6.3 迫近梯度法

SMaLL

[1] 中国石油大学（华东）
SMaLL 课题组
small.sem.upc.edu.cn
liangxijunsd@163.com

2023

迫近梯度法

迫近梯度法：动机

假设

$$f(x) = g(x) + h(x) \tag{1}$$

- g 是凸的，可微的，$\mathrm{dom}(g) = \mathbb{R}^n$
- h 是凸的，不一定是可微的

如果 f 是可微的，那么梯度下降更新将是：

$$x^+ = x - t \cdot \nabla f(x) \tag{2}$$

最小化 f 在 x 附近的近似值，用 $\frac{1}{t}I$ 代替 $\nabla^2 f(x)$

$$x^+ = \underset{z}{\mathrm{argmin}} \underbrace{f(x) + \nabla f(x)^T (z - x) + \frac{1}{2t}\|z - x\|_2^2}_{\bar{f}_t(z)} \tag{3}$$

迫近梯度法

模型.

$$\min_x \ f(x) = g(x) + h(x) \tag{4}$$

g 可微, f 不可微，不使用 h

Idea. 对 g 进行二次近似, 令 h 保持不变

$$\begin{aligned} x^+ &= \operatorname*{argmin}_z \ \bar{g}_t(z) + h(z) \\ &= \operatorname*{argmin}_z \ g(x) + \nabla g(x)^T(z-x) + \frac{1}{2t}\|z-x\|_2^2 + h(z) \qquad (5) \\ &= \operatorname*{argmin}_z \ \frac{1}{2t}\|z-(x-t\nabla g(x))\|_2^2 + h(z) \end{aligned}$$

$\frac{1}{2t}\|z-(x-t\nabla g(x))\|_2^2$: 对 g 梯度下降

$h(z)$: 使 h 变小

迫近梯度法♣

定义 迫近映射:

$$\text{prox}_{h,t}(x) = \underset{z}{\text{argmin}} \frac{1}{2t}\|x-z\|_2^2 + h(z) \tag{6}$$

迫近梯度下降: 选取初始点 $x^{(0)}$, 迭代:

$$x^{(k)} = \text{prox}_{h,t_k}\left(x^{(k-1)} - t_k \nabla g\left(x^{(k-1)}\right)\right), \quad k = 1, 2, 3, \ldots \tag{7}$$

使此更新步骤看起来熟悉，写为

$$x^{(k)} = x^{(k-1)} - t_k \cdot G_{t_k}\left(x^{(k-1)}\right) \tag{8}$$

其中 G_t: f 的广义梯度,

$$G_t(x) = \frac{x - \text{prox}_{h,t}(x - t\nabla g(x))}{t}$$

这有什么好处?

观察: 似乎一个最小化问题 $\rightarrow$ 另一个最小化问题

要点: $\mathrm{prox}_{h,t}(\cdot)$ 对许多重要函数 h 都有解析解

- 映射 $\mathrm{prox}_{h,t}(\cdot)$ 不依赖于 g, 只依赖于 h
- g 的平滑部分可能很复杂 $\rightarrow$ 只需要计算其梯度
- 每次迭代计算 $\mathrm{prox}_{h,t}(\cdot)$ 一次

迫近梯度法

Sec. 1 收敛性分析: g 是强凸的

Theorem 1 (收敛)

假设

- $g: \mathbb{R}^d \to \mathbb{R}$ 是连续可微的, 具有常数 $c > 0$ 的强凸
- ∇g: 常数 $L > 0$ *Lipschitz* 连续
- $\alpha_k = \alpha \in (0, 1/L]$ 对于任意的 $k \in \mathbb{N}$.

$\Rightarrow$ 对于所有 $k \in \mathbb{N}$, 产生的迭代点列的函数值满足

$$f\left(x^{k+1}\right) - f(x^*) \leq (1 - \alpha c)^k \left(f(x^1) - f(x^*)\right)$$

其中 $x^* \in \mathbb{R}^d$: f 的唯一全局最优解.

g 强凸时, $f\left(x^k\right)$ 线性收敛

Sec. 1 收敛性分析: g 是强凸的

证明.

