非光滑优化算法

于冰冰 21901037 数硕 1903

2020 年 9 月 21 日

目录

1 第一题

计算函数 $f:\mathfrak{R}^n \to \tilde{\mathfrak{R}}$ 的邻近映射 (proximal mapping):

- $f(X)=\|X\|_*$ 是 $X\in\mathbb{S}^n$ 的核范数;
- $f(X)=\delta_C^*(x)=max_{y\in C}<y,x>$, 其中 C 为闭凸集合;
- $f(x)=\inf_{y\in C}\|x-y\|_2$，其中 C 为闭凸集合。

1.1

核范数 (迹范数) 定义为:

$$\|X\|_*=\sum_{i=1}^{r}\sigma_i(X) \tag{1}$$

$\mathbb{S}^n$ 定义为全体 $n\times n$ 对称矩阵:

$$\mathbb{S}^n=\left\{\mathbf{A}\in\mathbb{R}^{n\times n}:\mathbf{A}=\mathbf{A}^T\right\} \tag{2}$$

假定 X 的奇异值分解为:

$$X=Udiag(\lambda(X))U^T$$

首先有 $\mathbb{S}^n$ 上的谱邻近公式:
假定 $F:\mathbb{S}\to(-\infty,\infty]$ 由 $F=f\circ\lambda$ 给定，其中 $f:\mathbb{R}^n\to(-\infty,\infty]$ 为置换对称的闭包凸函数，假定 $X=Udiag(\lambda(X))U^T$，其中 $U\in\mathbb{O}$, 则有:

$$\text{prox}_F(X)=Udiag\left(\text{prox}_f(\lambda(X))\right)U^T \tag{3}$$

证明. 首先有:

$$\text{prox}_F(X)=\text{argmin}_{Z\in\mathbb{S}^n}\left\{F(Z)+\frac{1}{2}\|Z-X\|_F^2\right\} \tag{4}$$

记 D 为 $D=diag(\lambda(X))$，注意到对于任意 $Z\in\mathbb{S}^n$:

$$F(Z)+\frac{1}{2}\|Z-X\|_F^2=F(Z)+\frac{1}{2}\left\|Z-UDU^T\right\|_F^2\overset{(*)}{=}F\left(U^TZU\right)+\frac{1}{2}\left\|U^TZU-D\right\|_F^2 \tag{5}$$

其中的转换 $(*)$ 是由于 $F(Z)=f(\lambda(Z))=f(\lambda(U^TZU))=F(U^TZU)$, 记 $W=U^TZU$，在 W 发生变化时，可以得到 (4) 的最优解由下式给定:

$$Z=UW^*U^T \tag{6}$$

其中 $W^*\in\mathbb{S}^n$ 为下式的最优解:

$$\min_{W\in\mathbb{S}^n}\left\{G(W)\equiv F(W)+\frac{1}{2}\|W-D\|_F^2\right\} \tag{7}$$

接下来我们证明 W^* 为对角阵，令 $i\in\{1,2,\cdots,n\}$. 记 V_i 为如下的对角阵: 在除 (i,i) 处的对角线上为 1，在 (i,i) 处为 -1。我们定义 $\widetilde{W}_i=V_iW^*V_i^T$。由于 $V_i\in\mathbb{O}^n$，显然有:

$$F(V_iW^*V_i^T)=f(\lambda(V_iW^*V_i^T)=f(\lambda(W^*))=F(W^*) \tag{8}$$

于是我们得到：

$$\begin{aligned} G\left(\widetilde{W}_i\right) &= F\left(\widetilde{W}_i\right) + \frac{1}{2}\left\|\widetilde{W}_i - D\right\|_F^2 \\ &= F\left(V_i W^* V_i^T\right) + \frac{1}{2}\left\|V_i W^* V_i^T - D\right\|_F^2 \\ &= F\left(W^*\right) + \frac{1}{2}\left\|W^* - V_i^T D V_i\right\|_F^2 \\ &\overset{**}{=} F\left(W^*\right) + \frac{1}{2}\left\|W^* - D\right\|_F^2 \\ &= G\left(W^*\right) \end{aligned} \tag{9}$$

