最优化理论和方法

第11讲
牛顿法

第3章 无约束问题算法(I)

一、 最速下降法

二、 Newton法及其修正形式

给定无约束问题

$$\min_{x \in R^n} f(x) \qquad (3.1)$$

下降算法的结构：$x_{k+1} = x_k + \alpha_k d_k$

（１）步长α_k计算方法：Armijo法，Wolfe-Powell法

（２）下降方向d_k的计算方法，在后面的几章里来介绍.

下降方向的计算是整个最优化方法的核心,不同计算方式对应不同的最优化算法,相应算法的收敛性理论和数值效果有很大区别

下降算法收敛性结果

定理2.4.1 - 2.4.2 设假设2.4.1成立, 序列$\{x_k\}$由算法2.1产生, 其中步长α_k由精确搜索或Wolfe - Powell搜索产生, 则

$$\sum_{k=0}^{\infty} \| \nabla f(x_k) \|^2 \cos^2 \theta_k < +\infty \qquad (2.13)$$

特别地, 若存在常数$\delta > 0$使得$\cos\theta_k \geq \delta$, 则

$$\lim_{k\to\infty} \| \nabla f(x_k) \| = 0 \qquad (2.14)$$

定理2.4.3 设假设2.4.1成立, 序列$\{x_k\}$由算法2.1产生, 其中步长α_k由Armijo搜索产生, 且存在常数$C > 0$, 使得

$$\| \nabla f(x_k) \| \leq C \| d_k \| \qquad (2.17)$$

则定理2.4.1的结论成立

下降算法收敛性结果

定理2.4.4 设假设2.4.1成立,序列$\{x_k\}$由算法2.1产生,其中步长α_k由精确搜索或Wolfe-Powell搜索或Armijo搜索确定且(2.17)成立. 若进一步

$$\liminf_{k\to\infty} \cos\theta_k > 0 \qquad (2.18)$$

则

$$\liminf_{k\to\infty} \|\nabla f(x_k)\| = 0 \qquad (2.19)$$

特别的，若存在常数$\eta > 0$,使得下式满足，则(2.18)成立

$$\prod_{i=0}^{k-1} \cos\theta_i \geq \eta^k \qquad (2.20)$$

什么影响算法的收敛性？

下降算法收敛速度结果

定理2.5.1　设假设2. 5. 1的条件成立，点列$\{x_k\}$由下降算法2.1产生其中步长α_k由Armijo搜索或Wolfe-Powell搜索确定。若存在常数$\eta > 0$，使得$\prod_{i=0}^{k-1} \cos\theta_i \geq \eta^k$,则存在常数$b > 0$，$r \in (0,1)$,使得当 k 充分大时，

$$\| \quad x_{k+1} - x^* \| \leq br^k$$

该结果说明什么 ？

下降算法的收敛速度是R线性收敛的。

下降算法收敛速度结果

定理2.5.2　设函数f二次连续可微，点列$\{x_k\}$由算法2.1产生其中步长α_k由Armijo搜索或Wolfe-Powell搜索确定，其中$\sigma_1 \in (0,1/2)$. 设$\{x_k\} \to x^*$,且$\nabla f(x^*)=0, \nabla^2 f(x^*)$正定. 若

$$\lim_{k\to\infty} \frac{\|\nabla f(x_k)+\nabla^2 f(x_k)d_k\|}{\|d_k\|}=0$$

则

(1)　当k充分大时,$\alpha_k=1$;

（2）序列$\{x_k\}$超线性收敛于x^*;

(3) 若$\nabla^2 f(x)$在x^*处Lipschitz连续，且

$$\delta_k=\frac{\|\nabla f(x_k)+\nabla^2 f(x_k)d_k\|}{\|d_k\|}=O(\|x_k-x^*\|)$$

则x_k二次收敛于x^*.