因为 $\alpha_k = \alpha \in (0, 1/L]$,

$$
\begin{aligned}
f\left(x^{k+1}\right) &= g\left(x^{k+1}\right) + h\left(x^{k+1}\right) \\
&\leq g\left(x^{k}\right) + \nabla g\left(x^{k}\right)^{T}\left(x^{k+1} - x^{k}\right) + \frac{1}{2}L\left\|x^{k+1} - x^{k}\right\|_2^2 + h\left(x^{k+1}\right) \\
&\leq g\left(x^{k}\right) + \nabla g\left(x^{k}\right)^{T}\left(x^{k+1} - x^{k}\right) + \frac{1}{2\alpha}\left\|x^{k+1} - x^{k}\right\|_2^2 + h\left(x^{k+1}\right) \\
&\leq g\left(x^{k}\right) + \nabla g\left(x^{k}\right)^{T}\left(w - x^{k}\right) + \frac{1}{2\alpha}\left\|w - x^{k}\right\|_2^2 + h(w)
\end{aligned}
$$

对于所有$w \in \mathbb{R}^d$

取 $w = x^k + d$, 得到

$$
\begin{aligned}
f\left(x^{k+1}\right) &\leq g\left(x^k\right) + \nabla g\left(x^k\right)^T d + \frac{1}{2\alpha}\|d\|_2^2 + h\left(x^k + d\right) \\
&\leq g\left(x^k\right) + \nabla g\left(x^k\right)^T d + \frac{1}{2}c\|d\|_2^2 - \frac{1}{2}c\|d\|_2^2 + \frac{1}{2\alpha}\|d\|_2^2 + h(x^k + d) \\
&\leq g\left(x^k + d\right) + h\left(x^k + d\right) - \frac{1}{2}c\|d\|_2^2 + \frac{1}{2\alpha}\|d\|_2^2 \\
&= f\left(x^k + d\right) + \frac{1}{2}\left(\frac{1}{\alpha} - c\right)\|d\|_2^2
\end{aligned}
$$

- 其中 $d = -\alpha c\left(x^k - x^*\right)$ 意味着
-

$$
\begin{aligned}
f\left(x^{k+1}\right) &\leq f\left(x^k - \alpha c\left(x^k - x^*\right)\right) + \frac{1}{2}\left(\frac{1}{\alpha} - c\right)\left\|\alpha c\left(x^k - x^*\right)\right\|_2^2 \\
&= f\left(x^k - \alpha c\left(x^k - x^*\right)\right) + \frac{1}{2}\alpha c^2(1 - \alpha c)\left\|x^k - x^*\right\|_2^2
\end{aligned} \tag{9}
$$

另一方面，f 是 c-强凸函数 $\rightarrow$

- $$f(\tau w+(1-\tau)\bar{w}) \leq \tau f(w)+(1-\tau)f(\bar{w})-\tfrac{1}{2}c\tau(1-\tau)\|w-\bar{w}\|_2^2$$
$$\text{对于所有}(w,\bar{w},\tau)\in\mathbb{R}^d\times\mathbb{R}^d\times[0,1] \tag{10}$$

$\Rightarrow$ (考虑 $\bar{w}=x^k, w=x^*$, 和 $\tau=\alpha c\in(0,1]$)

- $$\begin{aligned} f(x^k-\alpha c(x^k-x^*)) &\leq \alpha c f(x^*)+(1-\alpha c)f(x^k)-\frac{c}{2}\alpha c(1-\alpha c)\left\|x^k-x^*\right\|_2^2 \\ &= \alpha c f(x^*)+(1-\alpha c)f\left(x^k\right)-\frac{1}{2}\alpha c^2(1-\alpha c)\left\|x^k-x^*\right\|_2^2 \end{aligned} \tag{11}$$

合并并减去 $f(x^*) \rightarrow f(x^{k+1})-f(x^*) \leq (1-\alpha c)\left(f(x^k)-f(x^*)\right)$

□

对光滑的强凸函数：迫近梯度法与梯度下降法收敛速度相似.

