第4章　无约束最优化方法

- 最优性条件
- 最速下降法
- 牛顿法、阻尼牛顿法
- 共轭方向法、共轭梯度法
- 变尺度法（***DFP***算法和***BFGS***算法）

第4章　无约束最优化方法

无约束最优化问题：

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

无约束最优化方法：

求解无约束优化问题的计算方法

最优化方法中的基本方法---无约束最优化方法

无约束最优化方法：

- 应用广泛；
- 理论也比较成熟；
- 约束优化可转化为无约束优化

令 $F(x)=\begin{cases} f(x), x\in D \\ +\infty, x\notin D. \end{cases}$

则 $\min_{x\in D} f(x) = \min_{x\in R^n} F(x),$

最优化方法中的基本方法---无约束优化方法

无约束优化方法

本章介绍解析法

- 解析法:利用函数的一阶或二阶导数的方法
  收敛速度快，需要计算梯度或者***Hesse***矩阵
  可求得目标函数的梯度时使用解析法
- 直接法 :仅利用函数值的信息，寻找最优解
  不涉及导数，适用性强，但收敛速度慢

在不能求得目标函数的梯度或偏导数时使用直接法

本章内容

- 最优性条件
- 最速下降法
- 牛顿法、阻尼牛顿法
- 共轭方向法、共轭梯度法
- 变尺度法（***DFP***算法和***BFGS***算法）

最优性条件(Optimality Conditions)

最优性条件

- 优化问题的(局部或全局)最优解所必须满足的条件；
- 常见的有一阶必要条件和二阶充分条件；
- 是优化算法建立和分析的基础；
- 对于优化理论的研究具有重要意义。

无约束优化的最优性条件----一阶必要条件

定理(一阶必要条件)

设 $f:R^n\to R$, 若x*为 f 的局部极小点，且在$N_\varepsilon(x^*)$内连续可微，则 $\nabla f(x^*)=0$.

定理3 (必要条件)

设$\boldsymbol{f}:\boldsymbol{D}\subseteq\boldsymbol{R}^n\to\boldsymbol{R}$

(1) $\boldsymbol{x}$*为$\boldsymbol{D}$的一个内点;

(2) $\boldsymbol{f}$ 在$\boldsymbol{x}$* 可微;

(3) $\boldsymbol{x}$*为$\boldsymbol{f}$ 的极值点;

则 $\nabla\boldsymbol{f}\left(\boldsymbol{x}^*\right)=\boldsymbol{0}.$

无约束优化的最优性条件----二阶必要条件

定理（二阶必要条件）

若x*为 f 的局部极小点，且在$N_\varepsilon(x^*)$ 内 f 二次连续可微，则$\nabla f(x^*)=0,\nabla^2 f(x^*)$半正定。

无约束优化的最优性条件----二阶充分条件

定理（二阶充分条件）

设 $f: R^n \to R$, 若 $\nabla f(x^*)=0$, f 在 $N_\varepsilon(x^*)$ 内二次连续可微，且 $\nabla^2 f(x^*)$正定，则x^*为 f 的严格局部极小点。

如果 $\nabla^2 f(x^*)$ 负定，则x^*为 f 的严格局部极大点。

定理4 (充分条件) 设$f: D \subseteq R^n \to R$

(1) x^*为D的一个内点；

(2) f 在x^*二次连续可微；

(3) $\nabla f(x^*)=0$；

(4) $\nabla^2 f(x^*)$正定；

则x^*是f 的严格局部极小点。

无约束凸优化的最优性条件----一阶条件

定理（一阶充要条件）

设 $f: R^n \to R$ 是凸函数且在x^*处连续可微，则x^*为f 的全局极小点的充要条件是$\nabla f(x^*)=0$.

定理

设$f: R^n \to R$ 是严格凸函数且在x^*处连续可微，若 $\nabla f(x^*)=0$, 则x^*为 f 的唯一全局极小点。

无约束优化的最优性条件

例　利用最优性条件求解下列问题：

$$\min f(x)=\frac{1}{3}x_1^3+\frac{1}{3}x_2^3-x_2^2-x_1.$$

利用一阶条件求驻点

解　$\frac{\partial f}{\partial x_1}=x_1^2-1,\ \frac{\partial f}{\partial x_2}=x_2^2-2x_2,$

利用二阶条件判断驻点是否是极小点

令$\nabla f(x)=0$,即

$$\begin{cases} x_1^2-1=0 \\ x_2^2-2x_2=0, \end{cases}$$

得到驻点： $x_1=\begin{pmatrix}1\\0\end{pmatrix}, x_2=\begin{pmatrix}1\\2\end{pmatrix}, x_3=\begin{pmatrix}-1\\0\end{pmatrix}, x_4=\begin{pmatrix}-1\\2\end{pmatrix}.$

无约束优化的最优性条件

函数 f 的***Hesse***阵：

$$\nabla^2 f(x)=\begin{pmatrix}2x_1 & 0\\ 0 & 2x_2-2\end{pmatrix},$$

利用二阶条件判断驻点是否是极小点

在点 x_1,x_2,x_3,x_4 处的***Hesse***阵依次为：

$$\nabla^2 f(x_1)=\begin{pmatrix}2 & 0\\ 0 & -2\end{pmatrix},\ \nabla^2 f(x_2)=\begin{pmatrix}2 & 0\\ 0 & 2\end{pmatrix},$$

$$\nabla^2 f(x_3)=\begin{pmatrix}-2 & 0\\ 0 & -2\end{pmatrix},\ \nabla^2 f(x_4)=\begin{pmatrix}-2 & 0\\ 0 & 2\end{pmatrix}.$$

$$x_1=\begin{pmatrix}1\\0\end{pmatrix},\quad x_2=\begin{pmatrix}1\\2\end{pmatrix}, x_3=\begin{pmatrix}-1\\0\end{pmatrix}, x_4=\begin{pmatrix}-1\\2\end{pmatrix}.$$

无约束优化的最优性条件

$\nabla^2 f(x_1)=\begin{pmatrix}2 & 0\\ 0 & -2\end{pmatrix}, \nabla^2 f(x_4)=\begin{pmatrix}-2 & 0\\ 0 & 2\end{pmatrix}$ 的行列式小于0；

x_1, x_4 是鞍点；

$\nabla^2 f(x_2)=\begin{pmatrix}2 & 0\\ 0 & 2\end{pmatrix}$ 是正定矩阵；

x_2 是极小点；

$\nabla^2 f(x_3)=\begin{pmatrix}-2 & 0\\ 0 & -2\end{pmatrix}$ 是负定矩阵；

x_3 是极大点。

- 对某些较简单的函数，这样做有时是可行的；
- 但对一般n元函数f来说，由$\nabla f(x)=0$得到的是一个非线性方程组，求解相当困难。
- 常使用迭代法直接求解无约束优化问题。

根据迭代点是否沿某个方向产生
- 线搜索方法：迭代点沿某方向产生
- 信赖域方法：迭代点在某区域内搜索产生

线搜索迭代法的步骤

步骤1 选定初始点x^1，并令$k:=1$；

步骤2 检验x^k是否满足终止条件，若满足，则停止迭代，否则，转步骤3。

步骤3 确定搜索方向d^k；

步骤4 从x^k出发，沿方向d^k求步长λ_k，产生下一个迭代点x^{k+1}； 令$k:=k+1$，转步骤2.

步长可以是最佳步长、可接受步长、固定步长；

不同的搜索方向，对应着不同的算法。

最速下降法

负梯度方向

这是函数值减少最快的方向

假设f连续可微，取

$$d^k=-\nabla f(x^k),$$

$$\lambda_k=\arg\min_{\lambda\ge 0} f(x^k+\lambda d^k),$$

从而得到第$k+1$次迭代点，即

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

最速下降法是求多元函数极值的最古老的数值算法，早在1847年法国数学家Cauchy提出该算法，后来Curry作了进一步的研究。

最速下降法--算法步骤

步骤1　选定初始点 $x^1, \varepsilon>0$, 并令 $k=1$.

步骤2　若 $\|\nabla f(x^k)\| \le \varepsilon$, 算法终止，得到近似驻点 x^k, 否则转步骤3.

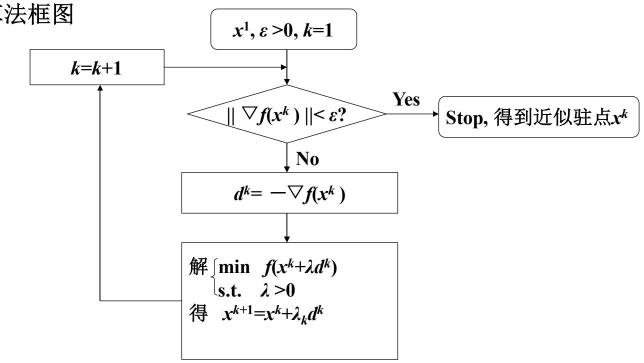
步骤3　令 $d^k = -\nabla f(x^k)$.

步骤4　确定最佳步长 $\lambda_k = \arg\min f(x^k + \lambda d^k)$,

令 $x^{k+1} = x^k + \lambda_k d^k, k := k+1$, 转步骤2.

最速下降法--算法框图

算法框图



最速下降法--算例

例　利用最速下降法求解 $\min f(x) = x_1^2 + 2x_2^2 - 2x_1x_2 - 4x_1$,

取 $x^1 = (1,1)^T, \varepsilon = 10^{-2}$.　　$x^* = (4,2)^T$.

解　函数的梯度为 $\nabla f(x) = \begin{pmatrix} 2x_1 - 2x_2 - 4 \\ -2x_1 + 4x_2 \end{pmatrix}$,

第1次迭代

$$\nabla f(x^1) = \begin{pmatrix} -4 \\ 2 \end{pmatrix},\ d^1 = -\nabla f(x^1) = \begin{pmatrix} 4 \\ -2 \end{pmatrix},$$

$$x^1 + \lambda d^1 = \begin{pmatrix} 1+4\lambda \\ 1-2\lambda \end{pmatrix},$$

$$\varphi(\lambda) = f(x^1 + \lambda d^1) = f(1+4\lambda, 1-2\lambda)$$

$$= (1+4\lambda)^2 + 2(1-2\lambda)^2 - 2(1+4\lambda)(1-2\lambda) - 4(1+4\lambda) = 40\lambda^2 - 20\lambda - 3$$

令 $0 = \varphi'(\lambda) = 80\lambda - 20$,　　得 $\lambda_1 = 1/4$,

最速下降法--算例

$$\nabla f(x) = \begin{pmatrix} 2x_1 - 2x_2 - 4 \\ -2x_1 + 4x_2 \end{pmatrix},\quad x^1 = (1,1)^T, \lambda_1 = 1/4, d^1 = \begin{pmatrix} 4 \\ -2 \end{pmatrix},$$

$$x^2 = x^1 + \lambda_1 d^1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} + 1/4 \begin{pmatrix} 4 \\ -2 \end{pmatrix} = \begin{pmatrix} 2 \\ 1/2 \end{pmatrix},\quad \nabla f(x^2) = \begin{pmatrix} -1 \\ -2 \end{pmatrix},$$

第2次迭代

$$d^2 = -\nabla f(x^2) = \begin{pmatrix} 1 \\ 2 \end{pmatrix},\quad x^2 + \lambda d^2 = \begin{pmatrix} 2+\lambda \\ 1/2+2\lambda \end{pmatrix},$$

$$\varphi(\lambda) = f(x^2 + \lambda d^2) = f(2+\lambda, 1/2+2\lambda)$$

$$= (2+\lambda)^2 + 2(1/2+2\lambda)^2 - 2(2+\lambda)(1/2+2\lambda) - 4(2+\lambda) = 5\lambda^2 - 5\lambda - 11/2$$

令 $0 = \varphi'(\lambda) = 10\lambda - 5$,　得 $\lambda_2 = 1/2$,

$$x^3 = x^2 + \lambda_2 d^2 = \begin{pmatrix} 2 \\ 1/2 \end{pmatrix} + 1/2 \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 5/2 \\ 3/2 \end{pmatrix},\quad \nabla f(x^3) = \begin{pmatrix} -2 \\ 1 \end{pmatrix},$$

最速下降法--算例

$$\nabla f(x)=\begin{pmatrix}2x_1-2x_2-4\\-2x_1+4x_2\end{pmatrix},\quad x^3=\begin{pmatrix}5/2\\3/2\end{pmatrix},\nabla f\left(x^3\right)=\begin{pmatrix}-2\\1\end{pmatrix},$$

第3次迭代

$$d^3=-\nabla f\left(x^3\right)=\begin{pmatrix}2\\-1\end{pmatrix},\quad x^3+\lambda d^3=\begin{pmatrix}5/2+2\lambda\\3/2-\lambda\end{pmatrix},$$

$$\phi(\lambda)=f\left(x^3+\lambda d^3\right)=f\left(5/2+2\lambda,3/2-\lambda\right)$$

$$=(5/2+2\lambda)^2+2(3/2-\lambda)^2-2(5/2+2\lambda)(3/2-\lambda)-4(5/2+2\lambda)=10\lambda^2-5\lambda-27/4$$

令 $0=\varphi'(\lambda)=20\lambda-5$, 得 $\lambda_3=1/4$,

$$x^4=x^3+\lambda_3 d^3=\begin{pmatrix}5/2\\3/2\end{pmatrix}+1/4\begin{pmatrix}2\\-1\end{pmatrix}=\begin{pmatrix}3\\5/4\end{pmatrix},\quad \nabla f\left(x^4\right)=\begin{pmatrix}-1/2\\-1\end{pmatrix},$$

经17次迭代，满足终止条件，得到近似极小点 $x^*=(3.99,1.99)^T$.

最速下降法--两个特征

1. 相邻两次迭代的搜索方向正交

令 $\varphi(\lambda)=f(x^k+\lambda d^k)$,

利用精确一维搜索，可得

$$\varphi'(\lambda_k)=\left(\nabla f(x^k+\lambda_k d^k)\right)^T d^k=0$$

由此得出

$$-\nabla f(x^k)=d^k$$

$$0=\left(\nabla f(x^k+\lambda_k d^k)\right)^T d^k=\left(\nabla f(x^{k+1})\right)^T d^k=-(d^{k+1})^T d^k$$

最速下降法在相邻两次迭代的搜索方向是正交的。

最速下降法--两个特征

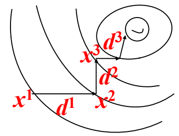
1. 相邻两次迭代的搜索方向正交 $(d^{k+1})^T d^k=0$

最速下降法在求 ***f*** 的极小点时，会发生锯齿现象

- 相邻两次迭代的搜索方向正交，
  使得迭代点向极小点的逼近是曲折前进的，
  这种现象称为锯齿现象；
- 对于一般的目标函数都会发生；
- 一些特殊的目标函数不会发生锯齿现象；
- 某些特殊的初始点不会发生锯齿现象。

最速下降法--两个特征

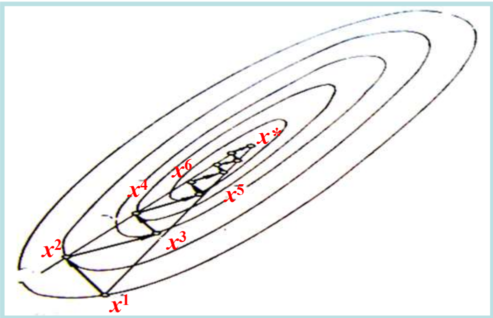
注：在最速下降法中，利用精确一维搜索求最佳步长，导致相邻两次迭代的搜索方向总是正交的，从而使得迭代点逼近极小点的过程是“之”字形。



从任何一个初始点开始，都可以很快到达极小点附近，但是越靠近极小点移动越慢，导致最速下降法的收敛速度很慢。

最速下降法--两个特征

2. 给定二元正定二次函数，用最速下降法求其极小点，算法产生点列：偶数点列均在一条直线上，奇数点列也均在一条直线上，且都过极小点.



