最优化理论

陈鸿峥

2019.06*

目录

*Build 20190615

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）

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

- 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→低秩矩阵补全
  矩阵横向为用户，纵向为电影，值为评分值($1 \sim 5$)，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]$\|\mathbf{z}\|_\star$

$$\begin{aligned} \min \quad & \|\mathbf{z}\|_\star := \|\mathbf{z}\|_1 \\ \text{s.t.} \quad & \mathbf{z}_{ij} = M_{ij},\ (i,j) \in \Omega \end{aligned}$$

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
- 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

[1] A的秩等于非零奇异值$\sqrt{eig(A^{\mathrm{T}}A)}$数目
[2] 矩阵所有奇异值之和

- 仿射集*(affine set)*

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

例 1. 用定义易证线性方程组的解集$\mathcal{C} = \{\mathbf{x} \mid A\mathbf{x} = \mathbf{b}\}$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）

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

- 仿射包*(hull)*：所有仿射组合的集合

$$\operatorname{aff}\mathcal{C} := \{\theta_1 x_1 + \cdots + \theta_k x_k \mid \forall x_1, \ldots, x_k \in \mathcal{C}, \theta_1 + \cdots + \theta_k = 1\}$$

- 凸集*(convex set)*

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

- 凸组合

$$\forall x_1, x_2, \ldots, x_k \in \mathcal{C}, \theta_1, \ldots, \theta_k \in [0,1], \theta_1 + \cdots + \theta_k = 1:\ \theta_1 x_1 + \cdots + \theta_k x_k \in \mathcal{C}$$

- 凸包：最小的凸集

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

- 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：

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

- 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

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

超平面将空间划分为两个部分，即半空间

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

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)).

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

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

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

例 2.
- 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 $\Longrightarrow$ $f(S) = \{f(\mathbf{x}) \mid \mathbf{x} \in S\}$为凸
- $C \subset \mathbb{R}^m$为凸 $\Longrightarrow$ $f^{-1}(C) = \{\mathbf{x} \in \mathbb{R}^n \mid f(\mathbf{x}) \in C\}$为凸

[3]零元、加法封闭、数乘封闭

例 3. 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$A \leftarrow \begin{bmatrix} I & I \end{bmatrix}, \mathbf{x} \leftarrow \begin{bmatrix} \mathbf{x} & \mathbf{y} \end{bmatrix}^{\mathrm{T}}, \mathbf{b} \leftarrow 0$，由仿射函数性质得证

定义 7 (透视(perspective)函数[4]). 透视函数$P : \mathbb{R}^{n+1} \mapsto \mathbb{R}^n, \operatorname{dom} P = \mathbb{R}^n \times \mathbb{R}_{++}$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

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

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

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

定义 8 (线性分数函数). 仿射函数

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

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

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

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. $f : \mathbb{R}^n \to \mathbb{R}$为凸 $\iff$ $\operatorname{dom} f$为凸，且

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

- 严格凸：$\theta \in (0,1)$，不等式不能取等

---

[4] +代表≥ 0，++代表> 0

- 凹函数：若$-f$为凸

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

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

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)*[5]

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

即*Taylor*公式一阶展开

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

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

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

例 4. $f(\mathbf{x}) = \mathbf{a}^{\mathrm{T}}\mathbf{x} + \mathbf{b}$

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

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

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。 $\mathbf{x} \in \operatorname{dom} f \subset \mathbb{R}^n, \operatorname{dom} \widetilde{f} = \mathbb{R}^n$，会有

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

注意指示/示性(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$\exists \mathbf{x} \in \operatorname{dom} f, \nabla f(\mathbf{x}) = 0$

例 5. 二次函数$f(\mathbf{x}) = \dfrac{1}{2}\mathbf{x}^{\mathrm{T}} P\mathbf{x} + \mathbf{q}^{\mathrm{T}}\mathbf{x} + r$，$P \in \mathbb{S}^n$（对称矩阵），$\mathbf{q} \in \mathbb{R}^n$，$r \in \mathbb{R}$

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

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

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

例 6. $f(x)=\dfrac{1}{x^2}$, $\operatorname{dom} f=\{x\in\mathbb{R}, x\neq 0\}$

分析. 注意$\operatorname{dom} f$不是凸集

常见的凸函数例子如下

- 仿射函数$f(x)=ax+b$
- 指数函数$f(x)=\mathrm{e}^{ax}$
- 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$f(x)=\max\{x_1,\dots,x_n\}, x\in\mathbb{R}^n$

例 7. 极大值函数的解析近似[6]是$f(x)=\log(\mathrm{e}^{x_1}+\cdots+\mathrm{e}^{x_n})$

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

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

求*Hessian*矩阵
$$H=\frac{1}{(\mathbb{1}^{\mathrm{T}}\mathbf{z})^2}((\mathbb{1}^{\mathrm{T}}\mathbf{z})\operatorname{diag}(\mathbf{z})-\mathbf{z}\mathbf{z}^{\mathrm{T}})$$

