非线性规划是一种求解目标函数或约束条件中有一个或几个非线性函数的最优化问题的方法。

非线性规划的一般形式：

$$\begin{aligned}&\min\ f(X)\\&\text{s.t.}\ \ h_i(X)=0,\ i=1,2,\cdots,m\\&\quad\ g_j(X)\le 0,\ j=1,2,\cdots,l\end{aligned}$$

其中 $X=(x_1,x_2,\cdots,x_n)^T\in R^n$

定以可行集为

$$\Omega=\left\{X\in R^n\,\middle|\,h_i(X)=0,\ 1\le i\le m,\ g_j(X)\le 0,1\le j\le l\right\}$$

上述一般形式可简写成 $\min\limits_{X\in\Omega} f(X)$

定义局部最优解和全局最优解：

局部最优解 $\hat{X}$：$\hat{X}\in\Omega$，且存在 $\varepsilon>0$ 满足

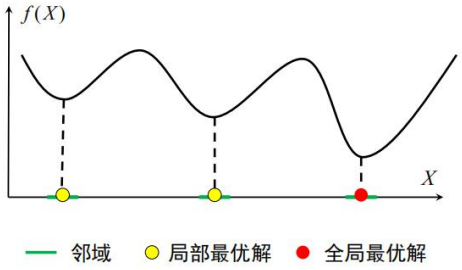
$$f(\hat{X})\le f(X),\ \forall X\in B(\hat{X},\varepsilon)\cap\Omega$$

全局最优解 $\hat{X}$：$\hat{X}\in\Omega$，$f(\hat{X})\le f(X),\ \forall X\in\Omega$

其中 X 的邻域可简单定义如下：

$\hat{X}$ 的 ε 邻域：$B(\hat{X},\varepsilon)=\left\{X\in R^n\,\middle|\,\|X-\hat{X}\|<\varepsilon\right\}$

如果在上面的定义中满足 $f(\hat{X})<f(X)$ ，则称为严格局部最优解和严格全局最优解



对于任意的函数而言，在众多的局部最优解中精确寻找它的全局最优解不是一个平凡的问题，然而有一种满足特定条件的函数族能够保证该问题的局部最优解即为所定义范围内的全局最优解，这个条件即为凸性，具体而言：

如果 Ω 是凸集，$f(X)$ 是其上的连续凸函数，称

$$\min_{X\in\Omega} f(X)$$

是凸规划问题

如果 $X^*\in\Omega$ 是凸规划问题的任意一个局部最优解，那么它也是该问题的全局最优解

在线性规划中，我们已经知道凸集的定义：

数学定义：Ω 是凸集当且仅当对任意实数 $0<\alpha<1$ 和任意的 $X_1,X_2\in\Omega$ 均成立

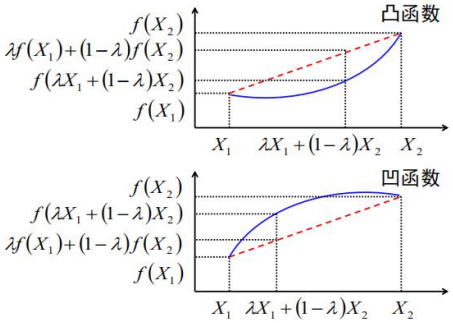
$$\alpha X_1+(1-\alpha)X_2\in\Omega$$

而凸函数则是指具有如下性质的函数：

设 $f(X)$ 是定义在集合 $\Omega\subset R^n$ 上的函数，如果 Ω 是凸集，并且对 Ω 中任意两点 X_1,X_2 以及闭区间 $[0,\ 1]$ 中任意一点 λ 都满足

$$f(\lambda X_1+(1-\lambda)X_2)\le\lambda f(X_1)+(1-\lambda)f(X_2)$$

一元凸（凹）函数的图象



凸函数全局最优值性质的证明如下：

证明：如果存在 $\hat{X}\in\Omega$ 满足 $f(\hat{X})<f(X^*)$

$$\Rightarrow\ \lambda f(\hat{X})<\lambda f(X^*),\quad\forall\lambda>0$$

又因为

$$f(\lambda\hat{X}+(1-\lambda)X^*)\le\lambda f(\hat{X})+(1-\lambda)f(X^*),\ \forall 0\le\lambda\le 1$$

$$\Rightarrow\ f(\lambda\hat{X}+(1-\lambda)X^*)<f(X^*),\ \forall 0<\lambda\le 1$$