给定二元正定二次函数 $f(x)=\frac{1}{2}x^TAx+b^Tx+c$, 用最速下降法求其极小点，产生点列：偶数点列均在一条直线上，奇数点列也均在一条直线上，且都过极小点.

分析：因为相邻方向正交，极小点为 $-A^{-1}b$,

则 $d^2//d^4//\dots//d^{2k}$

$\exists t,\ d^{2k}=td^2$ **t与k有关**

$$\therefore Ax^{2k}+b=t\left(Ax^2+b\right)\quad(\because d^k=-\nabla f(x^k)=-Ax^k-b)$$

$$\therefore A\left(x^{2k}+A^{-1}b\right)=tA\left(x^2+A^{-1}b\right)=A\left(t\left(x^2+A^{-1}b\right)\right)$$

$$\therefore x^{2k}+A^{-1}b=t\left(x^2+A^{-1}b\right),$$

故偶数点列共线，且过极小点。

最速下降法--收敛性分析

收敛性定理

设f连续可微，水平集 $L=\{x \mid f(x)\le f(x^1)\}$有界，则最速下降法或者在有限迭代步后得到驻点，或者得到点列$\{x^k\}$,其任何聚点都是f的驻点。

推论

在收敛定理的假设下，若f为凸函数，则最速下降法或在有限迭代步后达到极小点，或得到点列$\{x^k\}$,其任何聚点都是f的极小点。

用于二次函数时的收敛速度分析

定理 设二次函数 $f(x)=1/2x^TAx$, A对称正定，λ_1,λ_2分别为其最小和最大特征值，从任意初点 x^1 出发，用最速下降法求f的极小点，产生的序列为$\{x^k\}$, 对于 $k\ge2$ 有

$$f(x^k)\le\left(\frac{\lambda_2-\lambda_1}{\lambda_2+\lambda_1}\right)^2 f(x^{k-1}),\quad \|x^k\|\le\sqrt{\frac{\lambda_2}{\lambda_1}}\left(\frac{\lambda_2-\lambda_1}{\lambda_2+\lambda_1}\right)^k\|x^1\|,$$

由于 $\frac{\lambda_2-\lambda_1}{\lambda_2+\lambda_1}<1$,则$x^k\to0$.

函数 $f(x)=1/2x^TAx$ 的极小点恰好是 $x^*=0$. 故从任意初始点出发，最速下降法求解二次函数均收敛，且是线性收敛。

用于二次函数时的收敛速度分析

下面说明最速下降法收敛性的几何意义。

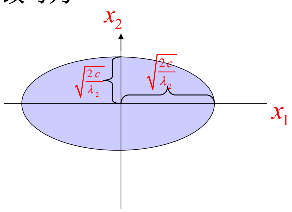
考虑A为正三角矩阵时的二次函数

$$f(x)=\frac{1}{2}x^T Ax=\frac{1}{2}\left(\lambda_1 x_1^2+\lambda_2 x_2^2\right),\quad A=\begin{pmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{pmatrix}$$

其中 $\lambda_2 \ge \lambda_1 > 0.$

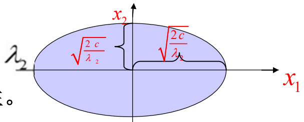
函数的等值线为 $f\left(x_1,x_2\right)=c, c>0,$ 可改写为

$$\frac{x_1^2}{\left(\sqrt{\frac{2c}{\lambda_1}}\right)^2}+\frac{x_2^2}{\left(\sqrt{\frac{2c}{\lambda_2}}\right)^2}=1,$$



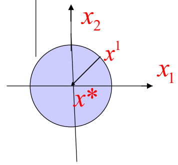
这是以 $\sqrt{\frac{2c}{\lambda_1}}$ 和 $\sqrt{\frac{2c}{\lambda_2}}$ 为半轴的椭圆。

从下面的分析可见 两个特征值 λ_1, λ_2 的相对大小决定最速下降法的收敛性。



（1）当 $\lambda_1 = \lambda_2$ 时，等值线变为圆，

此时 $\frac{\lambda_2-\lambda_1}{\lambda_2+\lambda_1}=0,$



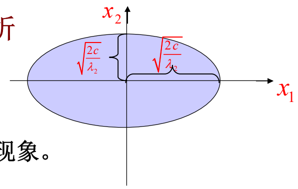
由上述定理可知: $x^2=0 \left(\because 0 \le \left\|x^2\right\| \le 0\right).$

故迭代一步就到了极小点，这表明最速下降法用于等值线为圆的目标函数时，只需迭代一步就到了极小点。

等值线为圆的目标函数，
不存在锯齿现象

$$\left\|x^k\right\| \le \sqrt{\frac{\lambda_2}{\lambda_1}}\left(\frac{\lambda_2-\lambda_1}{\lambda_2+\lambda_1}\right)^k\left\|x^1\right\|$$

用于二次函数时的收敛速度分析

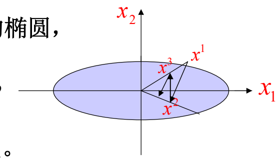


（2）当 $\lambda_1 < \lambda_2$ 时, 等值线为椭圆。

此时对于一般的初始点将产生锯齿现象。

（3）当 $\lambda_1 << \lambda_2$ 时，等值线是很扁的椭圆，

此时 $\frac{\lambda_2-\lambda_1}{\lambda_2+\lambda_1}\approx 1,$ 对于一般的初始点,



收敛速度十分缓慢，锯齿现象严重。

等值线为椭圆的目标函数，
存在锯齿现象；

椭圆越扁，锯齿现象越严重

$$\left\|x^k\right\| \le \sqrt{\frac{\lambda_2}{\lambda_1}}\left(\frac{\lambda_2-\lambda_1}{\lambda_2+\lambda_1}\right)^k\left\|x^1\right\|$$

最速下降法--优缺点

优点：理论明确，程序简单；

每次的计算量小，存储量小；

对初始点要求不严格。

缺点：收敛慢 (最速下降方向是某点的局部性质)。

最速下降法相邻两次搜索方向的正交性，决定了迭代全过程的搜索路线呈锯齿状，远快近慢。

最速下降法--优缺点

最速下降法

- 是无约束优化的基本方法之一；
- 不是好的实用算法；
- 一些有效算法是通过对它的改进
  或与其它收敛快的算法结合而得到的。

最速下降法--改进

(1) 选择不同初始点

例 $\min f(x)=x_1^2+25x_2^2$，取初始点 $x^1=(2,2)^T$.

第1次迭代

$$\nabla f(x^1)=(4,100)^T,\ d^1=-\nabla f(x^1)=(-4,-100)^T$$

$$x^1+\lambda d^1=(2-4\lambda,2-100\lambda)^T$$

$$\varphi(\lambda)=f(2-4\lambda,2-100\lambda)=(2-4\lambda)^2+25(2-100\lambda)^2$$

令 $0=\varphi'(\lambda)=-8(2-4\lambda)-5000(2-100\lambda)$，得 $\lambda_1=\frac{626}{31252}\approx 0.02003072$,

$$x^2=x^1+\lambda_1 d^1=(1.919877,-0.3071785\times10^{-2})^T$$

最速下降法--改进

$$x^2=(1.919877,-0.3071785\times10^{-2})^T$$

然后再从 x^2 开始新的迭代，

经过69次迭代，满足终止条件，

得到近似极小点 $x^*=(0.0042,0.0001)^T$.

最速下降法--改进

如果初始点不取 $x^1=(2,2)^T$ 而取 $x^1=(100,0)^T$,

第1次迭代

$$x^1=(100,0)^T,\ \nabla f(x^1)=(2x_1^1,50x_2^1)=(200,0)^T,$$

$$d^1=-\nabla f(x^1)=(-200,0)^T,$$

$$\varphi(\lambda)=f(100-200\lambda,0-0\lambda)=(100-200\lambda)^2+25(0-0\lambda)^2,$$

$$0=\varphi'(\lambda)=\frac{d}{d\lambda}\left[(100-200\lambda)^2+25(0-0\lambda)^2\right]=-400(100-200\lambda)$$

$$\lambda_1=\frac{1}{2},\quad x^2=x^1+\lambda_1 d^1=(100,0)^T+1/2(-200,0)^T=(0,0)^T.$$

一步迭代得到了极小点

最速下降法--改进

如果初始点不取$x^1=(2,2)^T$而取$x^1=(100,0)^T$，一步迭代得到极小点。

虽然$(100,0)^T$离极小点$x^*=(0,0)^T$更远，

但是迭代中没有出现锯齿现象。

锯齿现象的出现与初始点的选择有关，

选择合适的初始点可以不出现锯齿现象。

怎么选？ 很困难

最速下降法--改进

为了清除最速下降法中两个搜索方向正交的不良后果，提出了许多改进的方法，如：

（2）不采用精确一维搜索

可使$(d^{k+1})^T d^k \neq 0$, 从而改变最速下降法的收敛性。

- 采用非精确一维搜索求得的可接受步长
- 采用固定步长 → 固定步长最速下降法
    - 固定步长取几，1还是2，还是别的值，没有标准；
    - 步长取小了， 收敛慢；
    - 步长取大了，会漏掉极小点。

最速下降法--改进

（3）采用加速梯度法 负梯度方向和$d^k=x^k-x^{k-2}$结合

Shah等人于**1964**年提出了一种“平行切线法”（简记为**PARTAN**法），又称加速梯度法。

搜索方向可取$d^k=x^k-x^{k-2}$,

下两步继续用最速下降方向即负梯度方向,

这两种方向交替使用，效用要比最速下降法好的多。

Newton法

精确一维搜索中介绍了**Newton**法，即用目标函数的二阶**Tayloy**展开式近似代替目标函数，用二阶**Tayloy**展开式的极小点估计目标函数的极小点。

可推广到多维的情形

Newton法是求解无约束极小化问题的最古老的算法之一，现已发展成一类算法-----**Newton**型方法。

Newton法

精确一维搜索中的Newton迭代公式

$$x_{k+1}=x_k-\frac{f'(x_k)}{f''(x_k)},$$

多维无约束优化中的Newton迭代公式

$$x^{k+1}=x^k-\left(\nabla^2 f\left(x^k\right)\right)^{-1}\nabla f\left(x^k\right).$$

Newton法-- 基本思路

在x^k处对函数f 作二阶Tayloy展开：$f(x)=f(x^k)+\nabla f\left(x^k\right)^T\left(x-x^k\right)$

$$+\frac{1}{2}\left(x-x^k\right)^T\nabla^2 f\left(x^k\right)\left(x-x^k\right)+o\left(\left\|x-x^k\right\|^2\right)$$

略去高阶项，

$$f(x)\approx Q(x):=f(x^k)+\nabla f\left(x^k\right)^T\left(x-x^k\right)+\frac{1}{2}\left(x-x^k\right)^T\nabla^2 f\left(x^k\right)\left(x-x^k\right),$$

令 $\nabla Q(x)=\nabla f\left(x^k\right)+\nabla^2 f\left(x^k\right)\left(x-x^k\right)=0,$

若 $\nabla^2 f\left(x^k\right)\left(=\nabla^2 Q(x)\right)$正定，则二次函数的极小点为：

$$\boxed{x^{k+1}:=x^k-\left(\nabla^2 f\left(x^k\right)\right)^{-1}\nabla f\left(x^k\right),} \longrightarrow \text{Newton 迭代公式}$$

以x^{k+1}作为f 极小点的一个新的估计。

Newton法--算法步骤

求函数f 的极小点，给定误差限ε .

步骤1 选定初始点 x^1，计算 $f_1=f\left(x^1\right),k=1.$

步骤2 如果 $\left\|\nabla f\left(x^k\right)\right\|\le\varepsilon$, 停止，得到近似驻点 x^k,否则转步骤3.

步骤3 计算搜索方向 $d^k=-\left(\nabla^2 f\left(x^k\right)\right)^{-1}\nabla f\left(x^k\right)$ 牛顿方向

步骤4 令 $x^{k+1}=x^k+d^k, k=k+1,$ 转步骤2.

步长取常数1

Newton法--算法步骤

步骤1 选定初始点 x^1 ，计算 $f_1=f\left(x^1\right),k=1.$

步骤2 如果$\left\|\nabla f\left(x^k\right)\right\|\le\varepsilon$, 停止，得到近似驻点$x^k$, 否则转步骤3.

步骤3 计算搜索方向$d^k=-\left(\nabla^2 f\left(x^k\right)\right)^{-1}\nabla f\left(x^k\right).$

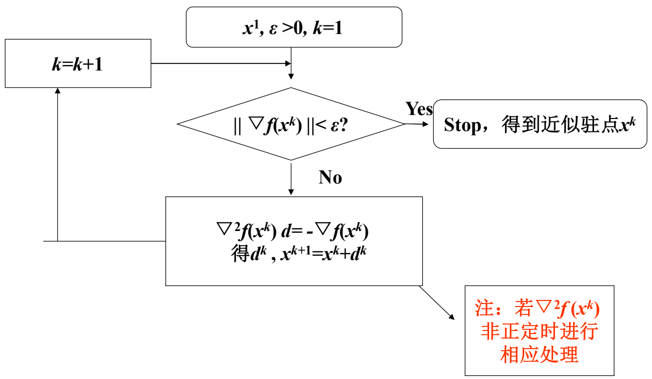
步骤4 令 $x^{k+1}=x^k+d^k, k=k+1,$ 转步骤2.