Sec. 2 收敛性分析: g 是凸函数

对于 $f(x) = g(x) + h(x)$,

Theorem 2 (定理)

假设

- g 是凸函数, 可微的, $\mathrm{dom}(g) = \mathbb{R}^n$, 且 ∇g 是 $Lipschitz$ 连续的, $Lipschitz$ 常数 $L > 0$
- h 是凸函数, $\mathrm{prox}_t(x) = \mathrm{argmin}_z \{\|x - z\|_2^2/(2t) + h(z)\}$ 可有效估计
- 固定步长的迫近梯度下降 $t \leq 1/L$

$$\Rightarrow \qquad f\left(x^{(k)}\right) - f^\star \leq \frac{\left\|x^{(0)} - x^\star\right\|_2^2}{2tk}$$

且同样的结果适用于回溯, t 将被 β/L 代替.

g 是凸函数: 迫近梯度下降法收敛速度 $O(1/k)$ 或 $O(1/\epsilon)$

迫近梯度法

示例: ISTA♣♣

取 $y \in \mathbb{R}^n, X \in \mathbb{R}^{n \times p}$, 回忆 Lasso 模型:

$$f(\beta) = \underbrace{\frac{1}{2}\|y - X\beta\|_2^2}_{g(\beta)} + \underbrace{\lambda\|\beta\|_1}_{h(\beta)} \tag{12}$$

迫近映射:♣♣

$$\begin{aligned} \text{prox}_t(\beta) &= \underset{z}{\text{argmin}} \frac{1}{2t}\|\beta - z\|_2^2 + \lambda\|z\|_1 \\ &= S_{\lambda t}(\beta) \end{aligned} \tag{13}$$

其中 $S_\lambda(\beta)$ 是软阈值算子,

$$[S_\lambda(\beta)]_i = \begin{cases} \beta_i - \lambda & \text{如果} \beta_i > \lambda \\ 0 & \text{如果} -\lambda \le \beta_i \le \lambda, \quad i = 1, \ldots, n \\ \beta_i + \lambda & \text{如果} \beta_i < -\lambda \end{cases}$$

分析

$$
\begin{aligned}
& \bar{\beta} \overset{\text{def}}{=} \text{prox}_t(\beta) = \underset{z}{\text{argmin}} \tfrac{1}{2t}\|\beta - z\|_2^2 + \lambda\|z\|_1 \\
\Leftrightarrow \quad & \bar{\beta}_i = \underset{z_i}{\text{argmin}} \tfrac{1}{2t}(\beta_i - z_i)^2 + \lambda|z_i| \overset{\text{def}}{=} \phi(z_i) \\
\Leftrightarrow \quad & 0 \in \partial\phi(\bar{\beta}_i) \\
\Leftrightarrow \quad & 0 \in -\tfrac{1}{t}(\beta_i - \bar{\beta}_i) + \lambda\partial|t|_{t=\bar{\beta}_i}
\end{aligned}
\tag{14}
$$

其中

$$
\partial|t|_{t=\bar{\beta}_i} = \left\{ \begin{array}{cc} 1, & \bar{\beta}_i > 0 \\ [-1,1], & \bar{\beta}_i = 0 \\ -1, & \bar{\beta}_i < 0 \end{array} \right\}
$$

$\Rightarrow$ 如果 $\bar{\beta}_i \neq 0,\ sgn(\beta_i - \bar{\beta}_i) = sgn(\bar{\beta}_i)$

(i) 如果 $\beta_i > \lambda t > 0$, 那么 $\bar{\beta}_i > 0$.

如 (14), $0 = -\frac{1}{t}(\beta_i - \bar{\beta}_i) + \lambda \cdot 1 \Rightarrow \bar{\beta}_i = \beta_i - \lambda t$

