运筹学

（无约束优化）

王焕钢

清华大学自动化系

要点：无约束优化的最优性条件

无约束优化问题 $\min\limits_{X\in R^n} f(X)$

基本假定：目标函数具有二阶导数

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

Hesse矩阵

$$\nabla^2 f(X)=\frac{\partial\nabla^T f(X)}{\partial X}=\begin{pmatrix}\dfrac{\partial^2 f(X)}{\partial x_1\partial x_1} & \cdots & \dfrac{\partial^2 f(X)}{\partial x_n\partial x_1}\\ \vdots & \ddots & \vdots\\ \dfrac{\partial^2 f(X)}{\partial x_n\partial x_1} & \cdots & \dfrac{\partial^2 f(X)}{\partial x_n\partial x_n}\end{pmatrix}$$

给定方向的二阶泰勒展开

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

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

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

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

$$\Rightarrow \quad X^* \text{ 不是局部最优解}$$

2）X^*是严格局部最优解的充分条件：

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

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

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

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

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

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

要点：下降方向法

1847年，法国数学家Cauchy提出梯度法

$$\min f(X),\ X \in \mathrm{R}^n$$

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

一维精确搜索：$X_{k+1} = X_k + \lambda_k^* D_k$

$$f(X_{k+1}) = \min f(X_k + \lambda_k D_k),\ \lambda_k > 0$$

梯度下降法：$D_k = -\nabla f(X_k)$

奥古斯丁·路易斯·柯西
Augustin Louis Cauchy
1789—1857

基本算法（下降方向法）：

1）任取 $X \in \boldsymbol{R}^n$

2）如果在 X 处找不到下降方向，停止，否则，确定

X 处的下降方向$D \in R^n$

3）直线搜索确定 t 满足 $f\left(X+tD\right)<f\left(X\right)$

4）用 $X+tD$ 替换 X ，回到 2）继续迭代

基本算法（下降方向法）：

1）任取 $X \in R^n$

2）如果在 X 处找不到下降方向，停止，否则，确定 X 处的下降方向$D \in R^n$

3）直线搜索确定 t 满足 $f(X+tD)< f(X)$

4）用 $X+tD$ 替换 X ，回到 2）继续迭代

实现算法的关键：如何确定下降方向 D ？

要点：梯度下降法

下降方向：负梯度方向

$$D = -\nabla f(X)$$

代入二阶泰勒展开

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

$$\Rightarrow f(X+tD) - f(X)$$

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

只要 $\nabla f(X) \neq 0$，就有 $\|\nabla f(X)\| > 0$ ，一定存在 $\bar{t} > 0$

满足 $f(X+tD) < f(X), \forall 0 < t \leq \bar{t}$

所以负梯度方向是下降方向

梯度下降法

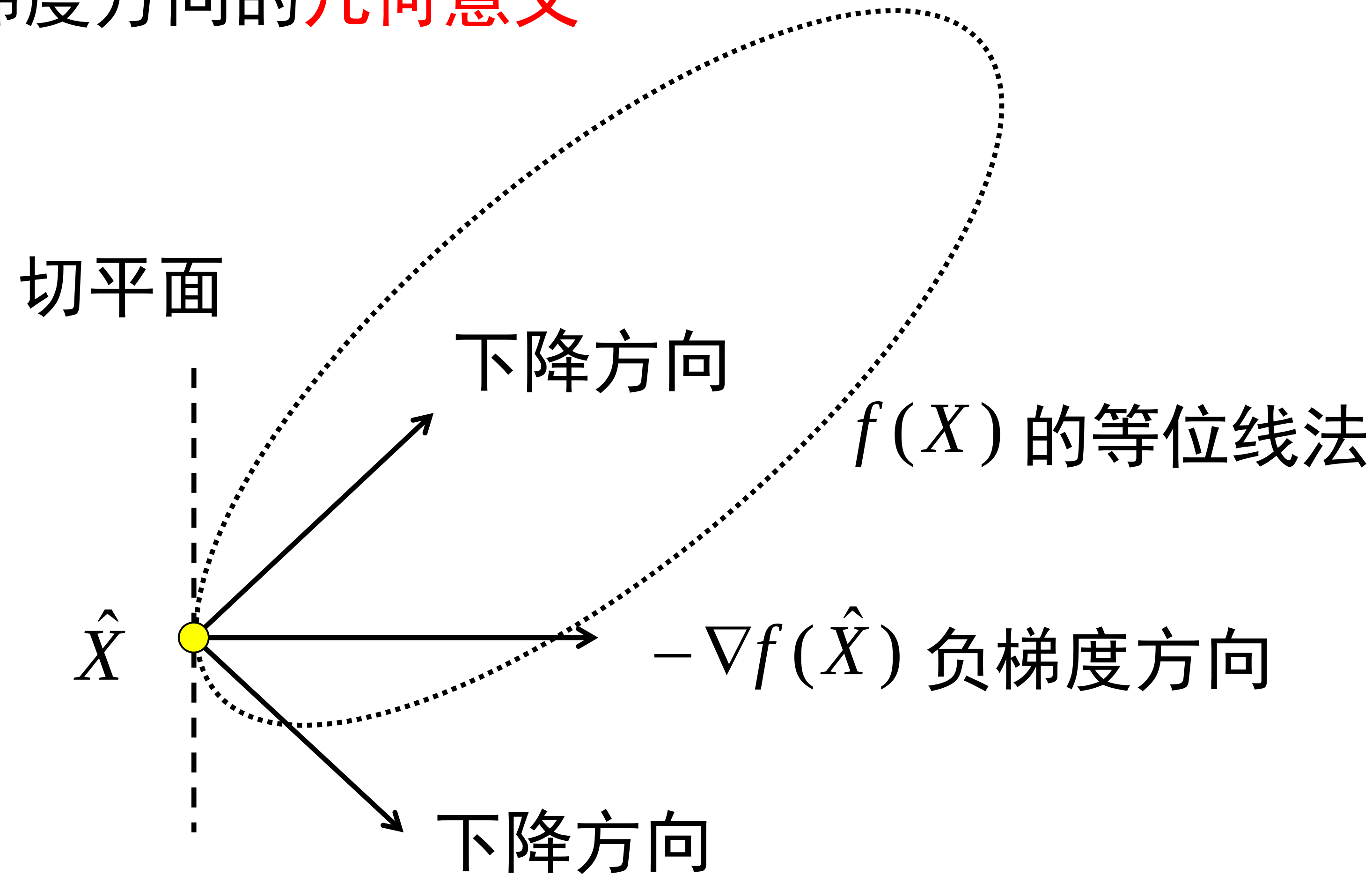
1）任取 $\hat{X} \in R^n$

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 $f\left(\hat{X} + \hat{t}D\right) < f\left(\hat{X}\right)$ 的 $\hat{t} > 0$