步骤3中的搜索方向 d^k， 可通过求解下列方程组得到

$$\nabla^2 f\left(x^k\right)d^k+\nabla f\left(x^k\right)=0.$$

- 已有标准程序解线性方程组；
- 减少计算量。

Newton法--算法框图



Newton法----算例

例 用Newton法求 $f(x_1,x_2)=x_1^2+25x_2^2$的极小点。

解 取初始点 $x^1=(2,2)^T$，

计算 $\nabla f(x^1)=\begin{pmatrix}2x_1\\50x_2\end{pmatrix}\Big|_{x^1}=\begin{pmatrix}4\\100\end{pmatrix},\nabla^2 f(x^1)=\begin{pmatrix}2&0\\0&50\end{pmatrix},$

代入Newton迭代公式，

$$x^2=x^1-\left(\nabla^2 f(x^1)\right)^{-1}\nabla f(x^1)=\begin{pmatrix}2\\2\end{pmatrix}-\begin{pmatrix}\frac{1}{2}&0\\0&\frac{1}{50}\end{pmatrix}\begin{pmatrix}4\\100\end{pmatrix}=\begin{pmatrix}0\\0\end{pmatrix},$$

得到问题的极小点$x^*=x^2$.

一步即达到最优解

利用Newton法求n元正定二次函数的极小点，从任意初始点出发，一步迭代即可达到极小点。

设$f(x)=1/2x^TAx+b^Tx+c$, 其中$A_{n\times n}$对称正定，$b\in R^n, c\in R$.

f的极小点是f等值面的中心，$x^*=-A^{-1}b$.

下面用Newton法求f的极小点：

$$\forall x^1\in R^n,\ \nabla f(x^1)=Ax^1+b,\ \nabla^2 f(x^1)=A,$$

$$x^2=x^1-\left(\nabla^2 f(x^1)\right)^{-1}\nabla f(x^1)=x^1-A^{-1}(Ax^1+b)=-A^{-1}b.$$

一步达到最优解

Newton法-- 基本思路

在x^k处对函数f 作二阶Taylor展开：

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

略去高阶项，

$$f(x)\approx Q(x):=f(x^k)+\nabla f(x^k)^T(x-x^k)+\frac{1}{2}(x-x^k)^T\nabla^2 f(x^k)(x-x^k),$$

令$\nabla Q(x)=\nabla f(x^k)+\nabla^2 f(x^k)(x-x^k)=0,$

若$\nabla^2 f(x^k)\left(=\nabla^2 Q(x)\right)$正定，则二次函数的极小点为：

$$\boxed{x^{k+1}:=x^k-\left(\nabla^2 f(x^k)\right)^{-1}\nabla f(x^k),}\ \Longrightarrow\ \text{Newton迭代公式}$$

以x^{k+1}作为f极小点的一个新的估计。

利用Newton法求n元正定二次函数的极小点，从任意初始点出发，一步迭代即可达到极小点。

Newton法具有二次收敛性

二次收敛性（二次终止性）

从任意初始点出发，经有限次迭代总可以达到n元正定二次函数的极小点，称这样的算法具有二次收敛性。

Newton法比最速下降法收敛快。

Newton法：局部二阶收敛

当初始点靠近极小点时，Newton法的收敛速度很快。

当初始点远离极小点时，Newton法可能不收敛，甚至连下降性都保证不了。

Newton法--优缺点

优点：算法收敛速度快 (初始点离极小点很近)；

不需要进行一维搜索；

对n元正定二次函数，迭代一次就可得到极小点。

Newton法--优缺点

缺点：(1) 对多数算法不具有全局收敛性；

(2) 每次迭代都要计算Hesse矩阵，计算量大；

(3) 每次迭代都要计算 $(\nabla^2 f(x^k))^{-1}$或者求解方程组

$$\nabla^2 f(x^k)d+\nabla f(x^k)=0,$$

- $(\nabla^2 f(x^k))^{-1}$可能不存在；
- 方程组是奇异的，病态的；
- $\nabla^2 f(x^k)$非正定，d^k 可能不是下降方向。

(4) 收敛于鞍点或极大点的可能性并不小。

当 $\nabla^2 f(x^k)$正定时，

$$\left(\nabla f(x^k)\right)^T d^k = -\left(\nabla f(x^k)\right)^T (\nabla^2 f(x^k))^{-1}\nabla f(x^k) < 0.$$

Newton法--改进

Newton法的优点和缺点都很突出，本身并不实用；

对Newton法进行改进，可得到求解无约束优化的有效方法.

怎么改进呢？保留Newton法的优点，克服部分缺点。

针对Newton法的缺点(1) 对多数算法不具有全局收敛性，

和 (4) 收敛于鞍点或极大点的可能性并不小，

步长不取固定值1，而是采用精确一维搜索

$$\min f(x^k+\lambda d^k)$$

找最佳步长λ_k，这就是阻尼Newton法。

Newton法的改进----阻尼Newton法

在Newton迭代公式中，

$$x^{k+1}=x^k-\left(\nabla^2 f\left(x^k\right)\right)^{-1}\nabla f\left(x^k\right),$$

加入精确一维搜索：$\min f\left(x^k+\lambda d^k\right)$

每次迭代目标函数值一定有所下降

求得最佳步长λ_k，得到下个迭代点 $x^{k+1}=x^k+\lambda_k d^k$.

这样修正之后通常可改进Newton法的缺点(1)和(4)。

缺点(1) 对多数算法不具有全局收敛性；

(4) 收敛于鞍点或极大点的可能性并不小；

Newton法的改进----阻尼Newton法

收敛性定理

设f存在二阶连续偏导数，$\nabla^2 f(x)\,(\forall x)$正定，水平集 $L=\left\{x \mid f(x)\le f(x^1)\right\}$有界，则阻尼Newton法或在有限迭代步后得到极小点，或得到无穷点列$\left\{x^k\right\}$，

(1) $\left\{f(x^k)\right\}$为严格单调下降序列且$\lim\limits_{k\to\infty} f(x^k)$存在；

(2) $\left\{x^k\right\}$有唯一聚点x^*，它是f的极小点。

特点：全局收敛

例 用阻尼Newton法求下列无约束优化的极小点，已知

$$f(x)=x_1^2+2x_2^2-2x_1x_2-4x_1,\ x^1=(1,1)^T,\varepsilon=10^{-2}.$$

解:

计算 $\nabla f(x^1)=\begin{pmatrix}2x_1-2x_2-4\\-2x_1+4x_2\end{pmatrix}\bigg|_{x=x^1}=\begin{pmatrix}-4\\2\end{pmatrix},$

$$\nabla^2 f(x^1)=\begin{pmatrix}2&-2\\-2&4\end{pmatrix}\bigg|_{x=x^1}=\begin{pmatrix}2&-2\\-2&4\end{pmatrix},$$

故 $d^1=-\left(\nabla^2 f(x^1)\right)^{-1}\nabla f(x^1)=-\begin{pmatrix}2&-2\\-2&4\end{pmatrix}^{-1}\begin{pmatrix}-4\\2\end{pmatrix}$

$$=-\begin{pmatrix}1&1/2\\1/2&1/2\end{pmatrix}\begin{pmatrix}-4\\2\end{pmatrix}=\begin{pmatrix}3\\1\end{pmatrix},$$

$$x^1+\lambda d^1=\begin{pmatrix}1+3\lambda\\1+\lambda\end{pmatrix},\qquad x^1=\begin{pmatrix}1\\1\end{pmatrix},d^1=\begin{pmatrix}3\\1\end{pmatrix},$$

$$\phi(\lambda)=f\left(x^1+\lambda d^1\right)=f\left(1+3\lambda,1+\lambda\right)$$

$$=\left(1+3\lambda\right)^2+2\left(1+\lambda\right)^2-2\left(1+3\lambda\right)\left(1+\lambda\right)-4\left(1+3\lambda\right)$$

令 $0=\varphi'(\lambda)=6\left(1+3\lambda\right)+4\left(1+\lambda\right)-6\left(1+\lambda\right)-2\left(1+3\lambda\right)-12$

得 $\lambda_1=1,\quad x^2=x^1+\lambda_1 d^1=\begin{pmatrix}4\\2\end{pmatrix},$

$\nabla f(x^2)=\begin{pmatrix}2x_1-2x_2-4\\-2x_1+4x_2\end{pmatrix}\Bigg|_{x=x^2}=\begin{pmatrix}0\\0\end{pmatrix},$ 又该问题是凸规划，

故得到极小点 $x^*=x^2=\begin{pmatrix}4\\2\end{pmatrix}.$

一步即达到最优解

利用阻尼**Newton**法求***n***元正定二次函数的极小点，从任意初始点出发，一步迭代即可达到极小点。

设 $f(x)=1/2x^TAx+b^Tx+c$,其中 $A_{n\times n}$ 正定，$b\in R^n,c\in R.$

f 的极小点是***f***等值面的中心，$x^*=-A^{-1}b.$

下面用阻尼**Newton**法求解***f***的极小点：

$$\forall x^1\in R^n,\ \nabla f(x^1)=Ax^1+b,\ \nabla^2 f(x^1)=A,$$

$$d^1=-\left(\nabla^2 f(x^1)\right)^{-1}\nabla f(x^1)=-A^{-1}g_1,$$

利用阻尼**Newton**法求***n***元正定二次函数的极小点，从任意初始点出发，一步迭代即可达到极小点。

$f(x)=1/2x^TAx+b^Tx+c,$

$d^1=-A^{-1}g_1,$

最佳步长 $\lambda_1=\arg\min f(x^1+\lambda d^1):=\arg\min\phi(\lambda)$

由 $\phi'(\lambda_1)=\left(\nabla f(x^1+\lambda_1 d^1)\right)^T d^1=0,$

$$0=\left(A(x^1+\lambda_1 d^1)+b\right)^T d^1=\left(g_1+\lambda_1 Ad^1\right)^T d^1,$$

故

$$\lambda_1=-\frac{g_1^Td^1}{(d^1)^TAd^1}=-\frac{g_1^Td^1}{(d^1)^TA\left(-A^{-1}g_1\right)}=\frac{g_1^Td^1}{(d^1)^Tg_1}=1,$$

$$x^2=x^1+\lambda_1 d^1=x^1+1*\left(-A^{-1}g_1\right)=x^1-A^{-1}(Ax^1+b)=-A^{-1}b.$$

一步即达到最优解

利用阻尼**Newton**法求***n***元正定二次函数的极小点，从任意初始点出发，一步迭代即可达到极小点。

阻尼**Newton**法具有二次终止性

二次终止性（二次收敛性）

从任意初始点出发，经有限次迭代总可以达到***n***元正定二次函数的极小点，称这样的算法具有二次终止性。

Newton法的进一步修正

阻尼Newton法改进了Newton法，但还是存在缺点：

(2) 每次迭代都要计算Hesse矩阵，计算量大；

(3) $\left(\nabla^2 f(x^k)\right)^{-1}$可能不存在，即使存在，也未必正定，因而

牛顿方向不一定是下降方向。

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

缺点(3)的改进方法之一：

当d^k为函数上升方向时，可向其负方向搜索，

但可能出现在$\pm d^k$上函数值都不变化的情况。

Newton法的改进---针对缺点(2)每次计算Hesse矩阵

(1) 为减小工作量，取m(正整数)，使每m次迭代使用同一个Hesse阵，迭代公式变为：

$$x^{k_m+j+1}=x^{k_m+j}-(\nabla^2 f(x^{k_m}))^{-1}\nabla f(x^{k_m+j}),$$

$$j=0,1,\ldots,m-1,\ k=0,1,2,\ldots$$

特点：收敛速度随m的增大而下降。

m=1时即Newton法，$m\to\infty$ 线性收敛。

牛顿法还有其它的修正方式

Newton法的改进---针对缺点(3)非正定和奇异的情况

(2) Goldstein-Price方法(G-P法)：

搜索方向: $d^k=\begin{cases}-(\nabla^2 f(x^k))^{-1}\nabla f(x^k), & \nabla^2 f(x^k)\text{ 正定}\\ -\nabla f(x^k), & \text{否则}\end{cases}$

步长：Armijo- Goldstein准则求λ_k，

$$\begin{cases}f(x^k+\lambda_k d^k)\le f(x^k)+\delta\lambda_k\left(\nabla f(x^k)\right)^T d^k\\ f(x^k+\lambda_k d^k)\ge f(x^k)+(1-\delta)\lambda_k\left(\nabla f(x^k)\right)^T d^k\end{cases},\ \delta\in\left(0,\frac{1}{2}\right),$$

特点：在一定条件下，G-P法全局收敛。

但当$\nabla^2 f(x^k)$非正定情况较多时，收敛速度降为接近线性。

Newton法的改进---针对缺点(3)非正定的情况

(3) Levenberg-Marguardt法（L-M法）：

主要思想：

找到尽可能小的 $\mu>0$使$\nabla^2 f(x^k)+\mu I$ 正定，

其中I为单位矩阵，

用 $\nabla^2 f(x^k)+\mu I$取代$\nabla^2 f(x^k)$进行迭代。

特点：全局二阶收敛。

作业 P 99 4.3

4.3 迭代14次得到近似解$(-0.6344,-0.0032)^T$，近似最优值为-0.4724.

共轭方向法和共轭梯度法

最速下降法，计算步骤简单，但收敛速度慢。

Newton法和阻尼**Newton**法收敛速度快，但需要计算**Hesse**矩阵及其逆矩阵，计算量、存储量很大。

需要寻找一种好的算法，这种算法能够兼有这两种方法的优点，又能克服它们的缺点，即收敛速度快同时计算简单。

这就是要讨论的共轭方向法和共轭梯度法。

共轭方向法和共轭梯度法

共轭方向法和共轭梯度法：收敛速度快同时计算简单。

共轭方向法计算效果好，应用广泛；

共轭梯度法是最著名的共轭方向法。

共轭方向法---共轭方向及其性质

定义 设 $A_{n\times n}$ 是对称正定矩阵，给定非**0**向量 $p^1, p^2, \ldots, p^m \in R^n$，

(1)如果 $\left(p^1\right)^T Ap^2=0$，则称 p^1 和 p^2 是 A 共轭(或 A 正交)的。

(2)如果 $(p^i)^T Ap^j = 0\ (\forall i \neq j)$，则称 $p^1, p^2, \ldots, p^m$ 是 A-共轭(或 A 正交)向量组，也称它们是一组 A 共轭方向。

注：

若 $A=I$, $\left(p^1\right)^T Ap^2=\left(p^1\right)^T p^2=0$, 则 p^1 与 p^2 是正交的。

共轭是正交的推广；

共轭向量组是正交向量组的推广。

共轭方向法---共轭方向及其性质

假设 f 是 n 元正定二次函数 $f(x)=\frac{1}{2}x^T Ax+b^T x+c$, A 正定，

二维情况（n=2）

任取初始点 x^1，沿某个下降方向 d^1 作精确一维搜索，得

$$x^2 = x^1 + \lambda_1 d^1.$$

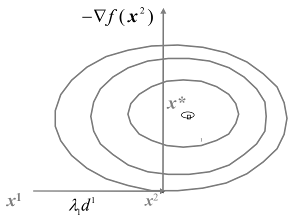
由精确一维搜索的性质，可知

$$\left(\nabla f(x^2)\right)^T d^1 = 0.$$

如果按最速下降法，选取

$d^2 = -\nabla f(x^2)$,

则将发生锯齿现象。



共轭方向法---共轭方向及其性质

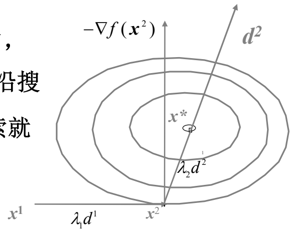
二维情况（n=2）

任取初始点 x^1，沿某个下降方向 d^1 作精确一维搜索，得

$$x^2 = x^1 + \lambda_1 d^1, \quad \left(\nabla f(x^2)\right)^T d^1 = 0.$$

如果希望下次迭代就得到最优解， d^2=?

