梯度下降算法

宋晓良

大连理工大学数学科学学院

教材《最优化：建模、算法与理论》配套电子教案

http://bicmr.pku.edu.cn/~wenzw/optbook.html

提纲

引言:无约束优化算法

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

- 线搜索
  1. 先确定搜索方向:梯度类算法、次梯度算法、牛顿算法、拟牛顿算法等.
  2. 按准则进行近似搜索.
- 信赖域
  1. 主要针对$f(x)$ 二阶可微的情形, 在一个给定的区域内使用二阶模型近似原问题.
  2. 不断直接求解该二阶模型从而找到最优值点.

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:
  1. 首先确定下一步该向哪一方向行走.
  2. 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:

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

  我们称d^k 为迭代点x^k 处的**搜索方向**, α_k 为相应的**步长**.这里要求d^k 是一个**下降方向**, 即$(d^k)^{\mathrm{T}}\nabla f(x^k) < 0$.
- 线搜索类算法的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

- 选取d^k 的方法千差万别, 但选取α_k 的方法却非常相似.
- 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. 这要求:
  1. α_k 应该使得f 充分下降
  2. 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

$$\alpha_k = \operatorname*{argmin}_{\alpha>0} \phi(\alpha),$$

即α_k 为最佳步长. 这种线搜索算法被称为**精确线搜索算法**

- 选取α_k 通常需要很大计算量, 在实际应用中较少使用

α_k 的选取:非精确线搜索算法

- 不要求α_k 是$\phi(\alpha)$ 的最小值点, 仅要求$\phi(\alpha_k)$ 满足某些不等式. 这种线搜索方法被称为**非精确线搜索算法**.
- 选取α_k 需要满足的要求被称为**线搜索准则**
- 线搜索准则的合适与否直接决定了算法的收敛性, 若选取不合适的线搜索准则将会导致算法无法收敛
- 例如, 若只要求选取的步长满足迭代点处函数值单调下降, 则函数值$f(x^k)$ 的下降量可能不够充分, 导致算法无法收敛到极小值点

例子:不合适的线搜索准则导致无法收敛

考虑一维无约束优化问题

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

迭代初始点$x^0 = 1$. 由于问题是一维的, 下降方向只有$\{-1, +1\}$ 两种. 我们选取$d^k = -\mathrm{sign}(x^k)$, 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

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

显然, 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

Armijo准则

定义 (Armijo 准则)

设d^k 是点x^k 处的下降方向，若

$$f(x^k + \alpha d^k) \leq f(x^k) + c_1 \alpha \nabla f(x^k)^{\mathrm{T}} d^k,$$

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

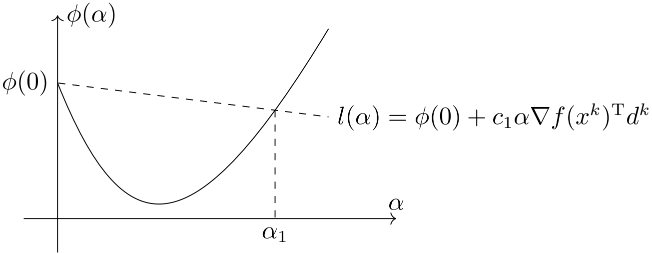


Figure: Armijo 准则

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

$$l(\alpha) = \phi(0) + c_1 \alpha \nabla f(x^k)^{\mathrm{T}} d^k$$

  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准则非常容易得到满足
- Armijo 准则需要配合其他准则以保证迭代的收敛性, 因为$\alpha = 0$ 显然满足Armijo准则, 此时迭代序列中的点固定不变

- 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

$$j_0 = \min\{j = 0, 1, \cdots \mid f(x^k + \gamma^j \hat{\alpha} d^k) \le f(x^k) + c_1 \gamma^j \hat{\alpha} \nabla f(x^k)^{\mathrm{T}} d^k\},$$

参数$\gamma \in (0, 1)$ 为一个给定的实数

Algorithm 1 线搜索回退法

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** $f(x^k + \alpha d^k) > f(x^k) + c\alpha \nabla f(x^k)^{\mathrm{T}} d^k$ **do**
3: 　令$\alpha \leftarrow \gamma\alpha$.
4: **end while**
5: 输出$\alpha_k = \alpha$.

