机器学习中的优化算法

Lecture06: 复合优化算法–近似点梯度法

张立平

清华大学数学科学系

办公室：理科楼#A302, Tel: 62798531

E-mail: lipingzhang@tsinghua.edu.cn

Contents and Acknowledgement

- 教材：最优化：建模、算法与理论

  http://bicmr.pku.edu.cn/ wenzw/bigdata2021.html

- 致谢：北京大学文再文教授

Outline of Lecture06

- 闭函数和共轭函数
- 邻近算子
- 投影
- 支撑函数、范数、距离
- 近似点梯度法
- 应用: LASSO问题、低秩矩阵恢复问题
- 收敛性分析

闭集、闭函数、共轭函数

▶**闭集:** 包含其边界的集合C, i.e.,

$$x^k \in C, \quad x^k \to \bar{x} \qquad \Longrightarrow \qquad \bar{x} \in C.$$

- 有限或无限个闭集的交集仍是闭集.
- 有限个闭集的并集仍是闭集.
- 线性映射的原象集: 若 C 闭, 则$\{x \mid Ax \in C\}$ 是闭集.
- 闭集C在线性映射下的像$AC = \{Ax \mid x \in C\}$不一定是闭的; e.g.,

$$C = \{(x_1, x_2) \in \mathbf{R}^2_+ \mid x_1 x_2 \geq 1\}, \quad A = \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad AC = \mathbf{R}_{++}.$$

▶闭集在线性映射下的像AC 为闭集的充分条件:

- C 是闭凸集
- A 的零空间中不包含 C 的回收方向 (recession direction), i.e.,

$$Ay = 0, \quad \hat{x} \in C, \quad \hat{x} + \alpha y \in C \quad \forall \alpha \geq 0 \qquad \Longrightarrow \qquad y = 0.$$

特别地, 若 C 有界, 则 AC 为闭集.

▶ 闭函数: 其上方图是闭集.

- 闭集的示性函数是闭函数, 不是闭集的集合的示性函数不是闭函数.
- f 是闭函数当且仅当 f 的所有 α-下水平集都是闭集.
- 若 f 是闭函数且存在有界的下水平集, 则 f 有最小值点.
- 函数 f 的共轭函数$f^*(y) = \sup_{x \in \mathbf{dom}\, f} \{y^T x - f(x)\}$ 恒为闭凸函数.

邻近算子

▶邻近算子:

$$\mathrm{prox}_h(x) = \arg\min_{u \in \mathbf{dom}\, h} \left\{ h(u) + \frac{1}{2}\|u - x\|_2^2 \right\}.$$

Theorem 1. *若 h 为适当的闭凸函数, 则对任意 x, $\mathrm{prox}_h(x)$ 存在且唯一.*

Proof. 由$h(u) + \frac{1}{2}\|u - x\|_2^2$ 是强凸函数知, 其所有的α-下水平集有界，故由Weierstrass 定理知最小值存在. 强凸函数最小值唯一. □

邻近算子与次梯度

Theorem 2. *若 h 为适当的闭凸函数, 则* $u = \mathrm{prox}_h(x) \iff x - u \in \partial h(u)$.

Proof. 若$u = \mathrm{prox}_h(x)$, 则由最优性条件得$0 \in \partial h(u) + (u - x)$, 因此有$x - u \in \partial h(u)$.

反之, 若$x - u \in \partial h(u)$, 则由次梯度的定义可得

$$h(v) \geq h(u) + (x-u)^{\mathrm{T}}(v-u), \quad \forall v \in \mathbf{dom}\, h.$$

因此, 对任意的$v \in \mathbf{dom}\, h$有

$$\begin{aligned} h(v) + \frac{1}{2}\|v - x\|^2 &\geq h(u) + (x-u)^{\mathrm{T}}(v-u) + \frac{1}{2}\|(v-u) - (x-u)\|^2 \\ &\geqslant h(u) + \frac{1}{2}\|u - x\|^2. \end{aligned}$$

根据定义可得$u = \mathrm{prox}_h(x)$. □

- $h(x)=\|x\|_1, \quad \mathrm{prox}_{th}(x)=\textsf{sign}(x)\max\{|x|-t,0\}$

邻近算子$u=\mathrm{prox}_{th}(x)$ 的最优性条件为

$$x-u\in t\partial\|u\|_1=\begin{cases}\{t\}, & u>0\\ [-t,t], & u=0\\ \{-t\}, & u<0\end{cases}$$

