

最优化理论

陈鸿峥

2019.06*

目录

1	简介	2
1.1	优化概述	2
1.2	分类	3
1.3	历史	3
2	凸集	4
2.1	基本概念	4
2.2	保凸运算	7
3	凸函数	8
3.1	基本概念与性质	8
3.2	常见例子	10
3.3	保凸运算	12
3.4	共轭函数	14
3.5	拟凸函数	15
4	凸优化问题	17
4.1	标准型	17
4.2	线性规划	19
4.3	二次规划	21
4.4	广义不等式约束	23
4.5	多目标优化问题	24
5	对偶理论	25
5.1	拉格朗日对偶	25
5.2	强弱对偶	28
5.3	对偶问题的几种解释	30
5.4	一般优化问题的对偶理论	33

*Build 20190617

6	优化算法	36
6.1	简介	36
6.2	梯度下降法	37
6.3	非光滑优化问题	42
6.4	二阶优化方法	49
6.5	有约束优化方法	51
6.6	对偶次梯度法	53
7	大数据中的优化问题与算法	57
7.1	并行优化	57
7.2	无中心分布式优化	59
7.3	有限和优化	61
7.4	方差消减	61
7.5	深度神经网络	62
7.6	在线优化	62
7.7	动态优化	63
7.8	Nesterov加速	63
A	线性代数基础	64
A.1	内积与范数	64
A.2	二次型	65
A.3	特征值分解	66
B	矩阵微积分	66
B.1	基本定义	66
B.2	实值函数对向量的导数	66
B.3	向量值函数对向量的导数	68
C	参考资料	68

1 简介

1.1 优化概述

优化(optimization): 从一个可行解的集合中寻找出最好的元素

- 最小二乘法 (凸问题)

$$\min \|A\mathbf{x} - \mathbf{b}\|_2^2$$

- 深度神经网络（非凸，见下）

$$\mathbf{x}_1^{(i)} = f_1(\mathbf{x}_0^{(i)}, \mathbf{w}_1)$$

... ..

$$\mathbf{x}_n^{(i)} = f_n(\mathbf{x}_{n-1}^{(i)}, \mathbf{w}_n)$$

$$\min \sum_{i=1}^m (\mathbf{y}^{(i)} - \mathbf{x}_n^{(i)})^2$$

- 图像处理，自然图像通常都是分块光滑的，原图 Φ_0 ，有噪声的新图 Φ
全变参(TV, Total Variation)范数，计算图像每个像素点左侧和下侧的差异

$$\|\Phi\|_{TV} = \sum_y \sum_x \sqrt{(\Phi(x, y) - \Phi(x, y-1))^2 + (\Phi(x, y) - \Phi(x-1, y))^2}$$

可得优化目标：近似自然图像，而且跟原图不能差太远

$$\min(\|\Phi\|_{TV} + \lambda \|\Phi - \Phi_0\|_F^2)$$

- 推荐系统：Netflix问题→低秩矩阵补全
矩阵横向为用户，纵向为电影，值为评分值(1 ~ 5)，问题是把矩阵补全，这样就可以做推荐了
电影很多，但类型不多，关联关系有限→近似低秩¹
低秩本来需要最小化 \mathbf{z} 的非零奇异值数目 $\|\mathbf{z}\|_0$ ，但是非凸的；转化为最小化和范数² $\|\mathbf{z}\|_*$

$$\min \quad \|\mathbf{z}\|_* := \|\mathbf{z}\|_1$$

$$\text{s.t.} \quad \mathbf{z}_{ij} = M_{ij}, (i, j) \in \Omega$$

1.2 分类

- 线性规划/非线性规划
- 凸规划/非凸规划（更好的分类）

目标函数凸函数，可行解集为凸集则是凸优化，一般容易求解

1.3 历史

- Newton-Raphson算法：求零点，等价于求 $\min f^2(x)$
- Gauss-Seidel算法：求解线性方程组 $A\mathbf{x} = \mathbf{b}$ ，等价于求 $\min \|A\mathbf{x} - \mathbf{b}\|_2^2$
- Lagrange：法国
- Kantoronc：苏联，线性规划，诺贝尔经济学奖
- Dantzig：美国，优化决策，线性规划单纯形

¹ A 的秩等于非零奇异值 $\sqrt{\text{eig}(A^T A)}$ 数目

² 矩阵所有奇异值之和

- Von Neumann: 线性规划问题对偶理论
- Karmarkar: 80年代, 线性规划内点法
- Nesterov: 后80年代, 非线性凸优化内点法
- 现代: 并行、随机算法

2 凸集

2.1 基本概念

定义 1. 一些集合概念如下

- 仿射集 (*affine set*)

C 为仿射集 \iff 过 C 内任意两点的直线都在 C 内

$$\iff \forall x_1, x_2 \in C, \theta \in \mathbb{R}, \theta x_1 + (1 - \theta)x_2 \in C$$

例 1. 用定义易证线性方程组的解集 $C = \{\mathbf{x} \mid A\mathbf{x} = \mathbf{b}\}$ 是仿射集; 反过来, 每一个仿射集都可以用线性方程组的解集表示

- 仿射组合 (多点扩展)

$$\forall x_1, x_2, \dots, x_k \in C, \theta_1, \dots, \theta_k \in \mathbb{R}, \theta_1 + \dots + \theta_k = 1: \theta_1 x_1 + \dots + \theta_k x_k \in C$$

- 仿射包 (*hull*): 所有仿射组合的集合

$$\text{aff } C := \{\theta_1 x_1 + \dots + \theta_k x_k \mid \forall x_1, \dots, x_k \in C, \theta_1 + \dots + \theta_k = 1\}$$

- 凸集 (*convex set*)

C 为凸集 \iff 过 C 内任意两点的线段都在 C 内

$$\iff \forall x_1, x_2 \in C, \theta \in [0, 1], \theta x_1 + (1 - \theta)x_2 \in C$$

- 凸组合

$$\forall x_1, x_2, \dots, x_k \in C, \theta_1, \dots, \theta_k \in [0, 1], \theta_1 + \dots + \theta_k = 1: \theta_1 x_1 + \dots + \theta_k x_k \in C$$

分析. 由两点推多点可用数学归纳法

$$(1 - \theta_{k+1}) \left(\frac{\theta_1}{1 - \theta_{k+1}} x_1 + \dots + \frac{\theta_k}{1 - \theta_{k+1}} x_k \right) + \theta_{k+1} x_{k+1} \in C$$

- 凸包：最小的凸集

$$\text{conv } \mathcal{C} := \{\theta_1 x_1 + \cdots + \theta_k x_k \mid \forall x_1, \dots, x_k \in \mathcal{C}, \theta_1, \dots, \theta_k \in [0, 1], \theta_1 + \cdots + \theta_k = 1\}$$

- 凸锥 (convex cone)

$$\mathcal{C} \text{ 为凸锥} \iff \forall x_1, x_2 \in \mathcal{C}, \theta_1, \theta_2 \geq 0, \theta_1 x_1 + \theta_2 x_2 \in \mathcal{C}$$

相当于 x_1 和 x_2 所夹的**正向区域**都在 \mathcal{C} 内，除了空集的凸锥都得包含**原点**（取 $\theta_1 = \theta_2 = 0$ ）

- 凸锥组合/非负线性组合：

$$\forall x_1, x_2, \dots, x_k \in \mathcal{C}, \theta_1, \dots, \theta_k \geq 0 : \theta_1 x_1 + \cdots + \theta_k x_k \in \mathcal{C}$$

- 凸锥包：类似前面定义

由上面的定义易知，仿射组合/凸锥组合（强条件）一定是凸组合。

定义 2 (超平面(hyperplane)与半空间(halfspace)). 超平面都是比原空间低一维

$$\{\mathbf{x} \mid \mathbf{a}^T \mathbf{x} = b, \mathbf{x}, \mathbf{a} \in \mathbb{R}^n, b \in \mathbb{R}, \mathbf{a} \neq 0\}$$

超平面将空间划分为两个部分，即半空间

$$\{\mathbf{x} \mid \mathbf{a}^T \mathbf{x} \leq b, \mathbf{a} \neq 0\}$$

若方程特解为 \mathbf{x}_0 ，则 $\mathbf{a} \perp (\mathbf{x} - \mathbf{x}_0)$

注：想一下二维的情况， $a_1 x_1 + a_2 x_2 = b$ 就是一条直线

Voronoi描述： $\{x \mid \|x - a\|_2 \leq \|x - b\|_2\}$ 为半空间，展开实际上为中位线描述

定义 3 (欧式球(Euclidean ball)).

$$B(\mathbf{x}_c, r) = \{\mathbf{x} \mid \|\mathbf{x} - \mathbf{x}_c\|_2 \leq r\}$$

范数(norm)球可类似定义

定义 4 (椭球(ellipsoid)).

$$\varepsilon(\mathbf{x}_c, P) = \{\mathbf{x} \mid (\mathbf{x} - \mathbf{x}_c)^T P^{-1} (\mathbf{x} - \mathbf{x}_c) \leq 1\}, P \succ 0$$

其中 $P \succ 0$ 表示 P 对称 ($P = P^T$) 且正定，或记为 $P \in \mathbb{S}_{++}$

分析. 定义内积 $\langle x^T, P^{-1} y \rangle$ (需证满足内积条件)，进而 $\|\mathbf{x}\|_p := \sqrt{\mathbf{x}^T P^{-1} \mathbf{x}}$ 是范数，而椭球不过是 p -范数意义下的球，由定理得椭球是凸的

定义 5 (多面体(polyhedron)). 由超平面和半空间组成的集合

$$P = \{\mathbf{x} \mid \mathbf{a}_i^T \mathbf{x} \leq b_i, \mathbf{c}_j^T \mathbf{x} = d_j, i = 1, \dots, m, j = 1, \dots, p\}$$

或用紧凑表示法³

$$P = \{\mathbf{x} \mid A\mathbf{x} \leq \mathbf{b}, C\mathbf{x} = \mathbf{d}\}$$

例 2. 常见几何图形凸性分类如下, 考虑直线、正向区域和线段很快可得出结论

	仿射	凸锥	凸集
空集	✓	✗	✓
点	✓	✗	✓
直线	✓	(过原点) ✓	✓
\mathbb{R}^n 空间	✓	✗	✓
\mathbb{R}^n 空间的子空间 ⁴	✓	✓	✓
超平面	✓	✗	✓
半空间	✗	✗	✓
欧式球	✗	✗	✓
椭球	✗	✗	✓
多面体	✗	✗	✓

定理 1. 一个集合是凸集当且仅当它与任意直线的交是凸的

分析. 充分性: 直线为凸, 凸集交凸集为凸

必要性: 取过 x_1, x_2 的直线, 集合与该直线交为凸, 则 x_1, x_2 的凸组合都在集合内, 由 x_1, x_2 的任意性, 该集合是凸的

例 3. 若 $A \succeq 0$, 则二次不等式的解集

$$C = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{x}^T A \mathbf{x} + \mathbf{b}^T \mathbf{x} + c \leq 0, A \in \mathbb{S}^n, \mathbf{b} \in \mathbb{R}^n, c \in \mathbb{R}\}$$

为凸集

分析. 考虑直线⁵ $\{\hat{\mathbf{x}} + t\mathbf{v} \mid t \in \mathbb{R}\}$, 有

$$(\hat{\mathbf{x}} + t\mathbf{v})^T A (\hat{\mathbf{x}} + t\mathbf{v}) + \mathbf{b}^T (\hat{\mathbf{x}} + t\mathbf{v}) + c = \alpha t^2 + \beta t + \gamma$$

交集为 $\{\hat{\mathbf{x}} + t\mathbf{v} \mid \alpha t^2 + \beta t + \gamma \leq 0\}$

由 $A \succeq 0 \implies \alpha = \mathbf{v}^T A \mathbf{v} \geq 0$, 即与任意直线交为凸, 故 C 为凸集

³这里 $\mathbf{u} \succeq \mathbf{v}$ 为向量/分量不等式, 即 $\forall i: u_i \geq v_i$

⁴零元、加法封闭、数乘封闭

⁵此为直线的参数方程

2.2 保凸运算

- 凸集的交：如超平面与半空间的交为多面体
- 仿射、逆仿射：如伸缩平移

定义 6 (仿射函数).

$$f: \mathbb{R}^n \mapsto \mathbb{R}^m \quad f(\mathbf{x}) = A\mathbf{x} + \mathbf{b}, A \in \mathbb{R}^{m \times n}, \mathbf{b} \in \mathbb{R}^m$$

性质如下:

- $S \subset \mathbb{R}^n$ 为凸 $\implies f(S) = \{f(\mathbf{x}) \mid \mathbf{x} \in S\}$ 为凸
- $C \subset \mathbb{R}^m$ 为凸 $\implies f^{-1}(C) = \{\mathbf{x} \in \mathbb{R}^n \mid f(\mathbf{x}) \in C\}$ 为凸

例 4. 两个集合的和 $S_1 + S_2 = \{\mathbf{x} + \mathbf{y} \mid \mathbf{x} \in S_1, \mathbf{y} \in S_2\}$ 保凸

分析. 直积 $S_1 \times S_2 = \{(\mathbf{x}, \mathbf{y}) \mid \mathbf{x} \in S_1, \mathbf{y} \in S_2\}$ 显然可以保凸 (相当于在两个集合同时画线)

令 $A \leftarrow \begin{bmatrix} I & I \end{bmatrix}, \mathbf{x} \leftarrow \begin{bmatrix} \mathbf{x} & \mathbf{y} \end{bmatrix}^T, \mathbf{b} \leftarrow \mathbf{0}$, 由仿射函数性质得证

例 5. 两个集合的部分和 $S = \{(x, y_1 + y_2) \mid (x, y_1) \in S_1, (x, y_2) \in S_2, x \in \mathbb{R}^n, y_i \in \mathbb{R}^m\}$ 保凸

- 透视函数: 对向量进行规范化, 使得最后一维分量为1并舍弃

定义 7 (透视(perspective)函数⁶). 透视函数 $P: \mathbb{R}^{n+1} \mapsto \mathbb{R}^n, \text{dom } P = \mathbb{R}^n \times \mathbb{R}_{++}$ 定义如下

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

反透视函数

$$P^{-1}(\mathcal{C}) := \left\{ (\mathbf{x}, t) \in \mathbb{R}^{n+1} \mid \frac{\mathbf{x}}{t} \in \mathcal{C}, t > 0 \right\}$$

凸集经过透视函数和反透视函数依然是凸集

分析. 考虑 \mathbb{R}^{n+1} 内的线段, $\mathbf{x} = (\tilde{\mathbf{x}} \in \mathbb{R}^n, \mathbf{x}_{n+1} \in \mathbb{R}_{++}), \mathbf{y} = (\tilde{\mathbf{y}}, \mathbf{y}_{n+1})$ 则经过透视函数仍是线段

$$\begin{aligned} P(\theta\mathbf{x} + (1-\theta)\mathbf{y}) &= \frac{\theta\tilde{\mathbf{x}} + (1-\theta)\tilde{\mathbf{y}}}{\theta\mathbf{x}_{n+1} + (1-\theta)\mathbf{y}_{n+1}} \\ &= \frac{\theta\mathbf{x}_{n+1}}{\theta\mathbf{x}_{n+1} + (1-\theta)\mathbf{y}_{n+1}} \frac{\tilde{\mathbf{x}}}{\mathbf{x}_{n+1}} + \frac{(1-\theta)\mathbf{y}_{n+1}}{\theta\mathbf{x}_{n+1} + (1-\theta)\mathbf{y}_{n+1}} \frac{\tilde{\mathbf{y}}}{\mathbf{y}_{n+1}} \\ &= \mu P(\mathbf{x}) + (1-\mu)P(\mathbf{y}) \end{aligned}$$

- 线性分数函数

定义 8 (线性分数函数). 仿射函数

$$g(\mathbf{x}) = \begin{bmatrix} A \\ \mathbf{c}^T \end{bmatrix} \mathbf{x} + \begin{bmatrix} \mathbf{b} \\ d \end{bmatrix}, A \in \mathbb{R}^{m \times n}, \mathbf{b} \in \mathbb{R}^m, \mathbf{c} \in \mathbb{R}^n, d \in \mathbb{R}$$

⁶ +代表 ≥ 0 , ++代表 > 0

线性分数函数 $f: \mathbb{R}^n \mapsto \mathbb{R}^m = p \circ g$

$$f(\mathbf{x}) = \frac{A\mathbf{x} + \mathbf{b}}{\mathbf{c}^T \mathbf{x} + d}, \text{dom } f = \{\mathbf{x} \mid \mathbf{c}^T \mathbf{x} + d > 0\}$$

3 凸函数

3.1 基本概念与性质

定义 9 (凸函数). 凸函数的几种基本定义如下

1. 原始定义: $f: \mathbb{R}^n \rightarrow \mathbb{R}$ 为凸 $\iff \text{dom } f$ 为凸, 且

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

- 严格凸: $\theta \in (0, 1)$, 不等式不能取等
- 凹函数: 若 $-f$ 为凸

2. 高维定义: $f: \mathbb{R}^n \mapsto \mathbb{R}$ 为凸 $\iff \text{dom } f$ 为凸, 且

$$\forall \mathbf{x} \in \text{dom } f, \mathbf{v} \in \mathbb{R}^n: g(t) := f(\mathbf{x} + t\mathbf{v}) \text{ 为凸, } \text{dom } g = \{t \mid \mathbf{x} + t\mathbf{v} \in \text{dom } f\}$$

相当于每一个剖面上的低维函数都是凸的 (固定 \mathbf{x} 和 \mathbf{v} 变化 t 看凸性, 则 $\mathbf{x} + t\mathbf{v}$ 在一条直线上/超平面上移动)

3. 一阶条件 (first-order condition)⁷: $f: \mathbb{R}^n \mapsto \mathbb{R}$ 为凸 $\iff \text{dom } f$ 为凸, 且

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

即 Taylor 公式一阶展开

分析. 用高维定义证明, 先证一维情况, 然后设 $g(t) = f(t\mathbf{y} + (1 - t)\mathbf{x})$, 并对 t 求导. 由于 $g(t)$ 是凸的, 故用一维情况得证

4. 二阶条件: $f: \mathbb{R}^n \mapsto \mathbb{R}$ 为凸 $\iff \text{dom } f$ 为凸, 且

$$\forall \mathbf{x} \in \text{dom } f: \nabla^2 f(\mathbf{x}) \succeq 0$$

- 凹函数: $\nabla^2 f(\mathbf{x}) \preceq 0$
- 严格凸: $\iff \nabla^2 f(\mathbf{x}) \succ 0$, 反例 $f(x) = x^4$ (在一个点斜率不变并不要紧)

分析. 同样先证明一维情况, 然后设 $g(t) = f(\mathbf{x} + t\mathbf{v})$, 并对 t 求二阶导, 类似可得证

例 6. $f(\mathbf{x}) = \mathbf{a}^T \mathbf{x} + \mathbf{b}$

⁷ $\nabla^T f(\mathbf{x}) = [\nabla f(\mathbf{x})]^T$, $\nabla^2 f(\mathbf{x})$ 为 $f(\mathbf{x})$ 的 Hessian 矩阵, 其他关于矩阵微积分的知识请见附录 B 节

分析. 有 $\nabla f(\mathbf{x}) = \mathbf{a}$, 进而

$$f(\mathbf{y}) = \mathbf{a}^T \mathbf{y} + \mathbf{b} \geq \mathbf{a}^T \mathbf{x} + \mathbf{b} + \mathbf{a}^T (\mathbf{y} - \mathbf{x}) = \mathbf{a}^T \mathbf{y} + \mathbf{b}$$

例 7. 假设 $f: \mathbb{R} \mapsto \mathbb{R}$ 为凸函数, $a < b$

$$1. \forall x \in [a, b]: f(x) \leq \frac{b-x}{b-a} f(a) + \frac{x-a}{b-a} f(b)$$

由 $f(\theta a + (1-\theta)b) \leq \theta f(a) + (1-\theta)f(b)$, 令

$$\theta a + (1-\theta)b = x \implies \theta = \frac{x-b}{a-b}$$

代入得证

$$2. \forall x \in [a, b]: \frac{f(x) - f(a)}{x-a} \leq \frac{f(b) - f(a)}{b-a} \leq \frac{f(b) - f(x)}{b-x}$$

展开等价于 (1) 式

3. 假设 f 可微,

$$f'(a) \leq \frac{f(b) - f(a)}{b-a} \leq f'(b)$$

由 (2) 式取极限可得

4. 假设 f 二次可微, 有 $f''(a) \geq 0$ 以及 $f''(b) \geq 0$

令 $b = x$, 由 (3) 式, $f'(x) - f'(a) \geq 0 \implies f''(a) \geq 0$

定义 10 (凸函数的扩展(extended-value)). 尽管凸函数的定义域为凸, 但往往不好处理, 那就将其扩展到全空间. $\mathbf{x} \in \text{dom } f \subset \mathbb{R}^n, \text{dom } \tilde{f} = \mathbb{R}^n$, 会有

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

例 8 (示性(indicator)函数). 扩展值延伸的示性函数是凸的

$$\tilde{I}_{\mathcal{C}}(\mathbf{x}) = \begin{cases} 0 & \mathbf{x} \in \mathcal{C} \\ +\infty & \mathbf{x} \notin \mathcal{C} \end{cases}$$

定理 2. 若 f 为凸, 可微, 则 $\exists \mathbf{x} \in \text{dom } f, \nabla f(\mathbf{x}) = 0$

例 9. 二次函数 $f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T P \mathbf{x} + \mathbf{q}^T \mathbf{x} + r$, $P \in \mathbb{S}^n$ (对称矩阵), $\mathbf{q} \in \mathbb{R}^n$, $r \in \mathbb{R}$

分析. $\nabla f(\mathbf{x}) = \frac{1}{2}(P + P^T)\mathbf{x} + \mathbf{q} \implies \nabla^2 f(\mathbf{x}) = P$

故 $P \in \mathbb{S}_+^n$, $f(\mathbf{x})$ 为凸; $P \in \mathbb{S}_{++}^n$, $f(\mathbf{x})$ 严格凸

例 10. $f(x) = \frac{1}{x^2}$, $\text{dom } f = \{x \in \mathbb{R}, x \neq 0\}$

分析. 注意 $\text{dom } f$ 不是凸集

3.2 常见例子

3.2.1 \mathbb{R} 上的函数

- 仿射函数 $f(x) = ax + b$
- 指数函数 $f(x) = e^{ax}$
- 幂函数 $f(x) = x^a$
- 绝对值的幂函数 $f(x) = |x|^p, x \in \mathbb{R}, p > 0$: $p \in [1, +\infty)$ 凸, $p \in (0, 1)$ 既不凸又不凹

分析.

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

- 对数函数 $f(x) = \log x$
- 熵 $f(x) = -x \log x$

3.2.2 \mathbb{R}^n 上的函数

- 任意范数 $f(\mathbf{x}) = \|\mathbf{x}\|$ (常被用来正则化!)