(ii) 类似的，如果 $\beta_i < -\lambda t < 0,\ \bar{\beta}_i = \beta_i + \lambda t$

(iii) if $-\lambda t \leq \beta_i \leq \lambda t,\ \bar{\beta}_i = 0$

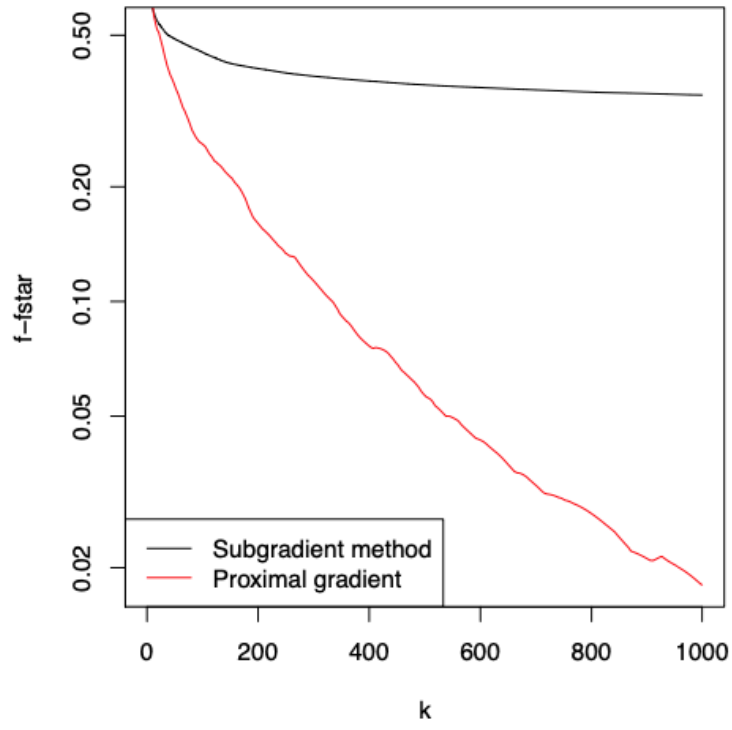
示例: ISTA♣♣

回忆 $\nabla g(\beta) = -X^T(y - X\beta) \rightarrow$ 迫近梯度更新:♣♣

$$\beta^+ = S_{\lambda t}\left(\beta + tX^T(y - X\beta)\right) \tag{15}$$

通常称为 迭代软阈值算法 (ISTA) $\beta^+ = S_{\lambda t}(\beta - t\nabla g(\beta))$

ISTA 和次梯度法的收敛曲线 (纵横: $f(x^k) - f^*$)



示例：矩阵填充

- 矩阵 $Y \in \mathbb{R}^{m \times n}$, 且已知元素 $Y_{ij}, (i,j) \in \Omega$
- 想要补上缺失的元素（例如，推荐系统）→矩阵填充问题:

$$\min_B \frac{1}{2} \sum_{(i,j) \in \Omega} (Y_{ij} - B_{ij})^2 + \lambda \|B\|_{\text{tr}} \tag{16}$$

- $\|B\|_{\text{tr}}$ 是 B 的迹范数 (或核范数):

$$\|B\|_{\text{tr}} = \sum_{i=1}^{r} \sigma_i(B) \tag{17}$$

其中 $r = \text{rank}(B)$, $\sigma_1(X) \geq \cdots \geq \sigma_r(X) \geq 0$ 是 X 的奇异值

示例：矩阵填充

定义 P_Ω 为观测集 Ω 上的投影算子:

$$[P_\Omega(B)]_{ij} = \begin{cases} B_{ij} & (i,j) \in \Omega \\ 0 & (i,j) \notin \Omega \end{cases} \tag{18}$$

那么

$$f(B) = \underbrace{\frac{1}{2}\|P_\Omega(Y) - P_\Omega(B)\|_F^2}_{g(B)} + \underbrace{\lambda\|B\|_{\mathrm{tr}}}_{h(B)} \tag{19}$$

迫近梯度下降法所需的两个运算:

- 梯度计算: $\nabla g(B) = -(P_\Omega(Y) - P_\Omega(B))$
- 迫近算子:

$$\mathrm{prox}_t(B) = \underset{Z}{\mathrm{argmin}} \frac{1}{2t}\|B - Z\|_F^2 + \lambda\|Z\|_{\mathrm{tr}} \tag{20}$$

示例：矩阵填充

定理 1

- $\text{prox}_t(B) = S_{\lambda t}(B)$,

  其中 $S_{\lambda t}(B)$：关于参数 λ 的软阈值矩阵算子，$S_\lambda(B)$ 定义为

$$S_\lambda(B) = U\Sigma_\lambda V^T \tag{21}$$

  $B = U\Sigma V^T$ 是 SVD，且 Σ_λ 是其对角阵

$$(\Sigma_\lambda)_{ii} = \max\{\Sigma_{ii} - \lambda, 0\} \tag{22}$$

示例：矩阵填充

证明.

