbf{x} \\ \text{s.t.} & \mathbf{x} \leq \mathbf{0} \\ & -\mathbf{x} \leq \mathbf{0} \end{cases}$$

对于非凸问题，其对偶问题通常没有强对偶性，但也有例外
例:

$$\begin{cases} \min & \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x} \\ \text{s.t.} & \mathbf{x}^T \mathbf{x} \leq 1 \end{cases} \quad \text{with } \mathbf{A} \in \mathbf{S}_n, \mathbf{b} \in \mathbf{R}^n$$

对偶问题的四种解释

1、多目标优化解释

考虑如下单目标优化问题与多目标优化问题

$$\textcircled{1} \begin{cases} \min & f_o(\mathbf{x}) \\ \text{s.t.} & f_i(\mathbf{x}) \leq 0 \quad i=1, \dots, m \end{cases} \quad \text{对偶问题: } \min \quad f_o(\mathbf{x}) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}) \quad \lambda_i \geq 0$$

$$\textcircled{2} \min \quad \{f_o(\mathbf{x}), f_1(\mathbf{x}), \dots, f_m(\mathbf{x})\} \quad \text{对各目标进行非负加权: } \min \quad f_o(\mathbf{x}) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}) \quad \lambda_i \geq 0$$

亦即通过拉格朗日函数可以把有约束的单目标优化问题和无约束的多目标优化问题联系在一起

2、几何解释

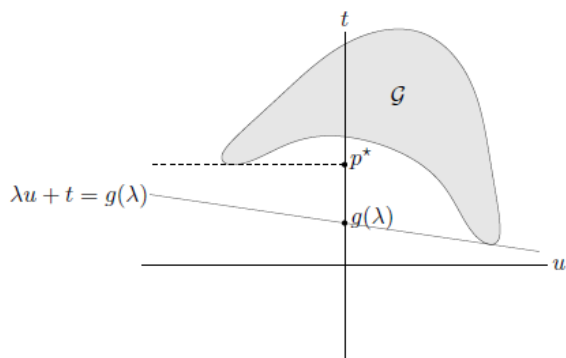
考虑如下单约束单目标优化问题

$$\begin{cases} \min & f_o(x) \\ \text{s.t.} & f_1(x) \leq 0 \end{cases}$$

$$\text{设 } \mathcal{G} = \{(f_1(x), f_o(x)) \mid x \in D\}$$

$$\text{则 } g(\lambda) = \inf \{t + \lambda u \mid (u, t) \in \mathcal{G}\}$$

$$p^* = \inf \{t \mid (u, t) \in \mathcal{G}, u \leq 0\}$$

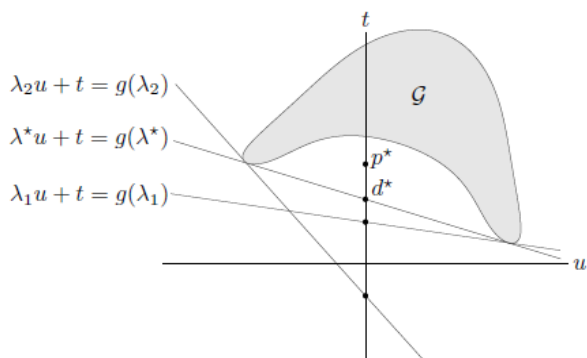


Geometric interpretation of dual function and lower bound $g(\lambda) \leq p^*$, for a problem with one (inequality) constraint.

Given λ , we minimize $(\lambda, 1)^T (u, t)$ over $\mathcal{G} = \{(f_1(x), f_o(x)) \mid x \in D\}$.

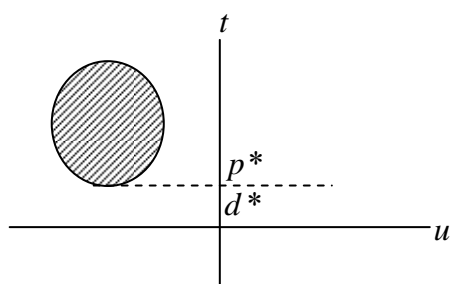
This yields a supporting hyperplane with slope $-\lambda$.

The intersection of this hyperplane with the $u = 0$ axis gives $g(\lambda)$.



Supporting hyperplanes corresponding to three dual feasible values of λ , including the optimum λ^* .

Strong duality does not hold; the optimal duality gap $p^* - d^*$ is positive.



强对偶性成立

3.经济学解释

$$\begin{cases} \min & f_o(\mathbf{x}) \\ \text{s.t.} & f_i(\mathbf{x}) \leq 0 \quad i = 1, \dots, m \end{cases}$$

\mathbf{x} : 运营条件 (Operating Condition)

$f_o(\mathbf{x})$: 运营条件为 \mathbf{x} 时的成本 (Cost) / $-f_o(\mathbf{x})$: 收益 (Profit)

$f_i(\mathbf{x})$: 运营条件为 \mathbf{x} 时的资源配额 (Resource Limit)

在共产主义条件下, 资源经配置后产生的最小成本为 p^*

现考虑自由市场条件下, 资源可买可卖, 设第 i 中资源的单价为 λ_i ($\lambda_i \geq 0$)

则成本可表示为: $f_o(\mathbf{x}) + \underbrace{\sum f_i(\mathbf{x})\lambda_i}_{buy} - \underbrace{\sum (-f_j(\mathbf{x}))\lambda_j}_{sale}$

若使成本最小化 (收益最大化), 则有: $\min f_o(\mathbf{x}) + \sum_{i=1}^m f_i(\mathbf{x})\lambda_i$

在自由市场条件下, 资源经配置后产生的最小成本为 d^* (此时对应的价格 λ^* 称为影子价格 Shadow Price)

由之前的结论: $d^* \leq p^*$

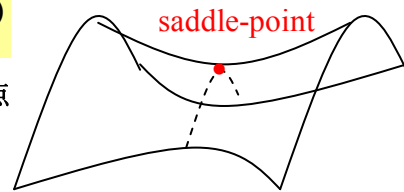
亦即自由市场条件下产生的最小成本必低于共产主义条件下产生的最小成本

4.鞍点解释

定理：考虑函数 $f(w, z)$, $w \in S_w$, $z \in S_z$, 有 $\sup_{z \in S_z} \inf_{w \in S_w} f(w, z) \leq \inf_{w \in S_w} \sup_{z \in S_z} f(w, z)$

定义：若 $\exists (\tilde{w}, \tilde{z}) \in \text{dom} f$, $\sup_{z \in S_z} \inf_{w \in S_w} f(\tilde{w}, \tilde{z}) = \inf_{w \in S_w} \sup_{z \in S_z} f(\tilde{w}, \tilde{z})$, 则 (\tilde{w}, \tilde{z}) 称为鞍点

性质：若 (\tilde{w}, \tilde{z}) 是鞍点, 则有 $f(\tilde{w}, z) \leq f(\tilde{w}, \tilde{z}) \leq f(w, \tilde{z})$, $\forall z \in S_z, \forall w \in S_w$



考虑

原问题：
$$\begin{cases} \min & f_o(\mathbf{x}) \\ \text{s.t.} & f_i(\mathbf{x}) \leq 0 \quad i=1, \dots, m \end{cases}$$

拉格朗日函数：
$$L(\mathbf{x}, \boldsymbol{\lambda}) = f_o(\mathbf{x}) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x})$$

$$\begin{cases} \sup_{\boldsymbol{\lambda} \geq \mathbf{0}} L(\mathbf{x}, \boldsymbol{\lambda}) = \sup_{\boldsymbol{\lambda} \geq \mathbf{0}} \left\{ f_o(\mathbf{x}) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}) \right\} = \begin{cases} f_o(\mathbf{x}) & f_i(\mathbf{x}) \leq 0, i=1, \dots, m \\ +\infty & \text{others} \end{cases} \\ \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}) = g(\boldsymbol{\lambda}) \end{cases}$$

$$\begin{cases} p^* = \inf_{\mathbf{x}} \{ f_o(\mathbf{x}) \mid f_i(\mathbf{x}) \leq 0, i=1, \dots, m \} = \inf_{\mathbf{x}} \sup_{\boldsymbol{\lambda} \geq \mathbf{0}} L(\mathbf{x}, \boldsymbol{\lambda}) \\ d^* = \sup_{\boldsymbol{\lambda} \geq \mathbf{0}} g(\boldsymbol{\lambda}) = \sup_{\boldsymbol{\lambda} \geq \mathbf{0}} \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}) \end{cases}$$

所以 $d^* \leq p^*$

当 $L(\mathbf{x}, \boldsymbol{\lambda})$ 有鞍点时, 则 $d^* = p^*$

若 $(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\lambda}})$ 为 $L(\mathbf{x}, \boldsymbol{\lambda})$ 的鞍点 \Leftrightarrow

对偶问题满足强对偶性, 且 $(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\lambda}})$ 为 Primal-Dual 最优解

$$\inf_{\mathbf{x}} \sup_{\boldsymbol{\lambda} \geq \mathbf{0}} L(\mathbf{x}, \boldsymbol{\lambda}) = \sup_{\boldsymbol{\lambda} \geq \mathbf{0}} \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}), \text{ 且 } \begin{cases} \tilde{\mathbf{x}} = \arg \inf_{\mathbf{x}} \sup_{\boldsymbol{\lambda} \geq \mathbf{0}} L(\mathbf{x}, \boldsymbol{\lambda}) & \text{primal optimal point} \\ \tilde{\boldsymbol{\lambda}} = \arg \sup_{\boldsymbol{\lambda} \geq \mathbf{0}} \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}) & \text{dual optimal point} \end{cases}$$

下面讨论最优解 $\tilde{\mathbf{x}}$ 与 $\tilde{\boldsymbol{\lambda}}$ 有哪些性质

$$(1) \begin{cases} f_i(\tilde{\mathbf{x}}) \leq 0, i=1, \dots, m \\ \tilde{\boldsymbol{\lambda}} \succeq \mathbf{0} \end{cases}$$

$$(2) \quad f_o(\tilde{\mathbf{x}}) = g(\tilde{\boldsymbol{\lambda}}) = \inf_{\mathbf{x}} \left\{ f_o(\mathbf{x}) + \sum_{i=1}^m \tilde{\lambda}_i f_i(\mathbf{x}) \right\}$$

$$\leq f_o(\tilde{\mathbf{x}}) + \sum_{i=1}^m \tilde{\lambda}_i f_i(\tilde{\mathbf{x}})$$

$$\leq f_o(\tilde{\mathbf{x}}) \quad \underline{\leq 0}$$

上式若成立，显然两个不等式必须取等号

$$(2-1) \quad \inf_{\mathbf{x}} \left\{ f_o(\mathbf{x}) + \sum_{i=1}^m \tilde{\lambda}_i f_i(\mathbf{x}) \right\} = f_o(\tilde{\mathbf{x}}) + \sum_{i=1}^m \tilde{\lambda}_i f_i(\tilde{\mathbf{x}})$$

$$\inf_{\mathbf{x}} L(\mathbf{x}, \tilde{\boldsymbol{\lambda}}) = L(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\lambda}})$$

$$(2-2) \quad f_o(\tilde{\mathbf{x}}) + \sum_{i=1}^m \tilde{\lambda}_i f_i(\tilde{\mathbf{x}}) = f_o(\tilde{\mathbf{x}}) = \sup_{\boldsymbol{\lambda} \succeq \mathbf{0}} \left\{ f_o(\tilde{\mathbf{x}}) + \sum_{i=1}^m \lambda_i f_i(\tilde{\mathbf{x}}) \right\}$$

$$L(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\lambda}}) = \sup_{\boldsymbol{\lambda} \succeq \mathbf{0}} L(\tilde{\mathbf{x}}, \boldsymbol{\lambda})$$

沿原变量去看，鞍点在最低点

沿对偶变量去看，鞍点在最高点

$$(2-3) \quad \sum_{i=1}^m \tilde{\lambda}_i f_i(\tilde{\mathbf{x}}) = 0$$

$$\begin{cases} f_i(\tilde{\mathbf{x}}) \neq 0 \\ \tilde{\lambda}_i = 0 \end{cases} \quad \text{或} \quad \begin{cases} f_i(\tilde{\mathbf{x}}) = 0 \\ \tilde{\lambda}_i \neq 0 \end{cases}$$

☆核心的核心

一般优化问题（可以是非凸问题）

$$\begin{cases} \min & f_o(\mathbf{x}) \\ \text{s.t.} & f_i(\mathbf{x}) \leq 0 \quad i=1, \dots, m \\ & h_j(\mathbf{x}) = 0 \quad j=1, \dots, p \end{cases}$$

$$g(\boldsymbol{\lambda}, \mathbf{v}) = \inf_{\mathbf{x} \in D} L(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{v}) = \inf_{\mathbf{x} \in D} \left(f_o(\mathbf{x}) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}) + \sum_{j=1}^p \nu_j h_j(\mathbf{x}) \right)$$

$$\begin{cases} \max & g(\boldsymbol{\lambda}, \mathbf{v}) \\ \text{s.t.} & \boldsymbol{\lambda} \succeq \mathbf{0} \end{cases}$$

假设原问题与对偶问题的对偶间隙为 0，则该问题存在 Primal-Dual 最优解

设 $(\mathbf{x}^*, \boldsymbol{\lambda}^*, \mathbf{v}^*)$ 是该问题的 Primal-Dual 最优解，下面讨论最优解的性质

$$(1) \begin{cases} f_i(\mathbf{x}^*) \leq 0 & i=1, \dots, m \\ h_j(\mathbf{x}^*) = 0 & j=1, \dots, p \quad (\text{约束条件/Constraint}) \\ \boldsymbol{\lambda}^* \succeq \mathbf{0} \end{cases}$$

$$\begin{aligned} (2) \quad f_o(\mathbf{x}^*) = g(\boldsymbol{\lambda}^*, \mathbf{v}^*) &= \inf_{\mathbf{x}} \left\{ f_o(\mathbf{x}) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}) + \sum_{j=1}^p \nu_j^* h_j(\mathbf{x}) \right\} \\ &\leq f_o(\mathbf{x}^*) + \underbrace{\sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}^*)}_{\leq 0} + \underbrace{\sum_{j=1}^p \nu_j^* h_j(\mathbf{x}^*)}_{=0} \\ &\leq f_o(\mathbf{x}^*) \end{aligned}$$

上式若成立，显然两个不等式必须取等号

$$(2-1) \quad \inf_{\mathbf{x}} \left\{ f_o(\mathbf{x}) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}) + \sum_{j=1}^p \nu_j^* h_j(\mathbf{x}) \right\} = f_o(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}^*) + \sum_{j=1}^p \nu_j^* h_j(\mathbf{x}^*)$$

$$\text{假设 } f_o, f_i, h_j \text{ 均可微, 则 } \left. \frac{\partial L(\mathbf{x}, \boldsymbol{\lambda}^*, \mathbf{v}^*)}{\partial \mathbf{x}} \right|_{\mathbf{x}=\mathbf{x}^*} = 0$$

$$\text{即 } \nabla f_o(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* \nabla f_i(\mathbf{x}^*) + \sum_{j=1}^p \nu_j^* \nabla h_j(\mathbf{x}^*) = 0 \quad (\text{稳定性条件/Stationarity})$$

$$(2-2) \quad f_o(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}^*) + \sum_{j=1}^p \nu_j^* h_j(\mathbf{x}^*) = f_o(\mathbf{x}^*)$$

$$\sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}^*) = 0$$

$$\lambda_i^* f_i(\mathbf{x}^*) = 0, \quad \forall i=1, \dots, m \quad (\text{互补松弛条件/Complementary Slackness})$$

★最核心

KKT 条件 (Karush-Kuhn-Tucker Conditions)

对于可微无对偶间隙优化问题，其最优解的必要条件是

$$\begin{aligned} f_i(\mathbf{x}^*) &\leq 0 & i=1, \dots, m \\ h_j(\mathbf{x}^*) &= 0 & j=1, \dots, p \end{aligned} \quad \text{---Primal feasibility}$$

$$\lambda_i^* \geq 0 \quad i=1, \dots, m \quad \text{---Dual feasibility}$$

$$\lambda_i^* f_i(\mathbf{x}^*) = 0 \quad i=1, \dots, m \quad \text{---Complementary Slackness}$$

$$\nabla f_o(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* \nabla f_i(\mathbf{x}^*) + \sum_{j=1}^p \nu_j^* \nabla h_j(\mathbf{x}^*) = 0 \quad \text{---Stationarity}$$

Tucker: Minsky, Nash

200 年前，拉格朗日退出了只有等式约束的 KKT 条件

$$\begin{cases} h_j(\mathbf{x}^*) = 0 & j=1, \dots, p \\ \nabla f_o(\mathbf{x}^*) + \sum_{j=1}^p \nu_j^* \nabla h_j(\mathbf{x}^*) = 0 \end{cases}$$

