信赖域算法

文再文

北京大学北京国际数学研究中心

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

致谢：本教案由邓展望协助准备

提纲

无约束优化问题

- 本节考虑如下无约束优化问题：

$$\min_{x\in\mathbb{R}^n} \quad f(x), \tag{1}$$

其中$f(x)$是$\mathbb{R}^n \to \mathbb{R}$的二次可微函数.

- 无约束优化问题是众多优化问题中最基本的问题，它对自变量x的取值范围不加限制，所以无需考虑x的可行性.

- 对于光滑函数，我们可以较容易地利用梯度和海瑟矩阵的信息来设计算法.无约束优化问题的优化算法主要分为两大类：

线搜索算法和**信赖域类算法**

信赖域算法简介

- 信赖域算法与线搜索算法是一般无约束光滑优化问题中的两种框架,二者类似的地方在于都是利用泰勒展开来对目标函数进行局部近似,但处理近似问题的方式不同.
- 线搜索算法是先求出下降方向,然后给定步长$x^{k+1} = x^k + \alpha d^k$.优点是方法可以很简单，比如最速下降法只需求一阶梯度即可,但它的缺点是对方向步长要求较高,比如最速下降法不适宜用精确步长而非线性共轭梯度法却需要精确步长.
- 在信赖域类算法中,我们直接在一个有界区域内求解这个近似模型,而后迭代到下一个点. 其优点是可简化问题，缺点是每次迭代近似模型效果不一定好.

信赖域算法框架

1. 在当前迭代点x^k建立局部模型
2. 求出局部模型的最优解
3. 更新模型信赖域的半径:
   - 模型足够好$\Rightarrow$增大半径
   - 模型比较差$\Rightarrow$缩小半径
   - 否则半径不变
4. 对模型进行评价:
   - 好$\Rightarrow$ 子问题的解即下一个迭代点
   - 差$\Rightarrow$迭代点不改变

信赖域算法的数学表达

- 根据带拉格朗日余项的泰勒展开

$$f(x^k+d)=f(x^k)+\nabla f(x^k)^T d+\frac{1}{2}d^T\nabla^2 f(x^k+td)d$$

其中$t\in(0,1)$为和d有关的正数.

- 和牛顿法相同，可利用$f(x)$的一个二阶近似来刻画$f(x)$在点x^k处的性质：

$$m_k(d)=f(x^k)+\nabla f(x^k)^T d+\frac{1}{2}d^T B^k d$$

其中B^k是对称矩阵，并且是海瑟矩阵的近似矩阵.

- 由于泰勒展开的局部性,需对上述模型添加约束：

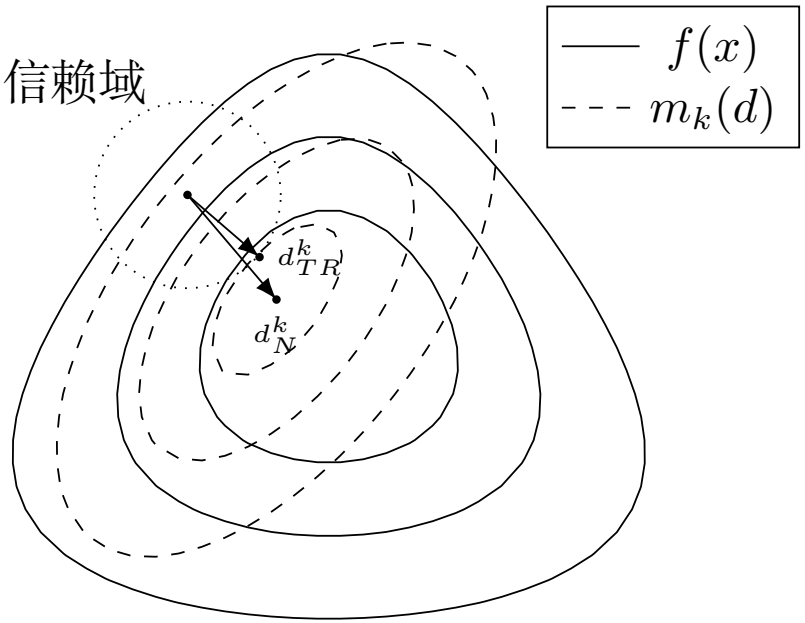
$$\Omega_k=\{x^k+d|\ ||d||\leqslant\Delta_k\},$$

其中$\Delta_k>0$是一个和迭代相关的参数。称Ω_k为信赖域，Δ_k为信赖域半径.

信赖域算法的数学表达

- 因此信赖域算法每一步都需要求解如下子问题:

$$\min_{d\in R^n} \quad m_k(d), \quad \text{s.t.} \quad \|d\| \le \Delta_k. \tag{2}$$



- 图中实线表示$f(x)$的等高线，虚线表示$m_k(d)$的等高线，d_N^k表示解无约束问题得到的下降方向，d_{TR}^k表示解信赖域子问题得到的下降方向.

模型近似程度好坏的的衡量

- 引入如下定义来衡量$m_k(d)$近似程度的好坏：

$$\rho_k = \frac{f\left(x^k\right) - f\left(x^k + d^k\right)}{m_k\left(0\right) - m_k\left(d^k\right)} \tag{3}$$

其中d^k为解信赖域子问题得到的迭代方向.根据ρ_k的定义可知，其为函数值实际下降量与预估下降量（即二阶近似模型下降量）的比值.

- 如果ρ_k接近1，说明$m_k(d)$来近似$f(x)$是比较成功的，则应该扩大Δ_k；如果ρ_k非常小甚至为负，就说明我们过分地相信了二阶模型$m_k(d)$，此时应该缩小Δ_k. 使用这个机制可以动态调节Δ_k，让二阶模型$m_k(d)$的定义域处于一个合适的范围.

Algorithm 1 信赖域算法

1: 给定最大半径$\Delta_{\max}$，初始半径Δ_0，初始点x^0，$k \leftarrow 0$.
2: 给定参数$0 \leq \eta < \bar{\rho}_1 < \bar{\rho}_2 < 1$，$\gamma_1 < 1 < \gamma_2$.
3: **while** 未达到收敛准则 **do**
4: 计算子问题(2)得到迭代方向d^k.
5: 根据(3)式计算下降率ρ_k.
6: 更新信赖域半径：

$$\Delta_{k+1} = \begin{cases} \gamma_1 \Delta_k, & \rho_k < \bar{\rho}_1, \\ \min\{\gamma_2 \Delta_k, \Delta_{\max}\}, & \rho_k > \bar{\rho}_2 \text{ 以及 } \|d^k\| = \Delta_k, \\ \Delta_k, & \text{其他}. \end{cases}$$

7: 更新自变量：

$$x^{k+1} = \begin{cases} x^k + d^k, & \rho_k > \eta, \\ x^k, & \text{其他}. \end{cases} \qquad /* \text{ 只有下降比例足够大才更新} */$$

8: $k \leftarrow k + 1$.
9: **end while**

提纲

二次问题最优性条件

引理 (二次问题最优性条件)

若$m(p)$是如下定义的二次问题

$$m(p) = g^T p + \frac{1}{2} p^T B p$$

其中B是任意对称矩阵，则下列命题成立:
(i) B为半正定矩阵时$m(p)$具有最小值当且仅当$g \in range(B)$,并且任意p满足$Bp = -g$均为$m(p)$的全局极小值.
(ii) $m(p)$有唯一极小值当且仅当B是正定的.

