牛顿法、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:本教案由陈乐恒、丁思哲、邓展望协助准备

提纲

经典牛顿法

- 对于可微二次函数$f(x)$, 考虑目标函数f在点x_k的二阶泰勒近似

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

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, 可构造迭代格式

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

当步长$\alpha_k=1$时迭代格式(2)被称为**经典牛顿法**.

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^*;
- $\left\{\left\|\nabla f(x^k)\right\|\right\}$ Q-二次收敛到 0.

定理证明

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

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

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, 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 $\left\|x^0-x^*\right\| \leqslant \min\left\{\delta, r, \frac{1}{2L\left\|\nabla^2 f\left(x^*\right)^{-1}\right\|}\right\}$ 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.

非精确牛顿法

当变量维数很大时，牛顿法可能有如下困难:

- 海瑟矩阵$\nabla^2 f(x)$ 本身的计算、存储存在困难;
- 对$\nabla^2 f(x)$求逆或者做Cholesky分解的代价很高.

非精确牛顿法

- 使用迭代法(如共轭梯度法)求解牛顿方程，在一定的精度下提前停机, 以提高求解效率.
- 引入向量r^k 来表示残差, 将上述方程记为

$$\nabla^2 f\left(x^k\right) d^k = -\nabla f\left(x^k\right) + r^k. \tag{6}$$

因此终止条件可设置为

$$\left\| r^k \right\| \leqslant \eta_k \left\| \nabla f\left(x^k\right) \right\|. \tag{7}$$

- 不同的$\{\eta_k\}$将导致不同的精度要求, 使算法有不同的收敛速度.

收敛性分析

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, $\eta_k = O\left(\left\|\nabla f(x^k)\right\|\right)$, 则梯度$\nabla f(x^k)$以*Q*-二次收敛速度收敛.

定理证明

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(6), 得(第二个不等式用到了$\eta_k < 1$)

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

利用泰勒展式和$\nabla^2 f$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(7), 得到

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

当x^k足够接近x^*时, $o(1)$项可被$(1-t)/2$控制, 则

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

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

提纲

信赖域算法框架

1. 在当前迭代点x^k建立局部模型

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

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

信赖域子问题

- 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，利用$f(x)$的二阶近似来刻画$f(x)$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,需对上述模型添加信赖域约束：

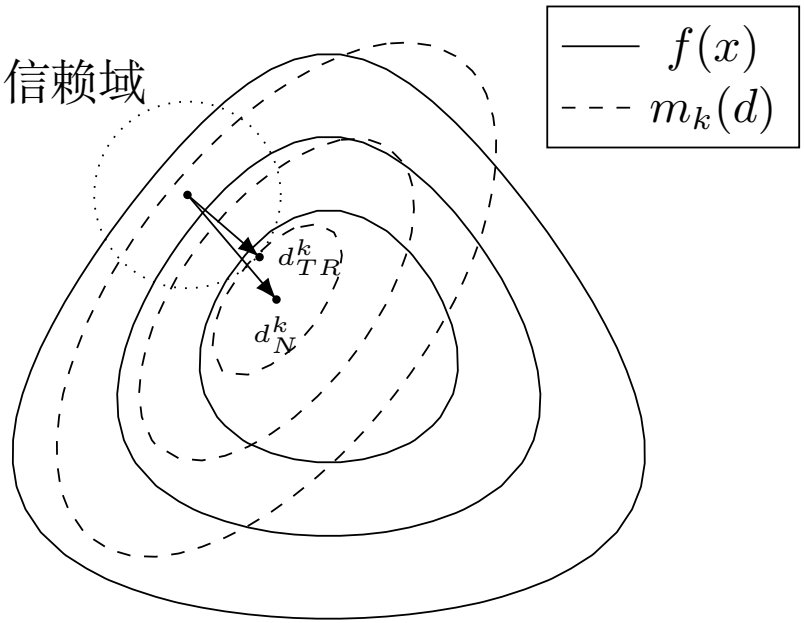
$$\Omega_k=\{x^k+d|\ ||d||\leqslant\Delta_k\},$$