当$x>t$ 时, $u=x-t$; 当$x<-t$ 时, $u=x+t$; 当$x\in[-t,t]$ 时, $u=0$. 因此,

$$u=\textbf{sign}(x)\max\{|x|-t,0\}.$$

- $h(x)=\|x\|_2, \quad \mathrm{prox}_{th}(x)=\begin{cases}\left(1-\frac{t}{\|x\|_2}\right)x, & \|x\|_2\geqslant t,\\ 0, & \text{otherwise.}\end{cases}$
- $h(x)=\frac{1}{2}x^{\mathrm{T}}Ax+b^{\mathrm{T}}x+c \ (A\succ 0), \quad \mathrm{prox}_{th}(x)=(I+tA)^{-1}(x-tb).$
- $h(x)=-\sum_{i=1}^{n}\ln x_i, \quad \mathrm{prox}_{th}(x)_i=\dfrac{x_i+\sqrt{x_i^2+4t}}{2}, \quad i=1,2,\cdots,n.$

邻近算子的计算法则

- 变量的常数倍放缩以及平移 $(\lambda \neq 0)$:

$$h(x) = g(\lambda x + a), \quad \mathrm{prox}_h(x) = \frac{1}{\lambda}\left(\mathrm{prox}_{\lambda^2 g}(\lambda x + a) - a\right).$$

- 函数及变量的常数倍放缩 $(\lambda > 0)$:

$$h(x) = \lambda g\left(\frac{x}{\lambda}\right), \quad \mathrm{prox}_h(x) = \lambda \mathrm{prox}_{\lambda^{-1} g}\left(\frac{x}{\lambda}\right).$$

- 加上线性函数:

$$h(x) = g(x) + a^{\mathrm{T}}x, \quad \mathrm{prox}_h(x) = \mathrm{prox}_g(x - a).$$

- 加上二次项 $(u > 0)$:

$$h(x) = g(x) + \frac{u}{2}\|x - a\|_2^2, \quad \mathrm{prox}_h(x) = \mathrm{prox}_{\theta g}(\theta x + (1 - \theta)a),$$

其中$\theta = \dfrac{1}{1 + u}$.

- 向量函数:

$$h\,(x,y)=\varphi_1(x)+\varphi_2(y),\quad \mathrm{prox}_h\,(x,y)=\begin{bmatrix}\mathrm{prox}_{\varphi_1}(x)\\ \mathrm{prox}_{\varphi_2}(y)\end{bmatrix}.$$

- 复合仿射映射: 设$h(x)=g(Ax+b)$, α为任意正常数．若$AA^{\mathrm{T}}=\dfrac{1}{\alpha}I$, 则

$$\mathrm{prox}_h(x)=(I-\alpha A^{\mathrm{T}}A)x+\alpha A^{\mathrm{T}}(\mathrm{prox}_{\alpha^{-1}g}(Ax+b)-b).$$

▶【例】 $h(x_1,x_2,\cdots,x_m)=g(x_1+x_2+\cdots+x_m)$ 的邻近算子为

$$\mathrm{prox}_h(x_1,x_2,\cdots,x_m)_i=x_i-\frac{1}{m}\left(\sum_{j=1}^m x_j-\mathrm{prox}_{mg}\left(\sum_{j=1}^m x_j\right)\right).$$

Moreau分解

- Moreau分解: $x = \mathrm{prox}_h(x) + \mathrm{prox}_{h^*}(x)$.

  - 共轭函数和次梯度的性质:

$$\begin{aligned} u = \mathrm{prox}_h(x) &\iff x - u \in \partial h(u) \\ &\iff u \in \partial h^*(x-u) \\ &\iff x - u = \mathrm{prox}_{h^*}(x). \end{aligned}$$

  - 子空间的正交投影的广义分解式: $x = P_L(x) + P_{L^\perp}(x)$, 其中$L$为一个子空间, $L^\perp$是它的正交补. 在Moreau分解中有$h = I_L$, $h^* = I_{L^\perp}$, 其中I表示示性函数.