如果能够选定这样的搜索方向，那么对于二元二次函数只需依次沿搜索方向d^1, d^2进行次精确一维搜索就可以求到极小点x^*，即



$$x^* = x^2 + \lambda_2 d^2.$$

共轭方向法---共轭方向及其性质

$$x^* = x^2 + \lambda_2 d^2.$$

x^*是f 的极小点，故x^*是f 的驻点，

$$\nabla f(x^*) = Ax^* + b = 0,$$

$$\left(\nabla f(x^2)\right)^T d^1 = 0$$

$$0 = \nabla f(x^*) = Ax^* + b = A(x^2 + \lambda_2 d^2) + b = (Ax^2 + b) + \lambda_2 Ad = \nabla f(x^2) + \lambda_2 Ad^2$$

将等式两边同时左乘 $(d^1)^T$ 得：

$$0 = \left(d^1\right)^T \nabla f(x^2) + \lambda_2 \left(d^1\right)^T Ad^2 = 0 + \lambda_2 \left(d^1\right)^T Ad^2,$$

即 $(d^1)^T Ad^2 = 0.$

两次迭代要得到二元二次函数的极小点，d^1必须满足的条件是：

$$(d^1)^T Ad^2 = 0,$$

搜索方向d^1和d^2是A共轭的。

d^1是某个下降方向

共轭方向法---共轭方向的性质

性质1 R^n中与n个线性无关的向量都正交的一定是零向量。

性质2 R^n中A共轭的向量组$p^1, p^2, \ldots, p^m$是线性无关的。

性质3 R^n中互相共轭的向量个数不超过n.

性质4 给定n元二次函数$f(x)=1/2x^TAx+b^Tx+c$, $A=A^T$正定，设n维向量组$p^1, p^2, \ldots, p^n$是A共轭向量组，从任意点x^1出发，依次以$p^1, p^2, \ldots, p^n$ 为搜索方向进行精确一维搜索，则

(1) $\nabla f(x^{k+1})$与$p^1, p^2, \ldots, p^k$ (k=1,2,…,n)正交；

(2) 最多n次迭代必达到n元二次函数f的极小点。

共轭方向法---共轭方向及其性质

性质1 R^n中与n个线性无关的向量都正交的一定是零向量。

证明：给定R^n 中的向量 $p^1, p^2, \ldots, p^n, d$，若$p^1, p^2, \ldots, p^n$线性无关，且d与 $p^1, p^2, \ldots, p^n$正交，证 d=0.

$p^1, p^2, \ldots, p^n$线性无关，故$p^1, p^2, \ldots, p^n$可作为R^n的一组基，故d可由$p^1, p^2, \ldots, p^n$线性表出，即存在一组实数$\alpha_1, \alpha_2, \ldots, \alpha_n$, $s.t.$

$$d = \alpha_1 p^1 + \alpha_2 p^2 + \ldots + \alpha_n p^n.$$

$$d^T d = d^T\left(\alpha_1 p^1 + \alpha_2 p^2 + \ldots + \alpha_n p^n\right) = \alpha_1 d^T p^1 + \alpha_2 d^T p^2 + \ldots + \alpha_n d^T p^n$$

$$= \alpha_1 \cdot 0 + \alpha_2 \cdot 0 + \ldots + \alpha_n \cdot 0 = 0,$$

故 d=0.

共轭方向法---共轭方向及其性质

性质2 R^n中A共轭的向量组$p^1,p^2,\ldots,p^m$是线性无关的。

证明 假设$\alpha_1 p^1+\alpha_2 p^2+\ldots+\alpha_m p^m=0,$ 要证$p^1,p^2,\ldots,p^m$线性无关，

只需证明$\alpha_1=\alpha_2=\ldots=\alpha_m=0.$

因为$p^1,p^2,\ldots,p^m$是A共轭向量组，两边同乘$\left(p^k\right)^T A,\forall k=1,\ldots,m,$

$$0=\left(p^k\right)^T A\left(\alpha_1 p^1+\alpha_2 p^2+\ldots+\alpha_m p^m\right)\quad \left(p^k\right)^T Ap^i=0,\forall i\neq k,$$

$$=0+0+\ldots+\alpha_k\left(p^k\right)^T Ap^k+\ldots+0$$

$$=\alpha_k\left(p^k\right)^T Ap^k,$$

因为A正定，$p^k\neq 0$，所以$\left(p^k\right)^T Ap^k>0$，故$\alpha_k=0,\forall k=1,\ldots,m,$

得证。

共轭方向法---共轭方向及其性质

性质3 R^n中互相共轭的向量个数不超过n.

证明 利用性质2即可。

性质2 R^n中A共轭的向量组$p^1,p^2,\ldots,p^m$是线性无关的。

R^n中线性无关的非零向量最多有n个。

共轭方向法---共轭方向的性质

性质4 给定n元函数$f(x)=1/2x^TAx+b^Tx+c, A=A^T$正定，设$n$维向量组$p^1,p^2,\ldots,p^n$是$A$共轭向量组，从任意点$x^1$出发，依次以$p^1,p^2,\ldots,p^n$为搜索方向进行精确一维搜索，则

(1) $\nabla f(x^{l+1})$与$p^1,p^2,\ldots,p^l$ $(l=1,2,\ldots,n)$正交；

(2) 最多n次迭代必达到n元二次函数f的极小点。

共轭方向法---共轭方向及其性质

性质4中(1) $\nabla f(x^{l+1})$与$p^1,p^2,\ldots,p^l$ $(l=1,2,\ldots,n)$正交；

证明

λ_l是按照精确一维搜索得到的，故$\left(\nabla f(x^{l+1})\right)^T p^l=0.$

下证 $\nabla f(x^{l+1})^T p^i=0,\ i=1,2,\ldots,l-1.$

$$f(x)=\frac{1}{2}x^TAx+b^Tx+c,\quad \nabla f(x)=Ax+b,$$

$$\nabla f(x^{l+1})=Ax^{l+1}+b=A(x^l+\lambda_l p^l)+b=(Ax^l+b)+\lambda_l Ap^l$$

$$=\nabla f(x^l)+\lambda_l Ap^l$$

$$\nabla f(x^{l+1})=\nabla f(x^l)+\lambda_l Ap^l=\nabla f(x^{l-1})+\lambda_{l-1}Ap^{l-1}+\lambda_l Ap^l$$

$$=\ldots$$

$$=\nabla f(x^{i+1})+\lambda_{i+1}Ap^{i+1}+\ldots+\lambda_{l-1}Ap^{l-1}+\lambda_l Ap^l,\quad i=1,2,\ldots,l-1,$$

共轭方向法---共轭方向及其性质

性质4中(1) $\nabla f(x^{l+1})$与$p^1, p^2,\ldots, p^l$ $(l=1,2,\ldots,n)$正交；

$$\nabla f(x^{l+1}) = \nabla f(x^{i+1}) + \lambda_{i+1}Ap^{i+1} + \ldots + \lambda_{l-1}Ap^{l-1} + \lambda_l Ap^l,$$

$$\left(\nabla f(x^{l+1})\right)^T p^i = \left(\nabla f(x^{i+1})\right)^T p^i + \lambda_{i+1}(Ap^{i+1})^T p^i + \ldots + \lambda_{l-1}(Ap^{l-1})^T p^i + \lambda_l (Ap^l)^T p^i$$

$$= \nabla f(x^{i+1})^T p^i + \lambda_{i+1}(p^{i+1})^T Ap^i + \ldots + \lambda_{l-1}(p^{l-1})^T Ap^i + \lambda_l (p^l)^T Ap^i$$

$$= 0,$$

故 $\nabla f(x^{l+1})$与$p^1, p^2,\ldots, p^l$ $(l=1,2,\ldots,n)$正交。

$P^1, P^2,\ldots, P^n$ 是A共轭向量组

共轭方向法---共轭方向及其性质

性质4中(2) 最多n次迭代必达到二次函数f的极小点。

证明　假设 $\nabla f(x^1), \nabla f(x^2), \ldots, \nabla f(x^n)$ 都不是0，下证

$$\nabla f(x^{n+1}) = 0.$$

利用性质4 (1)的结果，$\nabla f(x^{n+1})$与A共轭向量组$P^1, P^2,\ldots, P^n$都正交，由性质2和性质1可知，$\nabla f(x^{n+1}) = 0.$

性质2 R^n中A共轭的向量组$p^1,p^2,\ldots, p^m$是线性无关的。

性质4 (1) $\nabla f(x^{l+1})$与$p^1, p^2,\ldots, p^l$ $(l=1,2,\ldots,n)$正交；

性质1 R^n中与n个线性无关的向量都正交的一定是零向量。

共轭方向法---共轭方向的性质

性质4 给定n元函数$f(x)=1/2x^TAx+b^Tx+c$, $A=A^T$正定，设n维向量组$p^1, p^2,\ldots, p^n$是A共轭向量组，从任意点x^1出发，相继以$p^1, p^2,\ldots, p^n$为搜索方向进行精确一维搜索，则

(1) $\nabla f(x^{l+1})$与$p^1, p^2,\ldots, p^l$ $(l=1,2,\ldots,n)$正交；

(2) 最多n次迭代必达到n元二次函数f的极小点。

在迭代法中，若取搜索方向是共轭方向，得到的方法称为共轭方向法。

由性质4可知，若能找到一组A共轭向量$p^1, p^2,\ldots, p^n$，则结合最速下降法和Newton法优点的算法就找到了，就是共轭方向法。　这个算法具有二次收敛性。

怎么选取共轭方向？

定义　一个算法若能在有限步内求得正定二次函数的极小点，则称该算法具有二次收敛性(又称二次终止性)。

共轭方向的生成与共轭梯度法

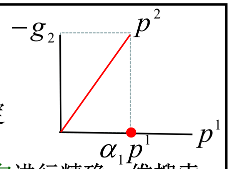
共轭方向的选取有很大任意性，共轭方向不同的选取方法对应着不同的共轭方向法。

作为一种迭代算法，自然希望共轭方向能在迭代过程中逐次生成。

下面先以正定二次函数为例，介绍一种产生共轭方向的方法，再将这种方法推广到非二次函数上。

这种方法中的每一个共轭向量都依赖于迭代点处的负梯度，因此称为共轭梯度法，它是共轭方向法中的一种。

共轭梯度法



令$f(x)=\frac{1}{2}x^T Ax+b^T x+c,\ A=A^T$ 正定

(1) 从任取初始点$\boldsymbol{x}^1$ 出发，沿负梯度方向进行精确一维搜索：

$$p^1=-\nabla f(x^1),\qquad x^2=x^1+\lambda_1 p^1,$$

(2) 若$\nabla f(x^2)=0$，停止，否则在$-\nabla f(x^2)$ 和 p^1张成的

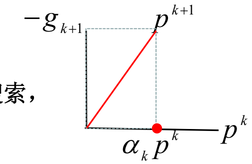
正交锥中找一个向量p^2,即令 $p^2=-\nabla f(x^2)+\alpha_1 p^1$

使得 p^1与p^2共轭，即$(p^2)^T Ap^1=0$.　　$\alpha_1=?$

由$(p^2)^T Ap^1=0=(-\nabla f(x^2)+\alpha_1 p^1)^T Ap^1$,可得

$$\alpha_1=\frac{\nabla f(x^2)^T Ap^1}{(p^1)^T Ap^1}$$

共轭梯度法



(3) 在$\boldsymbol{x}^2$ 处沿 $\boldsymbol{p}^2$方向进行精确一维搜索，

$$x^3=x^2+\lambda_2 p^2,$$

(4) 以此类推，$x^4, x^5, \ldots, x^k, x^{k+1}$,

(5) 若 $\nabla f(x^{k+1})=0$ ，停止，否则在$-\nabla f(x^{k+1})$ 和 p^k 张成的正交锥中找一个向量p^{k+1}，即令$p^{k+1}=-\nabla f(x^{k+1})+\alpha_k p^k$

使得p^{k+1}与 p^k 共轭，即 $(p^{k+1})^T Ap^k=0$.　　$\alpha_k=?$

由 $(p^{k+1})^T Ap^k=0=(-\nabla f(x^{k+1})+\alpha_k p^k)^T Ap^k$ 可得

$$\alpha_k=\frac{\nabla f(x^{k+1})^T Ap^k}{(p^k)^T Ap^k}$$

共轭梯度法

如此便构造了一组向量$p^1, p^2, \ldots, p^n$，

$$p^1=-\nabla f(x^1),$$　　相邻两个向量是共轭的

$$p^{k+1}=-\nabla f(x^{k+1})+\alpha_k p^k, k=1,2,\ldots,n-1,$$

$$\alpha_k=\frac{\nabla f(x^{k+1})^T Ap^k}{(p^k)^T Ap^k},$$

实际上，这组向量$p^1, p^2, \ldots, p^n$是一个$\boldsymbol{A}$共轭向量组。

共轭梯度法

定理1 设向量组 $p^1, p^2, \ldots, p^n$是由上述方法产生的向量组，向量组$g_1, g_2, \ldots, g_n$ 是由各点的梯度生成的向量组（$g_k=\nabla f(x^k)$），则

(1) $g_1, g_2, \ldots, g_n$ 是正交向量组；

(2) $p^1, p^2, \ldots, p^n$是$\boldsymbol{A}$共轭向量组。

共轭梯度法

定理1 设向量组 $p^1, p^2, \ldots, p^n$是由上述方法产生的向量组，向量组 $g_1, g_2, \ldots, g_n$是由各点的梯度生成的向量组（$g_k = \nabla f(x^k)$），则

(1) $g_1, g_2, \ldots, g_n$是正交向量组；

(2) $p^1, p^2, \ldots, p^n$是A共轭向量组。

证明 归纳法：

(i) n=2 由 p^i 的构造过程知，p^1, p^2是A共轭的，即

$\left(p^2\right)^T Ap^1$=0, 结论(2)成立；

利用精确一维搜索的性质知，$\left(g_2\right)^T p^1$=0，而$p^1 = -g_1$，

故 $\left(g_2\right)^T g_1$=0, 结论(1)成立。

(ii) 假设$n=k$时，结论(1)(2)成立，下证$n=k+1$时结论仍成立.