[6]即无穷阶可微

将前面常量丢弃

$$
\begin{aligned}
\mathbf{v}^{\mathrm{T}} H \mathbf{v} &= (\mathbb{1}^{\mathrm{T}} \mathbf{z}) \mathbf{v}^{\mathrm{T}} \operatorname{diag}(\mathbf{z}) \mathbf{v} - \mathbf{v}^{\mathrm{T}} \mathbf{z} \mathbf{z}^{\mathrm{T}} \mathbf{v} \\
&= (\sum_i z_i)(\sum_i v_i^2 z_i) - (\sum_i v_i z_i)^2 \\
&\quad \text{令} \mathbf{a}_i := \mathbf{v}_i \sqrt{\mathbf{z}_i} = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix}^{\mathrm{T}}, b_i := \sqrt{\mathbf{z}_i} \\
&= (\mathbf{b}^{\mathrm{T}} \mathbf{b})(\mathbf{a}^{\mathrm{T}} \mathbf{a}) - (\mathbf{a}^{\mathrm{T}} \mathbf{b})^2 \qquad \textit{Cauchy} \\
&\geq 0
\end{aligned}
$$

由二阶条件知$f(\mathbf{x})$为凸函数

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. 由三角不等式

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

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

分析. 用高维定义

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

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

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

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

保持函数凸性

- 非负加权和$f_1, \ldots, f_m$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），$w(y)\geq 0$

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

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

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

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

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

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

– 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

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

- 任意个凸函数极大值函数为凸

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

- 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 9. 点x到集合$\mathcal{C}$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. $\mathbf{x}\in\mathbb{R}^n$，$x[i]$为第$i$大元素，$x[1]\geq x[2]\geq\cdots\geq x[r]\geq\cdots\geq x[n]$

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

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

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

- 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$n=k=1, \operatorname{dom} g=\mathbb{R}^n, \operatorname{dom} h=\mathbb{R}^k, \operatorname{dom} f=\mathbb{R}$，$h,g$二阶可微

$$\begin{aligned} f'(x) &= h'(g(x))\cdot g'(x) \\ f''(x) &= h''(g(x))(g'(x))^2+h'(g(x))g''(x)>0 \end{aligned}$$

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

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

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

若$f(x)$为凸，则$g(x,t)$相对于(x,t)联合凸

例 13.
- $f(x)=x^{\mathrm{T}}x$, $g(x,t)=x^{\mathrm{T}}x/t$
- $f(x)=-\log x$, $g(x,t)=t\log(t/x)$
- $u,v\in\mathbb{R}^n_{++}$, $g(u,v)=\sum_{i=1}^n u_i\log(u_i/v_i)$，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 11 (α次水平集(α-sub level set))**.** $f:\mathbb{R}^n\mapsto\mathbb{R}$，$C_\alpha=\{\mathbf{x}\in\operatorname{dom} f\mid f(\mathbf{x})\le\alpha\}$

定义 12 (拟凸函数(quasi-convex))**.** α次水平集为凸集 $\Longleftrightarrow$ 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

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

- 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$\mathcal{D} = \bigcap_{i=0}^{m} \operatorname{dom} f_i \cap \bigcap_{i=1}^{p} \operatorname{dom} h_i$
- 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$P^\star = \inf\{f_0(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\}$
- 最优解$\mathbf{x}^\star \iff \forall \mathbf{z} \in \mathbb{R}^n, \mathbf{z} \in \mathcal{X} :\ f_0(\mathbf{z}) \geq f_0(\mathbf{x}^\star)$
- 最优解集$X^\star = \{\mathbf{x}^\star \mid f_0(\mathbf{x}^\star) = P^\star, \mathbf{x}^\star \in \mathcal{S}\}$
- ε-次优解集$X_\varepsilon = \{\mathbf{x} \mid f_0(\mathbf{x}) \leq P^\star + \varepsilon, \mathbf{x} \in \mathcal{X}\}$
- 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$x_{local} = \{\mathbf{x} \mid \mathbf{x}\text{为局部最优}\}$

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 14.

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

定理 3. 凸问题局部最优等价于全局最优

分析. 若$\mathbf{x}$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$\mathbf{z} = \theta\mathbf{x} + (1-\theta)\mathbf{y}$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

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

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

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

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

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$\min f_0(\mathbf{x})$，只需$\nabla f_0(\mathbf{x}^\star) = 0$

  分析. 由凸函数的性质

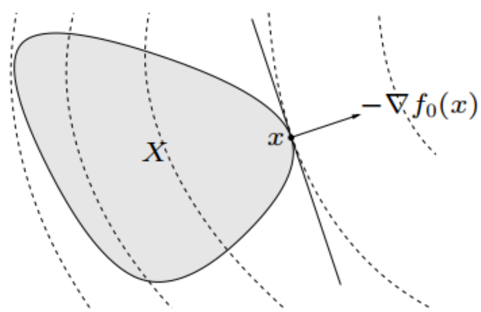
  $$\forall \mathbf{x},\mathbf{y}:\ f_0(\mathbf{y}) \ge f_0(\mathbf{x}) + \langle \nabla f_0(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle$$

