梯度法的困难

考虑无约束优化问题

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

使用梯度下降法，给出的迭代格式是

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

由于梯度下降的基本策略是沿一阶最速下降方向迭代。当$\nabla^2 f(x)$的条件数较大时，它的收敛速度比较缓慢(只用到一阶信息).

如果$f(x)$足够光滑，我们可以利用$f(x)$的二阶信息改进下降方向，以加速算法的迭代.

提纲

1 经典牛顿法

2 收敛性分析

3 修正牛顿法

4 非精确牛顿法

5 自适应的正则化牛顿算法

6 应用举例: 逻辑回归模型

经典牛顿法

对于可微二次函数$f(x)$, 对于第k步迭代,我们现在考虑目标函数 f 在点 x_k 的二阶Taylor近似

$$f\left(x^k+d^k\right)=f\left(x^k\right)+\nabla f\left(x^k\right)^{\mathrm{T}} d^k+\frac{1}{2}\left(d^k\right)^{\mathrm{T}} \nabla^2 f\left(x^k\right) d^k+o\left(\left\|d^k\right\|^2\right),$$

忽略高阶项$o\left(\left\|d^k\right\|^2\right)$, 并将等式右边视作 d^k 的函数并极小化, 得

$$\nabla^2 f\left(x^k\right) d^k=-\nabla f\left(x^k\right). \tag{1}$$

方程(1)被称为**牛顿方程**, d^k 被称为**牛顿方向**.
若 $\nabla^2 f\left(x^k\right)$ 非奇异, 则可得到迭代格式

$$x^{k+1}=x^k-\nabla^2 f\left(x^k\right)^{-1} \nabla f\left(x^k\right). \tag{2}$$

注意上式中步长$\alpha_k=1$. (2)这种迭代格式被称为经典牛顿法.

提纲

经典牛顿法的收敛性

定理

经典牛顿法的收敛性 假设 f 二阶连续可微,且存在 x^* 的一个邻域 $N_\delta(x^*)$ 及常数$L>0$ 使得

$$\left\|\nabla^2 f(x)-\nabla^2 f(y)\right\| \leqslant L\|x-y\|, \quad \forall x, y \in N_\delta(x^*)$$

如果 $f(x)$ 满足$\nabla f(x^*)=0, \nabla^2 f(x^*) \succ 0$, 则对于迭代格式(2)有:

- 如果初始点离x^* 足够近,则迭代点列$\{x^k\}$ 收敛到x^*;
- $\{x^k\}$ Q-二次收敛到x^*;
- $\{\|\nabla f(x^k)\|\}$ Q-二次收敛到0.

根据经典牛顿法定义以及$\nabla f\left(x^*\right)=0$,得

$$\begin{aligned} x^{k+1}-x^* &= x^k-\nabla^2 f\left(x^k\right)^{-1} \nabla f\left(x^k\right)-x^* \\ &= \nabla^2 f\left(x^k\right)^{-1}\left[\nabla^2 f\left(x^k\right)\left(x^k-x^*\right)-\left(\nabla f\left(x^k\right)-\nabla f\left(x^*\right)\right)\right], \end{aligned} \tag{3}$$

注意到

$$\nabla f\left(x^k\right)-\nabla f\left(x^*\right)=\int_0^1 \nabla^2 f\left(x^k+t\left(x^*-x^k\right)\right)\left(x^k-x^*\right) \mathrm{d} t,$$

由此

$$\begin{aligned} & \left\|\nabla^2 f\left(x^k\right)\left(x^k-x^*\right)-\left(\nabla f\left(x^k\right)-\nabla f\left(x^*\right)\right)\right\| \\ = & \left\|\int_0^1\left[\nabla^2 f\left(x^k+t\left(x^*-x^k\right)\right)-\nabla^2 f\left(x^k\right)\right]\left(x^k-x^*\right) \mathrm{d} t\right\| \\ \leqslant & \int_0^1\left\|\nabla^2 f\left(x^k+t\left(x^*-x^k\right)\right)-\nabla^2 f\left(x^k\right)\right\|\left\|x^k-x^*\right\| \mathrm{d} t \\ \leqslant & \left\|x^k-x^*\right\|^2 \int_0^1 L t \, \mathrm{d} t \text{(Lip. 连续性)} = \frac{L}{2}\left\|x^k-x^*\right\|^2 . \end{aligned} \tag{4}$$