分析. 由三角不等式

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

- 极大值函数 $f(\mathbf{x}) = \max\{x_1, \dots, x_n\}, x \in \mathbb{R}^n$

分析. 由原式定义

$$f(\theta \mathbf{x} + (1 - \theta) \mathbf{y}) = \max_i (\theta x_i + (1 - \theta) y_i) \leq \theta \max_i x_i + (1 - \theta) \max_i y_i = \theta f(\mathbf{x}) + (1 - \theta) f(\mathbf{y})$$

- 二次线性分式函数 $f(\mathbf{x}, y) = x^2/y, (x, y) \in \mathbb{R}^2, y > 0$

分析. 由二阶条件

$$\nabla^2 f(\mathbf{x}, y) = \frac{2}{y^3} \begin{bmatrix} y^2 & -xy \\ -xy & x^2 \end{bmatrix} = \frac{2}{y^3} \begin{bmatrix} y \\ -x \end{bmatrix} \begin{bmatrix} y \\ -x \end{bmatrix}^T \succeq 0$$

- 指数和的对数 $f(\mathbf{x}) = \log(e^{x_1} + \dots + e^{x_n})$

分析. 此即极大值函数的解析近似⁸, 因有下式成立

$$\max\{x_1, \dots, x_n\} \leq f(\mathbf{x}) \leq \max\{x_1, \dots, x_n\} + \log n$$

⁸即无穷阶可微

利用二阶条件证明凸性

$$\begin{aligned}\frac{\partial f}{\partial x_i} &= \frac{e^{x_i}}{e^{x_1} + \dots + e^{x_n}} \\ \frac{\partial^2 f}{\partial x_i \partial x_j} &= \begin{cases} \frac{-e^{x_i} e^{x_j}}{(e^{x_1} + \dots + e^{x_n})^2} & i \neq j \\ \frac{e^{x_i} (e^{x_1} + \dots + e^{x_{i-1}} + e^{x_{i+1}} + \dots + e^{x_n})}{(e^{x_1} + \dots + e^{x_n})^2} & i = j \end{cases} \\ \mathbf{z} &:= \begin{bmatrix} e^{x_1} & \dots & e^{x_n} \end{bmatrix}^T\end{aligned}$$

求 Hessian 矩阵

$$\nabla^2 f(\mathbf{x}) = \frac{1}{(\mathbf{1}^T \mathbf{z})^2} ((\mathbf{1}^T \mathbf{z}) \text{diag}(\mathbf{z}) - \mathbf{z} \mathbf{z}^T)$$

将前面常量丢弃

$$\begin{aligned}\mathbf{v}^T \nabla^2 f(\mathbf{x}) \mathbf{v} &= (\mathbf{1}^T \mathbf{z}) \mathbf{v}^T \text{diag}(\mathbf{z}) \mathbf{v} - \mathbf{v}^T \mathbf{z} \mathbf{z}^T \mathbf{v} \\ &= \left(\sum_i z_i \right) \left(\sum_i v_i^2 z_i \right) - \left(\sum_i v_i z_i \right)^2 \\ &\quad \text{令 } \mathbf{a}_i := \mathbf{v}_i \sqrt{z_i} = \begin{bmatrix} a_1 & \dots & a_n \end{bmatrix}^T, b_i := \sqrt{z_i} \\ &= (\mathbf{b}^T \mathbf{b})(\mathbf{a}^T \mathbf{a}) - (\mathbf{a}^T \mathbf{b})^2 \quad \text{Cauchy} \\ &\geq 0\end{aligned}$$

进而 $\nabla^2 f(\mathbf{x})$ 半正定, 即 $\nabla^2 f(\mathbf{x}) \succeq 0$

- 几何平均 $f(\mathbf{x}) = \left(\prod_{i=1}^n x_i \right)^{1/n}$ 为凹函数

分析.

$$\frac{\partial^2 f}{\partial x_k \partial x_l} = \begin{cases} -(n-1) \frac{(\prod_{i=1}^n x_i)^{1/n}}{n^2 x_k^2} & k = l \\ \frac{(\prod_{i=1}^n x_i)^{1/n}}{n^2 x_k x_l} & k \neq l \end{cases}$$

进而

$$\nabla^2 f(\mathbf{x}) = -\frac{(\prod_{i=1}^n x_i)^{1/n}}{n^2} \left(n \sum_{i=1}^n \frac{v_i^2}{x_i^2} - \left(\sum_{i=1}^n \frac{v_i}{x_i} \right)^2 \right) \leq 0$$

同样由 Cauchy 不等式可证得

- 对数行列式 $f(X) = \log \det(X)$, $\text{dom } f = \mathbb{S}_{++}^n$ 为凹函数

分析. 用高维定义

$$\begin{aligned}
g(t) &:= f(Z + tV) \\
&= \log \det(Z + tV) \\
&= \log \det(Z^{1/2}(I + tZ^{-1/2}VZ^{-1/2})Z^{1/2}), \quad Z^{1/2} \in \mathbb{S}_{++}^n, Z^{1/2}Z^{1/2} = Z \\
&= \log \det(Z) + \log \det(I + tZ^{-1/2}VZ^{-1/2}) \\
&= \log \det(Z) + \sum_{i=1}^n \log(1 + t\lambda_i), \quad \lambda_i \text{ 为 } Z^{-1/2}VZ^{-1/2} \text{ 的特征值}
\end{aligned}$$

因此下式成立

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

补充证明: 对对称阵进行特征值分解 $tZ^{1/2}VZ^{1/2} = tQ\Lambda Q^T$, 对角阵 Λ 即为 $QQ^T = I$, Q 为酉矩阵

$$\begin{aligned}
I + tZ^{-1/2}VZ^{-1/2} &= QQ^T + tQ\Lambda Q^T = Q(I + t\Lambda)Q^T \\
\log \det(I + tZ^{-1/2}VZ^{-1/2}) &= \log \det(Q) + \log \det(I + t\Lambda) + \log \det(Q^T)
\end{aligned}$$

3.3 保凸运算

- 非负加权和 f_1, \dots, f_m 为凸, 定义域 \mathbb{R}^n

$$f := \sum_{i=1}^m w_i f_i, w_i \geq 0$$

- 非负积分 $f(x, y)$ 对 $y \in A$ 均为凸 (A 不一定为凸), $w(y) \geq 0$

$$g(x) := \int_{y \in A} w(y) f(x, y) dy$$

- 仿射映射 $f: \mathbb{R}^n \mapsto \mathbb{R}$ 为凸, $A \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^m$, $\text{dom } g = \{\mathbf{x} \mid A\mathbf{x} + \mathbf{b} \in \text{dom } f\}$

$$g(\mathbf{x}) := f(A\mathbf{x} + \mathbf{b})$$

分析. $\text{dom } f$ 为凸, 则 $\text{dom } g$ 为凸

$$\begin{aligned}
\forall \mathbf{x}, \mathbf{y} \in \text{dom } g, \forall \theta \in [0, 1] : g(\theta\mathbf{x} + (1 - \theta)\mathbf{y}) &= f(A(\theta\mathbf{x} + (1 - \theta)\mathbf{y}) + \mathbf{b}) \\
&= f(\theta(A\mathbf{x} + \mathbf{b}) + (1 - \theta)(A\mathbf{y} + \mathbf{b})) \\
&\leq \theta f(A\mathbf{x} + \mathbf{b}) + (1 - \theta)f(A\mathbf{y} + \mathbf{b}) \\
&= \theta g(\mathbf{x}) + (1 - \theta)g(\mathbf{y})
\end{aligned}$$

其实只是在定义域上改变，而不是改变值域，因而函数凸性不会改变

- 两个函数的极大值函数/逐点(pointwise)最大函数， f_1, f_2 为凸

$$f(x) := \max\{f_1(x), f_2(x)\}, \text{dom } f = \text{dom } f_1 \cap \text{dom } f_2$$

例 11. $\mathbf{x} \in \mathbb{R}^n$, $x[i]$ 为第*i*大元素, $x[1] \geq x[2] \geq \dots \geq x[r] \geq \dots \geq x[n]$

$$f(\mathbf{x}) := \sum_{i=1}^r x[i]$$

* $r = 1$: $f(\mathbf{x}) = x[1] = \max\{x_1, \dots, x_n\}$, 每一项都是 $\mathbf{e}_i^T x_i$

* $r > 1$: $f(\mathbf{x}) = \max\{x_{i_1} + \dots + x_{i_r} \mid 1 \leq i_1 < i_2 < \dots < i_r \leq n\}$

即从 \mathbf{x} 的分量中选取 r 个分量进行求和的所有可能组合的最大值，即 $n!/(r!(n-r)!)$ 个线性函数的逐点最大，故为凸

- 任意个凸函数极大值函数为凸（分片线性函数）

$$f(x) = \max\{\mathbf{a}_1^T \mathbf{x} + \mathbf{b}_1, \dots, \mathbf{a}_m^T \mathbf{x} + \mathbf{b}_m\}$$

- 无限个凸函数， $y \in A$, $f(x, y)$ 对于 x 为凸，则 $g(x) = \sup_{y \in A} f(x, y)$ 为凸

例 12. 点 \mathbf{x} 到集合 C 的最远距离

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

位移对于范数凸性不会有影响

- 函数的组合: $h: \mathbb{R}^k \mapsto \mathbb{R}, g: \mathbb{R}^n \mapsto \mathbb{R}^k$

$$f := h \circ g: \mathbb{R}^n \mapsto \mathbb{R}$$

分析. 先考虑 $n = k = 1, \text{dom } g = \mathbb{R}^n, \text{dom } h = \mathbb{R}^k, \text{dom } f = \mathbb{R}$, h, g 二阶可微

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

$$f''(\mathbf{x}) = h''(g(\mathbf{x}))(g'(\mathbf{x}))^2 + h'(g(\mathbf{x}))g''(\mathbf{x}) > 0$$

即当满足以下任一条件时， $f(\mathbf{x})$ 为凸

- g 为凸， h 为凸且不降
- g 为凹， h 为凸且不增
- （若定义域非全空间） g 为凸， h 为凸，扩展值函数 \tilde{h} 不降
- （若定义域非全空间） g 为凹， h 为凸， \tilde{h} 不增

例 13. g 为凸, $\exp g(\mathbf{x})$ 为凸; g 为凹, $g > 0$, $\log g(\mathbf{x})$ 为凹; g 为凸, $g > 0$, $1/g(\mathbf{x})$ 为凸

例 14. $g(\mathbf{x}) = \mathbf{x}^2$, $\text{dom } g = \mathbb{R}$, $h(y) = 0$, $\text{dom } h = [1, 2]$, $f = h \circ g$, 注意 \tilde{h} 并非不降!

- 函数透视: $f: \mathbb{R}^n \mapsto \mathbb{R}$, $g: \mathbb{R}^n \times \mathbb{R}_{++} \mapsto \mathbb{R}$

$$g(\mathbf{x}, t) = tf\left(\frac{\mathbf{x}}{t}\right), \text{dom } g = \left\{(\mathbf{x}, t) \mid \frac{\mathbf{x}}{t} \in \text{dom } f\right\}$$

若 $f(\mathbf{x})$ 为凸, 则 $g(\mathbf{x}, t)$ 相对于 (\mathbf{x}, t) 联合凸

分析. 考虑点 (\mathbf{x}, t) 和 (\mathbf{y}, s) , 有

$$\begin{aligned} f\left(\frac{\theta\mathbf{x} + (1-\theta)\mathbf{y}}{\theta t + (1-\theta)s}\right) &= f\left(\frac{\theta t}{\theta t + (1-\theta)s} \frac{\mathbf{x}}{t} + \frac{(1-\theta)s}{\theta t + (1-\theta)s} \frac{\mathbf{y}}{s}\right) \\ &\leq \frac{\theta t}{\theta t + (1-\theta)s} f\left(\frac{\mathbf{x}}{t}\right) + \frac{(1-\theta)s}{\theta t + (1-\theta)s} f\left(\frac{\mathbf{y}}{s}\right) \end{aligned}$$

例 15. 考虑 \mathbb{R}_{++} 上的凸函数 $f(x) = -\log x$, 其透视函数为

$$g(x, t) = -t \log(x/t) = t \log(t/x) = t \log t - t \log x$$

在 \mathbb{R}_{++}^2 上为凸函数, 称 g 为关于 t 和 x 的相对熵。

进而可定义高维的相对熵 $\mathbf{u}, \mathbf{v} \in \mathbb{R}_{++}^n$

$$g(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^n u_i \log(u_i/v_i)$$

由于是一系列 u_i, v_i 相对熵的和, 故也是关于 (\mathbf{u}, \mathbf{v}) 的凸函数。

另一方面可以定义 $\mathbf{u}, \mathbf{v} \in \mathbb{R}_{++}^n$ 之间的 Kullback-Leibler/KL 散度

$$D_{KL} := \sum_{i=1}^n \left(u_i \log \frac{u_i}{v_i} - u_i + v_i \right)$$

由于是 (\mathbf{u}, \mathbf{v}) 相对熵和线性函数的和, 因此也是凸函数

分析. 设

$$f(\mathbf{u}) = \sum_{i=1}^n u_i \log u_i \quad (\nabla f(\mathbf{u}))_i = \log u_i + 1$$

由 $f(\mathbf{u})$ 的凸性可得

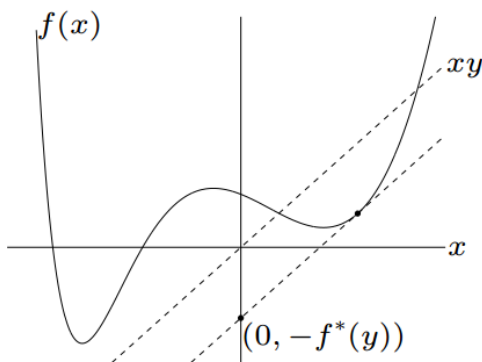
$$D_{KL}(\mathbf{u}, \mathbf{v}) = f(\mathbf{u}) - f(\mathbf{v}) - \nabla f(\mathbf{v})^T (\mathbf{u} - \mathbf{v}) \geq 0$$

3.4 共轭函数

定义 11 (函数的共轭(conjugate)). 设函数 $f: \mathbb{R}^n \mapsto \mathbb{R}$, $f^*: \mathbb{R}^n \mapsto \mathbb{R}$

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

几何意义即函数 $f(\mathbf{x})$ 到不同斜率直线的垂直距离最大值, \mathbf{y} 可粗略理解为斜率,且 $\mathbf{y}^T \mathbf{x}$ 一定过原点



由于 $f^*(\mathbf{x})$ 为一系列 \mathbf{y} 的凸函数的逐点上确界,故 f^* 为凸函数

例 16. 一些共轭函数的例子如下

- 仿射函数: $f(x) = ax + b$, 显然当斜率为 a 时 (即平行), 共轭函数才有界, 故共轭函数定义域为单点集 $\{a\}$, 且 $f^*(a) = -b$
- 最大熵函数: $f(\mathbf{x}) = \sum_{i=1}^n x_i \log x_i$

分析. 按照定义进行拆分计算即可

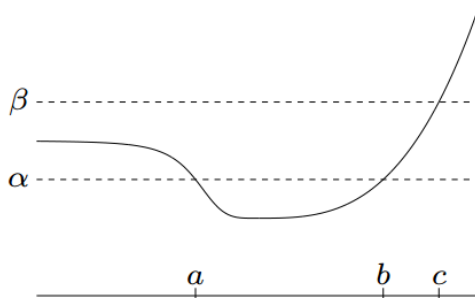
$$\begin{aligned} f^*(\mathbf{y}) &= \sup_{\mathbf{x}} \sum_{i=1}^n x_i \log x_i \\ &= \sum_{i=1}^n \sup_{x_i} y_i x_i - x_i \log x_i \quad \text{求导可得} \\ &= \sum_{i=1}^n e^{y_i - 1} \end{aligned}$$

3.5 拟凸函数

定义 12 (α 次水平集(α -sub level set)). $f: \mathbb{R}^n \mapsto \mathbb{R}$, $C_\alpha = \{\mathbf{x} \in \text{dom } f \mid f(\mathbf{x}) \leq \alpha\}$

定义 13 (拟凸函数(quasi-convex)). 所有 α 次水平集为凸集 $\iff f$ 为拟凸函数

拟凸函数有很好的性质 \rightarrow 单模态/单峰函数

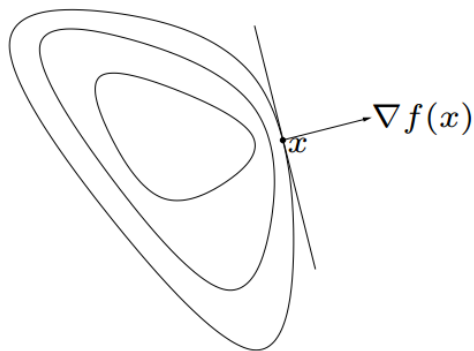


定理 3. 可微拟凸函数判定条件

- 一阶条件: $\text{dom } f$ 为凸集, 且

$$\forall \mathbf{x}, \mathbf{y} \in \text{dom } f : f(\mathbf{y}) \leq f(\mathbf{x}) \implies \nabla f(\mathbf{x})^T (\mathbf{y} - \mathbf{x}) \leq 0$$

当 $\nabla f(\mathbf{x}) \neq 0$ 时, 即 $\nabla f(\mathbf{x})$ 在 \mathbf{x} 处定义了水平集 $\{\mathbf{y} \mid f(\mathbf{y}) \leq f(\mathbf{x})\}$ 的支撑超平面

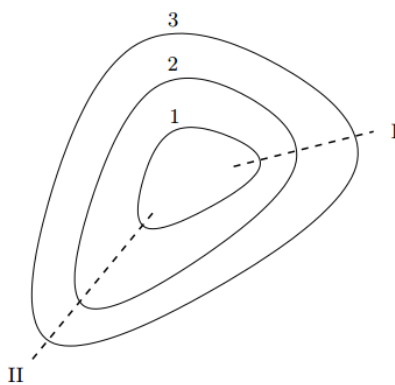


分析. 同样考虑一维的情况, 然后通过高维定义进行映射

- 二阶条件:

$$\forall \mathbf{x} \in \text{dom } f, \mathbf{y} \in \mathbb{R}^n : \mathbf{y}^T \nabla f(\mathbf{x}) = 0 \implies \mathbf{y}^T \nabla^2 f(\mathbf{x}) \mathbf{y} \geq 0$$

例 17 ((拟)凸函数的判定). 如下图, 水平集显然凸 (线段都在水平集内), 故为拟凸函数, 但对于直线 II , 水平集间距离并不是越来越密, 故不是凸函数



凸函数与凸集联系

- 凸函数定义域为凸集, 凸函数一定是拟凸函数
- 凸函数的 α 次水平集为凸集

4 凸优化问题

4.1 标准型

广义定义：极小化凸函数，约束为凸集

$$\begin{aligned} & \text{minimize} && f_0(\mathbf{x}) \\ & \text{subject to} && f_i(\mathbf{x}) \leq 0 \quad i = 1, \dots, m \\ & && h_j(\mathbf{x}) = 0 \quad j = 1, \dots, p \end{aligned}$$

- 优化变量 $\mathbf{x} \in \mathbb{R}^n$
- 目标/损失函数 $f_0 : \mathbb{R}^n \mapsto \mathbb{R}$
- 不等式约束函数 $f_i : \mathbb{R}^n \mapsto \mathbb{R}$
- 等式约束函数 $h_j : \mathbb{R}^n \mapsto \mathbb{R}$
- 定义域 $\mathcal{D} = \bigcap_{i=0}^m \text{dom } f_i \cap \bigcap_{j=1}^p \text{dom } h_j$ (注意不一定可行)
- 可行解 $\mathcal{X} = \{\mathbf{z} \mid f_i(\mathbf{z}) \leq 0, h_j(\mathbf{z}) = 0, i = 1, \dots, m, j = 1, \dots, p\}$
- 最优值(primal) $p^* = \inf\{f_0(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\}$
- 最优解 $\mathbf{x}^* \iff \forall \mathbf{z} \in \mathbb{R}^n, \mathbf{z} \in \mathcal{X} : f_0(\mathbf{z}) \geq f_0(\mathbf{x}^*)$
- 最优解集 $\mathcal{X}^* = \{\mathbf{x}^* \mid f_0(\mathbf{x}^*) = p^*, \mathbf{x}^* \in \mathcal{S}\}$
- ε -次优解集 $\mathcal{X}_\varepsilon = \{\mathbf{x} \mid f_0(\mathbf{x}) \leq p^* + \varepsilon, \mathbf{x} \in \mathcal{X}\}$
- 局部最优 $\exists R > 0, f_0(\mathbf{x}) = \inf\{f_0(\mathbf{z}) \mid \mathbf{x} \in \mathcal{X}, \mathbf{z} \in \mathcal{X}, \|\mathbf{x} - \mathbf{z}\| \leq R\}$, 即邻域内下确界
- 局部最优解集 $\mathbf{x}_{local} = \{\mathbf{x} \mid \mathbf{x} \text{ 为局部最优}\}$

狭义定义： $f_i(\mathbf{x}), i = 0, 1, \dots$ 为凸函数， $h_i(\mathbf{x})$ 为仿射函数

例 18. 将问题变换为标准型

$$\begin{aligned} \min \quad & f_0(x) = x_1^2 + x_2^2 \\ \text{s.t.} \quad & f_1(x) = \frac{x_1}{1 + x_2^2} \leq 0 \implies x_1 \leq 0 \\ & h_1(x) = (x_1 + x_2)^2 = 0 \implies x_1 + x_2 = 0 \end{aligned}$$

定理 4. 凸问题局部最优等价于全局最优

分析. 若 \mathbf{x} 为局部最优

$$\exists R > 0 : f_0(\mathbf{x}) = \inf\{f_0(\mathbf{z}) \mid \mathbf{z} \in \mathcal{X}, \mathbf{x} \in \mathcal{X}, \|\mathbf{x} - \mathbf{z}\|_2 \leq R\}$$

反证法, 设 \mathbf{x} 不是全局最优, \mathbf{y} 为全局最优, 即 $f_0(\mathbf{x}) > f_0(\mathbf{y})$

取 $\mathbf{z} = (1 - \theta)\mathbf{x} + \theta\mathbf{y}$ 为 \mathbf{x}, \mathbf{y} 连线上一点, 令 $\theta = \frac{R}{2\|\mathbf{y}-\mathbf{x}\|_2}$, 使 \mathbf{z} 能够落在 \mathbf{x} 的邻域内

$$\|\mathbf{z} - \mathbf{x}\|_2 = (1 - \theta) \|\mathbf{y} - \mathbf{x}\| = \frac{R}{2}$$

由 $\|\mathbf{z} - \mathbf{x}\|_2 \leq R \implies f_0(\mathbf{x}) \leq f_0(\mathbf{z})$, 又结合 $f_0(\mathbf{x}) > f_0(\mathbf{y})$, 有

$$f_0(\mathbf{z}) \leq \theta f_0(\mathbf{x}) + (1 - \theta)f_0(\mathbf{y}) < \theta f_0(\mathbf{z}) + (1 - \theta)f_0(\mathbf{z}) = f_0(\mathbf{z})$$

左边第一个不等号由凸函数定义, 第二个不等号由推导出的条件, 故矛盾

定理 5. 对于可微凸目标函数 $f_0(\mathbf{x})$, 最优解满足以下条件

- 无约束问题 $\min f_0(\mathbf{x})$, \mathbf{x}^* 为最优解, 当且仅当 $\mathbf{x}^* \in \mathcal{X}$, 且

$$\nabla f_0(\mathbf{x}^*) = 0$$

分析. 由凸函数的性质

$$\forall \mathbf{x}, \mathbf{y} : f_0(\mathbf{y}) \geq f_0(\mathbf{x}) + \langle \nabla f_0(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$$

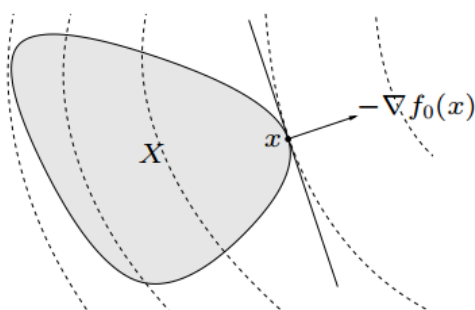
进而

$$f_0(\mathbf{y}) \geq f_0(\mathbf{x}^*) + \langle \nabla f_0(\mathbf{x}^*), \mathbf{y} - \mathbf{x}^* \rangle = f_0(\mathbf{x}^*)$$

- 有约束问题 $\min f_0(\mathbf{x}), s.t. \mathbf{x} \in \mathcal{X}$, \mathbf{x}^* 为最优解, 当且仅当 $\mathbf{x}^* \in \mathcal{X}$, 且

$$\forall \mathbf{y} \in \mathcal{X} : \langle \nabla f_0(\mathbf{x}^*), \mathbf{y} - \mathbf{x}^* \rangle \geq 0$$

相当于 $-\nabla f_0(\mathbf{x})$ 在 \mathbf{x} 上定义了可行解集的支撑超平面



例 19 (等式约束). 设 f_0 可微,

$$\begin{aligned} \min \quad & f_0(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{Ax} = \mathbf{b} \end{aligned}$$