对于可微无对偶间隙凸优化问题，KKT 条件等价于 Primal-Dual 最优解

必要性已证，下面证充分性

$$f_i(\mathbf{x}^*) \leq 0 \quad i=1, \dots, m$$

$$h_j(\mathbf{x}^*) = 0 \quad j=1, \dots, p$$

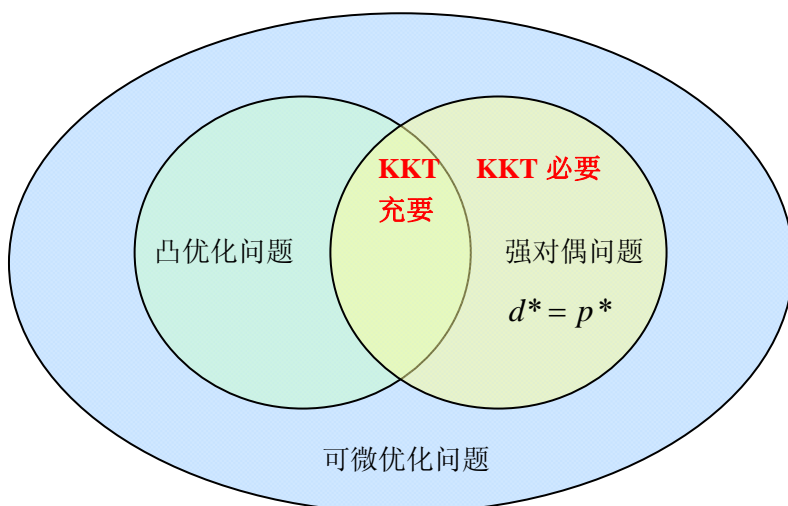
$$\lambda_i^* \geq 0 \quad i=1, \dots, m$$

$$\sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}^*) = 0$$

$$\begin{cases} L(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) \text{ is convex} \\ \left. \frac{\partial L(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)}{\partial \mathbf{x}} \right|_{\mathbf{x}=\mathbf{x}^*} = 0 \end{cases} \Rightarrow L(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) = \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$$

$$\begin{aligned} g(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) &= \inf_{\mathbf{x}} \left\{ f_o(\mathbf{x}) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}) + \sum_{j=1}^p \nu_j^* h_j(\mathbf{x}) \right\} \\ &= f_o(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}^*) + \sum_{j=1}^p \nu_j^* h_j(\mathbf{x}^*) \\ &= f_o(\mathbf{x}^*) \end{aligned}$$

亦即对偶间隙为 0



例：

$$\begin{cases} \min & \frac{1}{2} \mathbf{x}^T \mathbf{P} \mathbf{x} + \mathbf{q}^T \mathbf{x} + r \\ \text{s.t.} & \mathbf{A} \mathbf{x} = \mathbf{b} \end{cases} \text{ with } \mathbf{P} \in \mathbf{S}_{++}^n \text{ (QP 问题)}$$

$$\Rightarrow \begin{cases} \mathbf{A} \mathbf{x}^* = \mathbf{b} \\ \mathbf{P} \mathbf{x}^* + \mathbf{q} + \mathbf{A}^T \mathbf{v}^* = \mathbf{0} \end{cases} \text{ (KKT 条件)}$$

$$\Rightarrow \begin{pmatrix} \mathbf{P} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x}^* \\ \mathbf{v}^* \end{pmatrix} = \begin{pmatrix} -\mathbf{q} \\ \mathbf{b} \end{pmatrix}$$

求解此线性方程组即可求得最优解 $(\mathbf{x}^*, \mathbf{v}^*)$

例：非负象限约束优化问题（Minimization over the nonnegative orthant）

$$\begin{cases} \min & f_o(\mathbf{x}) \\ \text{s.t.} & \mathbf{x} \succeq \mathbf{0} \end{cases}$$

前面已经通过分析“可微目标函数下最优解的性质”得出最优解满足

$$\begin{cases} \mathbf{x}^* \succeq \mathbf{0} & (a) \\ \nabla f_o(\mathbf{x}^*) \succeq \mathbf{0} & (b) \\ \nabla^T f_o(\mathbf{x}^*) \mathbf{x}^* = 0 & (c) \end{cases}$$

由 KKT 条件，有

$$\begin{cases} \mathbf{x}^* \succeq \mathbf{0} & (1) \\ \boldsymbol{\lambda}^* \succeq \mathbf{0} & (2) \\ (\boldsymbol{\lambda}^*)^T (-\mathbf{x}^*) = 0 & (3) \\ \nabla f_o(\mathbf{x}^*) + (-\boldsymbol{\lambda}^*) = \mathbf{0} & (4) \end{cases}$$

将 (4) 代入 (2) 和 (3)，消去 $\boldsymbol{\lambda}^*$ ，即得 (a) (b) (c) 三式

例：Water Filling 算法

$$P: \begin{cases} \min & -\sum_{i=1}^n \log(\alpha_i + x_i) \\ \text{s.t.} & \mathbf{x} \succeq \mathbf{0} \\ & \mathbf{1}^T \mathbf{x} = 1 \end{cases}$$

$$\mathbf{x}^* \succeq \mathbf{0}$$

$$\mathbf{1}^T \mathbf{x}^* = 1$$

$$\Rightarrow \boldsymbol{\lambda}^* \succeq \mathbf{0}$$

$$(\boldsymbol{\lambda}^*)^T (-\mathbf{x}^*) = 0$$

$$\left. \frac{\partial L(\mathbf{x}, \boldsymbol{\lambda}^*, \nu^*)}{\partial \mathbf{x}} \right|_{\mathbf{x}=\mathbf{x}^*} = \frac{\partial \left(-\sum_{i=1}^n \log(\alpha_i + x_i) - (\boldsymbol{\lambda}^*)^T \mathbf{x} + \nu^* (\mathbf{1}^T \mathbf{x} - 1) \right)}{\partial \mathbf{x}} \bigg|_{\mathbf{x}=\mathbf{x}^*} = \begin{pmatrix} -\frac{1}{\alpha_1 + x_1^*} \\ \vdots \\ -\frac{1}{\alpha_n + x_n^*} \end{pmatrix} - \boldsymbol{\lambda}^* + \nu^* \mathbf{1} = \mathbf{0}$$

将三四五式写成标量的形式

$$\mathbf{x}^* \succeq \mathbf{0}$$

$$\mathbf{1}^T \mathbf{x}^* = 1$$

$$\Rightarrow \lambda_i^* \geq 0 \quad i=1, \dots, n$$

$$\lambda_i^* (-x_i^*) = 0 \quad i=1, \dots, n$$

$$-\frac{1}{\alpha_i + x_i} - \lambda_i^* + \nu^* = 0 \quad i=1, \dots, n$$

将第五式代入第三式和第四式，消去 $\boldsymbol{\lambda}^*$

$$\mathbf{x}^* \succeq \mathbf{0}$$

$$\mathbf{1}^T \mathbf{x}^* = 1$$

$$\Rightarrow \nu^* \geq \frac{1}{\alpha_i + x_i^*} \quad i=1, \dots, n$$

$$x_i^* (\nu^* - \frac{1}{\alpha_i + x_i^*}) = 0 \quad i=1, \dots, n$$

若 $\nu^* < \frac{1}{\alpha_i}$ ，根据第三式，有 $x_i^* > 0$ ；再根据第四式，有 $x_i^* = \frac{1}{\nu^*} - \alpha_i$

若 $\nu^* \geq \frac{1}{\alpha_i}$ ，根据第四式，有 $\nu^* - \frac{1}{\alpha_i + x_i^*} > 0$ ，即 $x_i^* = 0$

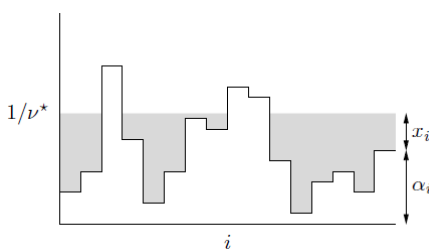


Illustration of water-filling algorithm.

The height of each patch is given by α_i .

The region is flooded to a level $1/\nu^*$ which uses a total quantity of water equal to one.

The height of the water (shown shaded) above each patch is the optimal value of x_i^* .

干扰 (Perturbation) 及敏感性分析 (Sensitivity Analysis)

原问题

$$\begin{cases} \min & f_o(\mathbf{x}) \\ \text{s.t.} & f_i(\mathbf{x}) \leq 0 \quad i=1, \dots, m \\ & h_j(\mathbf{x}) = 0 \quad j=1, \dots, p \end{cases}$$

干扰问题

$$\begin{cases} \min & f_o(\mathbf{x}) \\ \text{s.t.} & f_i(\mathbf{x}) \leq u_i \quad i=1, \dots, m \\ & h_j(\mathbf{x}) = w_j \quad j=1, \dots, p \end{cases}$$

记干扰问题的最优值为 $p^*(\mathbf{u}, \mathbf{w})$ ，则原问题的最优值为 $p^*(\mathbf{0}, \mathbf{0})$

性质 1: 若原问题为凸，则 $p^*(\mathbf{u}, \mathbf{w})$ 是 (\mathbf{u}, \mathbf{w}) 的凸函数

$$\text{证明: } p^*(\mathbf{u}, \mathbf{w}) = \inf_{\mathbf{x}} \left\{ f_o(\mathbf{x}) \mid \begin{matrix} f_i(\mathbf{x}) \leq u_i & i=1, \dots, m \\ h_j(\mathbf{x}) = w_j & j=1, \dots, p \end{matrix} \right\} \triangleq \inf_{\mathbf{x}} S(\mathbf{x}, \mathbf{u}, \mathbf{w})$$

$$\text{其中, } S(\mathbf{x}, \mathbf{u}, \mathbf{w}) = f_o(\mathbf{x}), \text{ 而 } \text{dom} S = \left\{ \mathbf{x} \in \text{dom} f_o, \begin{matrix} f_i(\mathbf{x}) \leq u_i & i=1, \dots, m \\ h_j(\mathbf{x}) = w_j & j=1, \dots, p \end{matrix} \right\}$$

对 $S(\mathbf{x}, \mathbf{u}, \mathbf{w})$ 而言，它的定义域是凸集，且是关于 $(\mathbf{x}, \mathbf{u}, \mathbf{w})$ 的联合凸函数

对 $S(\mathbf{x}, \mathbf{u}, \mathbf{w})$ 取极小后的 $p^*(\mathbf{u}, \mathbf{w})$ 而言，也必然是 (\mathbf{u}, \mathbf{w}) 的联合凸函数

性质 2: (全局敏感性) 若原问题为凸，且对偶间隙为零，有 $p^*(\mathbf{u}, \mathbf{w}) \geq p^*(\mathbf{0}, \mathbf{0}) - (\boldsymbol{\lambda}^*)^T \mathbf{u} - (\mathbf{v}^*)^T \mathbf{w}$

证明:

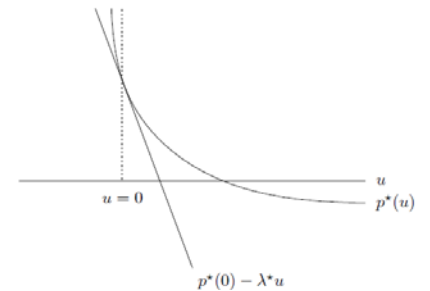
设 $(\mathbf{x}^*, \boldsymbol{\lambda}^*, \mathbf{v}^*)$ 是原问题的 Primal-Dual 最优解，则 $p^*(\mathbf{0}, \mathbf{0}) = f_o(\mathbf{x}^*)$

设 $\tilde{\mathbf{x}}$ 是干扰问题的最优解，则 $p^*(\mathbf{u}, \mathbf{w}) = f_o(\tilde{\mathbf{x}})$ ，且 $f_i(\tilde{\mathbf{x}}) \leq u_i$ ， $h_j(\tilde{\mathbf{x}}) = w_j$

因为原问题的对偶间隙为零，故有

$$\begin{aligned} p^*(\mathbf{0}, \mathbf{0}) = f_o(\mathbf{x}^*) & \stackrel{p^*=d^*}{=} g(\boldsymbol{\lambda}^*, \mathbf{v}^*) \leq f_o(\tilde{\mathbf{x}}) + \sum_{i=1}^m \lambda_i^* f_i(\tilde{\mathbf{x}}) + \sum_{j=1}^p v_j^* h_j(\tilde{\mathbf{x}}) \\ & \leq f_o(\tilde{\mathbf{x}}) + (\boldsymbol{\lambda}^*)^T \mathbf{u} + (\mathbf{v}^*)^T \mathbf{w} \\ & = p^*(\mathbf{u}, \mathbf{w}) + (\boldsymbol{\lambda}^*)^T \mathbf{u} + (\mathbf{v}^*)^T \mathbf{w} \end{aligned}$$

所以 $p^*(\mathbf{u}, \mathbf{w}) \geq p^*(\mathbf{0}, \mathbf{0}) - (\boldsymbol{\lambda}^*)^T \mathbf{u} - (\mathbf{v}^*)^T \mathbf{w}$



Optimal value $p^*(u)$ of a convex problem with one constraint $f_i(x) \leq u$.
For $u = 0$, we have the original unperturbed problem;
for $u < 0$ the constraint is tightened, and for $u > 0$ the constraint is loosened.
The affine function $p^*(0) - \lambda^* u$ is a lower bound on p^* .

Sensitivity interpretations

• If λ_i^* is large and we tighten the i^{th} constraint (i.e., choose $u_i < 0$), then the optimal value $p^*(\mathbf{u}, \mathbf{w})$ is guaranteed to increase greatly.

经济学解释 (λ_i^* 代表影子价格): 增加高价资源 (如黄金、稀土等) 的买入配额 (或减少高价资源的卖出配额), 则成本会急剧上升

• If V_j^* is large and positive and $w_j < 0$, or if V_j^* is large and negative and $w_j > 0$, then the optimal value $p^*(\mathbf{u}, \mathbf{w})$ is guaranteed to increase greatly.

• If λ_i^* is small, and we loosen the i^{th} constraint ($u_i > 0$), then the optimal value $p^*(\mathbf{u}, \mathbf{w})$ will not decrease too much.

经济学解释 (λ_i^* 代表影子价格): 减少低价资源 (如空气、用水等) 的买入配额 (或增加低价资源的卖出配额), 则成本不会明显下降

• If V_j^* is small and positive, and $w_j > 0$, or if V_j^* is small and negative and $w_j < 0$, then the optimal value $p^*(\mathbf{u}, \mathbf{w})$ will not decrease too much.

性质 3: (局部敏感性) 若 $p^*(\boldsymbol{u}, \boldsymbol{w})$ 在 $(\boldsymbol{u}, \boldsymbol{w}) = (\mathbf{0}, \mathbf{0})$ 处可微, 则有

$$\lambda_i^* = -\frac{\partial p^*(\mathbf{0}, \mathbf{0})}{\partial u_i}, \quad i = 1, \dots, m$$
$$v_j^* = -\frac{\partial p^*(\mathbf{0}, \mathbf{0})}{\partial v_j}, \quad j = 1, \dots, p$$

解释:

由性质 2, 有 $p^*(\boldsymbol{u}, \boldsymbol{w}) \geq p^*(\mathbf{0}, \mathbf{0}) - (\boldsymbol{\lambda}^*)^T \boldsymbol{u} - (\boldsymbol{v}^*)^T \boldsymbol{w}$

若 $(\boldsymbol{u}, \boldsymbol{w})$ 在 $(\mathbf{0}, \mathbf{0})$ 的邻域内, 则上述不等式左右两边近似取等号, 即 $p^*(\boldsymbol{u}, \boldsymbol{w}) \doteq p^*(\mathbf{0}, \mathbf{0}) - (\boldsymbol{\lambda}^*)^T \boldsymbol{u} - (\boldsymbol{v}^*)^T \boldsymbol{w}$

分别对 $\boldsymbol{u}, \boldsymbol{w}$ 求导, 即得性质 3

从 Primal-Dual 角度理解算法

例: Boolean LP

$$\text{Boolean LP: } \begin{cases} \min & \mathbf{c}^T \mathbf{x} \\ \text{s.t.} & \mathbf{Ax} \leq \mathbf{b} \\ & x_i \in \{0, 1\} \quad i = 1, \dots, n \end{cases} \quad (\text{非凸问题})$$

$$\Rightarrow \begin{cases} \min & \mathbf{c}^T \mathbf{x} \\ \text{s.t.} & \mathbf{Ax} \leq \mathbf{b} \\ & x_i(x_i - 1) = 0 \quad i = 1, \dots, n \end{cases}$$

该问题的拉格朗日函数

$$\begin{aligned} L(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{v}) &= \mathbf{c}^T \mathbf{x} + \boldsymbol{\lambda}^T (\mathbf{Ax} - \mathbf{b}) + \sum_{i=1}^n v_i (x_i^2 - x_i) \\ &= \sum_{i=1}^n v_i x_i^2 + (\mathbf{c} + \mathbf{A}^T \boldsymbol{\lambda} - \mathbf{v})^T \mathbf{x} - \boldsymbol{\lambda}^T \mathbf{b} \end{aligned}$$

该问题的对偶函数

$$\begin{aligned} g(\boldsymbol{\lambda}, \mathbf{v}) &= \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{v}) \\ &= \begin{cases} -\frac{1}{4} \sum_{i=1}^n (c_i + \mathbf{a}_i^T \boldsymbol{\lambda} - v_i)^2 / v_i - \boldsymbol{\lambda}^T \mathbf{b} & \mathbf{v} \succeq 0 \quad x_i = -(c_i + \mathbf{a}_i^T \boldsymbol{\lambda} - v_i) / 2v_i \\ -\infty & \text{others} \end{cases} \end{aligned}$$