注意到, $\exists r>0$, 当 $\|x-x^*\| \leqslant r$ 时有 $\left\|\nabla^2 f(x)^{-1}\right\| \leqslant 2\left\|\nabla^2 f\left(x^*\right)^{-1}\right\|$ 成立 (请思考为何?), 故结合(3)及(4),得到

$$
\begin{aligned}
& \left\|x^{k+1}-x^*\right\| \\
\leqslant & \left\|\nabla^2 f\left(x^k\right)^{-1}\right\|\left\|\nabla^2 f\left(x^k\right)\left(x^k-x^*\right)-\left(\nabla f\left(x^k\right)-\nabla f\left(x^*\right)\right)\right\| \\
\leqslant & \left\|\nabla^2 f\left(x^k\right)^{-1}\right\| \cdot \frac{L}{2}\left\|x^k-x^*\right\|^2 \\
\leqslant & L\left\|\nabla^2 f\left(x^*\right)^{-1}\right\|\left\|x^k-x^*\right\|^2 .
\end{aligned} \tag{5}
$$

当初始点 x^0 满足 $\left\|x^0-x^*\right\| \leqslant \min \left\{\delta, r, \frac{1}{2L\left\|\nabla^2 f(x^*)^{-1}\right\|}\right\}$ 时,迭代点列一直处于邻域 $N_{\hat{\delta}}\left(x^*\right)$ 中,故 $\left\{x^k\right\}$ Q-二次收敛到 x^*.

另一方面,由牛顿方程(1)可知

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

这证明梯度的范数Q-二次收敛到0.

收敛速度分析

牛顿法收敛速度非常快, 但实际使用中会存在若干限制因素.

- 初始点x^0 需要距离最优解充分近(因为局部收敛性)
  应用时常以梯度类算法先求得较低精度的解, 后用牛顿法加速.
- $\nabla^2 f(x^*)$ 需正定, 半正定条件下可能退化到Q-线性收敛.
- $\nabla^2 f$ 的条件数较高时, 将对初值的选择作出较严苛的要求.

提纲

修正牛顿法

经典牛顿法的基本格式如下:

$$x^{k+1} = x^k - \nabla^2 f\left(x^k\right)^{-1} \nabla f\left(x^k\right).$$

除了计算、存储代价昂贵之外, 经典牛顿法还存在如下问题:

- 海瑟矩阵可能非正定,导致牛顿方向其实并非下降方向;
- 初始点离最优值点较远时候迭代不稳定(步长固定), 因而算法可能不收敛.

为提高算法的稳定性, 从以上两方面考虑, 应该:

- 对$\nabla^2 f(x)$进行修正,使其正定(所有特征值大于0);
- 用线搜索确定步长来增加算法的稳定性(Wolfe, Goldstein, Armijo).

综上考虑, 我们提出下面带线搜索的牛顿方法, 并称其为**修正牛顿方法**.

带线搜索的修正牛顿法

Algorithm 1 带线搜索的修正牛顿法

给定初始点x^0.
for $k = 0, 1, 2, \cdots$ **do**
  确定矩阵E^k 使得矩阵$B^k \stackrel{\text{def}}{=} \nabla^2 f\left(x^k\right) + E^k$ 正定且条件数较小.
  求解修正的牛顿方程$B^k d^k = -\nabla f\left(x^k\right)$ 得方向d^k.
  使用任意一种线搜索准则确定步长α_k.
  更新$x^{k+1} = x^k + \alpha_k d^k$.
end for

在上述算法中, B^k (即E^k) 的选取较为关键, 需要注意:

- B^k 应具有较低的条件数(原因见收敛性定理);
- 对$\nabla^2 f\left(x\right)$ 的改动较小, 以保存二阶信息;
- B^k 本身的计算代价不应太高.

修正牛顿法的全局收敛性

定理

修正牛顿法全局收敛性定理　令 f 在开集$\mathcal{D}$ 上二阶连续可微,且初始点x^0 满足$\{x \in \mathcal{D} : f(x) \leq f(x_0)\}$ 为紧集.若算法1中B^k 的条件数上界存在,即

$$\kappa(B_k) = \|B_k\| \left\|B_k^{-1}\right\| \leq C, \quad \exists C > 0, \forall k = 0, 1, 2, \cdots \tag{6}$$

则

$$\lim_{k \to \infty} \nabla f(x_k) = 0,$$

即算法具有全局收敛性.

