

最优化理论

陈鸿峥

2019.05*

目录

1	简介	2
1.1	优化概述	2
1.2	分类	2
1.3	历史	3
2	凸集	3
3	凸函数	6
4	凸优化问题	11
4.1	标准型	11
4.2	线性规划	13
5	对偶理论	17
6	优化算法	27
6.1	简介	27
6.2	梯度下降法	28
6.3	非光滑优化问题	33
6.4	二阶优化方法	38
6.5	有约束优化方法	40
7	大数据中的优化问题与算法	49
7.1	方差消减	50
7.2	深度神经网络	50
7.3	在线优化	51
7.4	动态优化	51

*Build 20190523

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

例 1. • 最小二乘线性拟合 (凸问题)

- 深度神经网络 (非凸, 见下)

$$\begin{aligned} \mathbf{x}_1^{(i)} &= f_1(\mathbf{x}_0^{(i)}, \mathbf{w}_1) \\ &\dots \dots \\ \mathbf{x}_n^{(i)} &= f_n(\mathbf{x}_{n-1}^{(i)}, \mathbf{w}_n) \\ \min \sum_{i=1}^m (\mathbf{y}^{(i)} - \mathbf{x}_n^{(i)})^2 \end{aligned}$$

- 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

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→低秩矩阵补全

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 $\|\mathbf{z}\|_*$

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

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

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

¹A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

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
- Kantorovich: 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
- 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

定义 1. 一些集合概念如下

- 仿射集(*affine set*)

$$\begin{aligned} C \text{ 为仿射集} &\iff \text{过} C \text{ 内任意两点的直线都在} C \text{ 内} \\ &\iff \forall x_1, x_2 \in C, \theta \in \mathbb{R}, \theta x_1 + (1 - \theta)x_2 \in C \end{aligned}$$

例 2. 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

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

- 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*)

$$\begin{aligned} C \text{ 为凸集} &\iff \text{过} C \text{ 内任意两点的线段都在} C \text{ 内} \\ &\iff \forall x_1, x_2 \in C, \theta \in [0, 1], \theta x_1 + (1 - \theta)x_2 \in C \end{aligned}$$

- 凸组合

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

- 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)

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

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 $\mathbf{a} \perp (\mathbf{x} - \mathbf{x}_0)$

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 $\langle \mathbf{x}^T, P^{-1} \mathbf{y} \rangle$ (需证满足内积条件)，进而 $\|\mathbf{x}\|_p := \sqrt{\mathbf{x}^T P \mathbf{x}}$ 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)).

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

例 3. • 空集、点、 \mathbb{R}^n 空间均为仿射

- 任意直线为仿射；若过原点则为凸锥
- \mathbb{R}^n 空间的子空间³为仿射和凸锥
- 超平面为仿射
- 半空间、欧式球、椭球、多面体为凸集

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

定义 7 (透视(perspective)函数⁴). 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

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

若 $c \in \text{dom } P$ 为凸, 则 $P(c) := \{P(x), x \in c\}$ 为凸; 反透视函数仍保持 c 的凸性。

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

分析.

$$P(\theta\mathbf{x} + (1-\theta)\mathbf{y}) = \frac{\theta\tilde{\mathbf{x}} + (1-\theta)\tilde{\mathbf{y}}}{\theta\mathbf{x}_{n+1} + (1-\theta)\mathbf{y}_{n+1}} = \frac{\theta\mathbf{x}_{n+1}}{\theta\mathbf{x}_{n+1} + (1-\theta)\mathbf{y}_{n+1}} \frac{\tilde{\mathbf{x}}}{\mathbf{x}_{n+1}} + \frac{(1-\theta)\mathbf{y}_{n+1}}{\theta\mathbf{x}_{n+1} + (1-\theta)\mathbf{y}_{n+1}} \frac{\tilde{\mathbf{y}}}{\mathbf{y}_{n+1}}$$

定义 8 (线性分数函数). 仿射函数

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

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

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

³零元、加法封闭、数乘封闭

⁴+代表 ≥ 0 , ++代表 > 0

保凸性

- 凸集的交
- 仿射、逆仿射
- 透视函数
- 线性分数函数

3 凸函数

定义 9 (凸函数). 凸函数的几种基本定义如下

1. $f: \mathbb{R}^n \rightarrow \mathbb{R}$ 为凸 $\iff \text{dom } f$ 为凸且 $\forall x, y \in \text{dom } f, \theta \in [0, 1]$

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

- 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

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

相当于每一个剖面上的低维函数都是凸的

3. 一阶条件 (first-order condition)⁵

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

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

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

- 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)

例 5. $f(\mathbf{x}) = \mathbf{a}^T \mathbf{x} + \mathbf{b}$

分析. 有 $\nabla f(\mathbf{x}) = \mathbf{a}$, 进而⁶

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

⁵ $\nabla^T f(\mathbf{x}) = [\nabla f(\mathbf{x})]^T$, $\nabla^2 f(\mathbf{x})$ 为 $f(\mathbf{x})$ 的 Hessian 矩阵

⁶ $\langle \mathbf{a}, \mathbf{x} \rangle = \mathbf{a}_1 x_1 + \cdots + \mathbf{a}_n x_n \implies \nabla_{\mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_n \end{bmatrix} = \mathbf{a}$

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

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

指示/示信(indicator)函数不一定是凸的

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

定理 1. 若 f 为凸, 可微, 则 $\exists \mathbf{x} \in \text{dom } f, \nabla f(\mathbf{x}) = 0$

例 6. 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), $\mathbf{q}^T \in \mathbb{R}^n$, $r \in \mathbb{R}$

分析. $\nabla^2 f(\mathbf{x}) = P$, $P \in \mathbb{S}_+^n$ 凸, $P \in \mathbb{S}_{++}^n$ 严格凸⁷

例 7. $f(x) = \frac{1}{x^2}$, $\text{dom } f = \{x \in \mathbb{R}, x \neq 0\}$

分析. 注意 $\text{dom } f$ 不是凸集

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

定义 11 (解析近似). 无穷阶可微

极大值函数的解析近似是 $f(x) = \log(e^{x_1} + \dots + e^{x_n})$

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

分析.

$$\frac{\partial f}{\partial x_i} = \frac{e^{x_i}}{e^{x_1} + \dots + e^{x_n}}$$

⁷由二次型理论(Taylor展开), $\nabla^2(x^T P x) = P + P^T$, 又由于 $P \in \mathbb{S}^n$, $\nabla^2 f(\mathbf{x}) = \frac{1}{2} \cdot 2P = P$, 具体推导过程可见<https://math.stackexchange.com/questions/239207/hessian-matrix-of-a-quadratic-form>

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

$$\mathbf{z} := \begin{bmatrix} e^{x_1} & \dots & e^{x_n} \end{bmatrix}^T$$

求Hessian矩阵

$$H = \frac{1}{(\mathbb{1}^T \mathbf{z})^2} (-\mathbf{z} \cdot \mathbf{z}^T + (\mathbb{1}^T \mathbf{z}) \text{diag}(\mathbf{z}))$$

将前面常量丢弃⁸

$$\mathbf{a}_i := \mathbf{v}_i \sqrt{\mathbf{z}_i} = \begin{bmatrix} a_1 & \dots & a_n \end{bmatrix}^T, b_i = \sqrt{\mathbf{z}_i}$$

$$\begin{aligned} \mathbf{v}^T H \mathbf{v} &= (\mathbb{1}^T \mathbf{z}) \mathbf{v}^T \text{diag}(\mathbf{z}) \mathbf{v} - \mathbf{v}^T \mathbf{z} \mathbf{z}^T \mathbf{v} \\ &= \left(\sum_i z_i \right) \left(\sum_i v_i^2 z_i \right) - \left(\sum_i v_i z_i \right)^2 \\ &= (\mathbf{b}^T \mathbf{b})(\mathbf{a}^T \mathbf{a}) - (\mathbf{a}^T \mathbf{b})^2 \quad \text{Cauchy} \\ &\geq 0 \end{aligned}$$

定义 12 (范数). $\|\cdot\|$ 为范数需要满足以下三个条件

1. $\|a\mathbf{x}\| = |a| \|\mathbf{x}\|$
2. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$
3. $\|\mathbf{x}\| = 0 \iff \mathbf{x} = \mathbf{0}$

零范数 $\|\mathbf{x}\|_0$: 非零元素数目, 是伪范数 (不符合第一个定义)

定理 2. \mathbb{R}^n 中的范数都是凸函数, 因而常常用来正则化!