该问题的对偶问题

$$D: \begin{cases} \max_{\boldsymbol{\lambda}, \mathbf{v}} & -\frac{1}{4} \sum_{i=1}^n (c_i + \mathbf{a}_i^T \boldsymbol{\lambda} - v_i)^2 / v_i - \boldsymbol{\lambda}^T \mathbf{b} \\ \text{s.t.} & \boldsymbol{\lambda} \succeq 0 \\ & \mathbf{v} \succeq 0 \end{cases}$$

$$\max_{\boldsymbol{\lambda}, \mathbf{v}} g(\boldsymbol{\lambda}, \mathbf{v}) = \max_{\boldsymbol{\lambda}} \max_{\mathbf{v}} g(\boldsymbol{\lambda}, \mathbf{v})$$

$$\begin{aligned} \max_{\mathbf{v}} g(\boldsymbol{\lambda}, \mathbf{v}) &= \max_{\mathbf{v}} \left\{ -\frac{1}{4} \sum_{i=1}^n (c_i + \mathbf{a}_i^T \boldsymbol{\lambda} - v_i)^2 / v_i - \boldsymbol{\lambda}^T \mathbf{b} \right\} \\ &= \begin{cases} \sum_{i=1}^n (c_i + \mathbf{a}_i^T \boldsymbol{\lambda}) - \boldsymbol{\lambda}^T \mathbf{b} & c_i + \mathbf{a}_i^T \boldsymbol{\lambda} \leq 0 \quad v_i = -(c_i + \mathbf{a}_i^T \boldsymbol{\lambda}) \\ -\boldsymbol{\lambda}^T \mathbf{b} & c_i + \mathbf{a}_i^T \boldsymbol{\lambda} > 0 \end{cases} \\ &= \sum_{i=1}^n \min\{0, c_i + \mathbf{a}_i^T \boldsymbol{\lambda}\} - \boldsymbol{\lambda}^T \mathbf{b} \end{aligned}$$

对偶问题转化为

$$\begin{aligned} D &\Rightarrow \begin{cases} \max_{\boldsymbol{\lambda}} & \sum_{i=1}^n \min\{0, c_i + \mathbf{a}_i^T \boldsymbol{\lambda}\} - \boldsymbol{\lambda}^T \mathbf{b} \\ \text{s.t.} & \boldsymbol{\lambda} \succeq 0 \end{cases} \\ &\Rightarrow \begin{cases} \max_{\boldsymbol{\lambda}} & \sum_{i=1}^n w_i - \boldsymbol{\lambda}^T \mathbf{b} \\ \text{s.t.} & \boldsymbol{\lambda} \succeq 0 \\ & w_i = \min\{0, c_i + \mathbf{a}_i^T \boldsymbol{\lambda}\}, \quad \forall i \end{cases} \\ &\Rightarrow \begin{cases} \max_{\boldsymbol{\lambda}} & \sum_{i=1}^n w_i - \boldsymbol{\lambda}^T \mathbf{b} \\ \text{s.t.} & \boldsymbol{\lambda} \succeq 0 \\ & w_i \leq 0, \quad \forall i \\ & w_i \leq c_i + \mathbf{a}_i^T \boldsymbol{\lambda}, \quad \forall i \end{cases} \end{aligned}$$

$$\text{Relaxation of LP: } \begin{cases} \min & \mathbf{c}^T \mathbf{x} \\ \text{s.t.} & \mathbf{Ax} \leq \mathbf{b} \\ & 0 \leq x_i \leq 1 \quad i = 1, \dots, n \end{cases} \quad (\text{Boolean LP 松弛问题})$$

$$\Rightarrow \begin{cases} \min & \mathbf{c}^T \mathbf{x} \\ \text{s.t.} & \mathbf{Ax} \leq \mathbf{b} \\ & \mathbf{x} \succeq \mathbf{0} \\ & \mathbf{x} - \mathbf{1} \preceq \mathbf{0} \end{cases}$$

该问题的拉格朗日函数

$$\begin{aligned} L(\mathbf{x}, \mathbf{u}, \mathbf{v}, \mathbf{w}) &= \mathbf{c}^T \mathbf{x} + \mathbf{u}^T (\mathbf{Ax} - \mathbf{b}) - \mathbf{v}^T \mathbf{x} + \mathbf{w}^T (\mathbf{x} - \mathbf{1}) \\ &= (\mathbf{c} + \mathbf{A}^T \mathbf{u} - \mathbf{v} + \mathbf{w})^T \mathbf{x} - \mathbf{u}^T \mathbf{b} - \mathbf{1}^T \mathbf{w} \end{aligned}$$

该问题的对偶函数

$$\begin{aligned} g(\mathbf{u}, \mathbf{v}, \mathbf{w}) &= \inf_{\mathbf{x}} L(\mathbf{x}, \mathbf{u}, \mathbf{v}, \mathbf{w}) \\ &= \begin{cases} -\mathbf{u}^T \mathbf{b} - \mathbf{1}^T \mathbf{w} & \mathbf{c} + \mathbf{A}^T \mathbf{u} - \mathbf{v} + \mathbf{w} = \mathbf{0} \\ -\infty & \mathbf{c} + \mathbf{A}^T \mathbf{u} - \mathbf{v} + \mathbf{w} \neq \mathbf{0} \end{cases} \end{aligned}$$

该问题的对偶问题

$$D: \begin{cases} \max_{\mathbf{u}, \mathbf{v}, \mathbf{w}} & -\mathbf{u}^T \mathbf{b} - \mathbf{1}^T \mathbf{w} \\ \text{s.t.} & \mathbf{c} + \mathbf{A}^T \mathbf{u} - \mathbf{v} + \mathbf{w} = \mathbf{0} \\ & \mathbf{u} \succeq \mathbf{0} \\ & \mathbf{v} \succeq \mathbf{0} \\ & \mathbf{w} \succeq \mathbf{0} \end{cases}$$

消去 \mathbf{v} ，对偶问题转化为

$$D: \begin{cases} \max_{\mathbf{u}, \mathbf{w}} & -\mathbf{u}^T \mathbf{b} - \mathbf{1}^T \mathbf{w} \\ \text{s.t.} & \mathbf{c} + \mathbf{A}^T \mathbf{u} + \mathbf{w} \succeq \mathbf{0} \\ & \mathbf{u} \succeq \mathbf{0} \\ & \mathbf{w} \succeq \mathbf{0} \end{cases}$$

Boolean LP 是非凸问题，对偶间隙不为零；而 Relaxation of Boolean LP 是凸问题，对偶间隙为零。

Boolean LP 与 Relaxation of Boolean LP 的对偶问题是一样的。

例：罚函数法（Penalty Function Method）求解带等式约束的可微凸优化问题

$$\begin{cases} \min & f_o(\mathbf{x}) \\ \text{s.t.} & \mathbf{Ax} = \mathbf{b} \end{cases}$$

罚函数法：它将有约束优化问题转化为无约束的优化问题

$$\min \quad f_o(\mathbf{x}) + \alpha \|\mathbf{Ax} - \mathbf{b}\|_2^2$$

$$\begin{aligned} \frac{\partial \|\mathbf{Ax} - \mathbf{b}\|_2^2}{\partial \mathbf{x}} &= \frac{\partial (\mathbf{Ax} - \mathbf{b})^T (\mathbf{Ax} - \mathbf{b})}{\partial \mathbf{x}} \\ &= \frac{\partial ((\mathbf{Ax})^T (\mathbf{Ax}) - (\mathbf{Ax})^T \mathbf{b} - \mathbf{b}^T (\mathbf{Ax}) + \mathbf{b}^T \mathbf{b})}{\partial \mathbf{x}} \\ &= \frac{\partial (\mathbf{x}^T \mathbf{A}^T \mathbf{Ax} - \mathbf{x}^T \mathbf{A}^T \mathbf{b} - \mathbf{b}^T \mathbf{Ax} + \mathbf{b}^T \mathbf{b})}{\partial \mathbf{x}} \\ &= \frac{\partial (\mathbf{x}^T \mathbf{A}^T \mathbf{Ax} - 2\mathbf{b}^T \mathbf{Ax} + \mathbf{b}^T \mathbf{b})}{\partial \mathbf{x}} \\ &= 2\mathbf{A}^T \mathbf{Ax} - 2\mathbf{A}^T \mathbf{b} \end{aligned}$$

罚函数的最优解 $\tilde{\mathbf{x}}$ 满足 $\nabla f_o(\tilde{\mathbf{x}}) + 2\alpha \mathbf{A}^T (\mathbf{A}\tilde{\mathbf{x}} - \mathbf{b}) = \mathbf{0}$

罚函数的最优解 $\tilde{\mathbf{x}}$ 近似为原问题的最优解 \mathbf{x}^* ；且罚因子 α 越大， $\tilde{\mathbf{x}}$ 的近似程度越高

解释：

原问题的拉格朗日函数

$$L(\mathbf{x}, \mathbf{v}) = f_o(\mathbf{x}) + \mathbf{v}^T (\mathbf{Ax} - \mathbf{b})$$

原问题的对偶函数

$$g(\mathbf{v}) = \inf_{\mathbf{x}} L(\mathbf{x}, \mathbf{v}) = \inf_{\mathbf{x}} (f_o(\mathbf{x}) + \mathbf{v}^T (\mathbf{Ax} - \mathbf{b}))$$

原问题的对偶问题

$$\max_{\mathbf{v}} \quad g(\mathbf{v})$$

$$\begin{aligned} p^* = d^* = \max_{\mathbf{v}} g(\mathbf{v}) &\geq g(2\alpha(\mathbf{A}\tilde{\mathbf{x}} - \mathbf{b})) \\ &= \inf_{\mathbf{x}} (f_o(\mathbf{x}) + 2\alpha(\mathbf{A}\tilde{\mathbf{x}} - \mathbf{b})^T (\mathbf{Ax} - \mathbf{b})) \longleftarrow \frac{\partial (f_o(\mathbf{x}) + 2\alpha(\mathbf{A}\tilde{\mathbf{x}} - \mathbf{b})^T (\mathbf{Ax} - \mathbf{b}))}{\partial \mathbf{x}} = 0 \\ &= f_o(\tilde{\mathbf{x}}) + 2\alpha(\mathbf{A}\tilde{\mathbf{x}} - \mathbf{b})^T (\mathbf{A}\tilde{\mathbf{x}} - \mathbf{b}) \\ &= f_o(\tilde{\mathbf{x}}) + 2\alpha \|\mathbf{A}\tilde{\mathbf{x}} - \mathbf{b}\|_2^2 \\ &\geq f_o(\tilde{\mathbf{x}}) \end{aligned}$$

$\Rightarrow \nabla f_o(\mathbf{x}) + 2\alpha \mathbf{A}^T (\mathbf{A}\tilde{\mathbf{x}} - \mathbf{b}) = \mathbf{0}$
 $\Rightarrow \mathbf{x} = \tilde{\mathbf{x}}$

假如 $\tilde{\mathbf{x}}$ 不违反约束（即 $\mathbf{A}\tilde{\mathbf{x}} - \mathbf{b} = \mathbf{0}$ ），则 $f_o(\tilde{\mathbf{x}}) = p^*$ ， $\tilde{\mathbf{x}}$ 就是原问题的最优解

假如 $\tilde{\mathbf{x}}$ 违反约束（即 $\mathbf{A}\tilde{\mathbf{x}} - \mathbf{b} \neq \mathbf{0}$ ），则 $f_o(\tilde{\mathbf{x}}) < p^*$ ， $\tilde{\mathbf{x}}$ 近似于原问题的最优解

因为 $\tilde{\mathbf{x}}$ 满足 $\nabla f_o(\tilde{\mathbf{x}}) + 2\alpha \mathbf{A}^T (\mathbf{A}\tilde{\mathbf{x}} - \mathbf{b}) = \mathbf{0}$ ，罚因子 α 越大， $\tilde{\mathbf{x}}$ 越趋近于约束，也就越近似于原问题的最优解

例：对数障碍法（Log Barrier Method）求解带不等式约束的可微凸优化问题

$$\begin{cases} \min & f_o(\mathbf{x}) \\ \text{s.t.} & \mathbf{Ax} \succeq \mathbf{b} \end{cases}, \quad \mathbf{x} \in \mathbf{R}^n, \mathbf{A} \in \mathbf{R}^{m \times n}, \mathbf{b} \in \mathbf{R}^m$$

对数障碍法：它将有约束优化问题转化为无约束的优化问题

$$\min \quad f_o(\mathbf{x}) - \sum_{i=1}^m u_i \log(\mathbf{a}_i^T \mathbf{x} - b_i + \varepsilon)$$

对数障碍法的最优解 $\tilde{\mathbf{x}}$ 满足 $\nabla f_o(\tilde{\mathbf{x}}) - \sum_{i=1}^m u_i \frac{\mathbf{a}_i}{\mathbf{a}_i^T \tilde{\mathbf{x}} - b_i} = \mathbf{0}$ （不考虑 ε ）

解释：

原问题的拉格朗日函数

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f_o(\mathbf{x}) + \sum_{i=1}^m \lambda_i (b_i - \mathbf{a}_i^T \mathbf{x})$$

原问题的对偶函数

$$g(\mathbf{v}) = \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}) = \inf_{\mathbf{x}} \left(f_o(\mathbf{x}) + \sum_{i=1}^m \lambda_i (b_i - \mathbf{a}_i^T \mathbf{x}) \right)$$

原问题的对偶问题

$$\max_{\mathbf{v}} \quad g(\mathbf{v})$$

$$\begin{aligned} p^* = d^* &= \max_{\boldsymbol{\lambda}} g(\boldsymbol{\lambda}) \geq g\left(\tilde{\lambda}_i = \frac{u_i}{\mathbf{a}_i^T \tilde{\mathbf{x}} - b_i}\right) \\ &= \inf_{\mathbf{x}} \left(f_o(\mathbf{x}) + \sum_{i=1}^m \frac{u_i (b_i - \mathbf{a}_i^T \mathbf{x})}{\mathbf{a}_i^T \tilde{\mathbf{x}} - b_i} \right) \quad \leftarrow \frac{\partial \left(f_o(\mathbf{x}) + \sum_{i=1}^m \frac{u_i (b_i - \mathbf{a}_i^T \mathbf{x})}{\mathbf{a}_i^T \tilde{\mathbf{x}} - b_i} \right)}{\partial \mathbf{x}} = 0 \\ &= f_o(\tilde{\mathbf{x}}) + \sum_{i=1}^m \frac{u_i (b_i - \mathbf{a}_i^T \tilde{\mathbf{x}})}{\mathbf{a}_i^T \tilde{\mathbf{x}} - b_i} \\ &\Rightarrow \nabla f_o(\mathbf{x}) - \sum_{i=1}^m \frac{u_i \mathbf{a}_i}{\mathbf{a}_i^T \tilde{\mathbf{x}} - b_i} = 0 \\ &\Rightarrow \mathbf{x} = \tilde{\mathbf{x}} \end{aligned}$$

最优化算法之无约束优化

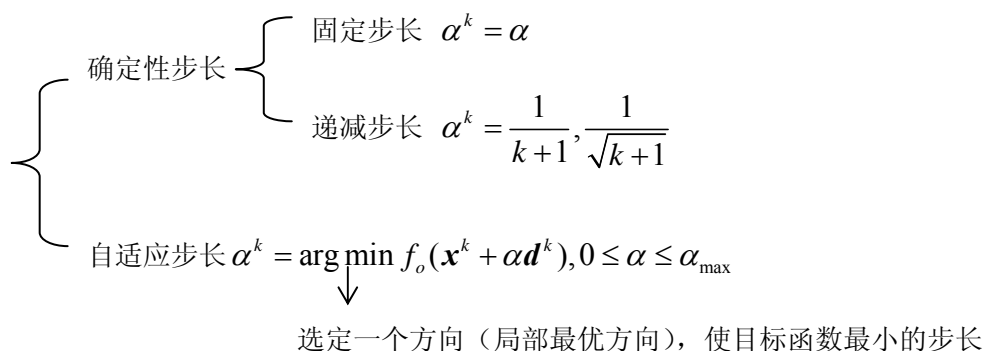
无约束优化: $\{\min f_o(\mathbf{x})\}$

#1 优化算法都是迭代的

基本结构: $\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha^k \mathbf{d}^k$

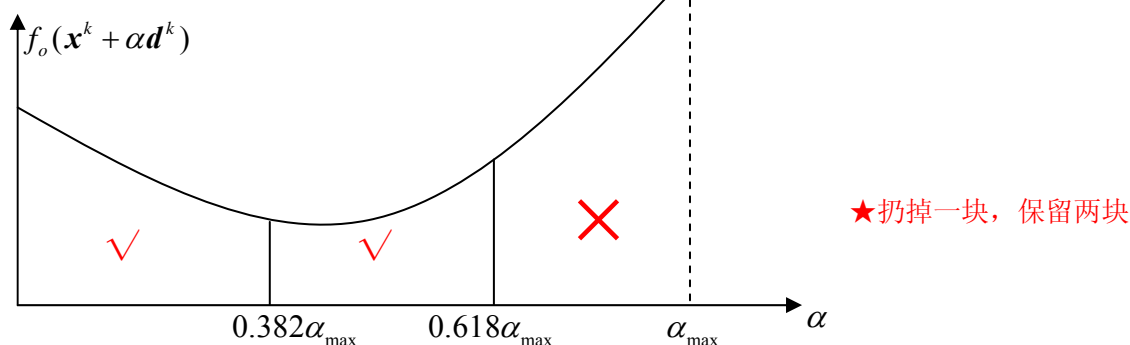
↓ ↘
步长 方向

步长选择:



#2 线搜索 (Line Search)

① 0.618 法 (黄金分割法, 优选法)



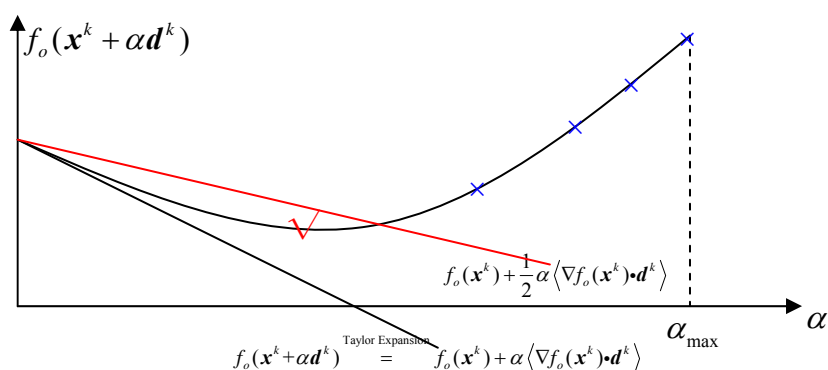
② Amijo Rule

$\alpha = \alpha_{\max}$ (γ 一般取 1/2)

若 $f_o(\mathbf{x}^k + \alpha \mathbf{d}^k) > f_o(\mathbf{x}^k) + \gamma \alpha \langle \nabla f_o(\mathbf{x}^k) \cdot \mathbf{d}^k \rangle$

则 $\alpha \leftarrow \alpha \beta$ ($\beta \in (0, 1)$: 缩减系数)

否则停止



无约束优化之梯度下降法 (Gradient Descent)

迭代结构: $\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha^k \mathbf{d}^k$

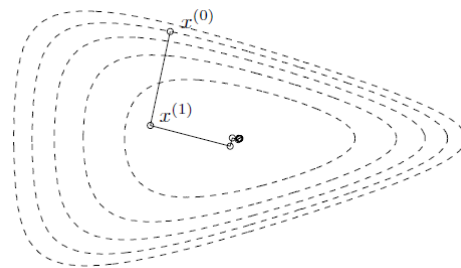
选定方向: $\mathbf{d}^k = -\nabla f_o(\mathbf{x}^k)$ —— 当前点负梯度方向

三个问题:

→ 能否收敛?

