§3.3 牛顿法

天津大学数学学院

2020年3月2日

本节假设目标函数 f 二阶连续可微.

牛顿法的基本思想：利用目标函数 f 在迭代点处的二阶Taylor展开式的极小值点序列去逼近目标函数 f 的极小值点.

§3.3.1 经典牛顿法

设x^k是 f 的极小值点x^*处的一个近似点，将f在 x^k附近进行二阶Taylor展开，且令

$$f(x) \approx q_k(x) = f(x^k) + g_k{}^{\mathrm{T}}(x - x^k) + \frac{1}{2}(x - x^k)^{\mathrm{T}} G_k (x - x^k)$$

若G_k正定，则$q_k(x)$是正定二次函数，必有唯一极小值点，将$q_k(x)$的极小值点取做x^*的下一次近似点x^{k+1}. 由一阶必要条件，x^{k+1}应满足

$$G_k(x^{k+1} - x^k) + g_k = 0,$$

牛顿方向

令$d^k = x^{k+1} - x^k$,则$G_k d^k = -g_k$ ----------牛顿方程

解得 $d^k = -G_k{}^{-1} g_k$

从而 $x^{k+1} = x^k - G_k{}^{-1} g_k.$ ----------牛顿迭代公式

经典牛顿法算法框架

注：当G_k正定时，牛顿方向$d^k=-G_k^{-1}g_k$是x^k处的一个下降方向.

算法3.3.1 （经典牛顿法）

选择初始点$x^0\in\mathbb{R}^n$, 置$k:=0$.

步1　计算g_k. 若$g_k=0$，算法终止.

步2　计算G_k, 并由牛顿方程$G_k d^k=-g_k$解出d^k.

步3　置$x^{k+1}:=x^k+d^k, k:=k+1$, 转步1.

步长为1

经典牛顿法不一定是下降算法

经典牛顿法的二次终止性

定理3.3.1 给定对称正定矩阵$Q\in\mathbb{R}^{n\times n}$及向量$q\in\mathbb{R}^n$，设

$$f(x)=\frac{1}{2}x^TQx+q^Tx,$$

那么，从任意初始点x^0出发，算法3.3.1至多一步迭代就可达到函数 f 的极小值点.

证明 通过简单计算可得 $g(x)=Qx+q$ 且 $G(x)=Q$. 由于矩阵 Q 对称正定，所以函数 f 为严格凸函数. 因此，由定理 3.1.3 可得，函数 f 的全局极小值点为 $x^*=-Q^{-1}q$.

另外，若 x^0 不是 f 的极小值点，则由算法 3.3.1 可知

$$x^1=x^0-G_0^{-1}g_0=x^0-Q^{-1}(Qx^0+q)=-Q^{-1}q.$$

所以，x^1 为函数 f 的极小值点. 定理得证. □

定义3.3.1 当一个算法用于求解严格凸二次函数极小值问题时，如果从任意初始点出发，算法经过有限步迭代后可达到函数的极小值点，则称该算法具有二次终止性.

结论：经典牛顿法具有二次终止性.

经典牛顿法的局部收敛性和二阶收敛速度

定理3.3.2 (局部二阶收敛性)

假设目标函数 f 二阶连续可微，x^*是 f的一个局部极小值点，G_*正定，且 f的Hesse阵$G(\cdot)$具有Lipschitz连续性，即存在$L>0,$使得对任意的$x,y\in\mathbb{R}^n,$有

$$\|G(x)-G(y)\|\le L\|x-y\|.$$

当初始点x^0充分靠近x^*时，对于所有的$k\in\mathbb{N},$经典牛顿迭代$x^{k+1}=x^k-G_k^{-1}g_k$有意义，算法3.3.1产生的迭代序列$\{x^k\}$收敛于x^*，并且具有二阶收敛性.

$$\lim_{k\to\infty}\frac{\|x^{k+1}-x^*\|}{\|x^k-x^*\|^2}=\xi>0.$$

证明 由于函数 f 二阶连续可微且其 Hesse 阵 $G(\cdot)$ 具有 Lipschitz 连续性，所以利用 Taylor 公式有

$$g(x^k + d) = g_k + G_k d + O(\|d\|^2).$$

从而，

$$0 = g(x^*) = g(x^k - (x^k - x^*)) = g_k - G_k(x^k - x^*) + O(\|x^k - x^*\|^2).$$

又因为 f 二阶连续可微且 G_* 对称正定，所以当 x^k 充分靠近 x^* 时，G_k 对称正定且 $\{\|G_k^{-1}\|\}$ 有界. 由此，用 G_k^{-1} 乘上式两边可得

$$O(\|x^k - x^*\|^2) = -G_k^{-1} g_k + (x^k - x^*) = d^k + (x^k - x^*) = x^{k+1} - x^*.$$

这表明：存在常数 $c > 0$ 使得

$$\|x^{k+1} - x^*\| \leqslant c\|x^k - x^*\|^2. \tag{3.3.4}$$

若 x^k 充分靠近 x^* 使得 $x^k \in \mathcal{N}_\varepsilon(x^*) := \{x \mid \|x - x^*\| \leqslant \varepsilon\}$, 其中 ε 满足 $c\varepsilon < 1$, 那么由 (3.3.4) 式可得

$$\|x^{k+1} - x^*\| \leqslant c\|x^k - x^*\|^2 \leqslant c\varepsilon\|x^k - x^*\| < \|x^k - x^*\|.$$

因此，$x^{k+1} \in \mathcal{N}_\varepsilon(x^*)$.

