章节 6.4 牛顿法

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

$$f(\mathbf{x}^k+\mathbf{s}) \approx q^k(\mathbf{s}) = f(\mathbf{x}^k) + {\mathbf{g}^k}^T\mathbf{s} + \frac{1}{2}\mathbf{s}^T G_k \mathbf{s} \tag{104}$$

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

将 $q^k(\mathbf{s})$ 极小化便得

$$\mathbf{s} = -G_k^{-1}\mathbf{g}^k. \tag{105}$$

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

牛顿法

在目标函数是正定二次函数

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

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

在第一次迭代里令 $H_0 = G^{-1}$, 则有

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

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

根据以上事实，可以认为即使对于一般的非线性函数 $f(\mathbf{x})$，在迭代中令搜索方向

$$\mathbf{d}^k = -\nabla^2 f(\mathbf{x}^k)^{-1} \nabla f(\mathbf{x}^k)$$

也是较合适的。

特别地，步长 $\alpha_k \equiv 1$ 的迭代公式为

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \mathbf{d}^k = \mathbf{x}^k - \nabla^2 f(\mathbf{x}^k)^{-1} \nabla f(\mathbf{x}^k). \tag{106}$$

这就是经典的牛顿迭代法

牛顿法为什么好?

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。

结论： 令 $\{x_k\}$ 是牛顿法对于 $f(x)$ 的序列，即

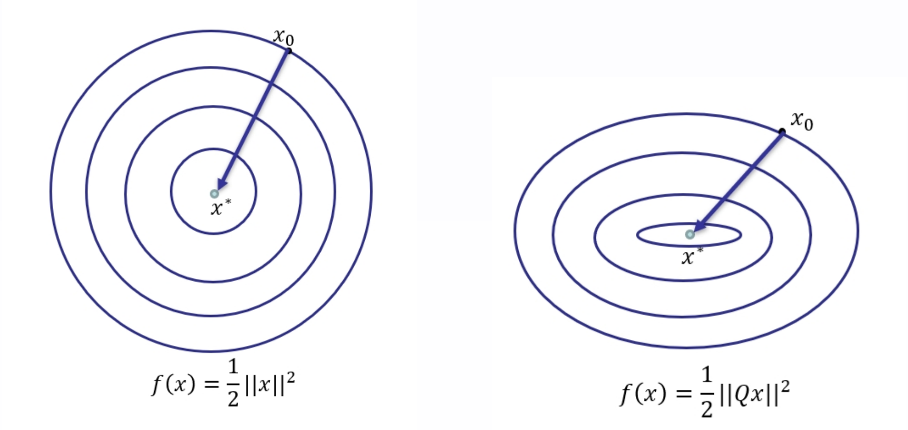
$$x_{k+1} = x_k - \nabla^2 f(x_k)^{-1} \nabla f(x_k);$$

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 $k \geq 1$, $y_k = A^{-1}x_k$ 。

证明：作业 6.6。
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。



(a) 牛顿法 1 (b) 牛顿法 2

Figure: 牛顿法对于正定二次问题，可以一步得到最优解。

牛顿法收敛定理:

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

证明:

根据牛顿法定义以及 $\nabla f(x^*) = 0$, 得

$$\begin{aligned} x^{k+1} - x^* &= x^k - \nabla^2 f\left(x^k\right)^{-1} \nabla f\left(x^k\right) - x^* \\ &= \nabla^2 f\left(x^k\right)^{-1} \left[\nabla^2 f\left(x^k\right)\left(x^k - x^*\right) - \left(\nabla f\left(x^k\right) - \nabla f(x^*)\right)\right], \end{aligned} \tag{107}$$

注意到

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

由此

$$\begin{aligned} &\left\|\nabla^2 f\left(x^k\right)\left(x^k - x^*\right) - \left(\nabla f\left(x^k\right) - \nabla f(x^*)\right)\right\| \\ =& \left\|\int_0^1 \left[\nabla^2 f\left(x^k + t\left(x^* - x^k\right)\right) - \nabla^2 f\left(x^k\right)\right]\left(x^k - x^*\right) \mathrm{d}t\right\| \\ \leqslant& \int_0^1 \left\|\nabla^2 f\left(x^k + t\left(x^* - x^k\right)\right) - \nabla^2 f\left(x^k\right)\right\| \left\|x^k - x^*\right\| \mathrm{d}t \\ \leqslant& \left\|x^k - x^*\right\|^2 \int_0^1 Lt \, \mathrm{d}t = \frac{L}{2}\left\|x^k - x^*\right\|^2. \end{aligned} \tag{108}$$

因为 $\nabla^2 f(x) \succ 0$, 由 Lipschitz 连续，所以 $\exists r > 0$, 当 $\|x - x^*\| \leqslant r$ 时有 $\left\|\nabla^2 f(x)^{-1}\right\| \leqslant 2\left\|\nabla^2 f(x^*)^{-1}\right\|$ 成立, 故结合 (107)和(108), 得到

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

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

在式(106)的牛顿迭代法里，如果选取的初始点 $\mathbf{x}^0$ 不在解 $\mathbf{x}^*$ 的附近，那么生成的点列 $\{\mathbf{x}^k\}$ 未必收敛于最优解。

为保证算法的全局收敛性，有必要对牛顿法作某些改进。

比如，在牛顿法中也可采用一维搜索来确定步长。

修正牛顿法:

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

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

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

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

牛顿法

牛顿法面临的主要困难是 Hesse 矩阵 $G_k = \nabla^2 f(\mathbf{x}^k)$ 不正定。这时二阶近似模型不一定有极小点，即二次函数 $q^k(\mathbf{s})$ 是无界的。

为了克服这些困难，人们提出了很多修正措施。

Goldstein & Price (1967)

$$\mathbf{d}^k = \begin{cases} -G_k^{-1}\mathbf{g}^k, & \text{if } \cos\theta_k > \eta \\ -\mathbf{g}^k, & \text{otherwise} \end{cases} \tag{109}$$

Levenberg(1944), Marquardt(1963), Goldfeld et. al(1966)

$$(G_k + \mu_k I)\mathbf{d}^k = -\mathbf{g}^k \tag{110}$$

设 $\mathbf{x}$ 是函数 f 的一个不定点，若方向 $\mathbf{d}$ 满足

$$\mathbf{d}^T\nabla^2 f(\mathbf{x})\mathbf{d} < 0,$$

则称 $\mathbf{d}$ 为 f 在 $\mathbf{x}$ 处的负曲率方向。

当 Hesse 矩阵 $\nabla^2 f(\mathbf{x}^k)$ 不正定时，负曲率方向法是修正牛顿法的另一种途径。