第五章 无约束优化算法

修贤超

https://xianchaoxiu.github.io

目录

引言: 无约束可微优化算法

- 考虑无约束优化问题

$$\min_{x\in\mathbb{R}^n} \quad f(x)$$

- 线搜索 $x^{k+1} = x^k + \alpha_k d^k$
  - 先确定下降方向: 负梯度、牛顿方向、拟牛顿方向等
  - 按某种准则搜索步长
- 信赖域 $z^k = x^k + d^k$

$$d^k = \arg\min_d \quad (g^k)^\top d + d^\top B d \quad \text{s.t.} \quad \|d\|_2 \le \Delta_k$$

  - 给定信赖域半径 (步长)Δ_k, 构造信赖域子问题求解方向 d^k
  - 如果 z^k 满足下降性条件，则 $x^{k+1} = z^k$, 否则 $x^{k+1} = x^k$ 更新 Δ_k

线搜索算法: 盲人下山

- 求解$f(x)$ 的最小值点如同盲人下山, 无法一眼望知谷底, 而是
  - 首先确定下一步该向哪一方向行走
  - 再确定沿着该方向行走多远后停下以便选取下一个下山方向
- 线搜索类算法的数学表述

$$x^{k+1} = x^k + \alpha_k d^k$$

  - α_k 为步长
  - d^k 为下降方向, 即 $(d^k)^\top \nabla f(x^k) < 0$
- 关键是如何选取一个好的方向 $d^k \in \mathbb{R}^n$ 以及合适的步长 α_k

α_k 的选取: 精确线搜索算法

- 首先构造一元辅助函数

$$\phi(\alpha) = f(x^k + \alpha d^k)$$

  其中 d^k 是给定的下降方向, $\alpha > 0$ 是该辅助函数的自变量

- 线搜索的目标是选取合适的 α_k 使得 $\phi(\alpha_k)$ 尽可能减小
  - α_k 应该使得 f 充分下降
  - 不应在寻找 α_k 上花费过多的计算量
- 一个自然的想法是寻找 α_k 使得

$$\alpha_k = \underset{\alpha>0}{} \phi(\alpha)$$

- 称为精确线搜索算法, 在实际应用中较少使用

例 5.1

- 考虑一维无约束优化问题

$$\min_x \quad f(x) = x^2$$

- 迭代初始点 $x^0 = 1$, 下降方向只有 $\{-1, +1\}$ 两种. 选取 $d^k = -\text{sign}(x^k)$, 且只要求选取的步长满足迭代点处函数值单调下降, 即 $f(x^k + \alpha_k d^k) < f(x^k)$
- 考虑选取如下两种步长

$$\alpha_{k,1} = \frac{1}{3^{k+1}}, \quad \alpha_{k,2} = 1 + \frac{2}{3^{k+1}}$$

- 通过简单计算可以得到

$$x_1^k = \frac{1}{2}(1 + \frac{1}{3^k}), \quad x_2^k = \frac{(-1)^k}{2}(1 + \frac{1}{3^k})$$

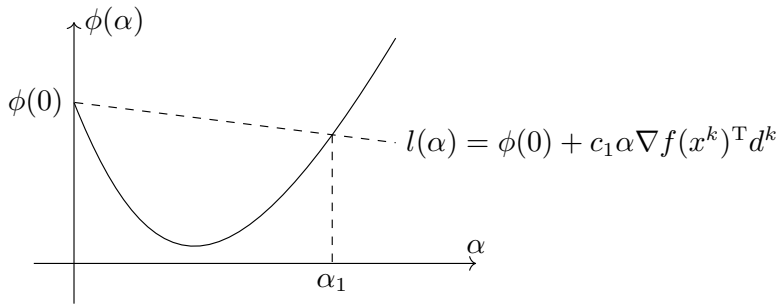
- 序列 $\{f(x_1^k)\}$ 和序列 $\{f(x_2^k)\}$ 均单调下降, 但序列 $\{x_1^k\}$ 收敛的点不是极小值点, 序列 $\{x_2^k\}$ 则在原点左右振荡, 不存在极限