→ 收敛到哪里?

→ 收敛速度如何? (不同算法比较、参数影响)



在分析该算法之前, 首先给出假设 0、1、2

假设 0 目标函数 $f_o(\mathbf{x})$ 为可微凸函数, \mathbf{x}^* 存在且有限, f_o^* 有限

假设 1 (Lipschitz Continuous Gradient)

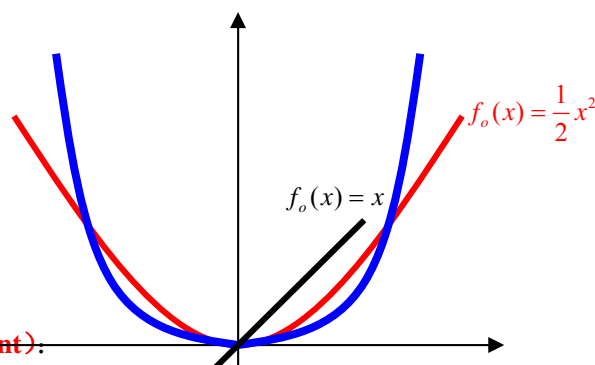
定义: Lipschitz Continuity 和 Lipschitz Continuous function

如果函数 $f(\mathbf{x})$ 满足如下条件:

对 $\forall \mathbf{x}, \mathbf{y}$, $\exists L > 0$, 使 $\|f(\mathbf{x}) - f(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\|$

则称函数 $f(\mathbf{x})$ 满足利普希茨连续性, 该函数称为利普希茨函数

意义: 函数值的变化不能快于点的变化的某一倍数 (即函数值不能变化太快)



目标函数 $f_o(\mathbf{x})$ 具有利普希茨连续梯度 (Lipschitz Continuous Gradient):

对 $\forall \mathbf{x}, \mathbf{y}$, $\exists L > 0$, 使 $\|\nabla f_o(\mathbf{x}) - \nabla f_o(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\|$

意义: 斜率的变化不能快于点的变化的某一倍数 (即斜率不能变化太快)

- 1) $f_o(x) = x$, $\nabla f_o(x) = 1$ 满足假设 1
- 2) $f_o(x) = \frac{1}{2}x^2$, $\nabla f_o(x) = x$ 满足假设 1
- 3) $f_o(x) = \frac{1}{4}x^4$, $\nabla f_o(x) = x^3$ 不满足假设 1

等价定义:

① 若 $f_o(\mathbf{x})$ 二阶可微, 则对 $\forall \mathbf{x}$, $\nabla^2 f_o(\mathbf{x}) \preceq L\mathbf{I}$

$$\textcircled{2} \frac{1}{L} \|\nabla f_o(\mathbf{x}) - \nabla f_o(\mathbf{y})\|^2 \leq \langle \nabla f_o(\mathbf{x}) - \nabla f_o(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \leq L \|\mathbf{x} - \mathbf{y}\|^2$$

由柯西施瓦茨不等式得到

$$\|a\| \leq L \|b\| \Rightarrow \begin{cases} a^T b \leq \|a\| \cdot \|b\| \leq L \|b\|^2 \\ \frac{1}{a^T b} \leq \left\| \frac{1}{a} \right\| \cdot \left\| \frac{1}{b} \right\| \leq L \left\| \frac{1}{a} \right\|^2 \end{cases}$$

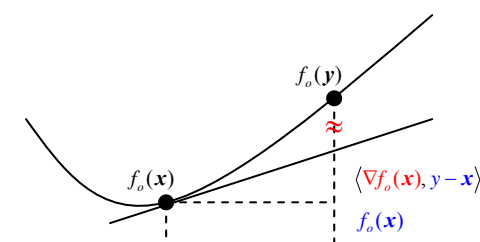
$$\textcircled{3} \frac{1}{2L} \|\nabla f_o(\mathbf{x}) - \nabla f_o(\mathbf{y})\|^2 \leq f_o(\mathbf{y}) - f_o(\mathbf{x}) - \langle \nabla f_o(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2$$

只证不等式右边, 构造 $g(\mathbf{x}) = \frac{L}{2} \mathbf{x}^T \mathbf{x} - f_o(\mathbf{x})$ 为凸函数 (因为 $\nabla^2 g(\mathbf{x}) = L\mathbf{I} - \nabla^2 f_o(\mathbf{x}) \succeq 0$)

由凸函数一阶条件 $g(\mathbf{y}) - g(\mathbf{x}) - \langle \nabla g(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \geq 0$

$$\left(\frac{L}{2} \mathbf{y}^T \mathbf{y} - f_o(\mathbf{y}) \right) - \left(\frac{L}{2} \mathbf{x}^T \mathbf{x} - f_o(\mathbf{x}) \right) - \langle L\mathbf{x} - \nabla f_o(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \geq 0$$

$$f_o(\mathbf{y}) - f_o(\mathbf{x}) - \langle \nabla f_o(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq \frac{L}{2} (\mathbf{x}^2 + \mathbf{y}^2 - 2\mathbf{x}^T \mathbf{y}) = \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2$$



假设 2 (Strong Convexity)

对 $\forall \mathbf{x}, \mathbf{y}$, $\exists u > 0$, 使 $f_o(\mathbf{y}) - f_o(\mathbf{x}) - \langle \nabla f_o(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \geq \frac{u}{2} \|\mathbf{x} - \mathbf{y}\|^2$

意义: 最优解唯一 (参见严格凸函数)

强凸必严格凸

反证: 若 $\exists \mathbf{x} \neq \mathbf{y}$, 使 $f_o(\mathbf{x}) = f_o(\mathbf{y}) = f_o^*$

$$\cancel{f_o(\mathbf{y}) - f_o(\mathbf{x}) - \langle \nabla f_o(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle} \geq \frac{u}{2} \|\mathbf{x} - \mathbf{y}\|^2, \text{ 矛盾}$$

故最优解唯一

等价定义:

① 若 $f_o(\mathbf{x})$ 二阶可微, 则对 $\forall \mathbf{x}$, $\nabla^2 f_o(\mathbf{x}) \succeq u\mathbf{I}$

$$\textcircled{2} u \|\mathbf{x} - \mathbf{y}\|^2 \leq \langle \nabla f_o(\mathbf{x}) - \nabla f_o(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \leq \frac{1}{u} \|\nabla f_o(\mathbf{x}) - \nabla f_o(\mathbf{y})\|^2$$

性质: 当假设 1、2 同时满足, 则有 $\langle \nabla f_o(\mathbf{x}) - \nabla f_o(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{uL}{u+L} \|\mathbf{x} - \mathbf{y}\|^2 + \frac{1}{u+L} \|\nabla f_o(\mathbf{x}) - \nabla f_o(\mathbf{y})\|^2$

证明:

i) 若 $L = u = \beta$

由假设 2, 有 $\langle \nabla f_o(\mathbf{x}) - \nabla f_o(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \beta \|\mathbf{x} - \mathbf{y}\|^2$

由假设 1, 有 $\langle \nabla f_o(\mathbf{x}) - \nabla f_o(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{1}{\beta} \|\nabla f_o(\mathbf{x}) - \nabla f_o(\mathbf{y})\|^2$

于是 $\langle \nabla f_o(\mathbf{x}) - \nabla f_o(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{\beta}{2} \|\mathbf{x} - \mathbf{y}\|^2 + \frac{1}{2\beta} \|\nabla f_o(\mathbf{x}) - \nabla f_o(\mathbf{y})\|^2$, 结论成立

ii) 若 $L > u$

$$\text{构造辅助函数 } \phi(\mathbf{x}) = f_o(\mathbf{x}) - \frac{u}{2} \|\mathbf{x}\|^2 \Rightarrow \begin{cases} \nabla \phi(\mathbf{x}) = \nabla f_o(\mathbf{x}) - u\mathbf{x} \\ \nabla^2 \phi(\mathbf{x}) = \nabla^2 f_o(\mathbf{x}) - u\mathbf{I} \end{cases}$$

由假设 2, $\nabla^2 \phi(\mathbf{x}) = \nabla^2 f_o(\mathbf{x}) - u\mathbf{I} \succeq \mathbf{0}$, 即 $\phi(\mathbf{x})$ 为凸函数

由假设 1, $\nabla^2 \phi(\mathbf{x}) = \nabla^2 f_o(\mathbf{x}) - u\mathbf{I} \preceq (L - u)\mathbf{I}$, 即 $\nabla \phi(\mathbf{x})$ 满足参数为 $L - u$ 的 Lipschitz 连续

$$\langle \nabla \phi(\mathbf{x}) - \nabla \phi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{1}{L - u} \|\nabla \phi(\mathbf{x}) - \nabla \phi(\mathbf{y})\|^2$$

$$\langle \nabla f_o(\mathbf{x}) - \nabla f_o(\mathbf{y}) - u(\mathbf{x} - \mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{1}{L - u} \|\nabla f_o(\mathbf{x}) - \nabla f_o(\mathbf{y}) - u(\mathbf{x} - \mathbf{y})\|^2$$

$$\langle \nabla f_o(\mathbf{x}) - \nabla f_o(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle - u \|\mathbf{x} - \mathbf{y}\|^2 \geq \frac{1}{L - u} \|\nabla f_o(\mathbf{x}) - \nabla f_o(\mathbf{y})\|^2 + \frac{u^2}{L - u} \|\mathbf{x} - \mathbf{y}\|^2 - \frac{2u}{L - u} \langle \nabla f_o(\mathbf{x}) - \nabla f_o(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$$

$$\frac{L + u}{L - u} \langle \nabla f_o(\mathbf{x}) - \nabla f_o(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{1}{L - u} \|\nabla f_o(\mathbf{x}) - \nabla f_o(\mathbf{y})\|^2 + \frac{Lu}{L - u} \|\mathbf{x} - \mathbf{y}\|^2$$

$$\langle \nabla f_o(\mathbf{x}) - \nabla f_o(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{uL}{u + L} \|\mathbf{x} - \mathbf{y}\|^2 + \frac{1}{u + L} \|\nabla f_o(\mathbf{x}) - \nabla f_o(\mathbf{y})\|^2$$

回到无约束优化算法之梯度下降法

迭代结构: $\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha^k \mathbf{d}^k$

选定方向: $\mathbf{d}^k = -\nabla f_o(\mathbf{x}^k)$ ——当前点负梯度方向

考虑固定步长: $\alpha^k = \alpha$

→非强凸情况, 分析函数值的收敛性质

→强凸情况, 分析点的收敛性质

定理: 若 $f_o(\mathbf{x})$ 满足**假设 0,1**, 步长 $\alpha \in \left(0, \frac{2}{L}\right)$, 则有

$$f_o(\mathbf{x}^k) - f_o^* \leq \frac{2(f_o(\mathbf{x}^0) - f_o^*) \|\mathbf{x}^0 - \mathbf{x}^*\|^2}{2\|\mathbf{x}^0 - \mathbf{x}^*\|^2 + \alpha(2 - L\alpha)(f_o(\mathbf{x}^0) - f_o^*)} \quad \text{即 } f_o(\mathbf{x}^k) \text{ 具有 } O\left(\frac{1}{k}\right) \text{ 的收敛性}$$

证明: ①点的单调性 (即: 迭代变量与最优解的距离单调减小)

$$\begin{aligned} \|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2 &= \|\mathbf{x}^k - \mathbf{x}^* - \alpha \nabla f_o(\mathbf{x}^k)\|^2 \\ &= \|\mathbf{x}^k - \mathbf{x}^*\|^2 - 2\alpha \langle \mathbf{x}^k - \mathbf{x}^*, \nabla f_o(\mathbf{x}^k) \rangle + \alpha^2 \|\nabla f_o(\mathbf{x}^k)\|^2 \\ &= \|\mathbf{x}^k - \mathbf{x}^*\|^2 - 2\alpha \langle \mathbf{x}^k - \mathbf{x}^*, \nabla f_o(\mathbf{x}^k) - \nabla f_o(\mathbf{x}^*) \rangle + \alpha^2 \|\nabla f_o(\mathbf{x}^k)\|^2 \quad \rightarrow \because \nabla f_o(\mathbf{x}^*) = 0 \\ &\leq \|\mathbf{x}^k - \mathbf{x}^*\|^2 - \frac{2\alpha}{L} \|\nabla f_o(\mathbf{x}^k)\|^2 + \alpha^2 \|\nabla f_o(\mathbf{x}^k)\|^2 \\ &= \|\mathbf{x}^k - \mathbf{x}^*\|^2 + \alpha\left(\alpha - \frac{2}{L}\right) \|\nabla f_o(\mathbf{x}^k)\|^2 \quad \checkmark \\ &\leq \|\mathbf{x}^k - \mathbf{x}^*\|^2 \end{aligned}$$

②值的单调性 (即: 目标函数值与最优值的距离单调减小)

$$\begin{aligned} f_o(\mathbf{x}^{k+1}) &\leq f_o(\mathbf{x}^k) + \langle \nabla f_o(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \frac{L}{2} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \\ &= f_o(\mathbf{x}^k) + \langle \nabla f_o(\mathbf{x}^k), -\alpha \nabla f_o(\mathbf{x}^k) \rangle + \frac{L}{2} \|\alpha \nabla f_o(\mathbf{x}^k)\|^2 \\ &= f_o(\mathbf{x}^k) - \alpha\left(1 - \frac{L\alpha}{2}\right) \|\nabla f_o(\mathbf{x}^k)\|^2 \\ f_o(\mathbf{x}^{k+1}) - f_o^* &\leq f_o(\mathbf{x}^k) - f_o^* - \alpha\left(1 - \frac{L\alpha}{2}\right) \|\nabla f_o(\mathbf{x}^k)\|^2 \\ \Delta^{k+1} &\leq \Delta^k - w \|\nabla f_o(\mathbf{x}^k)\|^2 \quad \checkmark \\ &\leq \Delta^k \end{aligned}$$

③为 $\Delta^{k+1} \leq \Delta^k - w \|\nabla f_o(\mathbf{x}^k)\|^2$ 中的 $\|\nabla f_o(\mathbf{x}^k)\|$ 寻找一个下界

$$\begin{aligned} f_o(\mathbf{x}^k) - f_o^* &\stackrel{\text{凸函数}}{\leq} \left\langle \nabla f_o(\mathbf{x}^k), \mathbf{x}^k - \mathbf{x}^* \right\rangle \stackrel{\text{柯西施瓦茨不等式}}{\leq} \|\nabla f_o(\mathbf{x}^k)\| \cdot \|\mathbf{x}^k - \mathbf{x}^*\| \stackrel{\text{点的单调性}}{\leq} \|\nabla f_o(\mathbf{x}^k)\| \cdot \|\mathbf{x}^0 - \mathbf{x}^*\| \\ \Rightarrow \|\nabla f_o(\mathbf{x}^k)\| &\geq \frac{\Delta^k}{\|\mathbf{x}^0 - \mathbf{x}^*\|} \end{aligned}$$

将结果代入 $\Delta^{k+1} \leq \Delta^k - w \|\nabla f_o(\mathbf{x}^k)\|^2$

$$\Rightarrow \Delta^{k+1} \leq \Delta^k - w \|\nabla f_o(\mathbf{x}^k)\|^2 \leq \Delta^k - w \frac{\|\Delta^k\|^2}{\|\mathbf{x}^0 - \mathbf{x}^*\|^2}$$

不等式两边同乘以 $\frac{1}{\Delta^k \Delta^{k+1}}$

$$\Rightarrow \frac{1}{\Delta^k} \leq \frac{1}{\Delta^{k+1}} - \frac{w}{\|\mathbf{x}^0 - \mathbf{x}^*\|^2} \frac{\Delta^k}{\Delta^{k+1}}$$