回退法:以Armijo准则为例

- 该算法被称为回退法是因为α 的试验值是由大至小的, 它可以确保输出的α_k 能尽量地大
- 算法1不会无限进行下去, 因为d^k 是一个下降方向, 当α 充分小时, Armijo 准则总是成立的
- 实际应用中我们通常也会给α 设置一个下界, 防止步长过小

- 为了克服Armijo 准则的缺陷，我们需要引入其他准则来保证每一步的α^k 不会太小
- Armijo准则只要求点$(\alpha, \phi(\alpha))$ 必须处在某直线下方，我们也可使用相同的形式使得该点必须处在另一条直线的上方．这就是Armijo-Goldstein 准则，简称Goldstein 准则

定义 (Goldstein 准则)

设d^k 是点x^k 处的下降方向，若

$$f(x^k + \alpha d^k) \le f(x^k) + c\alpha \nabla f(x^k)^{\mathrm{T}} d^k, \tag{1a}$$

$$f(x^k + \alpha d^k) \ge f(x^k) + (1-c)\alpha \nabla f(x^k)^{\mathrm{T}} d^k, \tag{1b}$$

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

Goldstein 准则有直观的几何含义，它指的是点$(\alpha, \phi(\alpha))$ 必须在两条直线

$$l_1(\alpha) = \phi(0) + c\alpha\nabla f(x^k)^{\mathrm{T}} d^k,$$

$$l_2(\alpha) = \phi(0) + (1-c)\alpha\nabla f(x^k)^{\mathrm{T}} d^k$$

之间. 区间$[\alpha_1, \alpha_2]$ 中的点均满足Goldstein 准则. 同时我们也注意到Goldstein 准则确实去掉了过小的α.

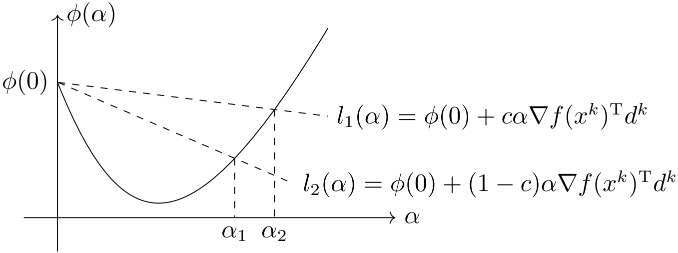


Figure: Goldstein 准则

Wolfe准则

- Goldstein 准则能够使得函数值充分下降, 但是它可能避开了最优的函数值. 上页图中的一维函数$\phi(\alpha)$ 的最小值点并不在满足Goldstein 准则的区间$[\alpha_1, \alpha_2]$ 中
- 在Wolfe准则中, 第一个不等式即是Armijo 准则, 而第二个不等式则是Wolfe 准则的本质要求

定义 (Wolfe 准则)

设d^k 是点x^k 处的下降方向, 若

$$f(x^k + \alpha d^k) \le f(x^k) + c_1 \alpha \nabla f(x^k)^{\mathrm{T}} d^k, \tag{2a}$$

$$\nabla f(x^k + \alpha d^k)^{\mathrm{T}} d^k \ge c_2 \nabla f(x^k)^{\mathrm{T}} d^k, \tag{2b}$$

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$\alpha^*$ 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

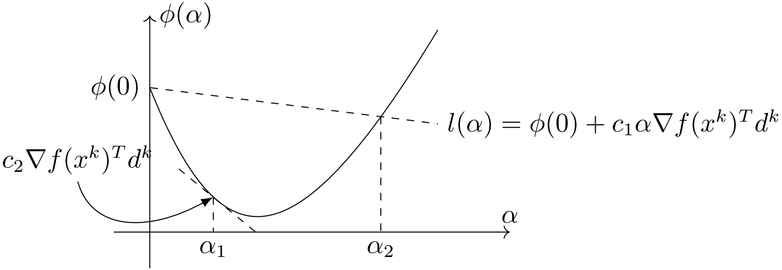


Figure: Wolfe 准则

非单调线搜索准则

定义 (Grippo)

设d^k 是点x^k 处的下降方向, $M > 0$ 为给定的正整数. 以下不等式可作为一种线搜索准则：

$$f(x^k + \alpha d^k) \leq \max_{0 \leq j \leq \min\{k,M\}} f(x^{k-j}) + c_1 \alpha \nabla f(x^k)^{\mathrm{T}} d^k,$$

其中$c_1 \in (0,1)$ 为给定的常数.