因为对充分小的 $\lambda>0$，$\lambda\hat{X}+(1-\lambda)X^*$ 能够充分接近 X^*，说明 X^* 不是局部最优解，矛盾！

为了方便应用上述性质，我们引入下面的凸函数判别定理：

多元可导凸函数的二阶充分条件

$$\nabla^2 f(X)\ge 0,\ \forall X\in\Omega$$

证明途径

$$g(t)=f(X_1+t(X_2-X_1))$$

$$g'(t)=\nabla f(X_1+t(X_2-X_1))^T(X_2-X_1)$$

$$g''(t)=(X_2-X_1)^T\nabla^2 f(X_1+t(X_2-X_1))(X_2-X_1)$$

$$\nabla^2 f(X)\ge 0,\ \forall X\in\Omega\quad\Rightarrow\quad g''(t)\ge 0$$

$$\Rightarrow\quad g(t)\text{ 是一元凸函数}$$

下面介绍针对非线性规划问题的基本途径“迭代算法”

1. 函数求极值问题

$$\min f(X),\ X\in\mathrm{R}^n\quad\Rightarrow\quad\nabla f(X^*)=0$$

2. 迭代算法

$$X_{k+1}=X_k+\lambda_k D_k$$

$\lambda_k\in\mathrm{R}^1$ 一维搜索步长、$D_k\in\mathrm{R}^n$ 寻优方向

3. 下降方向

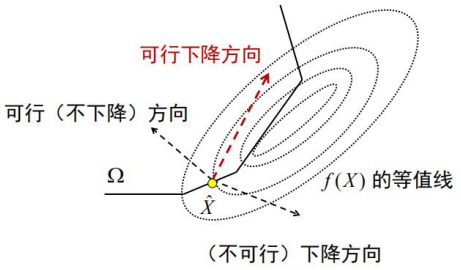
对于优化问题 $\min\limits_{X\in\Omega} f(X)$，给定可行解 $\hat{X}\in\Omega$ 以及向量 $D\in R^n$，如果存在 $\bar{t}>0$ 满足

$$\hat{X}+tD\in\Omega,\ \forall 0<t\le\bar{t}$$

称 D 是 $\hat{X}$ 处的可行方向，如果存在 $\bar{t}>0$ 满足

$$f\left(\hat{X}+tD\right)<f(\hat{X}),\ \forall 0<t\le\bar{t}$$

称 D 是 $\hat{X}$ 处的下降方向，既可行又下降的方向称为可行下降方向



4. 全部思路

$\min\limits_{X\in\Omega} f(X)$ 寻优算法的基本思路：可行下降迭代算法

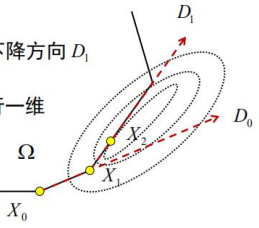
确定初始可行解 $X_0\in\Omega$ → 确定可行下降方向 D_0

⇒ 一维搜索确定 λ_0 满足 $X_1=X_0+\lambda_0 D_0\in\Omega$ 以及 $f(X_1)<f(X_0)$

⇒ 确定 X_1 处的可行下降方向 D_1

⇒ 在 X_1 处沿 D_1 进行一维搜索确定 X_2 ……

$X_k=X_{k-1}+\lambda_{k-1}D_{k-1}$



对于前面的凸优化问题，有下面的迭代终止方案：

对于凸规划 $\min\limits_{X\in\Omega} f(X)$（Ω 凸集，$f(X)$ 凸函数）$X^*\in\Omega$ 是最优解的充要条件是在该点不存在可行下降方向

必要性显然。充分性反证：如果有另一点更好，连接两点可得可行下降方向。

或者由微积分中的结论有下面最优性条件：

1）X^* 是局部最优解的必要条件：$\nabla f(X^*)=0$

理由：利用二阶泰勒展开

$$D=-\nabla f(X^*)\ \Rightarrow\ \begin{aligned} f(X^*+tD)-f(X^*) &= -\left\|\nabla^T f(X^*)\right\|^2 t+\frac{1}{2}D^T\nabla^2 f(X^*+\xi D)Dt^2 \\ &= -t\left(\left\|\nabla^T f(X^*)\right\|^2-\frac{1}{2}D^T\nabla^2 f(X^*+\xi D)Dt\right)\end{aligned}$$