其中$\Delta_k>0$是信赖域半径。

信赖域子问题

- 因此信赖域算法每一步都需要求解如下子问题:

$$\min_{d\in R^n} \quad m_k(d), \quad \text{s.t.} \quad \|d\| \le \Delta_k. \tag{8}$$



- 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：

$$\rho_k = \frac{f\left(x^k\right) - f\left(x^k + d^k\right)}{m_k\left(0\right) - m_k\left(d^k\right)} \tag{9}$$

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

信赖域子问题最优性条件

定理 (信赖域子问题最优性条件)

d^*是信赖域子问题

$$\min \quad m(d) = f + g^{\mathrm{T}} d + \frac{1}{2} d^{\mathrm{T}} B d, \quad \text{s.t.} \quad \|d\| \leq \Delta \tag{10}$$

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

$$(B + \lambda I) d^* = -g, \tag{11a}$$

$$\lambda(\Delta - \|d^*\|) = 0, \tag{11b}$$

$$B + \lambda I \succeq 0. \tag{11c}$$

信赖域子问题最优性条件

必要性：问题(10)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(11a) 式和(11b) 式.
- 为了证明(11c)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(11a)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.

信赖域子问题最优性条件

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(11c) 可知$\hat{m}(d)$关于d是凸函数. 根据条件(11a)，d^*满足凸函数一阶最优性条件，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(11b) 可知$\lambda(\Delta^2 - \|d^*\|^2) = 0$，代入上式消去$\|d^*\|^2$得

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

求解信赖域子问题迭代法

- 上述定理提供了一个寻找d的方法:
  - $\lambda = 0$，测试是否$B \succeq 0$，$d = -B^{-1}g$ 且$\|d\| \leq \Delta$ 满足.
  - 选择充分大的λ使得$B + \lambda I \succeq 0$ 且$\|d(\lambda)\| = \Delta$，其中

$$d(\lambda) = -(B + \lambda I)^{-1} g. \tag{12}$$

关键：求解关于λ的方程$\|d(\lambda)\| = \Delta$ 或者$1/\|d(\lambda)\| = 1/\Delta$.

- 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，$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$是$B$的特征值.

$$d(\lambda) = -Q(\Lambda + \lambda I)^{-1} Q^{\mathrm{T}} g = -\sum_{i=1}^{n} \frac{q_i^{\mathrm{T}} g}{\lambda_i + \lambda} q_i. \tag{13}$$

这正是$d(\lambda)$的正交分解，由正交性可容易求出

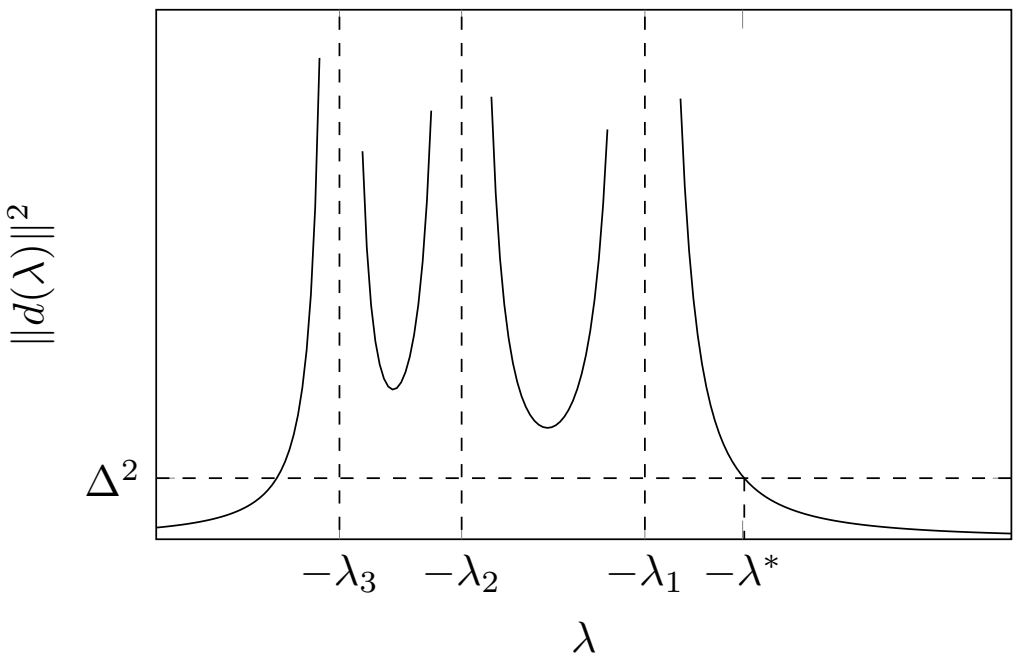
$$\|d(\lambda)\|^2 = \sum_{i=1}^{n} \frac{(q_i^{\mathrm{T}} g)^2}{(\lambda_i + \lambda)^2}. \tag{14}$$