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

负梯度方向的几何意义



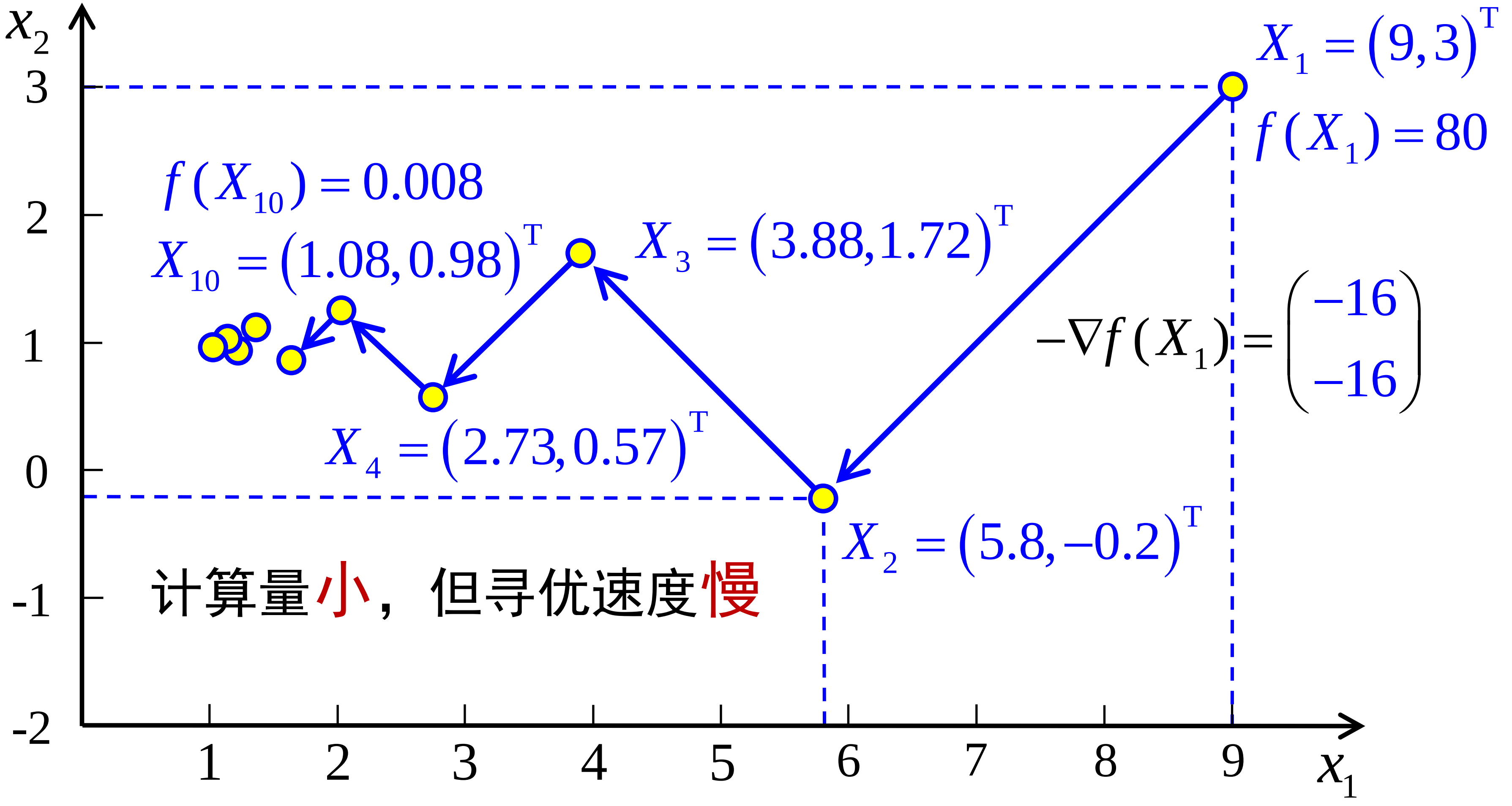
负梯度方向和切平面垂直

梯度下降法的寻优过程

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

梯度下降法

$$X_{k+1}=X_k-\lambda_k^*\nabla f(X_k)$$



计算量小，但寻优速度慢

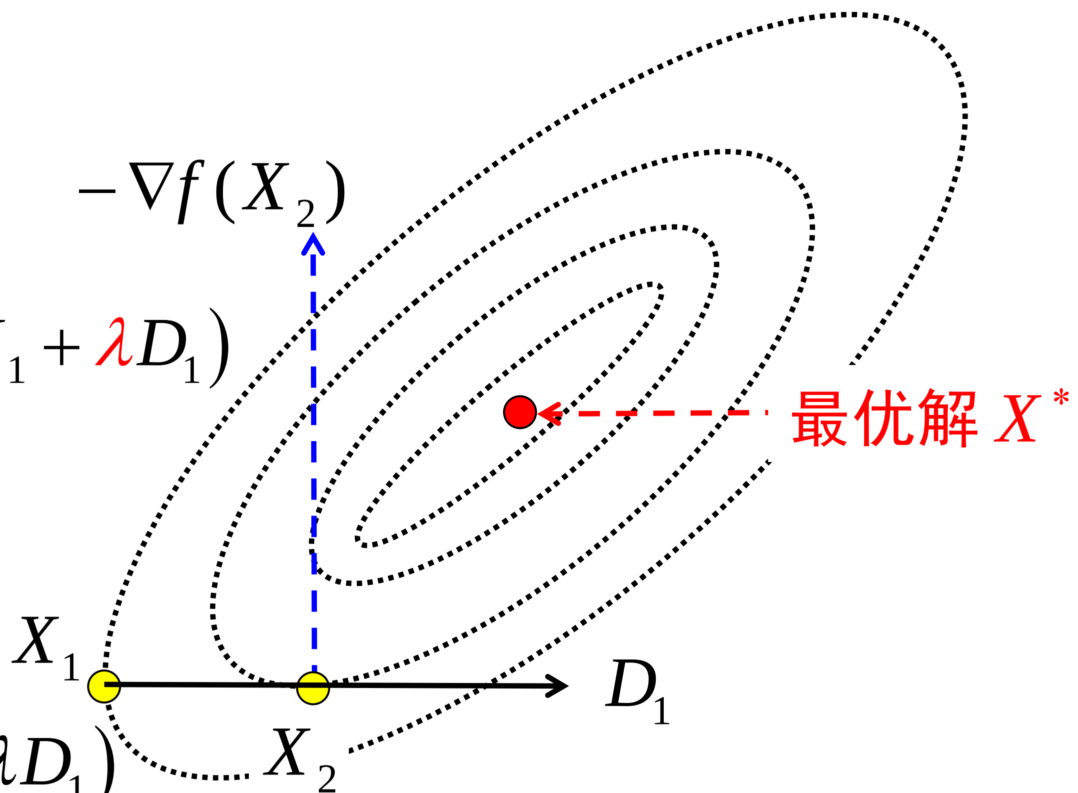
要点：负梯度方向的缺陷

一维精确搜索的特性

下降方向 D_1

$$X_2 = X_1 + \lambda_1^* D_1$$

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



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

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

精确搜索得到新点的梯度方向与搜索方向正交

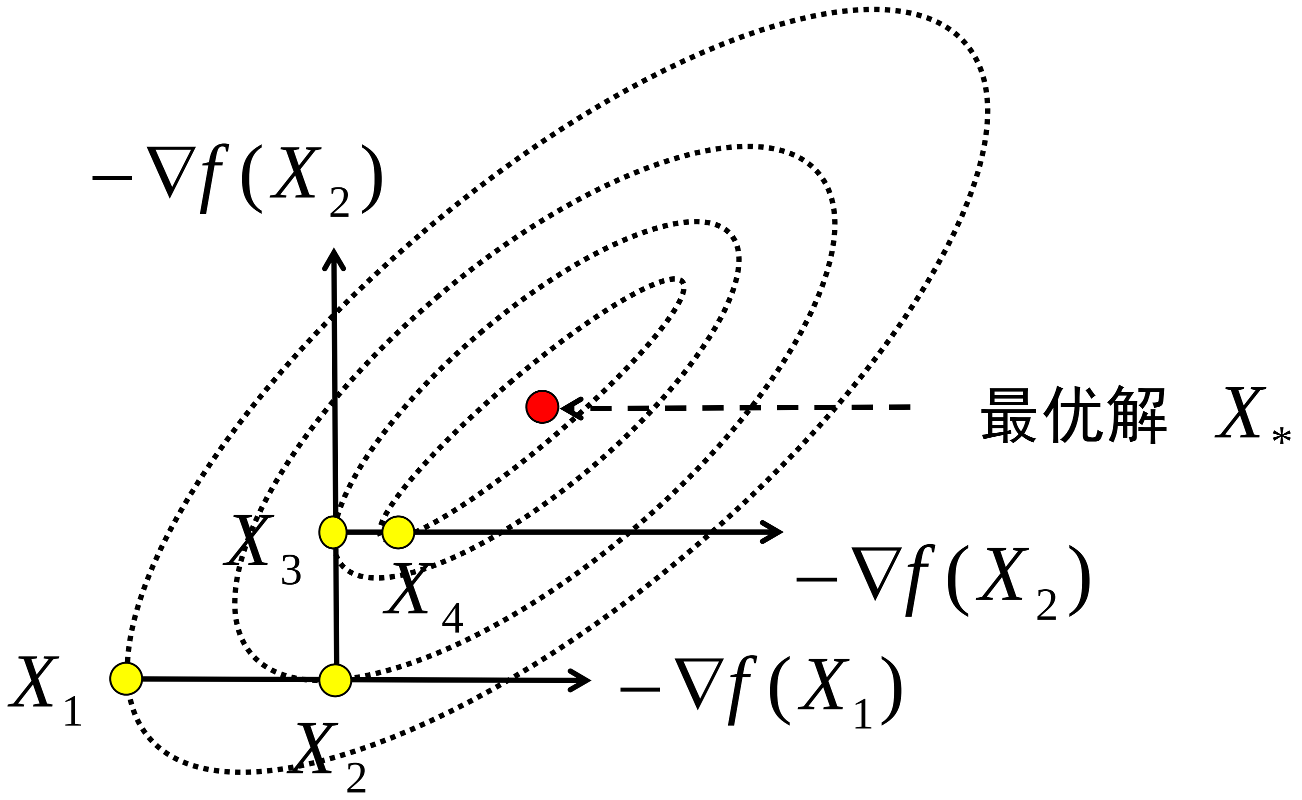
负梯度方向的特点

设 $\hat{X}'$ 是在 $\hat{X}$ 处沿负梯度方向 $D=-\nabla f(\hat{X})$ 进行一维搜索能得到的最好的点，由前面的结果可知

$$\nabla^T f(\hat{X}+\hat{t}D)\nabla f(\hat{X})=0$$

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

负梯度方向的缺陷