思 考

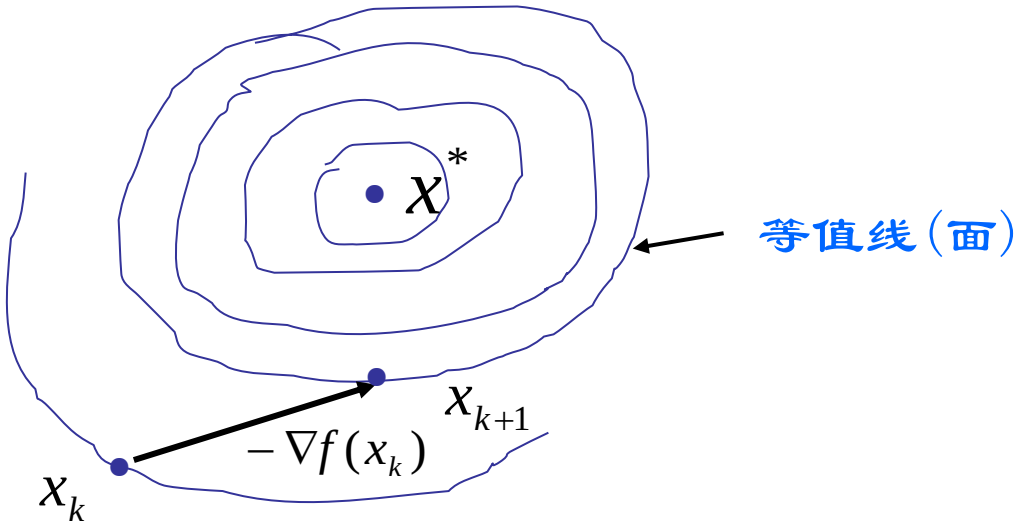
下降方向d_k如何选取?

一、 最速下降法

$$d_k = -\nabla f(x_k)$$

最古老的优化方法，十九世纪中叶由Cauchy提出

1、 思想 ：每次沿负梯度方向进行搜索



2、 算法步骤

算法3.1 (最速下降法)

步1 给定初始点$x_0 \in R^n$, 精度$\varepsilon > 0$.令$k = 0$;

步2 若$\| \nabla f(x_k) \| \leq \varepsilon$, 则得解$x_k$, 算法终止.否则

计算$d_k = -\nabla f(x_k)$, 然后转步3;

步3 由线性搜索计算步长α_k;

步4 令$x_{k+1} = x_k + \alpha_k d_k, k := k+1$, 转步2.

两点步长梯度法(BB法)

Barzilai 和 Borwein (1988) 提出两点步长梯度法，其基本思想是利用迭代当前点以及前一点的信息来确定步长因子. 迭代公式 $x_{k+1} = x_k - \alpha_k g_k$ 可以看成是

$$x_{k+1} = x_k - D_k g_k,$$

其中 $D_k = \alpha_k I$ 是一个矩阵. 为了使矩阵 D_k 具有拟牛顿性质 (拟牛顿法将在第五章讨论), 计算 α_k 使得

$$\min \|s_{k-1} - D_k y_{k-1}\|,$$

$$s_{k-1} = x_k - x_{k-1}, \quad y_{k-1} = g_k - g_{k-1}.$$

$$\alpha_k = s_{k-1}^T y_{k-1} / \|y_{k-1}\|^2 \qquad (3.1.37a)$$

步长采用上一步的最佳步长

$$x_k + \alpha_k(-g_k) = x_{k-1} + \alpha_k(-g_{k-1}),$$
$$g_k := -\nabla f(x_k)$$

算法 3.1.7 (两点步长梯度法)

步 1　给出 $x_0 \in R^n$, $0 \leqslant \varepsilon \ll 1$, $k := 0$;

步 2　计算 $d_k = -g_k$; 如果 $\|g_k\| \leqslant \varepsilon$, 则停止;

步 3　如果 $k = 0$, 利用一维搜索求 α_0; 否则, 利用 (3.1.37a) 或 (3.1.37b) 计算 α_k.