- 广义Moreau分解: $x = \mathrm{prox}_{\lambda f}(x) + \lambda \mathrm{prox}_{\lambda^{-1} f^*}(x/\lambda), \quad \forall \lambda > 0.$

  - 由共轭函数的性质: $(\lambda f)^*(y) = \lambda f^*(y/\lambda)$, 及对$\lambda f$应用Moreau 分解, 可得

$$\begin{aligned} x &= \mathrm{prox}_{\lambda f}(x) + \mathrm{prox}_{(\lambda f)^*}(x) \\ &= \mathrm{prox}_{\lambda f}(x) + \lambda \mathrm{prox}_{\lambda^{-1} f^*}(x/\lambda). \end{aligned}$$

非凸函数的邻近算子

▶ 适当闭函数的邻近算子: 设 h 是适当闭函数(可以非凸), 且具有有限的下界, $\inf\limits_{x\in\mathbf{dom}h} h(x) > -\infty$, 定义 h 的邻近算子为

$$\mathrm{prox}_h(x) = \arg\min_{u\in\mathbf{dom}\,h}\left\{h(u) + \frac{1}{2}\|u-x\|^2\right\}.$$

Theorem 3. *设h 是适当闭函数且$\inf_{x\in\mathbf{dom}\,h} h(x) > -\infty$, 则对任意的$x \in \mathbf{dom}\,h$, $\mathrm{prox}_h(x)$ 是$\mathbb{R}^n$ 上的非空紧集.*

Proof. 令

$$g(u) = h(u) + \frac{1}{2}\|u-x\|^2, \quad \ell = \inf_{x\in\mathbf{dom}h} h(x).$$

取$u_0 \in \mathbf{dom}\,h$, 由于$\frac{1}{2}\|u-x\|^2$ 无上界, 故$\exists R > 0$, 从而对任意的满足$\|u-x\| > R$ 的u, 有

$$\frac{1}{2}\|u-x\|^2 > g(u_0) - \ell \quad \Rightarrow \quad g(u) > g(u_0).$$

这说明下水平集$\{u \mid g(u) \leqslant g(u_0)\}$ 含于球$\|u-x\| \leqslant R$ 内, 故g 有一个非空

有界下水平集. 显然$g(u)$ 是闭函数, 由Weierstrass 定理可知, $g(u)$ 的最小值点集合$\mathrm{prox}_h(x)$ 是非空紧集. □

Theorem 4. *设h 是适当闭函数(可非凸)且有下界, $u \in \mathrm{prox}_h(x)$, 则$x - u \in \partial h(u)$.*

▶极限次微分:设$f : \mathbb{R}^n \to (-\infty, +\infty]$ 是适当下半连续函数.

- 对给定的$x \in \mathbf{dom}\ f$, 满足如下条件的所有向量$u \in \mathbb{R}^n$ 的集合定义为f 在点x 处的*Fréchet* 次微分:

$$\liminf_{y \to x, y \neq x} \frac{f(y) - f(x) - \langle u, y - x\rangle}{\|y - x\|} \geq 0,$$

  记为$\hat{\partial} f(x)$.当$x \notin \mathbf{dom}\ f$ 时, 将$\hat{\partial} f(x)$ 定义为空集$\varnothing$.

- f 在点$x \in \mathbb{R}^n$ 处的极限次微分(或简称为次微分)定义为

$$\partial f(x) = \{u \in \mathbb{R}^n\ :\ \exists\, x^k \to x, f(x^k) \to f(x), u^k \in \hat{\partial} f(x^k) \to u\}.$$

  极限次微分通过对x 附近的点处的*Fréchet* 次微分取极限得到.

闭凸集上的投影与示性函数的邻近算子

▶ 闭凸集 C的示性函数 I_C 的邻近算子为点 x 到 C 的投影$\mathcal{P}_C(x)$:

$$\begin{aligned}\mathrm{prox}_{I_C}(x) &= \arg\min_u \left\{ I_C(u) + \frac{1}{2}\|u-x\|^2 \right\} \\ &= \operatorname*{argmin}_{u\in C} \|u-x\|^2 = \mathcal{P}_C(x).\end{aligned}$$

几何意义: $u = \mathcal{P}_C(x) \Leftrightarrow (x-u)^{\mathrm{T}}(z-u) \leqslant 0, \quad \forall z \in C.$

- 超平面 $C = \{x | a^T x = b\}$ ($a \neq 0$)

$$P_C(x) = x + \frac{b - a^T x}{\|a\|_2^2} a.$$

- 仿射集 $C = \{x | Ax = b\}$ ($A \in \mathbb{R}^{p\times n}$ 且$\mathrm{rank}(A) = p$)

$$P_C(x) = x + A^T (AA^T)^{-1} (b - Ax).$$

- 半平面 $C = \{x | a^T x \leq b\}$ $(a \neq 0)$

$$P_C(x) = \begin{cases} x + \dfrac{b - a^T x}{\|a\|_2^2} a & \text{if } a^T x > b, \\ x & \text{if } a^T b \leq b. \end{cases}$$