分析.

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

例 8. 行列式的对数 $f(\mathbf{x}) = \log \det(\mathbf{x})$, $\text{dom } f = \mathbb{S}_{++}^n$, $n = 1$ 凹函数, 证明 $n > 1$ 也为凹

分析. 用高维定义

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

⁸ H 半正定, 则 $\forall \mathbf{v} \in \mathbb{R}^n: \mathbf{v}^T H \mathbf{v} \geq 0$

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

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

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

$$I + t\mathbf{z}^{-1/2}\mathbf{v}\mathbf{z}^{-1/2} = QQ^T + tQ\Lambda Q^T = Q(I + t\Lambda)Q^T$$

$$\log \det(I + t\mathbf{z}^{-1/2}\mathbf{v}\mathbf{z}^{-1/2}) = \log \det(Q) + \log \det(I + t\Lambda) + \log \det(Q^T)$$

保持函数凸性

- 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

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

$$\begin{aligned} g(\theta\mathbf{x} + (1 - \theta)\mathbf{y}) &= f(A(\theta\mathbf{x} + (1 - \theta)\mathbf{y}) + \mathbf{b}) \\ &= f(\theta(A\mathbf{x} + \mathbf{b}) + (1 - \theta)(A\mathbf{y} + \mathbf{b})) \\ &\leq \theta f(A\mathbf{x} + \mathbf{b}) + (1 - \theta)f(A\mathbf{y} + \mathbf{b}) \\ &= \theta g(\mathbf{x}) + (1 - \theta)g(\mathbf{y}) \end{aligned}$$

— 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， f_1, f_2 为凸

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

- 任意个凸函数极大值函数为凸

$$f(x) = \max\{a_1^T x + b_1, \dots, a_m^T x + b_m\}$$

- 无限个凸函数, $y \in A$, $f(x, y)$ 对于 x 为凸, 则

$$g(x) := \sup_{y \in A} f(x, y)$$

例 9. 点 x 到集合 C 的最远距离

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

位移对于范数凸性不会有影响

例 10. $x \in \mathbb{R}^n$, $x[i]$ 为第 i 大元素, $x[1] \geq x[2] \geq \dots \geq x[r] \geq \dots \geq x[n]$

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

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

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

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

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

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

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

$$f''(x) = h''(g(x))(g'(x))^2 + h'(g(x))g''(x) > 0$$

即当 g 为凸, h 为凸且不降; g 为凹, h 为凸且不增时, $f(x)$ 为凸

(若定义域非全空间) 当 g 为凸, h 为凸, 扩展值函数 \tilde{h} 不降; g 为凹, h 为凸, \tilde{h} 不增时, $f(x)$ 为凸

例 11. g 为凸, $\exp g(x)$ 为凸; g 为凹, $g > 0$, $\log g(x)$ 为凹; g 为凸, $g > 0$, $1/g(x)$ 为凸

例 12. $g(x) = x^2, \text{dom } g = \mathbb{R}, h(y) = 0, \text{dom } h = [1, 2], f = h \circ g$, 注意 \tilde{h} 并非不降!

- 函数透视: $P: \mathbb{R}^{n+1} \mapsto \mathbb{R}^n, \text{dom } P \in \mathbb{R}^n \times \mathbb{R}_{++}, P(z, t) = \frac{z}{t}$

$$f: \mathbb{R}^n \mapsto \mathbb{R}, g(x, t) = tf\left(\frac{x}{t}\right), \text{dom } g = \{(x, t) \mid \frac{x}{t} \in \text{dom } f\}, g: \mathbb{R}^n \times \mathbb{R}_{++} \mapsto \mathbb{R}$$

若 $f(x)$ 为凸, 则 $g(x, t)$ 相对于 (x, t) 联合凸

例 13. – $f(x) = x^T x, g(x, t) = x^T x/t$

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

– $u, v \in \mathbb{R}_{++}^n, g(u, v) = \sum_{i=1}^n u_i \log(u_i/v_i)$, 信息论常用, 衡量相似性, KL 散度

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

定义 13 (α 次水平集(α -sub level set)). $f : \mathbb{R}^n \mapsto \mathbb{R}$, $C_\alpha = \{x \in \text{dom } f \mid f(x) \leq \alpha\}$

定义 14 (拟凸函数(quasi-convex)). 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

凸函数与凸集联系

- 凸函数定义域为凸集
- 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 $h_j : \mathbb{R}^n \mapsto \mathbb{R}$
- 域 $\mathcal{D} = \bigcap_{i=1}^m \text{dom } f_i \cap \bigcap_{j=1}^p \text{dom } h_j$
- 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 $P^* = \inf\{f_0(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\}$
- 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 $\mathcal{X}^* = \{\mathbf{x}^* \mid f_0(\mathbf{x}^*) = P^*, \mathbf{x}^* \in \mathcal{X}\}$
- 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 $\exists R > 0, f_0(x) = \inf\{f_0(\mathbf{z}) \mid \mathbf{x} \in \mathcal{X}, \mathbf{z} \in \mathcal{X}, \|\mathbf{x} - \mathbf{z}\| \leq R\}$
- 局部最优解集 $\mathcal{X}_{local} = \{\mathbf{x} \mid \mathbf{x} \text{为局部最优}\}$

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

例 14.

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

定理 3. 凸问题局部最优等价于全局最优

分析. 若 x 为局部最优

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

反证法, 设 x 不是全局最优, y 为全局最优, $f_0(x) > f_0(y)$

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

$$\|z - x\|_2 = \frac{R\|y - x\|_2}{2\|y - x\|_2} = \frac{R}{2}$$

由 $\|z - x\|_2 \leq R \implies f_0(x) \leq f_0(z)$, 有

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

矛盾

可微凸目标函数

无约束 $\min f_0(x), \nabla f_0^*(x) = 0$

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

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

有约束 $\min f_0(x), s.t. x \in \mathcal{X}$

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

例 15. 等式约束 $\min f_0(x), \text{dom } f_0 \subset \mathbb{R}^n, f_0$ 可微, 使得 $Ax = b$

分析. x^* 最优, $Ax^* = b, \forall y \in \mathcal{X}, Ay = b$

$$\langle \nabla f_0(x^*), y - x^* \rangle \geq 0$$

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

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

1. $\text{Nul } A = \{0\}$

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

例 16. 正约束 $\min f_0(x), s.t. x \geq 0$

分析. 若 x^* 最优, $\iff x^* \geq 0, \forall y \geq 0, \langle \nabla f_0(x^*), y - x^* \rangle \geq 0$

$$\iff \langle \nabla f_0(x^*), y \rangle \geq \langle \nabla f_0(x^*), x^* \rangle$$

1. 若 $\nabla f_0(x^*) \not\geq 0$ 有矛盾 (负数行乘上正无穷), 故 $\nabla f_0(x^*) \geq 0$

2. 令 $y = 0$, 有 $0 \geq \langle \nabla f_0(x^*), x^* \rangle \implies \sum_{i=1}^n n(\nabla f_0(x^*)_i x^*) \leq 0$
 前面 ≥ 0 , 进而互补松弛条件
3. $x^* \geq 0$

4.2 线性规划

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

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

\mathbf{s} 为松弛变量(slack variable)