- 该准则和Armijo 准则非常相似, 区别在于Armijo 准则要求下一次迭代的函数值$f(x^{k+1})$ 相对于本次迭代的函数值$f(x^k)$ 有充分下降, 而该准则只需要下一步函数值相比前面至多M 步以内迭代的函数值有下降就可以了
- 这一准则的要求比Armijo 准则更宽, 它也不要求$f(x^k)$ 的单调性

提纲

回退法的优缺点

- 只要修改一下算法的终止条件，回退法就可以被用在其他线搜索准则之上，它是最常用的线搜索算法之一.
- 然而，回退法的缺点也很明显：
  1. 它无法保证找到满足Wolfe 准则的步长，即条件二不一定成立，但对一些优化算法而言，找到满足Wolfe 准则的步长是十分必要的
  2. 回退法以指数的方式缩小步长，因此对初值$\hat{\alpha}$ 和参数γ 的选取比较敏感，当γ 过大时每一步试探步长改变量很小，此时回退法效率比较低，当γ 过小时回退法过于激进，导致最终找到的步长太小，错过了选取大步长的机会

基于多项式插值的线搜索算法

- 设初始步长$\hat{\alpha}_0$ 已给定，如果经过验证，$\hat{\alpha}_0$ 不满足Armijo 准则，下一步就需要减小试探步长
- 基于$\phi(0), \phi'(0), \phi(\hat{\alpha}_0)$ 这三个信息构造一个二次插值函数$p_2(\alpha)$
- 寻找二次函数$p_2(\alpha)$ 满足

$$p_2(0) = \phi(0), \quad p_2'(0) = \phi'(0), \quad p_2(\hat{\alpha}_0) = \phi(\hat{\alpha}_0).$$

  由于二次函数只有三个参数，以上三个条件可以唯一决定$p_2(\alpha)$
- $p_2(\alpha)$ 的最小值点恰好位于$(0, \hat{\alpha}_0)$ 内
- 取$p_2(\alpha)$ 的最小值点$\hat{\alpha}_1$ 作为下一个试探点，利用同样的方式不断递归下去直至找到满足Armijo 准则的点
- 基于插值的线搜索算法可以有效减少试探次数，但仍然不能保证找到的步长满足Wolfe 准则

提纲

Zoutendijk定理

定理 (Zoutendijk定理)

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

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

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

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

这个不等式也被称为***Zoutendijk*** 条件.

Zoutendijk定理的证明

- 由Wolfe准则知$\nabla f(x^{k+1})^\mathrm{T} d^k \geq c_2 f(x^k)^\mathrm{T} d^k$, 故

$$\left(\nabla f(x^{k+1}) - \nabla f(x^k)\right)^\mathrm{T} d^k \geq (c_2 - 1)\nabla f(x^k)^\mathrm{T} d^k.$$

- 由柯西不等式和梯度L-利普希茨连续性质,

$$\left(\nabla f(x^{k+1}) - \nabla f(x^k)\right)^\mathrm{T} d^k \leq \|\nabla f(x^{k+1}) - \nabla f(x^k)\|\|d^k\| \leq \alpha_k L \|d^k\|^2.$$

- 结合上述两式可得

$$\alpha_k \geq \frac{c_2 - 1}{L} \frac{\nabla f(x^k)^\mathrm{T} d^k}{\|d^k\|^2}.$$