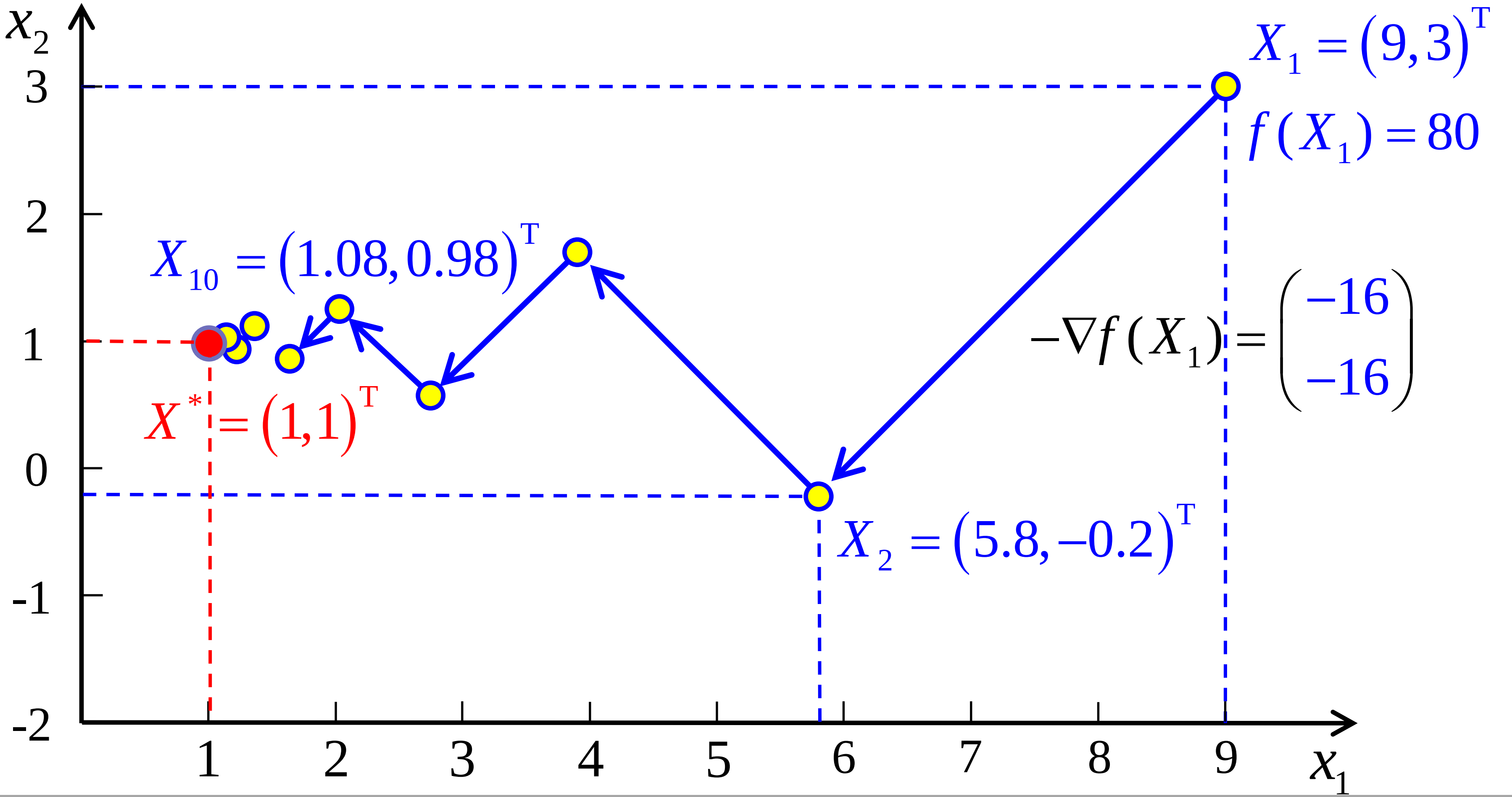
梯度下降法是沿锯齿状路线前进，接近最优解时一维搜索效率很低，前进速度很慢

改进梯度下降法的思路

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

梯度下降法

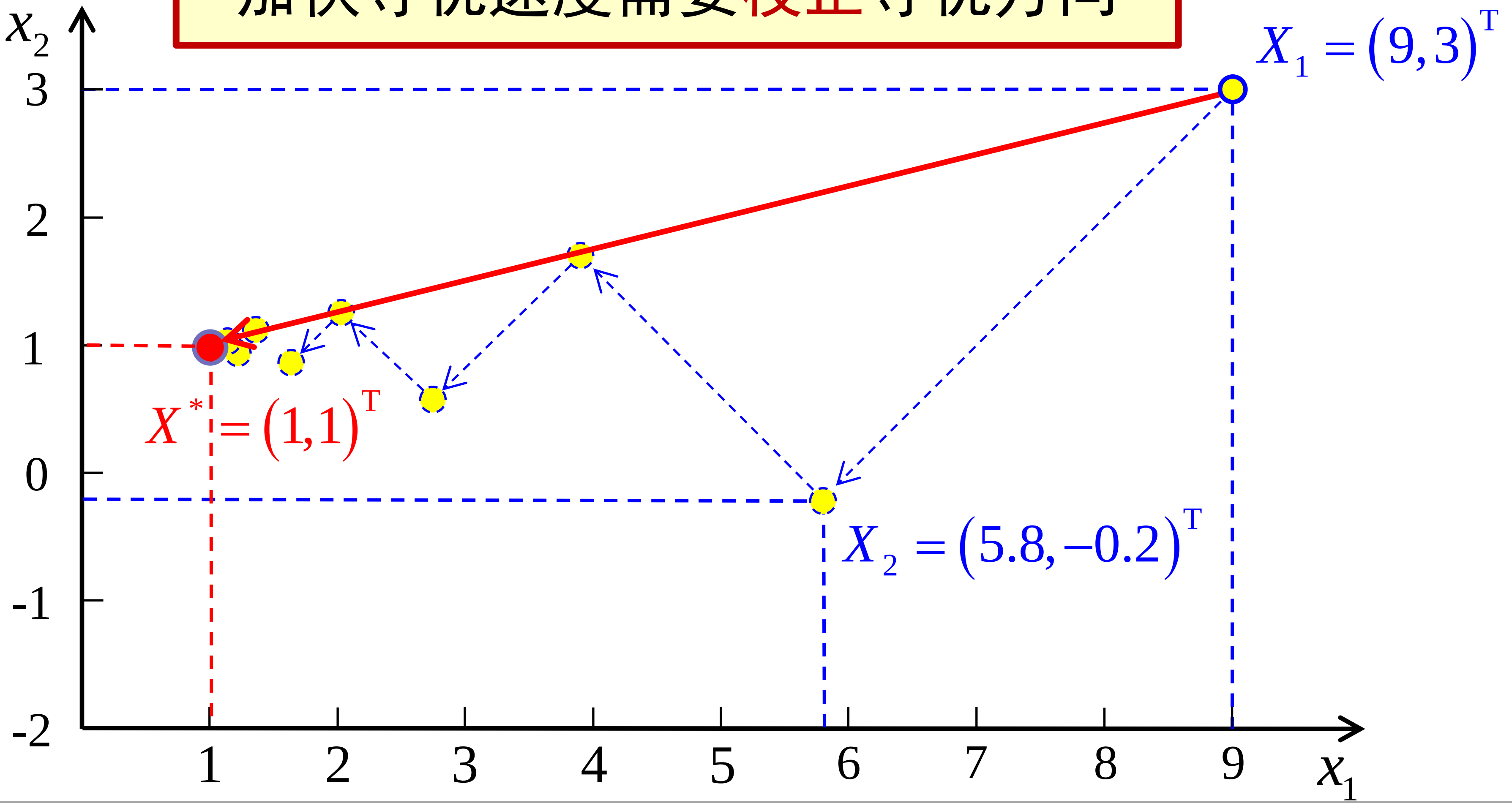
$$X_{k+1} = X_k - \lambda_k^* \nabla f(X_k)$$



改进梯度下降法的思路

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

加快寻优速度需要**校正**寻优方向



要点：利用梯度方向生成其它下降方向

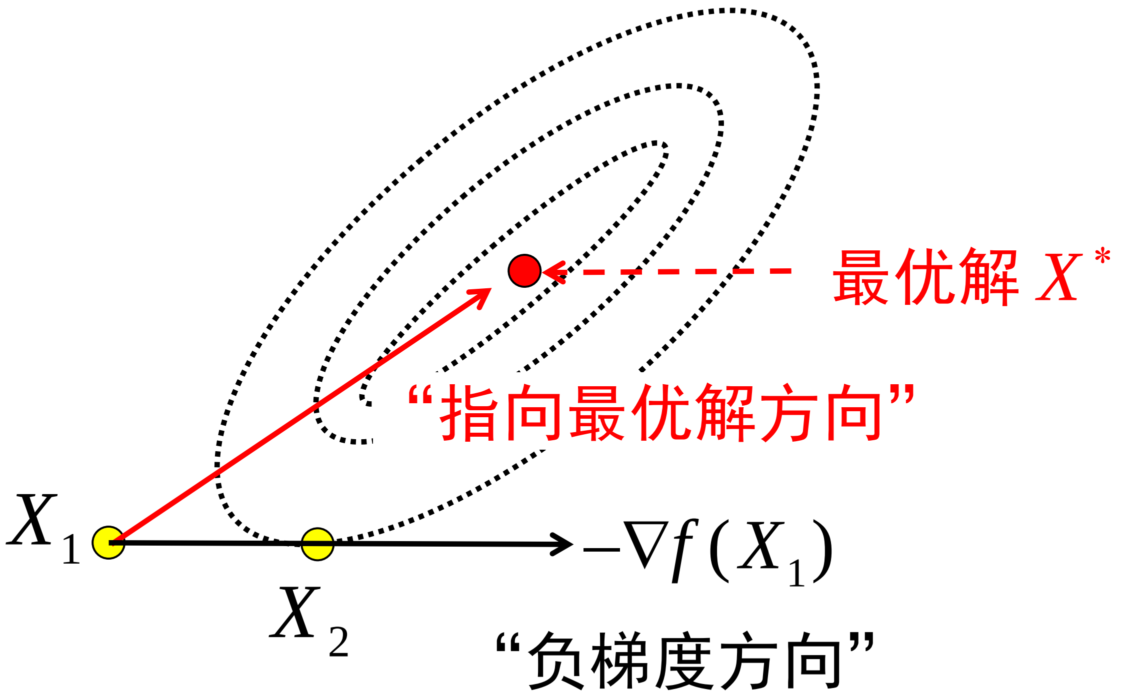
二次函数 $f(X)=\frac{1}{2}(X-X^*)^T A(X-X^*)$，其中$A\in R^{n\times n}$为对称正定矩阵，$X^*$是固定点

该函数等值面是以X^*为中心的椭球面，显然X^*为极小值点。

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

沿梯度方向：

$$X_2=X_1-\lambda_1\nabla f(X_1)$$



利用梯度方向生成其它下降方向

任取一个正定矩阵 $Q \in R^{n\times n}$，令 $D = -Q\nabla f(X)$

代入二阶泰勒展开可得

$$f(X+tD) - f(X)$$

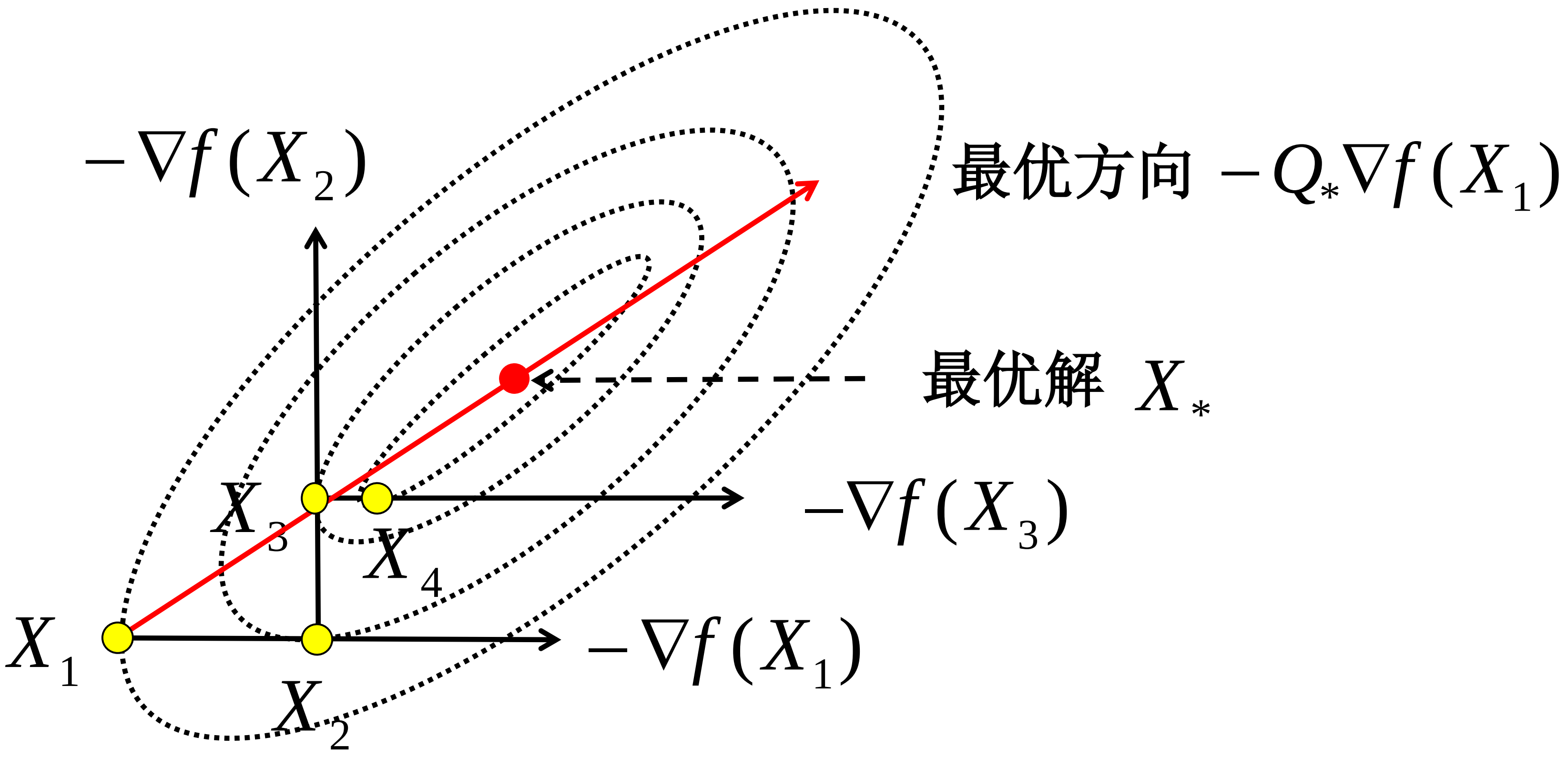
$$= -t\left(\nabla^T f(X)Q\nabla f(X) - \frac{1}{2}D^T\nabla^2 f(X+\xi D)Dt\right)$$