- 矩形 $C = [l, u] = \{l \preceq x \preceq u\}$

$$P_C(x)_i = \begin{cases} l_i & x_i \leq l_i, \\ x_i & l_i \leq x_i \leq u_i, \\ u_i & x_i \geq u_i. \end{cases}$$

- 非负象限$C = \mathbb{R}_+^n$: $P_C(x) = x_+$.
- 概率单纯形 $C = \{x | \mathbf{1}^T x = 1, x \succ 0\}$: $P_C(x) = (x - \lambda 1)_+$, 其中$\lambda$满足$\mathbf{1}^T (x - \lambda 1)_+ = \sum_{k=1}^{n} \max\{0, x_k - \lambda\} = 1$.
- 单纯形 $C = \{x | a^T x = b, l \preceq x \preceq u\}$: $P_c(x) = P_{[l,u]}(x - \lambda a)$, 其中$\lambda$满足$a^T P_{[l,u]}(x - \lambda a) = b$.

支撑函数、范数、距离的邻近算子

▶ 闭凸集C的支撑函数的共轭是其示性函数:

$$f(x) = S_C(x) = \sup_{y \in C} x^T y, \qquad f^*(y) = I_C(y).$$

支撑函数的邻近算子:

$$\begin{aligned} \text{prox}_{t\,f}(x) &= x - t\text{prox}_{t^{-1}\,f^*}(x/t) \\ &= x - tP_C(x/t). \end{aligned}$$

【例】 $f(x) = x_{[1]} + \cdots + x_{[r]} = S_C(x), \qquad C = \{y | 0 \preceq y \preceq 1, \mathbf{1}^T y = r\}.$

▶**范数的共轭**是对偶范数球的示性函数:

$$f(x) = \|c\|, \qquad f^*(x) = I_B(y), \qquad B = \{y | \|y\|_* \leq 1\}.$$

范数的邻近算子:

$$\begin{aligned} \text{prox}_{t\,f}(x) &= x - t\text{prox}_{t^{-1}\,f^*}(x/t) \\ &= x - tP_B(x/t) \\ &= x - P_{tB}(x). \end{aligned}$$

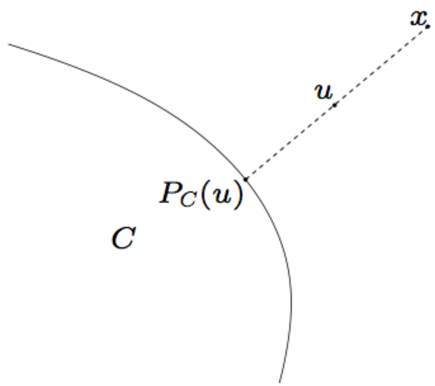
▶距离的邻近算子: 设C是闭凸集, $d(x)=\inf_{y\in C}\|x-y\|_2$的邻近算子

$$\mathrm{prox}_{td}(x)=\theta P_C(x)+(1-\theta)x, \qquad \theta=\begin{cases} t/d(x), & \text{if } d(x)\ge t,\\ 1, & \text{otherwise.}\end{cases}$$

- 若 $u=\mathrm{prox}_{td}(x)\notin C$, 则有

$$x-u=\frac{t}{d(u)}(u-P_C(u)).$$

由此可推出$P_C(u)=P_C(x)$, $d(x)\ge t$, 且 u 是 x 和 $P_C(x)$ 的加权平均



- 若 $u\in C$, 则$\min\{d(u)+\frac{1}{2t}\|u-x\|_2^2\} \;\Leftrightarrow\; \min\{\frac{1}{2t}\|u-x\|_2^2\}$. 故有 $u=P_C(x)$.