求解信赖域子问题迭代法

- 由(14)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.



求解信赖域子问题牛顿法

求解$\phi(\lambda) = 1/\Delta - 1/\|d(\lambda)\| = 0$ 且$B + \lambda I \succeq 0$.

- $\phi'(\lambda) = \|d(\lambda)\|^{-3} d(\lambda)^\top d'(\lambda)$. 对(11a) 关于$\lambda$ 求导数：

$$d(\lambda) + (B + \lambda I)d'(\lambda) = 0 \Longrightarrow d'(\lambda) = -(B + \lambda I)^{-1} d(\lambda).$$

- 假设$B + \lambda I = R^\top R$。因此

$$\begin{aligned} d(\lambda)^\top d'(\lambda) &= -d(\lambda)^\top (B + \lambda I)^{-1} d(\lambda) = -d(\lambda)^\top R^{-1} R^{-\top} d(\lambda) \\ &= -\|R^{-\top} d(\lambda)\|_2^2. \end{aligned}$$

- 牛顿法：

$$\begin{aligned} \lambda^{(l+1)} &= \lambda^{(l)} - \frac{\phi(\lambda^{(l)})}{\phi'(\lambda^{(l)})} \\ &= \lambda^{(l)} + \frac{\|d(\lambda)\|_2^2}{\|R^{-\top} d(\lambda)\|_2^2} \frac{\|d(\lambda)\|_2 - \Delta}{\Delta} \end{aligned}$$

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.

共轭梯度法介绍

- 标准共轭梯度法求解二次极小化问题

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

- 给定初值$s^0 = 0, r^0 = g, p^0 = -g$，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

- 截断共轭梯度法则是标准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.
- 矩阵B不一定正定，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. 遇到这种情况立即终止算法. 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(8)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

$$\min_s \quad q(s) \overset{\text{def}}{=\!=} g^{\mathrm{T}}s + \frac{1}{2}s^{\mathrm{T}}Bs, \ \text{s.t.} \ \|s\| \le \Delta.$$

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** $\|p^0\| \le \varepsilon$ **then**
3: 算法停止，输出$s = 0$.
4: **end if**
5: **LOOP**
6: **if** $(p^k)^{\mathrm{T}}Bp^k \le 0$ **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$\alpha_k = \frac{\|r^k\|^2}{(p^k)^{\mathrm{T}}Bp^k}$，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**