$$\nabla f(X^*)\ne 0\ \Rightarrow\ f(X^*+tD)-f(X^*)<0,\ \forall t\in(0,\varepsilon)$$

$$\Rightarrow\ X^*\text{ 不是局部最优解}$$

2）X^* 是严格局部最优解的充分条件：

$$\nabla f(X^*)=0\quad \nabla^2 f(X^*)>0$$

理由：$\nabla f(X^*)=0\ \Rightarrow\ \nabla^T f(X^*)D=0,\ \forall D\in R^n$

$$\Rightarrow\ f(X^*+tD)-f(X^*)=\frac{1}{2}D^T\nabla^2 f(X^*+\xi D)Dt^2,\ \forall D\in R^n$$

$$\nabla^2 f(X^*)>0\ \Rightarrow\ \nabla^2 f(X^*+\xi D)>0,\ \forall \xi\in(0,\hat{\varepsilon})$$

$$\Rightarrow\ f(X^*+tD)>f(X^*),\ \forall D\in R^n, t\in(0,\hat{\varepsilon})$$

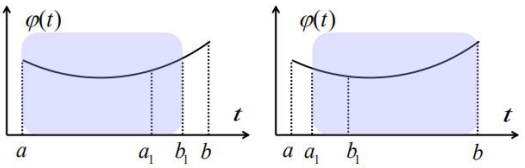
$$\Rightarrow\ f(X)>f(X^*),\ \forall X\in B(X^*,\varepsilon)$$

要实现上面的全部程序，需要明确下面两步的方法：

1. 一维最优解的搜索方法
2. 每点处优化方向的确定

精确搜索的办法可以通过压缩区间来逼近，逐步缩小包含局部最优解的区间直至区间长度小于给定阈值

已知闭区间 $[a,b]$ 是单谷区间，在其内部任取两点 $a_1<b_1$，计算 $\varphi(a_1),\varphi(b_1)$，如果 $\varphi(a_1)<\varphi(b_1)$，局部最优解在区间 $[a,b_1]$，否则局部最优解在区间 $[a_1,b]$，两种情况均能压缩区间



为了比值，可以采用黄金分割比 0.618 作为每次收缩的大小：

在单谷区间 $[a,b]$ 搜索局部最优解的0.618法

1）确定误差阈值 δ 及满足 $0.618^{n-1}(b-a)\le\delta$ 的 n

2）令 $a_0=a, b_0=b$

3）对于 $k=1,2,\cdots,n$ 依次完成以下运算

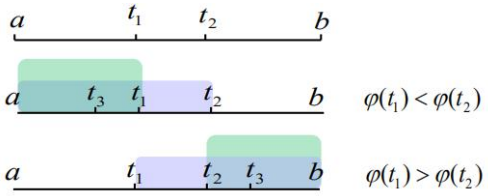
a）令 $t_k=a_{k-1}+0.618(b_{k-1}-a_{k-1})$

$t_k'=b_{k-1}+0.618(a_{k-1}-b_{k-1})$

b）计算 $\varphi(t_k)$ 和 $\varphi(t_k')$ 中未知的数值

c）比较 $\varphi(t_k)$ 和 $\varphi(t_k')$ 的大小确定 $[a_k,b_k]$

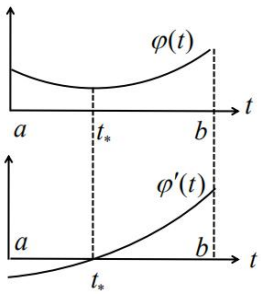
4）取所求局部最优解为 $0.5(a_n+b_n)$



当然我们还可以采用区间对分的方法来进行精确搜索，每次只需要依据一个点的信息进行搜索，所以还需要原函数的导数信息。如下图所示，求解原函数的极值点相当于求解导函数的零点：

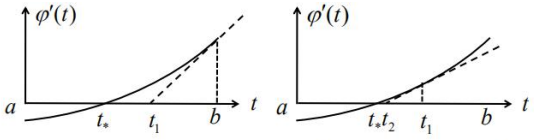
单谷区间 $[a,b]$ 一定满足 $\varphi'(a)<0,\varphi'(b)>0$