用 \mathbf{x}^+ 和 \mathbf{x}^- 拆分, 得到 $\mathbf{x} = \mathbf{x}^+ - \mathbf{x}^-$, $\mathbf{x}^+ \geq 0, \mathbf{x}^- \geq 0, \mathbf{s} \geq 0$

例 17 (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 $i = 1, \dots, m, j = 1, \dots, n$

线性分数规划

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

等价于

$$\begin{aligned}
 \min \quad & c^T y + dz \\
 \text{s.t.} \quad & Gy - hz \leq 0 \\
 & Ay - bz = 0 \\
 & e^T y + fz = 1 \\
 & z \geq 0
 \end{aligned}$$

分析. 证明两个问题等价, P_0 与 P_1

若 x 在 P_0 内可行

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

若 (y, z) 在 P_1 中可行

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

若 $z = 0$, x_0 为 P_0 的可行解

$$\begin{aligned}
 x &= x_0 + ty, t \geq 0 \\
 \lim_{t \rightarrow \infty} \frac{c^T(x_0 + ty) + d}{e^T(x_0 + ty) + f} &= c^T y
 \end{aligned}$$

代入看所有条件结论都相同

二次规划(Quadratic Programming)

$$\begin{aligned}
 \min \quad & \frac{1}{2} x^T p x + q^T x + r, \quad p \succ 0 \\
 \text{s.t.} \quad & Gx \leq h \\
 & Ax = b
 \end{aligned}$$

二次约束二次规划(QCQP)

$$\begin{aligned}
 \min \quad & \frac{1}{2} p_0 x + q_0^T x + r_0, \quad p \succ 0 \\
 \text{s.t.} \quad & \frac{1}{2} x^T p_i x + q_i^T x + r_i \leq 0, \quad i = 1, \dots, m, p_i \succ 0 \\
 & Ax = b
 \end{aligned}$$

最小二乘问题

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

$$\frac{1}{2} (x^T A^T A x - 2b^T A x + b^T b)$$

一范数规范化最小二乘

$$\min \frac{1}{2} \|Ax - b\|_2^2 + \lambda_1 \|x\|_1$$

本来用零范数，但用一范数拟合

改写

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

Basic Pursuit

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

原式很难平衡两者，下式只需考虑 $\|x\|_1$ 的影响

岭回归(Ridge): 所有 x 差距不要太大

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

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

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

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

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

$$\begin{aligned} \min \quad & x^T \Sigma x \\ \text{s.t.} \quad & p^T x \geq r_{\min} \\ & x_1 + \dots + x_n = B \\ & x_1, \dots, x_n \geq 0 \end{aligned}$$

半定规划(semi-definite programming, SDP) (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 18 (谱范数极小化问题). 矩阵多项式 $A(x) = A_0 + x_1 A_1 + \dots + x_n A_n, A_i \in \mathbb{R}^{p \times q}$

$$\min_x \|A(x)\|_2$$

谱范数代表 $A(x)$ 的最大奇异值⁹

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

例 19 (最快分布式线性平均).

$$\begin{aligned} x(t) &= Px(t-1) \\ P &= \begin{bmatrix} P_{11} & \dots & P_{1n} \\ \vdots & & \vdots \\ P_{n1} & \dots & P_{nn} \end{bmatrix}, P\mathbb{1} = \mathbb{1} \end{aligned}$$

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

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

$$P \succeq 0, P_{ij} \geq 0$$

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

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

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

$$\begin{aligned} \min \max\{|\lambda_2|, |\lambda_n|\} \\ \max\{|\lambda_2|, |\lambda_n|\} = \left\| P - \frac{1}{n} \mathbb{1} \mathbb{1}^T \right\| \end{aligned}$$

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

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

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

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

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

拉格朗日函数(Lagrangian function)

$$L(x, \lambda, v) = f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p v_i h_i(x), \text{dom } L = D \times \mathbb{R}^m \times \mathbb{R}^p$$

拉格朗日乘子(multiplier)

- 原变量(primal variable): $\lambda = [\lambda_1 \quad \dots \quad \lambda_m]^T$
- 对偶变量(dual variable): $v = [v_1 \quad \dots \quad v_p]^T$

拉格朗日对偶函数

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

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}

- $g(\lambda, v)$ 一定是关于 λ 和 v 的凹函数（关于 λ 和 v 的仿射函数，注意 x 为常数）

- $\forall \lambda \geq 0, \forall v, g(\lambda, v) \leq P^*$

对偶(dual)问题

$$\max \quad g(\lambda, v)$$

$$\text{s.t.} \quad \lambda \geq 0$$

其最优解记为 D^* ，则 $D^* \leq P^*$ ，即给出了原问题的一个最优下界

x^* 原问题最优解

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

$$L(x^*, \lambda, v) = f_0(x^*) + (\dots) \leq P^*$$

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

例 20.

$$\min \quad x^T x$$

$$\text{s.t.} \quad Ax = b$$

分析.

$$L(x, v) = x^T x + v^T (Ax - b)$$

$$\begin{aligned} g(v) &= \inf_{x \in \mathcal{D}} L(x, v) \\ &= \inf_{x \in \mathcal{D}} x^T x + v^T Ax - v^T b \\ &= \left(-\frac{A^T v}{2}\right)^T \left(-\frac{A^T v}{2}\right) + v^T A \left(-\frac{A^T v}{2}\right) - v^T b \\ &= -\frac{1}{4} v^T A A^T v - b^T v \end{aligned}$$

补充求梯度： $2x + A^T v = 0 \implies x = -\frac{A^T v}{2}$

因而得到对偶问题

$$\max_v \quad -\frac{1}{4} v^T A A^T v - b^T v$$

例 21.

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

分析. 注意 λ 前面符号, 要化为一般形式

$$L(x, \lambda, v) = c^T x - \lambda^T x + v^T (Ax - b)$$

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

对偶问题, 由于要极大, 故不考虑负无穷部分

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

逆过来求解

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

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

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

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

对偶的对偶不一定回去，线性规划才满足

例 22.

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

分析.

$$L(x, v) = x^T w x + \sum_{i=1}^n v_i (x_i^2 - 1)$$

$$\begin{aligned} g(v) &= \inf_x L(x, v) \\ &= \inf_x x^T w x + \sum_{i=1}^n v_i x_i^2 - \sum_{i=1}^n v_i \\ &= \inf_x x^T (w + \text{diag}(v)) x - \mathbb{1}^T v = \begin{cases} -\mathbb{1}^T v & w + \text{diag}(v) \succeq 0 \\ -\infty & \text{otherwise} \end{cases} \end{aligned}$$

补充求梯度： $2(w + \text{diag}(v))x = 0$

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

定义 15 (函数的共轭). $f: \mathbb{R}^n \mapsto \mathbb{R}, f^*(y) = \sup_{x \in \text{dom } f} (y^T x - f(x))$, 几何意义即到不同斜率直线的距离最大值

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

$$\begin{aligned} L(x, \lambda, v) &= f_0(x) + \lambda^T (Ax - b) + v^T (cx - d) \\ &= f_0(x) + (A^T \lambda + c^T v)^T x - \lambda^T b - v^T d \\ g(\lambda, v) &= \inf_x f_0(x) + (A^T \lambda + c^T v)^T x - \lambda^T b - v^T d \\ &= -\sup_x -(A^T \lambda + c^T v)^T x - f_0 \\ &= -f_0^*(-(A^T \lambda + c^T v)) - \lambda^T b - v^T d \end{aligned}$$

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slater条件满足, $p^* = d^*$

定义 16 (相对内点(relative interior)).

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

定理 4 (Slater条件).

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

$\exists x \in relint D$ 使得 $f_i(x) < 0, i = 1, \dots, m, Ax = b$

例 23. 二次规划(QP)

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

Slater条件 $\{x \mid Ax = b\}$ 非空

例 24. 二次约束二次规划(QCQP)

$$\begin{aligned} \min \quad & \frac{1}{2} x^T P_0 x + q_0^T x + r_0 \\ \text{s.t.} \quad & \frac{1}{2} x^T P_i x + q_i^T x + r_i \leq 0, \quad i = 1, \dots, m \end{aligned}$$

P_0, \dots, P_i 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

例 25.