步 4　$x_{k+1} = x_k + \alpha_k d_k$;

步 5　$k := k + 1$, 转步 2.　□

$$\alpha_k = s_{k-1}^T y_{k-1} / \|y_{k-1}\|^2 \tag{3.1.37a}$$

$$\alpha_k = \|s_{k-1}\|^2 / s_{k-1}^T y_{k-1}. \tag{3.1.37b}$$

超线性收敛

$$s_{k-1} = x_k - x_{k-1}, \quad y_{k-1} = g_k - g_{k-1}.$$

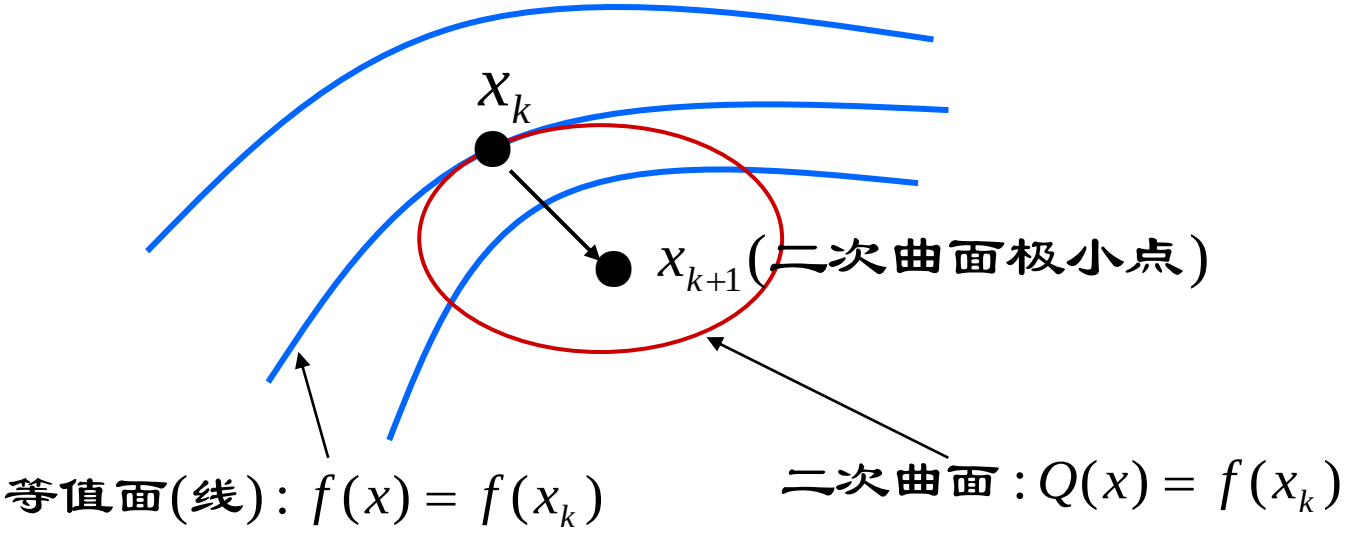
二、 Newton法及其修正形式

1、思想：用近似二次函数的极小点作为原问题的新的近似解

考虑从x_k到x_{k+1}的迭代过程，在x_k处对$f(x)$用与它最密切的二次函数$Q(x)$来近似，把二次函数的极小点作为x_{k+1}

$$d_k = -(\nabla^2 f(x_k))^{-1}\nabla f(x_k)$$

几何解释：



$f(x)$在x_k处二阶泰勒展开式为

$$f(x) = f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2}(x - x_k)^T \nabla^2 f(x_k)(x - x_k) + o(\| x - x_k \|^2)$$

二次近似函数

$$Q(x) = f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2}(x - x_k)^T \nabla^2 f(x_k)(x - x_k)$$

且 $\nabla Q(x) = \nabla f(x_k) + \nabla^2 f(x_k)(x - x_k)$

若$\nabla^2 f(x_k)$正定,则$Q(x)$的极小点$\bar{x}$为

$$\nabla f(x_k) + \nabla^2 f(x_k)(x - x_k) = 0$$

的解, 即 $\bar{x} = x_k - \nabla^2 f(x_k)^{-1} \nabla f(x_k)$

把二次函数的极小点作为x_{k+1},则

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

我们称迭代公式

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

为古典Newton法的迭代公式. 其中

$$d_k = -\nabla^2 f(x_k)^{-1} \nabla f(x_k)$$

称为x_k处的Newton方向.

