MATH 3011 运筹学 **秋季 2023**

Lecture 12: 无约束优化 牛顿法

Lecturer: 陈士祥 *Scribes:* 陈士祥

1 问题形式

无约束最优化问题

$$\min_{x\in\mathbb{R}^n} \quad f(x) \tag{12.1}$$

其目标函数 f 是定义在 $\mathbb{R}^n$ 上的实值函数，决策变量 x 的可取值之集合是全空间 $\mathbb{R}^n$. f 是二次可微的。

2 牛顿法

设 $f(x)$ 是二次可微实函数，在 x^k 附近作二阶 Taylor 展开近似

$$f(x^k+s) \approx q^k(s) = f(x^k) + {g^k}^T s + \frac{1}{2} s^T G_k s \tag{12.2}$$

其中 $g^k = \nabla f(x^k), G_k = \nabla^2 f(x^k)$.

将 $q^k(s)$ 极小化便得

$$s = -G_k^{-1} g^k. \tag{12.3}$$

上式给出的搜索方向 $-G_k^{-1}g^k$ 称为**牛顿方向 (Newton Direction).**

Example 12.1 在目标函数是正定二次函数

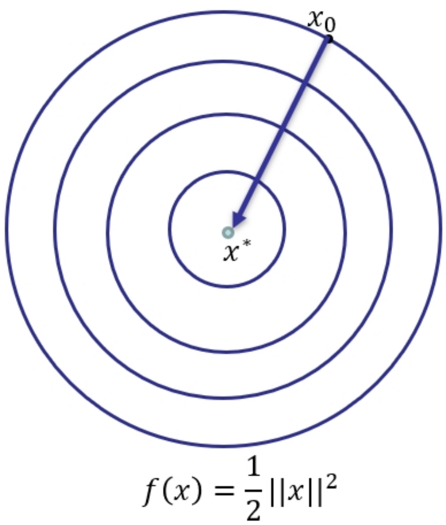
$$f(x) = \frac{1}{2}x^T G x - c^T x$$

的情况下 (G 为正定阵)，对任意的 x 有 $\nabla^2 f(x) = G$.

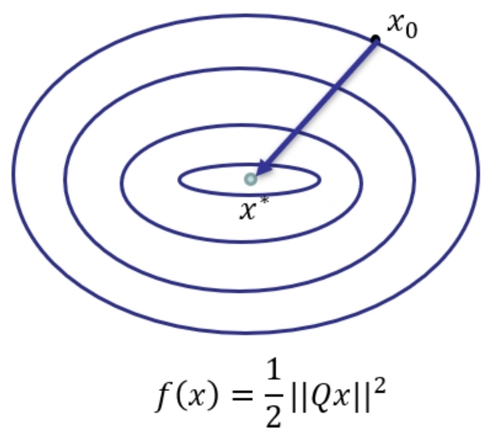
在第一次迭代里令 $H_0 = G^{-1}$, 则有

$$d^0 = -H_0 \nabla f(x^0) = -G^{-1}(Gx^0 - c) = -(x^0 - x^*).$$

这里，$x^* = G^{-1}c$ 是问题的最优解。若 $x^0 \neq x^*$, 取步长 $\alpha_0 = 1$, 于是得 $x^1 = x^0 + \alpha_0 d^0 = x^*$. 由此知道，不管初始点 x^0 如何取，在一次迭代后即可到达最优解 x^*.



(a) 牛顿法 1



(b) 牛顿法 2

图 12.1: 牛顿法对于正定二次问题，可以一步得到最优解。

选取步长 $\alpha_k \equiv 1$ 的迭代公式为

$$x^{k+1} = x^k + d^k = x^k - \nabla^2 f(x^k)^{-1} \nabla f(x^k). \tag{12.4}$$

这就是经典的牛顿迭代法。

2.1 Why is Newton's method good?

对于正定二次函数而言，牛顿法一步即可达到最优解。对于非二次函数，牛顿法并不能保证经有限次迭代求得最优解。但由于目标函数在极小点附近可用二次函数较好地近似，故当初始点靠近极小点时，牛顿法的收敛速度一般会很快。