Proof.

(i) "$\Leftarrow$": 由于$g \in range\,(B)$,所以存在p^*满足$Bp^* = -g$，所以对任意$\omega \in \mathbb{R}^n$，我们都有

$$m\,(p^* + \omega) = g^T\,(p^* + \omega) + \frac{1}{2}\,(p^* + \omega)^T\,B\,(p^* + \omega)$$

$$= \left(g^T p^* + \frac{1}{2} p^{*T} B p^*\right) + g^T \omega + (Bp^*)^T\,\omega + \frac{1}{2}\omega^T B \omega$$

$$= m\,(p^*) + \frac{1}{2}\omega^T B \omega \geqslant m\,(p^*)$$

由此可知p^*是$m(p)$的最小值

"$\Rightarrow$": 若p^*是$m(p)$的最小值，所以$\nabla m\,(p^*) = Bp^* + g = 0$,所以$g \in range\,(B)$,再由于$\nabla^2 m\,(p^*) = B$,为半正定矩阵,结果得证

□

Proof.

(ii) " $\Leftarrow$" : 由于对任意$\omega(\neq 0) \in \mathbb{R}^n$

$$m\left(p^* + \omega\right) = g^T\left(p^* + \omega\right) + \frac{1}{2}\left(p^* + \omega\right)^T B\left(p^* + \omega\right)$$

$$= \left(g^T p^* + \frac{1}{2} p^{*T} B p^*\right) + g^T\omega + \left(Bp^*\right)^T\omega + \frac{1}{2}\omega^T B\omega$$

$$= m\left(p^*\right) + \frac{1}{2}\omega^T B\omega \geqslant m\left(p^*\right)$$

由于B是正定矩阵,即不等号严格成立,由此可知p^*是$m(p)$的唯一最小值

"$\Rightarrow$": 根据题意B已为半正定矩阵，假设B不为正定矩阵，所以存在$\omega(\neq 0)$使得$B\omega = 0$,则$m(p^*) = m(p^* + \omega)$,矛盾.

最优性条件

定理 (约束优化问题的最优性条件)

d^*是信赖域子问题

$$\min \quad m(d) = f + g^{\mathrm{T}} d + \frac{1}{2} d^{\mathrm{T}} B d, \quad \text{s.t.} \quad \|d\| \leq \Delta \tag{4}$$

的全局极小解当且仅当d^*是可行的且存在$\lambda \geq 0$使得

$$(B + \lambda I) d^* = -g, \tag{5a}$$
$$\lambda(\Delta - \|d^*\|) = 0, \tag{5b}$$
$$(B + \lambda I) \succeq 0. \tag{5c}$$

Proof.

必要性：问题(4)的拉格朗日函数为

$$L(d,\lambda) = f + g^{\mathrm{T}}d + \frac{1}{2}d^{\mathrm{T}}Bd - \frac{\lambda}{2}(\Delta^2 - \|d\|^2),$$

其中乘子$\lambda \geq 0$.

- 由KKT 条件，d^*是可行解，且$\nabla_d L(d^*,\lambda) = (B+\lambda I)d^* + g = 0$.此外由互补条件$\frac{\lambda}{2}(\Delta^2 - \|d^*\|^2) = 0$,整理后就是(5a) 式和(5b) 式.
- 为了证明(5c)式，我们任取d满足$\|d\| = \Delta$，根据最优性，有

$$m(d) \geq m(d^*) = m(d^*) + \frac{\lambda}{2}(\|d^*\|^2 - \|d\|^2).$$

利用(5a)式消去g，代入上式整理有$(d-d^*)^{\mathrm{T}}(B+\lambda I)(d-d^*) \geq 0$,由$d$的任意性可知$B+\lambda I$半正定.

Proof.

再证明充分性. 定义辅助函数

$$\hat{m}(d) = f + g^{\mathrm{T}}d + \frac{1}{2}d^{\mathrm{T}}(B + \lambda I)d = m(d) + \frac{\lambda}{2}d^{\mathrm{T}}d,$$

由条件(5c) 可知$\hat{m}(d)$关于d是凸函数. 根据条件(5a)，d^*满足凸函数一阶最优性条件，结合引理1 可推出d^*是$\hat{m}(d)$的全局极小值点，进而对任意可行解d，我们有

$$m(d) \geq m(d^*) + \frac{\lambda}{2}(\|d^*\|^2 - \|d\|^2).$$

由互补条件(5b) 可知$\lambda(\Delta^2 - \|d^*\|^2) = 0$，代入上式消去$\|d^*\|^2$得

$$m(d) \geq m(d^*) + \frac{\lambda}{2}(\Delta^2 - \|d\|^2) \geq m(d^*).$$

□

提纲

迭代法

- 上述定理提供了一个问题维数n较小时寻找d^*的一个方法，定义$d(\lambda)$如下

$$d(\lambda) = -(B + \lambda I)^{-1} g, \tag{6}$$

根据互补条件(5b)，当$\lambda > 0$时必有$\|d(\lambda)\| = \Delta$；根据半正定条件(5c)，λ须大于等于B的最小特征值的相反数.