- 记 $\text{prox}_t(B) = Z$, 其中 Z 满足

$$0 \in Z - B + \lambda t \cdot \partial \|Z\|_{\text{tr}} \tag{23}$$

- 有用的结论: 若 $Z = U\Sigma V^T \Rightarrow$

$$\partial \|Z\|_{\text{tr}} = \left\{ UV^T + W : \|W\|_{\text{op}} \leq 1, U^T W = 0, WV = 0 \right\} \tag{24}$$

- 可以验证：(23) 成立

□

迫近梯度更新步骤:

$$B^+ = S_{\lambda t}\left(B + t(P_\Omega(Y) - P_\Omega(B))\right) \tag{25}$$

示例：矩阵填充

- 可验证 $\nabla g(B)$ Lipschitz 连续，Lipschitz 常数 $L = 1$
- 固定步长 $t = 1$
- $\rightarrow$ 更新公式:

$$B^{+} = S_{\lambda}\left(P_{\Omega}(Y) + P_{\Omega^{\perp}}(B)\right) \tag{26}$$

  其中 $\Omega^{\perp}$ 未观测集, $P_{\Omega}(B) + P_{\Omega^{\perp}}(B) = B$
- 简单有效迭代算法 $\rightarrow$ 实现矩阵填充

迫近梯度法

特殊情况

迫近梯度下降也称为 复合梯度下降或广义梯度下降

为什么” 广义”?

迫近梯度下降: $\min f = g + h$:

- $h = 0$: gradient descent (梯度下降法)
- $h = I_C$: gradient projection method (梯度投影法)
- $g = 0$: proximal minimization algorithm (迫近极小化算法)

梯度投影法 ♣

- 闭凸集 $C \in \mathbb{R}^n$, $h(x) \equiv 0$

$$\min_{x \in C} g(x) + 0 \iff \min_{x} g(x) + I_C(x) \tag{27}$$

其中 $I_C(x) = \begin{cases} 0 & x \in C \\ \infty & x \notin C \end{cases}$ 是 C 的指示函数

-

$$\begin{aligned} \text{prox}_t(x) &= \underset{z}{\text{argmin}} \frac{1}{2t}\|x - z\|_2^2 + I_C(z) \\ &= \underset{z \in C}{\text{argmin}} \|x - z\|_2^2 \end{aligned} \tag{28}$$

- $\rightarrow$ 此种情况下，$\text{prox}_t(x) = P_C(x)$ 到 C 的投影算子

$$\begin{aligned} x^{(k)} &= \text{prox}_{h, t_k}\left(x^{(k-1)} - t_k \nabla g\left(x^{(k-1)}\right)\right) \\ &= P_C\left(x^{(k-1)} - t_k \nabla g\left(x^{(k-1)}\right)\right) \end{aligned} \tag{29}$$

迫近梯度下降 $\rightarrow$ 梯度投影法

迫近极小化算法

- 设 $g = 0$, h 是凸的 (不一定是可微的):

$$\min_x h(x) \tag{30}$$

- 迫近梯度法更新步骤:

$$x^+ = \operatorname*{argmin}_z \ \frac{1}{2t}\|x - z\|_2^2 + h(z) \tag{31}$$

  称作 迫近极小化算法 (proximal minimization algorithm)

- 比次梯度方法快，但通常未必方便实现 $\rightarrow$ 除非知道 (31) 封闭形式的解

不能准确计算迫近算子?

迫近梯度法: $f = g + h$, 假设迫近算子是准确计算:

$$\mathrm{prox}_t(x) = \underset{z}{\mathrm{argmin}} \frac{1}{2t}\|x - z\|_2^2 + h(z) \tag{32}$$

可精确求解

Q. 如果只能计算近似解?

A. 如果能够控制逼近迫近算子的误差 $\rightarrow$ 原始的收敛速度

迫近梯度法

加速迫近梯度法

- 考虑

$$\min_x g(x) + h(x) \tag{33}$$

其中 g 是凸的, 可微的且 h 也是凸函数

- 加速迫近梯度法:

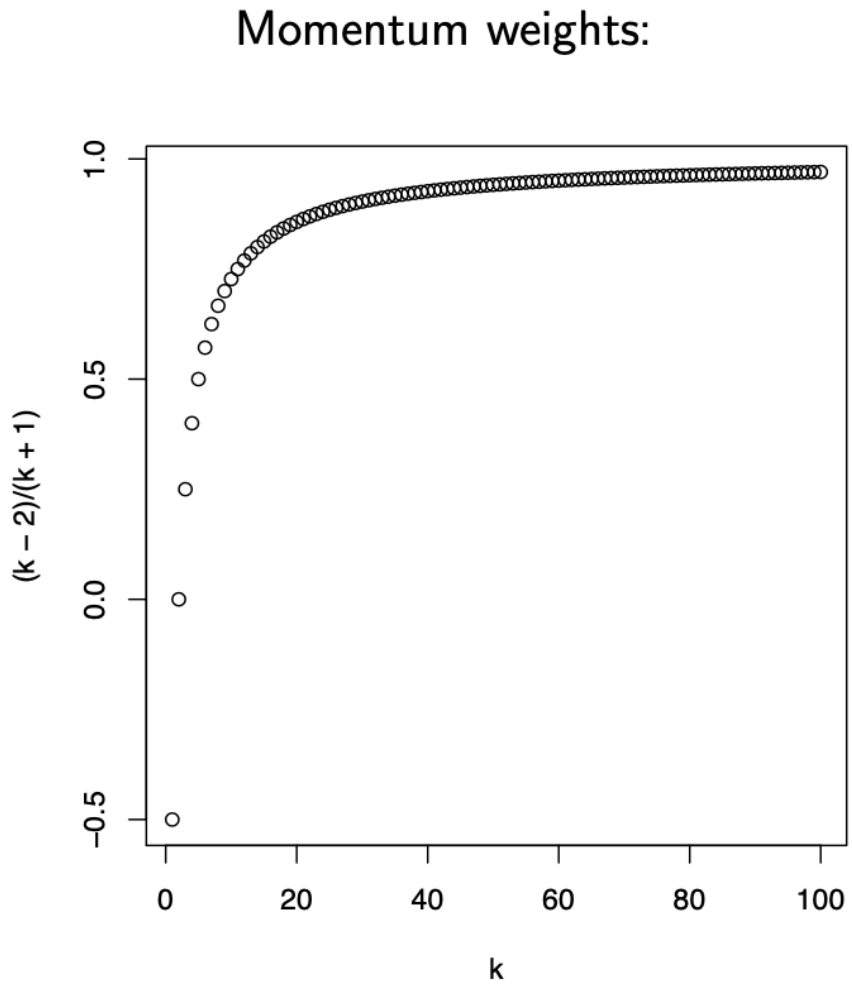
取初始点 $x^{(0)} = x^{(-1)} \in \mathbb{R}^n$,

For $k = 1, 2, 3, \ldots$

$$\begin{aligned} v &= x^{(k-1)} + \frac{k-2}{k+1}\left(x^{(k-1)} - x^{(k-2)}\right) \\ x^{(k)} &= \mathrm{prox}_{t_k}\left(v - t_k \nabla g(v)\right) \end{aligned} \tag{34}$$