若$\nabla^2 f(x_k)$正定, 则Newton方向是函数f在x_k处的一个下降方向, 并且是下列线性方程组的解

$$\nabla^2 f(x_k) d + \nabla f(x_k) = 0$$

为确保算法的下降性, 我们在古典Newton法中加入线性搜索, 得到Newton迭代公式如下：

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

2、 Newton法的算法步骤

算法3.2 (Newton法)

步1 给定初始点$x_0 \in R^n$, 精度$\varepsilon > 0$.令$k = 0$;

步2 若$\| \nabla f(x_k) \| \leq \varepsilon$, 则得解$x_k$, 算法终止. 否则

解线性方程组

$$\nabla^2 f(x_k) d + \nabla f(x_k) = 0 \qquad (3.3)$$

得解d_k;

步3 由线性搜索计算步长α_k;

步4 令$x_{k+1} = x_k + \alpha_k d_k$, $k := k+1$, 转步2.

步长α分别采用:

(1) 黄金分割法, (2) *Armijo*算法, (3) *Wolfe – Powell*法

例3.2.1 用精确搜索的Newton法求解下面的最优化问题：

$$\min f(x)=\frac{1}{2}x_1^2+x_2^2-x_1x_2-x_1$$

初始点分别取$x^{(0)}=(0,\ 0)^T$和$(1,\ 1)^T$. 该问题的最优解为

$x^*=(2,\ 1)^T$

解　经计算得：$\nabla f(x)=\begin{pmatrix}x_1-x_2-1\\ -x_1+2x_2\end{pmatrix}, Q=\nabla^2 f(x)=\begin{pmatrix}1 & -1\\ -1 & 2\end{pmatrix}$

采用精确搜索极小化二次函数的步长公式为

$$\alpha=-\frac{\nabla f(x)^T d}{d^T Q d}=-\frac{(x_1-x_2-1)d_1+(-x_1+2x_2)d_2}{d_1^2+2d_2^2-2d_1d_2}$$

又Newton方向的表达式为

$$d^{(k)}=-\nabla^2 f(x_k)^{-1}\nabla f(x_k)=-\begin{pmatrix}1 & -1\\ -1 & 2\end{pmatrix}^{-1}\begin{pmatrix}x_1^{(k)}-x_2^{(k)}-1\\ -x_1^{(k)}+2x_2^{(k)}\end{pmatrix}=-\begin{pmatrix}x_1^{(k)}-2\\ x_2^{(k)}-1\end{pmatrix}$$

表 3.1　例 3.2.1 的计算结果

$x^{(0)}$	k	$x^{(k)}$	$f(x^{(k)})$	$\nabla f(x^{(k)})$	$d^{(k)}$	α_k
$(0,0)^T$	0	$(0,0)^T$	0	$(-1,0)^T$	$(2,1)^T$	1
	1	$(2,1)^T$	-1	$(0,0)^T$		
$(1,1)^T$	0	$(1,1)^T$	$-1/2$	$(-1,1)^T$	$(1,0)^T$	1
	1	$(2,1)^T$	-1	$(0,0)^T$		

对不同的两个初始点，经一次迭代求出最优解，

这是偶然还是必然的呢？

定理3.2.1 设

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

其中Q是n阶对称正定矩阵.则从任意初始点x_0出发,采用精确搜索的Newton法最多经一次迭代即可达到f的最小值点.