- 由Wolfe准则的条件一知$f(x^{k+1}) \leq f(x^k) + c_1 \alpha_k \nabla f(x^k)^\mathrm{T} d^k$, 注意到$\nabla f(x^k)^\mathrm{T} d^k < 0$, 将上式代入得

$$f(x^{k+1}) \leq f(x^k) + c_1 \frac{c_2 - 1}{L} \frac{\left(\nabla f(x^k)^\mathrm{T} d^k\right)^2}{\|d^k\|^2}.$$

Zoutendijk定理的证明

- 根据θ_k 的定义，此不等式可等价表述为

$$f(x^{k+1}) \le f(x^k) + c_1 \frac{c_2 - 1}{L} \cos^2 \theta_k \|\nabla f(x^k)\|^2.$$

- 再关于k 求和，我们有

$$f(x^{k+1}) \le f(x^0) - c_1 \frac{1 - c_2}{L} \sum_{j=0}^{k} \cos^2 \theta_j \|\nabla f(x^j)\|^2.$$

- 又因为函数f 是下有界的，且由$0 < c_1 < c_2 < 1$ 可知$c_1(1 - c_2) > 0$，因此当$k \to \infty$ 时，

$$\sum_{j=0}^{\infty} \cos^2 \theta_j \|\nabla f(x^j)\|^2 < +\infty.$$

Zoutendijk 定理刻画了线搜索准则的性质, 配合下降方向d^k 的选取方式我们可以得到最基本的收敛性.

推论 (线搜索算法的收敛性)

对于迭代法$x^{k+1} = x^k + \alpha_k d^k$, 设$\theta_k$ 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

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

线搜索算法收敛性的证明

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

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

- 根据θ_k的假设, 对任意的k,

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

- 我们仅考虑Zoutendijk条件中第k_l项的和, 有

$$\begin{aligned}\sum_{k=0}^{\infty}\cos^2\theta_k\|\nabla f(x^k)\|^2 &\geq \sum_{l=1}^{\infty}\cos^2\theta_{k_l}\|\nabla f(x^{k_l})\|^2 \\ &\geq \sum_{l=1}^{\infty}(\sin^2\gamma)\cdot\delta^2 \to +\infty,\end{aligned}$$

- 这显然和Zoutendijk定理矛盾. 因此必有

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

- 线搜索算法收敛性建立在Zoutendijk 条件之上, 它的本质要求是$\theta_k < \frac{\pi}{2} - \gamma$, 即每一步的下降方向$d^k$ 和负梯度方向不能趋于正交.
- 几何直观：当下降方向d^k 和梯度正交时, 根据泰勒展开的一阶近似, 目标函数值$f(x^k)$ 几乎不发生改变. 因此我们要求d^k 与梯度正交方向夹角有一致的下界.
- 不涉及算法收敛速度的分析, 因为算法收敛速度极大地取决于d^k 的选取.

提纲

梯度下降法

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

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

- 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.
- 因此梯度法就是选取$d^k=-\nabla f(x^k)$ 的算法，它的迭代格式为

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

步长α_k 的选取可依赖于线搜索算法，也可直接选取固定的α_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$\alpha_k=0.085$. 我们使用梯度法$x^{k+1}=x^k-\alpha_k\nabla f(x^k)$ 进行15 次迭代.

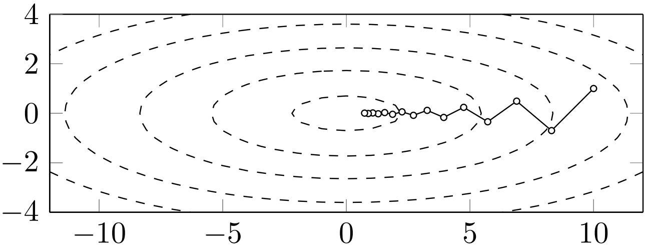


Figure: 梯度法的前15 次迭代

二次函数的收敛定理

定理 (二次函数的收敛定理)

考虑正定二次函数

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

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

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

则梯度法关于迭代点列$\{x^k\}$ 是Q-线性收敛的, 即

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

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

- 定理中线性收敛速度的常数和矩阵A 最大特征值与最小特征值之比有关.
- 从等高线角度来看，这个比例越大则$f(x)$ 的等高线越扁平，迭代路径折返频率会随之变高，梯度法收敛也就越慢.
- 这个结果其实说明了梯度法的一个很重大的缺陷：当目标函数的海瑟矩阵条件数较大时，它的收敛速度会非常缓慢.