分析. 设 \mathbf{x}^* 为最优解, 有 $A\mathbf{x}^* = \mathbf{b}$, 则最优解需满足

$$\forall \mathbf{y}, A\mathbf{y} = \mathbf{b} : \langle \nabla f_0(\mathbf{x}^*), \mathbf{y} - \mathbf{x}^* \rangle \geq 0$$

对 \mathbf{y} 进行改写

$$\begin{cases} \mathbf{y} = \mathbf{x}^* + \mathbf{v} \\ A\mathbf{v} = \mathbf{0} \end{cases}, \mathbf{v} \in \text{Nul } A$$

故最优解条件变为

$$\forall \mathbf{v} \in \text{Nul } A : \langle \nabla f_0(\mathbf{x}^*), \mathbf{v} \rangle \geq 0$$

那么

1. A 可逆, $\text{Nul } A = \{\mathbf{0}\}$
2. A 不可逆, $\nabla f_0(\mathbf{x}^*) \perp \text{Nul } A$

例 20 (非负约束). 设 f_0 可微,

$$\begin{aligned} \min \quad & f_0(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \succeq \mathbf{0} \end{aligned}$$

分析. 最优性条件为

$$\mathbf{x}^* \succeq \mathbf{0}, \forall \mathbf{y} \succeq \mathbf{0} : \langle \nabla f_0(\mathbf{x}^*), \mathbf{y} - \mathbf{x}^* \rangle \geq 0 \iff \langle \nabla f_0(\mathbf{x}^*), \mathbf{y} \rangle \geq \langle \nabla f_0(\mathbf{x}^*), \mathbf{x}^* \rangle$$

1. 若 $\nabla f_0(\mathbf{x}^*) \prec \mathbf{0}$, 则存在矛盾 ($\mathbf{y} - \mathbf{x}$ 全为正), 故 $\nabla f_0(\mathbf{x}^*) \succeq \mathbf{0}$
2. 取 $\mathbf{y} = \mathbf{0}$, 有

$$0 \geq \langle \nabla f_0(\mathbf{x}^*), \mathbf{x}^* \rangle \implies \sum_{i=1}^n (\nabla f_0(\mathbf{x}^*))_i x_i^* \leq 0 \implies (\nabla f_0(\mathbf{x}^*))_i x_i^* = 0$$

得到互补松弛条件, 即 \mathbf{x} 与 $\nabla f_0(\mathbf{x}^*)$ 的稀疏模式必须是互补的

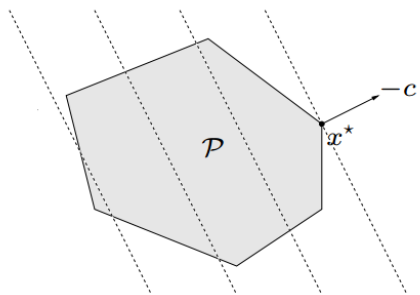
总结来说, 最优性条件为

$$\mathbf{x}^* \succeq \mathbf{0} \quad \nabla f_0(\mathbf{x}^*) \succeq \mathbf{0} \quad (\nabla f_0(\mathbf{x}^*))_i x_i^* = 0, i = 1, \dots, n$$

4.2 线性规划

当目标函数和约束函数都是仿射时, 问题称为线性规划(Linear Programming, LP)

$$\begin{aligned} \min \quad & \mathbf{c}^T \mathbf{x} + d \\ \text{s.t.} \quad & G\mathbf{x} \preceq \mathbf{h} \\ & A\mathbf{x} = \mathbf{b} \end{aligned}$$



通过引入松弛变量，对原问题进行变换

$$\begin{aligned} \min \quad & \mathbf{c}^T \mathbf{x} + d \\ \text{s.t.} \quad & G\mathbf{x} + \mathbf{s} = \mathbf{h} \\ & A\mathbf{x} = \mathbf{b} \\ & \mathbf{s} \succeq \mathbf{0} \end{aligned}$$

进一步，将 \mathbf{x} 表示为两个非负变量 \mathbf{x}^+ 和 \mathbf{x}^- 的差，即 $\mathbf{x} = \mathbf{x}^+ + \mathbf{x}^-$ ，得到标准型线性规划问题（不等式都是分量的非负约束）

$$\begin{aligned} \min \quad & \mathbf{c}^T \mathbf{x}^+ - \mathbf{c}^T \mathbf{x}^- + d \\ \text{s.t.} \quad & G\mathbf{x}^+ - G\mathbf{x}^- + \mathbf{s} = \mathbf{h} \\ & A\mathbf{x}^+ - A\mathbf{x}^- = \mathbf{b} \\ & \mathbf{x}^+ \succeq \mathbf{0} \\ & \mathbf{x}^- \succeq \mathbf{0} \\ & \mathbf{s} \succeq \mathbf{0} \end{aligned}$$

例 21 (食谱问题). m 种营养元素不小于 b_1, \dots, b_m , n 种食物, 单位含量 a_{1j}, \dots, a_{mj} , 食物量 x_1, \dots, x_n , 价格 c_1, \dots, c_n

$$\begin{aligned} \min \quad & \sum_{j=1}^n c_j x_j \\ \text{s.t.} \quad & \sum_{j=1}^n a_{ij} x_j \geq b_i \\ & x_j \geq 0 \end{aligned}$$

其中 $i = 1, \dots, m, j = 1, \dots, n$

例 22 (线性分数规划). 拟凸优化问题

$$\begin{aligned} \min \quad & f_0(\mathbf{x}) = \frac{\mathbf{c}^T \mathbf{x} + d}{\mathbf{e}^T \mathbf{x} + f}, \text{dom } \mathbf{f} = \{\mathbf{x} \mid \mathbf{e}^T \mathbf{x} + f > 0\} \\ \text{s.t.} \quad & G\mathbf{x} \preceq \mathbf{h} \\ & A\mathbf{x} = \mathbf{b} \end{aligned}$$

等价于

$$\begin{aligned} \min_{\mathbf{y}, z} \quad & \mathbf{c}^T \mathbf{y} + dz \\ \text{s.t.} \quad & G\mathbf{y} - \mathbf{h}z \preceq 0 \\ & A\mathbf{y} - \mathbf{b}z = 0 \\ & \mathbf{e}^T \mathbf{y} + fz = 1 \\ & z \geq 0 \end{aligned}$$

分析. 证明两个问题等价, P_0 与 P_1

若 \mathbf{x} 在 P_0 内可行

$$\mathbf{y} = \frac{\mathbf{x}}{\mathbf{e}^T \mathbf{x} + f}, z = \frac{1}{\mathbf{e}^T \mathbf{x} + f}$$

若 (\mathbf{y}, z) 在 P_1 中可行

$$\mathbf{x} = \frac{\mathbf{y}}{z} (z \neq 0)$$

若 $z = 0$, \mathbf{x}_0 为 P_0 的可行解

$$\begin{aligned} \mathbf{x} &= \mathbf{x}_0 + t\mathbf{y}, t \geq 0 \\ \lim_{t \rightarrow \infty} \frac{\mathbf{c}^T(\mathbf{x}_0 + t\mathbf{y}) + d}{\mathbf{e}^T(\mathbf{x}_0 + t\mathbf{y}) + f} &= \mathbf{c}^T \mathbf{y} \end{aligned}$$

代入看所有条件结论都相同

4.3 二次规划

当凸优化问题的目标函数为(凸)二次型且约束函数为仿射时, 该问题称为二次规划(Quadratic Programming, QP)

$$\begin{aligned} \min \quad & \frac{1}{2} \mathbf{x}^T P \mathbf{x} + \mathbf{q}^T \mathbf{x} + r, \quad P \succ 0 \\ \text{s.t.} \quad & G\mathbf{x} \preceq \mathbf{h} \\ & A\mathbf{x} = \mathbf{b} \end{aligned}$$

其中 $P \in \mathbb{S}_+^n, G \in \mathbb{R}^{m \times n}, A \in \mathbb{R}^{p \times n}$

若不等式约束也是（凸）二次型，则成为二次约束二次规划(QCQP)

$$\begin{aligned} \min \quad & \frac{1}{2} \mathbf{x}^T P_0 \mathbf{x} + \mathbf{q}_0^T \mathbf{x} + r_0, P_0 \succ 0 \\ \text{s.t.} \quad & \frac{1}{2} \mathbf{x}^T P_i \mathbf{x} + \mathbf{q}_i^T \mathbf{x} + r_i \leq 0, i = 1, \dots, m \\ & A\mathbf{x} = \mathbf{b} \end{aligned}$$

其中 $P_i \in \mathbb{S}_+^n$

例 23 (最小二乘问题的改写).

$$\begin{aligned} \min_{\mathbf{x}} \quad & \frac{1}{2} \|A\mathbf{x} - \mathbf{b}\|_2^2 \\ \text{s.t.} \quad & A\mathbf{x} + \mathbf{e} = \mathbf{b} \end{aligned}$$

分析. 一范数规范化最小二乘

$$\min \frac{1}{2} \|A\mathbf{x} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{x}\|_1$$

本来用零范数，但用一范数拟合，改写

$$\|\mathbf{x}\|_1 = \mathbf{1}^T \mathbf{x}^+ + \mathbf{1}^T \mathbf{x}^-$$

Basic Pursuit

$$\begin{aligned} \min \quad & \frac{1}{2} \|A\mathbf{x} - \mathbf{b}\|_2^2 \\ \text{s.t.} \quad & \|\mathbf{x}\|_1 \leq \varepsilon_1 \end{aligned}$$

原式很难平衡两者，下式只需考虑 $\|\mathbf{x}\|_1$ 的影响

采用岭回归 (Ridge): 所有 \mathbf{x} 差距不要太大

$$\min \frac{1}{2} \|A\mathbf{x} - \mathbf{b}\|_2^2 + \frac{1}{2} \lambda_2 \|\mathbf{x}\|_2^2$$

可得类似的问题改写

$$\begin{aligned} \min \quad & \frac{1}{2} \|A\mathbf{x} - \mathbf{b}\|_2^2 \\ \text{s.t.} \quad & \|\mathbf{x}\|_2^2 \leq \varepsilon_2 \end{aligned}$$

例 24 (投资组合问题(portfolio optimization)). 初始价格 x_1, \dots, x_n , 最终价格 $P_1 x_1, \dots, P_n x_n$

$$\begin{aligned} \max \quad & P_1 x_1 + \dots + P_n x_n \\ \text{s.t.} \quad & x_1 + \dots + x_n = B \\ & x_1, \dots, x_n \geq 0 \end{aligned}$$

分析. $\bar{P} = \mathbb{E}(P)$ 已知, $\Sigma = \mathbb{D}(P)$

$$\begin{aligned} \min \quad & \mathbf{x}^T \Sigma \mathbf{x} \\ \text{s.t.} \quad & \mathbf{p}^T \mathbf{x} \geq r_{\min} \\ & \mathbf{1}^T \mathbf{x} = 1, \mathbf{x} \succeq \mathbf{0} \end{aligned}$$

二次锥规划问题(SOCP)

$$\begin{aligned} \min \quad & \mathbf{f}^T \mathbf{x} \\ \text{s.t.} \quad & \|A_i \mathbf{x} + \mathbf{b}_i\|_2 \leq \mathbf{c}_i^T \mathbf{x} + \mathbf{d}_i \\ & F\mathbf{x} = \mathbf{g} \end{aligned}$$

4.4 广义不等式约束

半定规划(semi-definite programming, SDP)为矩阵意义下的线性规划问题: $X \in \mathbb{R}^{n \times n}, C \in \mathbb{R}^{n \times n}, A_i \in \mathbb{R}^{n \times n}, b_i \in \mathbb{R}$

$$\begin{aligned} \min \quad & \text{tr}(CX) \\ \text{s.t.} \quad & \text{tr}(A_i X) = b_i, i = 1, \dots, p \\ & X \succeq 0 \end{aligned}$$

例 25 (谱范数极小化问题). 矩阵多项式 $A(\mathbf{x}) = A_0 + x_1 A_1 + \dots + x_n A_n, A_i \in \mathbb{R}^{p \times q}$

$$\min_{\mathbf{x}} \|A(\mathbf{x})\|_2$$

谱范数代表 $A(\mathbf{x})$ 的最大奇异值⁹

分析. 这是一个 $q \times q$ 矩阵不等式约束的凸优化问题

$$\begin{aligned} \min_{x, s} \quad & s \\ \text{s.t.} \quad & A^T(\mathbf{x})A(\mathbf{x}) \preceq sI \end{aligned}$$

例 26 (最快分布式线性平均). 图的最速混合 *Markov* 链

$$\mathbf{x}(t) = P\mathbf{x}(t-1)$$

⁹谱范数是诱导范数, F-范数(Frobenias) $\|A(\mathbf{x})\|_F$ 才算是矩阵意义下的2范数

其中 P 为邻接矩阵

$$P = \begin{bmatrix} P_{11} & \cdots & P_{1n} \\ \vdots & \ddots & \vdots \\ P_{n1} & \cdots & P_{nn} \end{bmatrix}, P\mathbf{1} = \mathbf{1}$$

其中 $(i, j) \in E$ 或 $i = j$, $P_{ij} \neq 0$; 否则 $P_{ij} = 0$

$$P = P^T, P_{ij} = P_{ji}, P_{ij} > 0, P \succeq 0, P_{ij} \geq 0$$

分析. 只要图是连通图, 则一定会收敛

收敛速度与第二大[特征值绝对值]有关

$$1 = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq -1$$

即收敛速度由 $\max\{|\lambda_2|, |\lambda_n|\}$ 决定

$$\min \max\{|\lambda_2|, |\lambda_n|\}$$

$$\max\{|\lambda_2|, |\lambda_n|\} = \left\| P - \frac{1}{n} \mathbf{1}\mathbf{1}^T \right\|$$

$$\begin{aligned} \min \quad & t := \left\| P - \frac{1}{n} \mathbf{1}\mathbf{1}^T \right\|_2 \\ \text{s.t.} \quad & P\mathbf{1} = \mathbf{1} \\ & P = P^T \\ & P \succeq 0 \\ & P_{ij} = 0, \quad (i, j) \notin E \wedge i \neq j \\ & -tI \preceq P - \frac{1}{n} \mathbf{1}\mathbf{1}^T \preceq tI \end{aligned}$$

4.5 多目标优化问题

帕累托最优解: 若有另一解在某个指标上更好, 则必有指标更差

帕累托最优值/帕累托最优面: $\min f_{01}$ 与 $\min f_{02}$ 的交点为理想点(oracle)

若 $f_{01}(x), \dots, f_{0q}(x)$ 为凸, \mathcal{X} 为凸

$$\begin{aligned} \min \quad & \lambda_1 f_{01}(x) + \cdots + \lambda_q f_{0q}(x), \quad \lambda_1, \dots, \lambda_q \geq 0 \\ \text{s.t.} \quad & x \in \mathcal{X} \end{aligned}$$

1. 能找到一个Pareto最优解
2. 遍历 $\lambda_1, \dots, \lambda_q$, 可找到全部

岭回归的多目标优化表示

$$\begin{cases} \min \frac{1}{2} \|Ax - b\|_2^2 \\ \min \frac{1}{2} \|x\|_2^2 \end{cases}$$

5 对偶理论

5.1 拉格朗日对偶

定义 14 (拉格朗日函数(Lagrangian function)). $f_i(\mathbf{x})$ 和 $h_j(\mathbf{x})$ 含义同4.1节, 前者不等式约束(≤ 0), 后者等式约束。

$$\begin{aligned} L(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{v}) &= f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^p v_i h_i(\mathbf{x}) \\ &= f_0(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{f}(\mathbf{x}) + \mathbf{v}^T \mathbf{h}(\mathbf{x}), \text{ dom } L = \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p \end{aligned}$$

定义 15 (拉格朗日乘子(multiplier)). 对偶变量即拉格朗日乘子

- 原变量(*primal variable*): \mathbf{x}
- 对偶变量(*dual variable*): $\boldsymbol{\lambda} = [\lambda_1 \ \cdots \ \lambda_m]^T$, $\mathbf{v} = [v_1 \ \cdots \ v_p]^T$

定义 16 (拉格朗日对偶函数).

$$\begin{aligned} g(\boldsymbol{\lambda}, \mathbf{v}) &= \inf_{\mathbf{x} \in \mathcal{D}} L(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{v}) \\ &= \inf_{\mathbf{x} \in \mathcal{D}} \left(f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^p v_i h_i(\mathbf{x}) \right) \end{aligned}$$

注意遍历域是 $\mathcal{D} = \bigcap_{i=1}^m \text{dom } f_i \cap \bigcap_{i=1}^p \text{dom } h_i$, 而不是可行解集 \mathcal{X}

如果约束 f 和 h 为标量, 那么 v 和 λ 也对应改为标量; 如果有多个向量约束, 可对应添加对偶变量

有以下两点性质:

- $g(\boldsymbol{\lambda}, \mathbf{v})$ 一定是关于 $\boldsymbol{\lambda}$ 和 \mathbf{v} 的凹函数 (关于 $\boldsymbol{\lambda}$ 和 \mathbf{v} 的仿射函数, 注意 \mathbf{x} 为常数)
- $\forall \boldsymbol{\lambda} \succeq \mathbf{0}, \forall \mathbf{v} : g(\boldsymbol{\lambda}, \mathbf{v}) \leq p^*$

定义 17 (对偶(dual)问题).

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

其最优解记为 d^* , 则 $d^* \leq p^*$, 即给出了原问题(*primal*)的一个最优下界

分析. 记 \mathbf{x}^* 为原问题最优解, 由优化问题定义有

$$\sum_{i=1}^m \lambda_i f_0(\mathbf{x}^*) + \sum_{i=1}^p v_i h_i(\mathbf{x}^*) \leq 0$$

又 $p^* = f_0(\mathbf{x}^*)$ 为原问题最优解，将 \mathbf{x}^* 代入 *Lagrange* 函数中

$$L(\mathbf{x}^*, \boldsymbol{\lambda}, \mathbf{v}) = f_0(\mathbf{x}^*) + \left(\sum_{i=1}^m \lambda_i f_0(\mathbf{x}^*) + \sum_{i=1}^p v_i h_i(\mathbf{x}^*) \right) \leq p^*$$

进而推出

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

例 27 (最小二乘).

$$\begin{aligned} \min \quad & \|\mathbf{x}\|_2 = \mathbf{x}^T \mathbf{x} \\ \text{s.t.} \quad & A\mathbf{x} = \mathbf{b} \end{aligned}$$

分析.

$$\begin{aligned} L(\mathbf{x}, \mathbf{v}) &= \mathbf{x}^T \mathbf{x} + \mathbf{v}^T (A\mathbf{x} - \mathbf{b}) \\ g(\mathbf{v}) &= \inf_{\mathbf{x} \in \mathcal{D}} L(\mathbf{x}, \mathbf{v}) \\ &= \inf_{\mathbf{x} \in \mathcal{D}} (\mathbf{x}^T \mathbf{x} + \mathbf{v}^T A\mathbf{x} - \mathbf{v}^T \mathbf{b}) \quad \text{求最小值相当于求导/求梯度代入} \\ &= \left(-\frac{A^T \mathbf{v}}{2} \right)^T \left(-\frac{A^T \mathbf{v}}{2} \right) + \mathbf{v}^T A \left(-\frac{A^T \mathbf{v}}{2} \right) - \mathbf{v}^T \mathbf{b} \\ &= -\frac{1}{4} \mathbf{v}^T A A^T \mathbf{v} - \mathbf{b}^T \mathbf{v} \end{aligned}$$

$$\text{补充求梯度: } \nabla_{\mathbf{x}} L(\mathbf{x}, \mathbf{v}) = 2\mathbf{x} + (\mathbf{v}^T A)^T = 0 \implies \mathbf{x} = -\frac{A^T \mathbf{v}}{2}$$

因而得到对偶问题

$$\max_{\mathbf{v}} \left(-\frac{1}{4} \mathbf{v}^T A A^T \mathbf{v} - \mathbf{b}^T \mathbf{v} \right)$$

例 28 (标准线性规划).

$$\begin{aligned} \min \quad & \mathbf{c}^T \mathbf{x} \\ \text{s.t.} \quad & A\mathbf{x} = \mathbf{b} \\ & \mathbf{x} \succeq \mathbf{0} \end{aligned}$$

分析. 1° 注意 $\boldsymbol{\lambda}$ 前面符号，要化为一般形式

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

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

$$\text{一次式应直接分拆变量和常量, 由于 } \mathbf{x} \text{ 可以任取, 故会到负无穷} = \begin{cases} -\infty & \mathbf{c} - \boldsymbol{\lambda} + A^T \mathbf{v} \neq \mathbf{0} \\ -\mathbf{v}^T \mathbf{b} & \mathbf{c} - \boldsymbol{\lambda} + A^T \mathbf{v} = \mathbf{0} \end{cases}$$

由于要极大，故不考虑负无穷部分，得到对偶问题如下

$$\begin{aligned} \max_{\lambda, \mathbf{v}} \quad & -\mathbf{v}^T \mathbf{b} \\ \text{s.t.} \quad & \mathbf{c} - \lambda + A^T \mathbf{v} = 0 \\ & \lambda \succeq \mathbf{0} \end{aligned}$$

2° 考虑对偶问题的对偶，先将对偶问题变为极小化问题

$$\begin{aligned} \min \quad & \mathbf{b}^T \mathbf{v} \\ \text{s.t.} \quad & A^T \mathbf{v} + \mathbf{c} \succeq \mathbf{0} \end{aligned}$$

$$L(\mathbf{v}, \lambda) = \mathbf{b}^T \mathbf{v} - \lambda^T (A^T \mathbf{v} + \mathbf{c})$$

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

得到新的对偶问题

$$\begin{aligned} \max \quad & -\lambda^T \mathbf{c} \\ \text{s.t.} \quad & \mathbf{b} - A\lambda = 0 \\ & \lambda \succeq \mathbf{0} \end{aligned}$$

会发现，对偶的对偶不一定回去，线性规划才满足

例 29 (二路分划(two-way partitioning)). 非凸问题，考虑可行解集有 2^n 个离散点，将 $\{1, \dots, n\}$ 分划到两个集合中， W_{ij} 是将 i, j 指派到同一个集合的开销， $-W_{ij}$ 是将 i, j 指派到不同集合的开销

$$\begin{aligned} \min \quad & \mathbf{x}^T W \mathbf{x} \\ \text{s.t.} \quad & x_i = \pm 1, \quad i = 1, \dots, n \end{aligned}$$

分析. 变为平方等式约束

$$\begin{aligned}
L(\mathbf{x}, \mathbf{v}) &= \mathbf{x}^T W \mathbf{x} + \sum_{i=1}^n v_i (x_i^2 - 1) \\
g(\mathbf{v}) &= \inf_{\mathbf{x}} L(\mathbf{x}, \mathbf{v}) \\
&= \inf_{\mathbf{x}} \left(\mathbf{x}^T W \mathbf{x} + \sum_{i=1}^n v_i x_i^2 - \sum_{i=1}^n v_i \right) \quad \text{二次项放对角线} \\
&= \inf_{\mathbf{x}} \left(\mathbf{x}^T (W + \text{diag}(\mathbf{v})) \mathbf{x} - \mathbf{1}^T \mathbf{v} \right) \\
&= \begin{cases} -\mathbf{1}^T \mathbf{v} & W + \text{diag}(\mathbf{v}) \succeq \mathbf{0} \\ -\infty & \text{otherwise} \end{cases}
\end{aligned}$$

其中

$$\text{diag} \mathbf{v} = \begin{bmatrix} v_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & v_n \end{bmatrix}$$

得到对偶问题如下

$$\begin{aligned}
\max_{\mathbf{v}} \quad & -\mathbf{1}^T \mathbf{v} \\
\text{s.t.} \quad & W + \text{diag}(\mathbf{v}) \succeq \mathbf{0}
\end{aligned}$$

例 30 (共轭函数).

$$\begin{aligned}
\min \quad & f_0(\mathbf{x}) \\
\text{s.t.} \quad & A\mathbf{x} \leq \mathbf{b} \\
& C\mathbf{x} = \mathbf{d}
\end{aligned}$$