证明：由于$\nabla f(x)=Qx+q,\nabla^2 f(x)=Q$正定,即$f$严格凸

如果 $x_{k+1}=x_k+\alpha_k d_k$，则$\nabla f(x_{k+1})=\nabla f(x_k)+\alpha_k Qd_k$

若$\nabla f(x_0)=0$,则x_0已经是最优解,即f的最小点;

若$\nabla f(x_0)\neq 0$,则$d_0=-Q^{-1}\nabla f(x_0)\neq 0$且精确搜索的步长

$$\alpha_0=\frac{-\nabla f(x_0)^{\mathrm{T}}d_0}{d_0^TQd_0}=\frac{\nabla f(x_0)^{\mathrm{T}}Q^{-1}\nabla f(x_0)}{\nabla f(x_0)^{\mathrm{T}}Q^{-1}\nabla f(x_0)}=1$$

所以

$$\nabla f(x_1)=\nabla f(x_0)+\alpha_0 Qd_0=\nabla f(x_0)-QQ^{-1}\nabla f(x_0)=0$$

即x_1是最小点

定义3.2.1 若一个算法求解严格凸二次函数极小化问题时，从任意初始点出发,算法经有限次迭代后可达到函数的最小值点，则称该算法具有二次终止性

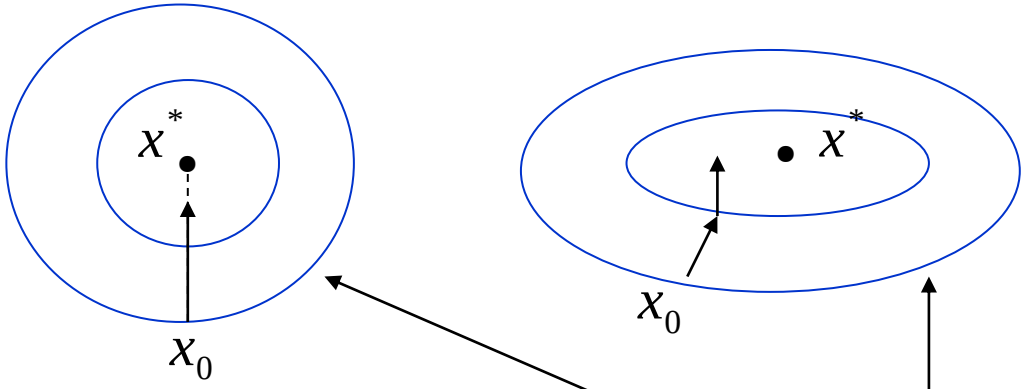
是否具有二次终止性可作为算法有效性的一个标准

最速下降法对一般的严格凸函数不具有二次终止性，而Newton法具有二次终止性. 我们后面要介绍的许多算法也具有二次终止性.

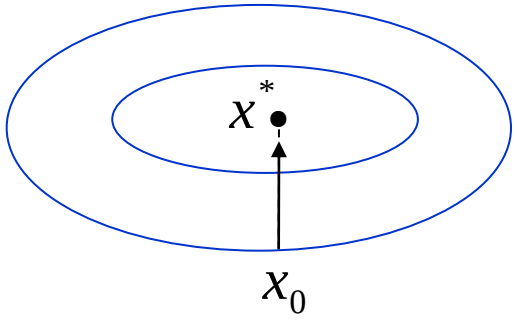
下面我们来看看最速下降法与Newton法求解二次函数的比较

考察R^2中的二次函数：注意等值线的形状

最速下降法



Newton法



3、Newton 的收敛性

由前面介绍的下降算法的收敛性，我们很容易得到Newton法关于非二次函数的全局收敛性结果如下：

定理3.2.2设函数f二次连续可微,且存在常数$m>0$,使得

$$d^T\nabla^2 f(x)d \geq m\|d\|^2, \forall d \in R^n, x \in \Omega \qquad (3.4)$$

其中：$\Omega=\{x|f(x)\leq f(x_0)\}$

设序列$\{x_k\}$由Newton算法3.2产生, 其中步长α_k由精确搜索,或Armijo型搜索，或Wolfe-Powell型搜索确定.则$\{x_k\}$收敛到f在Ω中的惟一全局最小点.