- 设B有特征值分解$B = Q\Lambda Q^{\mathrm{T}}$，其中$Q = [q_1, q_2, \cdots, q_n]$是正交矩阵，$\Lambda = \mathrm{Diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)$是对角矩阵，$\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$是$B$的特征值.

$$d(\lambda) = -Q(\Lambda + \lambda I)^{-1} Q^{\mathrm{T}} g = -\sum_{i=1}^{n} \frac{q_i^{\mathrm{T}} g}{\lambda_i + \lambda} q_i. \tag{7}$$

这正是$d(\lambda)$的正交分解，由正交性可容易求出

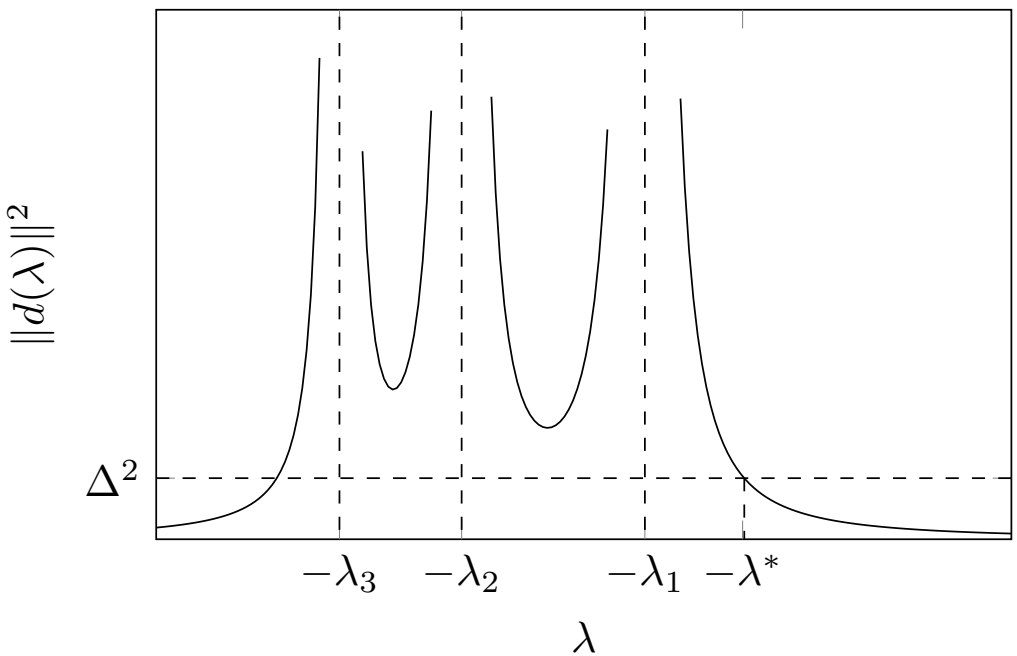
$$\|d(\lambda)\|^2 = \sum_{i=1}^{n} \frac{(q_i^{\mathrm{T}} g)^2}{(\lambda_i + \lambda)^2}. \tag{8}$$

迭代法

- 由(8)式，当$\lambda > -\lambda_1$且$q_1^{\mathrm{T}} g \neq 0$时，$\|d(\lambda)\|^2$是关于$\lambda$的严格减函数，且有

$$\lim_{\lambda \to \infty} \|d(\lambda)\| = 0, \qquad \lim_{\lambda \to -\lambda_1 +} \|d(\lambda)\| = +\infty.$$

- 由连续函数介值定理，$\|d(\lambda)\| = \Delta$的解必存在且唯一.
- 所以寻找λ^*已经转化为一个一元方程求根问题，可使用牛顿法求解.



困难情形

- 上面的分析中假定了$q_1^{\mathrm{T}}g \neq 0$，在实际中这个条件未必满足.
  当$q_1^{\mathrm{T}}g = 0$时，(8)式将没有和λ_1相关的项. 此时未必存在$\lambda^* > -\lambda_1$使得$\|d(\lambda^*)\| = \Delta$成立. 记

$$M = \lim_{\lambda \to -\lambda_1 +} \|d(\lambda)\|$$

- 当$M \geq \Delta$时，仍然可以根据介值定理得出$\lambda^*(> -\lambda_1)$的存在性；
- 当$M < \Delta$时，无法利用前面的分析求出λ^*和d^*，此时信赖域子问题变得比较复杂.
  实际上，$q_1^{\mathrm{T}}g = 0$且$M < \Delta$的情形被称为“困难情形(hard case)”.
  此情形发生时，区间$(-\lambda_1, +\infty)$中的点无法使得(5b) 成立，而定理15 的结果说明$\lambda^* \in [-\lambda_1, +\infty)$，因此必有$\lambda^* = -\lambda_1$.

困难情形

- 为了求出d^*，可以利用（奇异）线性方程组(5a)解的结构，其通解可以写为

$$d(\alpha) = -\sum_{i=2}^{n} \frac{q_i^{\mathrm{T}} g}{\lambda_i - \lambda_1} q_i + \alpha q_1, \quad \alpha \in \mathbb{R}.$$

- 由正交性，

$$\|d(\alpha)\|^2 = \alpha^2 + \sum_{i=2}^{n} \frac{(q_i^{\mathrm{T}} g)^2}{(\lambda_i - \lambda_1)^2}.$$

注意在困难情形中有$M = \sqrt{\sum_{i=2}^{n} \frac{(q_i^{\mathrm{T}} g)^2}{(\lambda_i - \lambda_1)^2}} < \Delta$，因此必存在$\alpha^*$使得$\|d(\alpha^*)\| = \Delta$. 这就求出了$d^*$的表达式.

截断共轭梯度法介绍

下面再介绍一种信赖域子问题的求解方法.

- 既然信赖域子问题的解不易求出，则求出其近似解，Steihaug在1983 年对共轭梯度法进行了改造，使其成为能求解子问题的算法. 此算法能够应用在大规模问题中，是一种非常有效的信赖域子问题的求解方法.

- 由于子问题和一般的二次极小化问题相差一个约束，如果先不考虑其中的约束$\|d\| \leq \Delta$而直接使用共轭梯度法求解，在迭代过程中找到合适的迭代点作为信赖域子问题的近似解，检测到负曲率或者到达信赖域边界$\|d\| = \Delta$即终止. 这就是截断共轭梯度法(见书P260)的基本思想.

截断共轭梯度法介绍

- 为了介绍截断共轭梯度法，我们简要回顾一下标准共轭梯度法的迭代过程. 对于二次极小化问题

$$\min_s \quad q(s) \xlongequal{\text{def}} g^{\mathrm{T}}s + \frac{1}{2}s^{\mathrm{T}}Bs,$$

- 给定初值$s^0 = 0, r^0 = g, p^0 = -g$，共轭梯度法的迭代过程为

$$\begin{aligned}
\alpha_{k+1} &= \frac{\|r^k\|^2}{(p^k)^{\mathrm{T}}Bp^k}, \\
s^{k+1} &= s^k + \alpha_k p^k, \\
r^{k+1} &= r^k + \alpha_k Bp^k, \\
\beta_k &= \frac{\|r^{k+1}\|^2}{\|r^k\|^2}, \\
p^{k+1} &= -r^{k+1} + \beta_k p^k,
\end{aligned}$$

其中迭代序列$\{s^k\}$最终的输出即为二次极小化问题的解，算法的终止准则是判断$\|r^k\|$是否足够小.