只要 $\nabla f(X) \neq 0$，就有 $\nabla^T f(X)Q\nabla f(X) > 0$，一定存在 $\bar{t} > 0$ 满足

$$f(X+tD) < f(X), \forall 0 < t \leq \bar{t}$$

所以 D 是下降方向

克服负梯度方向缺陷的途径



用适当的正定矩阵（尺度矩阵）乘负梯度方向，其作用是对后者进行适当的旋转，以获得更好的方向

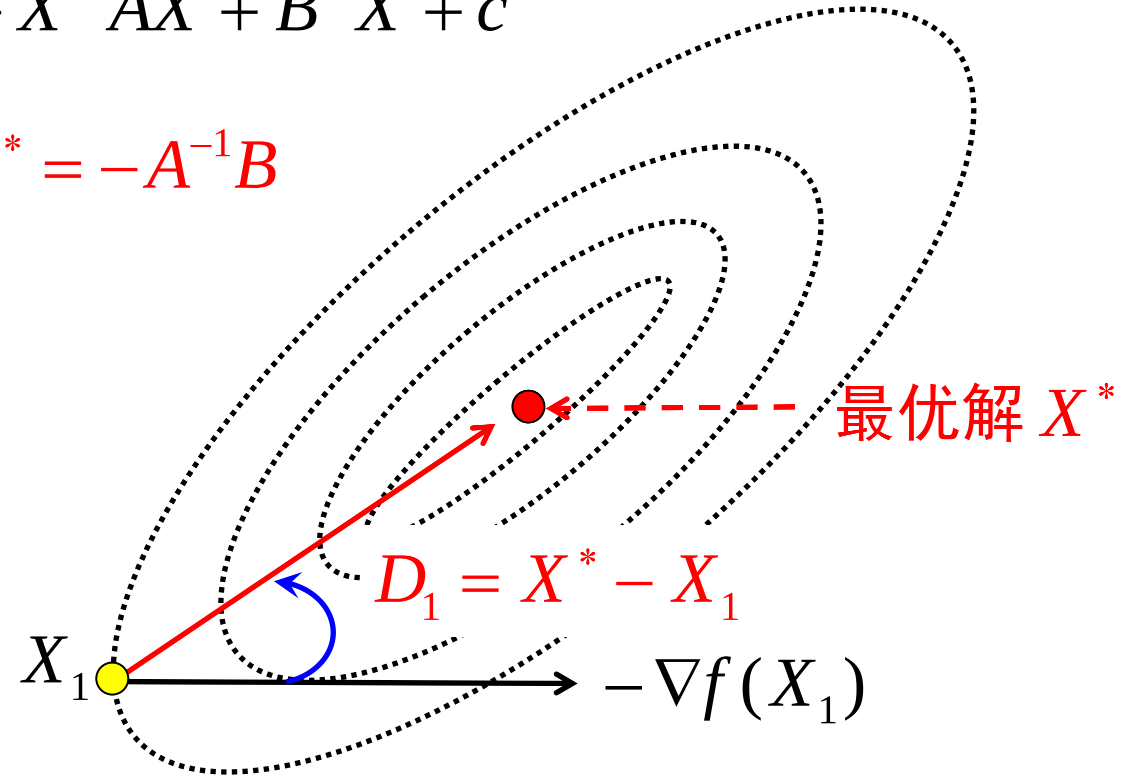
要点：牛顿方向（广义牛顿法）

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 $f(X)=\frac{1}{2}X^T AX+B^T X+c$

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

$$\begin{aligned} D_1 &= X^*-X_1 \\ &= -A^{-1}B-X_1 \\ &= -A^{-1}(AX_1+B) \end{aligned}$$

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



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

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

正定二次函数的最优方向

对正定二次函数 $f(X)=\frac{1}{2}X^TAX+B^TX+c$

$$\nabla f(X)=AX+B=0 \quad \Rightarrow \quad X_*=-A^{-1}B$$

$$\nabla^2 f(\hat{X})=A$$

从任何 $\hat{X}$ 出发，令

$$D_*=X_*-\hat{X}=-A^{-1}\left(A\hat{X}+B\right)=-\left(\nabla^2 f(\hat{X})\right)^{-1}\nabla f(\hat{X})$$

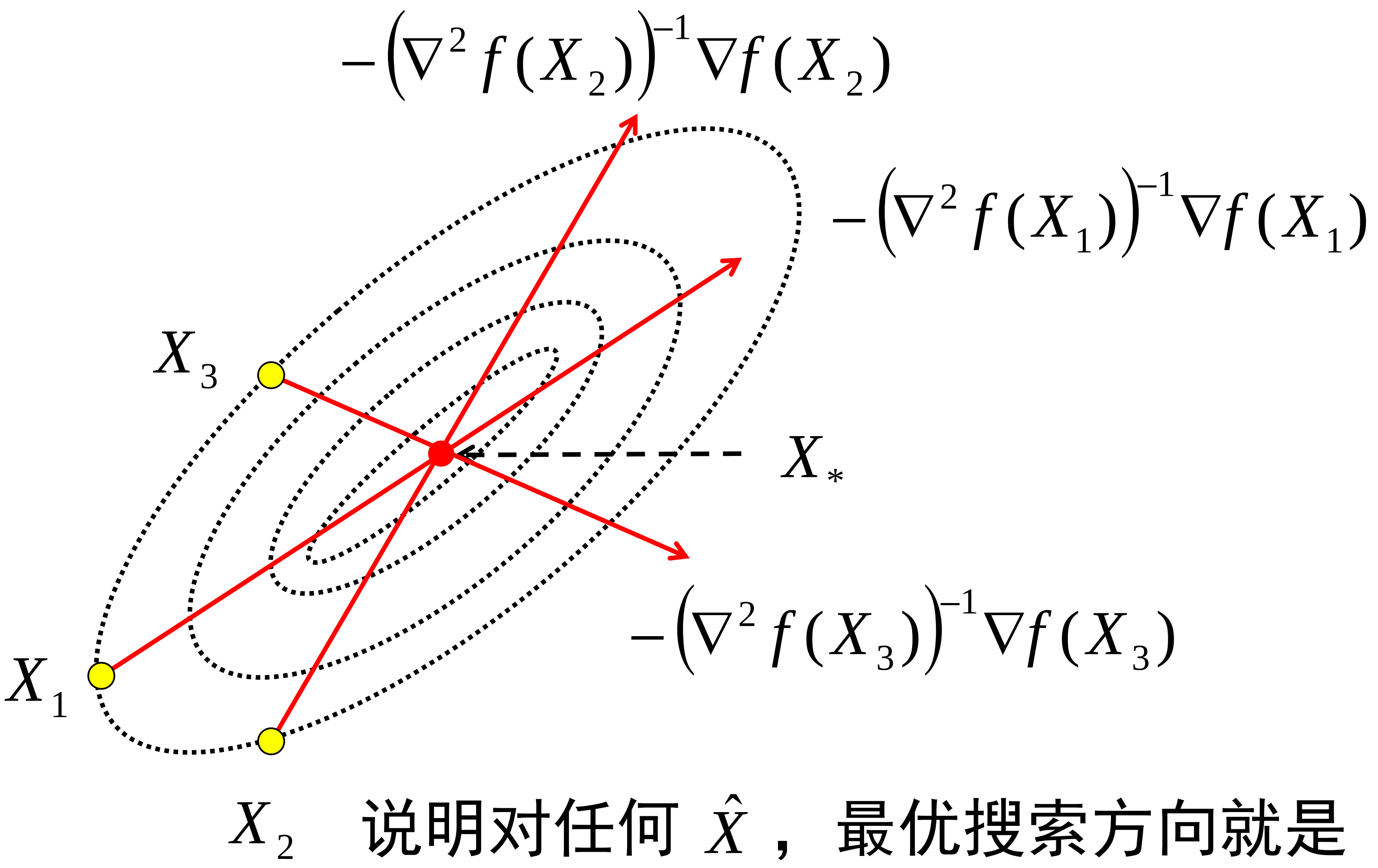
$$t_*=1$$

显然成立 $\min\limits_{t>0} f(\hat{X}+tD_*)=f(\hat{X}+t_*D_*)=f(X_*)$

说明对任何 $\hat{X}$，最优搜索方向就是

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

正定二次函数的牛顿方向



说明对任何 $\hat{X}$，最优搜索方向就是

$$D = -\left(\nabla^2 f(\hat{X})\right)^{-1}\nabla f(\hat{X})$$

广义牛顿法

1）任取 $\hat{X} \in R^n$

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

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

要点：牛顿方向的缺陷

牛顿方向的缺陷

用阻尼牛顿法求解下列问题

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

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，用阻尼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 $\left(\nabla^2 f(\hat{X})\right)^{-1}$，计算量大

2）$\left(\nabla^2 f(\hat{X})\right)^{-1}$ 可能不存在

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

要点：最速下降方向

给定方向的二阶泰勒展开

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

下降方向的充分条件 $\nabla^T f(X)D<0$

下降方向的必要条件 $\nabla^T f(X)D\le 0$

最速下降方向

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

最速下降方向

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

ℓ_1 范数 $\|D\|_1 = \sum_{i=1}^{n} |d_i|$

解决思路：

$$-\nabla^T f(X)D \le \sum_{i=1}^{n} \left| \frac{\partial f(X)}{\partial x_i} \right| |d_i| \le \max_{1\le i\le n} \left| \frac{\partial f(X)}{\partial x_i} \right| \sum_{i=1}^{n} |d_i|$$

$$\Downarrow$$

$$\|\nabla f(X)\|_\infty$$

最速下降方向 $\max\left\{ -\nabla^T f(X)D \,\middle|\, \text{s.t. } \|D\|=1 \right\}$

ℓ_1 范数 $\|D\|_1 = \sum_{i=1}^{n} |d_i|$

$$-\nabla^T f(X)D \le \|\nabla f(X)\|_\infty \|D\|_1$$

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

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

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

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

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

负梯度方向（ℓ_2 范数最速下降方向）

$$\hat{D} = -\nabla f(X)\left(\left\|\nabla f(X)\right\|_2\right)^{-1} \Leftrightarrow -\nabla f(X)$$

$$\nabla f(X)^T \hat{D} = -\left\|\nabla f(X)\right\|_2$$

牛顿方向（$\left\|D\right\|_{\nabla^2 f(X)} = \left(D^T \nabla^2 f(X) D\right)^{\frac{1}{2}}$ 的最速下降方向）

$$\hat{D} = \frac{-\left(\nabla^2 f(X)\right)^{-1} \nabla f(X)}{\left(\nabla f(X)^T \left(\nabla^2 f(X)\right)^{-1} \nabla f(X)\right)^{\frac{1}{2}}} \Leftrightarrow -\left(\nabla^2 f(X)\right)^{-1} \nabla f(X)$$

$$\nabla f(X)^T \hat{D} = -\nabla f(X)^T \left(\nabla^2 f(X)\right)^{-1} \nabla f(X)$$