  进而

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

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

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



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

分析. $\mathbf{x}^\star$最优，$A\mathbf{x}^\star = \mathbf{b}$，最优解需满足

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

从而

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

即

$$\forall \mathbf{v} \in \mathrm{Nul}\, A:\ \langle \nabla f_0(\mathbf{x}^\star), \mathbf{v}\rangle \ge 0$$

那么

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

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

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

1. 若$\nabla f_0(\mathbf{x}^\star) \ngeq 0$，则存在矛盾（负数行乘上正无穷），故$\nabla f_0(\mathbf{x}^\star) \geq 0$
2. 令$\mathbf{y} = 0$，有$0 \geq \langle \nabla f_0(\mathbf{x}^\star), \mathbf{x}^\star \rangle \implies \sum_{i=1}^n (\nabla f_0(\mathbf{x}^\star))_i x_i^\star \leq 0$
   前面≥ 0，进而互补松弛条件
3. $\mathbf{x}^\star \geq 0$

4.2 线性规划

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

例 17 (食谱问题)**.** m种营养元素不小于$b_1, \ldots, b_m$，n种食物，单位含量$a_{1j}, \ldots, a_{mj}$，食物量$x_1, \ldots, x_n$，价格$c_1, \ldots, c_n$

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

其中$i = 1, \ldots, m, j = 1, \ldots, n$

例 18 (线性分数规划)**.**

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

等价于

$$\begin{aligned}\min \quad & \mathbf{c}^{\mathrm{T}}\mathbf{y} + dz \\ \text{s.t.} \quad & G\mathbf{y} - hz \le 0 \\ & A\mathbf{y} - bz = 0 \\ & \mathbf{e}^{\mathrm{T}}\mathbf{y} + fz = 1 \\ & z \ge 0\end{aligned}$$

分析. 证明两个问题等价，P_0与P_1
若$\mathbf{x}$在P_0内可行

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

若$(\mathbf{y}, z)$在P_1中可行

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

若$z = 0$，x_0为P_0的可行解

$$\mathbf{x} = \mathbf{x}_0 + t\mathbf{y}, t \ge 0$$

$$\lim_{t\to\infty} \frac{\mathbf{c}^{\mathrm{T}}(\mathbf{x}_0 + t\mathbf{y}) + d}{\mathbf{e}^{\mathrm{T}}(\mathbf{x}_0 + t\mathbf{y}) + f} = \mathbf{c}^{\mathrm{T}}\mathbf{y}$$

代入看所有条件结论都相同

4.3 二次规划

二次规划(Quadratic Programming)

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

二次约束二次规划(QCQP)

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

例 19 (最小二乘问题)**.**

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

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

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

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

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

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

例 20 (投资组合问题(portfolio optimization))**.** 初始价格$x_1,\ldots,x_n$，最终价格$P_1x_1,\ldots,P_nx_n$

$$
\begin{aligned}
\max \quad & P_1x_1+\cdots+P_nx_n \\
\text{s.t.} \quad & x_1+\cdots+x_n=B \\
& x_1,\ldots,x_n \geq 0
\end{aligned}
$$

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

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

4.4 半定规划

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

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

例 21 (谱范数极小化问题). 矩阵多项式$A(\mathbf{x}) = A_0 + x_1 A_1 + \cdots + x_n A_n, A_i \in \mathbb{R}^{p\times q}$

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

谱范数代表$A(\mathbf{x})$的最大奇异值[7]

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

例 22 (最快分布式线性平均).

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

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

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

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

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

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

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

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

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

---

[7] 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

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

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

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

1. 能找到一个Pareto最优解
2. 遍历$\lambda_1, \ldots, \lambda_q$，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 13 (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，后者等式约束。

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

定义 14 (拉格朗日乘子(multiplier))**.** 包括原变量和对偶变量

- 原变量*(primal variable)*：$\boldsymbol{\lambda} = \begin{bmatrix}\lambda_1 & \cdots & \lambda_m\end{bmatrix}^{\mathrm{T}}$
- 对偶变量*(dual variable)*：$\mathbf{v} = \begin{bmatrix}v_1 & \cdots & v_p\end{bmatrix}^{\mathrm{T}}$

定义 15 (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$\mathcal{D} = \bigcap_{i=0}^{m} \operatorname{dom} f_i \cap \bigcap_{i=1}^{p} \operatorname{dom} h_i$，而不是可行解集$\mathcal{X}$。

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）
- $\forall \boldsymbol{\lambda} \geq 0, \forall \mathbf{v}:\ g(\boldsymbol{\lambda}, \mathbf{v}) \leq vp^\star$

定义 16 (对偶(dual)问题).

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

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

分析. 记$\mathbf{x}^\star$为原问题最优解，由优化问题定义有

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

又$P^\star = f_0(\mathbf{x}^\star)$为原问题最优解

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

进而推出

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

例 23 (最小二乘).