仿射不变性 (affine-invariant): 令 $A \in \mathbb{R}^{n\times n}$ 为一个可逆矩阵。$f(x)$ 为 $\mathbb{R}^n$ 上的一个函数。考虑如下函数

$$\phi(y) = f(Ay).$$

即对于原来的函数 f，我们选择了 $\mathbb{R}^n$ 新的一组基底 A，得到新坐标下的函数 $\phi(y)$. 牛顿法的关键性质可由下面的结论说明。

Lemma 12.1 令 $\{x_k\}$ 是牛顿法对于 $f(x)$ 的序列，即

$$x_{k+1} = x_k - \nabla^2 f(x_k)^{-1} \nabla f(x_k);$$

令 $\{y_k\}$ 是牛顿法对于 $\phi(y)$ 的序列，即

$$y_{k+1} = y_k - \nabla^2\phi(y_k)^{-1}\nabla\phi(y_k);$$

若 $y_0 = A^{-1}x_0$, 则对于任意 $k \geq 1$, $y_k = A^{-1}x_k$ 。

作业 12.1 证明：Lemma 12.1

该结论说明，牛顿法的迭代点不依赖于基底和度量的选择，因此只依赖于函数的拓扑性质。

2.2 牛顿法求解等式问题

牛顿法最初是为了求解一般等式问题。设 $F:\mathbb{R}^n \to \mathbb{R}^n$, 考虑如下问题：

$$F(x) = 0.$$

迭代为

$$x_{k+1} = x_k - JF(x_k)^{-1}F(x_k). \tag{12.5}$$

对于凸问题(12.1)来说，求解(12.1)的最小值等价于求解下面的等式：

$$\nabla f(x) = 0.$$

记 $F = \nabla f(x)$, 则(12.5)与(12.4)相同。

对于一维问题，即 $F:\mathbb{R}\to\mathbb{R}$，下面的例子展示牛顿法的迭代过程。

Example 12.2 用牛顿法求解 $F(x) = x^3 - 2x + 2 = 0$ 的根。在迭代点 x_k 处，作出函数图像的切线 $l(y) = F(x_k) + F'(x_k)(y - x_k)$，与 x 轴的交点得到下一个迭代点 x_{k+1}, 即 $x_{k+1} = x_k - \frac{F(x_k)}{F'(x_k)}$. 从初始点 $x_0 = -3$ 和 $x_0 = 1$ 出发, 牛顿法迭代分别如图 *12.2*和 *12.3*. 从 $x_0 = 1$ 出发的点，由于离 $F(x) = 0$ 的根太远，牛顿法不收敛。

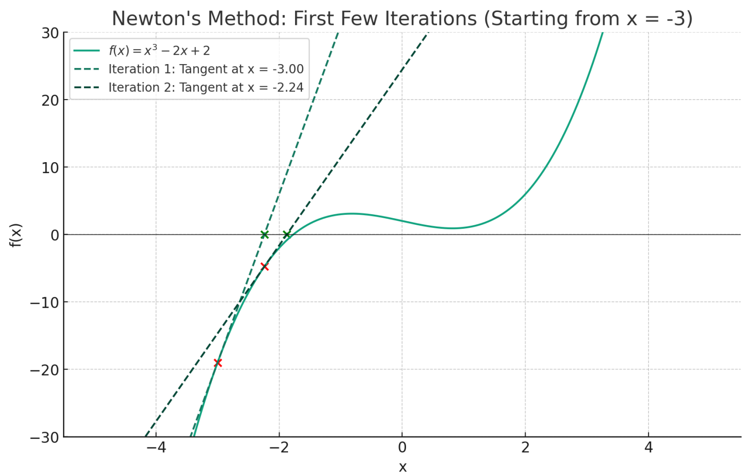


图 12.2: 从 $x_0=-3$ 出发，收敛到零点

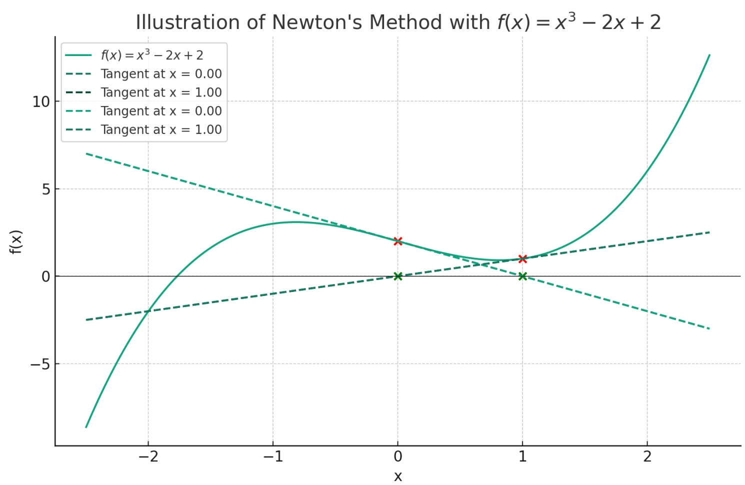


图 12.3: 从 $x_0=1$ 出发，牛顿法不收敛, 迭代点困于 0，1 两点。

上例我们知道，初始点若离最优点太远，牛顿法并不收敛。我们下面讨论牛顿法的局部收敛性质。

3 牛顿法的收敛性

Theorem 12.1 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 $\nabla f(x^*)=0, \nabla^2 f(x^*) \succ 0$, 则对于牛顿法有:

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

Proof: 根据牛顿法定义以及 $\nabla f\left(x^*\right)=0$, 得

$$\begin{aligned} x^{k+1}-x^* &= x^k-\nabla^2 f\left(x^k\right)^{-1} \nabla f\left(x^k\right)-x^* \\ &= \nabla^2 f\left(x^k\right)^{-1}\left[\nabla^2 f\left(x^k\right)\left(x^k-x^*\right)-\left(\nabla f\left(x^k\right)-\nabla f\left(x^*\right)\right)\right], \end{aligned} \tag{12.6}$$

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

$$\begin{aligned} & \left\|\nabla^2 f\left(x^k\right)\left(x^k-x^*\right)-\left(\nabla f\left(x^k\right)-\nabla f\left(x^*\right)\right)\right\| \\ = & \left\|\int_0^1\left[\nabla^2 f\left(x^k+t\left(x^*-x^k\right)\right)-\nabla^2 f\left(x^k\right)\right]\left(x^k-x^*\right) \mathrm{d} t\right\| \\ \leqslant & \int_0^1\left\|\nabla^2 f\left(x^k+t\left(x^*-x^k\right)\right)-\nabla^2 f\left(x^k\right)\right\|\left\|x^k-x^*\right\| \mathrm{d} t \\ \leqslant & \left\|x^k-x^*\right\|^2 \int_0^1 L t \ \mathrm{d} t=\frac{L}{2}\left\|x^k-x^*\right\|^2 . \end{aligned} \tag{12.7}$$

因为 $\nabla^2 f(x) \succ 0$, 由 Lipschitz 连续, 所以 $\exists r>0$, 当 $\left\|x-x^*\right\| \leqslant r$ 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 (12.6)和(12.7), 得到

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

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

另一方面, 由牛顿方程可知

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

这证明梯度的范数二次收敛到 0 .

■

4 修正牛顿法

在式(12.4)的牛顿迭代法里，如果选取的初始点 x^0 不在解 x^* 的附近，那么生成的点列 $\{x^k\}$ 未必收敛于最优解。为了保证算法的全局收敛性，有必要对牛顿法作某些改进。

线搜索牛顿法:

(0) 选取初始点 x^0, 设置终止误差 $\varepsilon > 0$, 令 $k := 0$.

(1) 计算 $g^k = \nabla f(x^k)$. 若 $\|g^k\| < \varepsilon$, 停止迭代并输出 x^k.
否则进行第 (2) 步。

(2) 解线性方程组 $\nabla^2 f(x^k) d = -g^k$, 求出牛顿方向 d^k.

(3) 采用一维搜索确定步长因子 α_k, 令 $x^{k+1} = x^k + \alpha_k d^k$, 置 $k := k+1$, 回到第 (1) 步。

牛顿法面临的另一个主要困难是 Hesse 矩阵 $G_k = \nabla^2 f(x^k)$ 不正定。这时二阶近似模型不一定有极小点，即二次函数 $q^k(s)$ 是无界的。另外，如果初始点离最优点较远，牛顿方向使用步长为 1 不一定能使得函数值减小。

为了克服这些困难，人们提出了很多修正措施。例如 Goldstein & Price (1967) 提出,

$$d^k = \begin{cases} -G_k^{-1} g^k, & \text{if } \cos\theta_k > \eta \\ \\ -g^k, & \text{otherwise} \end{cases} \tag{12.8}$$

上式中，θ_k 是 $-G_k^{-1}g^k$ 与 $-g^k$ 的夹角。即，如果牛顿方向与负梯度方向接近直角，则采用负梯度方向。

如果出现 G_k 非严格正定,或者为了保证牛顿法全局收敛,则有如下修正 Levenberg(1944), Marquardt(1963), Goldfeld et. al(1966)

$$(G_k + \mu_k I)d^k = -g^k \tag{12.9}$$

进一步的参考资料

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