上述定理的证明冗长,读者可具体参考: Newton's method, in Studies in Numerical Analysis, vol. 24 of MAA Studies in Mathematics, The Mathematical Association of America, 1984, pp. 29–82.

修正矩阵E^k的显式选取

进一步地，在修正步长后，为还使 B^k 正定，我们修正海瑟矩阵的特征值。首先做特征分解

$$\nabla^2 f(x_k) = Q\Lambda Q^T,$$

其中,Q 为正交矩阵,Λ 为对角矩阵.

取$E^k = \tau_k I$,即单位阵的常数倍，代入上式，则有

$$B^k = Q(\Lambda + \tau_k I) Q^T.$$

当正数τ_k 足够大的时候，可保证 B^k 正定，但 d^k 会接近负梯度方向，退化为一阶方法.
一种简单的技巧是对τ_k做截断，即

$$\tau_k = \max\left\{0, \delta - \lambda_{\min}(\nabla^2 f(x^k))\right\}, \quad (给定\delta > 0)$$

然而，此时$\lambda_{\min}(\nabla^2 f(x^k))$ 通常难以计算.我们通常使用Cholesky分解试探性取τ_k，从而避免这个问题.

下面利用Cholesky分解给出另一种算法 (a_{ii} 表示$\nabla^2 f\left(x^k\right)$ 的对角元素)

Algorithm 2 Cholesky分解(增加数量矩阵)

选取β, σ. (如$\beta = 10^{-3}, \sigma = 2$)
if $\min_i \{a_{ii}\} > 0$ **then**
  $\tau_0 = 0$
else
  $\tau_0 = -\min_i \{a_{ii}\} + \beta$
end if
for $t = 0, 1, 2, \cdots$ **do**
  尝试用Cholesky算法计算:$LL^T = \nabla^2 f\left(x^k\right) + \tau_t I$
  if Cholesky算法成功运行 **then**
    终止循环并返回L
  else
    $\tau_{t+1} = \max\{\sigma\tau_t, \beta\}$
  end if
end for

缺陷:可能需要多次的试验,而每一步都对 $\nabla^2 f\left(x^k\right) + \tau_t I$ 做分解的计算代价大(可考虑较大的σ,如$\sigma = 10$)

修正矩阵E^k的隐式选取

直接对$\nabla^2 f\left(x^k\right)$进行Cholesky分解可能会失败，因此考虑修正分解算法．回顾Cholesky 分解的定义：

$$A = LDL^{\mathrm{T}},$$

其中，$A=(a_{ij})$对称正定，$L=(l_{ij})$是对角线元素均为1 的下三角矩阵，$D=\mathrm{Diag}\left(d_1, d_2, \cdots, d_n\right)$ 是对角矩阵且对角线元素均为正．

Algorithm 3 标准Cholesky分解

1: 给定对称矩阵$A=(a_{ij})$.
2: **for** $j=1,2,\cdots,n$ **do**
3: $\quad c_{jj}=a_{jj}-\sum_{s=1}^{j-1} d_s l_{js}^2$
4: $\quad d_j=c_{jj}$
5: $\quad$**for** $i=j+1,j+2,\cdots,n$ **do**
6: $\quad\quad c_{ij}=a_{ij}-\sum_{s=1}^{j-1} d_s l_{is} l_{js}$
7: $\quad\quad l_{ij}=c_{ij}/d_j$.
8: $\quad$**end for**
9: **end for**

修正Cholesky分解

修正的目标:使 d_i 为正数,且 D 和 L 中元素值不太大.
具体地, 选取正的参数δ, β, 并要求在算法(3)中第j个外循环满足:

$$d_j \geq \delta, \quad \left| l_{ij}\sqrt{d_j} \right| \leq \beta, \quad i = j+1, j+2, \ldots, n \tag{7}$$

考虑下面的更新方式:

$$d_j = \max\left\{ |c_{jj}|, \left(\frac{\theta_j}{\beta}\right)^2, \delta \right\}, \quad \theta_j = \max_{i>j} |c_{ij}|$$