分析.

$$\begin{aligned}
L(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{v}) &= f_0(\mathbf{x}) + \boldsymbol{\lambda}^T (A\mathbf{x} - \mathbf{b}) + \mathbf{v}^T (C\mathbf{x} - \mathbf{d}) \\
&= f_0(\mathbf{x}) + (A^T \boldsymbol{\lambda} + C^T \mathbf{v})^T \mathbf{x} - \boldsymbol{\lambda}^T \mathbf{b} - \mathbf{v}^T \mathbf{d} \\
g(\boldsymbol{\lambda}, \mathbf{v}) &= \inf_{\mathbf{x}} (f_0(\mathbf{x}) + (A^T \boldsymbol{\lambda} + C^T \mathbf{v})^T \mathbf{x} - \boldsymbol{\lambda}^T \mathbf{b} - \mathbf{v}^T \mathbf{d}) \\
&= -\sup_{\mathbf{x}} \left(-(A^T \boldsymbol{\lambda} + C^T \mathbf{v})^T \mathbf{x} - f_0(\mathbf{x}) \right) - \boldsymbol{\lambda}^T \mathbf{b} - \mathbf{v}^T \mathbf{d} \\
&= -f_0^*(-(A^T \boldsymbol{\lambda} + C^T \mathbf{v})) - \boldsymbol{\lambda}^T \mathbf{b} - \mathbf{v}^T \mathbf{d}
\end{aligned}$$

5.2 强弱对偶

定义 18 (对偶间隙(duality gap)). $p^* - d^* \geq 0$

- 弱对偶: 严格大于 0

- 强对偶：对偶间隙为0

1. 对于非凸问题，通常 $p^* \neq d^*$
2. 对于凸问题，若满足Slater条件，则 $p^* = d^*$

定义 19 (相对内点(relative interior)). 存在一个邻域所有的点都落在集合内，即多元微积分中内点的概念。而相对内点则是考虑仿射包 $\text{aff } \mathcal{D}$ 的情况

$$\text{relint } \mathcal{D} = \{\mathbf{x} \in \mathcal{D} \mid B(\mathbf{x}, r) \cap \text{aff } \mathcal{D} \subset \mathcal{D}, \exists r > 0\}$$

定理 6 (Slater条件). 强对偶对于下列凸问题成立

$$\begin{aligned} \min \quad & f_0(\mathbf{x}) \\ \text{s.t.} \quad & f_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, m \\ & A\mathbf{x} = \mathbf{b} \end{aligned}$$

若它严格可行，即

$$\exists \mathbf{x} \in \text{relint } \mathcal{D}, \text{ s.t. } f_i(\mathbf{x}) < 0, \quad i = 1, \dots, m, \quad A\mathbf{x} = \mathbf{b}$$

例 31. 二次规划(QP)

$$\begin{aligned} \min \quad & \mathbf{x}^T \mathbf{x} \\ \text{s.t.} \quad & A\mathbf{x} = \mathbf{b} \end{aligned}$$

Slater条件 $\{\mathbf{x} \mid A\mathbf{x} = \mathbf{b}\}$ 非空

例 32 (二次约束二次规划(QCQP)).

$$\begin{aligned} \min \quad & \frac{1}{2} \mathbf{x}^T P_0 \mathbf{x} + \mathbf{q}_0^T \mathbf{x} + r_0 \\ \text{s.t.} \quad & \frac{1}{2} \mathbf{x}^T P_i \mathbf{x} + \mathbf{q}_i^T \mathbf{x} + r_i \leq 0, \quad i = 1, \dots, m \end{aligned}$$

P_0, \dots, P_m 半正定

凸问题+Slater条件 $\implies p^* = d^*$ ，但有可能不满足Slater条件也依然强对偶，如下面的例子。

例 33.

$$\begin{aligned} \min \quad & x, \quad x \in \mathbb{R} \\ \text{s.t.} \quad & x \leq 0 \\ & -x \leq 0 \end{aligned}$$

分析.

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

$$g(\lambda_1, \lambda_2) = \inf_{x \in \mathbb{R}} (1 + \lambda_1 - \lambda_2)x = \begin{cases} 0 & 1 + \lambda_1 - \lambda_2 = 0 \\ -\infty & \text{otherwise} \end{cases}$$

得到对偶问题

$$\begin{aligned} \max_{\lambda_1, \lambda_2} \quad & 0 \\ \text{s.t.} \quad & 1 + \lambda_1 - \lambda_2 = 0 \end{aligned}$$

可以推出 $p^* = d^* = 0$

例 34 (置信域问题).

$$\begin{aligned} \min \quad & \mathbf{x}^T A \mathbf{x} + \mathbf{b}^T \mathbf{x} \\ \text{s.t.} \quad & \mathbf{x}^T \mathbf{x} \leq 1 \\ & A \neq 0 \end{aligned}$$

依然可以得到 $p^* = d^*$

5.3 对偶问题的几种解释

5.3.1 几何解释

考虑问题只有一个约束 $f_1(\mathbf{x}) \leq 0$, 记 $\mathcal{G} = \{(f_1(\mathbf{x}), f_0(\mathbf{x})) \mid \mathbf{x} \in \mathcal{D}\}$, 则原问题最优解

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

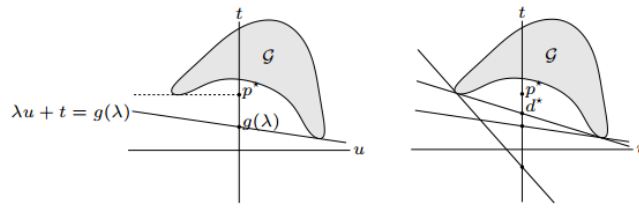
对偶函数如下, 相当于给定斜率, 在 \mathcal{G} 内找点使 $t + \lambda u$ 最小

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

对偶问题为

$$\begin{aligned} \max \quad & g(\lambda) \\ \text{s.t.} \quad & \lambda \geq 0 \end{aligned}$$

左图即在 $u \leq 0$ 的部分找 t 的最小值, 得到 p^* ; 右图得到对偶函数的最大值 d^*



注意问题必须要有可行解

5.3.2 经济学解释

满足原材料约束下，利润最多，价格 $\lambda_i \geq 0$ ，极小化成本

$$g(\boldsymbol{\lambda}) = \inf_{\mathbf{x}} (f_0(\mathbf{x}) + \lambda_1 f_1(\mathbf{x}) + \cdots + \lambda_m f_m(\mathbf{x})) = \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda})$$

则 $g(\boldsymbol{\lambda})$ 为对偶函数，市场 p^* 损失最小 ($g(\boldsymbol{\lambda}) \leq p^*$)

$$d^* = \sup_{\boldsymbol{\lambda}} g(\boldsymbol{\lambda})$$

市场平衡点，均衡市场 $p^* = d^*$ ，最优/影子价格 $\boldsymbol{\lambda}^*$

5.3.3 多目标优化解释

$$\begin{cases} \min f_0(\mathbf{x}) & 1 \\ \min f_1(\mathbf{x}) & \lambda_1 \\ \vdots & \vdots \\ \min f_m(\mathbf{x}) & \lambda_m \end{cases}$$

$$\min_{\mathbf{x}} (f_0(\mathbf{x}) + \lambda_1 f_1(\mathbf{x}) + \cdots + \lambda_m f_m(\mathbf{x}))$$

5.3.4 鞍点(saddle point)解释

考虑函数 $f(\mathbf{w}, \mathbf{z})$, $\mathbf{w} \in \mathcal{W}$, $\mathbf{z} \in \mathcal{Z}$ 极小极大不等式

$$\sup_{\mathbf{z} \in \mathcal{Z}} \inf_{\mathbf{w} \in \mathcal{W}} f(\mathbf{w}, \mathbf{z}) \leq \inf_{\mathbf{w} \in \mathcal{W}} \sup_{\mathbf{z} \in \mathcal{Z}} f(\mathbf{w}, \mathbf{z})$$

定义 20 (鞍点). 若有 $(\tilde{\mathbf{w}}, \tilde{\mathbf{z}})$ 使得

$$(\tilde{\mathbf{w}}, \tilde{\mathbf{z}}) = \arg \max_{\mathbf{z} \in \mathcal{Z}} \min_{\mathbf{w} \in \mathcal{W}} f(\mathbf{w}, \mathbf{z})$$

$$(\tilde{\mathbf{w}}, \tilde{\mathbf{z}}) = \arg \min_{\mathbf{w} \in \mathcal{W}} \max_{\mathbf{z} \in \mathcal{Z}} f(\mathbf{w}, \mathbf{z})$$

则 $(\tilde{\mathbf{w}}, \tilde{\mathbf{z}})$ 为鞍点, 即下面不等式成立

$$\forall \mathbf{z} \in \mathcal{Z}, \mathbf{w} \in \mathcal{W}: f(\tilde{\mathbf{w}}, \mathbf{z}) \leq f(\tilde{\mathbf{w}}, \tilde{\mathbf{z}}) \leq f(\mathbf{w}, \tilde{\mathbf{z}})$$

相当于从一个方向望过去是最小, 从另一个方向望过去是最大

$$\begin{aligned} L(\mathbf{x}, \boldsymbol{\lambda}) &= f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}) \\ \implies \sup_{\boldsymbol{\lambda} \succeq \mathbf{0}} L(\mathbf{x}, \boldsymbol{\lambda}) &= \sup_{\boldsymbol{\lambda} \succeq \mathbf{0}} \left\{ f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}) \right\} = \begin{cases} f_0(\mathbf{x}) & f_i(\mathbf{x}) \leq 0, i = 1, \dots, m \\ +\infty & \text{otherwise} \end{cases} \\ \implies p^* &= \inf_{\mathbf{x}} \{ f_0(\mathbf{x}) \mid f_i(\mathbf{x}) \leq 0, i = 1, \dots, m \} = \inf_{\mathbf{x}} \sup_{\boldsymbol{\lambda} \succeq \mathbf{0}} L(\mathbf{x}, \boldsymbol{\lambda}) \end{aligned}$$

由极小极大不等式可得

$$d^* = \sup_{\boldsymbol{\lambda} \succeq \mathbf{0}} \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}) \leq \inf_{\mathbf{x}} \sup_{\boldsymbol{\lambda} \succeq \mathbf{0}} L(\mathbf{x}, \boldsymbol{\lambda}) = p^*$$

鞍点在无约束优化问题中是很糟糕的点 (所有方向上梯度为0), 但是有约束优化问题中则是非常好的点。

定理 7. $(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\lambda}})$ 为拉格朗日函数鞍点 $\iff p^* = d^*$, 且 $(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\lambda}})$ 为原对偶问题的最优解

$$\begin{cases} \tilde{\mathbf{x}} = \arg \inf_{\mathbf{x}} \sup_{\boldsymbol{\lambda} \succeq \mathbf{0}} L(\mathbf{x}, \boldsymbol{\lambda}) \\ \tilde{\boldsymbol{\lambda}} = \arg \max_{\boldsymbol{\lambda} \succeq \mathbf{0}} \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}) \end{cases}$$

分析. 右推左, $(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\lambda}})$ 为原对偶问题可行解

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

因 $p^* = d^*$, 有

$$\begin{aligned} f_0(\tilde{\mathbf{x}}) &= g(\tilde{\boldsymbol{\lambda}}) \\ &= \inf_{\mathbf{x}} \left\{ f_0(\mathbf{x}) + \sum_{i=1}^m \tilde{\lambda}_i f_i(\mathbf{x}) \right\} \quad \text{最小值必然比其他值都小} \\ &\leq f_0(\tilde{\mathbf{x}}) + \sum_{i=1}^m \tilde{\lambda}_i f_i(\tilde{\mathbf{x}}) \quad \text{约束条件} \\ &\leq f_0(\tilde{\mathbf{x}}) \end{aligned}$$

进而不等号都得为等号, 还可得到

$$1. \inf_{\mathbf{x}} L(\mathbf{x}, \tilde{\boldsymbol{\lambda}}) = L(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\lambda}}) \quad (\text{由第一条不等式})$$

$$2. f_0(\tilde{\mathbf{x}}) = \sup_{\boldsymbol{\lambda} \succeq \mathbf{0}} \left\{ f_0(\tilde{\mathbf{x}}) + \sum_{i=1}^m \lambda_i f_i(\tilde{\mathbf{x}}) \right\} = \sup_{\boldsymbol{\lambda} \succeq \mathbf{0}} L(\tilde{\mathbf{x}}, \boldsymbol{\lambda}) = L(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\lambda}}) \quad (\text{由第二条不等式})$$

故右推左成立

5.4 一般优化问题的对偶理论

$$\begin{aligned} \min \quad & f_0(\mathbf{x}) \\ \text{s.t.} \quad & f_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, m \\ & h_i(\mathbf{x}) = 0, \quad i = 1, \dots, p \end{aligned}$$

不一定是凸问题，但 $p^* = d^*$ ，最优解满足什么条件？

设其对偶问题为下式

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

分析. 设 $(\mathbf{x}^*, \boldsymbol{\lambda}^*, \mathbf{v}^*)$ 为原对偶最优解，则 $(\mathbf{x}^*, \boldsymbol{\lambda}^*, \mathbf{v}^*)$ 为原对偶可行解

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

$$\begin{aligned} p^* = d^* &\implies f_0(\mathbf{x}^*) = g(\boldsymbol{\lambda}^*, \mathbf{v}^*) \\ &= \inf_{\mathbf{x}} \left\{ f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}) + \sum_{i=1}^p v_i^* h_i(\mathbf{x}) \right\} \quad \text{将inf拆掉} \\ &\leq f_0(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}^*) + \sum_{i=1}^p v_i^* h_i(\mathbf{x}^*) \quad \text{原问题约束} \\ &\leq f_0(\mathbf{x}^*) \end{aligned}$$

同上理，不等号全取等，有

1. $\lambda_i^* f_i(\mathbf{x}^*) = 0, \forall i = 1, \dots, m$
2. $\mathbf{x}^* = \arg \min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}^*, \mathbf{v}^*)$

若 f_0, f_i, h_i 均可微，则必要条件为

$$\left. \frac{\partial L(\mathbf{x}, \boldsymbol{\lambda}^*, \mathbf{v}^*)}{\partial \mathbf{x}} \right|_{\mathbf{x}=\mathbf{x}^*} = \mathbf{0}$$

定理 8 (KKT条件). 假设 $f_0, \dots, f_m, h_1, \dots, h_p$ 可微，无论这些函数是否凸，最优解必须满足KKT(Karush-Kuhn-Tucker, KKT)条件

- 原始可行性(primal feasibility): $f_i(\mathbf{x}^*) \leq 0, i = 1, \dots, m$
- 原始可行性(primal feasibility): $h_i(\mathbf{x}^*) = 0, i = 1, \dots, p$
- 对偶可行性(dual feasibility): $\boldsymbol{\lambda}^* \succeq \mathbf{0}$
- 对偶互斥条件(complementarity slackness): $\lambda_i^* f_i(\mathbf{x}^*) = 0, i = 1, \dots, m$
- 稳定性(stability): $\left. \frac{\partial L(\mathbf{x}, \boldsymbol{\lambda}^*, \mathbf{v}^*)}{\partial \mathbf{x}} \right|_{\mathbf{x}=\mathbf{x}^*} = \mathbf{0}$

若原问题为凸，则KKT条件为最优解的充要条件

分析. 必要性已证, 证明充分性

若 $(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\lambda}}, \tilde{v})$ 满足KKT条件 $\implies (\tilde{\mathbf{x}}, \tilde{\boldsymbol{\lambda}}, \tilde{v})$ 最优, 其中, $\tilde{\mathbf{x}}$ 为原问题可行解, $(\tilde{\boldsymbol{\lambda}}, \tilde{v})$ 为对偶问题可行解

证明思路: $g(\tilde{\boldsymbol{\lambda}}, \tilde{v}) = f_0(\tilde{\mathbf{x}})$

$L(\mathbf{x}, \tilde{\boldsymbol{\lambda}}, \tilde{v})$ 为 \mathbf{x} 的凸函数, 则 $\tilde{\mathbf{x}}$ 使 $L(\mathbf{x}, \tilde{\boldsymbol{\lambda}}, \tilde{v})$ 最小

$$\begin{aligned} g(\tilde{\boldsymbol{\lambda}}, \tilde{v}) &= \inf_{\mathbf{x}} L(\mathbf{x}, \tilde{\boldsymbol{\lambda}}, \tilde{v}) \\ &= L(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\lambda}}, \tilde{v}) \\ &= f_0(\tilde{\mathbf{x}}) + \sum_{i=1}^m \tilde{\lambda}_i f_i(\tilde{\mathbf{x}}) + \sum_{i=1}^p \tilde{v}_i h_i(\tilde{\mathbf{x}}) \\ &= f_0(\tilde{\mathbf{x}}) \end{aligned}$$

例 35 (Water-filling). 共 n 个信道(channel)

$$source \longleftrightarrow destination$$

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

分析. 拉格朗日函数, 注意 $\boldsymbol{\lambda}^T \mathbf{x}$ 项符号

$$L(\mathbf{x}, \boldsymbol{\lambda}, v) = -\sum_{i=1}^n \log(\alpha_i + x_i) - \boldsymbol{\lambda}^T \mathbf{x} + v(\mathbf{1}^T \mathbf{x} - 1)$$

KKT条件如下

- 原始可行性: $\mathbf{x}^* \succeq \mathbf{0}$
- 原始可行性: $\mathbf{1}^T \mathbf{x}^* = 1$
- 对偶可行性: $\boldsymbol{\lambda}^* \succeq \mathbf{0}$
- 对偶互斥条件: $\lambda_i^* x_i^* = 0, \forall i$
- 稳定性条件:

$$\begin{aligned} \left(\frac{\partial L(\mathbf{x}, \boldsymbol{\lambda}, v)}{\partial \mathbf{x}} \right)_i &= -\frac{1}{\alpha_i + x_i} - \lambda_i + v \\ &= -\frac{1}{\alpha_i + x_i^*} - \lambda_i^* + v^* = 0, \forall i \\ \implies v^* &= \frac{1}{\alpha_i + x_i^*} + \lambda_i^*, i = 1, \dots, n \\ \implies v^* &\geq \frac{1}{\alpha_i + x_i^*} \end{aligned}$$

或消去 λ_i , 由对偶互斥条件有

$$x_i^* \left(v^* - \frac{1}{\alpha_i + x_i^*} \right) = 0$$

若 $v^* \geq \frac{1}{\alpha_i} > \frac{1}{\alpha_i + x_i^*}$, 则

$$x_i^* = 0, \lambda_i = v - \frac{1}{\alpha_i}$$

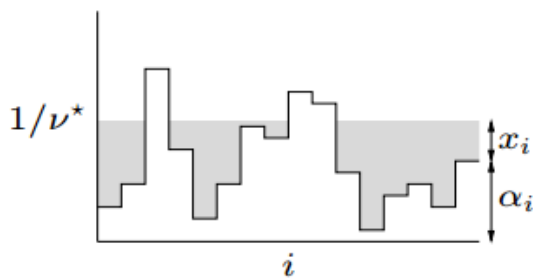
若 $\frac{1}{\alpha_i + x_i^*} \leq v^* < \frac{1}{\alpha_i}$, 则

$$\begin{aligned} x_i^* &> 0 \\ v^* &= \frac{1}{\alpha_i + x_i^*} \\ x_i^* &= \frac{1}{v^*} - \alpha_i \end{aligned}$$

综上有

$$x_i^* = \max \left\{ 0, \frac{1}{v^*} - \alpha_i \right\}$$

结合 $\sum_i x_i^* = 1$, 即注水算法



n 个块, 每块高度为 α_i , 用单位水量填充面积, 最终的高度为 $1/v^*$

定义 21 (扰动(perturbed)问题).

$$\begin{aligned} \min \quad & f_0(x) \\ \text{s.t.} \quad & f(x) \leq u_i, \quad i = 1, \dots, m \\ & h_i(x) = w_i, \quad i = 1, \dots, p \end{aligned}$$

新问题的最优解记为 $p^*(\mathbf{u}, \mathbf{w})$

定理 9. 若原始问题为凸, 则 $p^*(\mathbf{u}, \mathbf{w})$ 是 (u, w) 的凸函数

定理 10. 设 $(\lambda^*, \mathbf{v}^*)$ 为未被扰动的问题的对偶问题的最优解, 则

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

分析. 由强对偶性有

$$\begin{aligned} p^*(0,0) &= g(\boldsymbol{\lambda}^*, \mathbf{v}^*) \leq f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}) + \sum_{i=1}^p \mathbf{v}^* h_i(\mathbf{x}) \quad g(\boldsymbol{\lambda}^*, \mathbf{v}^*) \text{ 定义} \\ &\leq f_0(\mathbf{x}) + \boldsymbol{\lambda}^{*\top} \mathbf{u} + \mathbf{v}^{*\top} \mathbf{w} \quad \boldsymbol{\lambda}^* \succeq \mathbf{0} \end{aligned}$$

定理 11. 若原始问题为凸, 对偶间隙为 0, $p^*(\mathbf{u}, \mathbf{w})$ 在 $(0,0)$ 可微, 则

$$\lambda_i^* = -\frac{\partial p^*(0,0)}{\partial u_i} \quad \mathbf{v}_i^* = -\frac{\partial p^*(0,0)}{\partial w_i}$$

6 优化算法

6.1 简介

例 36.

$$\begin{aligned} \min \quad & f_0(\mathbf{x}) \\ \text{s.t.} \quad & A\mathbf{x} = \mathbf{b} \end{aligned}$$

分析. 1° 罚函数法 (penalty function)

$$\min F(\mathbf{x}) := f_0(\mathbf{x}) + \frac{\lambda}{2} \|A\mathbf{x} - \mathbf{b}\|_2^2$$

$$\tilde{\mathbf{x}} = \arg \min_{\mathbf{x}} F \implies \nabla f_0(\tilde{\mathbf{x}}) + \lambda A^\top (A\tilde{\mathbf{x}} - \mathbf{b}) = 0$$

2° 拉格朗日函数方法

$$\begin{aligned} L(\mathbf{x}, \mathbf{v}) &= f_0(\mathbf{x}) + \mathbf{v}^\top (A\mathbf{x} - \mathbf{b}) \\ \implies g(\mathbf{v}) &= \inf_{\mathbf{x}} (f_0(\mathbf{x}) + \mathbf{v}^\top (A\mathbf{x} - \mathbf{b})) \\ \mathbf{v} &= \lambda(A\tilde{\mathbf{x}} - \mathbf{b}) \\ \implies g(\lambda(A\tilde{\mathbf{x}} - \mathbf{b})) &= \inf_{\mathbf{x}} (f_0(\mathbf{x}) + \lambda(A\tilde{\mathbf{x}} - \mathbf{b})^\top (A\mathbf{x} - \mathbf{b})) \end{aligned}$$

例 37.

$$\begin{aligned} \min \quad & f_0(\mathbf{x}) \\ \text{s.t.} \quad & A\mathbf{x} \geq \mathbf{b} \end{aligned}$$

分析. log-barrier函数

$$\min f_0(\mathbf{x}) + \sum_{i=1}^m u_i \log(a_i^\top \mathbf{x} - b_i)$$

考虑这样的问题, $f_0(\mathbf{x})$ 可微, 凸, 无约束, 即

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

1. 所有算法都是迭代的

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha^{(k)} \mathbf{d}^{(k)}$$

$\alpha \geq 0$ 为步长, \mathbf{d} 为方向, 所有算法本质上都是选择方向与步长的问题

2. 如何选择步长 $\alpha^{(k)}$

$$\left\{ \begin{array}{l} \text{确定步长} \\ \text{搜索步长} \end{array} \right\} \left\{ \begin{array}{l} \text{固定步长} \\ \text{变化步长 (递减步长)} \end{array} \right.$$

最优步长: 线搜索问题, 给定当前点及方向

$$\alpha^{(k)} = \arg \min_{\alpha \geq 0} f_0(\mathbf{x}^{(k)} + \alpha \mathbf{d}^{(k)})$$

(a) 黄金分割法(0.618法)/优选法求解线搜索问题: 这样做的采样复杂度很低, 之前算过的点很容易被再用!