由归纳法可知: 如果 $x^0 \in \mathcal{N}_\varepsilon(x^*)$，那么对所有的 k，经典牛顿法都有意义，且

$$\|x^k - x^*\| \leqslant (c\varepsilon)^k\|x^0 - x^*\| \to 0 \quad (k \to \infty).$$

因此，由 (3.3.4) 式可知：算法 3.3.1 具有二阶收敛性. 定理得证. □

经典牛顿法的优点和缺点

优点：（1）收敛速度快（若G_*正定，$G(x)$是Lipschitz连续的，只要初始点选取合适，算法是二阶收敛的）；

（2）具有二次终止性.

缺点：（1）要求目标函数二阶连续可微;

（2）且每步需计算G_k,并求解方程组$G_k d^k = -g_k$, 工作量大;

（3）仅是局部收敛的（要求初始点充分靠近最优解，否则算法可能不收敛）；

（4）收敛于鞍点或极大点的可能性并不小(当$g_k = 0$但G_k不正定）；

注 当初始点远离最优解时，G_k不一定正定，牛顿方向不一定是下降方向，经典牛顿法不一定收敛.

§3.3.2 带线搜索的牛顿法

算法3.3.2 （带线搜索的牛顿法）

选择初始点$x^0 \in \mathbb{R}^n$, 置$k := 0.$

步1　计算g_k. 若$g_k = 0$，算法终止.

步2　计算G_k, 并由牛顿方程$G_k d^k = -g_k$解出d^k.

步3 由线搜索计算步长λ_k.

步4　置$x^{k+1} := x^k + \lambda_k d^k, k := k+1,$ 转步1.

迭代公式：$x^{k+1} = x^k - \lambda_k G_k^{-1} g_k.$　　　　(3.3.5)

采用精确一维线搜索计算步长时，称为阻尼牛顿法.

带线搜索的牛顿法的全局收敛性

定理3.3.3 算法3.3.2具有二次终止性.

定理3.1.7 假设f：$\mathbb{R}^n \to \mathbb{R}$二阶连续可微, 无穷序列$\{x^k\}$由算法3.1.1产生，其中迭代步长$\lambda_k$由Armijo线搜索得到(其中$\sigma \in (0, \frac{1}{2})$)或由Wolfe线搜索得到. 假设$x^k \to x^*$，$g_* = 0$且$G_*$正定. 如果

$$\lim_{k\to\infty} \frac{\|g_k + G_k d^k\|}{\|d^k\|} = 0,$$

则 (1) 当k充分大时，$\lambda_k = 1$; (2) 序列$\{x^k\}$局部超线性收敛于x^*.

带线搜索的牛顿法的全局收敛性

定理3.3.4 （全局收敛性）

设目标函数 f 二阶连续可微且存在常数$c>0,$使得

$$d^T G(x)d \ge c\|d\|^2, \forall d \in \mathbb{R}^n, \forall x \in \mathcal{L}(f) := \{x : f(x) \le f(x^0)\}.$$

设迭代序列$\{x^k\}$由算法3.3.2产生，其中步长λ_k由精确一维线搜索得到，或由*Armijo*线搜索得到，或由*Wolfe*线搜索得到，则或者算法3.3.2有限终止于f在$\mathcal{L}(f)$中的唯一全局极小点，或者算法3.3.2产生的迭代序列$\{x^k\}$收敛于f 在$\mathcal{L}(f)$中的唯一全局极小点.

证明 由定理的条件及定理 1.3.6 可知, f 在 $\mathscr{L}(f)$ 上为强凸函数. 根据函数 f 的强凸性，易证：水平集 $\mathscr{L}(f)$ 是一个有界闭集，且函数 f 的全局极小值点存在且唯一. 由定理 3.1.3 可知, f 的唯一全局极小值点是 $g(x)=0$ 的唯一解.

假设算法 3.3.2 不是有限终止于 f 在 $\mathscr{L}(f)$ 中的唯一全局极小值点的. 由算法 3.3.2 可知: 数列 $\{f(x^k)\}$ 是单调下降数列. 因此，对所有的 k, 有 $x^k \in \mathscr{L}(f)$. 另外，利用 (3.3.6) 式, 验证可得：定理 3.1.4~ 定理 3.1.6 的条件满足，因此，由定理 3.1.4~ 定理 3.1.6 可得

$$\lim_{k\to\infty} \|g(x^k)\| = 0,$$

即 $\{x^k\}$ 的任何聚点都是函数 f 的稳定点. 定理得证. □

附录

定理 1.2.19 设 f 是非空凸集 $D\in\mathbf{R}^n$ 上的连续凸函数，则水平集 $L_\alpha=\{x\in D\mid f(x)\leqslant\alpha\}$是闭凸集.