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

分析.

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

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

因而得到对偶问题

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

例 24 (标准线性规划)**.**

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

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

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

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

由于要极大，故不考虑负无穷部分，得到对偶问题如下

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

2° 考虑对偶问题的对偶，先将对偶问题变为极小化问题

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

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

得到新的对偶问题

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

对偶的对偶不一定回去，线性规划才满足

例 25 (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$\{1,\ldots,n\}$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

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

分析. 变为平方等式约束

$$
\begin{aligned}
L(\mathbf{x},\mathbf{v}) &= \mathbf{x}^{\mathrm{T}}W\mathbf{x} + \sum_{i=1}^{n} v_i(x_i^2-1) \\
g(\mathbf{v}) &= \inf_{x} L(\mathbf{x},\mathbf{v}) \\
&= \inf_{x}\left(\mathbf{x}^{\mathrm{T}}W\mathbf{x} + \sum_{i=1}^{n} v_i x_i^2 - \sum_{i=1}^{n} v_i\right) \\
&= \inf_{x}\left(\mathbf{x}^{\mathrm{T}}\left(W+\operatorname{diag}\mathbf{v}\right)\mathbf{x} - \mathbb{1}^{\mathrm{T}}\mathbf{v}\right) \\
&= \begin{cases} -\mathbb{1}^{\mathrm{T}}\mathbf{v} & W+\operatorname{diag}(\mathbf{v})\succeq 0 \\ -\infty & \textit{otherwise} \end{cases}
\end{aligned}
$$

其中

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

得到对偶问题如下

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

定义 17 (函数的共轭(conjugate))**.** $f:\mathbb{R}^n\mapsto\mathbb{R}, f^\star(\mathbf{y})=\sup_{\mathbf{x}\in\operatorname{dom}f}(\mathbf{y}^{\mathrm{T}}\mathbf{x}-f(\mathbf{x}))$，*几何意义即到不同斜率直线的距离最大值，例子如最大熵*

$$f_0(\mathbf{x})=\sum_{i=1}^n x_i\log x_i,\ f_0^\star(\mathbf{y})=\sum_{i=1}^n \mathrm{e}^{y_i-1}$$

例 26 (共轭函数)**.**

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

分析.

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

5.2 强弱对偶

定义 18 (对偶间隙(duality gap))**.** $\mathbf{p}^\star-\mathbf{d}^\star\ge 0$

- *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**$\mathbf{p}^\star\neq\mathbf{d}^\star$
2. 对于凸问题，若slater条件满足，$\mathbf{p}^\star=\mathbf{d}^\star$

定义 19 (相对内点(relative interior))**.**

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

定理 4 (Slater条件)**.** *强对偶对于下列凸问题成立*

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

若它严格可行，即

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

例 27. *二次规划(QP)*

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

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

例 28 (二次约束二次规划(QCQP))**.**

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

$P_0, \ldots, P_m$半正定

凸问题+Slater条件 $\Longrightarrow$ $\mathbf{p}^\star = \mathbf{d}^\star$，但有可能不满足Slater条件也依然强对偶

例 29.

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

分析.

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

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

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$\mathbf{p}^\star = \mathbf{d}^\star = 0$

例 30 (置信域问题)**.**

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

依然可以得到$\mathbf{p}^\star = \mathbf{d}^\star$

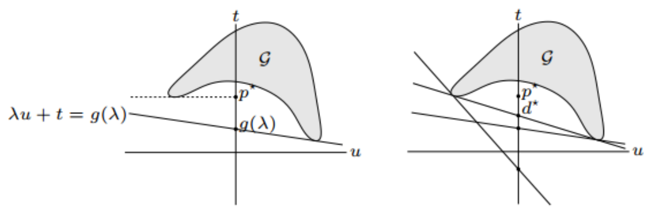
5.3 对偶问题的几种解释

5.3.1 几何解释

考虑问题只有一个约束$f_1(\mathbf{x}) \leq 0$，记$\mathcal{G} = \{(f_1(\mathbf{x}), f_0(\mathbf{x})) \mid \mathbf{x} \in \mathcal{D}\}$，则对偶函数

$$
\begin{aligned}
g(\boldsymbol{\lambda}) &= \inf\{t + \boldsymbol{\lambda} u \mid (u, t) \in G\} \\
\mathbf{p}^\star &= \inf\{t \mid (u, t) \in G, u \leq 0\}
\end{aligned}
$$

$$\max\ g(\boldsymbol{\lambda}),\ \boldsymbol{\lambda} \geq 0$$



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$\lambda_i \geq 0$

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

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

$$\mathbf{d}^\star = \sup_{\boldsymbol{\lambda} \geq 0} g(\boldsymbol{\lambda})$$