由假设可知，要证明$n=k+1$时结论成立，只需证明

g_{k+1} 与 $g_1, g_2, \ldots, g_k$ 正交，p^{k+1} 与 $p^1, p^2, \ldots, p^k$ A共轭。

(a) 证明 g_{k+1} 与 $g_1, g_2, \ldots, g_k$ 正交；

因为
$$p^i = \begin{cases} -g_i, & i = 1, \\ -g_i + \alpha_{i-1} p^{i-1}, & i = 2, \ldots, n, \end{cases}$$

所以
$$\boxed{\left(g_{k+1}\right)^T g_i} = \begin{cases} \left(g_{k+1}\right)^T \left(-p^i\right), & i = 1, \\ \left(g_{k+1}\right)^T \left(-p^i + \alpha_{i-1} p^{i-1}\right), & i = 2, \ldots, k, \end{cases} \quad \boxed{=0,}$$

$p^1, p^2, \ldots, p^k$是A共轭向量组，利用性质4(1)可知，

$\left(g_{k+1}\right)^T p^i = 0, i = 1, 2, \ldots, k,$

结论(a)成立，进而结论(1)成立。

性质4 给定n元函数$f(x)=1/2x^TAx+b^Tx+c$, $A=A^T$正定，设n维非零向量组$p^1, p^2, \ldots, p^n$是A共轭向量组，从任意点x^1出发，相继以$p^1, p^2, \ldots, p^n$为搜索方向进行精确一维搜索，则

(1) $\nabla f(x^{i+1})$与$p^1, p^2, \ldots, p^i$ $(i=1,2,\ldots,n)$正交；

(b) 证明 p^{k+1} 与 $p^1, p^2, \ldots, p^k$ 是A共轭的；

由 p^i 的构造过程知，p^{k+1} 与 p^k是A共轭的；

下证 p^{k+1} 与 $p^1, p^2, \ldots, p^{k-1}$是$A$共轭的；

$$\left(p^{k+1}\right)^T Ap^i = \left(-g_{k+1} + \alpha_k p^k\right)^T Ap^i \quad (i=1,2,\ldots,k-1)$$
$$= -\left(g_{k+1}\right)^T Ap^i \qquad \left(p^k\right)^T Ap^i = 0，i=1,2,\ldots,k-1$$

$$g_{i+1} - g_i = Ax^{i+1} + b - g_i = A(x^i + \lambda_i p^i) + b - g_i = \lambda_i Ap^i$$

$$\therefore \left(p^{k+1}\right)^T Ap^i = -\left(g_{k+1}\right)^T Ap^i = -\left(g_{k+1}\right)^T \frac{g_{i+1} - g_i}{\lambda_i}$$
$$= -1/\lambda_i \left(g_{k+1}\right)^T \left(g_{i+1} - g_i\right) \qquad \left(g_{k+1}\right)^T g_i = 0，i=1,2,\ldots,k$$
$$=0$$

结论(b)成立，进而结论(2)成立。

共轭梯度法

定理1 设向量组 $p^1, p^2, \ldots, p^n$是由上述方法产生的向量组，

向量组 $g_1, g_2, \ldots, g_n$ 是由各点的梯度生成的向量组

（$g_k = \nabla f(x^k)$），则

(1) $g_1, g_2, \ldots, g_n$ 是正交向量组；

(2) $p^1, p^2, \ldots, p^n$是A共轭向量组。

注 为保证方向的共轭性，初始方向取负梯度方向。

共轭梯度法--算例 初始方向为下降方向，但不是负梯度方向

$p^1 \neq -g_1$,

例 用共轭方向法求下列问题

$$\min f(x) = x_1^2 + \frac{x_2^2}{2} + \frac{x_3^2}{2}，\text{已知初始点 } x^1 = \begin{pmatrix}1\\1\\1\end{pmatrix}, \quad p^1 = \begin{pmatrix}-1\\-2\\0\end{pmatrix}.$$

解：第一次迭代：沿p^1方向进行精确一维搜索，得$\lambda_1 = 2/3$,

$$x^2 = x^1 + \lambda_1 p^1 = \left(\frac{1}{3}, -\frac{1}{3}, 1\right)^T, g_2 = \nabla f(x^2) = \left(\frac{2}{3}, -\frac{1}{3}, 1\right)^T,$$

第二次迭代：

$$g_1 = (2,1,1)^T,$$

$$\alpha_1 = \frac{g_2^T A p^1}{(p^1)^T A p^1} = -\left(\frac{2}{3}\right) \Big/ 6 = -\frac{1}{9}, \quad p^1 \neq -g_1,$$

$$p^2 = -g_2 + \alpha_1 p^1 = \left(-\frac{5}{9}, \frac{5}{9}, -1\right)^T.$$

共轭梯度法--算例

沿p^2方向进行精确一维搜索，得 $\lambda_2 = 21/26$,

$$x^3 = x^2 + \lambda_2 p^2 = \left(-\frac{3}{26}, \frac{3}{26}, \frac{5}{26}\right)^T, g_3 = \left(-\frac{3}{13}, \frac{3}{26}, \frac{5}{26}\right)^T,$$

第三次迭代：

$$\alpha_2 = \frac{g_3^T A p^2}{(p^2)^T A p^2} = \frac{45}{676},$$

$$p^3 = -g_3 + \alpha_2 p^2 = \frac{1}{676}(131, -53, -175)^T.$$

沿p^3方向进行精确一维搜索，得λ_3=910/1303，

$$x^4 = x^3 + \lambda_3 p^3 = \frac{1}{1303}(26, 79, 15)^T \neq (0,0,0)^T,$$

3次迭代没有到达3元二次函数的极小点$(0,0,0)^T$,为什么？

共轭梯度法--算例 p^1，p^2与p^3是不是A共轭向量组?

$$(p^1)^T A p^2 = (-1,-2,0)\,diag(2,1,1)\left(-\frac{5}{9}, \frac{5}{9}, -1\right)^T = 0,$$

$$f(\boldsymbol{x}) = x_1^2 + \frac{1}{2}x_2^2 + \frac{1}{2}x_3^2$$

$$(p^3)^T A p^2 = \frac{1}{676}(131,-53,-175)\,diag(2,1,1)\left(-\frac{5}{9}, \frac{5}{9}, -1\right)^T = 0,$$

$$A = \nabla^2 f(\boldsymbol{x}) = diag(2,1,1)$$

$$(p^3)^T A p^1 = \frac{1}{676}(131,-53,-175)\,diag(2,1,1)(-1,-2,0)^T = -\frac{3}{13},$$

$$p^1 = (-1,-2,0)^T, \quad p^2 = \left(-\frac{5}{9}, \frac{5}{9}, -1\right)^T,$$

$$p^3 = \frac{1}{676}(131,-53,-175)^T, \quad A = diag(2,1,1)$$

p^1，p^2与p^3不是A共轭向量组。

原因是初始方向不是负梯度方向。

共轭梯度法

每一个搜索方向都依赖迭代点处的负梯度，对应的算法称为共轭梯度法。

$$\begin{cases} p^1 = -\nabla f(x^1), \\ p^{k+1} = -\nabla f(x^{k+1}) + \alpha_k p^k, k = 1,2,\ldots,n-1, \\ \qquad \alpha_k = \dfrac{\nabla f(x^{k+1})^T A p^k}{(p^k)^T A p^k}. \end{cases} \quad (*)$$

$p^1, p^2, \ldots, p^n$ 是一个A共轭向量组

性质4 给定n元函数$f(x)=1/2x^TAx+b^Tx+c$, $A=A^T$正定，设n维非零向量组$p^1, p^2,\ldots, p^n$是A共轭向量组，从任意点x^1出发，相继以$p^1, p^2,\ldots, p^n$ 为搜索方向进行精确一维搜索，则

(1) $\nabla f(x^{l+1})$与$p^1, p^2,\ldots, p^l$ (l=1,2,…,n)正交；

(2) 最多n次迭代必达到正定二次函数f的极小点。

共轭梯度法

针对“$f(x)=1/2x^TAx+b^Tx+c$, $A=A^T$正定，最多n次迭代达到极小点”找到了一组共轭方向：

$$
\begin{cases}
p^1=-\nabla f(x^1),\\
p^{k+1}=-\nabla f(x^{k+1})+\alpha_k p^k,\ k=1,2,\dots,n-1,\\
\alpha_k=\dfrac{(\nabla f(x^{k+1}))^T Ap^k}{(p^k)^T Ap^k}.
\end{cases}
\quad (*)
$$

在正定二次函数的前提下，将α_k变形

针对一般函数，将这组方向进行推广，

直接对(*)式推广：$A\to\nabla^2 f(x^{k+1})$

能否将α_k中的A去掉？

存在问题：计算量、存储量都很大

怎么解决呢？

共轭梯度法

定理2 设 $f(x)=\frac{1}{2}x^TAx+b^Tx+c$, $A=A^T$正定，$p^1,p^2,\dots,p^n$ 是由上述方法构造的A共轭向量组，$g_k=\nabla f(x^k)$，利用前面所得的公式，得到几个等价的计算公式：

(1) $\alpha_k=\dfrac{(\nabla f(x^{k+1}))^T Ap^k}{(p^k)^T Ap^k}=\dfrac{(g_{k+1})^T Ap^k}{(p^k)^T Ap^k}$ (Daniel,1967)

$g_{k+1}-g_k=Ax^{k+1}+b-g_k=A(x^k+\lambda_k p^k)+b-g_k=\lambda_k Ap^k$

(2) $\alpha_k=\dfrac{(g_{k+1})^T(g_{k+1}-g_k)}{(p^k)^T(g_{k+1}-g_k)}$ (Sorenson−Wolfe,1972)

利用定理1，可知$g_1,g_2,\dots,g_n$是正交向量组，因此(2)中$g_{k+1}^Tg_k=0$；并且$g_{k+1}^Tp^k=0$.

(3) $\alpha_k=-\dfrac{(g_{k+1})^T g_{k+1}}{(p^k)^T g_k}$ (Dixon−Myers,1972)

共轭梯度法

(3) $\alpha_k=-\dfrac{(g_{k+1})^T g_{k+1}}{(p^k)^T g_k}$ (Dixon−Myers,1972)

在(3)中，由$p^k=-g_k+\alpha_{k-1}p^{k-1}$，得到$(p^k)^Tg_k=-(-g_k+\alpha_{k-1}p^{k-1})^Tg_k=-\|g_k\|^2$

(4) $\alpha_k=\dfrac{\|g_{k+1}\|^2}{\|g_k\|^2}$ (Flecher−Reeves,1964)

(2)中$(p^k)^T(g_{k+1}-g_k)=-(p^k)^Tg_k=-(-g_k+\alpha_{k-1}p^{k-1})^Tg_k=(g_k)^Tg_k$

(5) $\alpha_k=\dfrac{(g_{k+1})^T(g_{k+1}-g_k)}{(g_k)^T g_k}$ (Polyak−Polak−Ribiere,1969)

---

(1) $\alpha_k=\dfrac{(\nabla f(x^{k+1}))^T Ap^k}{(p^k)^T Ap^k}=\dfrac{(g_{k+1})^T Ap^k}{(p^k)^T Ap^k}$ (Daniel,1967)
Daniel共轭梯度法

(2) $\alpha_k=\dfrac{(g_{k+1})^T(g_{k+1}-g_k)}{(p^k)^T(g_{k+1}-g_k)}$ (Sorenson−Wolfe,1972)
SW共轭梯度法

(3) $\alpha_k=-\dfrac{(g_{k+1})^T g_{k+1}}{(p^k)^T g_k}$ (Dixon−Myers,1972)
DM共轭梯度法

(4) $\alpha_k=\dfrac{\|g_{k+1}\|^2}{\|g_k\|^2}$ (Flecher−Reeves,1964)介绍FR共轭梯度法
FR共轭梯度法 简洁，用的最多

(5) $\alpha_k=\dfrac{(g_{k+1})^T(g_{k+1}-g_k)}{(g_k)^T g_k}$ (Polyak−Polak−Ribiere,1969)
PPR共轭梯度法 更好用

对于正定二次函数，上面得到的5个计算公式是等价的；

这5种共轭梯度法也是完全等价的；

对于非二次函数，产生的搜索方向不再相同，

常利用公式(2)-(5)，通常不用公式 (1) ((1)中含有Hesse矩阵).

*FR*共轭梯度法--算法步骤

步骤1 选定初始点x^1.

步骤2 如果$\|g_1\|\le\varepsilon$,停止，得到近似驻点x^1,否则转步骤3。

步骤3 取$p^1=-g_1$,k=1.

步骤4 精确一维搜索找最佳步长λ_k，令 $x^{k+1}=x^k+\lambda_k p^k$.

步骤5 如果$\|g_{k+1}\|\le\varepsilon$,停止，得到近似驻点$x^{k+1}$,否则转步骤6。

步骤6 如果***k=n***, 令x^1=x^{k+1},p^1=$-g_{k+1}$, ***k*=1**， 转步骤4;

否则转步骤7。

步骤7 计算$\alpha_k=\dfrac{\|g_{k+1}\|^2}{\|g_k\|^2}$,$p^{k+1}=-g_{k+1}+\alpha_k p^k$, $k=k+1$, 转步骤4。

*FR*共轭梯度法--算法步骤

步骤4 精确一维搜索找最佳步长λ_k，令 $x^{k+1}=x^k+\lambda_k p^k$.

步骤5 如果$\|g_{k+1}\|\le\varepsilon$,停止，得到近似驻点 x^{k+1},否则转步骤6。

步骤6 如果***k=n***, 令x^1=x^{k+1},p^1=$-g_{k+1}$, ***k*=1**， 转步骤4;

否则转步骤7。

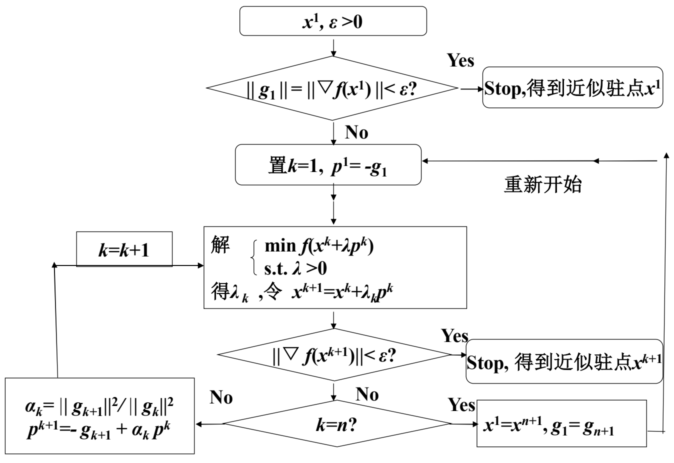
步骤7 计算$\alpha_k=\dfrac{\|g_{k+1}\|^2}{\|g_k\|^2}$,$p^{k+1}=-g_{k+1}+\alpha_k p^k$, $k=k+1$, 转步骤4。

误差可能会使***n***步迭代得不到正定二次函数的极小点。

R^n中共轭方向最多有***n***个，***n***步后构造的搜索方向不再是共轭的， 会降低收敛速度，

步骤6：重新开始技术：x^{n+1}作为新的x^1

*FR*共轭梯度法--算法框图



*FR*共轭梯度法--算例

例 用***FR***共轭梯度法求$\min f(x)=x_1^2+2x_2^2-4x_1-2x_1x_2$,

取 $x^1=(1,1)^T$,ε=10^{-2}. $x^*=(4,2)^T$.