梯度利普希茨连续

定义 (梯度利普希茨连续)

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

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|, \tag{3}$$

则称f是**梯度利普希茨连续的**，相应利普希茨常数为L. 有时也简记为**梯度L-利普希茨连续**或**L-光滑**.

引理 (二次上界)

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：

$$f(y) \le f(x) + \nabla f(x)^{\mathrm{T}}(y - x) + \frac{L}{2}\|y - x\|^2, \quad \forall\, x, y \in \mathrm{dom} f. \tag{4}$$

证明

可以证明:

$$
\begin{aligned}
& f(y)-f(x)-\nabla f(x)^{\mathrm{T}}(y-x) \\
= & \int_0^1 (\nabla f(x+t(y-x))-\nabla f(x))^{\mathrm{T}}(y-x)\mathrm{d}t \\
\leq & \int_0^1 \|\nabla f(x+t(y-x))-\nabla f(x)\|\|y-x\|\mathrm{d}t \\
\leq & \int_0^1 L\|y-x\|^2 t\mathrm{d}t=\frac{L}{2}\|y-x\|^2,
\end{aligned}
$$

其中最后一行的不等式利用了梯度利普希茨连续的条件(3). 整理可得(4) 式成立.

梯度法在凸函数上的收敛性

考虑梯度法

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

假设：

- 设函数$f(x)$ 为凸的梯度L -利普希茨连续函数
- 极小值$f^* = f(x^*) = \inf_x f(x)$ 存在且可达.
- 如果步长α_k 取为常数α 且满足$0 < \alpha < \frac{1}{L}$

结论：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$\mathcal{O}\left(\frac{1}{k}\right)$.

如果函数f 还是m -强凸函数，则梯度法的收敛速度会进一步提升为Q-线性收敛.

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,

$$f(x-\alpha\nabla f(x)) \le f(x) - \alpha\left(1-\frac{L\alpha}{2}\right)\|\nabla f(x)\|^2.$$

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

$$\begin{aligned}
f(\tilde{x}) &\le f(x) - \frac{\alpha}{2}\|\nabla f(x)\|^2 \\
&\le f^* + \nabla f(x)^{\mathrm{T}}(x - x^*) - \frac{\alpha}{2}\|\nabla f(x)\|^2 \\
&= f^* + \frac{1}{2\alpha}\left(\|x - x^*\|^2 - \|x - x^* - \alpha\nabla f(x)\|^2\right) \\
&= f^* + \frac{1}{2\alpha}(\|x - x^*\|^2 - \|\tilde{x} - x^*\|^2),
\end{aligned}$$

其中第一个不等式是因为$0 < \alpha < \frac{1}{L}$, 第二个不等式为f 的凸性.

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

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

- 由于$f(x^i)$ 是非增的，所以

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

凸函数性质

引理

设函数$f(x)$ 是$\mathbb{R}^n$ 上的凸可微函数，则以下结论等价：

1. f 的梯度为L -利普希茨连续的；
2. 函数$g(x) \xlongequal{\text{def}} \frac{L}{2}x^{\mathrm{T}}x - f(x)$ 是凸函数；
3. $\nabla f(x)$ 有**余强制性**，即对任意的$x, y \in \mathbb{R}^n$，有
$$(\nabla f(x) - \nabla f(y))^{\mathrm{T}}(x-y) \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2.$$

$(1) \implies (2)$ 即证$g(x)$ 的单调性. 对任意$x, y \in \mathbb{R}^n$,

$$\begin{aligned}(\nabla g(x) - \nabla g(y))^{\mathrm{T}}(x-y) &= L\|x-y\|^2 - (\nabla f(x) - \nabla f(y))^{\mathrm{T}}(x-y) \\ &\geq L\|x-y\|^2 - \|x-y\|\|\nabla f(x) - \nabla f(y)\| \geq 0.\end{aligned}$$

因此$g(x)$ 为凸函数.