非精确线搜索: Armijo 准则

- **定义 5.1** 设d^k 是点 x^k 处的下降方向, 若

$$f(x^k + \alpha d^k) \leq f(x^k) + c_1 \alpha \nabla f(x^k)^\top d^k$$

则称步长 α 满足Armijo 准则, 其中 $c_1 \in (0, 1)$ 是一个常数



Armijo 准则: 评注

- 引入 Armijo 准则的目的是保证每一步迭代充分下降
- Armijo 准则有直观的几何含义, 它指的是点 $(\alpha, \phi(\alpha))$ 必须在直线

$$l(\alpha) = \phi(0) + c_1 \alpha \nabla f(x^k)^\top d^k$$

  的下方, 上图中区间 $[0, \alpha_1]$ 中的点均满足 Armijo 准则
- 参数 c_1 通常选为一个很小的正数, 例如 $c_1 = 10^{-3}$
- Armijo 准则需要配合其他准则以保证迭代的收敛性, 反例 $\alpha = 0$

回退法: 以 Armijo 准则为例

- 给定初值 $\hat{\alpha}$, 回退法通过不断以指数方式缩小试探步长, 找到第一个满足 Armijo 准则的点
- 回退法选取

$$\alpha_k = \gamma^{j_0}\hat{\alpha}$$

其中 $j_0 = \min\{j \mid f(x^k + \gamma^j \hat{\alpha} d^k) \leq f(x^k) + c_1\gamma^j\hat{\alpha}\nabla f(x^k)^\top d^k\}, \gamma \in (0,1)$

==========

1 选择初始步长 $\hat{\alpha}$, 参数 $\gamma, c \in (0,1)$. 初始化 $\alpha \leftarrow \hat{\alpha}$
2 **while** $f(x^k + \alpha d^k) > f(x^k) + c\alpha\nabla f(x^k)^\top d^k$ **do**
3 令 $\alpha \leftarrow \gamma\alpha$
4 **end while**
5 输出 $\alpha_k = \alpha$

Wolfe 准则

- **定义 5.2** 设d^k 是点 x^k 处的下降方向, 若

$$f(x^k + \alpha d^k) \leq f(x^k) + c\alpha \nabla f(x^k)^\top d^k,$$
$$f(x^k + \alpha d^k) \geq f(x^k) + (1-c)\alpha \nabla f(x^k)^\top d^k$$

则称步长 α 满足 Goldstein 准则, 其中 $c \in (0, \frac{1}{2})$

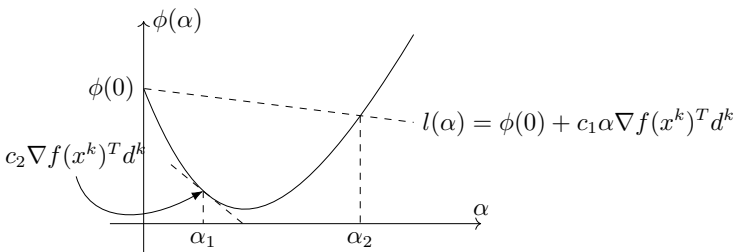
- **定义 5.3** 设 d^k 是点 x^k 处的下降方向, 若

$$f(x^k + \alpha d^k) \leq f(x^k) + c_1\alpha \nabla f(x^k)^\top d^k,$$
$$\nabla f(x^k + \alpha d^k)^\top d^k \geq c_2 \nabla f(x^k)^\top d^k$$

则称步长 α 满足 Wolfe 准则, 其中 $c_1, c_2 \in (0,1)$ 为给定的常数且 $c_1 < c_2$

Wolfe 准则

- $\nabla f(x^k + \alpha d^k)^{\mathrm{T}} d^k$ 恰好就是 $\phi(\alpha)$ 的导数, Wolfe 准则实际要求 $\phi(\alpha)$ 在点 α 处切线的斜率不能小于 $\phi'(0)$ 的 c_2 倍
- $\phi(\alpha)$ 的极小值点 α^* 处有 $\phi'(\alpha^*) = \nabla f(x^k + \alpha^* d^k)^{\mathrm{T}} d^k = 0$, 因此 α^* 永远满足条件二. 而选择较小的 c_1 可使得 α^* 同时满足条件一, 即 Wolfe 准则在绝大多数情况下会包含线搜索子问题的精确解