提纲

柯西点定义

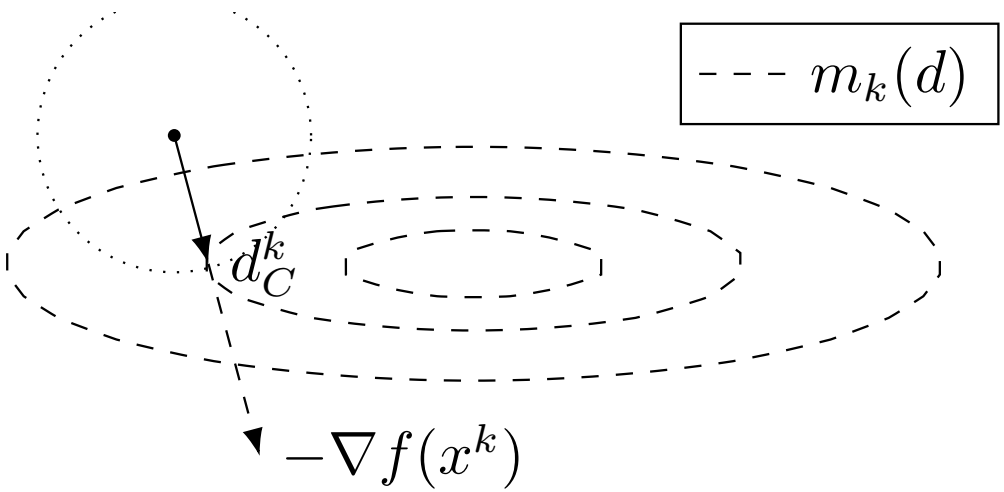
定义 (柯西点)

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$d_C^k = -\tau_k \nabla f(x^k)$.

柯西点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

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

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

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

$$m_k(0) - m_k(d_C^k) \geq c_1 \|g^k\| \min\left\{ \Delta_k, \frac{\|g^k\|}{\|B^k\|_2} \right\}. \tag{15}$$

其中$c_1 = \frac{1}{2}$

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(15)式成立.

柯西点性质

- 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(15)式成立.

- 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(15)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$\|B^k\|_2 \leq \beta, \forall\, k$，$f(x)$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(15)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

首先由ρ_k的定义可以得到：

$$\begin{aligned}|\rho_k - 1| &= \left| \frac{\left(f\left(x^k\right) - f\left(x^k + d^k\right)\right) - \left(m_k\left(0\right) - m_k\left(d^k\right)\right)}{m_k\left(0\right) - m_k\left(d^k\right)} \right| \\ &= \left| \frac{m_k\left(d^k\right) - f\left(x^k + d^k\right)}{m_k\left(0\right) - m_k\left(d^k\right)} \right|\end{aligned} \tag{16}$$

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

$$
\begin{aligned}
\left|m_k\left(d^k\right)-f\left(x^k+d^k\right)\right| &= \left|\frac{1}{2}d^{kT}B^k d^k-\int_0^1\left[g\left(x^k+td^k\right)-g\left(x^k\right)\right]^T d^k dt\right| \\
&\leqslant (\beta/2)\left\|d^k\right\|^2+\beta_1\left\|d^k\right\|^2,
\end{aligned}
\tag{17}
$$

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(15)式可知

$$
m_k(0)-m_k(d^k) \geq c_1\left\|g^k\right\| \min\left\{\Delta_k, \frac{\left\|g^k\right\|}{\left\|B^k\right\|_2}\right\} \geqslant c_1\varepsilon \min\left\{\Delta_k, \frac{\varepsilon}{\beta}\right\} \tag{18}
$$

全局收敛性1证明

由(17)和(18)可知

$$|\rho_k - 1| \leqslant \frac{\Delta_k^2 (\beta/2 + \beta_1)}{c_1 \epsilon \min(\Delta_k, \epsilon/\beta)}. \tag{19}$$

定义

$$\overline{\Delta} = \min\left((1 - 2\bar{\rho}_1)\frac{c_1\varepsilon}{(\beta/2 + \beta_1)}, R_0\right).$$