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

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

$$f(\mathbf{w},\mathbf{z}), \mathbf{w}\in S_w, \mathbf{z}\in S_z$$

极小极大不等式

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

若有$(\tilde{\mathbf{w}},\tilde{\mathbf{z}})$使得

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

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

则$(\tilde{\mathbf{w}},\tilde{\mathbf{z}})$为鞍点

有下面不等式成立

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

即从一个方向望过去是最小，从另一个方向望过去是最大

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

$$\Longrightarrow \sup_{\boldsymbol{\lambda}\ge 0} L(\mathbf{x},\boldsymbol{\lambda}) = \sup_{\boldsymbol{\lambda}\ge 0}\left\{f_0(\mathbf{x}) + \sum_{i=1}^{m}\lambda_i f_i(\mathbf{x})\right\} = \begin{cases} f_0(\mathbf{x}) & f_i(\mathbf{x})\le 0, i=1,\ldots,m \\ +\infty & \text{otherwise}\end{cases}$$

$$\Longrightarrow p^\star = \inf_x \{f_0(\mathbf{x}) \mid f_i(\mathbf{x}) \le 0, i=1,\ldots,m\} = \inf_x \sup_{\boldsymbol{\lambda}\ge 0} L(\mathbf{x},\boldsymbol{\lambda})$$

$$\mathbf{d}^\star = \sup_{\boldsymbol{\lambda}\ge 0} g(\boldsymbol{\lambda}) = \sup_{\boldsymbol{\lambda}\ge 0}\inf_{\mathbf{x}} L(\mathbf{x},\boldsymbol{\lambda}) \implies \mathbf{p}^\star \ge \mathbf{d}^\star$$

如果$L(\mathbf{x},\boldsymbol{\lambda})$有鞍点，则必有$\mathbf{p}^\star = \mathbf{d}^\star$

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。

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

$$\sup_{\boldsymbol{\lambda}\ge 0}\inf_{\mathbf{x}} L(\mathbf{x},\boldsymbol{\lambda}) = \inf_{\mathbf{x}}\sup_{\boldsymbol{\lambda}\ge 0} L(\mathbf{x},\boldsymbol{\lambda})$$

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

$$\tilde{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}\ge 0}\inf_{\mathbf{x}} L(\mathbf{x},\boldsymbol{\lambda})$$

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

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

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

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

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

因$\mathbf{p}^\star = \mathbf{d}^\star$，有

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

进而不等号都得为等号，还可得到

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

进而

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

故右推左成立

5.4 一般优化问题的对偶理论

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

不一定是凸问题，但$p^\star = d^\star$，最优解满足什么条件？

设其对偶问题为下式

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

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

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

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

同上理，不等号全取等，有

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

若f_0, f_i, h_i均可微，则必要条件为

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

定理 6 (KKT条件)**.** 可微优化问题的*KKT(Karush-Kuhn-Tucker)*条件如下

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

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

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

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

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

例 31 (Water-filling)**.** 共n个信道*(channel)*

$$
\textit{source} \longleftrightarrow \textit{destination}
$$

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

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

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

KKT条件如下

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

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

若$v^{\star} \geq \frac{1}{\alpha_i}$，则

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

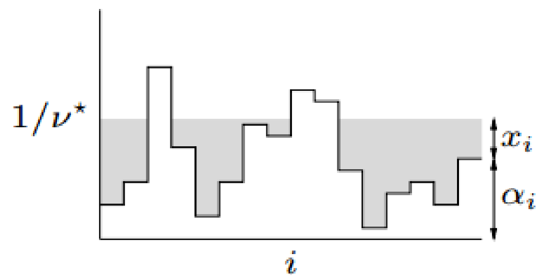
若$v^{\star} < \frac{1}{\alpha_i}$，则

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

进而

$$
\begin{aligned}
x_i^{\star} &> 0 \\
v^{\star} &= \frac{1}{\alpha_i + x_i^{\star}} \\
x_i^{\star} &= \frac{1}{v^{\star}} - \alpha_i \\
\Longrightarrow\ x_i^{\star} &= \max\left\{0, \frac{1}{v^{\star}} - \alpha_i\right\}
\end{aligned}
$$

结合$\sum_i x_i^{\star} = 1$，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$1/v^\star$

定义 20 (扰动(perturbed)问题)**.**

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

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

定理 7. 若原始问题为凸，则$\mathbf{p}^\star(\mathbf{u}, \mathbf{w})$是$(u, w)$的凸函数

6 优化算法

6.1 简介

例 32.

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

分析. *1° 罚函数法(penalty function)*

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

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

2° 拉格朗日函数方法

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

例 33.