▶平方距离的邻近算子: $f(x) = d(x)^2/2$的邻近算子为

$$\operatorname{prox}_{t\,f}(x) = \frac{1}{1+t}x + \frac{t}{1+t}P_C(x).$$

Proof.

$$\begin{aligned}\operatorname{prox}_{t\,f}(x) &= \arg\min_u \left(\frac{1}{2}d(u)^2 + \frac{1}{2t}\|u-x\|_2^2\right)\\ &= \arg\min_u \inf_{v\in C}\left(\frac{1}{2}\|u-v\|_2^2 + \frac{1}{2t}\|u-x\|_2^2\right).\end{aligned}$$

最优的 u 可以看成 v 的函数：

$$u = \frac{t}{t+1}v + \frac{1}{t+1}x.$$

最优的 v 在集合 C 上极小化

$$\frac{1}{2}\left\|\frac{t}{t+1}v + \frac{1}{t+1}x - v\right\|_2^2 + \frac{1}{2t}\left\|\frac{t}{t+1}v + \frac{1}{t+1}x - x\right\|_2^2 = \frac{1}{2(1+t)}\|v-x\|_2^2.$$

由此即得, $v = P_C(x)$. $\square$

复合优化

考虑如下复合优化问题:

$$\min_{x\in\mathbb{R}^n} \quad \psi(x) = f(x) + h(x).$$

- 函数f为可微函数, $\mathbf{dom}\, f = \mathbb{R}^n$.
- 函数h为凸函数, 可以是非光滑的, 其邻近算子容易计算.
- LASSO问题: $f(x) = \frac{1}{2}\|Ax-b\|^2$, $h(x) = \mu\|x\|_1$.
- 次梯度法计算的复杂度: $\mathcal{O}(1/\sqrt{k})$

是否可以设计复杂度为$\mathcal{O}(1/k)$的算法?

近似点梯度法

►近似点梯度法的迭代格式:

$$x^{k+1} = \mathrm{prox}_{t_k h}\left(x^k - t_k \nabla f(x^k)\right), \quad (1)$$

其中 $t_k > 0$ 为每次迭代的步长, 它可以是一个常数或者由线搜索得出.

- (1)式等价于

$$\begin{aligned} x^{k+1} &= \arg\min_u \left\{ h(u) + \frac{1}{2t_k}\|u - x^k + t_k \nabla f(x^k)\|^2 \right\} \\ &= \arg\min_u \left\{ h(u) + f(x^k) + \nabla f(x^k)^{\mathrm{T}}(u - x^k) + \frac{1}{2t_k}\|u - x^k\|^2 \right\}. \end{aligned}$$

- 复合优化问题的近似点梯度法是对光滑部分f做梯度下降, 对于非光滑部分h使用邻近算子.
- 近似点梯度法可看作对光滑部分做显式的梯度下降,关于非光滑部分做隐式的梯度下降:

$$x^{k+1} = x^k - t_k \nabla f(x^k) - t_k g^k, \quad g^k \in \partial h(x^{k+1}).$$

步长选取

- 当 f 为梯度L -利普希茨连续函数 时，可取固定步长 $t_k = t \leqslant \frac{1}{L}$.当$L$ 未知时可使用线搜索准则

$$f(x^{k+1}) \leqslant f(x^k) + \nabla f(x^k)^{\mathrm{T}}(x^{k+1} - x^k) + \frac{1}{2t_k}\|x^{k+1} - x^k\|^2.$$

- 利用BB 步长作为 t_k 的初始估计并用非单调线搜索进行校正：

$$\alpha_{\mathrm{BB1}}^{k} \stackrel{\text{def}}{=} \frac{\left(s^{k-1}\right)^{\mathrm{T}} y^{k-1}}{\left(y^{k-1}\right)^{\mathrm{T}} y^{k-1}} \quad \text{或} \quad \alpha_{\mathrm{BB2}}^{k} \stackrel{\text{def}}{=} \frac{\left(s^{k-1}\right)^{\mathrm{T}} s^{k-1}}{\left(s^{k-1}\right)^{\mathrm{T}} y^{k-1}},$$

其中$s^{k-1} = x^k - x^{k-1}$ 以及 $y^{k-1} = \nabla f\left(x^k\right) - \nabla f\left(x^{k-1}\right)$.

- 可构造如下适用于近似点梯度法的非单调线搜索准则：

$$\psi(x^{k+1}) \leq C^k - \frac{c_1}{2t_k}\|x^{k+1} - x^k\|^2,$$

$c_1 \in (0,1)$ 为正常数. 注意，定义C^k时需要使用整体函数值$\psi(x^k)$.