(b) 不精确线搜索(Armijo Rule): 一阶泰勒展开

Algorithm 1 不精确线搜索

```
1:  $\alpha^{(k)} = \alpha_{\max}$ 
2: if  $f_0(\mathbf{x}^{(k)} + \alpha^{(k)} \mathbf{d}^{(k)}) > f_0(\mathbf{x}^{(k)}) + \mu \alpha^{(k)} \langle \nabla f_0(\mathbf{x}^{(k)}), \mathbf{d}^{(k)} \rangle$  then
3:    $\alpha^{(k)} \leftarrow \alpha^{(k)} \beta, \beta \in (0, 1)$ 
4: else
5:   Stop
```

而实际上没有要求最优步长, 在该方向上的差异并没有太大

3. 关键问题是选方向

6.2 梯度下降法

取 $\mathbf{d}^{(k)} = -\nabla f_0(\mathbf{x}^{(k)})$, 则迭代格式为

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha^{(k)} \nabla f(\mathbf{x}^{(k)})$$

关注下面几个问题

- 能否收敛
- 收敛到哪里
- 收敛速度

6.2.1 前提假设

0. 基本假设: f 为可微的凸函数,

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

存在且有限, $f_0(\mathbf{x}^*)$ 有限

1. Lipschitz 连续梯度

$$\exists L \geq 0, \forall \mathbf{x}, \mathbf{y}: \|\nabla f_0(\mathbf{x}) - \nabla f_0(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\|$$

等价定义:

a. 若 $f_0(\mathbf{x})$ 二阶可微

$$\nabla^2 f_0(\mathbf{x}) \preceq LI, \forall \mathbf{x}$$

b. 下界

$$\langle \nabla f_0(\mathbf{x}) - \nabla f_0(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{1}{L} \|\nabla f_0(\mathbf{x}) - \nabla f_0(\mathbf{y})\|^2$$

c. 上界

$$\langle \nabla f_0(\mathbf{x}) - \nabla f_0(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \leq L \|\mathbf{x} - \mathbf{y}\|^2$$

d. 当函数为凸时

$$0 \leq f_0(\mathbf{y}) - f_0(\mathbf{x}) - \langle \nabla f_0(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2$$

2. 强凸性(strong convexity): 即加了正则化项

$$\exists \mu > 0, \forall \mathbf{x}, \mathbf{y}: f_0(\mathbf{y}) \geq f_0(\mathbf{x}) + \langle \nabla f_0(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{x} - \mathbf{y}\|^2$$

二阶可微情况下的等价定义

$$\nabla^2 f_0(\mathbf{x}) \succeq \mu I$$

例 38. 判断下列函数是否符合 Lipschitz 连续梯度及强凸性条件

$$\begin{array}{lll} f_0(\mathbf{x}) = \mathbf{1}^T \mathbf{x} & L = 0 & \times \\ f_0(\mathbf{x}) = \frac{1}{2} \|\mathbf{x}\|_2^2 & L = 1 & \mu = 1 \\ f_0(\mathbf{x}) = \frac{1}{4} \|\mathbf{x}\|_2^4 & \times & \times \end{array}$$

区别于严格凸(strictly convex), 强凸一定是严格凸

定理 12. 严格凸函数只有一个最小值点

分析. 反证法, 假设 \mathbf{x}, \mathbf{y} 均为最小值点, 且 $\mathbf{x} \neq \mathbf{y}$

$$f_0(\mathbf{y}) > f_0(\mathbf{x}) + \langle \nabla f_0(\mathbf{x}), \mathbf{x} - \mathbf{y} \rangle = f_0(\mathbf{x})$$

定理 13. 若 $f_0(\mathbf{x})$ 有 Lipschitz 连续梯度, 常数 L , 若 $\alpha \in (0, \frac{2}{L})$, 则有

$$f_0(\mathbf{x}^{(k)}) - f_0(\mathbf{x}^*) \leq \frac{2(f_0(\mathbf{x}^0) - f_0(\mathbf{x}^*)) \|\mathbf{x}^0 - \mathbf{x}^*\|^2}{2\|\mathbf{x}^0 - \mathbf{x}^*\|^2 + k\alpha(2 - L\alpha)(f_0(\mathbf{x}^0) - f_0(\mathbf{x}^*))}, \forall \mathbf{x}^*$$

即以 $O(\frac{1}{k})$ 速度收敛

分析. 1° 点的单调性: 与任意 \mathbf{x}^* 的距离在缩小

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\|^2 \leq \|\mathbf{x}^{(k)} - \mathbf{x}^*\|^2, \forall \mathbf{x}^*$$

$$\begin{aligned} LHS &= \|\mathbf{x}^{(k)} - \mathbf{x}^* - \alpha \nabla f_0(\mathbf{x}^{(k)})\|^2 \\ &= \|\mathbf{x}^{(k)} - \mathbf{x}^*\|^2 - 2\alpha \langle \mathbf{x}^{(k)} - \mathbf{x}^*, \nabla f_0(\mathbf{x}^{(k)}) \rangle + \alpha^2 \|\nabla f_0(\mathbf{x}^{(k)})\|^2 \\ &\leq \|\mathbf{x}^{(k)} - \mathbf{x}^*\|^2 + \alpha(\alpha - \frac{2}{L}) \|\nabla f_0(\mathbf{x}^{(k)})\|^2 \quad \text{注意到 } \nabla f_0(\mathbf{x}^*) = 0, \text{ 利用 Lipschitz 连续梯度} \\ &\leq \|\mathbf{x}^{(k)} - \mathbf{x}^*\|^2 \end{aligned}$$

2° 函数值的单调性: $f_0(\mathbf{x}^{(k+1)}) \leq f_0(\mathbf{x}^{(k)})$ (注意下降可能非常缓慢, 并不一定收敛)

$$\begin{aligned} f_0(\mathbf{x}^{(k+1)}) &\leq f_0(\mathbf{x}^{(k)}) + \langle \nabla f_0(\mathbf{x}^{(k)}), \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} \rangle + \frac{L}{2} \|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|^2 \\ &= f_0(\mathbf{x}^{(k)}) - \alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla f_0(\mathbf{x}^{(k)})\|^2 \\ &\leq f_0(\mathbf{x}^{(k)}) \end{aligned}$$

3° 函数值的充分下降 (即证明收敛性)

$$f_0(\mathbf{x}^{(k+1)}) - f_0(\mathbf{x}^*) \leq f_0(\mathbf{x}^{(k)}) - f_0(\mathbf{x}^*) - \omega \|\nabla f_0(\mathbf{x}^{(k)})\|^2$$

$$\begin{aligned} f_0(\mathbf{x}^{(k)}) - f_0(\mathbf{x}^*) &\leq \langle f_0(\mathbf{x}^{(k)}), \mathbf{x}^{(k)} - \mathbf{x}^* \rangle \\ &= \langle \nabla f_0(\mathbf{x}^{(k)}) - \nabla f_0(\mathbf{x}^*), \mathbf{x}^{(k)} - \mathbf{x}^* \rangle \\ &\leq \|\nabla f_0(\mathbf{x}^{(k)}) - \nabla f_0(\mathbf{x}^*)\| \|\mathbf{x}^{(k)} - \mathbf{x}^*\| \quad \text{Cauchy} \\ &\leq \|\nabla f_0(\mathbf{x}^{(k)})\| \|\mathbf{x}^{(k)} - \mathbf{x}^*\| \end{aligned}$$

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

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

错位相消可得结论 $O(\frac{1}{k})$ 收敛速度

定理 14. 若 f_0 有 *Lipschitz* 连续梯度, 常数 L , 强凸函数 n , 步长 $\alpha \in (0, \frac{2}{\mu+L}]$, 则

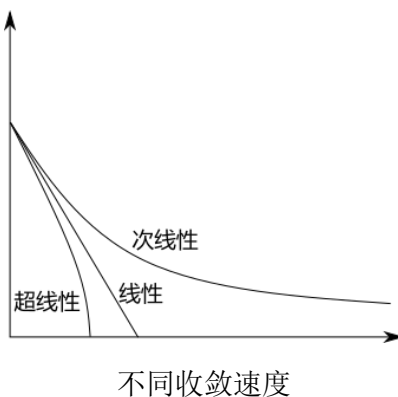
$$\|\mathbf{x}^{(k)} - \mathbf{x}^*\|^2 \leq \left(1 - \frac{2\alpha\mu L}{\mu + L}\right)^k \|\mathbf{x}^{(0)} - \mathbf{x}^*\|^2$$

分析.

$$\begin{aligned} \|\mathbf{x}^{(k)} - \mathbf{x}^*\|^2 &= \|\mathbf{x}^{(k)} - \alpha \nabla f_0(\mathbf{x}^{(k)}) - \mathbf{x}^*\|^2 \\ &= \|\mathbf{x}^{(k)} - \mathbf{x}^*\|^2 - 2\alpha \langle \mathbf{x}^{(k)} - \mathbf{x}^*, \nabla f_0(\mathbf{x}^{(k)}) \rangle + \alpha^2 \|\nabla f_0(\mathbf{x}^{(k)})\|^2 \\ &\leq \|\mathbf{x}^{(k)} - \mathbf{x}^*\|^2 - \frac{2\alpha}{\mu + L} \|\nabla f_0(\mathbf{x})\|^2 + \alpha^2 \|\nabla f_0(\mathbf{x}^{(k)})\|^2 \quad \text{内积不等式} \\ &\leq RHS \end{aligned}$$

$$1 - \frac{4\mu L}{(\mu + L)^2} = \frac{(L - \mu)^2}{(L + \mu)^2} = \frac{\left(\frac{L}{\mu} - 1\right)^2}{\left(\frac{L}{\mu} + 1\right)^2}$$

L 为 *Hessian* 矩阵的最大特征值, μ 为 *Hessian* 矩阵的最小特征值, 则 $\frac{L}{\mu}$ 为该矩阵的条件数



例 39.

$$f_0(\mathbf{x}) = \frac{1}{2} \|\mathbf{Ax} - \mathbf{b}\|_2^2$$

分析.

$$\mathbf{x}^{(0)} \rightarrow \mathbf{x}^{(1)}$$

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - \alpha(\mathbf{x}^{(0)} - \mathbf{b}) = \mathbf{b}$$

条件数糟糕的病态矩阵收敛速度是非常糟糕的, 会出现 zig-zag 的情况, 如下面的矩阵

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 10^{-4} \end{bmatrix}$$

可以通过预处理(precondition)来解决条件数糟糕的问题

$f_0(\mathbf{x})$, Lipschitz连续梯度(L), 强凸(μ), 考虑函数值收敛性

$$\begin{aligned}\tilde{f}_0(\alpha^{(k)}) &= f_0(\mathbf{x}^{(k+1)}) = f_0(\mathbf{x}^{(k)} - \alpha^{(k)} \nabla f_0(\mathbf{x}^{(k)})) \\ &\leq f_0(\mathbf{x}^{(k)}) + \langle \nabla f_0(\mathbf{x}^{(k)}), -\alpha^{(k)} \nabla f_0(\mathbf{x}^{(k)}) \rangle + \frac{L}{2} \left\| -\alpha^{(k)} \nabla f_0(\mathbf{x}^{(k)}) \right\|^2 \\ &= f_0(\mathbf{x}^{(k)}) + \frac{L(\alpha^{(k)})^2 - 2\alpha^{(k)}}{2} \left\| \nabla f_0(\mathbf{x}^{(k)}) \right\|^2\end{aligned}$$

- $\alpha^{(k)} = \alpha_{exact}^{(k)}$ 精确线搜索

$$\begin{aligned}\tilde{f}_0(\alpha_{exact}^{(k)}) &\leq \tilde{f}_0\left(\frac{1}{L}\right) = f_0(\mathbf{x}^{(k)}) - \frac{1}{2L} \left\| \nabla f_0(\mathbf{x}^{(k)}) \right\|^2 \\ \implies \tilde{f}_0(\alpha_{exact}^{(k)}) - f_0(\mathbf{x}^{(k)}) - f_0(\mathbf{x}^*) &- \frac{1}{2L} \left\| \nabla f_0(\mathbf{x}^{(k)}) \right\|^2 \leq \left(1 - \frac{\mu}{L}\right) (f_0(\mathbf{x}^{(k)}) + f_0(\mathbf{x}^*))\end{aligned}$$

$$\begin{aligned}f_0(\mathbf{x}^*) &\geq f_0(\mathbf{x}^{(k)}) + \langle \nabla f_0(\mathbf{x}^{(k)}), \mathbf{x}^* - \mathbf{x}^{(k)} \rangle + \frac{\mu}{2} \left\| \mathbf{x}^{(k)} - \mathbf{x}^* \right\|^2 \\ &\geq f_0(\mathbf{x}^{(k)}) - \frac{\mu}{2} \left\| \mathbf{x}^{(k)} - \mathbf{x}^* \right\|^2 - \frac{1}{2\mu} \left\| \nabla f_0(\mathbf{x}^{(k)}) \right\|^2 + \frac{\mu}{2} \left\| \mathbf{x}^{(k)} - \mathbf{x}^* \right\|^2 \quad ab \geq -\frac{\mu}{2}a^2 - \frac{1}{2\mu}b^2 \\ &= f_0(\mathbf{x}^{(k)}) - \frac{1}{2\mu} \left\| \nabla f_0(\mathbf{x}^{(k)}) \right\|^2\end{aligned}$$

$$f_0(\mathbf{x}^{(k)}) - f_0(\mathbf{x}^*) \leq \frac{1}{2\mu} \left\| \nabla f_0(\mathbf{x}^{(k)}) \right\|^2$$

- Armijo Rule

$$\begin{aligned}\tilde{f}_0(\alpha^{(k)}) &= f_0(\mathbf{x}^{(k+1)}) \leq f_0(\mathbf{x}^{(k)}) + \frac{L(\alpha^{(k)})^2 - 2\alpha^{(k)}}{2} \left\| \nabla f_0(\mathbf{x}^{(k)}) \right\|^2 \\ \tilde{f}_0(\alpha^{(k)}) &= f_0(\mathbf{x}^{(k+1)}) \leq f_0(\mathbf{x}^{(k)}) - \nabla \alpha^{(k)} \left\| \nabla f_0(\mathbf{x}^{(k)}) \right\|^2\end{aligned}$$

首先说明, 若 $0 \leq \alpha^{(k)} \leq \frac{1}{L}$ 时,

$$\tilde{f}_0(\alpha^{(k)}) \leq f_0(\mathbf{x}^{(k)}) - \gamma \alpha^{(k)} \left\| \nabla f_0(\mathbf{x}^{(k)}) \right\|^2$$

当 $\alpha^{(k)} \in [0, \frac{1}{2}]$ 时,

$$-\alpha^{(k)} + \frac{L}{2}(\alpha^{(k)})^2 \leq -\frac{\alpha^{(k)}}{2} \iff \frac{L}{2}(\alpha^{(k)})^2 \leq \frac{\alpha^{(k)}}{2} \iff L \cdot \alpha^{(k)} \leq 1$$

$$\begin{aligned}f_0(\mathbf{x}^{(k+1)}) &\leq f_0(\mathbf{x}^{(k)}) + \frac{L(\alpha^{(k)})^2 - 2\alpha^{(k)}}{2} \left\| \nabla f_0(\mathbf{x}^{(k)}) \right\|^2 \\ &\leq f_0(\mathbf{x}^{(k)}) - \frac{\alpha^{(k)}}{2} \left\| \nabla f_0(\mathbf{x}^{(k)}) \right\|^2 \\ &\leq f_0(\mathbf{x}^{(k)}) - \gamma \alpha^{(k)} \left\| \nabla f_0(\mathbf{x}^{(k)}) \right\|^2\end{aligned}$$

$$\begin{aligned}f_0(\mathbf{x}^{(k+1)}) &\leq f_0(\mathbf{x}^{(k)}) - \min \left\{ \gamma \alpha_{\max}, \frac{\gamma \beta}{L} \right\} \left\| \nabla f_0(\mathbf{x}^{(k)}) \right\|^2 \\ \implies f_0(\mathbf{x}^{(k+1)}) - f_0(\mathbf{x}^*) &\leq \left(1 - \min \left\{ 2\mu \gamma \alpha_{\max}, \frac{2\mu \gamma \beta}{L} \right\}\right) (f_0(\mathbf{x}^{(k)}) - f_0(\mathbf{x}^*))\end{aligned}$$

6.2.2 梯度下降法的解释

- 解释一

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha^{(k)} \nabla f_0(\mathbf{x}^{(k)})$$

将 f_0 在 $\mathbf{x}^{(k)}$ 处进行一阶Taylor展开

$$f_0(\mathbf{x}) \approx f_0(\mathbf{x}^{(k)}) + \langle \nabla f_0(\mathbf{x}^{(k)}), \mathbf{x} - \mathbf{x}^{(k)} \rangle + \frac{1}{2\alpha^{(k)}} \|\mathbf{x} - \mathbf{x}^{(k)}\|^2$$

求梯度

$$\begin{aligned} \nabla f_0(\mathbf{x}^{(k)}) + \frac{1}{\alpha^{(k)}}(\mathbf{x} - \mathbf{x}^{(k)}) &= 0 \\ \alpha^{(k)} \nabla f_0(\mathbf{x}^{(k)}) + \mathbf{x} - \mathbf{x}^{(k)} &= 0 \\ \mathbf{x} &= \mathbf{x}^{(k)} - \alpha^{(k)} \nabla f_0(\mathbf{x}^{(k)}) \end{aligned}$$

- 解释二

$$f_0(\mathbf{x}^{(k)} + \mathbf{v}) \approx f_0(\mathbf{x}^{(k)}) + \langle \nabla f_0(\mathbf{x}^{(k)}), \mathbf{v} \rangle$$

$$\mathbf{d}^{(k)} = \arg \min_{\mathbf{v}} \{ \langle \nabla f_0(\mathbf{x}^{(k)}), \mathbf{v} \rangle \mid \|\mathbf{v}\| = 1 \}$$

若采用2-范数，可得标准化的负梯度方向(normalized negative gradient)

$$\mathbf{d}^{(k)} = \frac{-\nabla f_0(\mathbf{x}^{(k)})}{\|\nabla f_0(\mathbf{x}^{(k)})\|_2}$$

通过改变不同的范数，有不同的特性

进而有坐标下降法/交替极小化(coordinate descent/alternating direction)

$$\mathbf{d}^{(k)} = \mathbf{e}_{k \bmod n}$$

注意，这里 $\mathbf{x} \in \mathbb{R}^n$, $n \bmod n = 0$

$$\alpha^{(k)} = \arg \min_{\alpha} f_0(\mathbf{x}^{(k)} + \alpha \mathbf{d}^{(k)}), \alpha_{\max} \geq \alpha \geq \alpha_{\min}$$

6.3 非光滑优化问题

6.3.1 次梯度法

同样考虑 f_0 连续，凸，不可微

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

梯度下降法→次梯度(subgradient)法

定义 22 (次梯度). 若 $g_0(\mathbf{x}) \in \partial f_0(\mathbf{x})$ (注意凹函数则对应的是 \supgradient) 为 $f_0(\mathbf{x})$ 的一个次梯度，则

$$\forall \mathbf{y} : f_0(\mathbf{y}) \geq f_0(\mathbf{x}) + \langle g_0(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$$

即过该点的直线都要在整条曲线的下方，则该直线的斜率范围为次梯度取值

如 $f(\mathbf{x}) = |\mathbf{x}|$ 在零点处次梯度为 $[-1, 1]$ 。

次梯度法¹⁰迭代格式如下

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha^{(k)} g_0(\mathbf{x}^{(k)})$$

只要有 $0 \in \partial f_0(\mathbf{x}_0)$ 就有最优解 $\mathbf{x} = \mathbf{x}_0$

对于梯度法来说，关键在于选择步长

- 固定步长 $\alpha^{(k)} = \alpha$
- 不可加但平方可加，如 $\frac{1}{k}$

$$\sum_{k=0}^{\infty} \alpha^{(k)} = \infty \quad \sum_{k=0}^{\infty} (\alpha^{(k)})^2 < \infty$$

- 不可加递减，如 $\frac{1}{\sqrt{k}}$

$$\sum_{k=0}^{\infty} \alpha^{(k)} = 0 \quad \lim_{k \rightarrow \infty} \alpha^{(k)} \rightarrow 0$$

考虑次梯度法的收敛速度

$$\inf_{i=0, \dots, k} (f_0(\mathbf{x}^{(i)}) - f_0(\mathbf{x}^*))$$

假设函数Lipschitz连续

$$\exists G > 0, \forall \mathbf{x}, \mathbf{y} : \|f_0(\mathbf{x}) - f_0(\mathbf{y})\| \leq G \|\mathbf{x} - \mathbf{y}\|$$

对任意最优解 \mathbf{x}^* ，有

$$\begin{aligned} & \|\mathbf{x}^{(k+1)} - \mathbf{x}^*\|^2 = \|\mathbf{x}^{(k)} - \alpha^{(k)} g_0(\mathbf{x}^{(k)}) - \mathbf{x}^*\|^2 \\ & = \|\mathbf{x}^{(k)} - \mathbf{x}^*\|^2 - 2\langle \alpha^{(k)} g_0(\mathbf{x}^{(k)}), \mathbf{x}^{(k)} - \mathbf{x}^* \rangle + (\alpha^{(k)})^2 \|g_0(\mathbf{x}^{(k)})\|^2 \quad \text{拆平方} \\ & \leq \|\mathbf{x}^{(k)} - \mathbf{x}^*\|^2 - 2(\alpha^{(k)})^2 (f_0(\mathbf{x}^{(k)}) - f_0(\mathbf{x}^*)) + (\alpha^{(k)})^2 G^2 \quad \text{多次迭代} \\ \Rightarrow & \|\mathbf{x}^{(k+1)} - \mathbf{x}^*\|^2 \leq \|\mathbf{x}^{(0)} - \mathbf{x}^*\|^2 - 2 \sum_{i=0}^k \alpha^{(i)} (f_0(\mathbf{x}^{(i)}) - f_0(\mathbf{x}^*)) + \sum_{i=0}^k (\alpha^{(i)})^2 G^2 \\ \Rightarrow & 2 \sum_{i=0}^k \alpha^{(i)} (f_0(\mathbf{x}^{(i)}) - f_0(\mathbf{x}^*)) \leq \|\mathbf{x}^{(0)} - \mathbf{x}^*\|^2 + \sum_{i=0}^k (\alpha^{(i)})^2 G^2 \\ & \sum_{i=0}^k \alpha^{(i)} (f_0(\mathbf{x}^{(i)}) - f_0(\mathbf{x}^*)) \geq \left(\sum_{i=0}^k \alpha^{(i)} \right) \inf_{i=0, \dots, k} (f_0(\mathbf{x}^{(i)}) - f_0(\mathbf{x}^*)) \\ \Rightarrow & \inf_{i=0, \dots, k} (f_0(\mathbf{x}^{(i)}) - f_0(\mathbf{x}^*)) \leq \frac{\|\mathbf{x}^{(0)} - \mathbf{x}^*\|^2 + G^2 \sum_{i=0}^k (\alpha^{(i)})^2}{2 \sum_{i=0}^k \alpha^{(i)}} \end{aligned}$$

¹⁰如果激活函数为非光滑的（如ReLU），那么出来的函数也是非光滑的，就要用次梯度

这是一个紧的界

- 固定步长得到上界 $\frac{G^2\alpha}{2}$, 以 $f(\mathbf{x}) = |\mathbf{x}|$ 为例
- 不可加平方可加一定收敛, 若步长为 $\frac{1}{k}$, 收敛速度为 $\frac{1}{\log k}$ (幂级数积分取上下界 $\sum_{i=0}^k \frac{1}{i} = O(\log k)$)
- 不可加平方不可加同样收敛, 若步长为 $\frac{1}{\sqrt{k}}$, 收敛速度为 $O\left(\frac{\log k}{\sqrt{k}}\right)$, 可以证明在该假设下该收敛速度最优

$$\lim_{k \rightarrow \infty} \inf_{i=0, \dots, k} (f_0(\mathbf{x}^{(i)}) - f_0(\mathbf{x}^*)) = 0$$

$$\forall \varepsilon > 0, \exists N_1 \in \mathbb{Z}, \forall i > N_1 : \alpha^{(i)} \leq \frac{\varepsilon}{G^2}$$

$$\exists N_2 \in \mathbb{Z}, \forall k > N_2 : \sum_{i=0}^k \alpha^{(i)} \geq \frac{1}{\varepsilon} \left(\|\mathbf{x}^{(0)} - \mathbf{x}^*\|^2 + G^2 \sum_{i=0}^{N_1} (\alpha^{(i)})^2 \right)$$