Zoutendijk 定理

- **定理 5.1** 考虑一般的迭代格式 $x^{k+1} = x^k + \alpha_k d^k$, 其中 d^k 是搜索方向, α_k 是步长, 且在迭代过程中 Wolfe 准则满足. 假设目标函数 f 下有界、连续可微且梯度 L -利普希茨连续, 即

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall\, x, y \in \mathbb{R}^n$$

那么

$$\sum_{k=0}^{\infty} \cos^2 \theta_k \|\nabla f(x^k)\|^2 < +\infty$$

其中 $\cos\theta_k$ 为负梯度 $-\nabla f(x^k)$ 和下降方向 d^k 夹角的余弦, 即

$$\cos\theta_k = \frac{-\nabla f(x^k)^\top d^k}{\|\nabla f(x^k)\|\|d^k\|}$$

这个不等式也被称为Zoutendijk 条件

线搜索算法的收敛性

- **推论 5.1** 对于迭代法 $x^{k+1} = x^k + \alpha_k d^k$, 设 θ_k 为每一步负梯度 $-\nabla f(x^k)$ 与下降方向 d^k 的夹角, 并假设对任意的 k, 存在常数 $\gamma > 0$, 使得

$$\theta_k < \frac{\pi}{2} - \gamma$$

则在 Zoutendijk 定理成立的条件下, 有

$$\lim_{karrow\infty} \nabla f(x^k) = 0$$

证明 假设结论不成立, 即存在子列 $\{k_l\}$ 和正常数 $\delta > 0$, 使得

$$\|\nabla f(x^{k_l})\| \geq \delta, \quad l = 1, 2, \cdots$$

线搜索算法收敛性的证明

根据 θ_k 的假设, 对任意的 k,

$$\cos\theta_k > \sin\gamma > 0$$

仅考虑 Zoutendijk 条件中第 k_l 项的和, 有

$$\sum_{k=0}^{\infty}\cos^2\theta_k\|\nabla f(x^k)\|^2 \geq \sum_{l=1}^{\infty}\cos^2\theta_{k_l}\|\nabla f(x^{k_l})\|^2$$

这显然和 Zoutendijk 定理矛盾. 因此必有

$$\lim_{karrow\infty}\nabla f(x^k) = 0$$

目录

梯度下降法

- 注意到 $\phi(\alpha) = f(x^k + \alpha d^k)$ 有泰勒展开

$$\phi(\alpha) = f(x^k) + \alpha \nabla f(x^k)^\top d^k + \mathcal{O}(\alpha^2 \|d^k\|^2)$$

- 由柯西不等式, 当 α 足够小时取 $d^k = -\nabla f(x^k)$ 会使函数下降最快

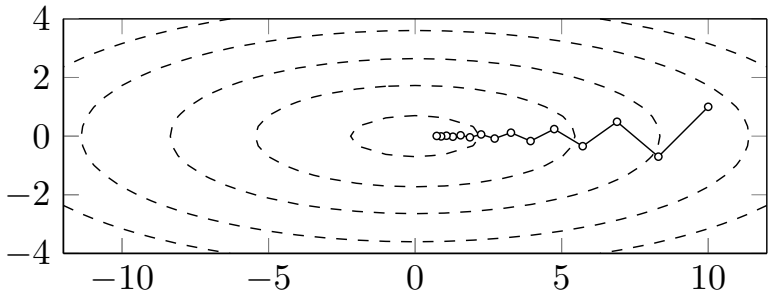
$$x^{k+1} = x^k - \alpha_k \nabla f(x^k)$$

- 另一种理解方式

$$\begin{aligned} x^{k+1} &= \arg\min_x \ f(x^k) + \nabla f(x^k)^\top (x - x^k) + \frac{1}{\alpha_k} \|x - x^k\|_2^2 \\ &= \arg\min_x \|x - (x^k - \alpha_k \nabla f(x^k))\|_2^2 \\ &= x^k - \alpha_k \nabla f(x^k) \end{aligned}$$