截断共轭梯度法

Algorithm 2 截断共轭梯度法1 (Steihaug-CG)

1: 给定精度$\varepsilon > 0$，初始化$s^0 = 0, r^0 = g, p^0 = -g$，$k \leftarrow 0$.
2: **if** $\|p^0\| \leq \varepsilon$ **then**
3: 算法停止，输出$s = 0$.
4: **end if**
5: **LOOP**
6: **if** $(p^k)^{\mathrm{T}} B p^k \leq 0$ **then**
7: 计算$\tau > 0$使得$\|s^k + \tau p^k\| = \Delta$.
8: 算法停止，输出$s = s^k + \tau p^k$.
9: **end if**
10: 计算$\alpha_k = \frac{\|r^k\|^2}{(p^k)^{\mathrm{T}} B p^k}$，更新$s^{k+1} = s^k + \alpha_k p^k$.
11: 接下页

截断共轭梯度法

Algorithm 3 截断共轭梯度法2 (Steihaug-CG)

1: **if** $\|s^{k+1}\| \geq \Delta$ **then**
2: 计算$\tau > 0$使得$\|s^k + \tau p^k\| = \Delta$.
3: 算法停止，输出$s = s^k + \tau p^k$.
4: **end if**
5: 计算$r^{k+1} = r^k + \alpha_k B p^k$.
6: **if** $\|r^{k+1}\| < \varepsilon \|r^0\|$ **then**
7: 算法停止，输出$s = s^{k+1}$.
8: **end if**
9: 计算$\beta_k = \frac{\|r^{k+1}\|^2}{\|r^k\|^2}$，更新$p^{k+1} = -r^{k+1} + \beta_k p^k$.
10: $k \leftarrow k + 1$.
11: **ENDLOOP**

截断共轭梯度法

- 截断共轭梯度法则是给标准的共轭梯度法增加了两条终止准则，并对最后一步的迭代点s^k进行修正来得到信赖域子问题的解．考虑到矩阵B不一定是正定矩阵，在迭代过程中可能会产生如下三种情况：
  1. $(p^k)^{\mathrm{T}}Bp^k \leq 0$，即$B$不是正定矩阵．我们知道共轭梯度法不能处理非正定的线性方程组，遇到这种情况应该立即终止算法．但根据这个条件也找到了一个负曲率方向，此时只需要沿着这个方向走到信赖域边界即可．
  2. $(p^k)^{\mathrm{T}}Bp^k > 0$但$\|s^{k+1}\| \geq \Delta$，这表示若继续进行共轭梯度法迭代，则点$s^{k+1}$将处于信赖域之外或边界上，此时必须马上停止迭代，并在s^k和s^{k+1}之间找一个近似解．
  3. $(p^k)^{\mathrm{T}}Bp^k > 0$且$\|r^{k+1}\|$充分小，这表示若共轭梯度法成功收敛到信赖域内．子问题(2)和不带约束的二次极小化问题是等价的．
- 从上述终止条件来看截断共轭梯度法仅仅产生了共轭梯度法的部分迭代点，这也是该方法名字的由来．

截断共轭梯度法

截断共轭梯度法的迭代序列$\{s^k\}$有非常好的性质，实际上我们可以证明如下定理：

定理

设$q(s)$是任意外迭代步信赖域子问题的目标函数，令$\{s^j\}$是由截断共轭梯度算法产生的迭代序列，则在算法终止前$q(s^j)$是严格单调递减的，即

$$q(s^{j+1}) < q(s^j). \tag{9}$$

并且$\|s^j\|$是严格单调递增的，即

$$0 = \|s^0\| < \|s^1\| < \cdots < \|s^{j+1}\| < \cdots \leq \Delta. \tag{10}$$

Proof.

设迭代在第t步终止. 根据算法2，在终止前，若$(p^j)^{\mathrm{T}}Bp^j > 0,\ j < t$一直成立. 此时算法即共轭梯度法，容易证明(9)式和(10)式.

又$q(s)$在点s^j处的梯度为r^j，由共轭梯度法性质$(r^j)^{\mathrm{T}}p^i = 0, i < j$，所以$(r^j)^{\mathrm{T}}p^j = (r^j)^{\mathrm{T}}(-r^j + \beta_{j-1}p^{j-1}) = -\|r^j\|^2 < 0$, 即$p^j$是下降方向. 而$\alpha_j$的选取为精确步长,因此有$q(s^{j+1}) < q(s^j)$. 此外由$s^j$的定义,$s^j = \sum_{i=0}^{j-1}\alpha_i p^i, \alpha_i > 0$.再根据共轭梯度法的性质:

$$(p^j)^{\mathrm{T}}s^j = \sum_{i=0}^{j-1}\alpha_i(p^j)^{\mathrm{T}}p^i = \sum_{i=0}^{j-1}\alpha_i\frac{\|r^j\|^2}{\|r^i\|^2}\|p^i\|^2 > 0.$$

结合以上表达式可得

$$\|s^{j+1}\|^2 = \|s^j + \alpha_j p^j\|^2 = \|s^j\|^2 + 2\alpha_j(p^j)^{\mathrm{T}}s^j + \alpha_j^2\|p^j\|^2 > \|s^j\|^2. \qquad \square$$

截断共轭梯度法

实际上，我们还可进一步说明截断共轭梯度算法的输出s满足如下关系：

$$q(s) \leq q(s^t), \quad \|s^t\| \leq \|s\|,$$

其中t为算法终止时的迭代数. 这只需要分别讨论三种终止条件即可.

1. 若$(p^t)^{\mathrm{T}} B p^t \leq 0$，则$p^t$是负曲率方向，沿着负曲率方向显然有$q(s) \leq q(s^t)$. 注意到此时$\|s\| = \Delta$，因此有$\|s^t\| \leq \|s\| = \Delta$.
2. 若$(p^t)^{\mathrm{T}} B p^t > 0$但$\|s^{t+1}\| \geq \Delta$，根据最速下降法的性质，$q(s^t + \alpha p^t)$关于$\alpha \in (0, \alpha_t]$单调下降，根据$\tau$的取法显然有$q(s) \leq q(s^t)$. 此时依然有$\|s\| = \Delta$，因此$\|s^t\| \leq \|s\| = \Delta$仍成立.
3. 若$(p^t)^{\mathrm{T}} B p^t > 0$且$\|r^{t+1}\| \leq \varepsilon \|r^0\|$，此时算法就是共轭梯度法，结论自然成立.