令 $N = \max\{N_1, N_2\}, \forall k > N$

$$\frac{\|\mathbf{x}^{(0)} - \mathbf{x}^*\|^2 + G^2 \sum_{i=0}^{N_1} (\alpha^{(i)})^2}{2 \sum_{i=0}^k \alpha^{(i)}} + \frac{\|\mathbf{x}^{(0)} - \mathbf{x}^*\|^2 + G^2 \sum_{i=N_1+1}^k (\alpha^{(i)})^2}{2 \sum_{i=0}^{N_1} \alpha^{(i)} + 2 \sum_{i=N_1+1}^k \alpha^{(i)}} \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

$$\text{第二项} \leq \frac{G^2 \sum_{i=N_1+1}^k \alpha^{(i)} \frac{\varepsilon}{G^2}}{2 \sum_{i=0}^{N_1} \alpha^{(i)} + 2 \sum_{i=N_1+1}^k \alpha^{(i)}} \leq \frac{\varepsilon}{2}$$

实际上这个假设一般情况下不成立, 但是我们只需保证在优化路径上成立即可, 也有设置 \mathbf{x} 有界的

6.3.2 邻近点梯度法(proximal gradient method)

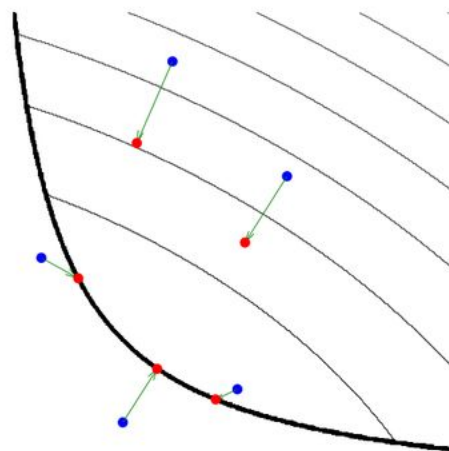
考虑有结构, 不可微的函数

$$\min f_0(\mathbf{x}) = s(\mathbf{x}) + r(\mathbf{x})$$

- s : smooth, 可微, 易求导
- r : regularization, 不可微, 易求邻近点投影

定义 23 (邻近点投影(proximal mapping)). 函数 r 和平衡参数 α 的近端算子

$$\text{prox}_r \hat{\mathbf{x}} = \arg \min_{\mathbf{x}} \left(r(\mathbf{x}) + \frac{1}{2\alpha} \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 \right)$$



注：邻近点/近端投影字面上理解即对点 $\hat{\mathbf{x}}$ 关于 r 下降方向的邻近点。如上图，黑细线为 r 的水平集，黑粗线为定义域边界，蓝色点为 $\hat{\mathbf{x}}$ ，红色为作用该算子后的点，这些点都统一朝着函数最小的方法前进¹¹。

邻近点梯度法迭代格式如下

$$\begin{cases} \mathbf{x}^{(k+\frac{1}{2})} = \mathbf{x}^{(k)} - \alpha^{(k)} \nabla s(\mathbf{x}^{(k)}) \\ \mathbf{x}^{(k+1)} = \text{prox}_r \mathbf{x}^{(k+\frac{1}{2})} = \arg \min_{\mathbf{x}} \left(r(\mathbf{x}) + \frac{1}{2\alpha} \|\mathbf{x} - \mathbf{x}^{(k+\frac{1}{2})}\|_2^2 \right) \end{cases}$$

例 40 (LASSO).

$$f(\mathbf{x}) = \frac{1}{2} \|A\mathbf{x} - \mathbf{b}\|_2^2 + p \|\mathbf{x}\|_1$$

分析. 设

$$\begin{cases} s(\mathbf{x}) := \frac{1}{2} \|A\mathbf{x} - \mathbf{b}\|_2^2 \\ r(\mathbf{x}) := p \|\mathbf{x}\|_1 \end{cases}$$

其中, $A \in \mathbb{R}^{m \times n}$, $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{b}, \mathbf{e} \in \mathbb{R}^m$ (在本题中 $m = 50, n = 100$), $s(\mathbf{x})$ 为光滑函数, $r(\mathbf{x})$ 为非光滑函数, 则原式

$$f(\mathbf{x}) = s(\mathbf{x}) + r(\mathbf{x})$$

先求 $r(\mathbf{x})$ 的邻近点投影

$$\text{prox} \hat{\mathbf{x}} = \arg \min_{\mathbf{x}} \left(p \|\mathbf{x}\|_1 + \frac{1}{2\alpha} \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 \right)$$

对上式右侧展开有

$$\arg \min_{\mathbf{x}} \left(p \sum_{i=1}^n |x_i| + \frac{1}{2\alpha} \sum_{i=1}^n (x_i - \hat{x}_i)^2 \right)$$

注意到上式对于下标 i 相互独立, 故要求上式的最小值, 等价于对每一个下标 i 求最小值后求和, 即

$$\arg \min_{x_i} \left(p|x_i| + \frac{1}{2\alpha} (x_i - \hat{x}_i)^2 \right), \forall i$$

¹¹详细见<https://zhuanlan.zhihu.com/p/37444622>

由不可微函数的极值判断条件有

$$0 \in \left(\partial_{x_i} p |x_i| + \frac{1}{\alpha} (x_i - \hat{x}_i) \right), \forall i$$

对每一个 x_i 进行分类讨论

- 若 $x_i > 0$, 则 $|x_i|$ 对于 $|x_i|$ 可微, 即 $\partial_{x_i} |x_i| = 1$, 有

$$p + \frac{1}{\alpha} (x_i - \hat{x}_i) = 0$$

整理得

$$x_i = \hat{x}_i - \alpha p$$

由于 $x_i > 0$, 故 $\hat{x}_i - \alpha p > 0$, 即 $\hat{x}_i > \alpha p$

- 若 $x_i < 0$, 则 $|x_i|$ 对于 $|x_i|$ 可微, 即 $\partial_{x_i} |x_i| = -1$, 有

$$-p + \frac{1}{\alpha} (x_i - \hat{x}_i) = 0$$

整理得

$$x_i = \hat{x}_i + \alpha p$$

由于 $x_i < 0$, 故 $\hat{x}_i + \alpha p < 0$, 即 $\hat{x}_i < -\alpha p$

- 若 $x_i = 0$, 则 $|x_i|$ 对于 $|x_i|$ 不可微, 需要求次梯度, $\partial_{x_i} |x_i| = [-1, 1]$, 即

$$0 \in \left[-p - \frac{\hat{x}_i}{\alpha}, p - \frac{\hat{x}_i}{\alpha} \right]$$

那么, 需要满足

$$\begin{cases} p - \frac{\hat{x}_i}{\alpha} \geq 0 \\ -p - \frac{\hat{x}_i}{\alpha} \leq 0 \end{cases}$$

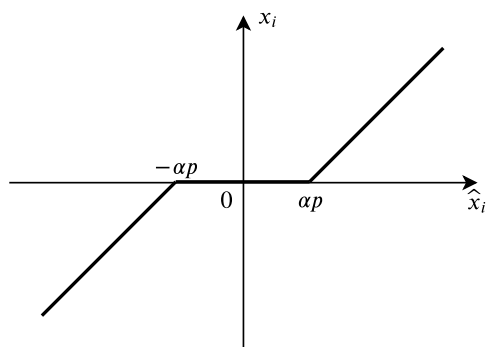
推得

$$\hat{x}_i \in [-\alpha p, \alpha p]$$

综上, 有

$$x_i = \begin{cases} \hat{x}_i + \alpha p & \hat{x}_i < -\alpha p \\ 0 & \hat{x}_i \in [-\alpha p, \alpha p] \\ \hat{x}_i - \alpha p & \hat{x}_i > \alpha p \end{cases}$$

可以得到下图的软门限 (*soft-thresholding*) 曲线



进一步，得到邻近点梯度下降法的迭代式如下

$$\begin{cases} \mathbf{x}^{(k+\frac{1}{2})} = \mathbf{x}^{(k)} - \alpha \nabla s(\mathbf{x}^{(k)}) = \mathbf{x}^{(k)} - \alpha A^T(A\mathbf{x}^{(k)} - \mathbf{b}) \\ \mathbf{x}^{(k+1)} = \text{prox } \mathbf{x}^{(k+\frac{1}{2})} = \arg \min_{\mathbf{x}} \left(p \|\mathbf{x}\|_1 + \frac{1}{2\alpha} \|\mathbf{x} - \mathbf{x}^{(k+\frac{1}{2})}\|_2^2 \right) \end{cases}$$

其中， $x^{(k+\frac{1}{2})}$ 可直接计算， $x^{(k+1)}$ 的显式解可由软门限求得。

此算法即迭代阈值收缩 (*Iterative shrinkage-thresholding algorithm, ISTA*) 算法

例 41 (盒限制(box constrained)优化问题).

$$f_0(\mathbf{x}) = s(\mathbf{x}) + \sum_{i=1}^n I(x_i \in [l_i, u_i])$$

分析. 构造邻近点投影

$$\begin{aligned} \arg \min \quad & \frac{1}{2\alpha} \|\mathbf{x} - \hat{\mathbf{x}}\|^2 \\ \text{subject to} \quad & x_i \in [l_i, u_i], \forall i \end{aligned}$$

如果有约束 $\mathbf{x} \in \mathcal{C}$

$$\begin{cases} \mathbf{x}^{(k+\frac{1}{2})} = \mathbf{x}^{(k)} - \alpha \nabla S(\mathbf{x}^{(k)}) \\ \mathbf{x}^{(k+1)} = \arg \min_{\mathbf{x}} \left(I_{\mathcal{C}}(\mathbf{x}^{(k+\frac{1}{2})}) + \frac{1}{2\alpha} \|\mathbf{x} - \mathbf{x}^{(k+\frac{1}{2})}\|^2 \right) = \arg \min_{\mathbf{x}} \frac{1}{2\alpha} \|\mathbf{x} - \mathbf{x}^{(k+\frac{1}{2})}\|^2, \mathbf{x} \in \mathcal{C} \end{cases}$$

相当于做投影，故称投影梯度法

$$\begin{aligned} 0 &\in \partial r(\mathbf{x}^{(k+1)}) + \frac{1}{\alpha}(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k+\frac{1}{2})}) \\ 0 &= \partial r(\mathbf{x}^{(k+1)}) + \frac{1}{\alpha}(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} + \alpha \nabla S(\mathbf{x}^{(k)})) \\ \mathbf{x}^{(k+1)} &= \mathbf{x}^{(k)} - \alpha \nabla S(\mathbf{x}^{(k)}) - \alpha \partial r(\mathbf{x}^{(k+1)}) \end{aligned}$$

数值计算：显式方法（次梯度法）→隐式方法（邻近点梯度法—需要先知道下一步信息，但是这是可以做的，因为有邻近点）

* 邻近点投影法的收敛性能与梯度下降法类似

例 42 (矩阵补全). $Y \in \mathbb{R}^{m \times n}, \{Y_{ij}, (i, j) \in \Omega\}$

$$\min_B \left(\frac{1}{2} \sum_{(i,j) \in \Omega} (Y_{ij} - B_{ij})^2 + \lambda \text{rank}(B) \right)$$

分析. 同 *LASSO*, 由于矩阵的秩 (奇异值向量 l_1 -范数) 不好求, 改为矩阵的和范数 $\|\cdot\|_*$ (奇异值向量 l_1 -范数), 即

$$\min_B \left(\frac{1}{2} \sum_{(i,j) \in \Omega} (Y_{ij} - B_{ij})^2 + \lambda \|B\|_* \right)$$

其中,

$$\|B\|_* := \sum_{i=1}^n \sigma_i(B)$$

$$\min \left(\frac{1}{2} \|P_\Omega(B - Y)\|_F^2 + \lambda \|B\|_* \right)$$

若原矩阵中该项不存在, P_Ω 为 0; 存在的话则保持不变

$$\nabla B \left(\frac{1}{2} \|P_\Omega(B - Y)\|_F^2 \right) = P_\Omega(B - Y)$$

对 B 做奇异值分解, U 为酉矩阵

$$\partial \|B\|_* = \{UDV^T, B = U\Sigma V^T, d = \partial \|\sigma\|_1\}$$

邻近点梯度迭代格式为

$$\begin{cases} B^{(k+\frac{1}{2})} = B^k - \alpha P_\Omega(B^k - Y) \\ B^{(k+1)} = \arg \min_B \lambda \|B\|_* + \frac{1}{2\alpha} \|B - B^{(k+\frac{1}{2})}\|_F^2 \end{cases}$$

求次梯度

$$0 \in \lambda \partial \|B\|_* + \frac{1}{\alpha} (B - B^{(k+\frac{1}{2})})$$

$$0 \in \left[\lambda UDV^T + \frac{1}{\alpha} (B - B^{(k+\frac{1}{2})}) \right], B = U\Sigma V^T, D = \partial \|\sigma\|_1$$

$$0 \in \left\{ \lambda UDV^T + \frac{1}{\alpha} (V\Sigma V^T - B^{(k+\frac{1}{2})}) \right\}$$

$$\exists V : 0 = \alpha \lambda UDV^T + V\Sigma V^T - B^{(k+\frac{1}{2})}$$

$$B^{(k+\frac{1}{2})} = U(\alpha \lambda D + \Sigma)V^T$$

对 $B^{(k+\frac{1}{2})}$ 进行奇异值分解

$$B^{(k+\frac{1}{2})} = U\Sigma^{(k+\frac{1}{2})}V$$

$$T_i = \alpha \lambda d_i + \sigma_i$$

若 $\sigma_i \neq 0 \implies \tau_i = \alpha \lambda + \sigma_i$

若 $\sigma_i = 0 \implies \tau_i \in [-\alpha \lambda + \sigma_i, \alpha \lambda + \sigma_i]$

$$\begin{cases} \sigma_i = \tau_i - \alpha \lambda & \tau_i > \alpha \lambda \\ \sigma_i = 0 & \tau_i \leq \alpha \lambda \end{cases}$$

该算法称为矩阵软门限算法

一阶方法总结:

- 梯度下降法
- 次梯度法: 在随机/不确定性优化问题中很有效
- 邻近点梯度法

6.4 二阶优化方法

6.4.1 牛顿法

牛顿法(Newton's method): 要求 $f_0(\mathbf{x})$ 二阶可微, 强凸, 考虑不同方向, 用Taylor展开有

$$\begin{aligned} \min_{\|\mathbf{v}\|=1} f_0(\mathbf{x}^{(k)} + \mathbf{v}) &\approx \min_{\|\mathbf{v}\|=1} f_0(\mathbf{x}^{(k)}) + \langle \nabla f_0(\mathbf{x}^{(k)}), \mathbf{v} \rangle \\ &\approx \min_{\mathbf{v}} f_0(\mathbf{x}^{(k)}) + \nabla f_0(\mathbf{x}^{(k)})^T \mathbf{v} + \frac{1}{2} \mathbf{v}^T \nabla^2 f_0(\mathbf{x}^{(k)}) \mathbf{v} \end{aligned}$$

求梯度有

$$\begin{aligned} \nabla f_0(\mathbf{x}^{(k)}) + \nabla^2 f_0(\mathbf{x}^{(k)}) \mathbf{v} &= 0 \\ \implies \mathbf{v} &= -(\nabla^2 f_0(\mathbf{x}^{(k)}))^{-1} \nabla f_0(\mathbf{x}^{(k)}) \rightarrow \text{牛顿方向} \\ \begin{cases} \mathbf{d}^{(k)} = -(\nabla^2 f_0(\mathbf{x}^{(k)}))^{-1} \nabla f_0(\mathbf{x}^{(k)}) \\ \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha^{(k)} (\nabla^2 f_0(\mathbf{x}^{(k)}))^{-1} \nabla f_0(\mathbf{x}^{(k)}) \end{cases} \end{aligned}$$

其中 $\alpha^{(k)}$ 为搜索步长。

看下降方向只要看其与负梯度方向是否小于90度

$$\begin{aligned} &\langle -\nabla f_0(\mathbf{x}^{(k)}), -(\nabla^2 f_0(\mathbf{x}^{(k)}))^{-1} \nabla f_0(\mathbf{x}^{(k)}) \rangle \\ &= \nabla^T f_0(\mathbf{x}^{(k)}) (\nabla^2 f_0(\mathbf{x}^{(k)}))^{-1} \nabla f_0(\mathbf{x}^{(k)}) \end{aligned}$$

假设 $\nabla^2 f_0(\mathbf{x})$ Lipschitz连续

- 若 $\|\nabla f_0(\mathbf{x}^{(k)})\|_2 > \eta$, 阻尼(damped)牛顿段
用Armijo Rule算步长, $\exists \gamma > 0, f_0(\mathbf{x}^{(k+1)}) - f_0(\mathbf{x}^{(k)}) \leq -\gamma$
- 若 $\|\nabla f_0(\mathbf{x}^{(k)})\|_2 \leq \eta$, 纯牛顿段
 $\alpha = 1, f_0(\mathbf{x}^{(k+1)}) - f_0(\mathbf{x}^*) \leq \Delta (\frac{1}{2})^{2^k}, \exists \Delta > 0$, 超线性收敛

多了二阶信息，往最优解跑的速度会越来越快

例 43.

$$\begin{aligned} \min \quad & \frac{1}{2} \mathbf{x}^T P \mathbf{x} + \mathbf{q}^T \mathbf{r} + c \\ \text{s.t.} \quad & P \succ 0 \end{aligned}$$

分析. 对于二阶强凸问题，只需1步到达最优解；但用梯度下降法，与条件数相关

与Newton-Raphson算法的联系，将其扩展至高维的凸问题

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \frac{g(\mathbf{x}^{(k)})}{g'(\mathbf{x}^{(k)})}$$

6.4.2 拟牛顿法

拟牛顿法(quasi-Newton): 希望像一阶算法一样好算，又像二阶算法一样收敛快

1. 构造 $(\nabla^2 f_0(\mathbf{x}^{(k)}))^{-1}$ 的近似矩阵 $G^{(k)}$ (直接的想法)
2. 构造 $\nabla^2 f_0(\mathbf{x}^{(k)})$ 的近似矩阵 $B^{(k)}$ ，且容易求逆

在 $\mathbf{x} = \mathbf{x}^{(k+1)}$ 点处做Taylor展开

$$\begin{aligned} f_0(\mathbf{x}) &\approx f_0(\mathbf{x}^{(k+1)}) + \langle \nabla f_0(\mathbf{x}^{(k+1)}), \mathbf{x} - \mathbf{x}^{(k+1)} \rangle + \frac{1}{2} (\mathbf{x} - \mathbf{x}^{(k+1)})^T \nabla^2 f_0(\mathbf{x}^{(k+1)}) (\mathbf{x} - \mathbf{x}^{(k+1)}) \\ \implies \nabla f_0(\mathbf{x}) &\approx \nabla f_0(\mathbf{x}^{(k+1)}) + \nabla^2 f_0(\mathbf{x}^{(k+1)}) (\mathbf{x} - \mathbf{x}^{(k+1)}) \\ \implies \nabla f_0(\mathbf{x}^{(k)}) &\approx \nabla f_0(\mathbf{x}^{(k+1)}) + \nabla^2 f_0(\mathbf{x}^{(k+1)}) (\mathbf{x}^{(k)} - \mathbf{x}^{(k+1)}) \\ \implies \nabla f_0(\mathbf{x}^{(k)}) - \nabla f_0(\mathbf{x}^{(k+1)}) &\approx \nabla^2 f_0(\mathbf{x}^{(k+1)}) (\mathbf{x}^{(k)} - \mathbf{x}^{(k+1)}) \end{aligned}$$

$$\begin{cases} \mathbf{q}^{(k)} = \nabla f_0(\mathbf{x}^{(k+1)}) - \nabla f_0(\mathbf{x}^{(k)}) \\ \mathbf{p}^{(k)} = \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} \end{cases} \implies \begin{cases} \mathbf{q}^{(k)} = B^{(k+1)} \mathbf{p}^{(k)} \\ \mathbf{p}^{(k)} = G^{(k+1)} \mathbf{q}^{(k)} \end{cases}$$

1. 近似 $G^{(k)}$

$$G^{(k+1)} = G^{(k)} + \Delta G^{(k)}$$

- a. 秩1校正 (希望 G 中不要有太多元素，故用低秩矩阵做近似)

$$\Delta G^{(k)} = \alpha^{(k)} \mathbf{z}^{(k)} (\mathbf{z}^{(k)})^T$$

$$\begin{aligned} \mathbf{p}^{(k)} &= G^{(k+1)} \mathbf{q}^{(k)} = G^{(k)} \mathbf{q}^{(k)} + \alpha^{(k)} \mathbf{z}^{(k)} (\mathbf{z}^{(k)})^T \mathbf{q}^{(k)} \\ \implies \Delta G^{(k)} &= \frac{(\mathbf{p}^{(k)} - G^{(k)} \mathbf{q}^{(k)}) (\mathbf{p}^{(k)} - G^{(k)} \mathbf{q}^{(k)})^T}{(\mathbf{q}^{(k)})^T (\mathbf{p}^{(k)} - G^{(k)} \mathbf{q}^{(k)})} \end{aligned}$$

稳定性很有问题，分母接近0的时候，越接近最优解越不稳定

b. 秩2校正(Dandon-Fletcher-Power, DFP)

$$\Delta G^{(k)} = \frac{\mathbf{p}^{(k)}(\mathbf{p}^{(k)})^T}{(\mathbf{p}^{(k)})^T \mathbf{q}^{(k)}} - \frac{G^{(k)} \mathbf{q}^{(k)}(\mathbf{q}^{(k)})^T G^{(k)}}{(\mathbf{q}^{(k)})^T G^{(k)} \mathbf{q}^{(k)}}$$

前后项都为秩1矩阵，数值稳定性强

2. 近似 $B^{(k)}$ (Broyden-Fletcher-Goldfarb-Shermo, BFGS)

$$B^{(k+1)} = B^{(k)} + \frac{\mathbf{q}^{(k)}(\mathbf{q}^{(k)})^T}{(\mathbf{q}^{(k)})^T \mathbf{p}^{(k)}} - \frac{B^{(k)} \mathbf{p}^{(k)}(\mathbf{p}^{(k)})^T B^{(k)}}{(\mathbf{p}^{(k)})^T B^{(k)} \mathbf{p}^{(k)}}$$

- 拟牛顿法以后可能很有用，因为结合一二阶优化优点
- 找核心问题（Hessian矩阵难算），然后就去解决
- 用结构信息，都对结构进行限制（一股脑就用Adam优化器，这是不对的，要分析问题结构）

有限内存(limited memory)—LM-BFGS

6.5 有约束优化方法

$$\begin{aligned} \min \quad & f_0(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \in \mathcal{C} \end{aligned}$$

可改写为

$$\begin{aligned} \min \quad & f_0(\mathbf{x}) \\ \text{s.t.} \quad & A\mathbf{x} = \mathbf{b} \end{aligned}$$

本质上都是在考虑它的KKT条件

$$\begin{cases} A\mathbf{x}^* = \mathbf{b} \\ \nabla f_0(\mathbf{x}^*) + A^T \mathbf{v}^* = 0 \end{cases}$$

例 44.

$$\begin{aligned} \min \quad & \frac{1}{2} \mathbf{x}^T P \mathbf{x} + \mathbf{q}^T \mathbf{x} + r, P \succeq 0 \\ \text{s.t.} \quad & A\mathbf{x} = \mathbf{b} \end{aligned}$$