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

分析.

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

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

$$\implies p^* = d^* = 0$$

置信域问题

$$\begin{aligned} \min \quad & x^T A x + b^T x \\ \text{s.t.} \quad & x^T x \leq 1 \\ & A \not\equiv 0 \end{aligned}$$

依然可以得到 $p^* = d^*$
几何解释

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

$$G = \{(f_1(x), f_0(x)) \mid x \in \mathcal{D}\}$$

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

$$L(x, \lambda) = f_0(x) + \lambda f_1(x)$$

$$g(\lambda) = \inf_{x \in \mathcal{D}} \{f_0(x) + \lambda f_1(x)\}$$

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

$$\lambda \geq 0, \max g(\lambda)$$

注意问题必须要有可行解

经济学解释：满足原材料约束下，利润最多价格 $\lambda_i \geq 0$

$$g(\lambda) = \inf_x f(x) + \lambda_1 f_1(x) + \dots + \lambda_m f_m(x) = \inf_x L(x, \lambda)$$

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

$$d^* = \sup_{\lambda \geq 0} g(\lambda)$$

市场平衡点，均衡市场 $p^* = d^*$ ，最优/影子价格 λ^*

多目标优化解释

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

$$\min_x f_0(x) + \lambda_1 f_1(x) + \cdots + \lambda_m f_m(x)$$

鞍点(saddle point)解释

$$f(w, z), w \in S_w, z \in S_z$$

极小极大不等式

$$\sup_{z \in S_z} \inf_{w \in S_w} f(w, z) \leq \inf_{w \in S_w} \sup_{z \in S_z} f(w, z)$$

若有 (\tilde{w}, \tilde{z}) 使得

$$(\tilde{w}, \tilde{z}) = \arg \max_{z \in S_z} \min_{w \in S_w} f(w, z)(\tilde{w}, \tilde{z}) = \arg \min_{w \in S_w} \max_{z \in S_z} f(w, z)$$

则 (\tilde{w}, \tilde{z}) 为鞍点

有下面不等式成立

$$f((\tilde{w}, z)) \leq f(\tilde{w}, \tilde{z}) \leq f(w, \tilde{z}), \forall z \in S_z, w \in S_w$$

即从一个方向望过去是最小，从另一个方向望过去是最大

$$\begin{aligned} L(x, \lambda) &= f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) \\ \implies \sup_{\lambda \geq 0} L(x, \lambda) &= \sup_{\lambda \geq 0} \{f_0(x) + \sum_{i=1}^m \lambda_i f_i(x)\} \\ &= \begin{cases} f_0(x) & f_i(x) \leq 0, i = 1, \dots, m \\ +\infty & \text{otherwise} \end{cases} \\ \implies p^* &= \inf_x \{f_0(x) \mid f_i(x) \leq 0, i = 1, \dots, m\} = \inf_x \sup_{\lambda \geq 0} L(x, \lambda) \\ d^* &= \sup_{\lambda \geq 0} g(\lambda) = \sup_{\lambda \geq 0} \inf_x L(x, \lambda) \implies p^* \geq d^* \end{aligned}$$

如果 $L(x, \lambda)$ 有鞍点，则必有 $p^* = d^*$

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则是非常好的点

若 $(\tilde{x}, \tilde{\lambda})$ 为 $L(x, \lambda)$ 鞍点 $\iff p^* = d^*$ 且 $\tilde{x}, \tilde{\lambda}$ 为原对偶问题最优解 \implies 若为鞍点， $p^* = d^*$

$$\sup_{\lambda \geq 0} \inf_x L(x, \lambda) = \inf_x \sup_{\lambda \geq 0} L(x, \lambda)$$

已知 $(\tilde{x}, \tilde{\lambda})$ 为左边最优

$$\tilde{\lambda} = \arg \max_{\lambda \geq 0} \inf_x L(x, \lambda)$$

$$\tilde{x} = \arg \inf_x \sup_{\lambda \geq 0} L(x, \lambda)$$

则 $\tilde{\lambda}$ 对偶最优, \tilde{x} 为原问题最优

$$\min f_0(x)$$

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

定理 5. $(\tilde{x}, \tilde{\lambda})$ 为拉格朗日函数鞍点 $\iff p^* = d^*$, 且 $(\tilde{x}, \tilde{\lambda})$ 为原对偶的最优解

分析. 右推左, $(\tilde{x}, \tilde{\lambda})$ 原对偶可行

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

因 $p^* = d^*$, 有

$$\begin{aligned} f_0(\tilde{x}) &= g(\tilde{\lambda}) \\ &= \inf_x \{f_0(x) + \sum_{i=1}^m \tilde{\lambda}_i f_i(x)\} \\ &\leq f_0(\tilde{x}) + \sum_{i=1}^m \tilde{\lambda}_i f_i(\tilde{x}) \\ &\leq f_0(\tilde{x}) \end{aligned}$$

进而不等号都得为等号

$$1. \inf_x L(x, \tilde{\lambda}) = L(\tilde{x}, \tilde{\lambda})$$

$$2. f_0(\tilde{x}) = \sup_{\lambda \geq 0} \{f_0(\tilde{x}) + \sum_{i=1}^m \lambda_i f_i(\tilde{x})\} = \sup_{\lambda \geq 0} L(\tilde{x}, \lambda)$$

$$\implies L(\tilde{x}, \tilde{\lambda}) = \sup_{\lambda \geq 0} L(\tilde{x}, \lambda)$$

$$\implies (\tilde{x}, \tilde{\lambda}) \text{ 是 } L(x, \lambda) \text{ 的鞍点}$$

一般优化问题的对偶理论

$$\min f_0(x)$$

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

$$h_i(x) = 0, \quad i = 1, \dots, p$$

不一定是凸问题, 但 $p^* = d^*$, 最优解满足什么条件?

对偶问题

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

分析. 设 (x^*, λ^*, v^*) 为原对偶最优解, 则 (x^*, λ^*, v^*) 为原对偶可行解

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

$$\begin{aligned} p^* = d^* &\implies f_0(x^*) = g(\lambda^*, v^*) \\ &= \inf_x \{f_0(x) + \sum_{i=1}^m \lambda_i^* f_i(x) + \sum_{i=1}^p v_i^* h_i(x)\} \\ &\leq f_0(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*) + \sum_{i=1}^p v_i^* h_i(x^*) \\ &\leq f_0(x^*) \end{aligned}$$

同上理, 不等号全取等

$$1. \lambda_i^* f_i(x^*) = 0, \forall i = 1, \dots, m$$

$$2. x^* = \arg \min_x L(x, \lambda^*, v^*)$$

若 f_0, f_i, h_i 均可微, 则必要条件为

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

可微优化问题的KKT(Karush-Kuhn-Tucker)条件

- $f_i(x^*) \leq 0, i = 1, \dots, m$ primal feasibility
- $h_i(x^*) = 0, i = 1, \dots, p$ primal feasibility
- $\lambda^* \geq 0$ dual feasibility
- $\lambda_i^* f_i(x^*) = 0, i = 1, \dots, m$ complementarity slackness(对偶互斥条件)
- $\left. \frac{\partial L(x, \lambda^*, v^*)}{\partial x} \right|_{x=x^*} = 0$ stability

定理 6. 若原问题为凸, 则KKT条件为充要条件

分析. 必要性已证, 证明充分性

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

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

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

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

例 26 (Waterfilling 算法). 共 n 个信道 (*channel*)

$$source \longleftrightarrow destination$$

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

分析. *KKT* 条件

- $x^* \geq 0$
- $\mathbb{1}^T x^* = 1$
- $\lambda^* \geq 0$
- $x_i^* \lambda_i^* = 0, \forall i$