提纲

柯西点定义

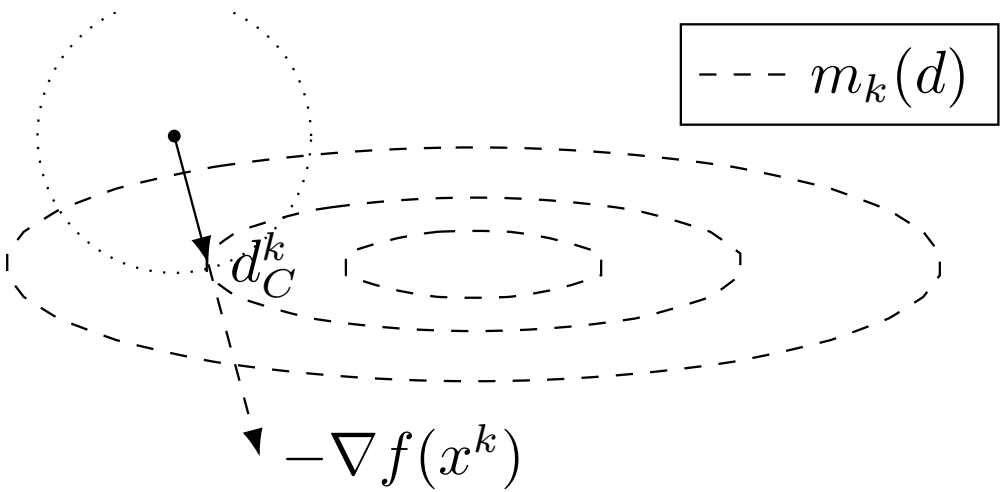
定义 (柯西点)

设$m_k(d)$是$f(x)$在点$x = x^k$处的二阶近似，τ_k 为如下优化问题的解：

$$\begin{aligned} \min \quad & m_k(-\tau \nabla f(x^k)), \\ \text{s.t.} \quad & \|\tau \nabla f(x^k)\| \le \Delta_k,\ \tau \ge 0. \end{aligned}$$

则称$x_C^k \xlongequal{\text{def}} x^k + d_C^k$为**柯西点**，其中$d_C^k = -\tau_k \nabla f(x^k)$.

由柯西点的定义，它实际上是在约束下对$m_k(d)$进行了一次精确线搜索的梯度法，



柯西点性质

实际上，给定$m_k(d)$，柯西点可以显式计算出来．为了方便我们用g^k表示$\nabla f(x^k)$，根据τ_k的定义，容易计算出其表达式为

$$\tau_k = \begin{cases} \dfrac{\Delta_k}{\|g^k\|}, & (g^k)^{\mathrm{T}} B^k g^k \leq 0, \\ \min\left\{ \dfrac{\|g^k\|^2}{(g^k)^{\mathrm{T}} B^k g^k}, \dfrac{\Delta_k}{\|g^k\|} \right\}, & \text{其他}. \end{cases}$$

引理 (柯西点的下降量)

设d_C^k为求解柯西点产生的下降方向，则

$$m_k(0) - m_k(d_C^k) \geq c_1 \|g^k\| \min\left\{ \Delta_k, \frac{\|g^k\|}{\|B^k\|_2} \right\}. \tag{11}$$

其中$c_1 = \frac{1}{2}$

Proof.

分三种情况证明该结论.(方便起见证明中忽略上标k)

- 首先考虑$g^TBg \leqslant 0$时的情况

$$\begin{aligned} m(d_C) - m(0) &= m(-\Delta g/||g||) - f \\ &= -\frac{\Delta}{||g||}||g||^2 + \frac{1}{2}\frac{\Delta^2}{||g||^2}g^TBg \\ &\leqslant -\Delta||g|| \\ &\leqslant -||g||\min\left(\Delta, \frac{||g||}{||B||}\right) \end{aligned}$$

即(11)式成立.

Proof.

- 然后考虑$g^TBg > 0, \frac{||g||^3}{\Delta g^TBg} \leqslant 1$时的情况

$$
\begin{aligned}
m(d_C) - m(0) &= -\frac{||g||^4}{g^TBg} + \frac{1}{2} g^TBg \frac{||g||^4}{(g^TBg)^2} \\
&= -\frac{1}{2} \frac{||g||^4}{g^TBg} \\
&\leqslant -\frac{1}{2} \frac{||g||^4}{||B||\ ||g||^2} \\
&= -\frac{1}{2} \frac{||g||^2}{||B||} \\
&\leqslant -\frac{1}{2} ||g|| \min\left(\Delta, \frac{||g||}{||B||}\right)
\end{aligned}
$$

即(11)式成立.

Proof.

- 最后考虑$\frac{||g||^3}{\Delta g^T B g} \geqslant 1$时的情况

$$
\begin{aligned}
m(d_C) - m(0) &= -\frac{\Delta}{||g||}||g||^2 + \frac{1}{2}\frac{\Delta^2}{||g||^2} g^T B g \\
&\leqslant -\Delta||g|| + \frac{1}{2}\frac{\Delta^2}{||g||^2}\frac{||g||^3}{\Delta} \\
&= -\frac{1}{2}\Delta||g|| \\
&\leqslant -\frac{1}{2}||g|| \min\left(\Delta, \frac{||g||}{||B||}\right)
\end{aligned}
$$

即(11)式成立. □

提纲

全局收敛性

回顾信赖域算法，我们引入了一个参数η来确定是否应该更新迭代点．这分为两种情况：当$\eta = 0$时，只要原目标函数有下降量就接受信赖域迭代步的更新；当$\eta > 0$时，只有当改善量ρ_k达到一定程度时再进行更新．在这两种情况下得到的收敛性结果是不同的，我们分别介绍这两种结果．在$\eta = 0$的条件下有如下收敛性定理：

定理 (全局收敛性1)

设近似海瑟矩阵B^k有界，即$\|B^k\|_2 \leq M, \forall\, k$，$f(x)$在下水平集$\mathcal{L} = \{x \,|\, f(x) \leq f(x^0)\}$上有下界，且$\nabla f(x)$在$\mathcal{L}$的一个开邻域$S(R_0)$内利普希茨连续．若$d^k$为信赖域子问题的近似解且满足(11)式，信赖域算法选取参数$\eta = 0$，则

$$\liminf_{k\to\infty} \|\nabla f(x^k)\| = 0,$$

即x^k的聚点中包含稳定点．

全局收敛性1证明

Proof.