近似点梯度法应用

►用近似点梯度法求解 LASSO 问题:

$$\min_x \quad \mu\|x\|_1 + \frac{1}{2}\|Ax - b\|^2.$$

令 $f(x) = \frac{1}{2}\|Ax - b\|^2, \quad h(x) = \mu\|x\|_1$,则

$$\nabla f(x) = A^{\mathrm{T}}(Ax - b)$$

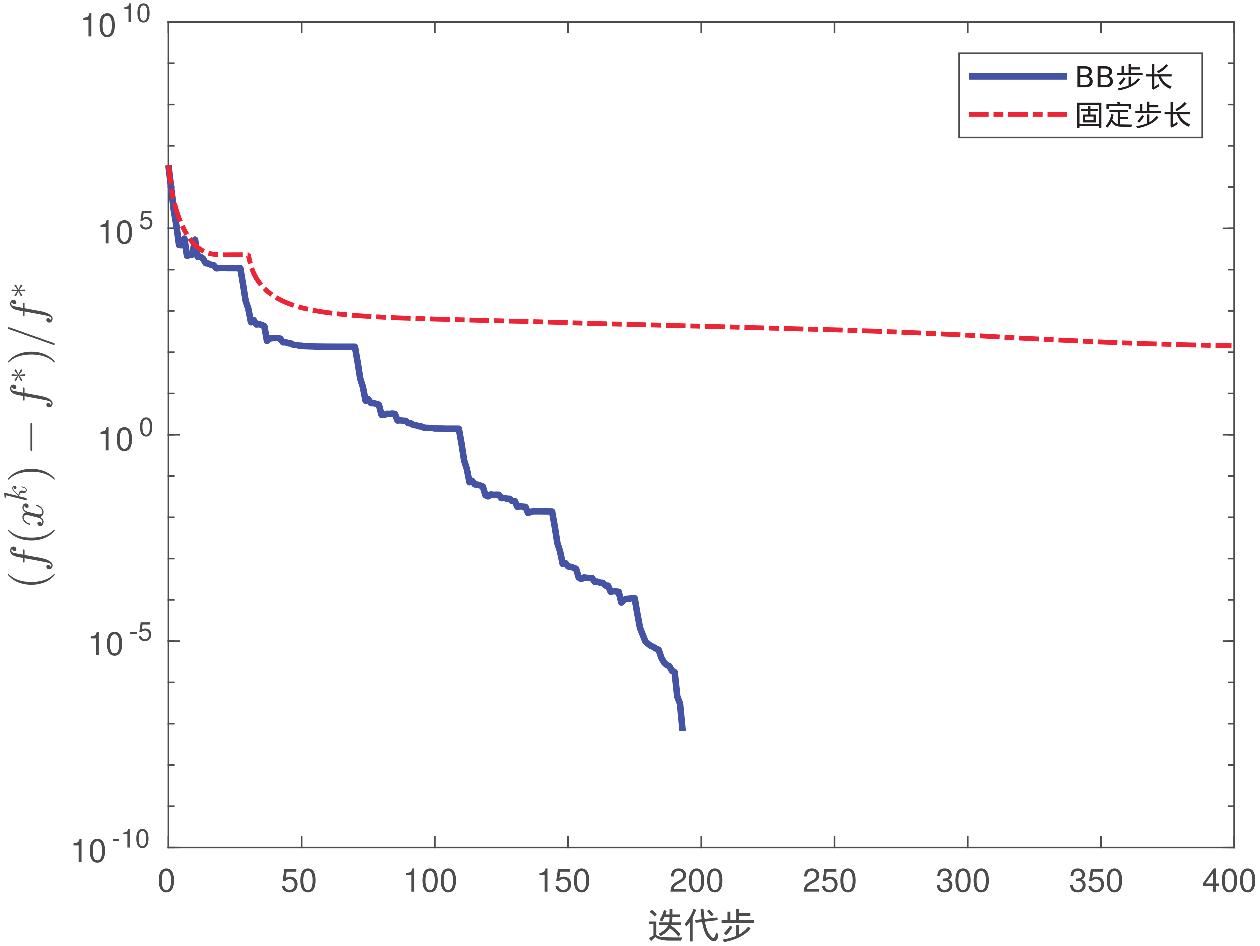
$$\operatorname{prox}_{t_k h}(x) = \operatorname{sign}(x) \max\{|x| - t_k\mu, 0\}$$

近似点梯度法求解LASSO 问题的迭代格式为:

$$y^k = x^k - t_k A^{\mathrm{T}}\left(Ax^k - b\right)$$

$$x^{k+1} = \operatorname{sign}\left(y^k\right) \max\left\{\left|y^k\right| - t_k\mu, 0\right\}$$