这里的 (**) 是由于 V_i 和 D 都是对角阵，因此有 $V_i^T D V_i = V_i^T V_i D = D$。我们得出结论：$\widetilde{W}_i$ 也是最优解。由 (7) 最优解的唯一性，我们可以得到 $W^* = V_i W^* V_i^T$ 。比较等式两边矩阵的第 i 行，可以看出对于任意 $i \neq i, W_{ij}^* = 0$。于是 W^* 为对角矩阵。(7) 的最优解由 $W^* = diag(w^*)$ 给定，其中 w^* 为下式的最优解：

$$\min_w \left\{ F(\operatorname{diag}(w)) + \frac{1}{2}\|\operatorname{diag}(w) - D\|_F^2 \right\} \tag{10}$$

由于 $F(\operatorname{diag}(w)) = f\left(w^{\downarrow}\right) = f(w)$, $\|\operatorname{diag}(w) - D\|_F^2 = \|w - \lambda(X)\|_2^2$，$w^*$ 可由下式给定：

$$\mathbf{w}^* = \operatorname{argmin}_w \left\{ f(w) + \frac{1}{2}\|w - \boldsymbol{\lambda}(X)\|_2^2 \right\} = \operatorname{prox}_f(\boldsymbol{\lambda}(X)) \tag{11}$$

因此有 $W^* = \operatorname{diag}(\operatorname{prox}_f(\lambda(X)))$，结合 (6)，证毕。 □

接下来计算一范数的邻近映射：若 $g : \mathbb{R}^n \to \mathbb{R}$ 由 $g(x) = \lambda\|x\|_1$ 定义，其中 $\lambda > 0$，则 $\operatorname{prox}_g(x) = \mathcal{T}_\lambda(x) = [|x| - \lambda e]_+ \odot \operatorname{sgn}(x)$

证明. 首先有

$$g(\mathbf{x}) = \sum_{i=1}^n \varphi\left(x_i\right)$$

其中 $\varphi(t) = \lambda|t|$。有 $\operatorname{prox}_\varphi(s) = \mathcal{T}_\lambda(s)$. 其中 $\mathcal{T}_\lambda$ 定义为：

$$\mathcal{T}_\lambda(y) = [|y| - \lambda]_+ \operatorname{sgn}(y) = \begin{cases} y - \lambda, & y \geq \lambda \\ 0, & |y| < \lambda \\ y + \lambda, & y \leq -\lambda \end{cases}$$

这里的 $\mathcal{T}_\lambda$ 称为软阈值函数。在此定义下，有

$$\operatorname{prox}_g(x) = \left(\mathcal{T}_\lambda\left(x_j\right)\right)_{j=1}^n$$

将软阈值函数的定义扩充到向量空间上，对于任意的 $x \in \mathbb{R}^n$，有

$$\mathcal{T}_\lambda(x) \equiv \left(\mathcal{T}_\lambda\left(x_j\right)\right)_{j=1}^n = [|x| - \lambda e]_+ \odot \operatorname{sgn}(x)$$

在此标记下，有

$$\operatorname{prox}_g^{(x)} = \mathcal{T}_\lambda(x)$$

证毕 □

根据上面两条定理易得，$\operatorname{prox}_f(x) = U \operatorname{diag}(\mathcal{T}_1(X)) U^T$

1.2

共轭函数：$f:\mathbb{E}\to[-\infty,\infty]$ 的共轭函数 $f^*:\mathbb{E}\to[-\infty,\infty]$ 定义为:

$$f^*(y)=\max_{x\in\mathbb{E}}\{\langle y,x\rangle-f(x)\},\quad y\in\mathbb{E}^*,$$

首先证明

$$\delta_C^*=\sigma_C \tag{12}$$

证明. 令 $f=\delta_C$，其中 $C\subset\mathbb{E}$ 非空，则对于任意 $y\in\mathbb{E}^*$:

$$f^*(y)=\max_{x\in\mathbb{E}}\{\langle y,x\rangle-\delta_C(\mathbf{x})\}=\max_{x\in C}\langle y,x\rangle=\sigma_C(y)$$

□

Moreau 分解公式：$f:\mathbb{E}\to[-\infty,\infty]$ 为封闭的、凸的，则对任意 $x\in\mathbb{E}$:

$$\mathrm{prox}_f(x)+\mathrm{prox}_{f^*}(x)=x \tag{13}$$

令 $g:\mathbb{E}\to[-\infty,\infty]$, 其中 $g(x)=\delta_C(x)$，C 非空，则

$$\mathrm{prox}_g(x)=\mathrm{argmin}_{u\in\mathbb{E}}\left\{\delta_C(u)+\frac{1}{2}\|u-x\|^2\right\}=\mathrm{argmin}_{u\in C}\|u-x\|^2=P_C(x) \tag{14}$$