注意到算法(3)中有$c_{ij} = l_{ij}d_j$, 根据下式可知条件(7)成立:

$$\left| l_{ij}\sqrt{d_j} \right| = \frac{|c_{ij}|}{\sqrt{d_j}} \leq \frac{|c_{ij}|\,\beta}{\theta_j} \leq \beta, \quad \forall i > j$$

- θ_j 可以在之前d_j 计算, 因为c_{ij} 在算法(3)内循环的计算不包含d_j.
- 修正算法其实是在计算$\nabla^2 f\left(x^k\right) + E^k$ 的Cholesky 分解 (E^k 为对角元素非负的对角阵). 当$\nabla^2 f\left(x^k\right)$ 正定且条件数足够小时有$E^k = 0$.
- 可证明, 由修正算法得到的B_k 的条件数存在上界, 即满足(6)

提纲

应用在大规模参数的场合时,牛顿法有以下问题:

- 参数规模大, 可能数以亿计, 直接使用牛顿法则存储和计算都存在困难;
- 无法存储海瑟矩阵, 同时对海瑟矩阵做Cholesky分解的代价过高.

为解决这些问题, 我们引入非精确牛顿法, 即通过解牛顿方程的形式以求出牛顿方向(非精确).

性质

非精确牛顿法的优势　此方法有优良的收敛性质和速度, 且能在海瑟矩阵不定的时候找到有效的下降方向, 甚至可以不显式计算、存储海瑟矩阵 *(Hessian-free)*.

下面介绍非精确牛顿法的基本想法.

首先回到牛顿方程: $\nabla^2 f\left(x^k\right) d^k=-\nabla f\left(x^k\right)$. 非精确牛顿法不精确求解牛顿方向$d^k$, 而是用迭代法(如共轭梯度法)求解线性方程组. 这两者的区别是, 使用迭代法求解时, 我们可以在一定的精度下提前停机, 以提高求解效率.

为控制精度, 我们引入向量r^k 来表示残差, 将上述方程记为

$$\nabla^2 f\left(x^k\right) d^k=-\nabla f\left(x^k\right)+r^k, \tag{8}$$

因此终止条件可设置为

$$\left\|r^k\right\| \leqslant \eta_k\left\|\nabla f\left(x^k\right)\right\|, \tag{9}$$

其中序列$\{\eta_k\}$的值可自由设置, 不同的设置将导致不同的精度要求, 从而使算法有不同的收敛速度.

我们叙述非精确牛顿法的局部收敛定理.

定理

非精确牛顿法的收敛定理 设函数$f(x)$ 二阶连续可微, 且$\nabla^2 f(x^*)$ 正定, 则在非精确牛顿法中,

(1) 若$\exists t < 1$,使得η_k 满足$0 < \eta_k < t, k = 1, 2, \cdots$, 且起始点$x_0$ 充分靠近x^* 并迭代最终收敛到x^*, 则梯度$\nabla f(x^k)$以*Q*-线性收敛速度收敛;

(2) 若$\lim_{k\to\infty} \eta_k = 0$成立, 则梯度$\nabla f(x^k)$以*Q*-超线性收敛速度收敛;

(3) 若*(1)*或*(2)*成立, 且$\nabla^2 f$在x^*附近*Lip.*连续, $\eta_k = \mathcal{O}\left(\left\|\nabla f(x^k)\right\|\right)$, 则梯度$\nabla f(x^k)$以*Q*-二次收敛速度收敛.

定理证明

我们只给出一个不严格的证明，请读者主要体会证明思路.

注意到$\nabla^2 f(x)$在x^*处正定，在x^*附近连续，故存在正常数L，使得

$$\left\| \left(\nabla^2 f(x^k) \right)^{-1} \right\| \leq L, \quad (\forall x^k \text{同} x^* \text{足够接近})$$

代入(8)，得(第二个不等式用到了$\eta_k < 1$)

$$\left\| d^k \right\| \leq L \left(\left\| \nabla f(x^k) \right\| + \left\| r^k \right\| \right) \leq 2L \left\| \nabla f(x^k) \right\|.$$