要点：共轭梯度方向

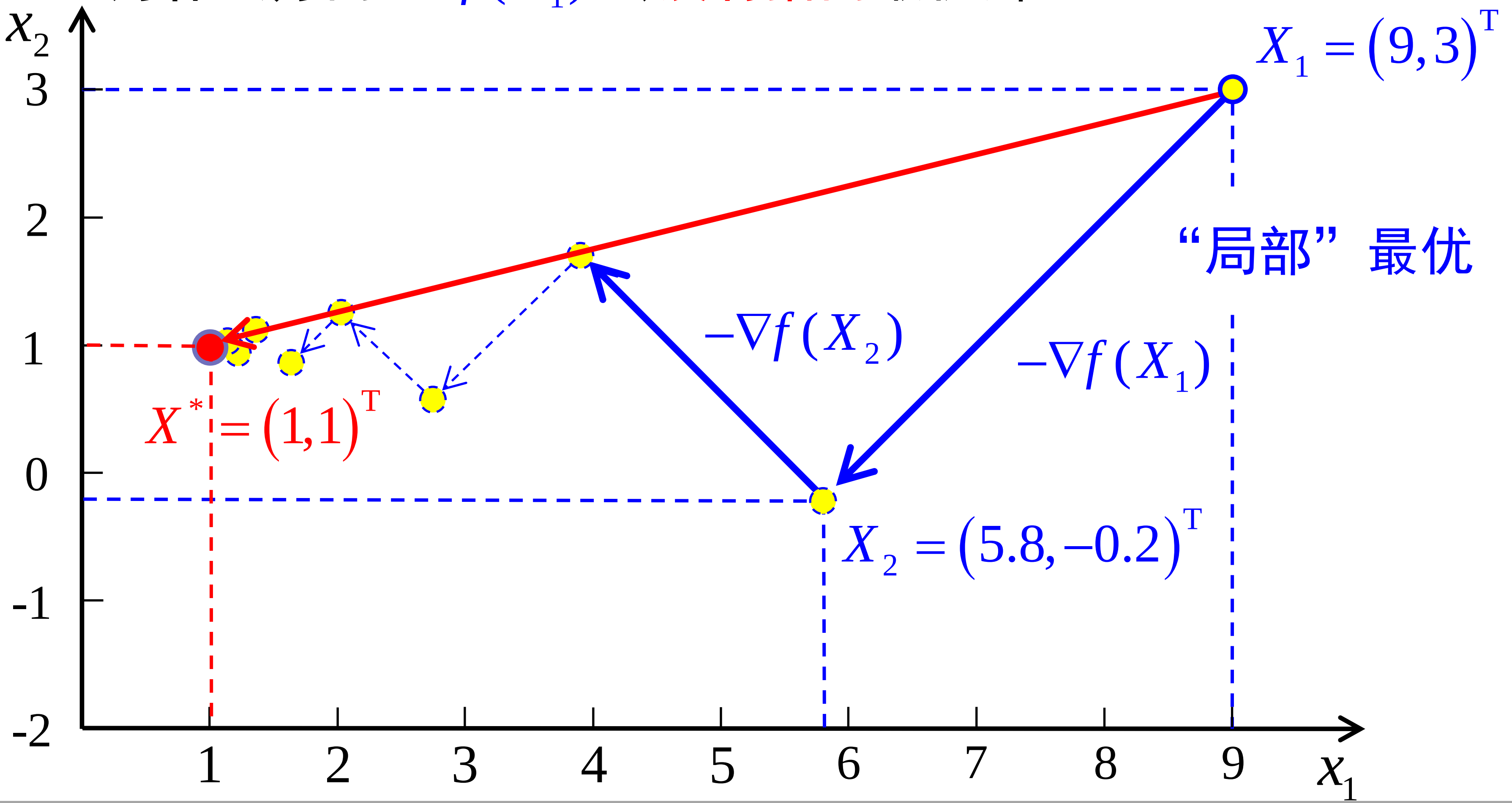
改进梯度下降法的思路

柯西：梯度下降法

$$X_{k+1} = X_k - \lambda_k^* \nabla f(X_k)$$

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

为什么方向 $-\nabla f(X_1)$ 会没有指向最优解 X^*？

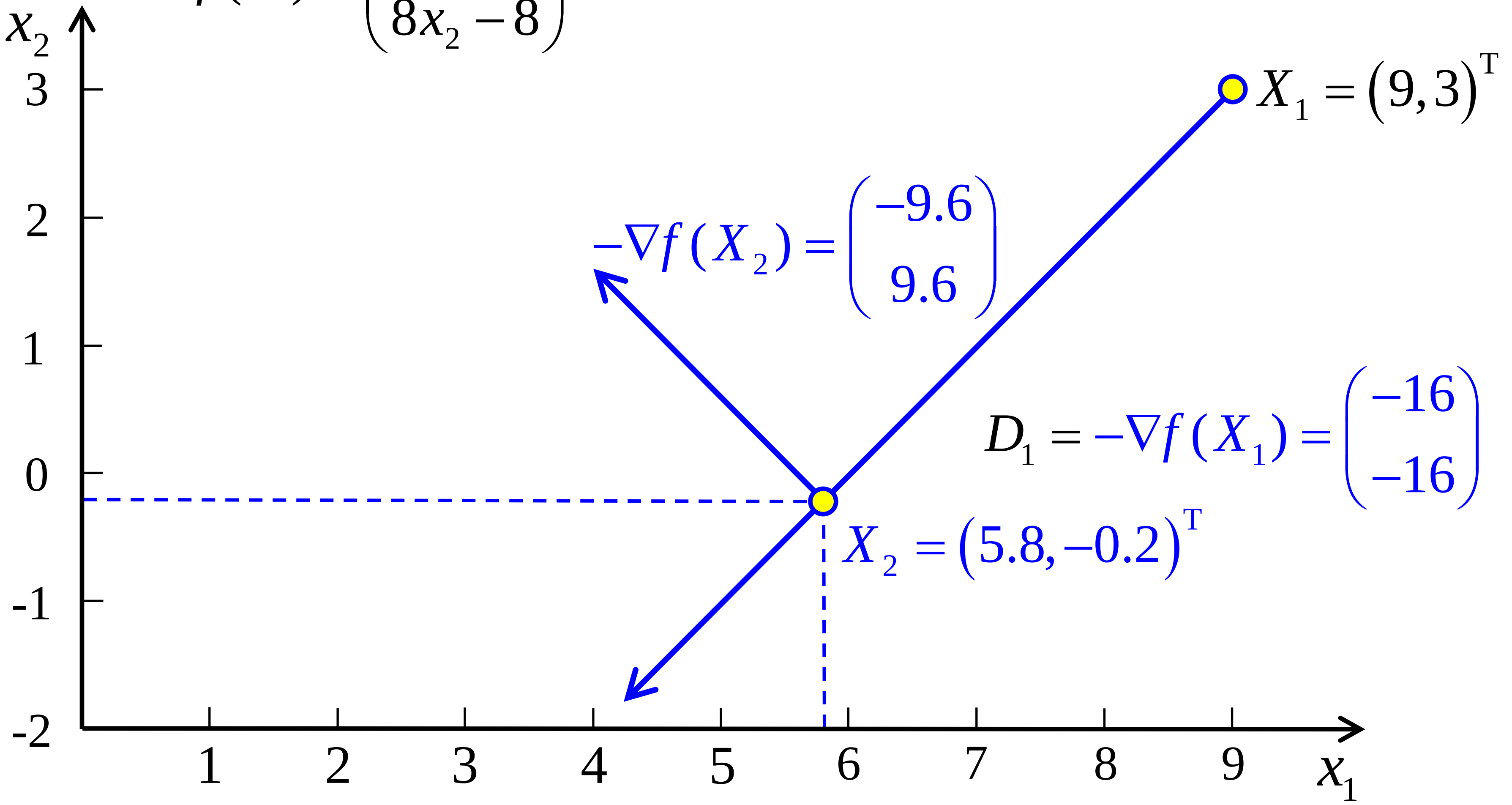


F-R 共轭梯度法 —— 利用共轭梯度寻优

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

第一步沿负梯度寻优

$$\nabla f(X) = \begin{pmatrix} 2x_1 - 2 \\ 8x_2 - 8 \end{pmatrix}$$

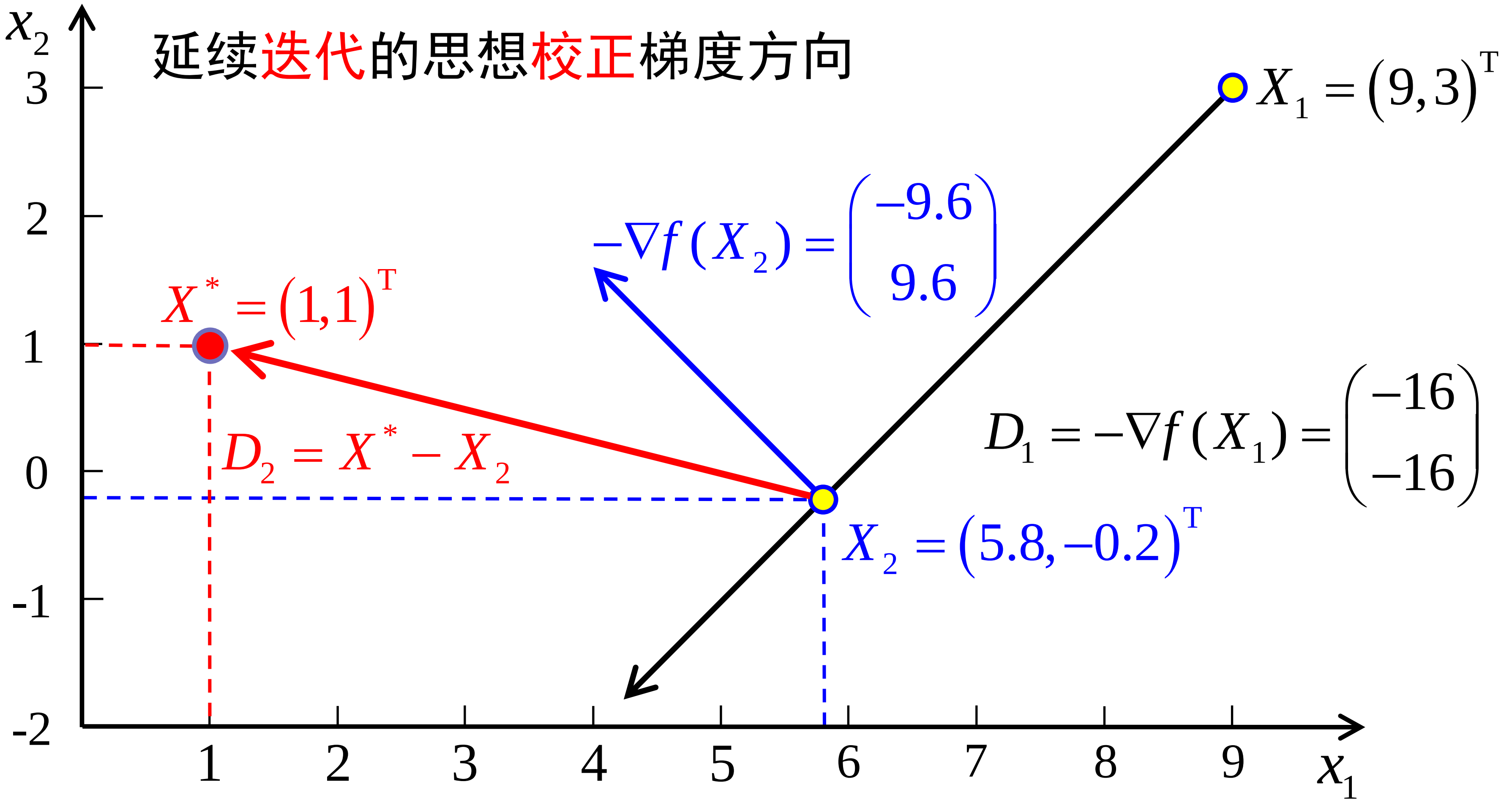


F-R 共轭梯度法 —— 利用共轭梯度寻优

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

$$D_2 = -\nabla f(x_2) + \alpha_1 D_1 \qquad X_{k+1} = X_k + \lambda_k D_k$$

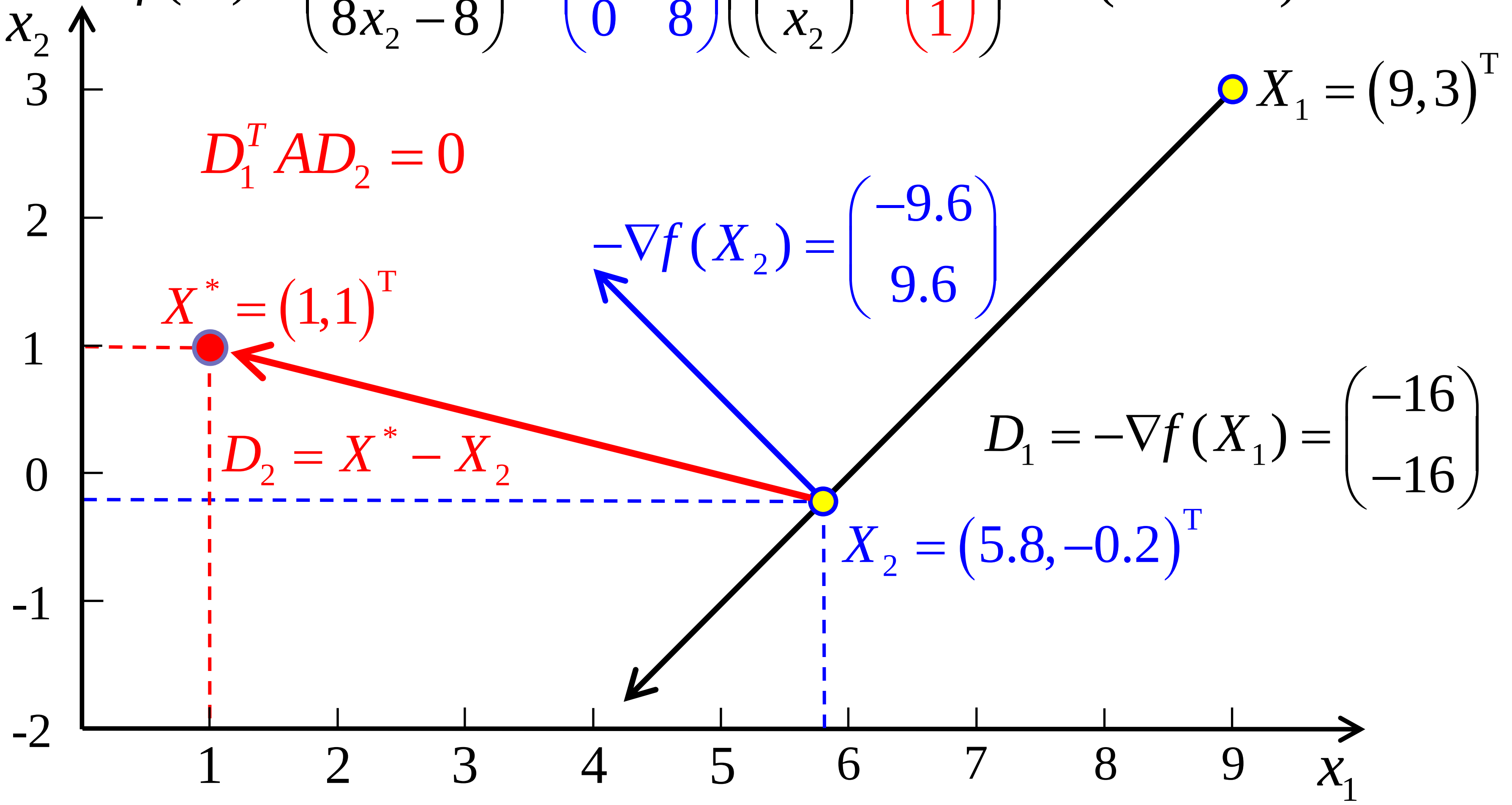
延续迭代的思想校正梯度方向



F-R 共轭梯度法 —— 利用共轭梯度寻优

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

$$\nabla f(X) = \begin{pmatrix} 2x_1 - 2 \\ 8x_2 - 8 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 8 \end{pmatrix}\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right) = A(X - X^*)$$