二次函数的梯度法

- 设二次函数 $f(x,y)=x^2+10y^2$, 初始点 (x^0,y^0) 取为 $(10,1)$, 取固定步长 $\alpha_k=0.085$, 使用梯度法 $x^{k+1}=x^k-\alpha_k\nabla f(x^k)$ 进行 15 次迭代



二次函数的收敛定理

- **定理 5.2** 考虑正定二次函数

$$f(x) = \frac{1}{2}x^\top Ax - b^\top x$$

其最优值点为 x^*. 若使用梯度法 $x^{k+1} = x^k - \alpha_k \nabla f(x^k)$ 并选取 α_k 为精确线搜索步长, 即

$$\alpha_k = \frac{\|\nabla f(x^k)\|^2}{\nabla f(x^k)^\top A \nabla f(x^k)}$$

则梯度法关于迭代点列 $\{x^k\}$ 是 Q-线性收敛的, 即

$$\|x^{k+1} - x^*\|_A^2 \leq (\frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n})^2 \|x^k - x^*\|_A^2$$

其中 λ_1, λ_n 分别为 A 的最大、最小特征值, $\|x\|_A = \sqrt{x^\top Ax}$ 为由正定矩阵 A 诱导的范数

梯度利普希茨连续

- 可微函数 f，若存在 $L > 0$，对任意的 $x, y \in \mathrm{dom}\ f$ 有

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

则称 f 是梯度利普希茨连续的，相应利普希茨常数为 L

- 设可微函数 $f(x)$ 的定义域 $\mathrm{dom}\ f = \mathbb{R}^n$，且为梯度 L-利普希茨连续的，则函数 $f(x)$ 有二次上界

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|^2, \quad \forall\ x, y \in \mathrm{dom}\ f$$

- 设可微函数 $f(x)$ 的定义域为 $\mathbb{R}^n$ 且存在一个全局极小点 x^*, 若 $f(x)$ 为梯度 L -利普希茨连续的, 则对任意的 x 有

$$\frac{1}{2L}\|\nabla f(x)\|^2 \leq f(x) - f(x^*)$$

梯度法在凸函数上的收敛性

- **定理 5.3** 设$f(x)$ 为凸的梯度 L -利普希茨连续函数，$f^* = f(x^*) = \inf_x f(x)$ 存在且可达，如果步长 α_k 取为常数 α 且满足 $0 < \alpha < \frac{1}{L}$，那么点列 $\{x^k\}$ 的函数值收敛到最优值, 且在函数值的意义下收敛速度为 $\mathcal{O}(\frac{1}{k})$

  证明 因为函数 f 是利普希茨可微函数, 对任意的 x, 根据二次上界引理,

$$f(x - \alpha\nabla f(x)) \leq f(x) - \alpha(1 - \frac{L\alpha}{2})\|\nabla f(x)\|^2$$

  记 $\tilde{x} = x - \alpha\nabla f(x)$ 并限制 $0 < \alpha < \frac{1}{L}$, 有

$$\begin{aligned} f(\tilde{x}) &\leq f(x) - \frac{\alpha}{2}\|\nabla f(x)\|^2 \\ &\leq f^* + \nabla f(x)^\top (x - x^*) - \frac{\alpha}{2}\|\nabla f(x)\|^2 \\ &= f^* + \frac{1}{2\alpha}(\|x - x^*\|^2 - \|\tilde{x} - x^*\|^2) \end{aligned}$$

证明

在上式中取 $x = x^{i-1}, \tilde{x} = x^i$ 并将不等式对 $i = 1, 2, \cdots, k$ 求和得到

$$\begin{aligned}\sum_{i=1}^{k}(f(x^i) - f^*) &\leq \frac{1}{2\alpha}\sum_{i=1}^{k}(\|x^{i-1} - x^*\|^2 - \|x^i - x^*\|^2) \\ &= \frac{1}{2\alpha}(\|x^0 - x^*\|^2 - \|x^k - x^*\|^2) \\ &\leq \frac{1}{2\alpha}\|x^0 - x^*\|^2\end{aligned}$$