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

分析. *log-barrier函数*

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

考虑这样的问题，$f_0(\mathbf{x})$可微，凸，无约束，即

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

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$\alpha^{(k)}$

$$\begin{cases} \text{确定步长} & \begin{cases} \text{固定步长} \\ \text{变化步长（递减步长）} \end{cases} \\ \text{搜索步长} \end{cases}$$

   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: $\quad\alpha^{(k)} \leftarrow \alpha^{(k)}\beta, \beta \in (0, 1)$
4: **else**
5: $\quad$ Stop

   而实际上没有必要求最优步长，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，

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

存在且有限，$f_0(\mathbf{x}^\star)$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

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

d. 当函数为凸时

$$0 \leq f_0(\mathbf{y}) - f_0(\mathbf{x}) - \langle \nabla, \mathbf{y} - \mathbf{x} \rangle \leq \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|_2^2$$

2. 强凸性(strong convexity)：即加了正则化项

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

二阶可微情况下的等价定义

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

例 34. 判断下列函数是否符合*Lipschitz*连续梯度及强凸性条件

$$
\begin{aligned}
f_0(\mathbf{x}) &= \mathbb{1}^{\mathrm{T}}\mathbf{x} & L&=0 & &✗\\
f_0(\mathbf{x}) &= \frac{1}{2}\|\mathbf{x}\|_2^2 & L&=1 & \mu&=1\\
f_0(\mathbf{x}) &= \frac{1}{4}\|\mathbf{x}\|_2^4 & &✗ & &✗
\end{aligned}
$$

区别于严格凸(strictly convex)，强凸一定是严格凸

定理 8. 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 9. 若$f_0(\mathbf{x})$有*Lipschitz*连续梯度，常数L，若$\alpha\in(0,\frac{2}{L})$，则有

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

即以$O(\frac{1}{k})$速度收敛

分析. 1° 点的单调性：与任意$\mathbf{x}^\star$的距离在缩小

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

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

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）

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

3° 函数值的充分下降（即证明收敛性）

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

$$
\begin{aligned}
f_0(\mathbf{x}^{(k)}) - f_0(\mathbf{x}^\star) &\le \langle f_0(\mathbf{x}^{(k)}, \mathbf{x}^{(k)} - \mathbf{x}^\star) \rangle \\
&= \langle \nabla f_0(\mathbf{x}^{(k)}) - \nabla f_0(\mathbf{x}^\star), \mathbf{x}^{(k)} - \mathbf{x}^\star \rangle \\
&\le \left\| \nabla f_0(\mathbf{x}^{(k)}) - \nabla f_0(\mathbf{x}^\star) \right\| \left\| \mathbf{x}^{(k)} - \mathbf{x}^\star \right\| \\
&\le \left\| \nabla f_0(\mathbf{x}^{(k)}) \right\| \left\| \mathbf{x}^{(k)} - \mathbf{x}^\star \right\|
\end{aligned}
$$

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

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

错位相消可得结论$O(\frac{1}{k})$收敛速度

定理 10. 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

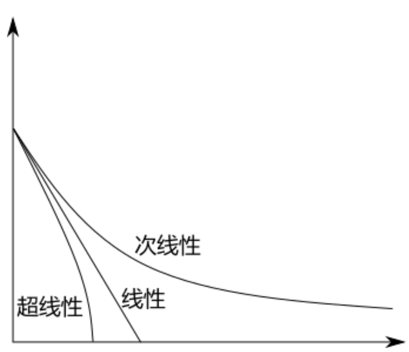
$$
\left\| \mathbf{x}^{(k)} - \mathbf{x}^\star \right\|^2 \le \left(1 - \frac{2\alpha\mu L}{\mu + L} \right)^k \left\| \mathbf{x}^{(0)} - \mathbf{x}^\star \right\|^2
$$

分析.

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

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

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

例 35.

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

分析.

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

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

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**

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

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

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

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

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

- Armijo Rule

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

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

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，

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

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

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

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

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

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

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

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

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

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

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)

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

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

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

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 21 (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[8]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$\frac{1}{k}$

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

- 不可加递减，如$\frac{1}{\sqrt{k}}$

$$\sum_{k=0}^{\infty}\alpha^{(k)}=0 \qquad \lim_{k\to\infty}\alpha^{(k)}\to 0$$

考虑收敛速度

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

假设函数Lipschitz连续

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

[8]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

$\forall \mathbf{x}^\star$最优

$$
\begin{aligned}
&\left\|\mathbf{x}^{(k+1)}-\mathbf{x}^\star\right\|^2=\left\|\mathbf{x}^{(k)}-\alpha^{(k)}g_0(\mathbf{x}^{(k)})-\mathbf{x}^\star\right\|^2\\
&=\left\|\mathbf{x}^{(k)}-\mathbf{x}^\star\right\|^2-2\langle\alpha^{(k)}g_0(\mathbf{x}^{(k)}),\mathbf{x}^{(k)}-\mathbf{x}^\star\rangle+(\alpha^{(k)})^2\left\|g_0(\mathbf{x}^{(k)})\right\|^2 \qquad \text{拆平方}\\
&\le\left\|\mathbf{x}^{(k)}-\mathbf{x}^\star\right\|^2-2(\alpha^{(k)})^2(f_0(\mathbf{x}^{(k)})-f_0(\mathbf{x}^\star))+(\alpha^{(k)})^2G^2\\
&\implies \left\|\mathbf{x}^{(k+1)}-\mathbf{x}^\star\right\|^2\le\left\|\mathbf{x}^{(0)}-\mathbf{x}^\star\right\|^2-2\sum_{i=0}^k\alpha^{(i)}(f_0(\mathbf{x}^{(i)})-f_0(\mathbf{x}^\star))+\sum_{i=0}^k(\alpha^{(i)})^2G^2\\
&\implies 2\sum_{i=0}^k\alpha^{(i)}(f_0(\mathbf{x}^{(i)})-f_0(\mathbf{x}^\star))\le\left\|\mathbf{x}^{(0)}-\mathbf{x}^\star\right\|^2+\sum_{i=0}^k(\alpha^{(i)})^2G^2
\end{aligned}
$$