解： 1）第一次迭代：沿负梯度方向搜寻

计算初始点处的梯度

$$g_1=\nabla f(x^1)=\begin{pmatrix}2x_1-2x_2-4\\4x_2-2x_1\end{pmatrix}\Bigg|_{x=x^1}=\begin{pmatrix}-4\\2\end{pmatrix},$$

$$p^1=-g_1=\begin{pmatrix}4\\-2\end{pmatrix}\qquad x^1+\lambda p^1=\begin{pmatrix}1\\1\end{pmatrix}+\lambda\begin{pmatrix}4\\-2\end{pmatrix}=\begin{pmatrix}1+4\lambda\\1-2\lambda\end{pmatrix}$$

FR共轭梯度法--算例

$$x^1+\lambda p^1=\begin{pmatrix}1+4\lambda\\1-2\lambda\end{pmatrix}$$

精确一维搜索求最佳步长，

$$\phi(\lambda)=f(x^1+\lambda p^1)=f(1+4\lambda,1-2\lambda)=40\lambda^2-20\lambda-3,$$

令 $0=\phi'(\lambda)=80\lambda-20,$

得 $\lambda_1=\frac{1}{4},\quad x^2=x^1+\lambda_1 p^1=\begin{pmatrix}2\\\frac{1}{2}\end{pmatrix},$

$$g_2=\nabla f(x^2)=\begin{pmatrix}-1\\-2\end{pmatrix},\quad \|g_2\|=\sqrt{5}>\varepsilon,$$

继续迭代；

FR共轭梯度法--算例

$$g_1=\begin{pmatrix}-4\\2\end{pmatrix},\quad g_2=\begin{pmatrix}-1\\-2\end{pmatrix},$$

2）第二次迭代

$$p^1=\begin{pmatrix}4\\-2\end{pmatrix},\quad x^2=\begin{pmatrix}2\\\frac{1}{2}\end{pmatrix},$$

$$\alpha_1=\frac{\|g_2\|^2}{\|g_1\|^2}=\frac{5}{20}=\frac{1}{4},$$

$$p^2=-g_2+\alpha_1 p^1=\begin{pmatrix}2\\\frac{3}{2}\end{pmatrix},\quad x^2+\lambda p^2=\begin{pmatrix}2\\\frac{1}{2}\end{pmatrix}+\lambda\begin{pmatrix}2\\\frac{3}{2}\end{pmatrix}=\begin{pmatrix}2+2\lambda\\\frac{1}{2}+\frac{3}{2}\lambda\end{pmatrix},$$

精确一维搜索求最佳步长，

$$\phi(\lambda)=f(x^2+\lambda p^2)=f(2+2\lambda,\frac{1}{2}+\frac{3}{2}\lambda)$$

$$=(2+2\lambda)^2+2(\frac{1}{2}+\frac{3}{2}\lambda)^2-2(2+2\lambda)(\frac{1}{2}+\frac{3}{2}\lambda)-4(2+2\lambda)$$

FR共轭梯度法--算例

$$\phi(\lambda)==(2+2\lambda)^2+2(\frac{1}{2}+\frac{3}{2}\lambda)^2-2(2+2\lambda)(\frac{1}{2}+\frac{3}{2}\lambda)-4(2+2\lambda)$$

令 $0=\phi'(\lambda),$

$$x^2+\lambda p^2=\begin{pmatrix}2+2\lambda\\\frac{1}{2}+\frac{3}{2}\lambda\end{pmatrix}$$

得 $\lambda_2=1,$

$$x^3=x^2+\lambda_2 p^2=\begin{pmatrix}4\\2\end{pmatrix},\quad g_3=\nabla f(x^3)=\begin{pmatrix}0\\0\end{pmatrix},$$

因 $\|g_3\|=0<\varepsilon$，算法终止，又该问题是凸规划，

故得到极小点 $x^*=x^3=\begin{pmatrix}4\\2\end{pmatrix}.$

FR共轭梯度法

共轭梯度法

- 具有二次终止性：从任意初始点出发，最多n步迭代达到n元正定二次函数的极小点；
- 具有较高的求解效率：共轭梯度法在一定条件下收敛，且收敛速度通常优于最速下降法；
- 对非二次函数，共轭梯度法产生的方向不再是共轭方向(目标函数的**Hesse**矩阵不再是常数矩阵)。

*FR*共轭梯度法

收敛性定理：

设凸函数 *f* 存在一阶连续偏导数，水平集 $L=\{x|f(x)\le f(x^1)\}$ 有界，则由共轭梯度法得到的无穷点列 $\{x^k\}$ 具有如下性质：

(1) $\{f(x^k)\}$ 为严格单调下降序列，且 $\lim_{k\to\infty} f(x^k)$ 存在；

(2) $\{x^k\}$ 的任意聚点 x^* 都是 *f* 的极小点。　特点：全局收敛

共轭梯度法在无约束优化方法中占有重要的地位，是目前最常用的方法之一。

共轭梯度法--算法特点

(1) 求解凸函数的极小点时，算法全局收敛；

(2) 适用于大规模问题

计算量小，算法的计算公式简单，不用求Hesse矩阵或者逆矩阵；

存储量小，每次迭代只需存储若干向量。

(3) 具有二次终止性

对于 *n* 元正定二次函数，至多 *n* 次迭代可达到极小点。

共轭梯度法--算法特点

(4) 共轭梯度法的收敛速率不坏于最速下降法(Crowder 和Wolfe证明)

如果初始方向不用负梯度方向，算法线性收敛；

(5) 共轭梯度法是目前求解无约束优化问题最常用的方法之一。

注：α_k 的不同公式，对于正定二次函数来说是等价的，

对于非正定二次函数，有不同的效果，

经验上 *PPR* 效果较好。

作业：

P 100　4.9, 4.10, 4.12, 4.13, 4.14, 4.17--4.19

变尺度法----一类特殊的拟Newton法

Newton法和阻尼Newton法收敛速度快，但需要计算Hesse矩阵的逆，$d^k=-\left(\nabla^2 f\left(x^k\right)\right)^{-1}\nabla f\left(x^k\right)$，计算量大，存储量也很大。

为减少计算量，用一个 *n* 阶对称正定矩阵 H_k 近似代替Hesse矩阵的逆 $\left(\nabla^2 f\left(x^k\right)\right)^{-1}$，即 $H_k\approx\left(\nabla^2 f\left(x^k\right)\right)^{-1}$，搜索方向是 $p^k=-H_k g_k$，由此产生的方法称为变尺度法，H_k 称为尺度矩阵，这是一种拟Newton法，是无约束优化中最有效的方法之一。

所谓拟Newton法是指由Newton法的思想出发产生的一类方法。

变尺度法----算法步骤

步骤1　任取初始点x^1,初始尺度矩阵H_1,令k=1.

步骤2　计算 $p^k=-H_k g_k$.

步骤3　利用精确一维搜索找最佳步长 λ_k,

$$x^{k+1}=x^k+\lambda_k p^k.$$

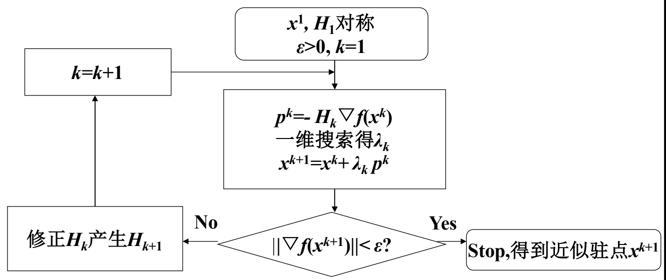
步骤4　如果 $\left\|\nabla f\left(x^{k+1}\right)\right\|\le\varepsilon$,停止，得到近似驻点$x^{k+1}$,否则转步骤5.

步骤5　令 $H_{k+1}=H_k+\Delta H_k, k=k+1$转步骤2.

其中 ΔH_k 称为修正矩阵。

不同的修正矩阵，对应着不同的变尺度法。

变尺度法--算法框架



变尺度法----一类特殊的拟Newton法

构造 H_k的原则

变尺度法的关键-----如何构造H_k,

为使算法有较快的收敛速度，H_k需满足几个原则：

拟牛顿性质，

二次终止性，

稳定性。

变尺度法----一类特殊的拟Newton法

构造 H_k的原则

拟牛顿性质

在 x^{k+1} 点处对函数 f 作二阶泰勒展开：

$$f(x)=f(x^{k+1})+\left(\nabla f\left(x^{k+1}\right)\right)^T\left(x-x^{k+1}\right)$$

$$+\frac{1}{2}\left(x-x^{k+1}\right)^T\nabla^2 f\left(x^{k+1}\right)\left(x-x^{k+1}\right)+o\left(\left\|x-x^{k+1}\right\|^2\right),$$

略去高阶项，

$$f(x)\approx f(x^{k+1})+\left(\nabla f\left(x^{k+1}\right)\right)^T\left(x-x^{k+1}\right)+\frac{1}{2}\left(x-x^{k+1}\right)^T\nabla^2 f\left(x^{k+1}\right)\left(x-x^{k+1}\right),$$

变尺度法----一类特殊的拟Newton法

$$f(x)\approx f(x^{k+1})+\left(\nabla f\left(x^{k+1}\right)\right)^T\left(x-x^{k+1}\right)+\frac{1}{2}\left(x-x^{k+1}\right)^T\nabla^2 f\left(x^{k+1}\right)\left(x-x^{k+1}\right),$$

在x^{k+1}附近，上式两端求导，得

$$\nabla f(x)\approx\nabla f\left(x^{k+1}\right)+\nabla^2 f\left(x^{k+1}\right)\left(x-x^{k+1}\right),$$

令 $x=x^k, g_k=\nabla f\left(x^k\right)$,可得 $g_k\approx g_{k+1}+\nabla^2 f\left(x^{k+1}\right)\left(x^k-x^{k+1}\right)$,

$$g_{k+1}-g_k\approx\nabla^2 f\left(x^{k+1}\right)\left(x^{k+1}-x^k\right),$$

记 $\Delta g_k=g_{k+1}-g_k, \Delta x_k=x^{k+1}-x^k$,

$$\Delta g_k\approx\nabla^2 f\left(x^{k+1}\right)\Delta x_k,$$

变尺度法----一类特殊的拟Newton法

构造 H_k的原则----拟牛顿性质

$$\Delta g_k\approx\nabla^2 f\left(x^{k+1}\right)\Delta x_k,$$

Hesse矩阵$\nabla^2 f\left(x^{k+1}\right)$对称正定，则 $H_k\approx\left(\nabla^2 f\left(x^k\right)\right)^{-1}$

$$\Delta x_k\approx\left(\nabla^2 f\left(x^{k+1}\right)\right)^{-1}\Delta g_k,$$

我们构造的 H_{k+1}使之满足与上式类似的等式，即

$$\boxed{\Delta x_k=H_{k+1}\Delta g_k,\qquad (1)}$$

拟牛顿性质（或条件或方程）

H_{k+1} 不唯一。

变尺度法----一类特殊的拟Newton法

构造 H_k的原则----二次终止性

$f(x)=1/2x^T Ax+b^T x+c$, A对称正定

把算法用于n元正定二次函数时，至多n次达到极小点。

> 构造的搜索方向 $p^1, p^2, \ldots, p^n$是一组A 共轭向量组且$H_{n+1}=A^{-1}$.

变尺度法----一类特殊的拟Newton法

构造 H_k的原则----稳定性

若忽略计算过程的舍入误差，任意给定k，在算法的迭代点x^k处可选择步长，使得下个迭代点x^{k+1}处的函数值下降，则称此算法是稳定的。

若$p^k=-H_k g_k$是f在点x^k处的下降方向，则总能找到一个充分小的正数作为步长，使得x^{k+1}处的函数值下降，从而算法是稳定的。

$p^k=-H_k g_k$是下降方向，可保证算法是稳定的。

变尺度法----一类特殊的拟Newton法

构造 H_k的原则----稳定性

$p^k = -H_k g_k$是下降方向，可保证算法是稳定的。

在什么条件下$p^k = -H_k g_k$是下降方向？

即在什么条件下，能使得

$$\left(\nabla f(x^k)\right)^T p^k = (g_k)^T p^k = -(g_k)^T H_k g_k < 0?$$

H_k对称正定

变尺度法----一类特殊的拟Newton法

H_k构造的要求

依据三个原则：

拟牛顿性质、二次收敛性和算法稳定性，

H_k满足下列条件即可：

拟牛顿性质 $H_{k+1}\Delta g_k = \Delta x_k$、

对称正定、

且当f是n元正定二次函数时，$p^1, p^2, \ldots, p^n$是共轭向量组，其中$p^k = -H_k g_k, k=1,\ldots,n.$

变尺度法----一类特殊的拟Newton法

H_k的构造策略

构造对称正定的尺度矩阵H_k的一般策略是：

(1) H_1取为任意一个n阶对称正定矩阵，

通常选取为n阶单位矩阵 I;

(2) 然后通过修正H_k, 给出 H_{k+1}，令

$$H_{k+1} = H_k + \Delta H_k$$

构造不同的H_k，也就是构造不同的修正矩阵ΔH_k，得到不同的变尺度法。

下面介绍*DFP*变尺度法中H_k的构造过程。

DFP算法 (*Davidon*(1959), *Fletcher and Powell* (1963))

DFP算法中H_k的构造方法

构造H_{k+1}的三原则：

拟牛顿性质、二次终止性和稳定性，

DFP算法构造H_k的过程，首先要求H_k

对称正定,满足拟牛顿性质 $H_{k+1}\Delta g_k = \Delta x_k$.