取 $t=0.5(a+b)$，若 $\varphi'(t)>0$，将区间压缩为 $[a,t]$，否则将区间压缩为 $[t,b]$



这种方法区间压缩比等于 0.5 ，比仅计算函数值的 0.618 法好，实际效果取决于**导数计算量**。

利用二阶导数的精确搜索算法也称为 Newton 法，它在一维时的情况如下所示：



如左图，在 b 点用切线近似 $\varphi'(t)$，求该切线的零点

切线方程：$g(t)=\varphi'(b)+\varphi''(b)(t-b)$　　如上所示

$$g(t_1)=0 \quad \Rightarrow \quad t_1=b-\frac{\varphi'(b)}{\varphi''(b)}$$　　单谷区间一定收敛

再在 t_1 点重复上述过程 $\Rightarrow$ $t_2=t_1-\dfrac{\varphi'(t_1)}{\varphi''(t_1)}$

确定下降方向的经典方法如梯度下降法：

$$f(X+tD)=f(X)-\|\nabla f(X)\|^2 t+\frac{1}{2}D^T\nabla^2 f(X+\xi D)Dt^2$$

$$\Rightarrow f(X+tD)-f(X)=-t\left(\|\nabla f(X)\|^2-\frac{1}{2}D^T\nabla^2 f(X+\xi D)Dt\right)$$

只要 $\nabla f(X)\neq 0$，就有 $\|\nabla f(X)\|>0$，一定存在 $\bar{t}>0$ 满足 $f(X+tD)<f(X),\forall 0<t\le\bar{t}$

所以负梯度方向是下降方向

梯度下降法

1）任取 $\hat{X}\in R^n$

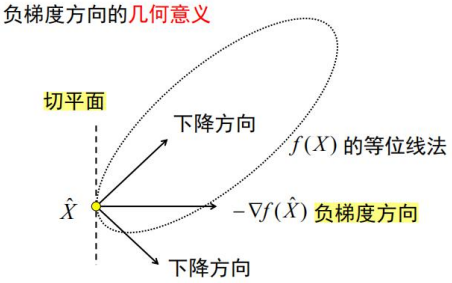
2）计算 $D=-\nabla f(\hat{X})$

3）如果 $\|D\|\le\delta$ 其中 δ 是预先设定的阈值，停止计算，以 $\hat{X}$ 为所求解，否则进行直线搜索，确定能够满足 $f(\hat{X}+\hat{t}D)<f(\hat{X})$ 的 $\hat{t}>0$

4）用 $\hat{X}+\hat{t}D$ 替换 $\hat{X}$，然后回到 2）继续迭代

它有清楚的几何意义：

负梯度方向的几何意义



负梯度方向和切平面垂直

函数的梯度相当于它的 l_2 范数，在实际中还有采用 l_1、l_∞ 范数的算法，他们都使用了不同范数下的“最速下降方向”

最速下降方向

$$\min\left\{\nabla^T f(X)D \,\middle|\, \text{s.t. } \|D\|=1\right\} \Leftrightarrow \max\left\{-\nabla^T f(X)D \,\middle|\, \text{s.t. } \|D\|=1\right\} \Rightarrow \hat{D}$$

ℓ_1 范数 $\|D\|_1=\sum_{i=1}^n|d_i|$

$$\hat{d}_i=\begin{cases}\operatorname{sgn}\left(-\dfrac{\partial f(X)}{\partial x_i}\right) & \text{if } \left|\dfrac{\partial f(X)}{\partial x_i}\right|=\|\nabla f(X)\|_\infty \\ 0 & \text{if } \left|\dfrac{\partial f(X)}{\partial x_i}\right|\neq\|\nabla f(X)\|_\infty\end{cases}$$

$$\nabla f(X)^T\hat{D}=-\|\nabla f(X)\|_\infty$$

ℓ_p 范数 $\|D\|_p=\left(\sum_i|d_i|^p\right)^{\frac{1}{p}},\ p>1$

$$\hat{d}_i=\operatorname{sgn}\left(-\frac{\partial f(X)}{\partial x_i}\right)\left|\frac{\partial f(X)}{\partial x_i}\right|^{q-1}\left(\|\nabla f(X)\|_q\right)^{-\frac{q}{p}},\forall i$$

$$\nabla f(X)^T\hat{D}=-\|\nabla f(X)\|_q \qquad \left(\frac{1}{q}=1-\frac{1}{p}\right)$$