对于所有满足$\Delta_k \leqslant \overline{\Delta}$的充分大$k$，下面给出$|\rho_k - 1|$的上界。注意到$c_1 \leqslant 1$，$\min(\Delta_k, \varepsilon/\beta) = \Delta_k$. 由此可得:

$$|\rho_k - 1| \leqslant \frac{\Delta_k^2 (\beta/2 + \beta_1)}{c_1 \varepsilon \Delta_k} \leqslant \frac{\overline{\Delta} (\beta/2 + \beta_1)}{c_1 \varepsilon} \leqslant 1 - 2\bar{\rho}_1.$$

因此$\rho_k > \bar{\rho}_1$. 再由算法1可知$\Delta_k \leqslant \overline{\Delta}$时$\Delta_{k+1} \geq \Delta_k$成立.

全局收敛性1证明

因此如果$k-1$失败且$\Delta_k \leq \bar{\Delta}$，则一定$\Delta_{k-1} \geq \bar{\Delta}$. 所以我们有

$$\Delta_k \geqslant \min(\Delta_K, \gamma_1 \overline{\Delta}), \quad \forall k \geqslant K. \tag{20}$$

假设有含有无穷项的子列$\mathcal{K}$指标集,使得$\rho_k \geqslant \bar{\rho}_1, k \in \mathcal{K}$. 则对于$k \in \mathcal{K}$,并且$k \geqslant K$,由(18)可知

$$f\left(x^k\right) - f\left(x^{k+1}\right) \geqslant \bar{\rho}_1 \left[m_k\left(0\right) - m_k\left(d^k\right)\right] \quad \geqslant \bar{\rho}_1 c_1 \varepsilon \min\left(\Delta_k, \varepsilon/\beta\right). \tag{21}$$

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(20)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$\lim\limits_{k \to \infty} \Delta_k = 0$,与(20)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(15) 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(18)可得：

$$
\begin{aligned}
f\left(x^m\right)-f\left(x^{l+1}\right) &= \sum_{k=m}^{l} f\left(x^k\right)-f\left(x^{k+1}\right) \\
&\geqslant \sum_{k=m,x^k\neq x^{k+1}}^{l} \eta\left[m_k\left(0\right)-m_k\left(d^k\right)\right] \qquad (22) \\
&\geqslant \sum_{k=m,x^k\neq x^{k+1}}^{l} \eta c_1 \min\left(\Delta_k,\frac{\epsilon}{\beta}\right)
\end{aligned}
$$

全局收敛性2证明

Proof.

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

$$f\left(x^{m}\right)-f\left(x^{l+1}\right) \geqslant \sum_{k=m, x^{k} \neq x^{k+1}}^{l} \eta c_{1} \epsilon \Delta_{k} \geqslant \eta c_{1} \epsilon R=\eta c_{1} \epsilon \min \left(\frac{\epsilon}{\beta_{1}}, R_{0}\right) \tag{23}$$

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

$$f\left(x^{m}\right)-f\left(x^{l+1}\right) \geqslant \eta c_{1} \epsilon \frac{\epsilon}{\beta} \tag{24}$$

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(23)与(24)式，则有

$$\begin{aligned} f\left(x^{m}\right)-f^{*} & \geqslant f\left(x^{m}\right)-f\left(x^{l+1}\right) \\ & \geqslant \eta c_{1} \varepsilon \min \left(\frac{\varepsilon}{\beta}, \frac{\varepsilon}{\beta_{1}}, R_{0}\right) \\ & =\frac{1}{2} \eta c_{1}\left\|g^{m}\right\| \min \left(\frac{\left\|g^{m}\right\|}{2 \beta}, \frac{\left\|g^{m}\right\|}{2 \beta_{1}}, R_{0}\right)>0 \end{aligned} \tag{25}$$

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

提纲

应用举例

- 考虑逻辑回归问题

$$\min_x \quad \frac{1}{m}\sum_{i=1}^{m}\ln(1+\exp(-b_i a_i^{\mathrm{T}}x))+\lambda\|x\|_2^2, \tag{26}$$

  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(26)，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．