共轭方向法原理之一

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

共轭方向线性无关性

若 $\vec{p}_0,\vec{p}_1,\cdots,\vec{p}_{n-1}$ 互为 $A>0$ 的共轭方向，则它们线性无关

理由：$\alpha_0\vec{p}_0+\alpha_1\vec{p}_1+\cdots+\alpha_{n-1}\vec{p}_{n-1}=0$

$$\Rightarrow\ \alpha_0\vec{p}_k^{T}A\vec{p}_0+\alpha_1\vec{p}_k^{T}A\vec{p}_1+\cdots+\alpha_{n-1}\vec{p}_k^{T}A\vec{p}_{n-1}=0$$

$$\Rightarrow\ \alpha_k\vec{p}_k^{T}A\vec{p}_k=0$$

$$\Rightarrow\ \alpha_k=0$$

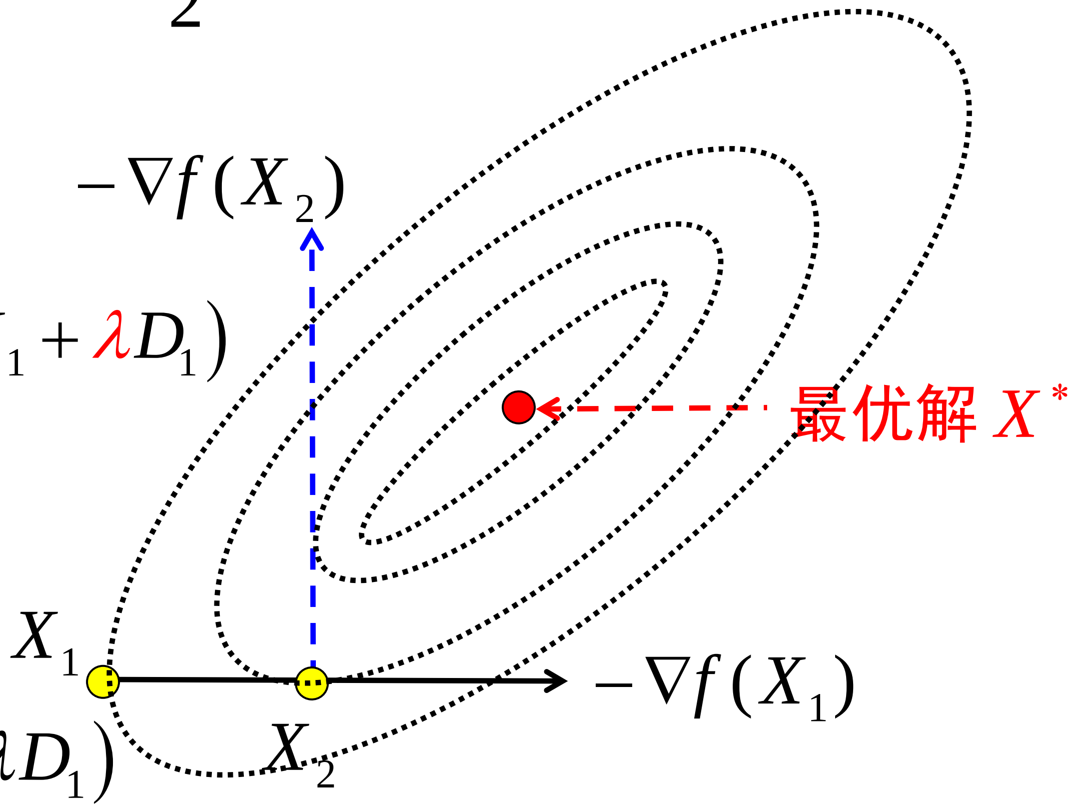
共轭梯度方向 $D_k = -\nabla f(X_k) + \alpha_{k-1} D_{k-1}$

对于二次正定函数 $f(X) = \dfrac{1}{2}(X - X^*)^T A(X - X^*)$

$D_1 = -\nabla f(X_1)$

$X_2 = X_1 + \lambda_1^* D_1$

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



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

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

$$f(X)=\frac{1}{2}(X-X^*)^T A(X-X^*) \qquad D_k=-\nabla f(X_k)+\alpha_{k-1}D_{k-1}$$