ℓ_∞ 范数 $\|D\|_\infty=\max_{1\le i\le n}|d_i|$

$$\hat{d}_i=\operatorname{sgn}\left(-\frac{\partial f(X)}{\partial x_i}\right),\forall i$$

$$\nabla f(X)^T\hat{D}=-\|\nabla f(X)\|_1$$

然而，负梯度等范数最速下降法下降法有其固有缺陷，这为共轭梯度法的提出作出了铺垫。精确搜索得到新点的梯度方向与搜索方向正交，即沿负梯度方向精确搜索前进时，相邻两点的梯度互相垂直：

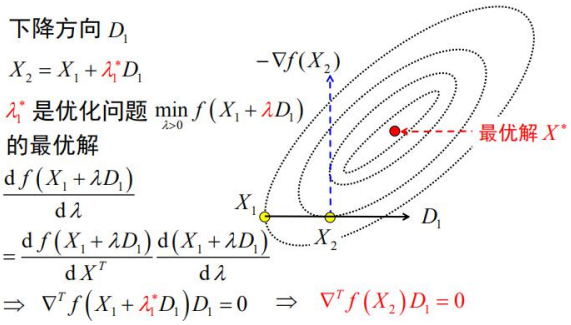
下降方向 D_1

$X_2=X_1+\lambda_1^*D_1$

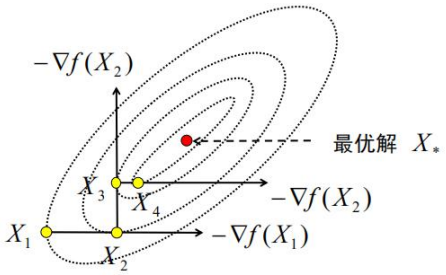
λ_1^* 是优化问题 $\min_{\lambda>0} f(X_1+\lambda D_1)$ 的最优解

$$\frac{\mathrm{d}f(X_1+\lambda D_1)}{\mathrm{d}\lambda}=\frac{\mathrm{d}f(X_1+\lambda D_1)}{\mathrm{d}X^T}\frac{\mathrm{d}(X_1+\lambda D_1)}{\mathrm{d}\lambda}$$

$$\Rightarrow \nabla^T f\left(X_1+\lambda_1^*D_1\right)D_1=0 \quad \Rightarrow \quad \nabla^T f\left(X_2\right)D_1=0$$



所以在解空间中梯度下降法沿锯齿状路线前进，接近最优解时一维搜索效率很低，前进速度很慢：



改进梯度下降法的思路可以是校正寻优方向以加快寻优速度。假设被优化的目标函数是一个二次函数

将二次正定函数 $f(X)=\frac{1}{2}(X-X^*)^T A(X-X^*)$ 改写为一般形式 $f(X)=\frac{1}{2}X^TAX+B^TX+c$

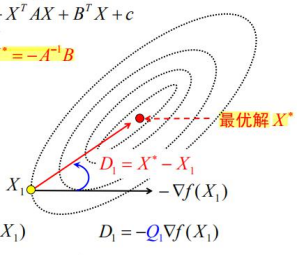
$\nabla f(X)=AX+B \Rightarrow X^*=-A^{-1}B$

$$D_1=X^*-X_1=-A^{-1}B-X_1=-A^{-1}(AX_1+B)$$

$\nabla^2 f(X)=A$

$D_1=-\left(\nabla^2 f(X_1)\right)^{-1}\nabla f(X_1)$　　$D_1=-Q_1\nabla f(X_1)$

$X_k=X_{k-1}-\lambda_{k-1}\left(\nabla^2 f(X_{k-1})\right)^{-1}\nabla f(X_{k-1})$ 牛顿法



其中 $D=-\left(\nabla^2 f(\hat{X})\right)^{-1}\nabla f(\hat{X})$ 被称为牛顿方向

广义牛顿法的流程如下：

1）任取 $\hat{X} \in R^n$

2）如果 $\|\nabla f(\hat{X})\|$ 不大于预先设定的阈值，停止计算，以 $\hat{X}$ 为所求解，否则到下一步

3）计算 $D=-(\nabla^2 f(\hat{X}))^{-1}\nabla f(\hat{X})$，进行一维搜索确定能够满足 $f(\hat{X}+\hat{t}D)<f(\hat{X})$ 的 $\hat{t}>0$

4）用 $\hat{X}+\hat{t}D$ 替换 $\hat{X}$，然后回到 2）继续迭代

类似于一维搜索中的牛顿法，这里用来求解优化方向的牛顿法也是利用目标函数和二次函数的近似特征，可以期望的是，当目标函数和用于近似的二次函数足够接近时，牛顿法能拥有比最速下降法更快的收敛速度。