为保证H_{k+1}的对称性，令

$$H_{k+1} = H_k + \alpha_k u^k (u^k)^T + \beta_k v^k (v^k)^T,$$

α_k, β_k是常数，u^k, v^k是n维列向量。

H_1是对称正定的矩阵。

DFP算法 (*Davidon*(1959), *Fletcher and Powell* (1963))

DFP算法中H_k的构造方法

DFP算法构造的H_k: 对称正定满足拟牛顿性质

$$H_{k+1}\Delta g_k = \Delta x_k$$

$$H_{k+1} = H_k + \alpha_k u^k (u^k)^T + \beta_k v^k (v^k)^T$$

α_k, β_k是常数，u^k, v^k是n维列向量。

H_1是对称正定的矩阵。

H_{k+1}满足拟牛顿性质，$H_{k+1}\Delta g_k = \Delta x_k$,

$$\alpha_k u^k (u^k)^T \Delta g_k + \beta_k v^k (v^k)^T \Delta g_k = \Delta x_k - H_k \Delta g_k,$$

DFP算法 (*Davidon(1959), Fletcher and Powell* (1963))

DFP算法中H_k的构造方法

满足 $\alpha_k u^k (u^k)^T \Delta g_k + \beta_k v^k (v^k)^T \Delta g_k = \Delta x_k - H_k \Delta g_k$ 的u^k和 v^k并不唯一。

简单的，取 $\alpha_k u^k (u^k)^T \Delta g_k = \Delta x_k$， $\beta_k v^k (v^k)^T \Delta g_k = -H_k \Delta g_k$，

即 $\alpha_k (u^k)^T \Delta g_k u^k = \Delta x_k$， $\beta_k (v^k)^T \Delta g_k v^k = -H_k \Delta g_k$，

令 $\alpha_k (u^k)^T \Delta g_k = 1$, $u^k = \Delta x_k$, $\beta_k (v^k)^T \Delta g_k = -1$, $v^k = H_k \Delta g_k$,

$$\alpha_k = \frac{1}{(\Delta x_k)^T \Delta g_k}, \qquad \beta_k = -\frac{1}{(\Delta g_k)^T H_k \Delta g_k}$$

DFP算法

DFP算法中H_k的构造方法

$$H_{k+1} = H_k + \alpha_k u^k \left(u^k\right)^T + \beta_k v^k \left(v^k\right)^T \text{ 中}$$

$$u^k = \Delta x_k, v^k = H_k \Delta g_k,$$

$$\alpha_k = \frac{1}{(\Delta x_k)^T \Delta g_k}, \beta_k = -\frac{1}{(\Delta g_k)^T H_k \Delta g_k},$$

故

$$H_{k+1} = H_k + \frac{\Delta x_k (\Delta x_k)^T}{(\Delta x_k)^T \Delta g_k} - \frac{H_k \Delta g_k (\Delta g_k)^T H_k}{(\Delta g_k)^T H_k \Delta g_k}$$

-------DFP公式

DFP算法

定理1 若 $g_i \neq 0, i = 1,2,\ldots,n$, 则DFP方法构造的矩阵$H_i$ (i=1,2,…,n)为对称正定矩阵。

证明 数学归纳法（Schwartz不等式，正定矩阵的性质，精确一维搜索的结果。）

参见P92定理4.12的证明。

注：此时搜索方向$p^k = -H_k g_k$一定是下降方向。

DFP算法

定理1 若 $g_i \neq 0, i=1,2,\dots,n$, 则**DFP**方法构造的矩阵$\boldsymbol{H_k}$ **(k=1,2,…,n)**为对称正定矩阵。

证明 数学归纳法

当 k=1 时，H_1=I是对称正定矩阵；

假设 H_k 是对称正定矩阵，证明H_{k+1}是对称正定的。

因为H_k 是对称正定矩阵，则由**DFP**公式可知

$$H_{k+1}=H_k+\frac{\Delta x_k(\Delta x_k)^T}{(\Delta x_k)^T\Delta g_k}-\frac{H_k\Delta g_k(\Delta g_k)^T H_k}{(\Delta g_k)^T H_k\Delta g_k}$$

H_{k+1}是对称的，$(H_{k+1})^T=H_{k+1}$,

下证 H_{k+1}是正定的：

$$H_{k+1}=H_k+\frac{\Delta x_k(\Delta x_k)^T}{(\Delta x_k)^T\Delta g_k}-\frac{H_k\Delta g_k(\Delta g_k)^T H_k}{(\Delta g_k)^T H_k\Delta g_k}$$

下证 H_{k+1}是正定的： 任意给定非零$\boldsymbol{x}$

$$0\overset{?}{<} x^T H_{k+1}x=x^T H_k x+\frac{x^T\Delta x_k(\Delta x_k)^T x}{(\Delta x_k)^T\Delta g_k}-\frac{x^T H_k\Delta g_k(\Delta g_k)^T H_k x}{(\Delta g_k)^T H_k\Delta g_k}$$

$$=\frac{x^T H_k x\left((\Delta g_k)^T H_k\Delta g_k\right)}{(\Delta g_k)^T H_k\Delta g_k}-\frac{x^T H_k\Delta g_k(\Delta g_k)^T H_k x}{(\Delta g_k)^T H_k\Delta g_k}+\frac{x^T\Delta x_k(\Delta x_k)^T x}{(\Delta x_k)^T\Delta g_k}$$

$$=\frac{(x^T H_k x)\left((\Delta g_k)^T H_k\Delta g_k\right)-(x^T H_k\Delta g_k)^2}{(\Delta g_k)^T H_k\Delta g_k}+\frac{(x^T\Delta x_k)^2}{(\Delta x_k)^T\Delta g_k}$$

$$=(1)+(2)$$

$$x^T H_{k+1}x=\frac{(x^T H_k x)\left((\Delta g_k)^T H_k\Delta g_k\right)-(x^T H_k\Delta g_k)^2}{(\Delta g_k)^T H_k\Delta g_k}+\frac{(x^T\Delta x_k)^2}{(\Delta x_k)^T\Delta g_k}$$

$$=(1)+(2)$$

先证 $(1)=\dfrac{(x^T H_k x)\left((\Delta g_k)^T H_k\Delta g_k\right)-(x^T H_k\Delta g_k)^2}{(\Delta g_k)^T H_k\Delta g_k}\geq 0$

再证 $(2)=\dfrac{(x^T\Delta x_k)^2}{(\Delta x_k)^T\Delta g_k}\geq 0$,

最后证$x^T H_{k+1}x=(1)+(2)>0$.

先证 $(1)=\dfrac{(x^T H_k x)\left((\Delta g_k)^T H_k\Delta g_k\right)-(x^T H_k\Delta g_k)^2}{(\Delta g_k)^T H_k\Delta g_k}\geq 0$

因为 H_k 是正定的,所以存在可逆矩阵**B**,使得 $H_k=B^TB$,

$$(x^T H_k\Delta g_k)^2=(x^TB^TB\Delta g_k)^2=\left((Bx)^T B\Delta g_k\right)^2$$

$$\leq (Bx)^T(Bx)\cdot(B\Delta g_k)^T(B\Delta g_k)$$

*Cauchy-Schwarz*不等式

$u^Tv\leq\|u\|\|v\|$,“=” $\Leftrightarrow$ u,v线性相关

$$=(x^TB^TBx)\cdot\left((\Delta g_k)^TB^TB\Delta g_k\right)$$

$$=(x^TH_kx)\cdot\left((\Delta g_k)^TH_k\Delta g_k\right)$$

$$(1)=\frac{(x^T H_k x)\left((\Delta g_k)^T H_k\Delta g_k\right)-(x^T H_k\Delta g_k)^2}{(\Delta g_k)^T H_k\Delta g_k}\geq 0$$

“=” $\Leftrightarrow$ $x,\Delta g_k$线性相关

再证 $(2)=\dfrac{\left(x^T\Delta x_k\right)^2}{\left(\Delta x_k\right)^T\Delta g_k}\ge 0,$

$$\left(\Delta x_k\right)^T\Delta g_k=\left(x^{k+1}-x^k\right)^T\left(g_{k+1}-g_k\right)$$

$$=\left(x^k+\lambda_k p^k-x^k\right)^T\left(g_{k+1}-g_k\right)$$

$$=\lambda_k\left(p^k\right)^T\left(g_{k+1}-g_k\right)\quad \text{精确一维搜索:}\left(g^{k+1}\right)^T p^k=0$$

$$=0-\lambda_k\left(p^k\right)^T g_k\quad p^k=-H_k g_k$$

$$=-\lambda_k\left(-H_k g_k\right)^T g_k$$

$$=\lambda_k\left(g_k\right)^T H_k g_k>0\quad g_k\ne 0, H_k\text{正定}$$

$$(2)=\frac{\left(x^T\Delta x_k\right)^2}{\left(\Delta x_k\right)^T\Delta g_k}\ge 0,\qquad \therefore x^T H_{k+1}x=(1)+(2)\ge 0,$$

下证 $x^T H_{k+1}x=(1)+(2)>0,$

即证当(1)、(2)中有一个为零时，另一个一定>0。

$$(1)=\frac{\left(x^T H_k x\right)\left(\left(\Delta g_k\right)^T H_k\Delta g_k\right)-\left(x^T H_k\Delta g_k\right)^2}{\left(\Delta g_k\right)^T H_k\Delta g_k}\ge 0$$

“=” $\Leftrightarrow x,\Delta g_k$ 线性相关

$$(2)=\frac{x^T\Delta x_k\left(\Delta x_k\right)^T x}{\left(\Delta x_k\right)^T\Delta g_k}=\frac{\left(x^T\Delta x_k\right)^2}{\left(\Delta x_k\right)^T\Delta g_k}\ge 0,$$

设存在 $\alpha\in R$，使得 $x=\alpha\Delta g_k$，

$$x^T\Delta x_k=x^T\left(x^{k+1}-x^k\right)=x^T\left(x^k+\lambda_k p^k-x^k\right)=\lambda_k x^T p^k\quad x=\alpha\Delta g_k$$

$$=\lambda_k\left(\alpha\Delta g_k\right)^T p^k=\lambda_k\alpha\left(g_{k+1}-g_k\right)^T p^k=-\lambda_k\alpha\left(g_k\right)^T p^k$$

$$=-\lambda_k\alpha\left(g_k\right)^T\left(-H_k g_k\right)\quad p^k=-H_k g_k$$

$$=\lambda_k\alpha\left(g_k\right)^T H_k g_k$$

下证 $x^T H_{k+1}x=(1)+(2)>0,$

即证当(1)、(2)中有一个为零时，另一个一定>0。

$$(1)=\frac{\left(x^T H_k x\right)\left(\left(\Delta g_k\right)^T H_k\Delta g_k\right)-\left(x^T H_k\Delta g_k\right)^2}{\left(\Delta g_k\right)^T H_k\Delta g_k}\ge 0$$

“=” $\Leftrightarrow x,\Delta g_k$ 线性相关

$$(2)=\frac{x^T\Delta x_k\left(\Delta x_k\right)^T x}{\left(\Delta x_k\right)^T\Delta g_k}=\frac{\left(x^T\Delta x_k\right)^2}{\left(\Delta x_k\right)^T\Delta g_k}\ge 0,$$

设存在 $\alpha\in R$，使得 $x=\alpha\Delta g_k$，

$$x^T\Delta x_k=x^T\left(x^{k+1}-x^k\right)=x^T\left(x^k+\lambda_k p^k-x^k\right)=\lambda_k x^T p^k\quad x=\alpha\Delta g_k$$

$$=\lambda_k\left(\alpha\Delta g_k\right)^T p^k=\lambda_k\alpha\left(g_{k+1}-g_k\right)^T p^k=-\lambda_k\alpha\left(g_k\right)^T p^k$$

$$=-\lambda_k\alpha\left(g_k\right)^T\left(-H_k g_k\right)\quad p^k=-H_k g_k$$

$$=\lambda_k\alpha\left(g_k\right)^T H_k g_k$$

下证 $x^T H_{k+1}x=(1)+(2)>0,$

即证当(1)、(2)中有一个为零时，另一个一定>0。

$$(1)=\frac{\left(x^T H_k x\right)\left(\left(\Delta g_k\right)^T H_k\Delta g_k\right)-\left(x^T H_k\Delta g_k\right)^2}{\left(\Delta g_k\right)^T H_k\Delta g_k}\ge 0$$

“=” $\Leftrightarrow x,\Delta g_k$ 线性相关

$$(2)=\frac{x^T\Delta x_k\left(\Delta x_k\right)^T x}{\left(\Delta x_k\right)^T\Delta g_k}=\frac{\left(x^T\Delta x_k\right)^2}{\left(\Delta x_k\right)^T\Delta g_k}\ge 0,$$

设存在 $\alpha\in R$，使得 $x=\alpha\Delta g_k$，则 $x^T\Delta x_k=\lambda_k\alpha\left(g_k\right)^T H_k g_k$，

当(1)=0时，$x,\Delta g_k$ 线性相关，则 $\alpha\ne 0$，注意 $g_k\ne 0, H_k$ 正定，

故 $x^T\Delta x_k=\lambda_k\alpha\left(g_k\right)^T H_k g_k\ne 0$，因此(2)>0.

当(2)=0时，由 $x^T\Delta x_k=\lambda_k\alpha\left(g_k\right)^T H_k g_k=0$ 可知，$\alpha=0$，

故 $x,\Delta g_k$ 线性无关，因此(1)>0.

DFP算法

定理2 设用DFP算法求解下列问题

$$\min \quad f(x)=\frac{1}{2}x^T Ax+b^T x+c$$

DFP算法最多n步必达到正定二次函数极小点，具有二次终止性。

其中A为n阶对称正定矩阵。由DFP方法产生的搜索方向为$p^1,p^2,\ldots,p^n$，对应的尺度矩阵为$H_1,H_2,\ldots,H_n$，则

(1) $\left(p^i\right)^T g_j=0, 1\le i<j\le n$

(2) $H_j Ap^i=p^i,\quad 1\le i<j\le n$

(3) $\left(p^i\right)^T Ap^j=0,\quad 1\le i<j\le n$

(3)表明DFP方法也是一种共轭方向法。

证明 参见P93定理4.14的证明

若H_1是单位矩阵，则$p^1=-H_1\nabla f\left(x^1\right)=-\nabla f\left(x^1\right)$，DFP方法也是一种共轭梯度法。

DFP算法--算法步骤

步骤1 选定初始点 x^1，初始矩阵$H_1=I_n,\varepsilon>0.$

步骤2 如果$\|g_1\|\le\varepsilon$，停止，得到近似驻点x^1，否则转步骤3。

步骤3 取 $p^1=-H_1 g_1, k=1.$

步骤4 精确一维搜索找最佳步长λ_k，令$x^{k+1}=x^k+\lambda_k p^k.$

步骤5 如果$\|g_{k+1}\|\le\varepsilon$，停止，得到近似驻点$x^{k+1}$，否则转步骤6。

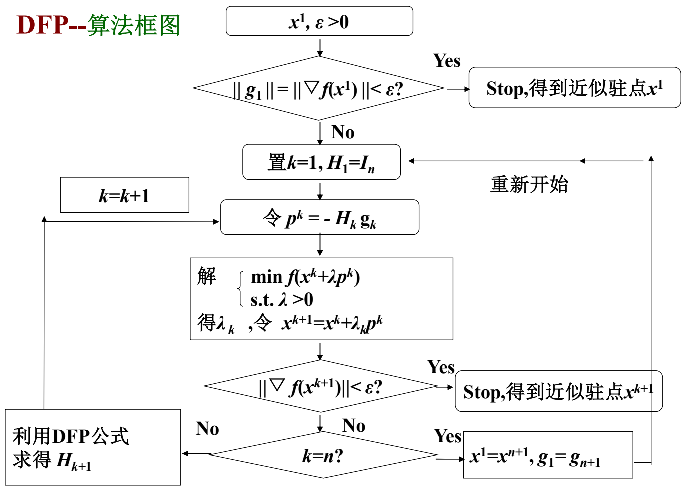
步骤6 如果$k=n$，令$x^1=x^{k+1},p^1=-g_{k+1},k=1$，转步骤4;否则转步骤7。

步骤7 令 $\Delta x_k=x^{k+1}-x^k,\Delta g_k=g_{k+1}-g_k,r_k=H_k\Delta g_k$，

计算 $$H_{k+1}=H_k+\frac{\Delta x_k\left(\Delta x_k\right)^T}{\left(\Delta x_k\right)^T\Delta g_k}-\frac{r_k\left(r_k\right)^T}{\left(\Delta g_k\right)^T H_k\Delta g_k}$$

和 $p^{k+1}=-H_{k+1}g_{k+1}$，令 $k=k+1$，转步骤4.