由值的单调性, $\frac{\Delta^k}{\Delta^{k+1}} \geq 1$

$$\Rightarrow \frac{1}{\Delta^k} \leq \frac{1}{\Delta^{k+1}} - \frac{w}{\|\mathbf{x}^0 - \mathbf{x}^*\|^2}$$

$$\begin{cases} \frac{1}{\Delta^0} \leq \frac{1}{\Delta^1} - \frac{w}{\|\mathbf{x}^0 - \mathbf{x}^*\|^2} \\ \vdots \\ \frac{1}{\Delta^k} \leq \frac{1}{\Delta^{k+1}} - \frac{w}{\|\mathbf{x}^0 - \mathbf{x}^*\|^2} \end{cases} \Rightarrow \frac{1}{\Delta^0} \leq \frac{1}{\Delta^k} - \frac{kw}{\|\mathbf{x}^0 - \mathbf{x}^*\|^2}, \text{ 得证}$$

根据该定理, $f_o(\mathbf{x}^k) - f_o^* \leq \frac{2(f_o(\mathbf{x}^0) - f_o^*)\|\mathbf{x}^0 - \mathbf{x}^*\|^2}{2\|\mathbf{x}^0 - \mathbf{x}^*\|^2 + k\alpha(2 - L\alpha)(f_o(\mathbf{x}^0) - f_o^*)}$, 可知 $\alpha(2 - L\alpha)$ 越大, 迭代步数 k 越小

极大化 $\alpha(2 - L\alpha)$, 可得最优步长 $\hat{\alpha} = \frac{1}{L}$

在最优步长下,

$$\begin{aligned} f_o(\mathbf{x}^k) - f_o^* &\leq \frac{2(f_o(\mathbf{x}^0) - f_o^*)\|\mathbf{x}^0 - \mathbf{x}^*\|^2}{2\|\mathbf{x}^0 - \mathbf{x}^*\|^2 + \frac{k}{L}(f_o(\mathbf{x}^0) - f_o^*)} \\ &\leq \frac{2 \cdot \frac{L}{2} \|\mathbf{x}^0 - \mathbf{x}^*\|^2 \cdot \|\mathbf{x}^0 - \mathbf{x}^*\|^2}{2\|\mathbf{x}^0 - \mathbf{x}^*\|^2 + \frac{k}{L} \frac{L}{2} \|\mathbf{x}^0 - \mathbf{x}^*\|^2} \\ &= \frac{2L}{k+4} \|\mathbf{x}^0 - \mathbf{x}^*\|^2 \end{aligned}$$

不等式右边关于 $f_o(\mathbf{x}^0) - f_o^*$ 递增

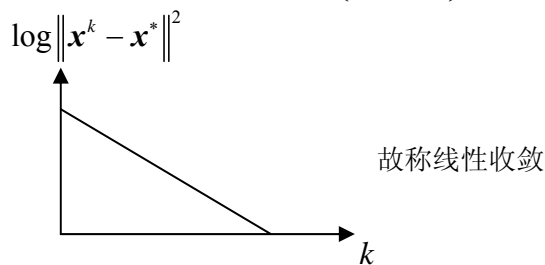
$$\begin{aligned} f_o(\mathbf{x}^0) &\leq f_o(\mathbf{x}^*) + \left\langle \nabla f_o(\mathbf{x}^*), \mathbf{x}^0 - \mathbf{x}^* \right\rangle + \frac{L}{2} \|\mathbf{x}^0 - \mathbf{x}^*\|^2 \\ &= f_o(\mathbf{x}^*) + \frac{L}{2} \|\mathbf{x}^0 - \mathbf{x}^*\|^2 \end{aligned}$$

定理：若 $f_o(\mathbf{x})$ 满足**假设 0,1,2**，Lipschitz 连续梯度常数为 L ，强凸常数为 u ；当步长 $\alpha \in \left(0, \frac{2}{u+L}\right)$ ，则有

$$\|\mathbf{x}^k - \mathbf{x}^*\|^2 = \left(1 - \frac{2\alpha u L}{u+L}\right)^k \|\mathbf{x}^0 - \mathbf{x}^*\|^2$$

意义：每迭代一次，当前解与最优解的距离便缩小一个比例

Because $1 - \frac{2\alpha u L}{u+L}$ is a constant that is less than 1



证明：

$$\begin{aligned} \|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2 &= \|\mathbf{x}^k - \mathbf{x}^* - \alpha \nabla f_o(\mathbf{x}^k)\|^2 \\ &= \|\mathbf{x}^k - \mathbf{x}^*\|^2 - 2\alpha \langle \mathbf{x}^k - \mathbf{x}^*, \nabla f_o(\mathbf{x}^k) \rangle + \alpha^2 \|\nabla f_o(\mathbf{x}^k)\|^2 \\ &= \|\mathbf{x}^k - \mathbf{x}^*\|^2 - 2\alpha \langle \mathbf{x}^k - \mathbf{x}^*, \nabla f_o(\mathbf{x}^k) - \nabla f_o(\mathbf{x}^*) \rangle + \alpha^2 \|\nabla f_o(\mathbf{x}^k)\|^2 \\ &\leq \|\mathbf{x}^k - \mathbf{x}^*\|^2 - 2\alpha \left(\frac{uL}{u+L} \|\mathbf{x}^k - \mathbf{x}^*\|^2 + \frac{1}{u+L} \|\nabla f_o(\mathbf{x})\|^2 \right) + \alpha^2 \|\nabla f_o(\mathbf{x}^k)\|^2 \\ &= \|\mathbf{x}^k - \mathbf{x}^*\|^2 - \frac{2\alpha u L}{u+L} \|\mathbf{x}^k - \mathbf{x}^*\|^2 + \alpha \left(\alpha - \frac{2}{u+L} \right) \|\nabla f_o(\mathbf{x}^k)\|^2 \\ &\leq \left(1 - \frac{2\alpha u L}{u+L} \right) \|\mathbf{x}^k - \mathbf{x}^*\|^2 \end{aligned}$$

由此递推关系，可得 $\|\mathbf{x}^k - \mathbf{x}^*\|^2 = \left(1 - \frac{2\alpha u L}{u+L}\right)^k \|\mathbf{x}^0 - \mathbf{x}^*\|^2$

极小化 $1 - \frac{2\alpha u L}{u+L}$ ，可得**最优步长** $\hat{\alpha} = \frac{2}{u+L}$

收敛速度 $1 - \frac{2\hat{\alpha} u L}{u+L} = 1 - \frac{4uL}{(u+L)^2} = \frac{(L-u)^2}{(L+u)^2}$

$$\|\mathbf{x}^k - \mathbf{x}^*\|^2 = \frac{(L-u)^{2k}}{(L+u)^{2k}} \|\mathbf{x}^0 - \mathbf{x}^*\|^2 = \frac{\left(\frac{L}{L+u}\right)^{2k}}{\left(\frac{L}{L+u}\right)^{2k}} \|\mathbf{x}^0 - \mathbf{x}^*\|^2$$

$\frac{L}{u}$ 即为 Hessian 矩阵 $\nabla^2 f_o(\mathbf{x}^k)$ 的**条件数**。

上式说明，**条件数**越小，收敛速度越快。

举个例子

$$f_o(\mathbf{x}) = \frac{1}{2} \|\mathbf{x}\|^2$$

$u=1$ ， $L=1$ 。最优步长 $\hat{\alpha}=1$ ，条件数为 1。

$$\mathbf{x}^1 = \mathbf{x}^0 - \hat{\alpha} \nabla f_o(\mathbf{x}^0) = \mathbf{x}^0 - \hat{\alpha} \mathbf{x}^0 = \mathbf{0}$$

亦即梯度下降法能一步达到最优解

In the field of numerical analysis, the **condition number** (条件数) of a **function** with respect to an argument measures how much the output value of the function can change for a small change in the input argument.

When the condition number associate with the linear equation $\mathbf{Ax} = \mathbf{b}$, the condition number of the matrix \mathbf{A} is defined as

$$\kappa(\mathbf{A}) = \|\mathbf{A}^{-1}\| \cdot \|\mathbf{A}\|$$

Of course, this definition depends on the choice of norm:

If $\|\cdot\|$ is the **norm** (usually noted as $\|\cdot\|_2$), then

$$\kappa(\mathbf{A}) = \frac{\sigma_{\max}(\mathbf{A})}{\sigma_{\min}(\mathbf{A})}$$

A problem with a low condition number is said to be **well-conditioned** (良态的), while a problem with a high condition number is said to be **ill-conditioned** (病态的).

该结论亦说明，若目标函数的 Hessian 矩阵**条件数很差**时，**梯度下降法的下降速度是很糟糕的**。

前面分析的都是固定步长的收敛性质，下面将讨论可变步长的收敛性质

迭代结构: $\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha^k \nabla f_o(\mathbf{x}^k)$

其中 α^k 随 k 的变化而变化

假定 $f_o(\mathbf{x})$ 满足假设 0,1,2

$$\begin{aligned}\tilde{f}_o(\alpha^k) &= f_o(\mathbf{x}^{k+1}) \leq f_o(\mathbf{x}^k) + \langle \nabla f_o(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \frac{L}{2} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \\ &= f_o(\mathbf{x}^k) + \langle \nabla f_o(\mathbf{x}^k), -\alpha^k \nabla f_o(\mathbf{x}^k) \rangle + \frac{L}{2} \|-\alpha^k \nabla f_o(\mathbf{x}^k)\|^2 \\ &= f_o(\mathbf{x}^k) + \frac{L \cdot (\alpha^k)^2 - 2\alpha^k}{2} \|\nabla f_o(\mathbf{x}^k)\|^2\end{aligned}$$

① 精确搜索 $\alpha^k = \alpha_{\text{exact}}^k$

当 $\alpha^k = \frac{1}{L}$ 时, $\frac{L \cdot (\alpha^k)^2 - 2\alpha^k}{2}$ 取得极小值 $-\frac{1}{2L}$

$$f_o(\mathbf{x}^{k+1}) \leq f_o(\mathbf{x}^k) - \frac{1}{2L} \|\nabla f_o(\mathbf{x}^k)\|^2$$

$$f_o(\mathbf{x}^{k+1}) - f_o^* \leq f_o(\mathbf{x}^k) - f_o^* - \frac{1}{2L} \|\nabla f_o(\mathbf{x}^k)\|^2$$

为梯度项 $\|\nabla f_o(\mathbf{x}^k)\|$ 寻找一个下界

$$\text{结论: } f_o^* \geq f_o(\mathbf{x}^k) - \frac{1}{2u} \|\nabla f_o(\mathbf{x}^k)\|^2$$

$$\text{证明: } f_o^* \geq f_o(\mathbf{x}^k) + \langle \nabla f_o(\mathbf{x}^k), \mathbf{x}^* - \mathbf{x}^k \rangle + \frac{u}{2} \|\mathbf{x}^* - \mathbf{x}^k\|^2$$

$$\begin{aligned}\text{Cauchy-Schwarz Inequality} \\ \geq \frac{-1}{2(a^2 + b^2) \leq a \cdot b \leq |a| \cdot |b|} f_o(\mathbf{x}^k) - \frac{1}{2u} \|\nabla f_o(\mathbf{x}^k)\|^2 - \frac{u}{2} \|\mathbf{x}^* - \mathbf{x}^k\|^2 + \frac{u}{2} \|\mathbf{x}^* - \mathbf{x}^k\|^2 \\ = f_o(\mathbf{x}^k) - \frac{1}{2u} \|\nabla f_o(\mathbf{x}^k)\|^2\end{aligned}$$

于是

$$f_o(\mathbf{x}^{k+1}) - f_o^* \leq (1 - \frac{u}{L})(f_o(\mathbf{x}^k) - f_o^*)$$

由此式观之, 函数值的收敛性依然与条件数 $\frac{L}{u}$ 有关, 条件数越小, 收敛速度越快。

迭代结构: $\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha^k \nabla f_o(\mathbf{x}^k)$

其中 α^k 随 k 的变化而变化

假定 $f_o(\mathbf{x})$ 满足 **假设 0,1,2**

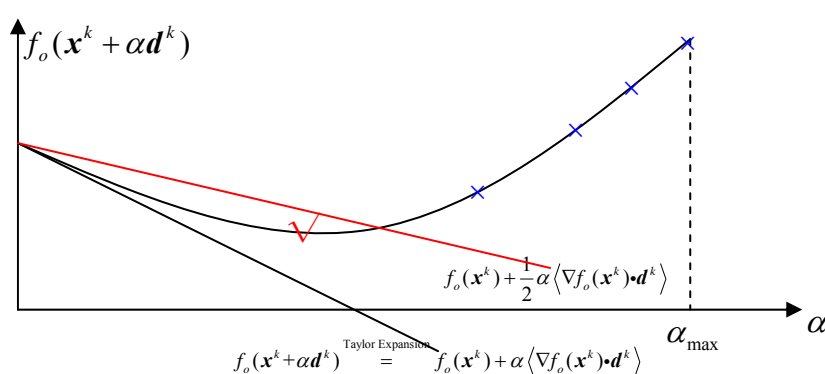
$$\begin{aligned}\tilde{f}_o(\alpha^k) &= f_o(\mathbf{x}^{k+1}) \leq f_o(\mathbf{x}^k) + \langle \nabla f_o(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \frac{L}{2} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \\ &= f_o(\mathbf{x}^k) + \langle \nabla f_o(\mathbf{x}^k), -\alpha^k \nabla f_o(\mathbf{x}^k) \rangle + \frac{L}{2} \|-\alpha^k \nabla f_o(\mathbf{x}^k)\|^2 \\ &= f_o(\mathbf{x}^k) + \frac{L \cdot (\alpha^k)^2 - 2\alpha^k}{2} \|\nabla f_o(\mathbf{x}^k)\|^2\end{aligned}$$

② Amijo Rule

$$\left\{ \begin{array}{l} \alpha = \alpha_{\max} \\ \text{若 } f_o(\mathbf{x}^{k+1}) > f_o(\mathbf{x}^k) + \gamma \alpha \langle \nabla f_o(\mathbf{x}^k), \mathbf{d}^k \rangle \\ \text{则 } \alpha \leftarrow \alpha \beta \\ \text{否则停止} \end{array} \right.$$

$\gamma \in (0, \frac{1}{2})$

$\beta \in (0, 1)$: 缩减系数



选定 α^k , 则 α^k 满足

$$\tilde{f}_o(\alpha^k) = f_o(\mathbf{x}^{k+1}) \leq f_o(\mathbf{x}^k) - \gamma \alpha^k \|\nabla f_o(\mathbf{x}^k)\|^2$$

若当前 α^k 足够小, α^k 必然被接受

$$\text{当 } 0 \leq \alpha^k \leq \frac{1}{L} \text{ 时, } \frac{L \cdot (\alpha^k)^2 - 2\alpha^k}{2} \leq -\frac{\alpha^k}{2}$$

$$f_o(\mathbf{x}^{k+1}) \leq f_o(\mathbf{x}^k) - \frac{\alpha^k}{2} \|\nabla f_o(\mathbf{x}^k)\|^2$$

$$\leq f_o(\mathbf{x}^k) - \gamma \alpha^k \|\nabla f_o(\mathbf{x}^k)\|^2$$

所选步长 α^k 满足 Amijo Rule, α^k 必然被接受

故 Amijo Rule 停止于 $\alpha^k = \alpha_{\max}$ 或 $\alpha^k \geq \frac{\beta}{L}$

$$\alpha^k \geq \min \left\{ \alpha_{\max}, \frac{\beta}{L} \right\}$$

$$f_o(\mathbf{x}^{k+1}) \leq f_o(\mathbf{x}^k) - \min \left\{ \gamma \alpha_{\max}, \frac{\gamma \beta}{L} \right\} \|\nabla f_o(\mathbf{x}^k)\|^2$$

$$\begin{aligned}f_o(\mathbf{x}^{k+1}) - f_o^* &\leq f_o(\mathbf{x}^k) - f_o^* - \min \left\{ \gamma \alpha_{\max}, \frac{\gamma \beta}{L} \right\} \|\nabla f_o(\mathbf{x}^k)\|^2 \\ &\leq \left(1 - \min \left\{ 2\gamma \alpha_{\max}, \frac{2\gamma \beta}{L} \right\} \right) (f_o(\mathbf{x}^k) - f_o^*)\end{aligned}$$

对于光滑的优化问题,
千万不要用递减步长,
收敛速度太慢

若刚开始选定的 $\alpha_{\max} \leq \frac{1}{L}$

则 α_{\max} 立马被接受

若 $\alpha^k > \frac{\beta}{L}$, 则上一步迭代的 $\alpha^k < \frac{1}{L}$

那么上一步迭代的 α^k 就被接受

亦即上一步迭代就终止, 矛盾

函数值的收敛性依然与条件数 $\frac{L}{u}$ 有关

正则化的梯度下降法

梯度下降法的推广

迭代结构: $\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha^k \mathbf{d}^k$

对方向的模做一个限制
否则达到负无穷

选定方向: $\mathbf{d}^k = \arg \min_{\mathbf{v}} \left\{ \nabla^T f_o(\mathbf{x}^k) \mathbf{v} \mid \|\mathbf{v}\| = 1 \right\}$

意义:

将 $f_o(\mathbf{x}^k + \mathbf{v})$ 泰勒展开, $f_o(\mathbf{x}^k + \mathbf{v}) \approx f_o(\mathbf{x}^k) + \nabla^T f_o(\mathbf{x}^k) \mathbf{v}$

本质就是极小化目标函数 f_o 在 $\mathbf{x}^k + \mathbf{v}$ 点的一阶增量

1) 若 $\mathbf{d}^k = \arg \min_{\mathbf{v}} \left\{ \nabla^T f_o(\mathbf{x}^k) \mathbf{v} \mid \|\mathbf{v}\|_2 = 1 \right\}$