4、 局部二次收敛性

定理3.2.3 设f在$x^* \in R^n$的某个邻域内二次连续可微且x^*满足$\nabla f(x^*)=0, \nabla^2 f(x^*)$正定.则存在常数$\delta>0$,使得当

$$x_0 \in U_\delta(x^*)=\{x \mid \| x-x^* \| < \delta\}$$

时,由单位步长Newton法(古典Newton法)

$$x_{k+1}=x_k-\nabla^2 f(x_k)^{-1}\nabla f(x_k), k=0,1,2,\cdots$$

产生的序列$\{x_k\}$超线性收敛于x^*.此外, 若$\nabla^2 f$在x^*Lipschtiz 连续,即存在常数$L>0$,使得

$$\| \nabla^2 f(x)-\nabla^2 f(x^*) \| \le L \| x-x^* \|, \forall x \in U_\delta(x^*)$$

则序列$\{x_k\}$二次收敛于x^*

定理的证明：（1）收敛性

证明：由于$\nabla f(x^*)=0$,且存在常数$M>0$ 使得

$$\|\nabla^2 f(x_k)^{-1}\| \le M, \forall x_k \in U_\delta(x^*)=U(x^*,\delta)$$

所以

$$\begin{aligned}
\|x_1-x^*\| &= \|x_0-x^*-\nabla^2 f(x_0)^{-1}[\nabla f(x_0)-\nabla f(x^*)] \\
&= \left\| \nabla^2 f(x_0)^{-1}\left[\nabla f(x_0)-\nabla f(x^*)-\nabla^2 f(x_0)(x_0-x^*)\right] \right\| \\
&\le M\left\|\nabla f(x_0)-\nabla f(x^*)-\nabla^2 f(x_0)(x_0-x^*)\right\| \\
&= M\left\|\int_0^1 \nabla^2 f(x_0+t(x^*-x_0))dt-\nabla^2 f(x_0)\right\| \left\|x_0-x^*\right\| \\
&\le \frac{1}{2}\|x_0-x^*\|
\end{aligned}$$

因为当x_0充分靠近x^*,$\|x_0-x^*\|\le\delta$，δ充分小时，

$$\|\nabla^2 f(x_0+t(x^*-x_0))-\nabla^2 f(x_0)\| \le \frac{1}{2M}$$

定理的证明：

(2) 收敛速度

证明：由于$\nabla f(x^*)=0$,且存在常数$M>0$使得

$$\| \nabla^2 f(x_k)^{-1} \| \le M, \forall x_k \in U_\delta(x^*) = U(x^*,\delta)$$

所以

$$\begin{aligned}
\| x_{k+1} - x^* \| &= \| x_k - x^* - \nabla^2 f(x_k)^{-1}[\nabla f(x_k) - \nabla f(x^*)] \\
&= \left\| \nabla^2 f(x_k)^{-1}\left[\nabla f(x_k) - \nabla f(x^*) - \nabla^2 f(x_k)(x_k - x^*)\right] \right\| \\
&\le M\left\|\nabla f(x_k) - \nabla f(x^*) - \nabla^2 f(x_k)(x_k - x^*)\right\| \\
&= M * o(\| x_k - x^* \|)
\end{aligned}$$

即x_k超线性收敛到到x^*

如果进一步$\nabla^2 f(x)$在该领域内Lipschitz连续，则有

$$\| x_{k+1} - x^* \| = \left\| \nabla^2 f(x_k)^{-1} \left[\nabla f(x_k) - \nabla f(x^*) - \nabla^2 f(x_k)(x_k - x^*) \right] \right\|$$

$$= \left\| \nabla^2 f(x_k)^{-1} \left[\int_0^1 \nabla^2 f(x^* + \theta (x_k - x^*)) \mathrm{d}\theta - \nabla^2 f(x_k) \right] (x_k - x^*) \right\|$$

$$\leq ML \left\| x_k - x^* \right\|^2$$

即x_k二次收敛到到x^*

由定理3.2.2, 我们看到，当函数 f 的Hessian矩阵在Ω上一致正定时,Newton法是收敛的. 然而当这一条件不满足时，Newton法可能失效.