凸函数性质

引理 (梯度L-利普希茨函数的性质)

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

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

(2) $\Longrightarrow$ (3)

- 构造辅助函数

$$f_x(z) = f(z) - \nabla f(x)^{\mathrm{T}} z,$$
$$f_y(z) = f(z) - \nabla f(y)^{\mathrm{T}} z,$$

容易验证f_x 和f_y 均为凸函数.

- $g_x(z) = \frac{L}{2} z^{\mathrm{T}} z - f_x(z)$ 关于z 是凸函数. 根据凸函数的性质, 我们有

$$g_x(z_2) \ge g_x(z_1) + \nabla g_x(z_1)^{\mathrm{T}}(z_2 - z_1), \quad \forall\, z_1, z_2 \in \mathbb{R}^n.$$

整理可推出$f_x(z)$ 有二次上界, 且对应的系数也为L.

凸函数性质

- 注意到$\nabla f_x(x)=0$, 这说明x 是$f_x(z)$ 的最小值点. 由上页引理,

$$
\begin{aligned}
f_x(y)-f_x(x) &= f(y)-f(x)-\nabla f(x)^{\mathrm{T}}(y-x) \\
&\geq \frac{1}{2L}\|\nabla f_x(y)\|^2 = \frac{1}{2L}\|\nabla f(y)-\nabla f(x)\|^2.
\end{aligned}
$$

- 同理, 对$f_y(z)$ 进行类似的分析可得

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

将以上两式不等号左右分别相加, 可得余强制性.

$(3) \implies (1)$ 由余强制性和柯西不等式,

$$
\begin{aligned}
\frac{1}{L}\|\nabla f(x)-\nabla f(y)\|^2 &\leq (\nabla f(x)-\nabla f(y))^{\mathrm{T}}(x-y) \\
&\leq \|\nabla f(x)-\nabla f(y)\|\|x-y\|,
\end{aligned}
$$

整理后即可得到$f(x)$ 是梯度L-利普希茨连续的.

梯度法在强凸函数上的收敛性

定理 (梯度法在强凸函数上的收敛性)

设函数$f(x)$ 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.

- 首先根据f 强凸且∇f 利普希茨连续, 可得

$$g(x) = f(x) - \frac{m}{2}x^{\mathrm{T}}x$$

为凸函数且$\frac{L-m}{2}x^{\mathrm{T}}x - g(x)$ 为凸函数.

- 由引理知函数$g(x)$ 是梯度$(L-m)$ -利普希茨连续的. 再次利用引理可得关于$g(x)$ 的余强制性

$$(\nabla g(x) - \nabla g(y))^{\mathrm{T}}(x-y) \geq \frac{1}{L-m}\|\nabla g(x) - \nabla g(y)\|^2.$$

梯度法在强凸函数上的收敛性

- 代入$g(x)$ 的表达式，可得

$$(\nabla f(x) - \nabla f(y))^{\mathrm{T}}(x-y) \geq \frac{mL}{m+L}\|x-y\|^2 + \frac{1}{m+L}\|\nabla f(x) - \nabla f(y)\|^2.$$

- 再估计固定步长下梯度法的收敛速度. 设步长$\alpha \in \left(0, \frac{2}{m+L}\right)$, 对$x^k, x^*$ 应用上式并注意到$\nabla f(x^*) = 0$ 得

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

$$\Rightarrow \|x^k - x^*\|^2 \leq c^k \|x^0 - x^*\|^2, \quad c = 1 - \alpha \frac{2mL}{m+L} < 1.$$

函数值收敛

强凸函数假设下

- 迭代点列$\{x^k\}$Q-线性收敛
- 如果取$t = \frac{2}{m+L}$，则有$c = \left(\frac{\gamma-1}{\gamma+1}\right)^2$且$\gamma = L/m$

如果$\mathbf{dom}\, f = \mathbf{R}^n$且$f$ 有极小点x^*, 则

$$\frac{1}{2L}\|\nabla f(x)\|_2^2 \le f(x) - f(x^*) \le \frac{L}{2}\|x - x^*\|_2^2 \quad \forall x$$

因此：

$$f(x^k) - f^* \le \frac{L}{2}\|x^k - x^*\|_2^2 \le \frac{c^k L}{2}\|x^0 - x^*\|_2^2$$

函数值的估计： 达到$f(x^k) - f^* \le \epsilon$ 的迭代步数是$O(\log(1/\epsilon))$

提纲

Barzilar-Borwein 方法

- Barzilar-Borwein (BB) 方法是一种特殊的梯度法, 经常比一般的梯度法有着更好的效果.
- BB 方法的下降方向仍是点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:

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

其中$D^k = \alpha_k I$.