分析. 等价于KKT条件

$$\begin{cases} A\mathbf{x}^* = \mathbf{b} \\ P\mathbf{x}^* + \mathbf{q} + A^T\mathbf{v}^* = 0 \end{cases}$$

等价于

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

6.5.1 约束满足的牛顿法

若方程组非线性, 那就做一个线性化

$$\begin{aligned} \arg \min_{\mathbf{d}} \quad & f_0(\mathbf{x}^{(k)} + \mathbf{d}) = \mathbf{d}^{(k)} \\ \text{subject to} \quad & A(\mathbf{x}^{(k)} + \mathbf{d}) = \mathbf{b} \end{aligned}$$

近似等价于 (二阶近似Taylor展开)

$$\begin{aligned} \arg \min_{\mathbf{d}} \quad & f_0(\mathbf{x}^{(k)}) + \langle \nabla f_0(\mathbf{x}^{(k)}), \mathbf{d} \rangle + \frac{1}{2} \mathbf{d}^T \nabla^2 f_0(\mathbf{x}^{(k)}) \mathbf{d} = \mathbf{d}^{(k)} \\ \text{subject to} \quad & A(\mathbf{x}^{(k)} + \mathbf{d}) = \mathbf{b} \end{aligned}$$

写出问题关于 \mathbf{d} 的KKT条件, 可得等价条件

$$\begin{bmatrix} \nabla f_0(\mathbf{x}^{(k)}) & A^T \\ A & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{d}^{(k)} \\ \mathbf{v}^{(k)} \end{bmatrix} = \begin{bmatrix} -\nabla f_0(\mathbf{x}^{(k)}) \\ \mathbf{b} - A\mathbf{x}^{(k)} \end{bmatrix}$$

若 $\mathbf{x}^{(0)}$ 可行, $A\mathbf{x}^{(0)} = \mathbf{b}$, 之后的 $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha^{(k)} \mathbf{d}^{(k)}$ 也可行。

6.5.2 拉格朗日乘子法/对偶分解法

$$L(\mathbf{x}, \mathbf{v}) = f_0(\mathbf{x}) + \langle A\mathbf{x} - \mathbf{b}, \mathbf{v} \rangle$$

更新原变量和对偶变量

$$\begin{cases} \mathbf{x}^{(k+1)} = \arg \min_{\mathbf{x}} L(\mathbf{x}, \mathbf{v}^{(k)}) \\ \mathbf{v}^{(k+1)} = \mathbf{v}^{(k)} + \alpha^{(k)} (A\mathbf{x}^{(k+1)} - \mathbf{b}) \end{cases}$$

$\alpha^{(k)}$ 可以是固定步长, 也可以是递减步长

即为找鞍点, \mathbf{x} 方向上找最小值, 本来 \mathbf{v} 方向上要找最大值, 但容易到正无穷。因此换种方法 $\mathbf{v}^{(k)}$ 做一个保守的计算, 每一步都走一个很小的步长。

例 45.

$$\begin{aligned} \min \quad & \frac{1}{2}x^2 \\ \text{s.t.} \quad & x = 1 \end{aligned}$$

分析.

$$\begin{aligned} L(x, v) &= \frac{1}{2}x^2 + v(x - 1) \\ &= \frac{1}{2}x^2 + vx - v \end{aligned}$$

6.6 对偶次梯度法

\mathbf{v} 才是最关键的，只是在寻找最优 \mathbf{v} 的时候顺带找到了 \mathbf{x} （收敛到 \mathbf{v}^* 的同时也找到了 \mathbf{x}^* ）

$$D(\mathbf{v}) = \inf_{\mathbf{x}} L(\mathbf{x}, \mathbf{v})$$

$D(\mathbf{v})$ 为凹函数，关注 $-D(\mathbf{v})$

$$\begin{cases} \mathbf{x}^{(k+1)} = \arg \min_{\mathbf{x}} L(\mathbf{x}, \mathbf{v}^{(k)}) \\ \mathbf{v}^{(k+1)} = \mathbf{v}^{(k)} - \alpha^{(k)}(-(A\mathbf{x}^{(k+1)}) - \mathbf{b}) \end{cases}$$

若 $f_0(\mathbf{x})$ 为凸，若 $\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} L(\mathbf{x}, \hat{\boldsymbol{\lambda}})$ ，则 $-(A\hat{\mathbf{x}} - \mathbf{b})$ 为 $-D(\boldsymbol{\lambda})$ 在 $\hat{\boldsymbol{\lambda}}$ 的次梯度

$$\forall \mathbf{v} : -D(\boldsymbol{\lambda}) \geq -D(\hat{\boldsymbol{\lambda}}) + \langle \mathbf{v} - \hat{\boldsymbol{\lambda}}, g(\hat{\boldsymbol{\lambda}}) \rangle$$

$$\begin{aligned} D(\mathbf{v}) &= \inf_{\mathbf{x}} L(\mathbf{x}, \mathbf{v}) \\ &= \inf_{\mathbf{x}} f_0(\mathbf{x}) + \langle \mathbf{v}, A\mathbf{x} + \mathbf{b} \rangle \\ &\leq f_0(\hat{\mathbf{x}}) + \langle \mathbf{v}, A\hat{\mathbf{x}} - \mathbf{b} \rangle \\ &= f_0(\hat{\mathbf{x}}) + \langle \boldsymbol{\lambda}, A\hat{\mathbf{x}} - \mathbf{b} \rangle + \langle \mathbf{v} - \hat{\boldsymbol{\lambda}}, A\hat{\mathbf{x}} - \mathbf{b} \rangle \\ &= D(\hat{\boldsymbol{\lambda}}) + \langle \mathbf{v} - \hat{\boldsymbol{\lambda}}, A\hat{\mathbf{x}} - \mathbf{b} \rangle \\ -D(\mathbf{v}) &\geq -D(\hat{\boldsymbol{\lambda}}) + \langle \mathbf{v} - \hat{\boldsymbol{\lambda}}, -(A\hat{\mathbf{x}} - \mathbf{b}) \rangle \end{aligned}$$

进而得到 $-(A\hat{\mathbf{x}} - \mathbf{b})$ 就是一个次梯度

这个算法一般来说性能不好，在机器学习里面很多时候都被乱用，有时候可以，有时候不行。

- 在什么情况下它是好用的？对偶函数是可微的，采用固定步长。
- 对偶函数 $D(\mathbf{v})$ 何时可微？任何 $-D(\boldsymbol{\lambda})$ 都具有 $-(A\hat{\mathbf{x}} - \mathbf{b})$ 的形式，得到当 $f_0(\mathbf{x})$ 严格凸时， $f_0(\mathbf{x}) + \langle \hat{\boldsymbol{\lambda}}, A\mathbf{x} - \mathbf{b} \rangle$ 严格凸，进而 $D(\mathbf{v})$ 可微

6.6.1 原对偶次梯度法

计算量出在 $\mathbf{x}^{(k+1)}$ ，那么想办法近似

$$\begin{cases} \mathbf{x}^{(k+1)} &= \mathbf{x}^{(k)} - \alpha^{(k)} \partial_x L(\mathbf{x}, \mathbf{v}^{(k)}) \\ \mathbf{v}^{(k+1)} &= \mathbf{v}^{(k)} + \alpha^{(k)} \partial_v L(\mathbf{x}^{(k+1)}, \mathbf{v}) \\ &= \mathbf{v}^{(k)} + \alpha^{(k)} (A\mathbf{x}^{(k+1)} - \mathbf{b}) \end{cases}$$

$\mathbf{v}^{(k+1)}$ 需要等待 $\mathbf{x}^{(k+1)}$ ，将其换成下式可以不用等待

$$\mathbf{v}^{(k)} + \alpha^{(k)} (A\mathbf{x}^{(k)} - \mathbf{b})$$

由于两个方向都不精确，故收敛性质糟糕。

6.6.2 增广(augmented)拉格朗日法

当函数不是严格凸时，依然能得到很好的效果

$$\begin{aligned} \min \quad & f_0(\mathbf{x}) \\ \text{s.t.} \quad & A\mathbf{x} = \mathbf{b} \end{aligned}$$

$$L(\mathbf{x}, \mathbf{v}) = f_0(\mathbf{x}) + \langle \mathbf{v}, A\mathbf{x} - \mathbf{b} \rangle = f_0(\mathbf{x}) + \mathbf{v}^T (A\mathbf{x} - \mathbf{b})$$

增广拉格朗日函数

$$L_c(\mathbf{x}, \mathbf{v}) = f_0(\mathbf{x}) + \langle \mathbf{v}, A\mathbf{x} - \mathbf{b} \rangle + \frac{c}{2} \|A\mathbf{x} - \mathbf{b}\|^2, c > 0$$

增广拉格朗日函数是下面问题的拉格朗日函数

$$\begin{aligned} \min \quad & f_0(\mathbf{x}) + \frac{c}{2} \|A\mathbf{x} - \mathbf{b}\|^2 \\ \text{s.t.} \quad & A\mathbf{x} = \mathbf{b} \end{aligned}$$

两个问题的原对偶最优解相同

设 $(\mathbf{x}^*, \mathbf{v}^*)$ 为原问题最优解，分别有以下两个KTT条件

$$\begin{cases} A\mathbf{x}^* = \mathbf{b} \\ \left. \frac{\partial L(\mathbf{x}, \mathbf{v}^*)}{\partial \mathbf{x}} \right|_{\mathbf{x}=\mathbf{x}^*} = 0 \end{cases} \quad \begin{cases} A\mathbf{x}^* = \mathbf{b} \\ \left. \frac{\partial L_c(\mathbf{x}, \mathbf{v}^*)}{\partial \mathbf{x}} \right|_{\mathbf{x}=\mathbf{x}^*} = 0 \end{cases}$$

对于原问题有

$$\left. \nabla_x (f_0(\mathbf{x}) + \langle \mathbf{v}^*, A\mathbf{x} - \mathbf{b} \rangle) \right|_{\mathbf{x}=\mathbf{x}^*} = 0$$

对于对偶问题有

$$\begin{aligned}
 & \nabla_x \left(f_0(\mathbf{x}) + \langle \mathbf{v}^*, A\mathbf{x} - \mathbf{b} \rangle + \frac{c}{2} \|A\mathbf{x} - \mathbf{b}\|^2 \right) \Big|_{\mathbf{x}=\mathbf{x}^*} \\
 &= \nabla_x \left(\frac{c}{2} \|A\mathbf{x} - \mathbf{b}\|^2 \right) \Big|_{\mathbf{x}=\mathbf{x}^*} \quad \text{原问题最优解代入} \\
 &= cA^T(A\mathbf{x}^* - \mathbf{b}) = 0
 \end{aligned}$$

得到增广拉格朗日法的迭代格式

$$\begin{cases} \mathbf{x}^{(k+1)} = \arg \min_{\mathbf{x}} L_c(\mathbf{x}, \mathbf{v}^{(k)}) \\ \mathbf{v}^{(k+1)} = \mathbf{v}^{(k)} + c(A\mathbf{x}^{(k+1)} - \mathbf{b}) \end{cases}$$

只要原问题是凸问题，无论 c 怎么取（ c 刚好就是固定步长），该算法总是可以收敛（不考虑计算精度的问题），只是收敛速度不同

例 46.

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

分析.

$$\begin{aligned}
 L(\mathbf{x}, v) &= \frac{1}{2}x_1^2 + \frac{1}{2}x_2^2 + v(x_1 - 1) \\
 \frac{\partial L(\mathbf{x}, v^*)}{\partial \mathbf{x}} \Big|_{\mathbf{x}=\mathbf{x}^*} &= 0 = \begin{bmatrix} x_1 + v^* \\ x_2 \end{bmatrix}
 \end{aligned}$$

增广拉格朗日法

$$\begin{aligned}
 L_{c^{(k)}}(\mathbf{x}, v) &= \frac{1}{2}x_1^2 + \frac{1}{2}x_2^2 + v(x_1 - 1) + \frac{c^{(k)}}{2}(x_1 - 1)^2 \\
 \mathbf{x}^{(k+1)} &= \arg \min L_{c^{(k)}}(\mathbf{x}, v^{(k)}) \\
 \begin{cases} x_1 + v^{(k)} + c^{(k)}(x_1 - 1) = 0 \\ x_2 = 0 \end{cases} &\implies \mathbf{x}^{(k+1)} = \begin{bmatrix} \frac{c^{(k)} - v^{(k)}}{c^{(k)} + 1} \\ 0 \end{bmatrix} \\
 v^{(k+1)} &= v^{(k)} + c^{(k)}(x_1^{(k+1)} - 1) \\
 &= v^{(k)} + c^{(k)} \left(\frac{c^{(k)} - v^{(k)}}{c^{(k)} + 1} - 1 \right) \\
 &= \frac{v^{(k)}}{c^{(k)} + 1} - \frac{c^{(k)}}{c^{(k)} + 1} \\
 v^{(k+1)} - v^* &= v^{(k+1)} + 1 = \frac{v^{(k)}}{c^{(k)} + 1} + \frac{1}{c^{(k)} + 1} = \frac{v^{(k)} - v^*}{c^{(k)} + 1}
 \end{aligned}$$

可以看出取一个固定步长，且大于零，增广拉格朗日的收敛是非常好的（线性收敛）

对于特殊的一些非凸问题，增广拉格朗日也是有效的，如把问题改成

$$\min \left(-\frac{1}{2}x_1^2 + \frac{1}{2}x_2^2 \right)$$

6.6.3 交替方向乘子法

交替方向乘子法(alternating direction method of multipliers, ADMM)同样探究有结构的优化问题。

$$\begin{aligned} \min_{\mathbf{x}, \mathbf{y}} \quad & f(\mathbf{x}) + g(\mathbf{y}) \\ \text{s.t.} \quad & A\mathbf{x} + B\mathbf{y} = 0 \end{aligned}$$

考虑增广拉格朗日函数

$$\begin{aligned} L_c(\mathbf{x}, \mathbf{y}, \mathbf{v}) &= f(\mathbf{x}) + g(\mathbf{y}) + \langle \mathbf{v}, A\mathbf{x} + B\mathbf{y} \rangle + \frac{c}{2} \|A\mathbf{x} + B\mathbf{y}\|_2^2 \\ &= f(\mathbf{x}) + g(\mathbf{y}) + \frac{c}{2} \left(\left\| A\mathbf{x} + B\mathbf{y} + \frac{\mathbf{v}}{c} \right\|_2^2 - \left\| \frac{\mathbf{v}}{c} \right\|_2^2 \right) \end{aligned}$$

若直接用下面的迭代格式

$$\begin{cases} (\mathbf{x}^{(k+1)}, \mathbf{y}^{(k+1)}) = \arg \min_{\mathbf{x}, \mathbf{y}} L_c(\mathbf{x}, \mathbf{y}, \mathbf{v}^{(k)}) \\ \mathbf{v}^{(k+1)} = \mathbf{v}^{(k)} + c(A\mathbf{x}^{(k+1)} + B\mathbf{y}^{(k+1)}) \end{cases}$$

由于在 $\|A\mathbf{x} + B\mathbf{y}\|_2^2$ 中， \mathbf{x} 和 \mathbf{y} 结合在一起，不好优化，故用交替的方法（选主元）来解决

$$\begin{cases} \mathbf{x}^{(k+1)} = \arg \min_{\mathbf{x}} L_c(\mathbf{x}, \mathbf{y}^{(k)}, \mathbf{v}^{(k)}) \\ \quad = \arg \min_{\mathbf{x}} f(\mathbf{x}) + \langle \mathbf{v}^{(k)}, A\mathbf{x} \rangle + \frac{c}{2} \|A\mathbf{x} + B\mathbf{y}^{(k)}\|_2^2 \\ \quad \iff \arg \min_{\mathbf{x}} f(\mathbf{x}) + \frac{c}{2} \|A\mathbf{x} + B\mathbf{y}^{(k)} + \frac{\mathbf{v}^{(k)}}{c}\|_2^2 \quad \text{配方, 关于}\mathbf{y}^{(k)}\text{的项为常数项, 可忽略} \\ \mathbf{y}^{(k+1)} = \arg \min_{\mathbf{y}} L_c(\mathbf{x}^{(k+1)}, \mathbf{y}, \mathbf{v}^{(k)}) \\ \quad \iff \arg \min_{\mathbf{y}} g(\mathbf{y}) + \frac{c}{2} \|A\mathbf{x}^{(k+1)} + B\mathbf{y} + \frac{\mathbf{v}^{(k)}}{c}\|_2^2 \\ \mathbf{v}^{(k+1)} = \mathbf{v}^{(k)} + c(A\mathbf{x}^{(k+1)} + B\mathbf{y}^{(k+1)}) \end{cases}$$

两块的算法依然具有很好的收敛性，但是多块的交替方向乘子法不一定可以收敛。

例 47 (LASSO).

$$\min \frac{1}{2} \|A\mathbf{x} - \mathbf{b}\|_2^2 + p \|\mathbf{x}\|_1$$

分析. 对原问题进行变形，等价于下述约束问题

$$\begin{aligned} \min \quad & \frac{1}{2} \|A\mathbf{x} - \mathbf{b}\|_2^2 + p \|\mathbf{y}\|_1 \\ \text{s.t.} \quad & \mathbf{x} - \mathbf{y} = 0 \end{aligned}$$

构造增广拉格朗日函数

$$L_c(\mathbf{x}, \mathbf{y}, \mathbf{v}) = \frac{1}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + p \|\mathbf{y}\|_1 + \mathbf{v}^T(\mathbf{x} - \mathbf{y}) + \frac{c}{2} \|\mathbf{x} - \mathbf{y}\|_2^2$$

可以得到交替方向乘子法的迭代格式

$$\begin{cases} \mathbf{x}^{(k+1)} = \arg \min_{\mathbf{x}} L_c(\mathbf{x}, \mathbf{y}^{(k)}, \mathbf{v}^{(k)}) \\ \mathbf{y}^{(k+1)} = \arg \min_{\mathbf{y}} L_c(\mathbf{x}^{(k+1)}, \mathbf{y}, \mathbf{v}^{(k)}) \\ \mathbf{v}^{(k+1)} = \mathbf{v}^{(k)} + c(\mathbf{x}^{(k+1)} - \mathbf{y}^{(k+1)}) \end{cases}$$

对 $\mathbf{x}^{(k+1)}$ 展开并配方，并将非主元项忽略，可求得上式与下面的式子等价

$$\begin{cases} \mathbf{x}^{(k+1)} = \arg \min_{\mathbf{x}} \left(\frac{1}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \frac{c}{2} \left\| \mathbf{x} - \mathbf{y}^{(k)} + \frac{\mathbf{v}^{(k)}}{c} \right\|_2^2 \right) \\ \mathbf{y}^{(k+1)} = \arg \min_{\mathbf{y}} \left(p \|\mathbf{y}\|_1 + \frac{c}{2} \left\| \mathbf{x}^{(k+1)} - \mathbf{y} + \frac{\mathbf{v}^{(k)}}{c} \right\|_2^2 \right) \\ \mathbf{v}^{(k+1)} = \mathbf{v}^{(k)} + c(\mathbf{x}^{(k+1)} - \mathbf{y}^{(k+1)}) \end{cases}$$

对于 $\mathbf{x}^{(k+1)}$ ，可直接通过求梯度的方法得到显式解，得到

$$\mathbf{x}^{(k+1)} = (\mathbf{A}^T \mathbf{A} + c\mathbf{I})^{-1}(\mathbf{A}^T \mathbf{b} + c\mathbf{y}^{(k)} - \mathbf{v}^{(k)})$$

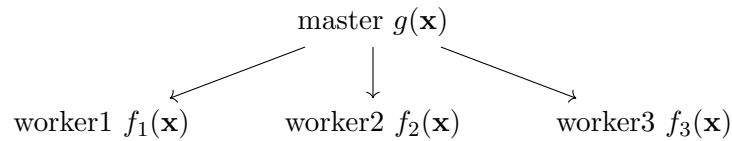
对于 $\mathbf{y}^{(k+1)}$ ，由于涉及一范数，故需要求次微分，设 $z_i = x_i^{(k+1)} + \frac{v_i^{(k)}}{c}$ ，可得到类似的软门限表达式

$$y_i = \begin{cases} z_i - \frac{p}{c} & z_i > \frac{p}{c} \\ 0 & z_i \in \left[-\frac{p}{c}, \frac{p}{c}\right] \\ z_i + \frac{p}{c} & z_i < -\frac{p}{c} \end{cases}$$

进而可以迭代求解。

7 大数据中的优化问题与算法

7.1 并行优化



$$\min_{\mathbf{x}} g(\mathbf{x}) + \sum_{i=1}^N f_i(\mathbf{x})$$

针对LASSO问题，每个人都有一个样本 (A_i, \mathbf{b}_i) ，最小化样本之和，以及正则化项

$$\begin{cases} (A_1, \mathbf{b}_1) \implies \frac{1}{2} \|\mathbf{A}_1 \mathbf{x} - \mathbf{b}_1\|_2^2 \\ \vdots \\ (A_n, \mathbf{b}_n) \implies \frac{1}{2} \|\mathbf{A}_n \mathbf{x} - \mathbf{b}_n\|_2^2 \\ g(\mathbf{x}) = v \|\mathbf{x}\|_1 \end{cases}$$

原问题即为

$$\min_x v \|\mathbf{x}\|_1 + \frac{c}{2} \left\| \begin{bmatrix} A_1 & \cdots & A_N \end{bmatrix} \mathbf{x} - \begin{bmatrix} \mathbf{b}_1 & \cdots & \mathbf{b}_N \end{bmatrix} \right\|_2^2$$

7.1.1 并行梯度下降法

$$\begin{cases} \mathbf{x}^{(k+\frac{1}{2})} = \mathbf{x}^{(k)} - \alpha \sum_{i=1}^N \nabla f_i(\mathbf{x}^{(k)}) \\ \mathbf{x}^{(k+1)} = \arg \min_{\mathbf{x}} g(\mathbf{x}) + \frac{1}{2\alpha} \left\| \mathbf{x} - \mathbf{x}^{(k+\frac{1}{2})} \right\|_2^2 \end{cases}$$

计算简单，只需求梯度，但所有梯度类问题都依赖于条件数。通信开销大。

7.1.2 对偶分解法

$$\begin{aligned} \min \quad & \sum_{i=1}^N f_i(\mathbf{x}_i) + g(\mathbf{z}) \\ \text{s.t.} \quad & \mathbf{x}_i = \mathbf{z}, \forall i \end{aligned}$$

$$L(\mathbf{x}, \mathbf{z}, \mathbf{v}) = \sum_{i=1}^N f_i(\mathbf{x}_i) + g(\mathbf{z}) + \sum_{i=1}^N \langle \mathbf{v}_i, \mathbf{x}_i - \mathbf{z} \rangle$$

$$(\mathbf{x}^{(k+1)}, \mathbf{z}^{(k+1)}) = \arg \min_{\mathbf{x}, \mathbf{z}} L(\mathbf{x}, \mathbf{z}, \mathbf{v}^{(k)})$$

$$\implies \begin{cases} \mathbf{x}_i^{(k+1)} = \arg \min_{\mathbf{x}_i} f_i(\mathbf{x}_i) + \langle \mathbf{v}_i^{(k)}, \mathbf{x}_i - \mathbf{z} \rangle \\ \mathbf{z}^{(k+1)} = \arg \min_{\mathbf{z}} g(\mathbf{z}) - \sum_{i=1}^N \langle \mathbf{v}_i^{(k)}, \mathbf{z} \rangle \\ \mathbf{v}_i^{(k+1)} = \mathbf{v}_i^{(k)} + \alpha \langle \mathbf{x}_i^{(k+1)} - \mathbf{z}^{(k+1)}, \mathbf{v}_i^{(k)} \rangle \end{cases}$$

不依赖于条件数，但每一步都需要求解一个最优解，不一定好求。通信开销小，但拉格朗日类方法收敛性差。

7.1.3 增广拉格朗日函数

$$\sum_{i=1}^n f_i(\mathbf{x}_i) + g(\mathbf{z}) + \sum_{i=1}^n \langle \boldsymbol{\lambda}_i, \mathbf{x}_i - \mathbf{z} \rangle + \frac{c}{2} \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{z}\|^2$$