而牛顿法也有其固有缺点，例如，分析如下问题时：

$$\min f(x)=x_1^4+x_1x_2+(1+x_2)^2$$

初始点 $x^{(1)}=(0,0)^T$。

在初始点的梯度和Hessian矩阵分别为

$$\nabla f(x)=\begin{bmatrix}4x_1^3+x_2\\ x_1+2(1+x_2)\end{bmatrix}\quad \nabla^2 f(x)=\begin{bmatrix}12x_1^2 & 1\\ 1 & 2\end{bmatrix}$$

$$\nabla f(x^{(1)})=\begin{bmatrix}0\\ 2\end{bmatrix}\qquad \nabla^2 f(x^{(1)})=\begin{bmatrix}0 & 1\\ 1 & 2\end{bmatrix}$$

在初始点的牛顿方向为

$$d^{(1)}=-\nabla^2 f(x^{(1)})^{-1}\nabla f(x^{(1)})=\begin{bmatrix}-2\\ 0\end{bmatrix}$$

在初始点沿牛顿方向进行一维精确搜索

$$\min \varphi(t)=f\left(x^{(1)}+td^{(1)}\right)$$
$$=16t^4+1$$

可以得到

$$\varphi'(t)=64t^3=0$$
$$t^{(1)}=0$$

显然，用牛顿法不能产生新的点，而初始点并不是无约束优化问题的极小点。牛顿方向失效的原因在于**初始点的 Hessian 矩阵非正定**！

牛顿方向的缺陷

1）每步迭代要计算 $(\nabla^2 f(\hat{X}))^{-1}$，计算量大

2）$(\nabla^2 f(\hat{X}))^{-1}$ 可能不存在

3）$(\nabla^2 f(\hat{X}))^{-1}$ 可能不正定，$D=-(\nabla^2 f(\hat{X}))^{-1}\nabla f(\hat{X})$ 不是下降方向

要克服牛顿法的缺陷，显然需要找办法绕过 Hessian 矩阵的计算，这便是共轭梯度法所采用的路径。所谓共轭方向是指：

共轭方向定义：$A\in R^{n\times n}$ 对称矩阵，$\vec{p},\vec{q}\in R^n$ 非零向量，若 $\vec{p}^T A\vec{q}=0$，称 $\vec{p},\vec{q}$ 为 A 共轭方向

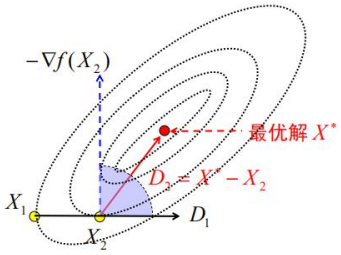
在这里依然考察在牛顿法中近似处理过的多元二次函数：

$$f(X)=\frac{1}{2}(X-X^*)^T A(X-X^*)$$

对每次梯度下降的方向作出如下的改进使之更快地收敛到极值点：

$$D_k=-\nabla f(X_k)+\alpha_{k-1}D_{k-1}$$

$$D_1=-\nabla f(X_1)$$
$$D_2=-\nabla f(X_2)+\alpha_1 D_1$$
$$\nabla f(X_2)=A(X_2-X^*)$$
$$=-AD_2$$
$$D_1^T\nabla f(X_2)=0$$
$$\Rightarrow\ -D_1^T AD_2=0$$
$$\Rightarrow\ D_1^T AD_2=0\ \Rightarrow\ -D_1^T A\nabla f(X_2)+\alpha_1 D_1^T AD_1=0$$



D_1 与 D_2 为 A 的共轭方向！

$$\alpha_1=\frac{D_1^T A\nabla f(X_2)}{D_1^T AD_1}$$

参数 α 中矩阵 A 的消除方法

由 $X_2=X_1+\lambda_1^* D_1\Rightarrow D_1=(X_2-X_1)/\lambda_1^*$

$$\alpha_1=\frac{D_1^T A\nabla f(X_2)}{D_1^T AD_1}=\frac{\nabla^T f(X_2)AD_1}{D_1^T AD_1}$$
$$=\frac{\nabla^T f(X_2)A(X_2-X_1)/\lambda_1^*}{D_1^T A(X_2-X_1)/\lambda_1^*}=\frac{\nabla^T f(X_2)A(X_2-X_1)}{D_1^T A(X_2-X_1)}$$
$$=\frac{\nabla^T f(X_2)(\nabla f(X_2)-\nabla f(X_1))}{D_1^T(\nabla f(X_2)-\nabla f(X_1))}=\frac{\nabla^T f(X_2)(\nabla f(X_2)+D_1)}{D_1^T(\nabla f(X_2)+D_1)}$$
$$=\frac{\nabla^T f(X_2)\nabla f(X_2)}{D_1^T D_1}=\frac{\|\nabla f(X_2)\|^2}{\|\nabla f(X_1)\|^2}$$