则 $\mathbf{d}^k = \frac{-\nabla f_o(\mathbf{x}^k)}{\|\nabla f_o(\mathbf{x}^k)\|_2}$ 与梯度下降法方向一样, 幅度不一样

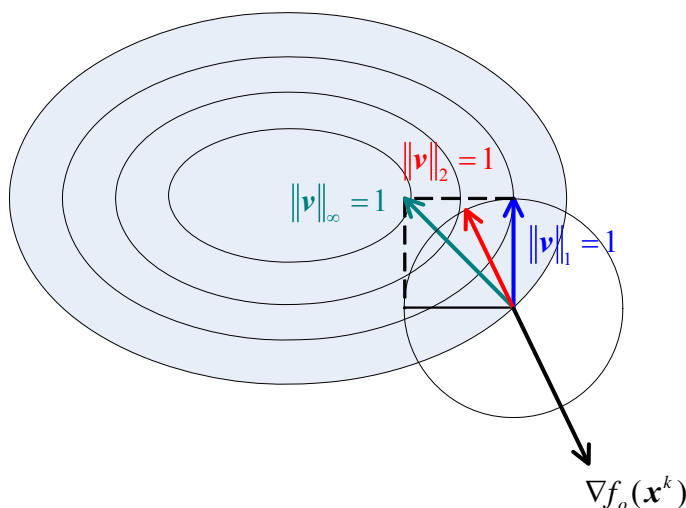
2) 若 $\mathbf{d}^k = \arg \min_{\mathbf{v}} \left\{ \nabla^T f_o(\mathbf{x}^k) \mathbf{v} \mid \|\mathbf{v}\|_\infty = 1 \right\}$

则 $\begin{cases} (\mathbf{d}^k)_i = 1 & (\nabla f_o(\mathbf{x}^k))_i \leq 0 \\ (\mathbf{d}^k)_i = -1 & (\nabla f_o(\mathbf{x}^k))_i > 0 \end{cases}$ \mathbf{d}^k 的元素要么为 1, 要么为 -1

3) 若 $\mathbf{d}^k = \arg \min_{\mathbf{v}} \left\{ \nabla^T f_o(\mathbf{x}^k) \mathbf{v} \mid \|\mathbf{v}\|_1 = 1 \right\}$

则 $\begin{cases} (\mathbf{d}^k)_j = \begin{cases} 1 & \text{when } (\nabla f_o(\mathbf{x}^k))_j < 0 \text{ and } |(\nabla f_o(\mathbf{x}^k))_j| \text{ is the maximum element} \\ 0 & \text{others} \end{cases} \\ (\mathbf{d}^k)_j = \begin{cases} -1 & \text{when } (\nabla f_o(\mathbf{x}^k))_j > 0 \text{ and } |(\nabla f_o(\mathbf{x}^k))_j| \text{ is the maximum element} \\ 0 & \text{others} \end{cases} \end{cases}$

\mathbf{d}^k 只有一个元素为 1 或 -1, 其余为 0, 非零元素集中力量专攻 $\nabla f_o(\mathbf{x}^k)$ 中绝对值最大的元素



坐标转换法（Coordinate Descent）

前提：目标函数一阶可微

算法：

$$\mathbf{x} \in \mathbf{R}^n$$

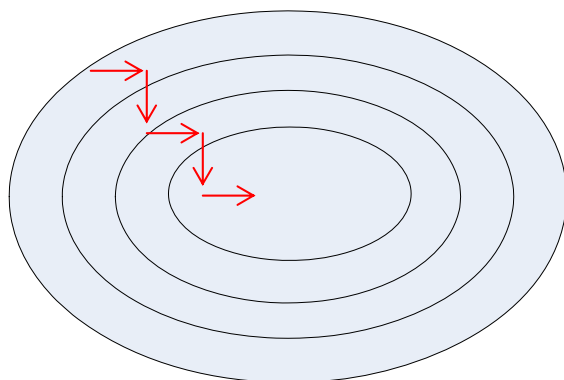
迭代步数

$$\mathbf{d}^k = \vec{e} \bmod (k, n) \longrightarrow \text{维数}$$

只有 n 个方向可供选择；每迭代 n 步，方向循环轮换一次

$$\alpha^k = \arg \min f_o(\mathbf{x}^k + \alpha \mathbf{d}^k), \quad -\alpha_{\max} \leq \alpha \leq \alpha_{\max}$$

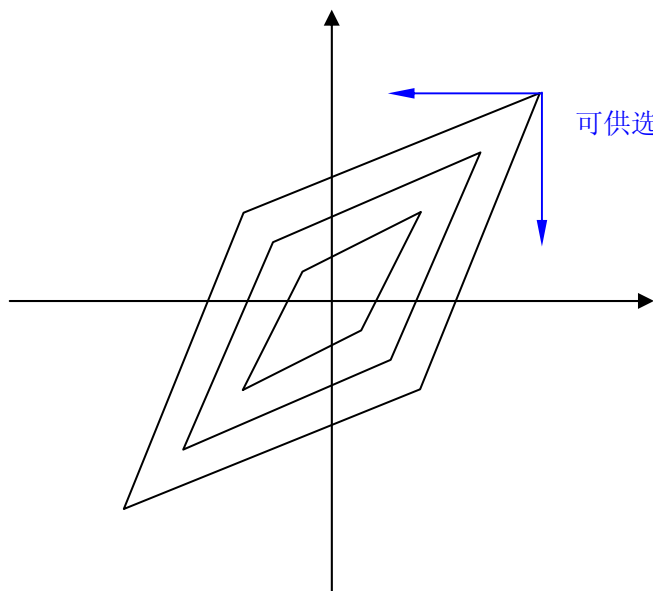
修正方向选择的错误，通过负步长弥补



注意：

坐标轮换法千万不能用固定步长

坐标轮换法不适用于不可微的目标函数



可供选择的方向都不能使函数值下降

总结：梯度下降法的收敛速度决定于条件数，既然改步长的方式被条件数所限制，那么就采用改方向的方法这是后面牛顿法所要讨论的内容

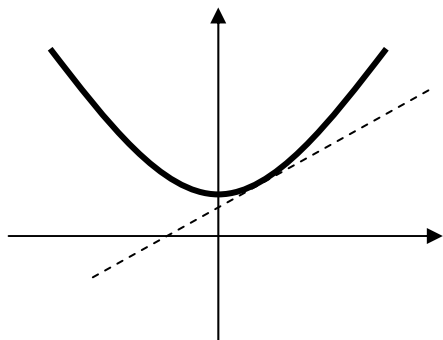
次梯度方法 (Sub-gradient Method)

将梯度下降法应用于不光滑的目标函数上

次梯度 $\mathbf{g}_o(\mathbf{x}) \in \partial f_o(\mathbf{x})$ 次微分

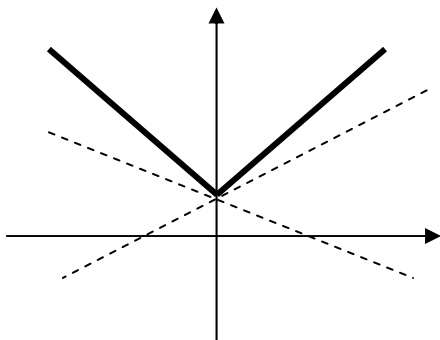
对 $\forall \mathbf{y}$, 有 $f_o(\mathbf{y}) \geq f_o(\mathbf{x}) + \mathbf{g}_o^T(\mathbf{x})(\mathbf{y} - \mathbf{x})$, 则称 $\mathbf{g}_o(\mathbf{x})$ 称为次梯度 (subgradient)

所有次梯度的集合 $\partial f_o(\mathbf{x})$ 称为次微分 (subdifferential)



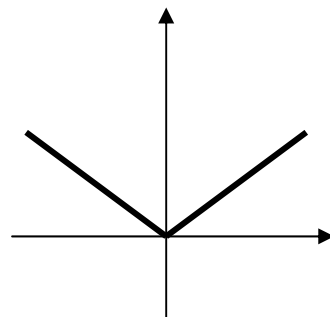
对于光滑曲线的任一点，
次梯度只有一个，即

$$\nabla f_o(\mathbf{x})$$



对于曲线上某一不可微点，
所有次梯度构成一个集合，

$$\text{即 } \partial f_o(\mathbf{x})$$



例: $y = |x|$

当 $x = 0$ 时, $\partial |x| = [-1, 1]$

次梯度法

迭代结构: $\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha^k \mathbf{d}^k$

选定方向: $\mathbf{d}^k = -\mathbf{g}_o(\mathbf{x}^k)$ —— 当前点负次梯度方向

步长规则:

有以下三种情况:

① 固定步长: $\alpha^k = \alpha$

② 步长序列不可加, 但平方可加: $\sum_{k=0}^{\infty} \alpha^k = \infty$, $\sum_{k=0}^{\infty} (\alpha^k)^2 < \infty$

$$\text{例: } \alpha^k = \frac{a}{k+b}$$

③ 步长序列不可加、平方不可加, 但趋于零: $\lim_{k \rightarrow \infty} \alpha^k = 0$

$$\text{例: } \alpha^k = \frac{a}{\sqrt{k+b}}$$

前面已经说过, 对于光滑的优化问题, 千万不要用递减步长, 收敛速度太慢;
但对于不可微问题, 不可避免地要使用递减步长。

理论分析

不再假设目标函数 $f_o(\mathbf{x})$ 满足 **Lipschitz 连续梯度**和**强凸性**

假设目标函数 $f_o(\mathbf{x})$ 满足 **Lipschitz 连续**，即

$$\text{对 } \forall \mathbf{x}, \mathbf{y}, \exists G > 0, \text{ 使 } \|f_o(\mathbf{x}) - f_o(\mathbf{y})\| \leq G \|\mathbf{x} - \mathbf{y}\|$$

意义：函数值的变化不能快于点的变化的某一倍数（即次梯度有界： $\|\mathbf{g}_o(\mathbf{x})\| \leq G$ ）

设 \mathbf{x}^* 为最优解

$$\begin{aligned}\|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2 &= \|\mathbf{x}^k - \mathbf{x}^* - \alpha^k \mathbf{g}_o(\mathbf{x}^k)\|^2 \\&= \|\mathbf{x}^k - \mathbf{x}^*\|^2 - 2\alpha^k \langle \mathbf{x}^k - \mathbf{x}^*, \mathbf{g}_o(\mathbf{x}^k) \rangle + (\alpha^k)^2 \|\mathbf{g}_o(\mathbf{x}^k)\|^2 \\&\leq \|\mathbf{x}^k - \mathbf{x}^*\|^2 - 2\alpha^k (f_o(\mathbf{x}^k) - f_o^*) + (\alpha^k)^2 \|\mathbf{g}_o(\mathbf{x}^k)\|^2 \\&\leq \|\mathbf{x}^0 - \mathbf{x}^*\|^2 - 2 \sum_{i=0}^k \alpha^i (f_o(\mathbf{x}^i) - f_o^*) + \sum_{i=0}^k (\alpha^i)^2 \|\mathbf{g}_o(\mathbf{x}^i)\|^2 \\&\Rightarrow \|\mathbf{x}^0 - \mathbf{x}^*\|^2 + \sum_{i=0}^k (\alpha^i)^2 \|\mathbf{g}_o(\mathbf{x}^i)\|^2 \geq 2 \sum_{i=0}^k \alpha^i (f_o(\mathbf{x}^i) - f_o^*) \geq 2 \left(\sum_{i=0}^k \alpha^i \right) \min_i (f_o(\mathbf{x}^i) - f_o^*) \\&\Rightarrow \min_i (f_o(\mathbf{x}^i) - f_o^*) \leq \frac{\|\mathbf{x}^0 - \mathbf{x}^*\|^2 + \sum_{i=0}^k (\alpha^i)^2 \|\mathbf{g}_o(\mathbf{x}^i)\|^2}{2 \sum_{i=0}^k \alpha^i} \leq \frac{\sum_{i=0}^k (\alpha^i)^2}{2 \sum_{i=0}^k \alpha^i} G^2\end{aligned}$$

第一种步长 $\alpha^k = \alpha$

$$\text{当 } k \rightarrow \infty \text{ 时, } \min_i (f_o(\mathbf{x}^i) - f_o^*) \leq \frac{\alpha}{2} G^2$$

即固定步长是没有办法保证严格收敛性的

$$\text{第二种步长 } \alpha^k = \frac{1}{k+1}, \text{ 分母 } \sum_{i=0}^k \frac{1}{k+1} \sim 1 + \log(k+1), \text{ 分子 } \sum_{i=0}^k \left(\frac{1}{k+1}\right)^2 \sim \frac{1}{k+1}$$

$$\text{当 } k \rightarrow \infty \text{ 时, } \min_i (f_o(\mathbf{x}^i) - f_o^*) \rightarrow 0$$

可保证收敛性

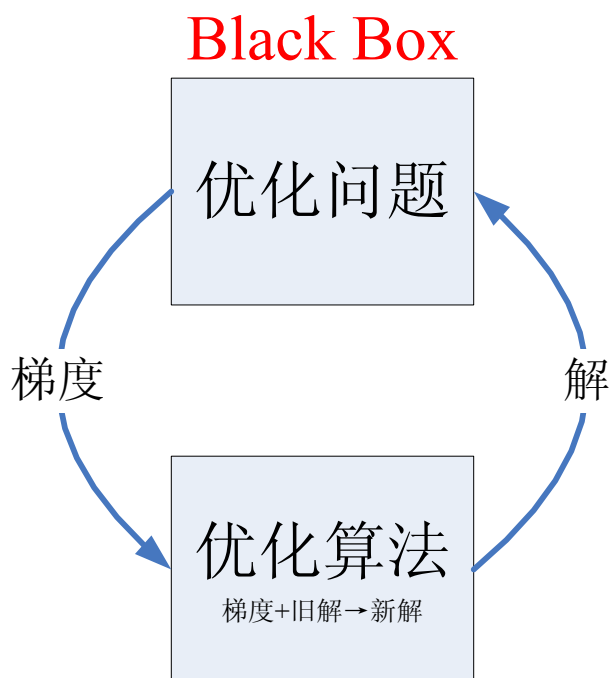
$$\text{第一种步长 } \alpha^k = \frac{1}{\sqrt{k+1}}, \text{ 分母 } \sum_{i=0}^k \frac{1}{\sqrt{k+1}} \sim \sqrt{k+1}, \text{ 分子 } \sum_{i=0}^k \frac{1}{k+1} \sim 1 + \log(k+1)$$

$$\text{当 } k \rightarrow \infty \text{ 时, } \min_i (f_o(\mathbf{x}^i) - f_o^*) \rightarrow 0$$

可保证收敛性

对偶平均法（Dual Averaging）/Nesterov

优化问题的黑箱模型



次梯度法迭代结构： $\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha^k \mathbf{g}_o(\mathbf{x}^k)$

由于采用**递减步长**，越老的信息越重要，越新的信息越不重要，这是一个逻辑硬伤。

由此 Nesterov 提出了**对偶平均法**

对偶平均法

\mathbf{S}^k 为所有点次梯度的加权和

$$\begin{cases} \mathbf{S}^{k+1} = \mathbf{S}^k + \alpha^k \mathbf{g}_o(\mathbf{x}^k) & \alpha^k > 0 \\ \mathbf{x}^{k+1} = \prod_{\beta^{k+1}}(-\mathbf{S}^{k+1}) & \beta^{k+1} > \beta^k, \quad \prod_{\beta}(-\mathbf{S}) = \arg \min_{\mathbf{x}} \{ \langle \mathbf{S}, \mathbf{x} \rangle + \beta d(\mathbf{x}) \} \end{cases}$$

β 为一个递增的参量 $d(\mathbf{x})$ 为一个距离函数

分析上述迭代结构代表的含义（取 $d(\mathbf{x}) = \frac{1}{2} \|\mathbf{x}\|^2$ ）

$$\mathbf{x}^{k+1} = \prod_{\beta^{k+1}}(-\mathbf{S}^{k+1}) = \arg \min_{\mathbf{x}} \left\{ \langle \mathbf{S}^{k+1}, \mathbf{x} \rangle + \frac{\beta^{k+1}}{2} \|\mathbf{x}\|^2 \right\} = -\frac{\mathbf{S}^{k+1}}{\beta^{k+1}}$$

$$\begin{cases} \mathbf{S}^{k+1} = \sum_{i=0}^k \alpha^i \mathbf{g}_o(\mathbf{x}^i) \\ \mathbf{x}^{k+1} = -\frac{1}{\beta^{k+1}} \sum_{i=0}^k \alpha^i \mathbf{g}_o(\mathbf{x}^i) \end{cases}$$

保证收敛

几种参数选择方式

1) α^k 为常数, $\beta^k \sim \sqrt{k+1}$

$$\min_i (f_o(\mathbf{x}^i) - f_o^*) \sim O\left(\frac{1}{\sqrt{k+1}}\right)$$

2) $\alpha^k \sim \frac{1}{\sqrt{k+1}}$, β^k 为常数

$$\min_i (f_o(\mathbf{x}^i) - f_o^*) \sim O\left(\frac{1 + \log(k+1)}{\sqrt{k+1}}\right)$$

3) $\alpha^k \sim \frac{1}{k+1}$, β^k 为常数

$$\min_i (f_o(\mathbf{x}^i) - f_o^*) \sim O\left(\frac{1}{1 + \log(k+1)}\right)$$

2) 3) 相当于次梯度法, 均劣于 1)

牛顿法 (Newton's Method)