首先由ρ_k的定义可以得到：

$$\begin{aligned} |\rho_k - 1| &= |\frac{\left(f\left(x^k\right) - f\left(x^k + d^k\right)\right) - \left(m_k\left(0\right) - m_k\left(d^k\right)\right)}{m_k\left(0\right) - m_k\left(d^k\right)}| \\ &= |\frac{m_k\left(d^k\right) - f\left(x^k + d^k\right)}{m_k\left(0\right) - m_k\left(d^k\right)}| \end{aligned} \tag{12}$$

再由泰勒展式可得:$(g(x^k) = \nabla f(x^k))$

$$f\left(x^k + d^k\right) = f\left(x^k\right) + g\left(x^k\right)^T d^k + \int_0^1 \left[g\left(x^k + td^k\right) - g\left(x^k\right)\right]^T d^k dt$$

其中$t \in (0,1)$,再由m的定义可得(接下页)

全局收敛性1证明

Proof.

$$
\begin{aligned}
|m_k\left(d^k\right)-f\left(x^k+d^k\right)| &= |\frac{1}{2}d^{kT}B^k d^k-\int_0^1\left[g\left(x^k+td^k\right)-g\left(x^k\right)\right]^T d^k dt| \\
&\leqslant (\beta/2)\left\|d^k\right\|^2+\beta_1\left\|d^k\right\|^2,
\end{aligned}
\tag{13}
$$

其中β_1为Lipschitz常数,并假设$\left\|d^k\right\| \leqslant R_0$以保证$x^k$和$x^k+td^k$均位于$S(R_0)$中. 再用反证法，反设结论不成立,则存在$\varepsilon$和指标$K$有:

$$
\left\|g^k\right\| \geqslant \varepsilon, \forall k \geqslant K
$$

从(11)式可知

$$
m_k(0)-m_k(d^k) \geq c_1\left\|g^k\right\| \min\left\{\Delta_k, \frac{\left\|g^k\right\|}{\left\|B^k\right\|_2}\right\} \geqslant c_1\varepsilon \min\left\{\Delta_k, \frac{\varepsilon}{\beta}\right\} \tag{14}
$$

全局收敛性1证明

Proof.

由(13)和(14)可知

$$|\rho_k - 1| \leqslant \frac{\Delta_k^2 \left(\beta/2 + \beta_1\right)}{c_1 \epsilon \min\left(\Delta_k, \epsilon/\beta\right)} \tag{15}$$

下面给出k足够大时Δ的一个上界$\overline{\Delta}$ 的定义，即$\Delta_k \leqslant \overline{\Delta}$:

$$\overline{\Delta} = \min\left(\frac{1}{2}\frac{c_1\varepsilon}{\left(\beta/2 + \beta_1\right)}, R_0\right)$$

并注意到$c_1 \leqslant 1$所以$\min(\Delta_k, \varepsilon/\beta) = \Delta_k$,由此可得:

$$|\rho_k - 1| \leqslant \frac{\Delta_k^2 \left(\beta/2 + \beta_1\right)}{c_1 \varepsilon \Delta_k} \leqslant \frac{\overline{\Delta}\left(\beta/2 + \beta_1\right)}{c_1\varepsilon} \leqslant \frac{1}{2}$$

全局收敛性1证明

Proof.

因此$\rho_k > \frac{1}{4}$,再由于算法(1)可知$\Delta_{k+1} \leqslant \Delta_k$只可能在$\Delta_k \leqslant \overline{\Delta}$时成立.所以我们有

$$\Delta_k \geqslant \min(\Delta_K, \overline{\Delta}/4), \quad \forall k \geqslant K \tag{16}$$

假设有含有无穷项的子列$\mathcal{K}$指标集,使得$\rho_k \geqslant \frac{1}{4}, k \in \mathcal{K}$,
则对于$k \in \mathcal{K}$,并且$k \geqslant K$,由(14)可知

$$f\left(x^k\right) - f\left(x^{k+1}\right) \geqslant \frac{1}{4}\left[m_k(0) - m_k\left(d^k\right)\right] \quad \geqslant \frac{1}{4} c_1 \varepsilon \min\left(\Delta_k, \varepsilon/\beta\right) \tag{17}$$

再由于f的下水平集有界，由该不等式可以得到

$$\lim_{k \in \mathcal{K}, k \to \infty} \Delta_k = 0$$

与(16)矛盾,由此可知$\mathcal{K}$不存在，所以k足够大时,Δ_k将会在每次迭代中缩小γ_1倍,所以$\lim_{k \to \infty} \Delta_k = 0$,与(16)矛盾，因此原假设不成立. □

全局收敛性

定理(全局收敛性1)表明若无条件接受信赖域子问题的更新，则信赖域算法仅仅有子序列的收敛性，迭代点序列本身不一定收敛．根据下面的定理则说明选取$\eta > 0$可以改善收敛性结果．

定理 (全局收敛性2)

在定理*(全局收敛性1)*的条件下，若信赖域算法选取参数$\eta > 0$, 且信赖域子问题近似解d^k满足(11) 式，则

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

和牛顿类算法不同，信赖域算法具有全局收敛性，因此它对迭代初值选取的要求比较弱．而牛顿法的收敛性极大地依赖初值的选取．

全局收敛性2证明

Proof.

取$g^m \neq 0(g^m = \nabla f(x^m))$,否则,$g^m \equiv 0$,矛盾,设$\beta_1$为Lipschitz常数,则有

$$||g(x) - g^m|| \leqslant \beta_1 ||x - x^m||$$

对任意$x \in S(R_0)$均成立,再定义ϵ和R如下:

$$\epsilon = \frac{1}{2}||g^m||,\ R = \min\left(\frac{\epsilon}{\beta_1}, R_0\right)$$

并且注意到$\mathcal{B}(x^m, R) = \{x|\ ||x - x^m|| \leqslant R\}$包含在$S(R_0)$中,故Lipschitz 连续性成立,则有

$$x \in \mathcal{B}(x^m, R) \Rightarrow ||g(x)|| \geqslant ||g^m|| - ||g(x) - g^m|| \geqslant \frac{1}{2}||g^m|| = \epsilon$$

全局收敛性2证明

Proof.

若$\left\{x^k\right\}_{k\geqslant m}$均在$\mathcal{B}\left(x^m,R\right)$中,则有$\left\|g^k\right\|\geqslant\epsilon>0$对所有$k\geqslant m$成立，则由全局收敛性1的证明可知此结论不成立，所以设$l\geqslant m$,且x^{l+1}是第一个离开$\mathcal{B}\left(x^m,R\right)$的迭代点.则对于$k=m,m+1,...,l$由(14)可得：

$$
\begin{aligned}
f\left(x^m\right)-f\left(x^{l+1}\right) &= \sum_{k=m}^{l} f\left(x^k\right)-f\left(x^{k+1}\right) \\
&\geqslant \sum_{k=m,x^k\neq x^{k+1}}^{l} \eta\left[m_k\left(0\right)-m_k\left(d^k\right)\right] \qquad (18) \\
&\geqslant \sum_{k=m,x^k\neq x^{k+1}}^{l} \eta c_1 \min\left(\Delta_k,\frac{\epsilon}{\beta}\right)
\end{aligned}
$$

全局收敛性2证明

Proof.