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

$$\left(\frac{\partial L(x, \lambda, v)}{\partial x} \right)_i = - \frac{1}{\alpha_i + x_i} - \lambda_i + v$$

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

$$\implies v^* \frac{1}{\alpha_i + x_i^*}, i = 1, \dots, n$$

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

$$\text{若 } v^* > \frac{1}{\alpha_i} \implies x_i^* = 0$$

$$\text{若 } v^* < \frac{1}{\alpha_i}$$

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

进而

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

结合 $\sum_i x_i^* = 1$, 即注水算法

Motivation: 误差, 调整参数测灵敏度

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

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

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

布尔线性规划问题做松弛(relaxation)

$$x_i \in \{0, 1\} \implies 1 \geq x_i \geq 0$$

6 优化算法

6.1 简介

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

罚函数法

$$\begin{aligned} \min \quad & f_0(x) + \frac{\lambda}{2} \|A\mathbf{x} - \mathbf{b}\|_2^2 \\ \tilde{x} &= \arg \min_x F \\ \nabla f_0(\tilde{\lambda}) + \lambda A^T(A\tilde{\mathbf{x}} - \mathbf{b}) &= 0 \end{aligned}$$

$$\begin{aligned}
L(x, v) &= f_0(x) + v^T(A\mathbf{x} - \mathbf{b}) \\
\Rightarrow g(v) &= \inf_x f_0(x) + v^T(A\mathbf{x} - \mathbf{b}) \\
v &= \lambda(A\tilde{\mathbf{x}} - \mathbf{b}) \\
\Rightarrow g(\lambda(A\tilde{\mathbf{x}} - \mathbf{b})) &= \inf_x f_0(x) + \lambda(A\tilde{\mathbf{x}} - \mathbf{b})^T(A\mathbf{x} - \mathbf{b}) \\
\nabla f_0(x) + \lambda A^T(A\tilde{\mathbf{x}} - \mathbf{b}) &= 0
\end{aligned}$$

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

log-barrier

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

$\min f_0(x)$ 可微, 凸, 无约束

1. 所有算法都是迭代的

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

$\alpha \geq 0$ 为步长, 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

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

3. 关键问题是选方向

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!

不精确线搜索(Armijo Rule): 一阶泰勒展开

实际上没有必要要求最优步长, 在该方向上的差异并没有太大

6.2 梯度下降法

$$d^{(k)} = -\nabla f_0(x^{(k)})$$

- 能否收敛
- 收敛到哪里
- 收敛速度

假设

0. 基本假设: f 为可微的凸函数,

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

存在且有限, $f_0(x^*)$ 有限

1. Lipschitz 连续梯度

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

等价定义:

a. 若 $f_0(x)$ 二阶可微

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

b. 下界

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

c. 上界

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

d. 当函数为凸时

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

2. 强凸性(strong convexity)

$$\exists \mu > 0: f_0(y) \geq f_0(x) + \langle \nabla f_0(x), y - x \rangle + \frac{\mu}{2} \|x - y\|_2^2, \forall x, y$$

二阶可微情况下的等价定义

$$\nabla^2 f(x) \succeq \mu I$$

例 27.

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

区别于严格凸(strictly convex), 强凸一定是严格凸

定理 7. 严格凸函数只有一个最小值点

分析. 反证法, 假设 x, y 均为最小值点, 且 $x \neq y$

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

定理 8. 若 $f_0(x)$ 有 *Lipschitz* 连续梯度, 常数 L , 若 $\alpha \in (0, \frac{2}{L})$, 则有

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

即以 $O(\frac{1}{k})$ 速度收敛

分析. 1° 点的单调性: 与任意 x^* 的距离在缩小

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

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

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

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

3° 函数值的充分下降 (即证明收敛性)

$$\begin{aligned} f_0(x^{(k+1)}) - f_0(x^*) &\leq f_0(x^{(k)}) - f_0(x^*) - \omega \|\nabla f_0(x^{(k)})\|^2 \\ f_0(x^{(k)}) - f_0(x^*) &\leq \langle f_0(x^{(k)}), x^{(k)} - x^* \rangle \\ &= \langle \nabla f_0(x^{(k)}) - \nabla f_0(x^*), x^{(k)} - x^* \rangle \\ &\leq \|\nabla f_0(x^{(k)}) - \nabla f_0(x^*)\| \|x^{(k)} - x^*\| \\ &\leq \|\nabla f_0(x^{(k)})\| \|x^{(k)} - x^*\| \\ \Delta^{(k+1)} &\leq \Delta^{(k)} - \frac{\omega}{\|x^0 - x^*\|^2} (\Delta^{(k)})^2 \\ \frac{1}{\Delta^{(k+1)}} &\leq \frac{1}{\Delta^{(k)}} - \frac{\omega}{\|x^0 - x^*\|^2} \frac{\Delta^{(k)}}{\Delta^{(k+1)}} \end{aligned}$$

错位相消可得结论 $O(\frac{1}{k})$ 收敛速度

定理 9. 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

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

分析.

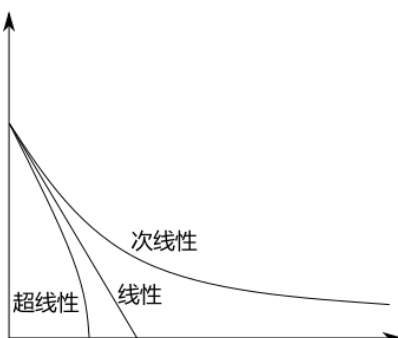
$$\begin{aligned}
 \|x^{(k)} - x^*\|^2 &= \|x^{(k)} - \alpha \nabla f_0(x^{(k)}) - x^*\|^2 \\
 &= \|x^{(k)} - x^*\|^2 - 2\alpha \langle x^{(k)} - x^*, \nabla f_0(x^{(k)}) \rangle + \alpha^2 \|\nabla f_0(x^{(k)})\|^2 \\
 &\leq \|x^{(k)} - x^*\|^2 - \frac{2\alpha}{\mu + L} \|\nabla f_0(x^{(k)})\|^2 + \alpha^2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

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

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

不同收敛速度

- 次线性收敛
- 线性收敛
- 超线性收敛



例 28.

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

分析.

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

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

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

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

可以通过预处理(precondition)来解决条件数糟糕的问题

f_0 , Lipschitz连续梯度(L), 强凸(μ), 函数值收敛性

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

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

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

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

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

Armijo Rule

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

$$\tilde{f}_0(\alpha^{(k)}) = f_0(x^{(k+1)}) \leq f_0(x^{(k)}) - \gamma \alpha^{(k)} \left\| \nabla f_0(x^{(k)}) \right\|^2$$

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

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

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

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

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

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

梯度下降法的解释1

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

将 f_0 在 $x^{(k)}$ 处进行一阶Taylor展开

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

求梯度

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

解释2

$$\begin{aligned}
f_0(x^{(k)} + v) &\approx f_0(x^{(k)}) + \langle \nabla f_0(x^{(k)}), v \rangle \\
d^{(k)} &= \arg \min_v \{ \langle \nabla f_0(x^{(k)}), v \rangle \mid \|v\| = 1 \}
\end{aligned}$$

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

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

通过改变不同的范数，有不同的特性

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

$$d^{(k)} = \mathbf{e}_{\text{mod}(k, n)}$$

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

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

6.3 非光滑优化问题

6.3.1 次梯度法

$$\min f_0(x), \quad f_0 \text{连续, 凸, 不可微}$$

梯度下降法→次梯度(subgradient)法

$g_0(x) \in \partial f_0(x)$ (注意凹函数则对应的是supgradient)

$$f_0(y) \geq f_0(x) + \langle g_0(x), y - x \rangle, \forall y$$

$f(x) = |x|$ 在零点处次梯度为 $[-1, 1]$

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

只要有 $0 \in \partial f_0(x_0)$ 就有最优解 $x = x_0$

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
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 $\frac{1}{\sqrt{k}}$