$$D_1=-\nabla f(X_1)$$

$$D_2=-\nabla f(X_2)+\alpha_1 D_1$$

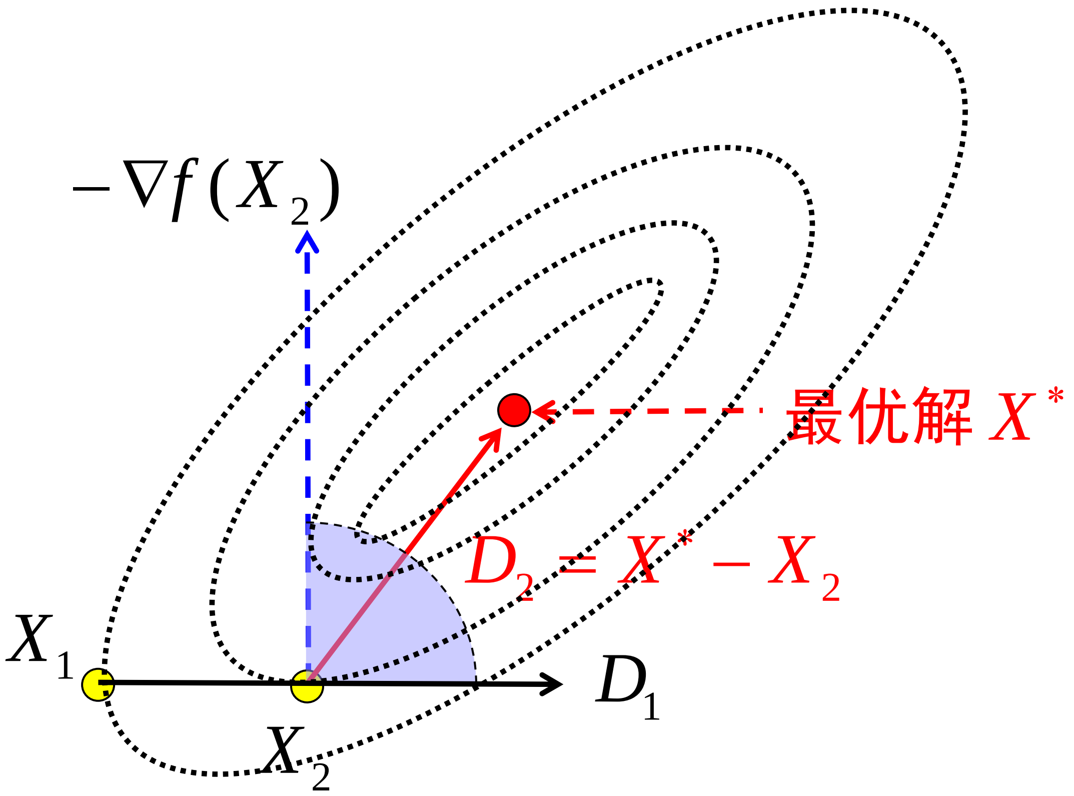
$$\nabla f(X_2)=A(X_2-X^*)$$

$$=-AD_2$$

$$D_1^T\nabla f(X_2)=0$$

$$\Rightarrow \quad -D_1^T AD_2=0$$

$$\Rightarrow \quad D_1^T AD_2=0$$



D_1与 D_2为 A 的共轭方向！

要点：F-R共轭梯度法

F-R 共轭梯度法 —— 利用共轭梯度寻优

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

$$A = \begin{pmatrix} 2 & 0 \\ 0 & 8 \end{pmatrix}$$

"长记性"寻优方向 D_2 和 D_1 是两个 A 的共轭方向

$$D_2 = -\nabla f(x_2) + \alpha_1 D_1$$

共轭梯度

只需要解决如何计算出合适的参数 α_1

1952年Hestenes和Stiefel提出利用共轭梯度

求解线性方程组 $AX = b,\ \ X \in \mathrm{R}^n$

$$\min\left(X^{\mathrm{T}}AX - b^{\mathrm{T}}X\right),\ \ X \in \mathrm{R}^n$$

Fletcher：用"简单"解决"复杂"

R. Fletcher 英国

皇家科学院院士

F-R 共轭梯度法 —— F-R共轭梯度法

1964年，Fletcher 和 Reeves提出了适用于一般无约束最优化问题的求解方法：F-R 共轭梯度法

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

$$D_k = \begin{cases} -\nabla f(X_k) & k = 1 \\ -\nabla f(X_k) \boxed{+ \alpha_{k-1} D_{k-1}} & k \geq 2 \end{cases}$$

梯度下降法

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

$$D_k = -\nabla f(X_k)$$

相邻两步寻优方向共轭性 $D_k^{\mathrm{T}} A D_{k-1} = 0$ 和精确搜索的特点

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

F-R 共轭梯度法计算简单、寻优速度快，在国际上开启了共轭梯度法求解非线性规划的研究先河！

要点：参数 α 的计算

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

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

$$D_1=-\nabla f(X_1)$$

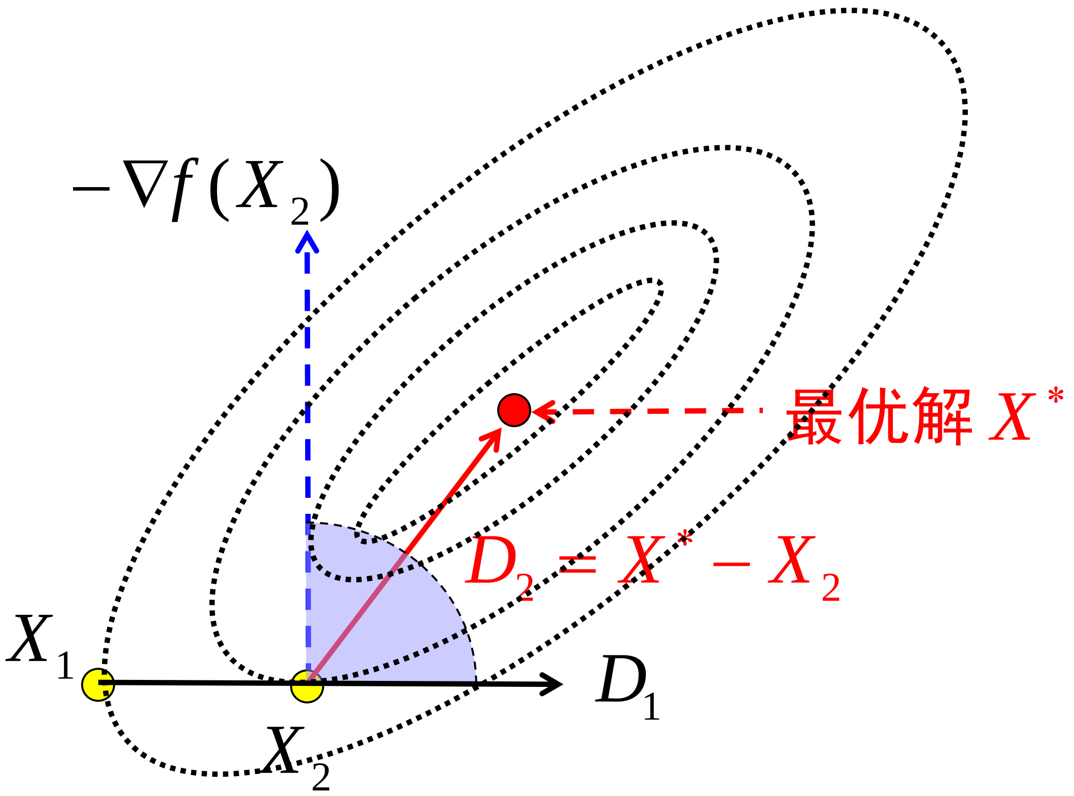
$$D_2=-\nabla f(X_2)+\alpha_1 D_1$$

$$\nabla f(X_2)=A(X_2-X^*)$$
$$=-AD_2$$

$$D_1^T\nabla f(X_2)=0$$

$$\Rightarrow\quad -D_1^T AD_2=0$$

$$\Rightarrow\quad D_1^T AD_2=0\quad\Rightarrow\quad -D_1^T A\nabla f(X_2)+\alpha_1 D_1^T AD_1=0$$



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

D_1与 D_2为 A 的共轭方向！

参数 α 中矩阵 A 的消除方法

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

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

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

$$= \frac{\nabla^T f(X_2)\left(\nabla f(X_2) - \nabla f(X_1)\right)}{D_1^T \left(\nabla f(X_2) - \nabla f(X_1)\right)} = \frac{\nabla^T f(X_2)\left(\nabla f(X_2) + D_1\right)}{D_1^T \left(\nabla f(X_2) + D_1\right)}$$

$$= \frac{\nabla^T f(X_2) \nabla f(X_2)}{D_1^T D_1} = \frac{\left\|\nabla f(X_2)\right\|^2}{\left\|\nabla f(X_1)\right\|^2}$$

要点：F-R共轭梯度法计算示例

F-R 共轭梯度法 —— 寻优速度对比

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

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

$$D_k = \begin{cases} \boxed{-\nabla f(X_k) \qquad k = 1} \\ \boxed{-\nabla f(X_k) + \alpha_{k-1} D_{k-1} \qquad k \geq 2} \end{cases}, \quad \boxed{\alpha_k = \frac{\|\nabla f(X_{k+1})\|^2}{\|\nabla f(X_k)\|^2}}$$

$$X_1 = (9,3)^{\mathrm{T}}$$

F-R 法计算步骤

① $D_1 = -\nabla f(X_1)$

② $\min\limits_{\lambda_1 > 0} f(X_1 + \lambda_1 D_1)$, $X_2 = X_1 + \lambda_1^* D_1$

③ $\alpha_1 = \dfrac{\|\nabla f(X_2)\|^2}{\|\nabla f(X_1)\|^2}$

④ $D_2 = -\nabla f(X_2) + \alpha_1 D_1$

⑤ $\min\limits_{\lambda_2 > 0} f(X_2 + \lambda_2 D_2)$, $X_3 = X_2 + \lambda_2^* D_2$