证明 我们只要证明 L_α 是闭的即可. 设 x^* 是 L_α 的极限点，则存在序列 $\{x_k\}\subset L_\alpha$，$\{x_k\}\to x^*$. 于是，由 f 的连续性，

$$f(x^*)=\lim_{k\to\infty}f(x_k)\leqslant\alpha,$$

从而 $x^*\in L_\alpha$. 这表明 L_α 是闭集.

附录

定理 1.2.20 设 f 在非空凸集 $D\in\mathbf{R}^n$ 上二次连续可微，且存在常数 $m>0$，使得

$$u^{\mathrm{T}}\nabla^2 f(x)u\geqslant m\|u\|^2,\quad \forall x\in L_\alpha,\quad u\in\mathbf{R}^n,\tag{1.2.15}$$

则水平集 $L(x_0)=\{x\in D\mid f(x)\leqslant f(x_0)\}$ 是有界闭凸集.

证明 由上面两个定理知 $L(x_0)$ 是闭凸集. 现在证明 $L(x_0)$ 的有界性.

因为水平集 $L(x_0)$ 是凸的，由(1.2.15)有，$\forall x,y\in L(x_0)$，

$$m\|y-x\|^2\leqslant(y-x)^{\mathrm{T}}\nabla^2 f(x+\alpha(y-x))(y-x),\quad \alpha\in(0,1).$$

由于 f 在 D 上二次连续可微，故

$$\begin{aligned}f(y)&=f(x)+\nabla f(x)^{\mathrm{T}}(y-x)+\frac{1}{2}(y-x)^{\mathrm{T}}\nabla^2 f(\xi)(y-x)\\&\geqslant f(x)+\nabla f(x)^{\mathrm{T}}(y-x)+\frac{1}{2}m\|y-x\|^2,\end{aligned}$$

其中 $\xi=x+\alpha(y-x)\in L(x_0)$，$m$ 与 x,y 无关. 因此，对任意 $y\in L(x_0)$，$y\neq x_0$，

$$\begin{aligned}f(y)-f(x_0)&\geqslant\nabla f(x_0)^{\mathrm{T}}(y-x_0)+\frac{1}{2}m\|y-x_0\|^2\\&\geqslant-\|\nabla f(x_0)\|\,\|y-x_0\|+\frac{1}{2}m\|y-x_0\|^2.\end{aligned}\tag{1.2.16}$$

由于 $f(y)\leqslant f(x_0)$，故

$$\|y-x_0\|\leqslant\frac{2}{m}\|\nabla f(x_0)\|.$$

这表明水平集 $L(x_0)=\{x\in D\mid f(x)\leqslant f(x_0)\}$ 有界. □

定理3.3.5 （局部二阶收敛性）

设定理3.3.4的条件成立，且 f 的Hesse阵$G(x)$具有Lipschitz连续性. 设迭代序列$\{x^k\}$由算法3.3.2产生，其中步长λ_k由Armijo线搜索得到$(\sigma \in (0, \frac{1}{2}))$或由Wolfe线搜索得到. 假设$x^k \to x^*$且$g_* = 0$,以及$G_*$正定，那么迭代序列$\{x^k\}$二阶收敛到$x^*$.

证明　根据定理 3.3.4 可知, 序列 $\{x^k\}$ 收敛于 x^* 且 $g(x^*) = 0$. 又由牛顿方程 (3.3.3) 式, 可得 (3.1.4) 式成立. 因此，从定理 3.1.7 的结论可知, 当 k 充分大时，$\lambda_k = 1$ 满足线搜索条件. 所以，由定理 3.3.2 可得, 迭代序列 $\{x^k\}$ 二阶收敛于 x^*. 定理得证. □

例1 利用阻尼牛顿法求 $f(x_1,x_2)=x_1^2+2x_2^2-4x_1-2x_1x_2$ 的极小值，取$x^0=(1,1)^{\mathrm{T}}$.

解 $g(x)=[2x_1-4-2x_2,4x_2-2x_1]^{\mathrm{T}}, G(x)=\begin{bmatrix} 2 & -2 \\ -2 & 4 \end{bmatrix}, \quad G^{-1}=\begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}.$

$g_0=g(x^0)=[-4,2]^{\mathrm{T}}\neq \mathbf{0}$, 于是

$d^0=-G^{-1}g_0=[3,1]^{\mathrm{T}}.$

$\varphi(\lambda)=f(x^0+\lambda d^0)=f(1+3\lambda,1+\lambda)=5\lambda^2-10\lambda-3, \quad \varphi'(\lambda)=10\lambda-10.$

令$\varphi'(\lambda)=0$, 得$\lambda_0=1$, 所以

$x^1=x^0+\lambda_0 d^0=[1,1]^{\mathrm{T}}+1\cdot[3,1]^{\mathrm{T}}=[4,2]^{\mathrm{T}}.$

$g_1=g(x^1)=\mathbf{0}$, 即迭代一次就达到了$f(x)$的极小值点$x^*=[4,2]^{\mathrm{T}}$，对应的极小值为$f(x^*)=-8.$

谢谢观看！