可以使用BB步长加速收敛



▶用近似点梯度法求解低秩矩阵恢复问题:

$$\min_{X\in\mathbb{R}^{m\times n}} \quad \mu\|X\|_* + \frac{1}{2}\sum_{(i,j)\in\Omega}(X_{ij}-M_{ij})^2 .$$

令$f(X)=\frac{1}{2}\sum_{(i,j)\in\Omega}(X_{ij}-M_{ij})^2$, $h(X)=\mu\|X\|_*$. 定义矩阵$P\in\mathbb{R}^{m\times n}$:

$$P_{ij}=\begin{cases}1, & (i,j)\in\Omega,\\ 0, & \text{otherwise.}\end{cases}$$

则

$$f(X)=\frac{1}{2}\|P\odot(X-M)\|_F^2,\quad \nabla f(X)=P\odot(X-M),$$

$$\text{prox}_{t_k h}(X)=U\text{Diag}\left(\max\{|d|-t_k\mu,0\}\right)V^{\mathrm{T}},$$

其中$X=U\,\text{Diag}(d)V^{\mathrm{T}}$为矩阵 X 的约化的奇异值分解.

近似点梯度法求解低秩矩阵恢复问题的迭代格式:

$$Y^k=X^k-t_kP\odot\left(X^k-M\right),$$

$$X^{k+1}=\text{prox}_{t_kh}\left(Y^k\right).$$

近似点梯度法收敛性分析

▶基本假设:

- f 在$\mathbb{R}^n$ 上是凸的; ∇f 为 L-利普希茨连续;
- h 是适当的闭凸函数 (因此 prox_{th} 的定义是合理的);
- 函数 $\psi(x) = f(x) + h(x)$ 的最小值 ψ^* 是有限的, 且在点x^* 处可取到(并不要求唯一).

▶梯度映射: 在基本假设的基础上, 定义**梯度映射**

$$G_t(x) = \frac{1}{t}\left(x - \text{prox}_{th}(x - t\nabla f(x))\right) \quad (t > 0) \tag{2}$$

$$\Rightarrow \quad \begin{cases} x^{k+1} = \text{prox}_{th}\left(x^k - t\nabla f(x^k)\right) = x^k - tG_t(x^k), \\ G_t(x) - \nabla f(x) \in \partial h\left(x - tG_t(x)\right), \\ G_t(x) = 0 \iff x \text{ 为 } \psi(x) = f(x) + h(x) \text{ 的最小值点.} \end{cases}$$

▶固定步长近似点梯度法的收敛性

Theorem 5. *取定步长为* $t_k = t \in \left(0, \dfrac{1}{L}\right]$, *设* $\{x^k\}$ *由迭代格式*(1)*产生, 则*

$$\psi(x^k) - \psi^* \leqslant \frac{1}{2kt} \left\|x^0 - x^*\right\|^2.$$

Proof. 根据L-光滑有二次上界, 得到

$$f(y) \leqslant f(x) + \nabla f(x)^{\mathrm{T}}(y-x) + \frac{L}{2}\|y-x\|^2, \quad \forall x, y \in \mathbb{R}^n.$$

令 $y = x - tG_t(x)$, 则由$t \in \left(0, \dfrac{1}{L}\right]$,

$$\begin{aligned} f\left(x - tG_t(x)\right) &\leqslant f(x) - t\nabla f(x)^{\mathrm{T}} G_t(x) + \frac{t^2 L}{2}\left\|G_t(x)\right\|^2 \\ &\leqslant f(x) - t\nabla f(x)^{\mathrm{T}} G_t(x) + \frac{t}{2}\left\|G_t(x)\right\|^2. \end{aligned} \tag{3}$$

由 $f(x), h(x)$ 为凸函数,

$$h\left(x-tG_t(x)\right) \leqslant h(z)-\left(G_t(x)-\nabla f(x)\right)^{\mathrm{T}}\left(z-x+tG_t(x)\right), \tag{4}$$

$$f(x) \leqslant f(z)-\nabla f(x)^{\mathrm{T}}(z-x). \tag{5}$$

将(3)(4)(5)式相加可得, 对任意 $z \in \mathbf{dom}\, \psi$ 有

$$\psi\left(x-tG_t(x)\right) \leqslant \psi(z)+G_t(x)^{\mathrm{T}}(x-z)-\frac{t}{2}\left\|G_t(x)\right\|^2. \tag{6}$$