DFP--算法框图



DFP算法--算例

例 用DFP算法求 $\min f(x)=2x_1^2+x_2^2-4x_1+2$，取 $x^1=(2,1)^T,\varepsilon=10^{-3}.$

解 计算函数f的梯度函数： $g(x)=\begin{pmatrix}4(x_1-1)\\2x_2\end{pmatrix}$，

1）第一次迭代

在初始点x^1处的梯度为:$g_1=\begin{pmatrix}4\\2\end{pmatrix}$，取 $H_1=\begin{pmatrix}1&0\\0&1\end{pmatrix}$，

$$p^1=-H_1 g_1=-\begin{pmatrix}1&0\\0&1\end{pmatrix}\begin{pmatrix}4\\2\end{pmatrix}=\begin{pmatrix}-4\\-2\end{pmatrix},$$

$$x^1+\lambda p^1=\begin{pmatrix}2\\1\end{pmatrix}+\lambda\begin{pmatrix}-4\\-2\end{pmatrix}=\begin{pmatrix}2-4\lambda\\1-2\lambda\end{pmatrix}$$

DFP算法--算例

$$x^1+\lambda p^1=\begin{pmatrix}2-4\lambda\\1-2\lambda\end{pmatrix}\quad g(x)=\begin{pmatrix}4(x_1-1)\\2x_2\end{pmatrix},$$

精确一维搜索求最佳步长，

$$\phi(\lambda)=f(x^1+\lambda p^1)=f(2-4\lambda,1-2\lambda)$$
$$=2(2-4\lambda)^2+(1-2\lambda)^2-4(2-4\lambda)+2$$

令 $0=\phi'(\lambda)=4(2-4\lambda)(-4)+2(1-2\lambda)(-2)-4\times(-4),$

得 $\lambda_1=\dfrac{5}{18},$

$$x^2=x^1+\lambda_1 p^1=\begin{pmatrix}2-4\lambda\\1-2\lambda\end{pmatrix}\Bigg|_{\lambda=\lambda_1=\frac{5}{18}}=\begin{pmatrix}8/9\\4/9\end{pmatrix},\quad g_2=g(x_2)=\begin{pmatrix}-4/9\\8/9\end{pmatrix},$$

$\|g_2\|=\dfrac{4\sqrt{5}}{9}>\varepsilon,$ 继续迭代；

DFP算法--算例

$$H_1=\begin{pmatrix}1&0\\0&1\end{pmatrix},$$

2）第二次迭代

$$\Delta x_1=x^2-x^1=\lambda_1 p^1=\frac{5}{18}\begin{pmatrix}-4\\-2\end{pmatrix}=\begin{pmatrix}-10/9\\-5/9\end{pmatrix},$$

$$\Delta g_1=g_2-g_1=\begin{pmatrix}-4/9\\8/9\end{pmatrix}-\begin{pmatrix}4\\2\end{pmatrix}=\begin{pmatrix}-40/9\\-10/9\end{pmatrix},$$

$$H_2=H_1+\frac{\Delta x_1\Delta x_1^T}{\Delta g_1^T\Delta x_1}-\frac{H_1\Delta g_1\Delta g_1^T H_1}{\Delta g_1^T H_1\Delta g_1}$$

$$=\begin{pmatrix}1&0\\0&1\end{pmatrix}+\frac{\begin{pmatrix}-10/9\\-5/9\end{pmatrix}(-10/9,-5/9)}{(-10/9,-5/9)\begin{pmatrix}-40/9\\-10/9\end{pmatrix}}-\frac{\begin{pmatrix}1&0\\0&1\end{pmatrix}\begin{pmatrix}-40/9\\-10/9\end{pmatrix}(-40/9,-10/9)\begin{pmatrix}1&0\\0&1\end{pmatrix}}{(-40/9,-10/9)\begin{pmatrix}1&0\\0&1\end{pmatrix}\begin{pmatrix}-40/9\\-10/9\end{pmatrix}}$$

DFP算法--算例

$$H_2=H_1+\frac{\Delta x_1\Delta x_1^T}{\Delta g_1^T\Delta x_1}-\frac{H_1\Delta g_1\Delta g_1^T H_1}{\Delta g_1^T H_1\Delta g_1}$$

$$=\begin{pmatrix}1&0\\0&1\end{pmatrix}+\frac{1}{18}\begin{pmatrix}4&2\\2&1\end{pmatrix}-\frac{1}{17}\begin{pmatrix}16&4\\4&1\end{pmatrix}=\frac{1}{306}\begin{pmatrix}86&-38\\-38&305\end{pmatrix}$$

令 $p^2=-H_2g_2=-\dfrac{1}{306}\begin{pmatrix}86&-38\\-38&305\end{pmatrix}\begin{pmatrix}-4/9\\8/9\end{pmatrix}=\dfrac{12}{51}\begin{pmatrix}1\\-4\end{pmatrix},$

$$x^2+\lambda p^2=\begin{pmatrix}8/9\\4/9\end{pmatrix}+\lambda\times\frac{12}{51}\begin{pmatrix}1\\-4\end{pmatrix}=\begin{pmatrix}8/9+\frac{12}{51}\lambda\\4/9-4\times\frac{12}{51}\lambda\end{pmatrix},$$

DFP算法--算例

$$x^2+\lambda p^2=\begin{pmatrix}8/9+\frac{12}{51}\lambda\\4/9-4\times\frac{12}{51}\lambda\end{pmatrix},$$

精确一维搜索求最佳步长，

$$\phi(\lambda)=f(x^2+\lambda p^2)=f(8/9+\frac{12}{51}\lambda,4/9-4\times\frac{12}{51}\lambda)$$

令 $0=\phi'(\lambda),$ 得 $\lambda_2=17/36,$

$$x^3=x^2+\lambda_2 p^2=\begin{pmatrix}8/9+\frac{12}{51}\times\frac{17}{36}\\4/9-4\times\frac{12}{51}\times\frac{17}{36}\end{pmatrix}=\begin{pmatrix}1\\0\end{pmatrix},\ g_3=\nabla f(x^3)=\begin{pmatrix}0\\0\end{pmatrix},$$

$\|g_3\|=0<\varepsilon,$

算法终止，又该问题是凸规划，得到最优解: $x^*=x^3=\begin{pmatrix}1\\0\end{pmatrix}.$

DFP算法--收敛性分析

收敛性定理：

设凸函数 f 存在一阶连续偏导数，水平集 $L=\{x \mid f(x)\le f(x^1)\}$ 有界，则由DFP法得到的无穷点列 $\{x^k\}$ 具有如下性质：

(1) $\{f(x^k)\}$ 为严格单调下降序列，且 $\lim_{k\to\infty} f(x^k)$ 存在；

(2) $\{x^k\}$ 的任意聚点 x^* 都是 f 的极小点。特点：全局收敛性

DFP法在无约束优化方法中占有重要的地位，是目前最常用的方法之一。

定理(精确一维搜索迭代法的收敛性) θ_k 是 d^k 与 $-\nabla f(x^k)$ 的夹角

设 $\lambda_k=\arg\min f(x^k+\lambda d^k)$，$\nabla f(x)$ 在 $L=\{x\in R^n: f(x)\le f(x^1)\}$ 上存在且一致连续，$\theta_k\le\frac{\pi}{2}-\mu\ (\exists\mu>0)$，则或者存在某个 k 使得 $\nabla f(x^k)=0$，或者 $\nabla f(x^k)\to 0$，或者 $f(x^k)\to-\infty$

DFP算法--优缺点

优点：

(1) DFP算法具有二次收敛性

当 $f(x)=1/2x^TAx+b^Tx+c$ (A对称正定)时，由DFP算法产生的方向 $p^1, p^2, \ldots, p^n$ 是A-共轭向量组，故DFP算法最多 n 次迭代就可达到 f 的极小点。

DFP算法实质上是一种共轭方向法。

DFP算法--优缺点

优点：

（2）求解凸函数的极小点时，DFP算法全局收敛；

（3）对非二次函数，DFP算法的效果也很好，收敛速度是超线性的，

它比最速下降法和共轭梯度法要有效的多。

DFP算法--优缺点

缺点：

(1) DFP算法的计算量、存储量要比共轭梯度法大，

对大规模优化问题，用共轭梯度法更方便。

(2) 实际运算中，舍入误差和一维搜索的不精确，都会对DFP算法的稳定性和计算效率产生很大的影响；但BFGS算法受到的影响要小得多。

BFGS算法

由前面的推导$\Delta x_k \approx \left(\nabla^2 f\left(x^{k+1}\right)\right)^{-1} \Delta g_k$，则

$$\Delta g_k \approx \nabla^2 f\left(x^{k+1}\right)\Delta x_k,$$

令 B_{k+1}满足

$$\Delta g_k = B_{k+1}\Delta x_k, \qquad (2)$$

公式(2)称为另一种拟牛顿性质(或称拟牛顿条件或方程)。

上面的公式(2)只需要交换 $\Delta g_k, \Delta x_k$ 就可以得到前面的拟牛顿性质：

$$\Delta x_k = H_{k+1}\Delta g_k \qquad (1)$$

BFGS算法

因此只需要在H_k的递推公式

$$H_{k+1} = H_k + \frac{\Delta x_k (\Delta x_k)^T}{(\Delta x_k)^T \Delta g_k} - \frac{H_k \Delta g_k (\Delta g_k)^T H_k}{(\Delta g_k)^T H_k \Delta g_k}$$

互换 $\Delta g_k, \Delta x_k$，并用B_{k+1}，B_k 分别取代H_{k+1}，H_k，就得到B_k的递推公式，

$$B_{k+1} = B_k + \frac{\Delta g_k (\Delta g_k)^T}{(\Delta g_k)^T \Delta x_k} - \frac{B_k \Delta x_k (\Delta x_k)^T B_k}{(\Delta x_k)^T B_k \Delta x_k} \qquad (3)$$

该公式称为关于矩阵B_k的BFGS修正公式，有时也称为DFP的对偶公式。

BFGS算法

设B_{k+1}可逆，则由(2)

$$\Delta g_k = B_{k+1}\Delta x_k \qquad (2)$$

可知 $$\Delta x_k = \left(B_{k+1}\right)^{-1}\Delta g_k \qquad (4)$$

所以$\left(B_{k+1}\right)^{-1}$满足拟牛顿条件 $\Delta x_k = H_{k+1}\Delta g_k$，令 $H_{k+1} = \left(B_{k+1}\right)^{-1}$

对(3)两边求导，求导方法见文（陈宝林），

$$B_{k+1} = B_k + \frac{\Delta g_k (\Delta g_k)^T}{(\Delta g_k)^T \Delta x_k} - \frac{B_k \Delta x_k (\Delta x_k)^T B_k}{(\Delta x_k)^T B_k \Delta x_k} \qquad (3)$$

得到

BFGS算法

$$H_{k+1}^{BFGS} = H_k + \left(1 + \frac{(\Delta g_k)^T H_k \Delta g_k}{\Delta x_k{}^T \Delta g_k}\right)\frac{\Delta x_k \Delta x_k{}^T}{\Delta x_k{}^T \Delta g_k} - \frac{\Delta x_k (\Delta g_k)^T H_k + H_k \Delta g_k \Delta x_k{}^T}{(\Delta x_k)^T \Delta g_k} \qquad (5)$$

-------BFGS公式

这个重要公式由Broyden(布洛伊登), Fletcher(弗莱彻), Goldfard(戈德法布)和Shanno于1970年提出。

BFGS公式应用广泛，数值计算实例表明，它比DFP公式的效果要好。

BFGS算法具有变尺度法的全部优点，在一定条件下，使用非精确一维搜索的BFGS算法具有全局收敛性。

Broyden （布洛伊登 ）族变尺度法

将DFP公式记为

$$H_{k+1}^{DFP}=H_k+\frac{\Delta x_k(\Delta x_k)^T}{(\Delta x_k)^T\Delta g_k}-\frac{H_k\Delta g_k(\Delta g_k)^T H_k}{(\Delta g_k)^T H_k\Delta g_k} \quad (6)$$

结合BFGS公式，

$$H_{k+1}^{BFGS}=H_k+\left(1+\frac{(\Delta g_k)^T H_k\Delta g_k}{\Delta x_k{}^T\Delta g_k}\right)\frac{\Delta x_k\Delta x_k{}^T}{\Delta x_k{}^T\Delta g_k}-\frac{\Delta x_k(\Delta g_k)^T H_k+H_k\Delta g_k\Delta x_k{}^T}{(\Delta x_k)^T\Delta g_k} \quad (5)$$

引入参数 φ，有

Broyden族变尺度法

$$H_{k+1}^{\varphi}=(1-\varphi)H_{k+1}^{DFP}+\varphi H_{k+1}^{BFGS} \quad (7)$$

将(5)和(6)代入(7)，得到

$$H_{k+1}^{\varphi}=H_{k+1}^{DFP}+\varphi v^k(v^k)^T, \quad (8)$$

其中

$$v^k=\left((\Delta g_k)^T H_k\Delta g_k\right)^{1/2}\left(\frac{\Delta x_k}{(\Delta x_k)^T\Delta g_k}-\frac{H_k\Delta g_k}{(\Delta g_k)^T H_k\Delta g_k}\right)$$

将(7)或者(8)给出的修正公式的全体称为Broyden族。

当 $\varphi=0$ 时，即为DFP公式；当 $\varphi=1$ 时，即为BFGS公式。

Broyden族变尺度法

由于DFP和BFGS公式都满足拟牛顿性质，因此Broyden族的所有成员也满足拟牛顿性质。

Broyden族的任何一个成员都具有一般变尺度法的优点；

DFP算法所具有的许多性质， Broyden族算法也有。

在拟Newton法的每次迭代中，可用Broyden族的任意一个成员作为修正公式。

Broyden族含有一个参数，给出了一类拟Newton算法；

Huang族含有三个参数， 也给出了一类拟Newton算法；

Broyden族是Huang族的一个子族。

第四章作业

P99 4.1 4.2 4.4 4.5 4.9 4.10 4.12--4.14
4.17--4.19 4.23 4.26(初始点取$(1,1)^T$)

4.3 建议编程
迭代14次得到近似解$(-0.6344，-0.0032)^T$,近似最优值为-0.4724.