无约束优化: $\{\min f_o(\mathbf{x})\}$

假设 $f_o(\mathbf{x})$ 二阶可微, 强凸

梯度下降法与牛顿法的比较

①梯度下降法的迭代方向: $\mathbf{d}^k = \arg \min_{\mathbf{v}} \left\{ \nabla^T f_o(\mathbf{x}^k) \mathbf{v} \mid \|\mathbf{v}\| = 1 \right\} \longrightarrow \mathbf{d}^k = -\nabla f_o(\mathbf{x}^k)$

将 $f_o(\mathbf{x}) = f_o(\mathbf{x}^k + \mathbf{v})$ 泰勒展开, $f_o(\mathbf{x}) \simeq f_o(\mathbf{x}^k) + \nabla^T f_o(\mathbf{x}^k) \mathbf{v}$

本质就是极小化目标函数 $f_o(\mathbf{x})$ 在 \mathbf{v} 方向的一阶展开

因为 $\nabla^2 f_o(\mathbf{x}^k)$ 强凸 \rightarrow 正定
所以无需对 \mathbf{v} 做约束

②牛顿法的迭代方向: $\mathbf{d}^k = \arg \min_{\mathbf{v}} \left\{ \nabla^T f_o(\mathbf{x}^k) \mathbf{v} + \frac{1}{2} \mathbf{v}^T \nabla^2 f_o(\mathbf{x}^k) \mathbf{v} \right\} \longrightarrow \mathbf{d}^k = -\left(\nabla^2 f_o(\mathbf{x}^k) \right)^{-1} \nabla f_o(\mathbf{x}^k)$

将 $f_o(\mathbf{x}) = f_o(\mathbf{x}^k + \mathbf{v})$ 泰勒展开, $f_o(\mathbf{x}) \simeq f_o(\mathbf{x}^k) + \nabla^T f_o(\mathbf{x}^k) \mathbf{v} + \frac{1}{2} \mathbf{v}^T \nabla^2 f_o(\mathbf{x}^k) \mathbf{v}$

本质就是极小化目标函数 $f_o(\mathbf{x})$ 在 \mathbf{v} 方向的二阶展开

$\left\{ \begin{array}{l} \text{梯度下降法的迭代方向: } \mathbf{d}^k = -\nabla f_o(\mathbf{x}^k) \\ \text{牛顿法的迭代方向: } \mathbf{d}^k = -\left(\nabla^2 f_o(\mathbf{x}^k) \right)^{-1} \nabla f_o(\mathbf{x}^k) \end{array} \right.$

牛顿法通过修正方向 (而不是步长) 来降低梯度下降法的收敛速度对条件数的敏感性

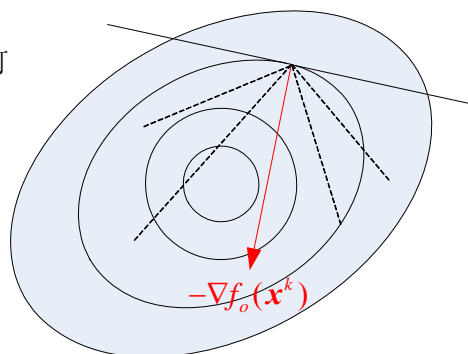
对于梯度下降法, 只要步长足够小, 其方向 \mathbf{d}^k 必然的是下降的方向

对于牛顿法, 其方向 \mathbf{d}^k 是否一定时下降的呢?

如右图所示, 只要求证牛顿法的方向与负梯度方向夹角小于 90° 即可

$$\begin{aligned} & \left\langle -\left(\nabla^2 f_o(\mathbf{x}^k) \right)^{-1} \nabla f_o(\mathbf{x}^k), -\nabla f_o(\mathbf{x}^k) \right\rangle \\ &= \left(\nabla^2 f_o(\mathbf{x}^k) \right)^{-1} \left\langle -\nabla f_o(\mathbf{x}^k), -\nabla f_o(\mathbf{x}^k) \right\rangle \geq 0 \end{aligned}$$

得证



牛顿法的收敛性分析

牛顿法迭代结构： $\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha^k \left(\nabla^2 f_o(\mathbf{x}^k) \right)^{-1} \nabla f_o(\mathbf{x}^k)$

假设 $\nabla^2 f_o(\mathbf{x})$ 上下有界，且 $\nabla^2 f_o(\mathbf{x})$ 满足 Lipschitz 连续

最大最小特征值有界 $\forall \mathbf{x}, \mathbf{y}, \quad \left\| \nabla^2 f_o(\mathbf{x}) - \nabla^2 f_o(\mathbf{y}) \right\| \leq L \left\| \mathbf{x} - \mathbf{y} \right\|$

若在牛顿法中步长采用 Armijo Rule

• 若 $\left\| \nabla f_o(\mathbf{x}) \right\|_2 > \eta$ ，阻尼牛顿段（Damped Newton Phase）： $f_o(\mathbf{x}^{k+1}) - f_o(\mathbf{x}^k) \leq -\gamma \qquad \gamma > 0$

• 若 $\left\| \nabla f_o(\mathbf{x}) \right\|_2 \leq \eta$ ，纯牛顿段（Pure Newton Phase）： $f_o(\mathbf{x}^{k+1}) - f_o^* \leq \sigma \left(\frac{1}{2} \right)^{2^k} \qquad \sigma > 0$
二次收敛

对比 $f_o(\mathbf{x}^{k+1}) - f_o^* \leq \sigma \varepsilon^k \qquad \sigma > 0$
线性收敛

对牛顿法，函数为二次函数时，采用合适的步长，可一步达到最优解

$$f_o(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{P}_0 \mathbf{x} + \mathbf{q}_0^T \mathbf{x} + r_0, \quad \mathbf{P}_0 \in \mathbf{S}_{++}^n$$

取步长 $\hat{\alpha} = 1$

$$\begin{aligned} \mathbf{x}^1 &= \mathbf{x}^0 - \hat{\alpha} \left(\nabla^2 f_o(\mathbf{x}^0) \right)^{-1} \nabla f_o(\mathbf{x}^0) \\ &= \mathbf{x}^0 - \hat{\alpha} \mathbf{P}_0^{-1} (\mathbf{P}_0 \mathbf{x}^0 + \mathbf{q}_0) \\ &= -\mathbf{P}_0^{-1} \mathbf{q}_0 \end{aligned}$$

即一步到达最优解

本专题所讲的牛顿法（Newton Method）与牛顿拉普森算法（Newton Raphson Algorithm）的联系

所有的无约束优化问题本质上都是求解这样一个非线性方程组： $\nabla f_o(\mathbf{x}) = \mathbf{0}$

用牛顿法求解非线性方程组 $\nabla f_o(\mathbf{x}) = \mathbf{0}$ 和用牛顿拉普森算法求解非线性方程组 $\mathbf{h}(\mathbf{x}) = \mathbf{0}$ 可以对应起来

牛顿法的缺陷（no-free-lunch principle）：求解牛顿方向十分繁杂！

$\nabla^2 f_o(\mathbf{x}^k) \mathbf{d}^k = -\nabla f_o(\mathbf{x}^k)$ ——涉及二阶梯度及求逆运算

拟牛顿法（Quasi-Newton Method）

$\mathbf{B}^k \mathbf{d}^k = -\nabla f_o(\mathbf{x}^k)$

BFGS 算法

- ①用两个一阶梯度之差，逼近二阶梯度
- ②避免求逆

有线性等式约束的凸优化问题

$$\begin{cases} \min & f_o(\mathbf{x}) \\ \text{s.t.} & \mathbf{Ax} = \mathbf{b} \end{cases}$$

KKT 条件

$$\begin{cases} \mathbf{Ax}^* = \mathbf{b} \\ \nabla f_o(\mathbf{x}^*) + \mathbf{A}^T \mathbf{v}^* = \mathbf{0} \end{cases}$$

如果 $\nabla f_o(\mathbf{x}^*)$ 这一项也为线性（即 f_o 为二次函数），则上述 KKT 条件转化为线性方程组

即

$$\begin{cases} \min & \frac{1}{2} \mathbf{x}^T \mathbf{Px} + \mathbf{q}^T \mathbf{x} + r \\ \text{s.t.} & \mathbf{Ax} = \mathbf{b} \end{cases} \text{ with } \mathbf{P} \in \mathbf{S}_{++}^n \text{ (QP 问题)}$$

$$\Rightarrow \begin{cases} \mathbf{Ax}^* = \mathbf{b} \\ \mathbf{Px}^* + \mathbf{q} + \mathbf{A}^T \mathbf{v}^* = \mathbf{0} \end{cases} \text{ (KKT 条件)}$$

$$\Rightarrow \begin{pmatrix} \mathbf{P} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x}^* \\ \mathbf{v}^* \end{pmatrix} = \begin{pmatrix} -\mathbf{q} \\ \mathbf{b} \end{pmatrix}$$

求解线性方程组的算法：**Jacobi 算法**, **Gauss-Seidel 算法**

约束满足的牛顿法

若 f_o 并非二次函数 \Rightarrow 线性化

$$\begin{cases} \min & f_o(\mathbf{x}^k + \mathbf{d}) \\ \text{s.t.} & \mathbf{A}(\mathbf{x}^k + \mathbf{d}) = \mathbf{b} \end{cases}$$

$$\mathbf{d}^k = \arg \min_{\mathbf{d}} \begin{cases} \min & \nabla^T f_o(\mathbf{x}^k) \mathbf{d} + \frac{1}{2} \mathbf{d}^T \nabla^2 f_o(\mathbf{x}^k) \mathbf{d} \\ \text{s.t.} & \mathbf{Ax}^k + \mathbf{Ad} = \mathbf{b} \end{cases}$$

$$\Rightarrow \begin{pmatrix} \nabla^2 f_o(\mathbf{x}^k) & \mathbf{A}^T \\ \mathbf{A} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{d}^k \\ \mathbf{v}^k \end{pmatrix} = \begin{pmatrix} -\nabla f_o(\mathbf{x}^k) \\ \mathbf{b} - \mathbf{Ax}^k \end{pmatrix}$$

如何保证新解 $\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha^k \mathbf{d}^k$ 亦满足约束

下面分析如果 \mathbf{x}^k 满足约束，则新解 $\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha^k \mathbf{d}^k$ 亦满足约束

$$\mathbf{Ax}^k = \mathbf{b}$$

$$\mathbf{Ax}^{k+1} = \mathbf{A}(\mathbf{x}^k + \alpha^k \mathbf{d}^k) = \mathbf{Ax}^k + \alpha^k \mathbf{Ad}^k = \mathbf{b}$$

算法步骤：

① 寻找 \mathbf{x}^0 ，满足 $\mathbf{Ax}^0 = \mathbf{b}$

② 寻找 \mathbf{d}^k

$$\begin{pmatrix} \nabla^2 f_o(\mathbf{x}^k) & \mathbf{A}^T \\ \mathbf{A} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{d}^k \\ \mathbf{v}^k \end{pmatrix} = \begin{pmatrix} -\nabla f_o(\mathbf{x}^k) \\ \mathbf{b} - \mathbf{Ax}^k \end{pmatrix}$$

③ 寻找 α^k

$$\alpha^k = \arg \min_{\alpha} f_o(\mathbf{x}^k + \alpha \mathbf{d}^k)$$

④ 迭代更新 $\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha^k \mathbf{d}^k$ ，重复②③

拉格朗日法

有线性约束的凸优化问题

$$\begin{cases} \min & f_o(\mathbf{x}) \\ \text{s.t.} & \mathbf{Ax} = \mathbf{b} \end{cases}$$

KKT 条件

$$\begin{cases} \nabla_{\mathbf{v}} L(\mathbf{x}^*, \mathbf{v}^*) = \mathbf{0} & \Rightarrow \mathbf{Ax}^* = \mathbf{b} \\ \nabla_{\mathbf{x}} L(\mathbf{x}^*, \mathbf{v}^*) = \mathbf{0} & \Rightarrow \nabla f_o(\mathbf{x}^*) + \mathbf{A}^T \mathbf{v}^* = \mathbf{0} \end{cases}$$

拉格朗日法的迭代结构

$$\begin{cases} \mathbf{x}^{k+1} = \mathbf{x}^k - \alpha^k (\nabla f_o(\mathbf{x}^k) + \mathbf{A}^T \mathbf{v}^k) \\ \mathbf{v}^{k+1} = \mathbf{v}^k + \alpha^k (\mathbf{Ax}^k - \mathbf{b}) \end{cases} \quad \text{如果无约束就蜕变成梯度下降法}$$

当前解离最优解差多少，就更新多少

一旦达到最优解， \mathbf{x}^k ， \mathbf{v}^k 就不动了

解释：

原优化问题的拉格朗日函数

$$L(\mathbf{x}, \mathbf{v}) = f_o(\mathbf{x}) + \mathbf{v}^T (\mathbf{Ax} - \mathbf{b})$$

$(\mathbf{x}^*, \mathbf{v}^*)$ 为鞍点

$$\begin{cases} \mathbf{x}^* = \arg \min_{\mathbf{x}} L(\mathbf{x}, \mathbf{v}^*) \\ \mathbf{v}^* = \arg \max_{\mathbf{v}} L(\mathbf{x}^*, \mathbf{v}) \end{cases} \quad \begin{array}{l} \text{梯度下降法: } \mathbf{x}^{k+1} = \mathbf{x}^k - \alpha^k \nabla L_{\mathbf{x}}(\mathbf{x}^k, \mathbf{v}^*) \\ \text{转化为无约束的优化问题，可以用梯度法求解} \\ \text{梯度上升法: } \mathbf{v}^{k+1} = \mathbf{v}^k + \alpha^k \nabla L_{\mathbf{v}}(\mathbf{x}^*, \mathbf{v}^k) \end{array}$$

拉格朗日法

$$\begin{cases} \mathbf{x}^{k+1} = \mathbf{x}^k - \alpha^k \nabla L_{\mathbf{x}}(\mathbf{x}^k, \mathbf{v}^k) \\ \mathbf{v}^{k+1} = \mathbf{v}^k + \alpha^k \nabla L_{\mathbf{v}}(\mathbf{x}^k, \mathbf{v}^k) \end{cases}$$

$$\begin{cases} \mathbf{x}^{k+1} = \mathbf{x}^k - \alpha^k \nabla L_{\mathbf{x}}(\mathbf{x}^k, \mathbf{v}^k) = \mathbf{x}^k - \alpha^k (\nabla f_o(\mathbf{x}^k) + \mathbf{A}^T \mathbf{v}^k) \\ \mathbf{v}^{k+1} = \mathbf{v}^k + \alpha^k \nabla L_{\mathbf{v}}(\mathbf{x}^k, \mathbf{v}^k) = \mathbf{v}^k + \alpha^k (\mathbf{Ax}^k - \mathbf{b}) \end{cases}$$

Jacobi 迭代法 不能用 Gauss-Seidel 迭代法

原因：优化问题可进一步化为

$$\begin{cases} \min_{\mathbf{x}} f_o(\mathbf{x}) + \mathbf{v}^T \mathbf{Ax} \\ \max_{\mathbf{v}} \mathbf{v}^T (\mathbf{Ax} - \mathbf{b}) \end{cases}$$

由于对偶目标函数是线性的。

当约束满足时，最优值为 0；当约束不满足时，最优值将达到 $+\infty$ 。

亦即在对偶域上， \mathbf{v} 非常敏感。

所以用 Jacobi 迭代法。

增广拉格朗日法 (Augmented Lagrangian Method)

问题 1:

原问题

$$\begin{cases} \min & f_o(\mathbf{x}) \\ \text{s.t.} & \mathbf{Ax} = \mathbf{b} \end{cases}$$

拉格朗日函数

$$L(\mathbf{x}, \mathbf{v}) = f_o(\mathbf{x}) + \mathbf{v}^T (\mathbf{Ax} - \mathbf{b})$$

问题 2:

增广问题

$$\begin{cases} \min & f_o(\mathbf{x}) + \frac{c}{2} \|\mathbf{Ax} - \mathbf{b}\|_2^2 \\ \text{s.t.} & \mathbf{Ax} = \mathbf{b} \end{cases}$$

增广拉格朗日函数

$$L_c(\mathbf{x}, \mathbf{v}) = f_o(\mathbf{x}) + \mathbf{v}^T (\mathbf{Ax} - \mathbf{b}) + \frac{c}{2} \|\mathbf{Ax} - \mathbf{b}\|_2^2$$

两个优化问题的最优解是一致的，对偶变量的最优解也是一样的。

设 $(\mathbf{x}^*, \mathbf{v}^*)$ 是问题 1 的 Primal-Dual 的最优解，则由 KKT 条件，有

$$\begin{cases} \mathbf{Ax}^* = \mathbf{b} \\ \nabla_{\mathbf{x}} L(\mathbf{x}^*, \mathbf{v}^*) = \mathbf{0} \end{cases} \Rightarrow \nabla f_o(\mathbf{x}^*) + \mathbf{A}^T \mathbf{v}^* = \mathbf{0}$$

下面验证 $(\mathbf{x}^*, \mathbf{v}^*)$ 也是问题 2 的 Primal-Dual 的最优解

$$\begin{cases} \mathbf{Ax}^* = \mathbf{b} \\ \nabla_{\mathbf{x}} L_c(\mathbf{x}^*, \mathbf{v}^*) = \mathbf{0} \end{cases} \Rightarrow \nabla f_o(\mathbf{x}^*) + \mathbf{A}^T \mathbf{v}^* + c \mathbf{A}^T (\mathbf{Ax}^* - \mathbf{b}) = \mathbf{0}$$