利用 (12)(13)(14) 可得：

$$\mathrm{prox}_f(x)=x-P_C(x)$$

1.3

注意到 $f(x)=\inf_{y\in C}\|x-y\|_2=d_C(x)$ 下面证明：若 $C\subset\mathbb{E}$ 为闭的、凸的非空集，$\lambda>0$，则对于任意 $x\in\mathbb{E}$, 有

$$\operatorname{prox}_{\lambda d_C}(x)=\begin{cases}(1-\theta)x+\theta P_C(x), & d_C(x)>\lambda\\ P_C(x), & d_C(x)\le\lambda\end{cases}\tag{15}$$

其中

$$\theta=\frac{\lambda}{d_C(x)}\tag{16}$$

证明. 记 $u=\operatorname{prox}_{\lambda d_C}(x)$, 由邻近映射第二定理，有

$$x-u\in\lambda\partial d_C(u)\tag{17}$$

接下来分两种情况进行讨论：
Case1 $U\notin C$,(17) 等价为

$$x-u=\lambda\frac{u-P_C(u)}{d_C(u)}\tag{18}$$

记 $\alpha=\frac{\lambda}{d_C(u)}$，公式也可以写为：

$$u=\frac{1}{\alpha+1}x+\frac{\alpha}{\alpha+1}P_C(u)\tag{19}$$

或

$$x-P_C(u)=(\alpha+1)\left(u-P_C(u)\right)\tag{20}$$

由第二投影定理，为证明 $P_C(u)=P_C(x)$，只需要证明：

$$\langle\mathbf{x}-P_C(\mathbf{u}),\mathbf{y}-P_C(\mathbf{u})\rangle\le 0\text{ for any }\mathbf{y}\in C\tag{21}$$

利用 (20), 我们可以证明 (21) 等价于：

$$(\alpha+1)\langle u-P_C(u),y-P_C(u)\rangle\le 0\text{ for any }y\in C\tag{22}$$

由第二投影定理，这是一个有效的不等式，因此 $P_C(u)=P_C(x)$，我们在 (20) 等式两边同时取范数，有：

$$d_C(x)=(\alpha+1)d_C(u)=d_C(u)+\lambda\tag{23}$$

由于 $d_C(u)>0$，所以 $d_C(x)>\lambda$，于是

$$\frac{1}{\alpha+1}=\frac{d_C(u)}{\lambda+d_C(u)}=\frac{d_C(x)-\lambda}{d_C(x)}=1-\theta\tag{24}$$

其中 θ 由 (16) 给定。于是 (19) 也可以表示为：

$$\operatorname{prox}_{\lambda d_C}(x)=(1-\theta)x+\theta P_C(x)\tag{25}$$

Case2: 若 $u\in C$，则 $u=P_C(x)$。
令 $v\in C$，由于 $u=\operatorname{prox}_{\lambda d_C}(x)$, 它遵循如下公式：

$$\lambda d_C(u)+\frac{1}{2}\|u-x\|^2\le\lambda d_C(v)+\frac{1}{2}\|v-x\|^2\tag{26}$$

由于 $d_C(u)=d_C(v)=0$,

$$\|u-x\|\le\|v-x\|$$

因此

$$u = \arg\min_{v \in C} \|v - x\| = P_C(x)$$

优化条件 (17) 变为:

$$\frac{x - P_C(x)}{\lambda} \in N_C(u) \cap B[0, 1]$$

这通常意味着

$$\|\frac{x - P_C(x)}{\lambda}\| \leq 1$$

即

$$d_C(x) = \|P_C(x) - x\| \leq \lambda$$

□

从而

$$\text{prox}_{d_C}(x) = \begin{cases} (1-\theta)x + \theta P_C(x), & d_C(x) > 1 \\ P_C(x), & d_C(x) \leq 1 \end{cases}$$

其中 $\theta = \frac{1}{d_C(x)}$
也可记为

$$\text{prox}_{d_C}(x) = x + \min\{\frac{1}{d_C(x)}, 1\}(P_C(x) - x)$$

2 第二题

考虑下述等式约束二次规划问题

$$\begin{aligned} \min \quad & f(x) = c^T x + \tfrac{1}{2} x^T G x \\ \text{s.t.} \quad & Ax - b = 0 \end{aligned}$$

其中 $G \in \mathbb{S}^n$ 是 $n \times n$ 的对称矩阵，$A \in \Re^{m \times n}$ 是行满秩矩阵，$b \in \Re^m$