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

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

$$\bar{x}^{(k)} := \frac{\sum_{i=0}^k \alpha^{(i)} x^{(i)}}{\sum_{i=0}^k \alpha^{(i)}}$$

$$f_0(\bar{x}^{(k)}) - f_0(x^*)$$

假设函数Lipschitz连续

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

$\forall x^*$ 最优

$$\begin{aligned} & \|x^{(k+1)} - x^*\|^2 = \|x^{(k)} - \alpha^{(k)} g_0(x^{(k)}) - x^*\|^2 \\ &= \|x^{(k)} - x^*\|^2 - 2\langle \alpha^{(k)} g_0(x^{(k)}), x^{(k)} - x^* \rangle + (\alpha^{(k)})^2 \|g_0(x^{(k)})\|^2 \\ &\leq \|x^{(k)} - x^*\|^2 - 2(\alpha^{(k)})^2 (f_0(x^{(k)}) - f_0(x^*)) + (\alpha^{(k)})^2 G^2 \\ &\implies \|x^{(k+1)} - x^*\|^2 \leq \|x^{(0)} - x^*\|^2 - 2 \sum_{i=0}^k \alpha^{(i)} (f_0(x^{(i)}) - f_0(x^*)) + \sum_{i=0}^k (\alpha^{(i)})^2 G^2 \\ &\implies 2 \sum_{i=0}^k \alpha^{(i)} (f_0(x^{(i)}) - f_0(x^*)) \leq \|x^{(0)} - x^*\|^2 + \sum_{i=0}^k (\alpha^{(i)})^2 G^2 \end{aligned}$$

$$\begin{aligned} \sum_{i=0}^k \alpha^{(i)} (f_0(x^{(i)}) - f_0(x^*)) &\geq \left(\sum_{i=0}^k \alpha^{(i)} \right) \inf_{i=0, \dots, k} (f_0(x^{(i)}) - f_0(x^*)) \\ \implies \inf_{i=0, \dots, k} (f_0(x^{(i)}) - f_0(x^*)) &\leq \frac{\|x^{(0)} - x^*\|^2 + G^2 \sum_{i=0}^k (\alpha^{(i)})^2}{2 \sum_{i=0}^k \alpha^{(i)}} \end{aligned}$$

这是一个紧的界

- 固定步长得到上界 $\frac{G^2 \alpha}{2}$, 以 $f(x) = |x|$ 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 $O(\frac{\log k}{\sqrt{k}})$, 可以证明在该假设下该收敛速度最优

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

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

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

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

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

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

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 x 有界的

6.3.2 邻近点梯度法(proximal gradient method)

有结构, 不可微

$$\min f_0(x) = s(x) + r(x)$$

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

$$r(x), \hat{x} \cdot \alpha > 0$$

邻近点投影(proximal mapping)

$$\min_x r(x) + \frac{1}{2\alpha} \|x - \hat{x}\|^2$$

例 29 (LASSO).

$$f_0(x) = \frac{1}{2} \|y - Ax\|^2 + \lambda \|x\|_1$$

分析. 本来想优化 l_2 -范数, 但因为不好做, 故变为优化 l_1 -范数

$$\arg \min \lambda \|x\|_1 + \frac{1}{2\alpha} \|x - \hat{x}\|^2$$

Hessian矩阵是单位阵, 好解得多

$$\lambda \sum_{i=1}^n \|x_i\| + \frac{1}{2\alpha} \sum_{i=1}^n (x_i - \hat{x}_i)^2$$

由于每一维并没有耦合在一起, 因此相当于每一个维度都最优化

$$\arg \min \lambda |x_i| + \frac{1}{2\alpha} (x_i - \hat{x}_i)^2$$

$$0 \in \lambda \partial |x_i| + \frac{1}{\alpha} (x_i - \hat{x}_i)$$

1. 若 $x_i > 0$, 则

$$\begin{aligned} 0 &= \lambda + \frac{1}{\alpha} (x_i - \hat{x}_i) \\ \implies x_i &= \hat{x}_i - \alpha\lambda, \hat{x}_i > \alpha\lambda \end{aligned}$$

2. 若 $x_i < 0$, 则

$$\begin{aligned} 0 &= -\lambda + \frac{1}{\alpha} (x_i - \hat{x}_i) \\ \implies x_i &= \hat{x}_i + \alpha\lambda, \hat{x}_i < \alpha\lambda \end{aligned}$$

3. 若 $x_i = 0$, 则

$$\begin{aligned} 0 \in [-\lambda, \lambda] - \frac{1}{\alpha} \hat{x}_i &\implies 0 \in [-\lambda - \frac{1}{\alpha} \hat{x}_i, \lambda - \frac{1}{\alpha} \hat{x}_i] \\ 0 \geq -\lambda - \frac{1}{\alpha} \hat{x}_i &\implies \hat{x}_i \geq -\lambda\alpha, \hat{x}_i \leq \lambda\alpha \end{aligned}$$

画软门限(*soft thresholding*)图, 横轴 \hat{x}_i , 纵轴 x_i

- $x^{(k+\frac{1}{2})} = x^{(k)} - \alpha A^T (Ax^{(k)} - y)$
- $x^{(k)} = \arg \min_x \lambda \|x\|_1 + \frac{1}{2\alpha} \|x - x^{(\frac{1}{2})}\|^2$

即ISTA算法

例 30 (盒限制(box constrained)优化问题).

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

分析.

$$\begin{aligned} \arg \min r(x) + \frac{1}{2\alpha} \|x - \hat{x}\|^2 \\ = \arg \min \frac{1}{2\alpha} \|x - \hat{x}\|^2 \\ s.t. \forall i, x_i \in [l_i, u_i] \end{aligned}$$

如果有约束 $x \in \mathcal{C}$

- $x^{(k+\frac{1}{2})} = x^{(k)} - \alpha \nabla S(x^{(k)})$
- $x^{(k+1)} = \arg \min_x I_{\mathcal{C}}(x^{(k+\frac{1}{2})}) + \frac{1}{2\alpha} \|x - x^{(k+\frac{1}{2})}\|^2 = \arg \min_x \frac{1}{2\alpha} \|x - x^{(k+\frac{1}{2})}\|^2, x \in \mathcal{C}$
- 相当于做投影，故称投影梯度法

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

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）

收敛性能与梯度下降法类似

例 31 (矩阵补全). $Y \in \mathbb{R}^{m \times n}, \{Y_{ij}, (i, j) \in \Omega\}$

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

分析. 同LASSO，由于矩阵的秩（奇异值向量 l -范数）不好求，改为矩阵的和范数 $\|\cdot\|_*$ （奇异值向量 1 -范数），即

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

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

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

其中 P_{Ω} 为0，若原矩阵中该项不存在；存在的话则保持不变

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

对 B 做奇异值分解, U 为酉矩阵

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

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

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

$$0 \in [\lambda UDV^T + \frac{1}{\alpha} (B - B^{(k+\frac{1}{2})})]$$

$$B = U\Sigma V^T, d = \partial \|\sigma\|_1$$

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

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

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

对 $B^{(k+\frac{1}{2})}$ 进行奇异值分解

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

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

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

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

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

该算法称为矩阵软门限(*soft thresholding*)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 $f_0(x)$ 二阶可微, 强凸

$$\min f_0(x^{(k)} + v) \approx \min_{\|v\|=1} f_0(x^{(k)}) + \langle \nabla f_0(x^{(k)}), v \rangle$$

$$\begin{aligned}
&\approx \min_v f_0(x^{(k)}) + \langle \nabla f_0(x^{(k)}), v \rangle + \frac{1}{2} v^T \nabla^2 f_0(x^{(k)}) v \\
&\implies \nabla f_0(x^{(k)}) + \nabla^2 f_0(x^{(k)}) v = 0 \\
&\implies v = -(\nabla^2 f_0(x^{(k)}))^{-1} \nabla f_0(x^{(k)}) \rightarrow \text{牛顿方向}
\end{aligned}$$

$$\begin{aligned}
d^{(k)} &= -(\nabla^2 f_0(x^{(k)}))^{-1} \nabla f_0(x^{(k)}) \\
x^{(k+1)} &= x^{(k)} - \alpha^{(k)} (\nabla^2 f_0(x^{(k)}))^{-1} \nabla f_0(x^{(k)})
\end{aligned}$$