即 $(\mathbf{x}^*, \mathbf{v}^*)$ 满足问题 2 的 KKT 条件，因而也是问题 2 的 Primal-Dual 的最优解

增广拉格朗日法

$$\begin{cases} \mathbf{x}^{k+1} = \arg \min_{\mathbf{x}} L_c(\mathbf{x}, \mathbf{v}^k) = \arg \min_{\mathbf{x}} f_o(\mathbf{x}) + (\mathbf{v}^k)^T (\mathbf{Ax} - \mathbf{b}) + \frac{c}{2} \|\mathbf{Ax} - \mathbf{b}\|_2^2 \\ \mathbf{v}^{k+1} = \mathbf{v}^k + \underline{c^k} (\mathbf{Ax}^{k+1} - \mathbf{b}) \end{cases}$$

既是引入二次项的惩罚系数

又是对偶变量的迭代步长

计算复杂程度受制于目标函数 $f_o(\mathbf{x})$

例:

$$\begin{cases} \min & \frac{1}{2}x_1^2 + \frac{1}{2}x_2^2 \\ \text{s.t.} & x_1 = 1 \end{cases}$$

先求一下该问题的 Primal-Dual 最优解

1) 原问题的最优解 $\mathbf{x}^* = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

2) 拉格朗日函数

$$L(\mathbf{x}, \nu) = \frac{1}{2}x_1^2 + \frac{1}{2}x_2^2 + \nu(x_1 - 1)$$

$$L_x(\mathbf{x}^*, \nu^*) = x_1^* + x_2^* + \nu^* = 0$$

$$\text{对偶问题最优解 } \nu^* = -1$$

下面用增广拉格朗日法求解这个问题

$$\text{增广拉格朗日函数: } L_{c^k}(\mathbf{x}, \nu) = \frac{1}{2}x_1^2 + \frac{1}{2}x_2^2 + \nu(x_1 - 1) + \frac{c^k}{2}(x_1 - 1)^2$$

增广拉格朗日法迭代式

$$\begin{cases} x_1^{k+1} = \arg \min_{x_1} \frac{1}{2}x_1^2 + \nu^k x_1 + \frac{c^k}{2}(x_1 - 1)^2 \stackrel{x_1 + \nu^k + c^k(x_1 - 1) = 0}{=} \frac{c^k - \nu^k}{c^k + 1} \\ x_2^{k+1} = \arg \min_{x_2} \frac{1}{2}x_2^2 = 0 \\ \nu^{k+1} = \nu^k + c^k(x_1^{k+1} - 1) = \nu^k + c^k \left(\frac{c^k - \nu^k}{c^k + 1} - 1 \right) \end{cases}$$

一旦达到最优解, \mathbf{x}^k , ν^k 就不动了, 下面验证迭代式的收敛性

$$\begin{aligned} \nu^{k+1} - \nu^* &= \nu^k - \nu^* + c^k \left(\frac{c^k - \nu^k}{c^k + 1} - 1 \right) \\ &= \left(1 - \frac{c^k}{c^k + 1} \right) \nu^k - \nu^* + \frac{(c^k)^2}{c^k + 1} - c^k \\ &= \frac{1}{c^k + 1} \nu^k + \left(1 - c^k + \frac{(c^k)^2}{c^k + 1} \right) - \nu^* \\ &= \frac{1}{c^k + 1} \nu^k + \frac{1}{c^k + 1} \\ &= \frac{\nu^k - \nu^*}{c^k + 1} \end{aligned}$$

对于非凸问题, 也可以用拉格朗日法
只要 c 取大一点

可以验证如下优化问题

$$\begin{cases} \min & -\frac{1}{2}x_1^2 + \frac{1}{2}x_2^2 \\ \text{s.t.} & x_1 = 1 \end{cases}$$

只要 $c^k > 0$, 每一步迭代都会使对偶变量当前解离对偶变量最优解越来越近

当 ν^k 趋近于 ν^* 时, \mathbf{x}^k 自然趋近于 \mathbf{x}^*

综上, 增广拉格朗日法迭代式一定会收敛到原变量与对偶变量的最优解里面

交替方向乘子法 (Alternating Direction Method of Multipliers)

优化问题

$$\begin{cases} \min & f_1(\mathbf{x}) + f_2(\mathbf{y}) \\ \text{s.t.} & \mathbf{Ax} + \mathbf{By} = \mathbf{0} \end{cases} \longrightarrow \begin{array}{l} \text{目标函数中, 优化变量 } \mathbf{x} \text{ 和 } \mathbf{y} \text{ 可分离成和的形式} \\ f_1(\mathbf{x}), f_2(\mathbf{y}) \text{ 足够简单, 极小化它们有显式解} \end{array}$$

增广拉格朗日函数

$$L_c(\mathbf{x}, \mathbf{y}, \mathbf{v}) = f_1(\mathbf{x}) + f_2(\mathbf{y}) + \mathbf{v}^T (\mathbf{Ax} + \mathbf{By}) + \frac{c}{2} \|\mathbf{Ax} + \mathbf{By}\|_2^2$$

增广拉格朗日法迭代式

$$\begin{cases} (\mathbf{x}^{k+1}, \mathbf{y}^{k+1}) = \arg \min_{\mathbf{x}, \mathbf{y}} f_1(\mathbf{x}) + f_2(\mathbf{y}) + (\mathbf{v}^k)^T (\mathbf{Ax} + \mathbf{By}) + \frac{c}{2} \|\mathbf{Ax} + \mathbf{By}\|_2^2 \\ \mathbf{v}^{k+1} = \mathbf{v}^k + c(\mathbf{Ax}^{k+1} + \mathbf{By}^{k+1}) \end{cases} \quad \downarrow$$

\mathbf{x} 和 \mathbf{y} 纠缠在一起

策略: 各个击破

交替方向乘子法

① 固定 \mathbf{y}, \mathbf{v} , 更新 \mathbf{x}

$$\mathbf{x}^{k+1} = \arg \min_{\mathbf{x}} f_1(\mathbf{x}) + (\mathbf{v}^k)^T \mathbf{Ax} + \frac{c}{2} \|\mathbf{Ax} + \mathbf{By}^k\|_2^2$$

② 固定 \mathbf{x}, \mathbf{v} , 更新 \mathbf{y}

$$\mathbf{y}^{k+1} = \arg \min_{\mathbf{y}} f_2(\mathbf{y}) + (\mathbf{v}^k)^T \mathbf{By} + \frac{c}{2} \|\mathbf{Ax}^{k+1} + \mathbf{By}\|_2^2$$

③ 更新拉格朗日乘子 \mathbf{v}

$$\mathbf{v}^{k+1} = \mathbf{v}^k + c(\mathbf{Ax}^{k+1} + \mathbf{By}^{k+1})$$

①②③不断迭代

另一种算法:

①②之间不断迭代, 稳定后执行③

坐标轮换法

增广拉格朗日法

例 1:

经典的**稀疏优化**（可用 LASSO 思想求解，略）

$$\min_{\mathbf{x}} \quad \frac{1}{2} \|\mathbf{Ax} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{x}\|_1$$

一般来说，该问题没有解析解
但当 \mathbf{A} 为单位阵时，有解析解

梯度下降法不能用
次梯度法收敛性差

⇔

$$\begin{cases} \min_{\mathbf{x}, \mathbf{y}} & \frac{1}{2} \|\mathbf{Ax} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{y}\|_1 \\ \text{s.t.} & \mathbf{x} = \mathbf{y} \end{cases}$$

增广拉格朗日函数

$$L_c(\mathbf{x}, \mathbf{y}, \mathbf{v}) = \frac{1}{2} \|\mathbf{Ax} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{y}\|_1 + \mathbf{v}^T (\mathbf{x} - \mathbf{y}) + \frac{c}{2} \|\mathbf{x} - \mathbf{y}\|_2^2$$

① \mathbf{x} 的子问题

$$\begin{aligned} \mathbf{x}^{k+1} &= \arg \min_{\mathbf{x}} \frac{1}{2} \|\mathbf{Ax} - \mathbf{b}\|_2^2 + (\mathbf{v}^k)^T \mathbf{x} + \frac{c}{2} \|\mathbf{x} - \mathbf{y}^k\|_2^2 \\ &= (\mathbf{A}^T \mathbf{A} + c\mathbf{I})^{-1} (c\mathbf{y}^k - \mathbf{v}^k + \mathbf{A}^T \mathbf{b}) \leftarrow \mathbf{A}^T (\mathbf{Ax} - \mathbf{b}) + (\mathbf{v}^k)^T \mathbf{x} + c(\mathbf{x} - \mathbf{y}^k) = 0 \end{aligned}$$

② \mathbf{y} 的子问题

$$\begin{aligned} \mathbf{y}^{k+1} &= \arg \min_{\mathbf{y}} \lambda \|\mathbf{y}\|_1 - (\mathbf{v}^k)^T \mathbf{y} + \frac{c}{2} \|\mathbf{y} - \mathbf{x}^{k+1}\|_2^2 \\ &= \arg \min_{\mathbf{y}} \lambda \|\mathbf{y}\|_1 + \frac{c}{2} \left\| \mathbf{y} - \left(\mathbf{x}^{k+1} + \frac{\mathbf{v}^k}{c} \right) \right\|_2^2 \\ &= \arg \min_{\mathbf{y}} \lambda \|\mathbf{y}\|_1 + \frac{c}{2} \|\mathbf{y} - \mathbf{z}^k\|_2^2 \end{aligned}$$

前面已经说过，

当二次项的二阶系数为单位阵时，该优化问题是有解析解的

$$y_i^{k+1} = \begin{cases} 0 & |z_i^k| \leq \frac{\lambda}{c} \\ z_i^k - \frac{\lambda}{c} \text{sgn}(z_i^k) & |z_i^k| > \frac{\lambda}{c} \end{cases}$$

③ \mathbf{v} 的子问题

$$\mathbf{v}^{k+1} = \mathbf{v}^k + c(\mathbf{x}^{k+1} - \mathbf{y}^{k+1})$$

最优解必须使 $\lambda \|\mathbf{y}\|_1 + \frac{c}{2} \|\mathbf{y} - \mathbf{z}^k\|_2^2$ 的次梯度为 0

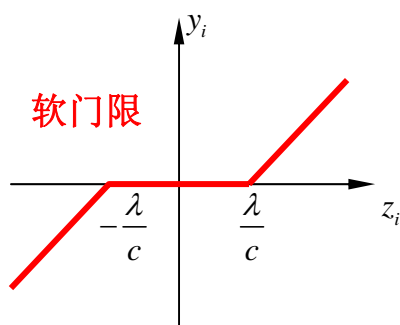
$$\text{设 } g(y_i) \in \partial |y_i| = \begin{cases} -1 & y_i < 0 \\ [-1, 1] & y_i = 0 \\ 1 & y_i > 0 \end{cases}, \text{ 于是}$$

$$\lambda g(y_i) + c(y_i - z_i^k) = 0$$

当 $y_i < 0$ 时， $y_i = z_i^k + \frac{\lambda}{c}$ ，此时 $z_i^k < -\frac{\lambda}{c}$

当 $y_i > 0$ 时， $y_i = z_i^k - \frac{\lambda}{c}$ ，此时 $z_i^k > \frac{\lambda}{c}$

当 $y_i = 0$ 时，此时 $|z_i^k| \leq \frac{\lambda}{c}$



例 2:
分布式优化

$$\min_{\mathbf{x}} \sum_{i=1}^n f_i(\mathbf{x})$$

\Leftrightarrow

$$\begin{cases} \min_{\{\mathbf{x}_i\}, \mathbf{y}} \sum_{i=1}^n f_i(\mathbf{x}_i) \\ \text{s.t.} \quad \mathbf{x}_i = \mathbf{y}, \quad \forall i \end{cases}$$

所有局部拷贝 \mathbf{x}_i 都是 \mathbf{y} (即所有局部拷贝都是一样) 的时候

此时的 \mathbf{x} (或 \mathbf{y}) 即为最优解

极小化本地目标函数 $f_i(\mathbf{x})$
得到拷贝 \mathbf{x}_i

增广拉格朗日函数

$$L_c(\{\mathbf{x}_i\}, \mathbf{y}, \{\mathbf{v}_i\}) = \sum_{i=1}^n f_i(\mathbf{x}_i) + \sum_{i=1}^n \mathbf{v}_i^T (\mathbf{x}_i - \mathbf{y}) + \frac{c}{2} \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{y}\|_2^2$$

第 i 个结点负责两个变量 \mathbf{x}_i 和 \mathbf{v}_i 的更新

中心结点负责 \mathbf{y} 的更新

因为该函数关于 \mathbf{x}_i 是可分的, 可用交替方向乘子法

① 关于 \mathbf{x}_i 的更新

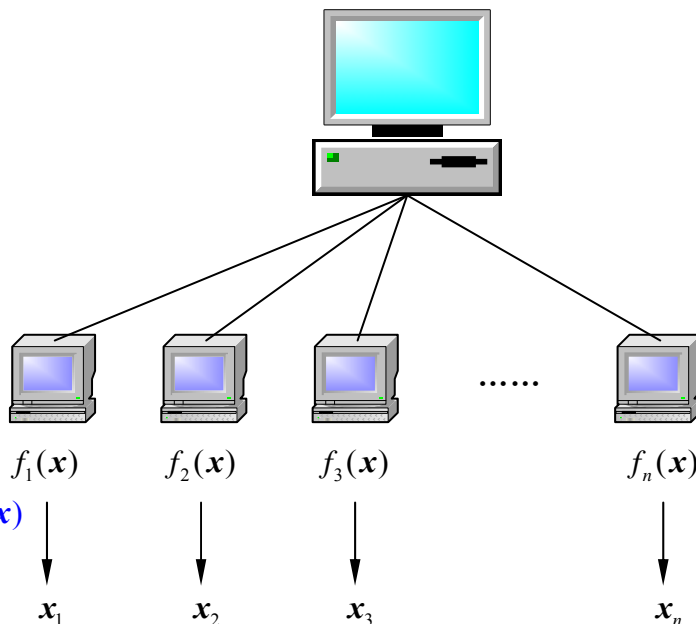
$$\begin{aligned} \mathbf{x}_i^{k+1} &= \arg \min_{\mathbf{x}_i} \left\{ f_i(\mathbf{x}_i) + (\mathbf{v}_i^k)^T \mathbf{x}_i + \frac{c}{2} \|\mathbf{x}_i - \mathbf{y}^k\|_2^2 \right\} \\ &= \arg \min_{\mathbf{x}_i} \left\{ f_i(\mathbf{x}_i) + \frac{c}{2} \left\| \mathbf{x}_i - \left(\mathbf{y}^k + \frac{\mathbf{v}_i^k}{c} \right) \right\|_2^2 \right\}, \quad \forall i \end{aligned}$$

② 关于 \mathbf{y} 的更新

$$\begin{aligned} \mathbf{y}^{k+1} &= \arg \min_{\mathbf{y}} \left\{ -\sum_{i=1}^n (\mathbf{v}_i^k)^T \mathbf{y} + \frac{c}{2} \sum_{i=1}^n \|\mathbf{x}_i^{k+1} - \mathbf{y}\|_2^2 \right\} \\ &= \frac{1}{cn} \sum_{i=1}^n (c\mathbf{x}_i^{k+1} + \mathbf{v}_i^k), \quad \forall i \end{aligned}$$

③ 关于 \mathbf{v}_i 的更新

$$\mathbf{v}_i^{k+1} = \mathbf{v}_i^k + c(\mathbf{x}_i^{k+1} - \mathbf{y}^{k+1})$$



有线性不等式约束的凸优化问题

$$\begin{cases} \min & f_o(\mathbf{x}) \\ \text{s.t.} & \mathbf{Ax} \leq \mathbf{b} \end{cases}$$

拉格朗日函数

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f_o(\mathbf{x}) + \boldsymbol{\lambda}^T (\mathbf{Ax} - \mathbf{b}) \quad \boldsymbol{\lambda} \succeq \mathbf{0}$$

方法一：

增广拉格朗日函数

$$L_c(\mathbf{x}, \boldsymbol{\lambda}) = f_o(\mathbf{x}) + \boldsymbol{\lambda}^T (\mathbf{Ax} - \mathbf{b}) + \frac{c}{2} \|\mathbf{Ax} - \mathbf{b}\|_2^2 \quad \boldsymbol{\lambda} \succeq \mathbf{0}$$

增广拉格朗日法

$$\begin{cases} \mathbf{x}^{k+1} = \arg \min_{\mathbf{x}} L_c(\mathbf{x}, \boldsymbol{\lambda}^k) = \arg \min_{\mathbf{x}} f_o(\mathbf{x}) + (\boldsymbol{\lambda}^k)^T (\mathbf{Ax} - \mathbf{b}) + \frac{c}{2} \|\mathbf{Ax} - \mathbf{b}\|_2^2 \\ \boldsymbol{\lambda}^{k+1} = \left[\boldsymbol{\lambda}^k + c^k (\mathbf{Ax}^{k+1} - \mathbf{b}) \right]^+ \end{cases} \quad \text{即对对偶变量做正投影}$$

方法二：

引入松弛变量

$$\begin{cases} \min_{\mathbf{x}, \mathbf{s}} & f_o(\mathbf{x}) \\ \text{s.t.} & \mathbf{Ax} - \mathbf{b} + \mathbf{s} = \mathbf{0} \\ & \mathbf{s} \succeq \mathbf{0} \end{cases}$$

增广拉格朗日函数

$$L_c(\mathbf{x}, \mathbf{v}) = f_o(\mathbf{x}) + \mathbf{v}^T (\mathbf{Ax} - \mathbf{b} + \mathbf{s}) + \frac{c}{2} \|\mathbf{Ax} - \mathbf{b} + \mathbf{s}\|_2^2 \quad \{\mathbf{s} / \mathbf{s} \succeq \mathbf{0}\}$$