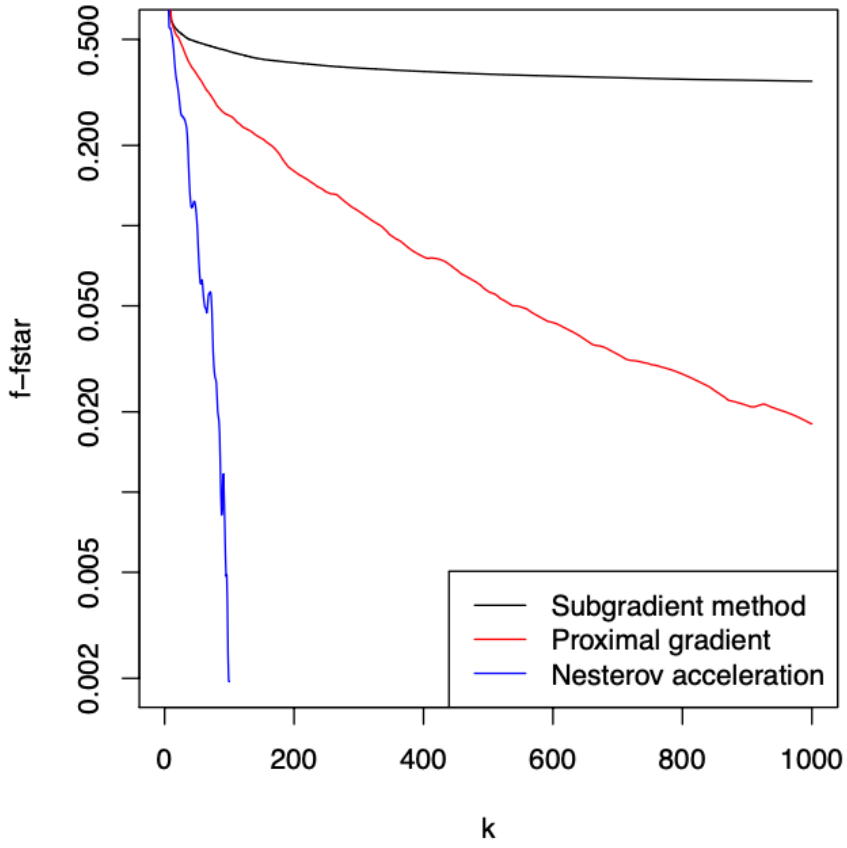
  - $v = x^{(k-1)} + \frac{k-2}{k+1}\left(x^{(k-1)} - x^{(k-2)}\right)$: 从以前的迭代中获得一些“动力”
  - 当 $h = 0 \to$ 加速梯度法.

加速迫近梯度法



加速迫近梯度法

回到 lasso 的例子：加速真的很有帮助!



注. 加速迫近梯度法的迭代函数值未必是单调下降的

收敛性分析

对于 $f(x) = g(x) + h(x)$, 假设同前:

- g 是凸的, 可微的, $\mathrm{dom}(g) = \mathbb{R}^n$, 且 ∇g Lipschitz 连续, Lipschitz 常数 $L > 0$
- h 是凸的, $\mathrm{prox}_t(x) = \mathrm{argmin}_z \{\|x - z\|_2^2/(2t) + h(z)\}$ 可有效估计

Theorem 3

固定步长的加速迫近梯度法 $t \leq 1/L$ 满足

$$f\left(x^{(k)}\right) - f^\star \leq \frac{2\left\|x^{(0)} - x^\star\right\|_2^2}{t(k+1)^2}$$

且同样的结果适用于回溯, t 被 β/L 取代

- 对于一阶方法实现 最佳速率 $O\left(1/k^2\right)$ 或 $O(1/\sqrt{\epsilon})$

FISTA

- 回到 lasso 问题:

$$\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \tag{35}$$

- 回忆 ISTA (迭代软阈值算法):

$$\beta^{(k)} = S_{\lambda t_k}\left(\beta^{(k-1)} + t_k X^T\left(y - X\beta^{(k-1)}\right)\right), \quad k = 1, 2, 3, \ldots \tag{36}$$