- BB 方法选取的α_k 是如下两个最优问题之一的解:

$$\begin{aligned} &\min_{\alpha} \quad \|\alpha y^{k-1} - s^{k-1}\|^2, \\ &\min_{\alpha} \quad \|y^{k-1} - \alpha^{-1} s^{k-1}\|^2, \end{aligned}$$

其中引入记号$s^{k-1} \stackrel{\text{def}}{=\!=} x^k - x^{k-1}$ 以及$y^{k-1} \stackrel{\text{def}}{=\!=} \nabla f(x^k) - \nabla f(x^{k-1})$.

Barzilar-Borwein 方法

- 容易验证问题的解分别为

$$\alpha_{\mathrm{BB1}}^k \xlongequal{\mathrm{def}} \frac{(s^{k-1})^{\mathrm{T}} y^{k-1}}{(y^{k-1})^{\mathrm{T}} y^{k-1}} \ \text{和}\ \alpha_{\mathrm{BB2}}^k \xlongequal{\mathrm{def}} \frac{(s^{k-1})^{\mathrm{T}} s^{k-1}}{(s^{k-1})^{\mathrm{T}} y^{k-1}},$$

- 因此可以得到BB 方法的两种迭代格式：

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

- 计算两种BB 步长的任何一种仅仅需要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即可选取算法步长.
- BB方法计算出的步长可能过大或过小，因此我们还需要将步长做上界和下界的截断，即选取$0 < \alpha_m < \alpha_M$ 使得

$$\alpha_m \le \alpha_k \le \alpha_M.$$

- BB 方法本身是非单调方法，有时也配合非单调收敛准则使用以获得更好的实际效果.

Algorithm 2 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$f(x,y)=x^2+10y^2$，并使用*BB* 方法进行迭代，初始点为$(-10,-1)$.
- *BB* 方法的收敛速度较快，在经历*15* 次迭代后已经接近最优值点. 从等高线也可观察到*BB* 方法是非单调方法.
- 实际上，对于正定二次函数，BB方法有R-线性收敛速度.

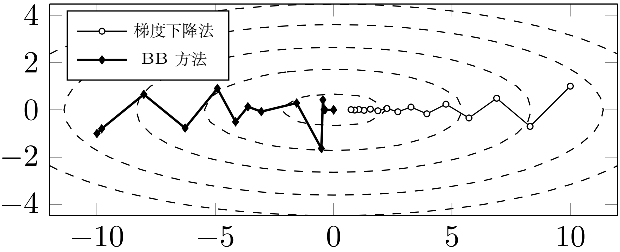


Figure: 梯度法与BB 方法的前15 次迭代

提纲

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

- LASSO 问题的目标函数$f(x)$ 不光滑，在某些点处无法求出梯度，因此不能直接对原始问题使用梯度法求解
- 不光滑项为$\|x\|_1$，它实际上是x 各个分量绝对值的和，考虑如下一维光滑函数：

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

- 上述定义实际上是Huber 损失函数的一种变形，当$\delta \to 0$ 时，光滑函数$l_\delta(x)$ 和绝对值函数$|x|$ 会越来越接近.

LASSO问题求解

光滑化LASSO 问题为

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

δ 为给定的光滑化参数.