所以若$\Delta \leqslant \frac{\epsilon}{\beta}$恒成立,则有

$$f\left(x^{m}\right)-f\left(x^{l+1}\right) \geqslant \sum_{k=m, x^{k} \neq x^{k+1}}^{l} \eta c_{1} \epsilon \Delta_{k} \geqslant \eta c_{1} \epsilon R=\eta c_{1} \epsilon \min \left(\frac{\epsilon}{\beta_{1}}, R_{0}\right) \tag{19}$$

否则,可得$\Delta_k > \epsilon/\beta$对某些$k = m, m+1, \ldots l$成立,则有

$$f\left(x^{m}\right)-f\left(x^{l+1}\right) \geqslant \eta c_{1} \epsilon \frac{\epsilon}{\beta} \tag{20}$$

由于$\left\{f\left(x^{k}\right)\right\}_{k=0}^{\infty}$单调递减并且有下界,则有

$$f\left(x^{k}\right) \downarrow f^{*}$$

全局收敛性2证明

Proof.

因此,利用(19)与(20)式，则有

$$\begin{aligned} f\left(x^m\right)-f^* &\geqslant f\left(x^m\right)-f\left(x^{l+1}\right) \\ &\geqslant \eta c_1 \varepsilon \min\left(\frac{\varepsilon}{\beta}, \frac{\varepsilon}{\beta_1}, R_0\right) \\ &= \frac{1}{2}\eta c_1 \left\|g^m\right\| \min\left(\frac{\left\|g^m\right\|}{2\beta}, \frac{\left\|g^m\right\|}{2\beta_1}, R_0\right) > 0 \end{aligned} \tag{21}$$

再由于$f\left(x^k\right)\downarrow f^*$,必有$g^m \to 0$,得证. □

局部收敛性

- 在构造信赖域子问题时利用了$f(x)$的二阶信息，它在最优点附近应该具有牛顿法的性质．特别地，当近似矩阵B^k取为海瑟矩阵$\nabla^2 f(x^k)$时，根据信赖域子问题的更新方式，二次模型$m_k(d)$将会越来越逼近原函数$f(x)$.

定理

设$f(x)$在最优点$x = x^*$的一个邻域内二阶连续可微，且$\nabla f(x)$利普希茨连续，在最优点x^*处二阶充分条件成立，即$\nabla^2 f(x) \succ 0$．若迭代点列$\{x^k\}$收敛到x^*，并且在迭代中选取B^k为海瑟矩阵$\nabla^2 f(x^k)$，且对充分大的k，任意满足(11)式的信赖域子问题算法产生的迭代方向d^k均满足

$$\|d^k - d_N^k\| = o(\|d_N^k\|), \tag{22}$$

其中d_N^k为第k步迭代的牛顿方向且满足假设$\|d_N^k\| \leq \frac{\Delta_k}{2}$.则当$k$足够大时,信赖域约束$\Delta_k$将未被激活.

局部收敛性证明

Proof.

我们证明在k充分大时$\|d_N^k\| \leqslant \frac{1}{2}\Delta_k$与$\|d^k\| \leqslant \Delta_k$.则结论得证.
设k充分大时有$o(\|d_N^k\|) < \|d_N^k\|$,当$\|d_N^k\| \geqslant \frac{1}{2}\Delta_k$时,则有$\|d^k\| \geqslant 2\|d_N^k\|$,但若$\|d_N^k\| > \frac{1}{2}\Delta_k$,我们有$\|d^k\| \geqslant \Delta_k \geqslant 2\|d_N^k\|$,两种情况均有：$\|d^k\| \leqslant 2\|d_N^k\| \leqslant 2\|\nabla^2 f(x^k)^{-1}\|\|g^k\|$ 再由(11)可知:

$$m_k(0) - m_k(d^k) \geqslant c_1 \left\|g^k\right\| \min\left(\Delta_k, \frac{\left\|g^k\right\|}{\left\|\nabla^2 f(x^k)\right\|}\right)$$

$$\geqslant c_1 \frac{\left\|d^k\right\|}{2\left\|\nabla^2 f(x^k)^{-1}\right\|} \min\left(\left\|d^k\right\|, \frac{\left\|d^k\right\|}{2\left\|\nabla^2 f(x^k)\right\|\left\|\nabla^2 f(x^k)^{-1}\right\|}\right)$$

$$\geqslant c_1 \frac{\left\|d^k\right\|^2}{4\left\|\nabla^2 f(x^k)^{-1}\right\|^2\left\|\nabla^2 f(x^k)\right\|}$$

局部收敛性证明

Proof.

再由于$x^k \to x^*$,并利用$\nabla^2 f(x)$的连续性与$\nabla^2 f(x^*)$的正定性知当k充分大时

$$\frac{c_1}{4\left\|\nabla^2 f\left(x^k\right)^{-1}\right\|^2\left\|\nabla^2 f\left(x^k\right)\right\|} \geqslant \frac{c_1}{8\left\|\nabla^2 f\left(x^*\right)^{-1}\right\|^2\left\|\nabla^2 f\left(x^*\right)\right\|} \overset{\text{def}}{=\!=} c_3$$

因此可得当k足够大时有：

$$m_k(0) - m_k\left(d^k\right) \geqslant c_3\left\|d^k\right\|^2$$

再利用$\nabla^2 f(x)$的Lipschitz连续性并利用Taylor展式可得

$$\left|\left(f\left(x^k\right) - f\left(x^k + d^k\right)\right) - \left(m_k(0) - m_k\left(d^k\right)\right)\right|$$

局部收敛性证明

Proof.

$$
\begin{aligned}
&= \left|\frac{1}{2} d^{k} T \nabla^{2} f\left(x^{k}\right) d^{T} - \frac{1}{2} \int_{0}^{1} d^{kT} \nabla^{2} f\left(x^{k} + t d^{k}\right) d^{k} dt\right| \\
&\leqslant \frac{L}{4} \left\| d^{k} \right\|^{3}
\end{aligned}
$$

其中L为Lipschitz常数，因此由ρ_k的定义可得k足够大时:

$$
\left|\rho_{k} - 1\right| \leqslant \frac{\left\| d^{k} \right\|^{3} (L/4)}{c_{3} \left\| d^{k} \right\|^{2}} = \frac{L}{4 c_{3}} \left\| d^{k} \right\| \leqslant \frac{L}{4 c_{3}} \Delta_{k}
$$

又由于信赖域半径只可能在$\rho_k < \frac{1}{4}$时减小,所以由上式可得序列Δ_k不收敛到0,又由于$x^k \to x^*$,所以我们有$\|d_N^k\| \to 0$,因此$\|d^k\| \to 0$,所以信赖域半径约束在k足够大时未激活,并且$\|d_N^k\| \leqslant \frac{1}{2}\Delta_k$最终一直成立

局部收敛性

推论 (信赖域算法的局部收敛速度)

在定理5的条件下，信赖域算法产生的迭代序列$\{x^k\}$具有Q-超线性收敛速度.