  $S_\lambda(\cdot)$ 是向量软阈值算子

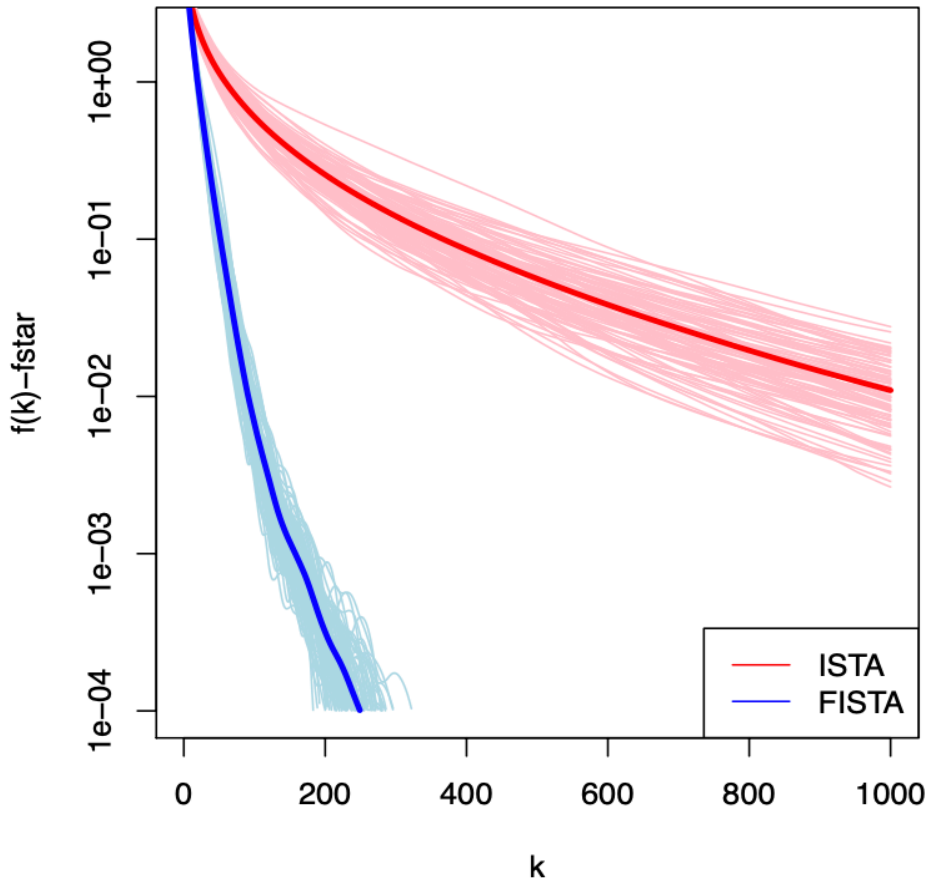
- 使用加速技巧 FISTA (F: Fast)

  对于 $k = 1, 2, 3, \ldots$

$$\begin{aligned} v &= \beta^{(k-1)} + \frac{k-2}{k+1}\left(\beta^{(k-1)} - \beta^{(k-2)}\right) \\ \beta^{(k)} &= S_{\lambda t_k}\left(v + t_k X^T(y - Xv)\right) \end{aligned} \tag{37}$$

FISTA

Lasso 回归: 100 个样本 (n = 100, p = 500):



加速总是有用的吗?

有时回溯和加速可能是 不利的!

回顾矩阵填充问题: 迫近梯度更新:

$$B^+ = S_\lambda\left(B + t(P_\Omega(Y) - P_{\Omega^\perp}(B))\right) \tag{38}$$

其中 S_λ 是矩阵软阈值算子 ... 要实施奇异值分解

加速总是有用的吗?

- 一个回溯循环在要 t 的不同值上估计 prox $\rightarrow$ 对于矩阵填充，意味着多个 SVD
- 加速技巧改变了传递给 prox 的参数: $v - t\nabla g(v)$ 代替了 $x - t\nabla g(x)$. 对于矩阵填充 (且 $t = 1$)，

$$B - \nabla g(B) = \underbrace{P_\Omega(Y)}_{\text{sparse}} + \underbrace{P_{\Omega^\perp}(B)}_{\text{low rank}} \Rightarrow \text{ fast SVD}$$

$$V - \nabla g(V) = \underbrace{P_\Omega(Y)}_{\text{sparse}} + \underbrace{P_{\Omega^\perp}(V)}_{\text{not necessarily low rank}} \Rightarrow \text{ slow SVD} \tag{39}$$

作业

1. 记矩阵 $Y \in \mathbb{R}^{m \times n}$，针对矩阵填充问题，自主设定已知元素 $Y_{ij}, (i,j) \in \Omega$，编程实现迫近梯度算法，求解相应模型，实现矩阵缺失元素填充。

参考文献和进一步阅读

Extensions and/or analyses:

- A. Beck and M. Teboulle (2008), "A fast iterative shrinkage-thresholding algorithm for linear inverse problems"
- S. Becker and J. Bobin and E. Candes (2009), "NESTA: a fast and accurate first-order method for sparse recovery"
- P. Tseng (2008), "On accelerated proximal gradient methods for convex-concave optimization"

参考文献和进一步阅读

Helpful lecture notes/books:

- E. Candes, Lecture notes for Math 301, Stanford University, Winter 2010 – 2011
- Y. Nesterov (1998), "Introductory lectures on convex optimization: a basic course", Chapter 2
- L. Vandenberghe, Lecture notes for EE 236C, UCLA, Spring 2011 – 2012