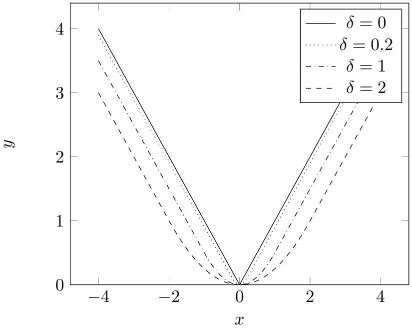


Figure: 当 δ 取不同值时 $l_\delta(x)$ 的图形

- $f_\delta(x)$ 的梯度为

$$\nabla f_\delta(x) = A^{\mathrm{T}}(Ax-b) + \mu \nabla L_\delta(x),$$

其中$\nabla L_\delta(x)$ 是逐个分量定义的：

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

- $f_\delta(x)$ 的梯度是利普希茨连续的，且相应常数为$L = \|A^{\mathrm{T}}A\|_2 + \frac{\mu}{\delta}$.
- 根据梯度法在凸函数上的收敛性定理，固定步长需不超过$\frac{1}{L}$ 才能保证算法收敛，如果δ 过小，那么我们需要选取充分小的步长α_k 使得梯度法收敛.

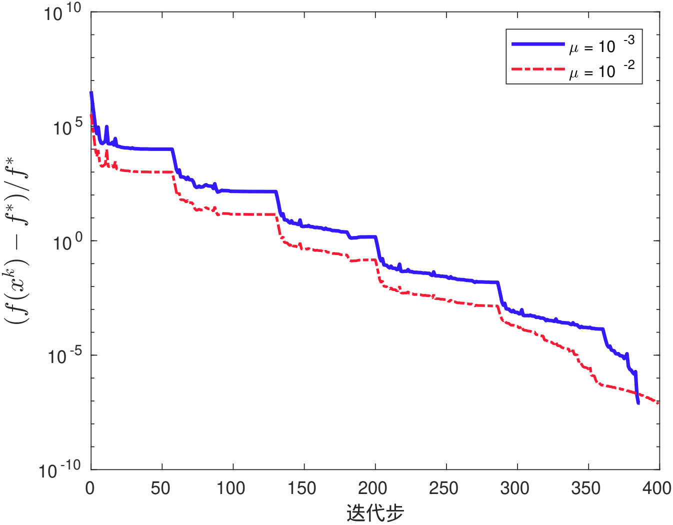
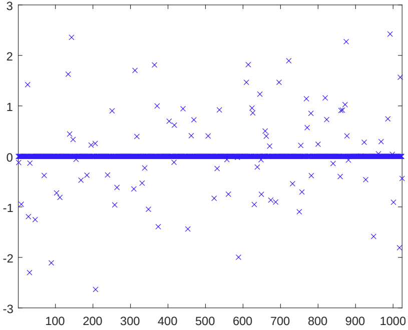
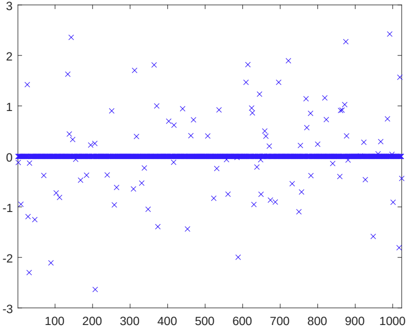


Figure: 光滑化LASSO 问题求解迭代过程

LASSO问题求解



(a) 精确解



(b) 梯度法解

Figure: 光滑化LASSO 问题求解结果

Tikhonov 正则化模型求解

- x 表示真实图像, 它是一个$m \times n$ 点阵, 用y 来表示带噪声图像

$$y = x + e,$$

其中e 是高斯白噪声.

- 为了从有噪声的图像y 中还原出原始图像x, 利用Tikhonov 正则化的思想可以建立如下模型：

$$\min_{x} \quad f(x) = \frac{1}{2}\|x - y\|_F^2 + \lambda(\|D_1 x\|_F^2 + \|D_2 x\|_F^2),$$

其中$D_1 x, D_2 x$ 分别表示对x 在水平方向和竖直方向上做向前差分,

$$(D_1 x)_{ij} = \frac{1}{h}(x_{i+1,j} - x_{ij}), \quad (D_2 x)_{ij} = \frac{1}{h}(x_{i,j+1} - x_{ij}),$$

其中h 为给定的离散间隔.