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

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

假设 $\nabla^2 f(x)$ Lipschitz连续

- 若 $\|\nabla f_0(x^{(k)})\|_2 > \eta$, 阻尼(damped)牛顿段
用Armijo Rule算步长, $\exists \gamma > 0, f_0(x^{(k+1)}) - f_0(x^{(k)}) \leq -\gamma$
- 若 $\|\nabla f_0(x^{(k)})\|_2 \leq \eta$, 纯牛顿段
 $\alpha = 1, f_0(x^{(k+1)}) - f_0(x^*) \leq \Delta(\frac{1}{2})^{2^k}, \exists \Delta > 0$, 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 32. $\min \frac{1}{2} x^T P x + q^T r + c, P \succ 0$

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

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

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

在 $x = k + 1$ 点处做Taylor展开

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

$$\begin{aligned}q^{(k)} &= \nabla f_0(x^{(k+1)}) - \nabla f_0(x^{(k)}) \\p^{(k)} &= x^{(k+1)} - x^{(k)}\end{aligned}$$

$$\begin{cases} q^{(k)} = B^{(k+1)} p^{(k)} \\ p^{(k)} = G^{(k+1)} q^{(k)} \end{cases}$$

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）

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

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

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)

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

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)

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

- 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

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

变为

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

本质上都是在考虑它的KKT条件

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

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

等价于KKT条件

$$\begin{cases} Ax^* = b \\ Px^* + q + A^T v^* = 0 \end{cases}$$

等价于

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

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

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

近似等价于（二阶近似Taylor展开）

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

写出问题关于d的KKT条件，可得等价条件

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

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

6.5.2 拉格朗日乘子法/对偶分解法

$$L(x, v) = f_0(x) + \langle Ax - b, v \rangle$$

更新原变量和对偶变量

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

$\alpha^{(k)}$ 可以是固定步长，也可以是递减步长

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

例 33.

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

分析.

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

对偶次梯度法： v 才是最关键的，只是在寻找最优 v 的时候顺带找到了 x （收敛到 v^* 的同时也找到了 x^* ）

$$D(v) = \inf_x L(x, v)$$

$D(v)$ 为凹函数，关注 $-D(v)$

$$\begin{cases} -(Ax^{(k+1)} - b) \\ x^{(k+1)} = \arg \min_x L(x, v^{(k)}) \end{cases}$$

$$v^{(k+1)} = v^{(k)} - \alpha^{(k)}(-(Ax^{(k+1)}) - b)$$

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

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

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

$$-D(v) \geq -D(\hat{\lambda}) + \langle v - \hat{\lambda}, -(A\hat{x} - b) \rangle$$

进而得到 $-(A\hat{x} - b)$ 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 $D(v)$ 何时可微？

任何 $-D(\lambda)$ 都具有 $-(A\hat{x} - b)$ 的形式，得到当 $f_0(x)$ 严格凸时， $f_0(x) + \langle \hat{\lambda}, Ax - b \rangle$ 严格凸，进而 $D(v)$ 可微

原对偶次梯度法：计算量出在 $x^{(k+1)}$ ，那么想办法近似

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

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

$$v^{(k)} + \alpha^{(k)} (Ax^{(k)} - b)$$

由于两个方向都不精确，故收敛性质糟糕。

6.5.3 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

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

$$L(x, v) = f_0(x) + \langle v, Ax - b \rangle$$

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

增广拉格朗日函数是另一个问题的拉格朗日函数

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

两个问题的原对偶最优解相同

设 (x^*, v^*) 为原问题最优解

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

$$\begin{aligned}
& \begin{cases} Ax^* = b \\ \frac{\partial L_c(x, v^*)}{\partial x} \Big|_{x=x^*} = 0 \end{cases} \\
& \nabla_x(f_0(x) + \langle v^*, Ax - b \rangle) \Big|_{x=x^*} = 0 \\
& \nabla_x(f_0(x) + \langle v^*, Ax - b \rangle + \frac{c}{2} \|Ax - b\|^2) \Big|_{x=x^*} = 0 \\
& = \nabla_x(\frac{c}{2} \|Ax - b\|^2) \\
& = cA^T(Ax - b) \Big|_{x=x^*} = 0 \\
& \begin{cases} x^{(k+1)} = \arg \min_x L_c(x, v^{(k)}) \\ v^{(k+1)} = v^{(k)} + c(Ax^{(k+1)} - b) \end{cases}
\end{aligned}$$

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 34.

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

分析.

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

增广拉格朗日法

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

$$\begin{aligned}
v^{(k+1)} &= v^{(k)} + c^{(k)}(x_1^{(k+1)} - 1) \\
&= v^{(k)} + c^{(k)} \left(\frac{c^{(k)} - v^{(k)}}{c^{(k)} + 1} - 1 \right) \\
&= \frac{v^{(k)}}{c^{(k)} + 1} - \frac{c^{(k)}}{c^{(k)} + 1} \\
v^{(k+1)} - v^* &= v^{(k+1)} + 1 = \frac{v^{(k)}}{c^{(k)} + 1} + \frac{1}{c^{(k)} + 1} = \frac{v^{(k)} - v^*}{c^{(k)} + 1}
\end{aligned}$$

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

$$\min -\frac{1}{2}x_1^2 + \frac{1}{2}x_2^2$$

6.5.4 交替方向乘法

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

$$\begin{aligned}
\min \quad & f_1(x) + f_2(y) \\
\text{s.t.} \quad & Ax + By = 0
\end{aligned}$$

$$L_c(x, y, v) = f_1(x) + f_2(y) + \langle v, Ax + By \rangle + \frac{c}{2} \|Ax + By\|_2^2$$

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

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

$$\begin{cases} x^{(k+1)} = \arg \min_x L_c(x, y^{(k)}, v^{(k)}) \\ \quad = \arg \min_x f_1(x) + \langle v^{(k)}, Ax \rangle + \frac{c}{2} \|Ax + By^{(k)}\|_2^2 \\ \quad \iff \arg \min_x f_1(x) + \frac{c}{2} \|Ax + By^{(k)} + \frac{v^{(k)}}{c}\|_2^2 & \text{配方，关于 } y^{(k)} \text{ 的项为常数项，可忽略} \\ y^{(k+1)} = \arg \min_y L_c(x^{(k+1)}, y, v^{(k)}) \\ \quad \iff \arg \min_y f_2(y) + \frac{c}{2} \|Ax^{(k+1)} + By + \frac{v^{(k)}}{c}\|_2^2 \\ v^{(k+1)} = v^{(k)} + c(Ax^{(k+1)} + By^{(k+1)}) \end{cases}$$

两块算法依然具有很好的收敛性，但是多块的交替方向乘法不一定可以收敛。

例 35 (LASSO).

$$\min \frac{1}{2} \|Ax - b\|_2^2 + v \|x\|_1$$

分析. 写成交替方向乘子法的格式

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

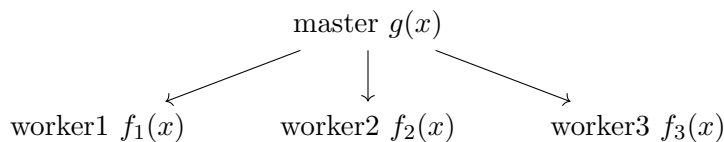
$$\begin{cases} x^{(k+1)} = \arg \min_x \frac{1}{2} \|Ax - b\|_2^2 + \frac{c}{2} \left\| x - y^{(k)} + \frac{v^{(k)}}{c} \right\|_2^2 \\ x^{(k+1)} = \arg \min_y v \|y\|_1 + \frac{c}{2} \left\| x^{(k+1)} - y + \frac{v^{(k)}}{c} \right\|_2^2 \\ v^{(k+1)} = v^{(k)} + c(x^{(k+1)} - y^{(k+1)}) \end{cases}$$