因$x^i=x^{i-1}-tG_t(x^{i-1})$, 在(6) 中, 取 $z=x^*, x=x^{i-1}$ 得到

$$\begin{aligned}
\psi(x^i)-\psi^* &\leqslant G_t(x^{i-1})^{\mathrm{T}}\left(x^{i-1}-x^*\right)-\frac{t}{2}\left\|G_t(x^{i-1})\right\|^2 \\
&=\frac{1}{2t}\left(\left\|x^{i-1}-x^*\right\|^2-\left\|x^{i-1}-x^*-tG_t(x^{i-1})\right\|^2\right) \\
&=\frac{1}{2t}\left(\left\|x^{i-1}-x^*\right\|^2-\left\|x^i-x^*\right\|^2\right).
\end{aligned} \tag{7}$$

取 $i=1,2,\cdots,k$ 并累加, 得

$$\begin{aligned}\sum_{i=1}^{k}\left(\psi\left(x^{i}\right)-\psi^{*}\right) &\leqslant \frac{1}{2t}\sum_{i=1}^{k}\left(\left\|x^{i-1}-x^{*}\right\|^{2}-\left\|x^{i}-x^{*}\right\|^{2}\right)\\ &=\frac{1}{2t}\left(\left\|x^{0}-x^{*}\right\|^{2}-\left\|x^{k}-x^{*}\right\|^{2}\right)\\ &\leqslant \frac{1}{2t}\left\|x^{0}-x^{*}\right\|^{2}.\end{aligned}$$

在不等式(6)中, 取$z=x^{i-1}$得:

$$\psi(x^{i})\leqslant\psi(x^{i-1})-\frac{t}{2}\left\|G_t(x^{i-1})\right\|^{2}.$$

故$\{\psi(x^{i})\}$ 不增, 因此

$$\psi(x^{k})-\psi^{*}\leqslant\frac{1}{k}\sum_{i=1}^{k}\left(\psi(x^{i})-\psi^{*}\right)\leqslant\frac{1}{2kt}\left\|x^{0}-x^{*}\right\|^{2}.$$

$\square$

►**注：** 定理5中要求 $t \leq \frac{1}{L}$, 而根据定理5的证明过程, 也可以用线搜索准则：

- 从某个 $t = \hat{t} > 0$ 开始进行回溯($t \leftarrow \beta t$), 直到满足不等式

$$f\left(x - tG_t(x)\right) \leqslant f(x) - t\nabla f(x)^{\mathrm{T}} G_t(x) + \frac{t}{2}\left\|G_t(x)\right\|^2. \tag{8}$$

- 这等价于算法部分提到的线搜索准则：

$$f\left(x^{k+1}\right) \leqslant f\left(x^k\right) + \nabla f\left(x^k\right)^{\mathrm{T}}\left(x^{k+1} - x^k\right) + \frac{1}{2t_k}\left\|x^{k+1} - x^k\right\|^2.$$

由此我们解释了该线搜索准则的合理性.

▶非固定步长近似点梯度法的收敛性

Theorem 6. *从某个 $t=\hat{t}>0$ 开始进行回溯($t \leftarrow \beta t$) 直到满足不等式(8), 设 $\{x^k\}$ 是由迭代格式(1) 产生的序列, 则*

$$\psi\left(x^k\right)-\psi^* \leqslant \frac{1}{2k\min\{\hat{t},\beta/L\}}\left\|x^0-x^*\right\|^2.$$

Proof. 由定理5 的证明, 当$0<t \leqslant \frac{1}{L}$ 时, 不等式(8)成立, 故由线搜索所得的步长 t 应满足 $t \geqslant t_{\min}=\min\left\{\hat{t},\frac{\beta}{L}\right\}$. 同理,我们有$\psi\left(x^i\right)$单调不增,且

$$\psi\left(x^i\right)-\psi^* \leqslant \frac{1}{2t_{\min}}\left(\left\|x^{i-1}-x^*\right\|^2-\left\|x^i-x^*\right\|^2\right).$$

取$i=1,2,\cdots,k$并累加, 利用$\psi\left(x^i\right)$ 不增, 可得

$$\psi\left(x^k\right)-\psi^* \leqslant \frac{1}{2kt_{\min}}\left\|x^0-x^*\right\|^2.$$

□