Proof.

根据牛顿法收敛性定理，对牛顿方向，

$$\|x^k + d_N^k - x^*\| = \mathcal{O}(\|x^k - x^*\|^2),$$

因此得到估计$\|d_N^k\| = \mathcal{O}(\|x^k - x^*\|)$. 又根据牛顿法的局部收敛性定理，

$$\|x^k + d^k - x^*\| \leq \|x^k + d_N^k - x^*\| + \|d^k - d_N^k\| = o(\|x^k - x^*\|).$$

这说明信赖域算法是Q-超线性收敛的. □

收敛性

- 容易看出，若在迭代后期$d^k = d_N^k$能得到满足，则信赖域算法是Q-二次收敛的.
- 很多算法都会有这样的性质，例如前面提到的截断共轭梯度法和dogleg 方法. 因此在实际应用中，截断共轭梯度法是最常用的信赖域子问题的求解方法，使用此方法能够同时兼顾全局收敛性和局部Q-二次收敛性.

提纲

应用举例

- 考虑逻辑回归问题

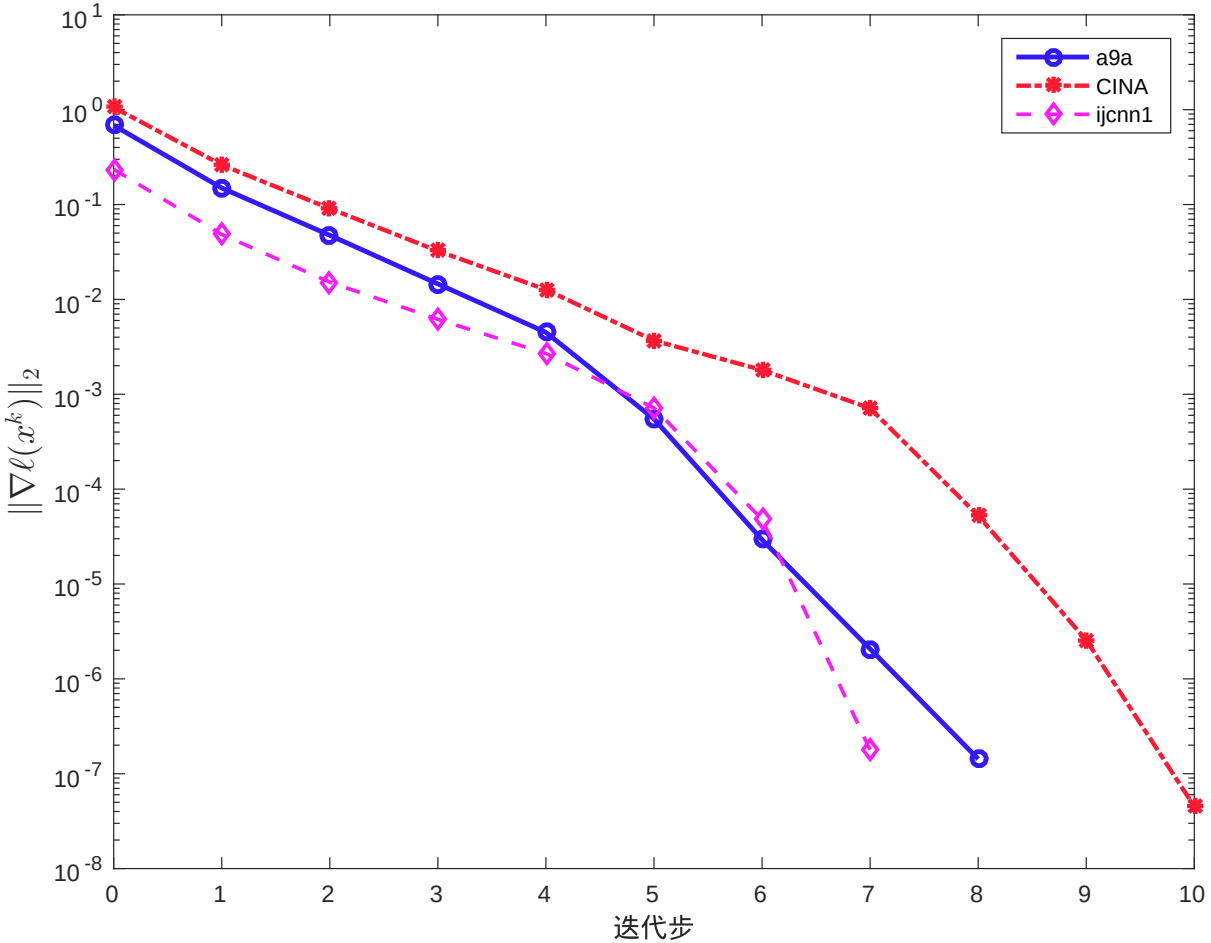
$$\min_x \quad \frac{1}{m}\sum_{i=1}^{m}\ln(1+\exp(-b_i a_i^{\mathrm{T}}x))+\lambda\|x\|_2^2, \tag{23}$$

  这里选取$\lambda=\frac{1}{100m}$.

- 同样地，我们选取LIBSVM上的数据集，调用信赖域算法求解代入数据集后的问题(23)，其迭代收敛过程见图1.
- 其中使用截断共轭梯度法来求解信赖域子问题，精度设置同牛顿法一致.

应用举例

从图中可以看到，在精确解附近梯度范数具有Q-超线性收敛性质．由于这个问题是强凸的，所以选取一个较大的初始信赖域半径（$\sqrt{n}$）．在数据集a9a和ijcnn1的求解中，信赖域子问题的求解没有因为超出信赖域边界而停机，因此牛顿法的数值表现一致．



参考文献

J. Nocedal and S. J. Wright.

Numerical Optimization.

Springer, second edition, 2006.

P. E. Gill, W. Murray, and Wright M. H.

Practical Optimization.

Academic Press, 1981.

Powell MJ D.

A new algorithm for unconstrained optimization[G]//Nonlinear Programming

Amsterdam:Elsevier,1970:31-65