例题： 考察从初始点$x^{(0)}=(1,1)^{\mathrm{T}}$出发的Newton法求解问题

$$\min_{x\in R^2} f(x)=x_1^3+x_1x_2-x_1^2x_2^2$$

时, 一个精心编制的Newton法计算机程序不成功, 试分析失败的大致原因.

解： $\nabla f(x)=\begin{pmatrix}3x_1^2+x_2-2x_1x_2^2\\ x_1-2x_1^2x_2\end{pmatrix}$

$$\nabla^2 f(x)=\begin{pmatrix}6x_1-2x_2^2 & 1-4x_1x_2\\ 1-4x_1x_2 & -2x_1^2\end{pmatrix},\quad \nabla^2 f(1,1)=\begin{pmatrix}4 & -3\\ -3 & -2\end{pmatrix}$$

由于$\nabla^2 f(1,1)$是不定的, 故在$x^{(0)}$处Newton方向不一定是下降方向, 所以…

5、 Newton法的修正形式：如何有效计算下降方向

Newton法要求$\nabla^2 f(x_k)$正定，否则失效

修正Newton法：$\nabla^2 f(x_k) + v_k I \Rightarrow \nabla^2 f(x_k), v_k > 0$

最速下降 — Newton法：用最速下降方向

替换Newton方向

算法3.3 (修正Newton法)

步1 给定初始点$x_0 \in R^n$, 精度$\varepsilon > 0$.令$k = 0$;

步2 若$\| \nabla f(x_k) \| \leq \varepsilon$, 则得解$x_k$, 算法终止. 否则

解线性方程组

$$A_k d + \nabla f(x_k) = 0 \qquad (3.3)$$

得解d_k, 其中$A_k = \nabla^2 f(x_k) + v_k I, v_k > 0$使得$A_k$正定;

步3 由线性搜索计算步长α_k;

步4 令$x_{k+1} = x_k + \alpha_k d_k, k := k+1$, 转步2.

在修正Newton法中, 为确保A_k的正定性而要求$v_k > 0$足够大, 而为确保算法的收敛性又要求v_k不能太大, 即要求

$$v_k \leq C \| \nabla f(x_k) \|,$$

C为一常数, 但C的确定是一件困难的事情.

算法3.4 (Newton—最速下降法)

步1 给定初始点$x_0 \in R^n$, 精度$\varepsilon > 0$.令$k = 0$;

步2 若$\| \nabla f(x_k) \| \leq \varepsilon$, 则得解$x_k$, 算法终止. 否则

解线性方程组

$$\nabla^2 f(x_k) d + \nabla f(x_k) = 0 \qquad (3.5)$$

若有解d_k且满足$\nabla f(x_k)^{\mathrm{T}} d_k < 0$, 转步3,

否则取$d_k = -\nabla f(x_k)$;

步3 由线性搜索计算步长α_k;

步4 令$x_{k+1} = x_k + \alpha_k d_k, k := k+1$, 转步2.

该算法有较好的稳定性及较快的收敛速度

上面的Newton法的两种修正形式，在较弱的条件下具有超线性收敛性或二次收敛性.

还有很多其他的修正形式

注意：当x_k为鞍点时，即

$$\nabla f(x_k)=0, \text{但} \nabla^2 f(x_k) \text{不定}$$

所有修正失效.这时d_k可取负曲率方向，即d_k满足

$$d_k^T \nabla^2 f(x_k) d_k < 0$$

沿着此方向搜索目标函数值必下降？

Newton法 的优点：收敛快；

缺点：对初始点要求很高，而且计算量大

从Newton法出发进行修改，利用其优点，克服其缺点，产生很多效果非常好的其他新算法

作 业

- 习题3（李董辉编著）

  1，8, 12， 15,

- 思考题: 14 (d(k)满足条件该如何修改才有结论?)
- 练习Matlab编程，分别编写最速下降法,BB法和Newton法的程序，计算11，以及Rosenbock函数的极小值点：$\min f(x)=100(x_2-x_1^2)^2+(1-x_1)^2$

  初始点分别取（1.2，1.2）和（-1.2，1）。

  列出每一步的步长。比较三种算法的优劣。