每前后两次寻优方向 D_k、D_{k+1} 之间互为 A 的共轭方向，所以采用上述修正寻优方法的算法也被称为共轭梯度法。上面我们已经推导出了经典的 Fletcher-Reeves 共轭梯度法。

共轭梯度法（Fletcher-Reeves）

1）任取 $X_0\in R^n$，令 $k=0$

2）如果 $\|\nabla f(X_k)\|\le\delta$，停止计算

3）如果 k/n 等于0或整数，令 $D_k=-\nabla f(X_k)$

否则令 $D_k=-\nabla f(X_k)+\alpha_k D_{k-1}$，其中

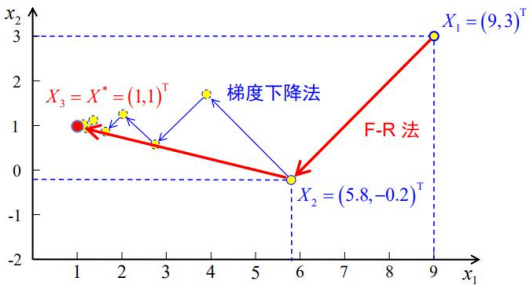
$$\alpha_k=\frac{\|\nabla f(X_k)\|^2}{\|\nabla f(X_{k-1})\|^2}$$

4）进行精确搜索获得 $X_{k+1}=X_k+t_k D_k$

5）用 $k+1$ 替换 k，回到 2）继续迭代

F-R 共轭梯度法 —— 寻优轨迹对比

$$\min f(X)=x_1^2+4x_2^2-2x_1-8x_2+5$$



利用相似的方法可以求得其它两种共轭梯度法：

F-R 共轭梯度法

Fletcher-Reeves $\alpha_k^{FR}=\dfrac{\|\nabla f(X_{k+1})\|^2}{\|\nabla f(X_k)\|^2}$

P-R-P 共轭梯度法

Polak-Ribiere
Polyak $\alpha_k^{PRP}=\dfrac{\nabla^{\mathrm{T}} f(X_{k+1})(\nabla f(X_{k+1})-\nabla f(X_k))}{\|\nabla f(X_k)\|^2}$

H-S 共轭梯度法

Beale-Sorenson
Hestenes-Stiefel $\alpha_k^{HS}=\dfrac{\nabla^{\mathrm{T}} f(X_{k+1})(\nabla f(X_{k+1})-\nabla f(X_k))}{D_k^{\mathrm{T}}(\nabla f(X_{k+1})-\nabla f(X_k))}$

有趣的是，第四种共轭梯度算法是通过对偶性来发现的：

几种著名的共轭梯度法 ——“缺失的一角”

分母 \ 分子	$\nabla^{\mathrm{T}} f(X_{k+1})(\nabla f(X_{k+1})-\nabla f(X_k))$	$\|\nabla f(X_{k+1})\|^2$
$\|\nabla f(X_k)\|^2$	P-R-P法 α_k^{PRP}	F-R法 α_k^{FR}
$D_k^{\mathrm{T}}(\nabla f(X_{k+1})-\nabla f(X_k))$	H-S法 α_k^{HS}	?

在“缺失的一角”有另外一种未知的共轭梯度法

第四种著名的共轭梯度法：D-Y法

$$\alpha_k=\frac{\|\nabla f(X_{k+1})\|^2}{D_k^{\mathrm{T}}(\nabla f(X_{k+1})-\nabla f(X_k))}$$

