3.3 牛顿法 — 经典牛顿法

对于无约束优化问题(3.0.1), 最速下降法本质上是使用线性函数去近似目标函数. 为了得到快速算法，可考虑目标函数的高阶逼近. 本节假设目标函数f二次连续可微, 并且利用目标函数f在迭代点处的二次Taylor展开式的极小值点序列去逼近目标函数的极小值点. 这种方法被称为牛顿(Newton)法.

§3.3.1 经典牛顿法

假设点x^k是函数f在极小值点x^*处的一个近似点，将f在x^k附近进行二次Taylor展开，且令

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

若G_k对称正定，则二次函数$q_k(\cdot)$有唯一极小值点，将它取为x^*的下一次近似点x^{k+1}.

3.3 牛顿法 — 经典牛顿法

由一阶必要条件可知, x^{k+1}应满足$G_k(x^{k+1}-x^k)+g_k=0$. 令

$$x^{k+1}=x^k+d^k, \tag{3.3.1}$$

其中, 方向d^k称为牛顿方向且满足

$$G_k d^k=-g_k. \tag{3.3.2}$$

称(3.3.2)式为牛顿方程，解之可得$d^k=-G_k^{-1}g_k$, 并代入(3.3.1)式, 于是有

$$x^{k+1}=x^k-G_k^{-1}g_k. \tag{3.3.3}$$

称(3.3.3)式为牛顿迭代公式. 容易证明：牛顿方向$d^k=-G_k^{-1}g_k$是目标函数f在椭球范数$\|\cdot\|_{G_k}$ (注：$\|x\|_{G_k}=\sqrt{x^\top G_k x}$, $\forall x\in\mathbb{R}^n$)下的最速下降方向，即易证：$d^k=-G_k^{-1}g_k$是极小化问题

$$\min_{d\in\mathbb{R}^n\setminus\{0\}}\frac{g_k^\top d}{\|d\|_{G_k}}$$

的最优解.

3.3 牛顿法 — 经典牛顿法

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. 如果$g_k = 0$, 算法终止.

步*2* 计算G_k, 并由*(3.3.2)*式求解出d^k.

步*3* 置$x^{k+1} := x^k + d^k$和$k := k + 1$, 转步*1*.

3.3 牛顿法 — 经典牛顿法

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

$$f(x) = \frac{1}{2}x^\top Qx + q^\top x.$$

那么，从任意初始点x^0出发，算法3.3.1至多一步迭代可达到函数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，因而函数f为严格凸函数. 所以，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：

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

因此，x^1为函数f的极小值点. 定理得证. □

3.3 牛顿法 — 经典牛顿法

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，若从任意初始点出发，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.

由定理3.3.1可知：经典牛顿法具有二次终止性.

定理 3.3.2 假设目标函数f二次连续可微，x^*是f的一个局部极小值点，G_*对称正定，且函数f的*Hesse* 阵$G(\cdot)$ 具有*Lipschitz*连续性，即存在$\mathcal{L} > 0$使得对任意的$x, y \in \mathbb{R}^n$，有$\|G(x) - G(y)\| \leq \mathcal{L}\|x - y\|$. 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*(3.3.3)*式都有意义，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.

证明 由于函数f二次连续可微且其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，

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

3.3 牛顿法 — 经典牛顿法

又因为f二次连续可微且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

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

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

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

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

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

$$\|x^k - x^*\| \leq (c\epsilon)^k\|x^0 - x^*\| \to 0, \quad (k \to \infty).$$

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. □

§3.3.2　带线搜索的牛顿法

定理3.3.2表明：经典牛顿法3.3.1具有很快的收敛速度，但它只是局部收敛的，即只有当初始点x^0充分接近极小值点x^*时，才能保证其好的收敛性. 另外，该算法收敛于鞍点或极大值点的可能性与收敛于极小值点的可能性是一样的. 为了克服这些缺陷，人们考虑带线搜索的牛顿法，即牛顿迭代公式(3.3.3)式改变为

$$x^{k+1} = x^k - \lambda_k G_k^{-1} g_k, \tag{3.3.5}$$

其中步长λ_k由某个线搜索得到. 显然，当$\lambda_k = 1$时，(3.3.5)式即为(3.3.3)式. 具体算法如下：

3.3 牛顿法 — 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. 如果$g_k = 0$, 算法终止.

步*2*　计算G_k, 并由*(3.3.2)*式求解出d^k.

步*3*　由线搜索计算步长λ_k.

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

类似于定理3.3.1的证明，易得如下结论.

定理 3.3.3　使用精确一维线搜索的牛顿法具有二次终止性.

3.3 牛顿法 — 带线搜索的牛顿法

定理 3.3.4 设目标函数f二次连续可微且存在常数$c > 0$使得

$$d^\top G(x)d \geq c\|d\|^2, \quad \forall d \in \mathbb{R}^n, \quad \forall x \in \mathscr{L}(f) := \{x \mid f(x) \leq f(x^0)\}\tag{3.3.6}$$

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. 那么，或者算法3.3.2有限终止于f在$\mathscr{L}(f)$中的唯一全局极小值点，或者迭代序列$\{x^k\}$收敛于f在$\mathscr{L}(f)$中的唯一全局极小值点.

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，易证可得：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3.1.5以及定理3.1.6的条件满足，因此，由定理3.1.4、定理3.1.5以及定理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. □

3.3 牛顿法 — 带线搜索的牛顿法

定理 3.3.5 设定理*3.3.4*的条件成立并且f的*Hesse*阵$G(x)$具有*Lipschitz*连续性. 令序列$\{x^k\}$由算法*3.3.2*产生, 其中步长λ_k由*Armijo*线搜索得到(其中$\sigma \in (0, 1/2)$)或由*Wolfe*线搜索得到. 假设$x^k \to x^*$, 且$g_* = 0$以及G_*对称正定. 那么, 迭代序列$\{x^k\}$二阶收敛于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. □

3.3 牛顿法 — 作业

3.11 假设函数$f:\mathbb{R}^n \to \mathbb{R}$二次连续可微，且对任意的$x \in \mathbb{R}^n$，函数$f$的Hesse阵$G(x)$正定. 试证明: 牛顿方向$d^k = -G_k^{-1} g_k$是极小化问题

$$\min_{d \in \mathbb{R}^n \setminus \{0\}} \frac{g_k^\top d}{\|d\|_{G_k}}$$

的最优解.