由于 $f(x^i)$ 是非增的, 所以

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

凸函数性质

- **引理 5.1** 设函数$f(x)$ 是 $\mathbb{R}^n$ 上的凸可微函数, 则以下结论等价
  - f 的梯度为 L -利普希茨连续的
  - 函数 $g(x) = \frac{L}{2}x^\top x - f(x)$ 是凸函数
  - $\nabla f(x)$ 有**余强制性**, 即对任意的 $x, y \in \mathbb{R}^n$, 有

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2$$

- **定理 5.4** 设$f(x)$ 为 m -强凸的梯度 L -利普希茨连续函数, $f(x^*) = \inf_x f(x)$ 存在且可达. 如果步长 α 满足 $0 < \alpha < \frac{2}{m+L}$, 那么由梯度下降法迭代得到的点列 $\{x^k\}$ 收敛到 x^*, 且为 Q-线性收敛

Barzilar-Borwein 方法

- Barzilar-Borwein (BB) 方法是一种特殊的梯度法, 下降方向仍是点 x^k 处的负梯度方向 $-\nabla f(x^k)$, 但步长 α_k 并不是直接由线搜索算法给出的
- 考虑梯度下降法的格式

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k) \quad \Leftrightarrow \quad x^{k+1} = x^k - D^k \nabla f(x^k), D^k = \alpha_k I$$

- BB 方法选取的 α_k 是如下两个最优问题之一的解

$$\min_{\alpha} \quad \|\alpha y^{k-1} - s^{k-1}\|^2$$
$$\min_{\alpha} \quad \|y^{k-1} - \alpha^{-1} s^{k-1}\|^2$$

其中记号 $s^{k-1} = x^k - x^{k-1}$ 以及 $y^{k-1} = \nabla f(x^k) - \nabla f(x^{k-1})$

Barzilar-Borwein 方法

- 容易验证问题的解分别为

$$\alpha_{\mathrm{BB1}}^k = \frac{(s^{k-1})^\top y^{k-1}}{(y^{k-1})^\top y^{k-1}} \quad \text{和} \quad \alpha_{\mathrm{BB2}}^k = \frac{(s^{k-1})^\top s^{k-1}}{(s^{k-1})^\top y^{k-1}},$$

- 得到 BB 方法的两种迭代格式

$$x^{k+1} = x^k - \alpha_{\mathrm{BB1}}^k \nabla f(x^k) \quad \text{和} \quad x^{k+1} = x^k - \alpha_{\mathrm{BB2}}^k \nabla f(x^k)$$

- 仅需函数相邻两步的梯度信息和迭代点信息, 不需要任何线搜索算法
- BB 方法计算出的步长可能过大或过小, 将步长做上界和下界的截断, 即

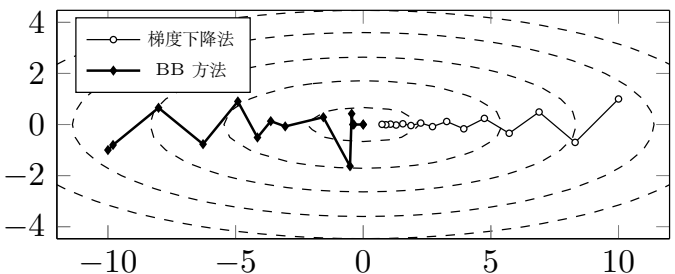
$$\alpha_m \leq \alpha_k \leq \alpha_M$$

非单调线搜索的 BB 方法

1 给定x^0，选取初值 $\alpha > 0$，整数 $M \geq 0$，$c_1, \beta, \varepsilon \in (0,1)$，$k = 0$
2 **while** $\|\nabla f(x^k)\| > \varepsilon$ **do**
3 **while** $f(x^k - \alpha \nabla f(x^k)) \geq \max_{0 \leq j \leq \min(k,M)} f(x^{k-j}) - c_1 \alpha \|\nabla f(x^k)\|^2$ **do**
4 令 $\alpha \leftarrow \beta \alpha$
5 **end while**
6 令 $x^{k+1} = x^k - \alpha \nabla f(x^k)$
7 根据 BB 步长公式之一计算 α，并做截断使得 $\alpha \in [\alpha_m, \alpha_M]$
8 $k \leftarrow k + 1$
9 **end while**