对于 x 可以求出显式解

$$\begin{aligned} & A^T(Ax - b) + c\left(x - y^{(k)} + \frac{v^{(k)}}{c}\right) \\ \implies & (A^T A + cI)x = A^T b + cy^{(k)} - v^{(k)} \end{aligned}$$

同样对于 y 也可以求出显式解

6.5.5 并行优化



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

针对LASSO问题, 每个人都有一个样本 (A_i, b_i) , 最小化样本之和, 以及正则化项

$$\begin{cases} (A_1, b_1) \implies \frac{1}{2} \|A_1 x - b_1\|_2^2 \\ \vdots \\ (A_n, b_n) \implies \frac{1}{2} \|A_n x - b_n\|_2^2 \\ g(x) = v \|x\|_1 \end{cases}$$

原问题即为

$$\min_x v \|x\|_1 + \frac{c}{2} \left\| \begin{bmatrix} A_1 & \cdots & A_N \end{bmatrix} x - \begin{bmatrix} b_1 & \cdots & b_N \end{bmatrix} \right\|_2^2$$

并行梯度下降法

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

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。

对偶分解法

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

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

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

$$\Rightarrow \begin{cases} x_i^{(k+1)} = \arg \min_{x_i} f_i(x_i) + \langle v_i^{(k)}, x_i \rangle \\ z^{(k+1)} = \arg \min_{z_i} g(z) - \sum_{i=1}^N \langle v_i^{(k)}, z \rangle \\ v_i^{(k+1)} = v_i^{(k)} + \alpha \langle x_i^{(k+1)}, z^{(k+1)} \rangle \end{cases}$$

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。

增广拉格朗日函数

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

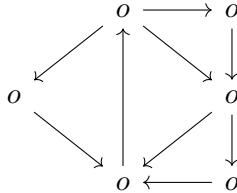
正则项会产生 x_i 和 z 的交叉项，不好处理

注意到 x_i 之间是没有依赖的，故采用交替方向乘法，加了增广项，可用固定步长达到最优解

$$\begin{cases} x_i^{(k+1)} = \arg \min f_i(x_i) + \langle \lambda_i^k, x_i \rangle + \frac{c}{2} \|x_i - z^{(k)}\|^2 \\ z^{(k+1)} = \arg \min g(z) - \langle \sum_{i=1}^n \lambda_i^{(k)}, z \rangle + \frac{c}{2} \sum_{i=1}^n \|x_i^{(k+1)} - z\|^2 \\ \lambda^{(k+1)} = \lambda^{(k)} + c(x_i^{(k+1)} - z^{(k+1)}) \end{cases}$$

每次要多传一倍的变量，以通信量开销换性能提升

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

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

梯度下降法，但由于去中心， $x^{(k)}$ 无处摆放

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

那就每一个点分配一个本地变量 x_i ，对邻居的更新做一个加权平均

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

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

$$x_i^{(k+1)} = \sum_{i=1}^n \omega_{ij} x_j^{(k)} - \alpha^{(k)} \nabla f_i(x_i^{(k)})$$

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 $x_i^{(k)} \rightarrow x^*$ ，将 x^* 代入

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

但原问题最优解

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

与上面的式子不等价

改成有约束优化的形式

$$\begin{aligned} \min \quad & \sum_{i=1}^n f_i(x_i) \\ \text{s.t.} \quad & x_1 = x_2 = \dots = x_n \end{aligned}$$

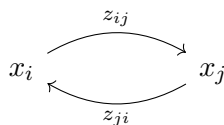
写出拉格朗日函数，对偶分解法

$$\sum_{i=1}^n f_i(x_i) + \sum_{(i,j) \in E} \langle \lambda_{ij}, x_i - x_j \rangle$$

别人的东西都在对偶变量中体现，求同存异

$$\begin{cases} x_i^{(k+1)} = \arg \min f_i(x_i) + \sum_{(i,j) \in E} (\lambda_{ij}^{(k)}, x_i) - \sum_{(j,i) \in E} \langle \lambda_{ij}, x_i \rangle \\ \lambda_{ij}^{(k+1)} = \lambda_{ij}^{(k)} + \alpha^{(k)} (x_i^{(k+1)} - x_j^{(k+1)}) - \sum_{(j,i) \in E} \langle \lambda_{ij}^{(k)}, x_i \rangle \end{cases}$$

依然要采用递减步长，才能保证收敛



$$\begin{aligned}
\min \quad & \sum_{i=1}^n f_i(x_i) \\
\text{s.t.} \quad & x_i = z_{ij}, \forall (i, j) \in E \\
& x_j = z_{ji}, \forall (i, j) \in E
\end{aligned}$$

可分线性约束，进而可以用交替方向乘法

$$\begin{aligned}
& \sum_{i=1}^n f_i(x_i) + \sum_{(i,j) \in E} (\langle \alpha_{ij}, x_i - z_j \rangle + \langle \beta_{ij}, x_j - z_{ji} \rangle) \\
& + \sum_{(i,j) \in E} \frac{c}{2} (\|x_i - z_{ij}\|^2 + \|x_j - z_{ji}\|^2) \\
\left\{ \begin{array}{l} x_i^{(k+1)} = \arg \min f_i(x_i) + \sum_{(i,j) \in E} \langle \alpha_{ij}^{(k)}, x_i \rangle + \sum_{(i,j) \in E} \langle \beta_{ji}^{(k)}, x_i \rangle \\ \quad + \frac{c}{2} \sum_{(i,j) \in E} \|x_i - z_j^{(k)}\|^2 + \frac{c}{2} \sum_{(i,j) \in E} \|x_i - z_{ji}^{(k)}\|^2 \\ z_j^{(k+1)} = \dots \\ \alpha_{ij}^{(k+1)} = \dots \\ \beta_{ij}^{(k+1)} = \dots \end{array} \right.
\end{aligned}$$

7 大数据中的优化问题与算法

n 个样本，每个样本为 $f_i(x)$

有限和优化问题

$$\min_x \frac{1}{n} \sum_{i=1}^n f_i(x)$$

等价于期望极小化问题

$$\min_x \mathbb{E}(f_i(x, \xi))$$

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

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]

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

注意这里需要采用变步长，否则无法收敛到最优解

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

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

7.1 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

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

对于每一个样本都存储一个梯度值

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

当时间足够长, 每一个里面都存在最优梯度

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

用空间换时间

7.2 深度神经网络

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

其中,

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

为损失函数, x_n 为第 n 层的网络输出 $f_n(x_{n-1}, \omega_n)$, 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 x_n 和 w_n 的梯度

$$\begin{cases} \frac{\partial E^{(i)}}{\partial x_n^{(i)}} = x_n^{(i)} - Y^{(i)} \\ \frac{\partial E^{(i)}}{\partial w_n} = \frac{\partial E^{(i)}}{\partial x_n^{(i)}} \frac{\partial x_n^{(i)}}{\partial w_n} = \frac{\partial E^{(i)}}{\partial x_n^{(i)}} \frac{\partial f_n(x_n^{(i)}, w_n)}{\partial w_n} \end{cases}$$

7.3 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

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

$$x_{t+1} = x_t - \alpha_t \nabla f_t(x_t)$$

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

7.4 动态优化

动态优化问题

$$\min \quad f_t(x)$$

$$x_t = x_{t-1} - \alpha \nabla f_t(x_{t-1})$$