利用Taylor展式和$\nabla^2 f$的连续性，得到

$$\begin{aligned} \nabla f(x^{k+1}) &= \nabla f(x^k) + \nabla^2 f(x^k) d^k + \int_0^1 \left[\nabla^2 f\left(x^k + t d^k\right) - \nabla^2 f\left(x^k\right) \right] d^k dt \\ &= \nabla f(x^k) + \nabla^2 f(x^k) d^k + o\left(\left\| d^k \right\| \right) \\ &= \nabla f(x^k) - \left(\nabla f(x^k) - r^k \right) + o\left(\left\| \nabla f(x^k) \right\| \right) \\ &= r^k + o\left(\left\| \nabla f(x^k) \right\| \right). \end{aligned}$$

定理证明

上式中两边取范数, 结合精度控制式(9), 得到

$$\left\|\nabla f(x^{k+1})\right\| \leq \eta_k \left\|\nabla f(x^k)\right\| + o\left(\left\|\nabla f(x^k)\right\|\right) \leq (\eta_k + o(1)) \left\|\nabla f(x^k)\right\|.$$

当x^k足够接近x^*时,$o(1)$项可被$(1-t)/2$控制,则

$$\left\|\nabla f(x^{k+1})\right\| \leq (\eta_k + (1-t)/2) \left\|\nabla f(x^k)\right\| \leq \frac{1+t}{2} \left\|\nabla f(x^k)\right\|.$$

由于$t < 1$,故梯度梯度$\nabla f(x^k)$以Q-线性收敛速度收敛.

推论

从以上证明过程可以看出如下的结果:

$$\frac{\|\nabla f(x^{k+1})\|}{\|\nabla f(x^k)\|} \leq \eta_k + o(1).$$

若$\lim_{k\to\infty} \eta_k = 0$成立,可有Q-超线性收敛的结论

$$\lim_{k\to\infty} \frac{\|\nabla f(x^{k+1})\|}{\|\nabla f(x^k)\|} = 0.$$

若$\nabla^2 f$ 在x^* 附近Lip.连续, 令$\eta_k = \mathcal{O}\left(\|\nabla f(x^k)\|\right)$, 可有二次收敛的结论

$$\|\nabla f(x^{k+1})\| = \mathcal{O}\left(\|\nabla f(x^k)\|^2\right).$$

因此, 在实际应用时:

- 取$\eta_k = \min\left(0.5, \sqrt{\|\nabla f(x^k)\|}\right)$, 可成立局部的超线性收敛;
- 取$\eta_k = \min\left(0.5, \|\nabla f(x^k)\|\right)$, 可成立局部的二次收敛.

逻辑回归模型

考虑二分类的逻辑回归模型

$$\min_{x} \quad \ell(x) \stackrel{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m} \ln\left(1+\exp\left(-b_i a_i^{\mathrm{T}} x\right)\right) + \lambda\|x\|_2^2.$$

为使用牛顿法,需要计算目标函数的梯度与海瑟矩阵:

$$\begin{aligned}\nabla \ell(x) &= \frac{1}{m}\sum_{i=1}^{m} \frac{1}{1+\exp\left(-b_i a_i^{\mathrm{T}} x\right)} \cdot \exp\left(-b_i a_i^{\mathrm{T}} x\right) \cdot \left(-b_i a_i\right) + 2\lambda x \\ &= -\frac{1}{m}\sum_{i=1}^{m} \left(1-p_i(x)\right) b_i a_i + 2\lambda x,\end{aligned}$$

其中$p_i(x) = \frac{1}{1+\exp\left(-b_i a_i^{\mathrm{T}} x\right)}$.

进一步对$\nabla \ell(x)$求导，成立

$$\begin{aligned}\nabla^2 \ell(x) &= \frac{1}{m}\sum_{i=1}^{m} b_i \cdot \nabla p_i(x) a_i^{\mathrm{T}} + 2\lambda I \\ &= \frac{1}{m}\sum_{i=1}^{m} b_i \frac{-1}{\left(1+\exp\left(-b_i a_i^{\mathrm{T}} x\right)\right)^2} \cdot \exp\left(-b_i a_i^{\mathrm{T}} x\right) \cdot \left(-b_i a_i a_i^{\mathrm{T}}\right) + 2\lambda I \\ &= \frac{1}{m}\sum_{i=1}^{m} (1-p_i(x))\, p_i(x) a_i a_i^{\mathrm{T}} + 2\lambda I \quad (b_i^2 = 1).\end{aligned}$$