二次函数的 BB 方法

- 设二次函数 $f(x, y) = x^2 + 10y^2$, 初始点为 $(-10, -1)$
- BB 方法的收敛速度较快, 但非单调
- 对于正定二次函数, BB 方法有 R-线性收敛速度



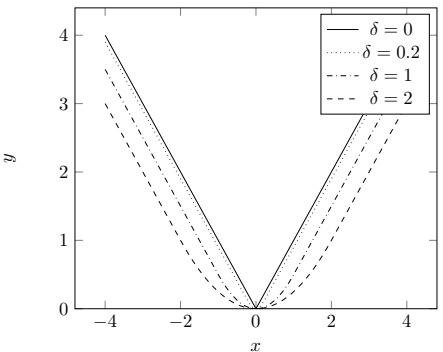
LASSO 问题求解

- 考虑

$$\min \quad f(x) = \frac{1}{2}\|Ax - b\|^2 + \mu\|x\|_1$$

- 由于 $\|x\|_1$ 不光滑, 考虑 Huber 光滑函数

$$l_\delta(x) = \begin{cases} \frac{1}{2\delta}x^2, & |x| < \delta \\ |x| - \frac{\delta}{2}, & \text{其他} \end{cases}$$



LASSO 问题求解

- 光滑化 LASSO 问题为

$$\min \quad f_\delta(x) = \frac{1}{2}\|Ax-b\|^2 + \mu L_\delta(x), \quad \text{其中} \quad L_\delta(x) = \sum_{i=1}^{n} l_\delta(x_i),$$

- $f_\delta(x)$ 的梯度为

$$\nabla f_\delta(x) = A^\top(Ax-b) + \mu \nabla L_\delta(x),$$

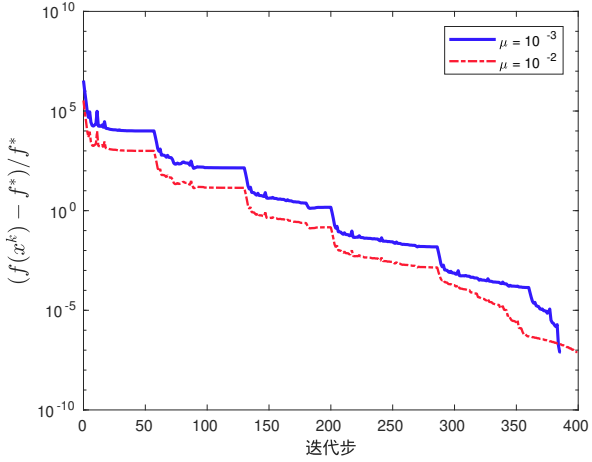
其中

$$(\nabla L_\delta(x))_i = \begin{cases} \operatorname{sign}(x_i), & |x_i| > \delta \\ \frac{x_i}{\delta}, & |x_i| \le \delta \end{cases}$$