实际中三种基于梯度的搜索方向各有其特点，在实际问题的解决中要注意合理运用：

	计算量	效率		鲁棒性
		解附近	远离解	
负梯度	A	C	A	A
共轭梯度	B	B	B	B
牛顿方向	C	A	C	C

附录：多元函数及矩阵的求导方法如下

1. 标量函数求偏导数（梯度）

$$\nabla f(X)=\frac{\partial f(X)}{\partial X}=\begin{pmatrix}\frac{\partial f(X)}{\partial x_1}\\ \frac{\partial f(X)}{\partial x_2}\\ \vdots\\ \frac{\partial f(X)}{\partial x_n}\end{pmatrix}$$

$$\nabla^T f(X)=\frac{\partial f(X)}{\partial X^T}=\left(\frac{\partial f(X)}{\partial x_1},\frac{\partial f(X)}{\partial x_2},\cdots,\frac{\partial f(X)}{\partial x_n}\right)$$

2. 向量函数求偏导数

$$F(X)=(f_1(X),f_2(X),\cdots,f_m(X))^T$$

$$\frac{\partial F^T(X)}{\partial X}=\left(\frac{\partial f_1(X)}{\partial X},\frac{\partial f_2(X)}{\partial X},\cdots,\frac{\partial f_m(X)}{\partial X}\right)_{n\times m}$$

$$=(\nabla f_1(X),\nabla f_2(X),\cdots,\nabla f_m(X))_{n\times m}$$

$$\frac{\partial F(X)}{\partial X^T}=\begin{pmatrix}\frac{\partial f_1(X)}{\partial X^T}\\ \frac{\partial f_2(X)}{\partial X^T}\\ \vdots\\ \frac{\partial f_m(X)}{\partial X^T}\end{pmatrix}_{m\times n}=\begin{pmatrix}\nabla^T f_1(X)\\ \nabla^T f_2(X)\\ \vdots\\ \nabla^T f_m(X)\end{pmatrix}_{m\times n}$$

3. 对向量函数的点积求偏导数

$$F(X)=(f_1(X),f_2(X),\cdots,f_m(X))^T$$

$$G(X)=(g_1(X),g_2(X),\cdots,g_m(X))^T$$

$$\frac{\partial(F^T(X)G(X))}{\partial X}=\frac{\partial F^T(X)}{\partial X}G(X)+\frac{\partial G^T(X)}{\partial X}F(X)$$

$$\frac{\partial(F^T(X)G(X))}{\partial X^T}=F^T(X)\frac{\partial G(X)}{\partial X^T}+G^T(X)\frac{\partial F(X)}{\partial X^T}$$

4. 对常数矩阵和向量函数的乘积求偏导数

$$A\in R^{m\times m}\quad F(X)=(f_1(X),f_2(X),\cdots,f_m(X))^T$$

$$\frac{\partial(AF(X))^T}{\partial X}=\frac{\partial(F^T(X)A^T)}{\partial X}=\frac{\partial F^T(X)}{\partial X}A^T$$

$$\frac{\partial AF(X)}{\partial X^T}=A\frac{\partial F(X)}{\partial X^T}$$

5. 对二次函数求偏导数

$$\frac{\partial(X^TAX)}{\partial X}=\frac{\partial X^T}{\partial X}AX+\frac{\partial(AX)^T}{\partial X}X=(A+A^T)X=2AX$$

$$\frac{\partial(X^TAX)}{\partial X^T}=X^T\frac{\partial(AX)}{\partial X^T}+(AX)^T\frac{\partial X}{\partial X^T}=X^T(A+A^T)=2X^TA$$

6. 海赛（Hesse）矩阵

$$\nabla^2 f(X)=\frac{\partial\nabla f(X)}{\partial X^T}=\begin{bmatrix}\frac{\partial^2 f(X)}{\partial x_1x_1} & \frac{\partial^2 f(X)}{\partial x_1x_2} & \cdots & \frac{\partial^2 f(X)}{\partial x_1x_n}\\ \frac{\partial^2 f(X)}{\partial x_2x_1} & \frac{\partial^2 f(X)}{\partial x_2x_2} & \cdots & \frac{\partial^2 f(X)}{\partial x_2x_n}\\ \vdots & \vdots & \ddots & \vdots\\ \frac{\partial^2 f(X)}{\partial x_nx_1} & \frac{\partial^2 f(X)}{\partial x_nx_2} & \cdots & \frac{\partial^2 f(X)}{\partial x_nx_n}\end{bmatrix}$$

7. 多元函数在给定点沿给定方向的二阶泰勒展开

$$f(\hat{X}+tD)=f(\hat{X})+c_1t+\frac{1}{2}c_2(\xi)t^2$$

$$=f(\hat{X})+\nabla^T f(\hat{X})Dt+\frac{1}{2}D^T\nabla^2 f(\hat{X}+\xi D)Dt^2$$