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

这是一个紧的界

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

$$\lim_{k\to\infty}\inf_{i=0,\dots,k}\left(f_0(\mathbf{x}^{(i)})-f_0(\mathbf{x}^\star)\right)=0$$

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

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

令$N=\max\{N_1,N_2\},\forall k>N$

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

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

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)

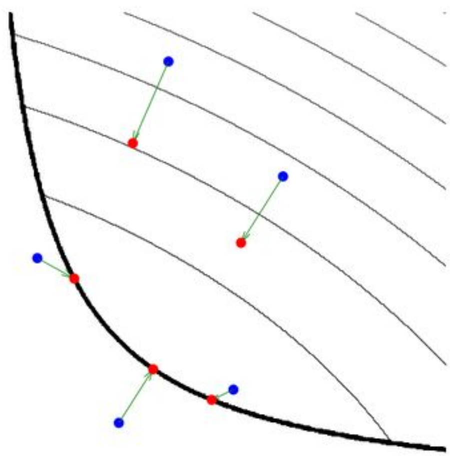
考虑有**结构**，不可微的函数

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

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

定义 22 (邻近点投影(proximal mapping)). 函数r和平衡参数λ的近端算子

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



注：邻近点/近端投影字面上理解即对点$\hat{x}$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[9]。

邻近点梯度法迭代格式如下

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

例 36 (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

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

[9]详细见https://zhuanlan.zhihu.com/p/37444622

先求$r(\mathbf{x})$的邻近点投影

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

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

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

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

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

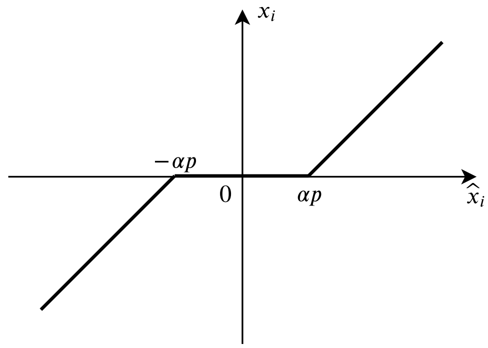
  推得

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

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

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

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 37 (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

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

如果有约束$\mathbf{x} \in \mathcal{C}$

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

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

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

分析. 同*LASSO*，由于矩阵的秩（奇异值向量*0*-范数）不好求，改为矩阵的和范数$\|\cdot\|$（奇异值向量*1*-范数），即

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

其中，

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

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

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

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

邻近点梯度迭代格式为

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

求次梯度

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

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

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

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

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

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

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

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

求梯度有

$$\nabla f_0(\mathbf{x}^{(k)}) + \nabla^2 f_0(\mathbf{x}^{(k)})\mathbf{v} = 0$$

$$\implies \mathbf{v} = -(\nabla^2 f_0(\mathbf{x}^{(k)}))^{-1}\nabla f_0(\mathbf{x}^{(k)}) \to \text{牛顿方向}$$

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

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

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

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

- 若$\left\|\nabla f_0(\mathbf{x}^{(k)})\right\|_2 > \eta$，阻尼(damped)牛顿段
  用Armijo Rule算步长，$\exists\gamma > 0$，$f_0(\mathbf{x}^{(k+1)}) - f_0(\mathbf{x}^{(k)}) \le -\gamma$
- 若$\left\|\nabla f_0(\mathbf{x}^{(k)})\right\|_2 \le \eta$，纯牛顿段
  $\alpha = 1, f_0(\mathbf{x}^{(k+1)}) - f_0(\mathbf{x}^\star) \le \Delta(\frac{1}{2})^{2^k}, \exists\Delta > 0$，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 39.

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

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

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

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

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）

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

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

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)

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

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)

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

- 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

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

例 40.

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

分析. 等价于KKT条件

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

等价于

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

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

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

近似等价于（二阶近似Taylor展开）

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

写出问题关于$\mathbf{d}$的KKT条件，可得等价条件

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

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**，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 41.

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

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$\mathbf{v}^\star$的同时也找到了$\mathbf{x}^\star$）

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

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

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

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

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

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

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

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

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

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

增广拉格朗日函数

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

增广拉格朗日函数是下面问题的拉格朗日函数

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

两个问题的原对偶最优解相同

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

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

对于原问题有

$$\nabla_x (f_0(\mathbf{x}) + \langle \mathbf{v}^\star, A\mathbf{x} - \mathbf{b} \rangle)\Big|_{\mathbf{x}=\mathbf{x}^\star} = 0$$

对于对偶问题有

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

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 42.