引入矩阵$A=[a_1,a_2,\cdots,a_m]^{\mathrm{T}}\in\mathbb{R}^{m\times n}$, 向量$b=(b_1,b_2,\cdots,b_m)^{\mathrm{T}}$, 以及

$$p(x)=(p_1(x),p_2(x),\cdots,p_m(x))^{\mathrm{T}},$$

则可重写梯度和海瑟矩阵为

$$\nabla\ell(x)=-\frac{1}{m}A^{\mathrm{T}}(b-b\odot p(x))+2\lambda x,$$
$$\nabla^2\ell(x)=\frac{1}{m}A^{\mathrm{T}}W(x)A+2\lambda I,$$

其中$W(x)$为由$\{p_i(x)(1-p_i(x))\}_{i=1}^m$生成的对角矩阵.

则最终牛顿法迭代格式可以写作:

$$x^{k+1}=x^k+\left(\frac{1}{m}A^{\mathrm{T}}W\left(x^k\right)A+2\lambda I\right)^{-1}\left(\frac{1}{m}A^{\mathrm{T}}\left(b-b\odot p\left(x^k\right)\right)-2\lambda x^k\right).$$

我们得到了牛顿法的迭代格式, 因此可以调用牛顿法直接求解逻辑回归问题. 正如我们对牛顿方程处理思路的不同, 若变量规模不大, 则可尝试利用正定矩阵的Cholesky分解求解牛顿方程; 若变量规模较大, 则可以使用共轭梯度法对方程进行不精确的求解.

我们使用LIBSVM网站的数据集(具体数据集见下表), 对不同的数据集均调用非精确CG-牛顿法求解, 设置精度条件为

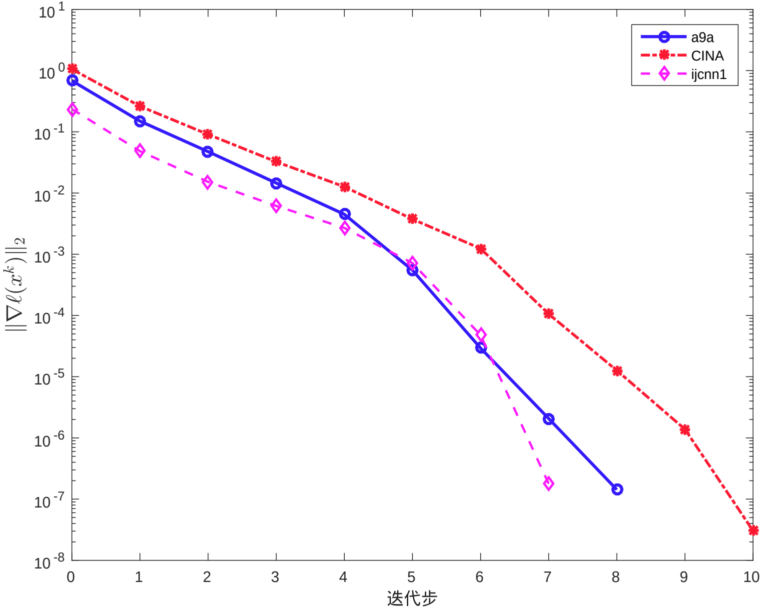
$$\left\|\nabla^2 \ell\left(x^k\right) d^k+\nabla \ell\left(x^k\right)\right\|_2 \leqslant \min \left\{\left\|\nabla \ell\left(x^k\right)\right\|_2^2, 0.1\left\|\nabla \ell\left(x^k\right)\right\|_2\right\},$$

Table: LIBSVM数据集(部分)

名称 \ 维数	m	n
a9a	16281	122
ijcnn1	91701	22
CINA	3206	132

在数据集中进行算法测试, 数值结果如下图所示. 从图中可以看到, 精确解附近梯度范数具有Q-超线性收敛性.

数值结果



a9a
CINA
ijcnn1
$\|\nabla \ell(x^k)\|_2$
迭代步
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
0
1
2
3
4
5
6
7
8
9
10

参考文献

[1] Nocedal J, Wright S. Numerical optimization[M]. Springer Science and Business Media, 2006.

[2] Hu J, Milzarek A, Wen Z, et al. Adaptive quadratically regularized Newton method for Riemannian optimization[J]. SIAM Journal on Matrix Analysis and Applications, 2018, 39(3): 1181-1207.