正则项会产生 \mathbf{x}_i 和 \mathbf{z} 的交叉项，不好处理

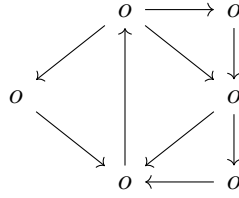
注意到 \mathbf{x}_i 之间是没有依赖的，故采用交替方向乘子法，加了增广项，可用固定步长达到最优解

$$\begin{cases} \mathbf{x}_i^{(k+1)} &= \arg \min f_0(\mathbf{x}_i) + \langle \boldsymbol{\lambda}_i^{(k)}, \mathbf{x}_i \rangle + \frac{c}{2} \|\mathbf{x}_i - \mathbf{z}^{(k)}\|^2 \\ \mathbf{z}^{(k+1)} &= \arg \min g(\mathbf{z}) - \langle \sum_{i=1}^n \boldsymbol{\lambda}_i^{(k)}, \mathbf{z} \rangle + \frac{c}{2} \sum_{i=1}^n \|\mathbf{x}_i^{(k+1)} - \mathbf{z}\|^2 \\ \boldsymbol{\lambda}^{(k+1)} &= \boldsymbol{\lambda}^{(k)} + c(\mathbf{x}_i^{(k+1)} - \mathbf{z}^{(k+1)}) \end{cases}$$

每次要多传一倍的变量，以通信量开销换性能提升

7.2 无中心分布式优化

考虑无向图



每个结点自己优化，协同决策

$$\min \sum_{i=1}^n f_i(\mathbf{x})$$

梯度下降法，但由于去中心， $\mathbf{x}^{(k)}$ 无处摆放

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha^{(k)} \sum_{i=1}^n \nabla f_i(\mathbf{x}^{(k)})$$

那就每一个点分配一个本地变量 \mathbf{x}_i ，对邻居的更新做一个加权平均

$$\mathbf{x}^{(k+1)} = \sum_{i=1}^n \omega_{ij} \mathbf{x}_j^{(k)} - \alpha^{(k)} \sum_{i=1}^n \omega_{ij} \nabla f_j(\mathbf{x}_j^{(k)})$$

其中

$$\begin{cases} \omega_{ij} \neq 0 & (i, j) \in E, i = j \\ \omega_{ij} = 0 & \text{otherwise} \end{cases}$$

$$W = [\omega_{ij}], W = W^T, W\mathbf{1} = \mathbf{1}$$

在不可信的系统里面，存在个人隐私等信息，故更激进些，采用自己的梯度进行更新（在本地进行梯度下降），在无人机系统中非常常见

$$\mathbf{x}_i^{(k+1)} = \sum_{i=1}^n \omega_{ij} \mathbf{x}_j^{(k)} - \alpha^{(k)} \nabla f_i(\mathbf{x}_i^{(k)})$$

非常糟糕的算法，如果采用固定步长，则找不到最优解

分析. 反证法, 假设 $\mathbf{x}_i^{(k)} \rightarrow \mathbf{x}^*$, 将 \mathbf{x}^* 代入

$$\mathbf{x}^* = \sum_{j=1}^n \omega_{ij} \mathbf{x}^* - \alpha \nabla f_i(\mathbf{x}^*) \iff \nabla f_i(\mathbf{x}^*) = 0$$

但原问题最优解

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

与上面的式子不等价

改成有约束优化的形式

$$\begin{aligned} \min \quad & \sum_{i=1}^n f_i(\mathbf{x}_i) \\ \text{s.t.} \quad & \mathbf{x}_1 = \mathbf{x}_2 = \cdots = \mathbf{x}_n \end{aligned}$$

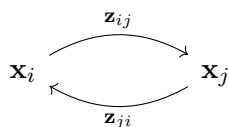
写出拉格朗日函数, 对偶分解法

$$\sum_{i=1}^n f_i(\mathbf{x}_i) + \sum_{(i,j) \in E} \langle \boldsymbol{\lambda}_{ij}, \mathbf{x}_i - \mathbf{x}_j \rangle$$

别人的东西都在对偶变量中体现, 求同存异

$$\begin{cases} \mathbf{x}_i^{(k+1)} = \arg \min f_i(\mathbf{x}_i) + \sum_{(i,j) \in E} \langle \boldsymbol{\lambda}_{ij}^{(k)}, \mathbf{x}_i \rangle - \sum_{(j,i) \in E} \langle \boldsymbol{\lambda}_{ij}^{(k)}, \mathbf{x}_i \rangle \\ \boldsymbol{\lambda}_{ij}^{(k+1)} = \boldsymbol{\lambda}_{ij}^{(k)} + \alpha^{(k)} (\mathbf{x}_i^{(k+1)} - \mathbf{x}_j^{(k+1)}) - \sum_{(j,i) \in E} \langle \boldsymbol{\lambda}_{ij}^{(k)}, \mathbf{x}_i \rangle \end{cases}$$

依然要采用递减步长, 才能保证收敛



$$\begin{aligned} \min \quad & \sum_{i=1}^n f_i(\mathbf{x}_i) \\ \text{s.t.} \quad & \mathbf{x}_i = \mathbf{z}_{ij}, \forall (i,j) \in E \\ & \mathbf{x}_j = \mathbf{z}_{ji}, \forall (i,j) \in E \end{aligned}$$

可分线性约束, 进而可以用交替方向乘子法

$$\begin{aligned} \sum_{i=1}^n f_i(\mathbf{x}_i) + \sum_{(i,j) \in E} (\langle \boldsymbol{\alpha}_{ij}, \mathbf{x}_i - \mathbf{z}_j \rangle + \langle \boldsymbol{\beta}_{ij}, \mathbf{x}_j - \mathbf{z}_{ji} \rangle) \\ + \sum_{(i,j) \in E} \frac{c}{2} (\|\mathbf{x}_i - \mathbf{z}_{ij}\|^2 + \|\mathbf{x}_j - \mathbf{z}_{ji}\|^2) \end{aligned}$$

$$\begin{cases} \mathbf{x}_i^{(k+1)} &= \arg \min f_i(\mathbf{x}_i) + \sum_{(i,j) \in E} \langle \alpha_{ij} k^{(\cdot)} \mathbf{x}_i \rangle + \sum_{(i,j) \in E} \langle \beta_{ji}^{(k)}, \mathbf{x}_i \rangle \\ &+ \frac{c}{2} \sum_{(i,j) \in E} \|\mathbf{x}_i - \mathbf{z}_j^{(k)}\|^2 + \frac{c}{2} \sum_{(i,j) \in E} \|\mathbf{x}_i - \mathbf{z}_{ji}^{(k)}\|^2 \\ \mathbf{z}_j^{(k+1)} &= \dots \\ \alpha_{ij}^{(k+1)} &= \dots \\ \beta_{ij}^{(k+1)} &= \dots \end{cases}$$

7.3 有限和优化

n 个样本，每个样本为 $f_i(\mathbf{x})$

$$\min_{\mathbf{x}} \quad \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{x})$$

等价于期望极小化问题

$$\min_{\mathbf{x}} \quad \mathbb{E}(f_i(\mathbf{x}, \xi))$$

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha \frac{1}{n} \sum_{i=1}^n \nabla f_i(\mathbf{x}^{(k)})$$

将 k 改为 $i^{(k)}$ ，随机梯度下降(Stochastic gradient descent, SGD)，取了一个无偏的估计[Bottou, NIPS 2010]

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha^{(k)} \nabla f_{i^{(k)}}(\mathbf{x}^{(k)})$$

注意这里需要采用变步长，否则无法收敛到最优解

$$\begin{cases} \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha^{(k)} \nabla f_{i^{(k)}}(\mathbf{x}^{(k)}) \\ \mathbf{x}^* = \mathbf{x}^* - \alpha^{(k)} \nabla f_{i^{(k)}}(\mathbf{x}^*) \end{cases} \implies \nabla f_{i^{(k)}}(\mathbf{x}^*) = 0$$

若问题强凸， $O(\frac{1}{k}) \rightarrow O(\frac{1}{k})$ ；凸， $O(\frac{1}{\sqrt{k}}) \rightarrow O(\frac{1}{\sqrt{k}})$

梯度噪声的问题：选的随机梯度与真正的全梯度不同

7.4 方差消减

方差消减(Variance Reduction)：挑选样本数目增大时，方差会减小

1. 小批量(mini-batch)
2. SURG、SAG、SAGA

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \frac{\alpha}{n} \sum_{i=1}^n y_i^{(k)}$$

对于每一个样本都存储一个梯度值

$$y_i^{(k)} = \begin{cases} \nabla f_i(\mathbf{x}^{(k)}) & i = i^{(k)} \\ y_i^{(k-1)} & i \neq i^{(k)} \end{cases}$$

当时间足够长，每一个里面都存在最优梯度

$$\mathbf{x}^* = \mathbf{x}^* - \frac{\alpha}{n} \sum_{i=1}^n \nabla f_i(\mathbf{x}^*)$$

相当于用空间换时间

7.5 深度神经网络

$$\min \sum_{i=1}^S E^{(i)}$$

其中，

$$E^{(i)} = \frac{1}{2} \|\mathbf{x}_n^{(i)} - Y^{(i)}\|^2$$

为损失函数， \mathbf{x}_n 为第 n 层的网络输出 $f_n(\mathbf{x}_{n-1}, \mathbf{w}_n)$ ，与有限和优化问题相同

反向传播算法(Back propagation): 自底向上求出 E 相对于 \mathbf{x}_n 和 \mathbf{w}_n 的梯度

$$\begin{cases} \frac{\partial E^{(i)}}{\partial \mathbf{x}_n^{(i)}} = \mathbf{x}_n^{(i)} - Y^{(i)} \\ \frac{\partial E^{(i)}}{\partial \mathbf{w}_n} = \frac{\partial E^{(i)}}{\partial \mathbf{x}_n^{(i)}} \frac{\partial \mathbf{x}_n^{(i)}}{\partial \mathbf{w}_n} = \frac{\partial E^{(i)}}{\partial \mathbf{x}_n^{(i)}} \frac{\partial f_n(\mathbf{x}_n^{(i)}, \mathbf{w}_n)}{\partial \mathbf{w}_n} \end{cases}$$

7.6 在线优化

在线优化(Online Learning): 样本不是已有的，而是依照时间给出的

$$\min \frac{1}{T} \sum_{t=1}^T f_t(\mathbf{x})$$

迭代格式为

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \alpha_t \nabla f_t(\mathbf{x}_t)$$

Regret分析: 将当前值丢进下一刻的优化函数中，如果优化效果好，说明有预测能力

7.7 动态优化

动态优化问题

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

$$\mathbf{x}_t = \mathbf{x}_{t-1} - \alpha \nabla f_t(\mathbf{x}_{t-1})$$

7.8 Nesterov加速

Nesterov加速 $\min f(\mathbf{x})$: $O\left(\frac{1}{k^2}\right)$

$$\begin{cases} \mathbf{x}^{(k+1)} = \mathbf{y}^{(k)} - \frac{1}{L} \nabla f(\mathbf{y}^{(k)}) \\ \mathbf{y}^{(k+1)} = (1 - \gamma^{(k)}) \mathbf{x}^{(k+1)} + \gamma^{(k)} \mathbf{x}^{(k)} \\ \beta^{(k)} = \frac{1 + \sqrt{1 + 4(\beta^{(k-1)})^2}}{2}, \beta^{(0)} = 0 \\ \gamma^{(k)} = \frac{1 - \beta^{(k)}}{\beta^{(k+1)}} \end{cases}$$

构造两个序列， \mathbf{y} 为辅助序列，利用问题本身历史信息，做一个凸组合（先跳一步，从 $\mathbf{y}^{(k)}$ 开始做梯度下降）。权重为 γ ，不同时刻权重不同，引入 β 系数。

Trick：为避免权重趋于0（ \mathbf{x} 和 \mathbf{y} 趋同），加速了 n 步后重新设置 β 为0。

梯度下降相当于对 f 做一个二阶近似，二阶Taylor展开。

$$\xrightarrow{\mathbf{x}^{(k)}} f(\mathbf{x}) \longrightarrow \nabla f(\mathbf{x}^{(k)})$$

Nesterov加速是针对确定性优化问题，而机器学习是随机优化问题。

A 线性代数基础

A.1 内积与范数

定义 24 (标准内积). 标准向量内积

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n: \langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y} = \sum_{i=1}^n x_i y_i$$

标准矩阵内积

$$\forall X, Y \in \mathbb{R}^{m \times n} = \text{tr}(X^T Y) = \sum_{i=1}^m \sum_{j=1}^n X_{ij} Y_{ij}$$

定义 25 (向量范数). $\|\cdot\|$ 为范数需要满足以下三个条件

1. 齐次: $\|a\mathbf{x}\| = |a| \|\mathbf{x}\|$
2. 正定: $\|\mathbf{x}\| \geq 0 \iff \mathbf{x} = \mathbf{0}$ 时取等
3. 三角不等式: $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$

常见的向量范数如下, 注意每个元素都要加绝对值

- 0-范数: 非零元素数目, 是伪范数 (不符合第一个定义)
- 1-范数: 绝对值之和

$$\|\mathbf{x}\|_1 = |x_1| + |x_2| + \cdots + |x_n|$$

- 2-范数: 欧几里得距离

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^n |x_i|^2}$$

- p-范数: p次方之和的p次根

$$\|\mathbf{x}\|_p = \sqrt[p]{\sum_{i=1}^n |x_i|^p}$$

- 无穷范数(Chebyshev): 最大值

$$\|\mathbf{x}\|_\infty = \max\{|x_1|, \dots, |x_n|\}$$

定义 26 (矩阵的范数). 对于n阶方阵A, 若对应非负实数 $\|A\|$, 满足

1. $\|A\| \geq 0$, 当且仅当 $A = \mathbf{0}$ 时等号成立
2. 对任意 $\alpha \in \mathbb{R}$, $\|\alpha A\| = |\alpha| \cdot \|A\|$
3. 对任意两个n阶方阵A和B, 满足三角不等式 $\|A + B\| \leq \|A\| + \|B\|$
4. 对任意两个n阶方阵A和B, 满足矩阵乘法要求 $\|AB\| \leq \|A\| \cdot \|B\|$

则称 $\|A\|$ 为方阵A的矩阵范数。

记 $\rho(A) = \max_{1 \leq i \leq n} |\lambda_i|$ 为A的谱半径, 这里 λ_i 为A的特征值, 则常见的矩阵范数定义如下

- 1-范数：绝对值之和

$$\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^n |a_{ij}|$$

- 2-范数/谱范数： $A^T A$ 的谱半径

$$\|A\|_2 = \sqrt{\rho(A^T A)}$$

- 无穷范数：最大值

$$\|A\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}|$$

- Frobenius(F-)范数：注意F-范数才是向量2-范数的直接推广

$$\|X\|_F = (\text{tr}(X^T X))^{1/2} = \left(\sum_{i=1}^m \sum_{j=1}^n X_{ij}^2 \right)^{1/2}$$

A.2 二次型

定义 27 (二次型). $Q(\mathbf{x}) = \mathbf{x}^T A \mathbf{x}$, 其中 A 是对称矩阵

将二次型合并为矩阵的写法, 平方项放对角线, 交叉项取一半对称写. 以三元二次型为例, 观察下面各个元素的去向.

$$\begin{aligned} Q(\mathbf{x}) = \mathbf{x}^T A \mathbf{x} &= \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \\ &= a_{11}x_1^2 + a_{22}x_2^2 + a_{33}x_3^2 + (a_{12} + a_{21})x_1x_2 + (a_{13} + a_{31})x_1x_3 + (a_{23} + a_{32})x_2x_3 \end{aligned}$$

定义 28 (正定(positive definite)矩阵). 若矩阵 A 满足

$$\forall \mathbf{z} \neq 0 : \mathbf{z}^T A \mathbf{z} > 0$$

则称 A 为正定矩阵。

定义 29 (合同(congruent)矩阵). 若存在可逆矩阵 C 使得 $C^T A C = B$, 则称 A 为 B 合同, 记作 $A \simeq B$

定义 30 (主子式). 从 n 阶矩阵中选取行号和列号相同的 i 列, 行列交汇处的元素形成的行列式称为 n 阶矩阵的一个 i 阶主子式。如果挑选 $1 \sim i$ 行和 $1 \sim i$ 列, 则成为该矩阵的 i 阶顺序主子式。

正定矩阵 A 的等价命题如下

- A 的所有顺序主子式均为正
- A 的所有主子式均为正
- A 的特征值均为正

关于可逆矩阵的性质如下:

- 可逆矩阵即非奇异(non-singular)矩阵
- 行列式为0
- 满秩

A.3 特征值分解

定义 31 (谱分解/特征值分解). 假设 $A \in \mathbb{S}^n$, 则 A 可被分解为 $A = Q\Lambda Q^T$, 其中 $Q \in \mathbb{R}^{n \times n}$ 为正交矩阵, 满足 $Q^T Q = I$, 而 $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$

利用特征值可将行列式和迹表示成

$$\det A = \prod_{i=1}^n \lambda_i \quad \text{tr } A = \sum_{i=1}^n \lambda_i$$

而谱范数和F-范数同样可表示为

$$\|A\|_2 = \max_{i=1, \dots, n} |\lambda_i| = \max\{\lambda_1, -\lambda_n\} \quad \|A\|_F = \left(\sum_{i=1}^n \lambda_i^2 \right)^{1/2}$$

最大和最小特征值满足

$$\lambda_{\max}(A) = \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\mathbf{x}^T A \mathbf{x}}{\mathbf{x}^T \mathbf{x}} \quad \lambda_{\min}(A) = \inf_{\mathbf{x} \neq \mathbf{0}} \frac{\mathbf{x}^T A \mathbf{x}}{\mathbf{x}^T \mathbf{x}}$$

特别地, 对于任意 \mathbf{x} , 有

$$\lambda_{\min}(A) \mathbf{x}^T \mathbf{x} \leq \mathbf{x}^T A \mathbf{x} \leq \lambda_{\max}(A) \mathbf{x}^T \mathbf{x}$$

令 $A \in \mathbb{S}_+^n$ 的特征值分解为 $A = Q \text{diag}(\lambda_1, \dots, \lambda_n) Q^T$, 则定义 A 的 (对称) 平方根为

$$A^{1/2} = Q \text{diag}(\lambda_1^{1/2}, \dots, \lambda_n^{1/2}) Q^T$$

平方根 $A^{1/2}$ 是矩阵方程 $X^2 = A$ 的唯一的对称半正定的解

B 矩阵微积分

B.1 基本定义

定义 32 (Taylor公式). 对 $f(\mathbf{x})$ 在点 \mathbf{x}_0 处展开有

$$f(\mathbf{x}) = f(\mathbf{x}_0) + (\mathbf{x} - \mathbf{x}_0)^T \nabla f(\mathbf{x}_0) + \frac{1}{2!} (\mathbf{x} - \mathbf{x}_0)^T \nabla^2(\mathbf{x})(\mathbf{x} - \mathbf{x}_0) + \dots$$

B.2 实值函数对向量的导数

设 $f(\mathbf{x}) : \mathbb{R}^n \mapsto \mathbb{R}$, 劈形算子 ∇ 默认对 \mathbf{x} 求导, 可以得到以下公式。

1. $\nabla(\mathbf{v}^T \mathbf{x}) = \mathbf{v}$

分析. 展开, 对每一个元素讨论

$$\frac{\partial \mathbf{v}^T \mathbf{x}}{\partial x_i} = \frac{\partial \sum_{i=1}^n v_i x_i}{\partial x_i} = v_i$$

2. $\nabla \|\mathbf{x}\|_2^2 = \nabla(\mathbf{x}^T \mathbf{x}) = 2\mathbf{x}$

分析. 法一: 考虑每一元素

$$\frac{\partial \mathbf{x}^T \mathbf{x}}{\partial x_i} = \frac{\partial \sum_{i=1}^n x_i^2}{\partial x_i} = 2x_i$$

法二: 变量多次出现的求导法则, 下标 c 代表视为常数

$$\nabla(\mathbf{x}^T \mathbf{x}) = \nabla(\mathbf{x}_c^T \mathbf{x}) + \nabla(\mathbf{x}^T \mathbf{x}_c) = 2\nabla(\mathbf{x}_c^T \mathbf{x}) = 2\mathbf{x}_c = 2\mathbf{x}$$

3. $\nabla(\mathbf{x}^T A \mathbf{x}) = (A + A^T)\mathbf{x}$

分析. 变量多次出现的求导法则

$$\begin{aligned} LHS &= \nabla(\mathbf{x}_c^T A \mathbf{x}) + \nabla(\mathbf{x}^T A \mathbf{x}_c) \\ &= \nabla((A^T \mathbf{x}_c)^T \mathbf{x}) + \nabla((A \mathbf{x}_c)^T \mathbf{x}) \\ &= A^T \mathbf{x}_c + A \mathbf{x}_c \\ &= RHS \end{aligned}$$

4. $\nabla \left(\frac{1}{2} \|A\mathbf{x} - \mathbf{b}\|_2^2 \right) = A^T(A\mathbf{x} - \mathbf{b})$

分析. 法一: 展开括号, 逐一求导

$$\begin{aligned} LHS &= \frac{1}{2} (\nabla((A\mathbf{x} - \mathbf{b})^T (A\mathbf{x} - \mathbf{b}))) \\ &= \frac{1}{2} (\nabla(\mathbf{x}^T A^T A \mathbf{x}) - \nabla(\mathbf{x}^T A^T \mathbf{b}) - \nabla(\mathbf{b}^T A \mathbf{x}) + \nabla(\mathbf{b}^T \mathbf{b})) \\ &= \frac{1}{2} ((A^T A + (A^T A)^T) \mathbf{x} - A^T \mathbf{b} - A^T \mathbf{b}) \\ &= A^T(A\mathbf{x} - \mathbf{b}) \\ &= RHS \end{aligned}$$

法二: 线性变换的求导公式

$$\begin{aligned} LHS &= \frac{1}{2} A^T \nabla_{A\mathbf{x} - \mathbf{b}} \|A\mathbf{x} - \mathbf{b}\|_2^2 \\ &= A^T(A\mathbf{x} - \mathbf{b}) \\ &= RHS \end{aligned}$$

B.3 向量值函数对向量的导数

设 $\mathbf{f}: \mathbb{R}^n \mapsto \mathbb{R}^m$, 则劈形算子

$$\nabla_{\mathbf{x}} \mathbf{f} = \frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \left[\frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right]_{ij} = \left[\frac{\partial f_i}{\partial x_j} \right]$$

得到一个 $m \times n$ 的矩阵, 即为雅可比(Jacobi)矩阵

1. $\nabla_{\mathbf{x}}(A\mathbf{x}) = A$

分析. 基于此式, 由乘法法则可以推出

$$\begin{aligned}\nabla \mathbf{x}^T \mathbf{x} &= (\nabla \mathbf{x}^T) \mathbf{x} + \mathbf{x}^T (\nabla \mathbf{x}) \\ &= (\nabla(I\mathbf{x}))^T \mathbf{x} + \mathbf{x}^T (\nabla I\mathbf{x}) \\ &= I\mathbf{x} + \mathbf{x}^T I \\ &= 2\mathbf{x}\end{aligned}$$

2. $\nabla^2 f = \nabla(\nabla f)$: Hessian矩阵

分析. Hessian矩阵其实是 \mathbf{x} 到 ∇f 的 Jacobi 矩阵

$$(\nabla^2 f)_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial \mathbf{x} \partial \mathbf{x}^T} = \frac{\partial^2 f}{\partial \mathbf{x}^T \partial \mathbf{x}}$$

C 参考资料

1. *Convex optimization*, Stephen Boyd and Lieven Vandenberghe
2. 凸优化 (2018年秋季-北京大学), <http://bicmr.pku.edu.cn/~wenzw/opt-2018-fall.html>
3. 机器学习中的矩阵/向量求导, <https://zhuanlan.zhihu.com/p/25063314>
4. 矩阵求导术, <https://zhuanlan.zhihu.com/p/24709748>
5. *The Matrix Cookbook*, Kaare Brandt Petersen, Michael Syskind Pedersen