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

分析.

$$L(\mathbf{x}, v) = \frac{1}{2}x_1^2 + \frac{1}{2}x_2^2 + v(x_1 - 1)$$

$$\frac{\partial L(\mathbf{x}, v^\star)}{\partial \mathbf{x}}\Big|_{\mathbf{x}=\mathbf{x}^\star} = 0 = \begin{bmatrix} x_1 + v^\star \\ x_2 \end{bmatrix}$$

增广拉格朗日法

$$L_{c^{(k)}}(\mathbf{x}, v) = \frac{1}{2}x_1^2 + \frac{1}{2}x_2^2 + v(x_1 - 1) + \frac{c^{(k)}}{2}(x_1 - 1)^2$$

$$\mathbf{x}^{(k+1)} = \arg\min L_{c^{(k)}}(\mathbf{x}, v^{(k)})$$

$$\begin{cases}
x_1 + v^{(k)} + c^{(k)}(x_1 - 1) = 0 \\
x_2 = 0
\end{cases} \implies \mathbf{x}^{(k+1)} = \begin{bmatrix} \frac{c^{(k)} - v^{(k)}}{c^{(k)}+1} \\ 0 \end{bmatrix}$$

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

$$v^{(k+1)} - v^\star = v^{(k+1)} + 1 = \frac{v^{(k)}}{c^{(k)} + 1} + \frac{1}{c^{(k)} + 1} = \frac{v^{(k)} - v^\star}{c^{(k)} + 1}$$

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

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

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$\|Ax+By\|_2^2$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

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

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 43 (LASSO)**.**

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

分析. 对原问题进行变形，等价于下述约束问题

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

构造增广拉格朗日函数

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

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

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

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

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

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

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

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

$$\begin{tikzcd}[column sep=small] & \text{master } g(\mathbf{x}) \arrow[dl] \arrow[d] \arrow[dr] & \\ \text{worker1 } f_1(\mathbf{x}) & \text{worker2 } f_2(\mathbf{x}) & \text{worker3 } f_3(\mathbf{x})\end{tikzcd}$$

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

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

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

原问题即为

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

7.1.1 并行梯度下降法

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

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

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

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

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

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

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

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

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

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

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

$$\begin{tikzcd} & o \arrow[r] \arrow[dl] \arrow[dr] & o \arrow[d] \\ o \arrow[dr] & & o \arrow[dl] \arrow[d] \\ & o \arrow[uu] & o \arrow[l] \end{tikzcd}$$

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

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

$$W = \left[\omega_{ij}\right], W = W^{\mathrm{T}}, W\mathbb{1} = \mathbb{1}$$

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

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

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$\mathbf{x}_i^{(k)} \to \mathbf{x}^\star$，将$\mathbf{x}^\star$代入

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

但原问题最优解

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

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

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

依然要采用递减步长，才能保证收敛

$$\begin{tikzcd} \mathbf{x}_i \arrow[r, bend left, "\mathbf{z}_{ij}"] & \mathbf{x}_j \arrow[l, bend left, "\mathbf{z}_{ji}"] \end{tikzcd}$$

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

可分线性约束，进而可以用交替方向乘子法

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

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

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

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

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

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

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

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

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

相当于用空间换时间

7.5 深度神经网络

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

其中，

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

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

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

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

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

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

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

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

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

7.8 Nesterov加速

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

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

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$\boldsymbol{\gamma}$，不同时刻权重不同，引入$\boldsymbol{\beta}$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$\boldsymbol{\beta}$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　矩阵微积分

A.1　基本矩阵知识

定义 23 (正定(positive definite)矩阵)**.** 若矩阵A满足

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

则称A为正定矩阵。

定义 24 (合同(congruent)矩阵)**.** 若存在可逆矩阵C使得$C^{\mathrm{T}}AC = B$，则称A为B合同，记作$A \backsimeq B$

定义 25 (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$1-i$行和$1-i$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

定义 26 (向量范数)**.** $\|\cdot\|$为范数需要满足以下三个条件

1. 数乘：$\|a\mathbf{x}\| = |a|\,\|\mathbf{x}\|$
2. 零元：$\|\mathbf{x}\| = 0 \iff \mathbf{x} = \mathbf{0}$
3. 三角不等式：$\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$

- 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
- 2-范数：欧几里得距离
- p-范数：p次方之和的p次根
- 无穷范数：最大值

定义 27 (Taylor公式)**.** 对$f(\mathbf{x})$在点$\mathbf{x}_0$处展开有

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

A.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$\nabla$默认对$\mathbf{x}$求导，可以得到以下公式。

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

   分析. 展开，对每一个元素讨论

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

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

分析. 法一：考虑每一元素

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

法二：变量多次出现的求导法则，下标c代表视为常数

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

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

分析. 变量多次出现的求导法则

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

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

分析. 法一：展开括号，逐一求导

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

法二：线性变换的求导公式

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

A.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

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

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

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

B 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