- 写出增广 Lagrange 方法的 (x^k, λ^k) 迭代格式
- 分析 G 与 A 满足什么条件时，增广 Lagrange 方法是收敛的
- 用 $\theta_r : \Re^m \to \bar{\Re}$ 记增广 Lagrange 函数对偶的目标函数，即

  $$\theta_r(\lambda) = \inf_x L_r(x, \lambda)$$

  其中

  $$L_r(x, \lambda) = c^T x + \frac{1}{2} x^T G x + \lambda^T (Ax - b) + \frac{r}{2} \|Ax - b\|^2$$

  根据 $\nabla^2_{\lambda\lambda} \theta_r(\bar{\lambda})$ 特征值说明增广 Lagrange 的收敛速度，当 r 充分大时接近 Newton 方法的收敛速度。

2.1

定义增广拉格朗日函数

$$\begin{aligned} L_r(x, \lambda) &= f(x) + \frac{1}{2r} \sum_{i=1}^{n} [(\lambda_i + r(Ax_i - b_i))^2 - \lambda_i^2] \\ &= f(x) + \sum_{i=1}^{n} \lambda_i (Ax_i - b_i) + \frac{r}{2} \sum_{i=1}^{n} (Ax_i - b_i)^2 \\ &= f(x) + \lambda^T (Ax - b) + \frac{r}{2} (Ax - b)^T (Ax - b) \\ &= c^T x + \frac{1}{2} x^T G x + \lambda^T (Ax - b) + \frac{r}{2} (Ax - b)^T (Ax - b) \end{aligned}$$

其迭代格式为：

$$\begin{aligned} x^{k+1} &= \arg\min_x L_r(x, \lambda^k) \\ \lambda^{k+1} &= \lambda^k + \alpha (Ax^{k+1} - b) \end{aligned}$$

2.2

G 和 A 应该满足：对于 $Ax = b$ 的任一非零解 z，存在某个正数 r' 使得当 $r \geq r'$ 时，

$$\nabla^2_{xx} L_r(x^*, \lambda^*) \succ 0$$

2.3

$$\frac{\partial L_r(x, \lambda)}{\partial x} = c^T + x^T G + \lambda^T A + r x^T A^T A - r b^T A$$

于是，当 $x_* = (rA^TA + G)^{-1}(rA^Tb - A^T\lambda - c)$ 时，$L_r(x, \lambda)$ 取得最小值，此时有

$$\begin{aligned}\theta_r(\lambda) &= \inf_x L_r(x, \lambda) \\ &= c^Tx_* + \frac{1}{2}x_*^TGx_* + \lambda^TAx_* - \lambda^Tb + \frac{r}{2}x_*^TA^TAx_* - rb^TAx + \frac{r}{2}b^Tb \\ &= \frac{1}{2}(-rb^TA + \lambda^TA + c^T)x_* + \frac{r}{2}b^Tb - \lambda^Tb \\ &= -\frac{1}{2}(rA^Tb - A^T\lambda - c)^T(rA^TA + G)^{-1}(rA^Tb - A^T\lambda - c) + \frac{r}{2}b^Tb - \lambda^Tb\end{aligned}$$

计算有

$$\nabla^2_{\lambda\lambda}\theta_r(\bar{\lambda}) = -A(rA^TA + G)^{-1}A^T$$

增广 Lagrange 方法为线性收敛，当 r 充分大时，为超线性收敛。

3 第三题

阅读论文“Ying Cui, Chao Ding and Xinyuan Zhao, Quadratic growth conditions for convex matrix optimization problems associated with spectral functions, SIAM J. Optim. Vol. 27, No. 4, 2017, pp. 2332–2355”，详细论述 Rockafellar 两篇经典论文在其中起到的作用。

A：Augmented Lagrangians and applications of the proximal point algorithm in convex programming, Mathematics of Operations Research。

B:Monotone operators and the proximal point algorithm, SIAM Journal on Control and Optimization,

在 A 中，Rockafellar 论述了求解凸优化问题中的 ALM 方法的收敛速度，表明 ALM 方法是非精确双近点算法 (PPA) 的一种特殊情况。在证明命题 19 时，应用了 A 中命题六的证明思路，表明了 ALM 方法和 PPA 方法生成的迭代序列之间的关系。在证明定理 20 时，可以从 A 中的定理 4 获得整个序列的有界性和收敛性。在 B 中，Rockafellar 确定了 Lipschitz 连续的 $\mathcal{T}_\phi^{-1}$ 在原点处不精确 PPA 的收敛速度。