- $f_\delta(x)$ 的梯度是利普希茨连续的, 且相应常数为 $L = \|A^\top A\|_2 + \frac{\mu}{\delta}$

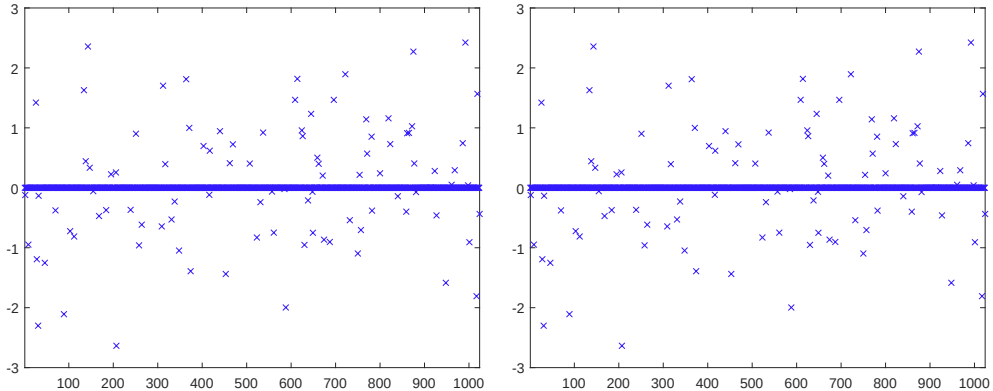
LASSO 问题求解

- 光滑化 LASSO 问题求解迭代过程



LASSO 问题求解

- 精确解 (左) v.s. 梯度法解 (右)



目录

回顾：梯度下降算法

- 设$f(x)$ 是可微凸函数且 $\mathrm{dom}\ f = \mathbb{R}^n$，考虑如下问题

$$\min_x \ f(x)$$

- 梯度下降法, 选择初始点 $x^0 \in \mathbb{R}^n$，然后重复

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k), \quad k = 0, 1, 2, \cdots$$

- 若 $\nabla f(x)$ 利普西茨连续，则梯度下降法的收敛速度是 $\mathcal{O}(\frac{1}{k})$

如果 $f(x)$不可微呢？

次梯度算法结构

- 回顾一阶充要条件

$$x^* \text{是一个全局极小点} \quad \Leftrightarrow \quad 0 \in \partial f(x^*)$$

- 类似梯度法构造如下次梯度算法的迭代格式

$$x^{k+1} = x^k - \alpha_k g^k, \quad g^k \in \partial f(x^k)$$

  - 固定步长 $\alpha_k = \alpha$
  - 固定 $\|x^{k+1} - x^k\|$，即 $\alpha_k \|g^k\|$ 为常数
  - 消失步长 $\alpha_k arrow 0$ 且 $\sum_{k=0}^{\infty} \alpha_k = +\infty$
  - 选取 α_k 使其满足某种线搜索准则

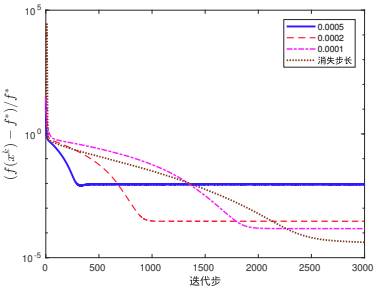
例: LASSO 问题求解

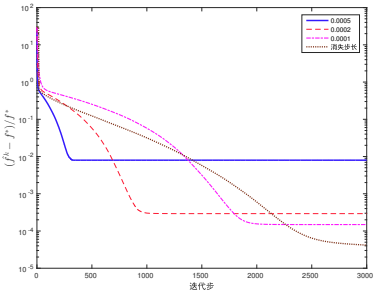
- 考虑 LASSO 问题

$$\min \quad f(x) = \frac{1}{2}\|Ax - b\|^2 + \mu\|x\|_1$$

- 次梯度算法

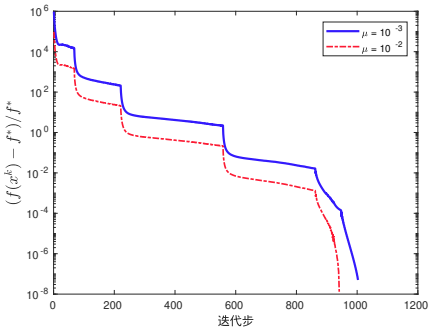
$$x^{k+1} = x^k - \alpha_k(A^\top(Ax^k - b) + \mu(x^k))$$





例: LASSO 问题求解

- 若$\mu_t > \mu$，则取固定步长 $\frac{1}{\lambda_{\max}(A^\top A)}$
- 若 $\mu_t = \mu$，则取步长 $\frac{1}{\lambda_{\max}(A^\top A)\cdot(\max\{k,100\}-99)}$



Q&A

Thank you!

感谢您的聆听和反馈