计算结果

$D_1 = -(16,16)^{\mathrm{T}}$

$\lambda_1^* = 0.2$, $X_2 = (5.8, -0.2)^{\mathrm{T}}$

$\nabla f(X_2) = (9.6, -9.6)^{\mathrm{T}}$

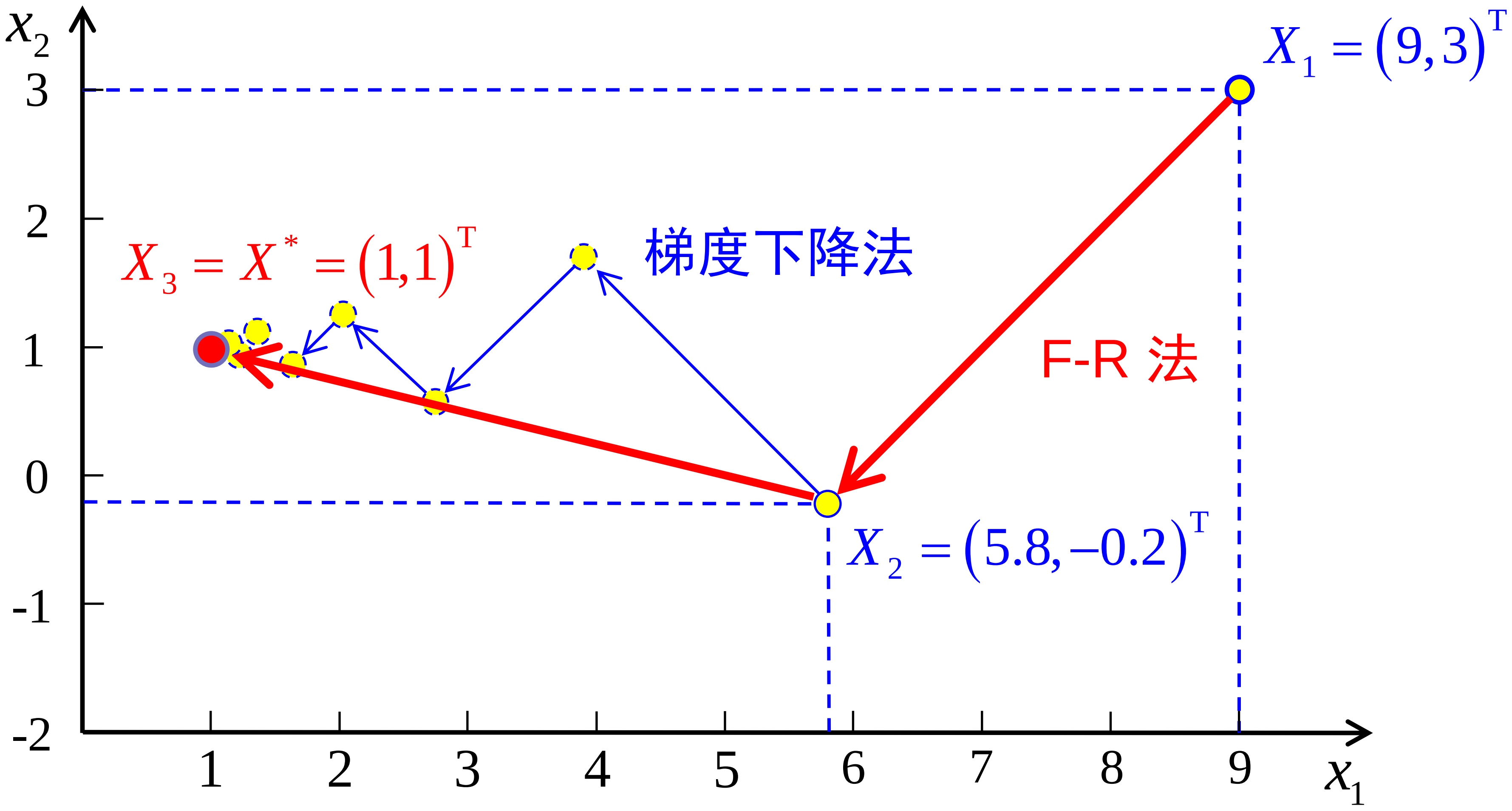
$\alpha_1 = 0.36$

$D_2 = (-15.36, 3.84)^{\mathrm{T}}$

$\lambda_2^* = 0.3125$, $X_3 = (1,1)^{\mathrm{T}}$

F-R 共轭梯度法 —— 寻优轨迹对比

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



要点：与一维最优解的梯度的正交性

共轭方向和一维最优解的梯度的正交性

条件： $f(X)=0.5X^TAX+B^TX+C$， $A>0$

$\vec{p}_0,\vec{p}_1,\cdots,\vec{p}_{n-1}$ 为 A 的共轭方向

$X_0\in R^n$ 是任意的出发点

由下述一维搜索依次确定 $X_1,X_2,\cdots,X_n$

$$f(X_{k+1})=f(X_k+t_k\vec{p}_k)=\min_{t\in R} f(X_k+t\vec{p}_k)$$

$$k=0,1,\cdots,n-1$$

结论： $\vec{p}_j^T\nabla f(X_k)=0,\ \forall 0\le j<k$

理由： $\min_{t>0} f(X_k + t\vec{p}_k) \quad \Rightarrow \quad t_k = -\dfrac{\vec{p}_k^T \nabla f(X_k)}{\vec{p}_k^T A\vec{p}_k}, \forall 0 \le k \le n-1$

$$X_k = X_{k-1} + t_{k-1}\vec{p}_{k-1} \quad \Rightarrow \quad X_k = X_0 + \sum_{i=0}^{k-1} t_i \vec{p}_i$$

$$\Rightarrow \quad \nabla f(X_k) = \nabla f(X_0) + \sum_{i=0}^{k-1} t_i A\vec{p}_i$$

$$\begin{aligned} \vec{p}_j^T \nabla f(X_k) &= \vec{p}_j^T \nabla f(X_0) + t_j \vec{p}_j^T A\vec{p}_j \\ \Rightarrow \quad &= \vec{p}_j^T \nabla f(X_0) - \vec{p}_j^T \nabla f(X_j), \ \forall 0 \le j < k \end{aligned}$$

$$\vec{p}_j^T \nabla f(X_j) = \vec{p}_j^T \nabla f(X_0), \ \forall j$$

$$\Rightarrow \quad \vec{p}_j^T \nabla f(X_k) = 0, \ \forall 0 \le j < k$$

推论：沿共轭方向寻优的每个 $X_k, k=1,2,\cdots,n$ 都满足

$$f\left(X_k\right)=\min\left\{ f\left(X\right) \middle| \text{ s.t. } X = X_0+\sum_{j=0}^{k-1}\beta_j \vec{p}_j \right\}$$

理由： $X = X_0+\sum_{j=0}^{k-1}\beta_j \vec{p}_j,\ X_k = X_0+\sum_{j=0}^{k-1}\beta_{kj} \vec{p}_j$

$$\Rightarrow \quad \nabla f\left(X_k\right)^T\left(X-X_k\right)=\sum_{j=0}^{k-1}\nabla f\left(X_k\right)^T \vec{p}_j\left(\beta_j-\beta_{kj}\right)$$

利用 $\nabla f\left(X_k\right)^T \vec{p}_j=0,\ j=0,1,\cdots,k-1$

可得 $\nabla f\left(X_k\right)^T\left(X-X_k\right)=0$

再利用凸函数一阶充要条件可得结论

要点：共轭方向二次函数有限终止性

共轭方向二次函数有限终止性

条件： $f(X)=0.5X^TAX+B^TX+c$，A对称正定

$\vec{p}_0,\vec{p}_1,\cdots,\vec{p}_{n-1}$ 为 A 的共轭方向

$X_0\in R^n$ 是任意的出发点

由下述直线搜索依次确定 $X_1,X_2,\cdots,X_n$

$$f(X_{k+1})=f(X_k+t_k\vec{p}_k)=\min_{t>0} f(X_k+t\vec{p}_k)$$

$$k=0,1,\cdots,n-1$$

结论： $f(X_n)=\min_{X\in R^n} f(X)$

理由：1）由推论可知

$$f\left(X_n\right)=\min\left\{f\left(X\right)\middle|\ \text{s.t. } X = X_0+\sum_{j=0}^{n-1}\vec{p}_j\beta_j\right\}$$

2）由原理之一可知 $R^n=\left\{X\middle| X = X_0+\sum_{j=0}^{n-1}\vec{p}_j\beta_j\right\}$

理由：从共轭方向的几个特点出发：

1、共轭方向线性无关 $\Rightarrow \vec{p}_0, \vec{p}_1, \cdots, \vec{p}_{n-1}$ 是R^n的一组基

2、$\nabla f(X) = AX + B$ 是 R^n 的列向量，则对于任意的 $\hat{X}$

若 $0 \neq \nabla f(\hat{X}) = \alpha_0 \vec{p}_0 + \alpha_1 \vec{p}_1 + \cdots + \alpha_{n-1} \vec{p}_{n-1}$

3、从 X_i 出发沿 $\vec{p}_i$ 直线搜索，则有 $\nabla^T f(X_k + t_k \vec{p}_k) \vec{p}_k = 0$

即 $\nabla^T f(X_{k+1}) \vec{p}_k = 0$

4、$\nabla f(X_{k+1}) = AX_{k+1} + B = A(X_k + t_k \vec{p}_k) + B = \nabla f(X_k) + t_k A\vec{p}_k$，

则有 $\vec{p}_{k-1}^T \nabla f(X_{k+1}) = \vec{p}_{k-1}^T \nabla f(X_k) + \vec{p}_{k-1}^T t_k A\vec{p}_k$，进而由3

和共轭方向性质有 $\vec{p}_{k-1}^T \nabla f(X_{k+1}) = 0$，依此类推得到

$\vec{p}_i^T \nabla f(X_{k+1}) = 0,\ i = 0,1,\cdots,k$

如果 $\nabla f(X_n) \neq 0$，则引发如下矛盾

$$\nabla^T f(X_n) \nabla f(X_n) = \nabla^T f(X_n) \left(\alpha_0 \vec{p}_0 + \alpha_1 \vec{p}_1 + \cdots + \alpha_{n-1} \vec{p}_{n-1} \right) = 0$$

要点：共轭方向的生成

共轭方向的生成

用Gram-Schmidt 正交化方法顺序生成 A 共轭方向

利用 A 共轭性确定下面方程组中所有待定系数

$$\vec{p}_0 = -\nabla f(X_0)$$

$$\vec{p}_1 = -\nabla f(X_1) + \alpha_{10}\vec{p}_0$$

$$\vdots$$

$$\vec{p}_{n-1} = -\nabla f(X_{n-1}) + \alpha_{n-10}\vec{p}_0 + \alpha_{n-11}\vec{p}_1 + \cdots + \alpha_{n-1n-2}\vec{p}_{n-2}$$

例如： $\vec{p}_0^T A\vec{p}_1 = 0 \quad \Rightarrow \quad 0 = -\vec{p}_0^T A\nabla f(X_1) + \alpha_{10}\vec{p}_0^T A\vec{p}_0$

$$\Rightarrow \quad \alpha_{10} = \frac{\vec{p}_0^T A\nabla f(X_1)}{\vec{p}_0^T A\vec{p}_0}$$

解前面方程组最终可得

$$\vec{p}_0 = -\nabla f(X_0)$$

$$\vec{p}_1 = -\nabla f(X_1) + \alpha_{10}\vec{p}_0$$

$$\vdots$$

$$\vec{p}_{n-1} = -\nabla f(X_{n-1}) + \alpha_{n-10}\vec{p}_0 + \alpha_{n-11}\vec{p}_1 + \cdots + \alpha_{n-1n-2}\vec{p}_{n-2}$$

其中

$$\alpha_{kj} = \frac{\nabla^T f(X_k) A \vec{p}_j^T}{\vec{p}_j^T A \vec{p}_j}, \quad 1 \le k \le n-1,\ 0 \le j \le k-1$$

为了应用于一般性的非线性函数，需要消除 A

消除 A 的基本途经：

$$X_k = X_{k-1} + t_{k-1}\vec{p}_{k-1} \quad \Rightarrow \quad \nabla f\left(X_k\right) = \nabla f\left(X_{k-1}\right) + t_{k-1}A\vec{p}_{k-1}$$

由推论，$\vec{p}_{k-1}^T\nabla f\left(X_k\right) = \vec{p}_{k-1}^T\nabla f\left(X_{k-1}\right) + t_{k-1}\vec{p}_{k-1}^T A\vec{p}_{k-1} = 0$

$$t_j = -\frac{\vec{p}_j^T\nabla f(X_j)}{\vec{p}_j^T A\vec{p}_j} \qquad \alpha_{kj} = \frac{\nabla^T f(X_k)A\vec{p}_j^T}{\vec{p}_j^T A\vec{p}_j} = \frac{t_j\nabla^T f(X_k)A\vec{p}_j^T}{-\vec{p}_j^T\nabla f(X_j)}$$

$$\Rightarrow \qquad = \frac{\nabla^T f(X_k)\left(\nabla f\left(X_{j+1}\right) - \nabla f\left(X_j\right)\right)}{-\vec{p}_j^T\nabla f(X_j)}$$

由于 $j \le k-1$ ，上式已经可以应用于一般性函数，再利用梯度和共轭方向的关系，可进一步简化系数表达式

$$\vec{p}_j = -\nabla f(X_j) + \alpha_{j0}\vec{p}_0 + \alpha_{j1}\vec{p}_1 + \cdots + \alpha_{j\,j-1}\vec{p}_{j-1}$$

$$\Rightarrow \quad \nabla f(X_j) = -\vec{p}_j + \alpha_{j0}\vec{p}_0 + \alpha_{j1}\vec{p}_1 + \cdots + \alpha_{j\,j-1}\vec{p}_{j-1}$$

$$\nabla^T f(X_k)\vec{p}_j = 0, \forall 0 \le j < k$$

$$\Rightarrow \quad \nabla^T f(X_k)\nabla f(X_j)$$

$$= \nabla^T f(X_k)\left(-\vec{p}_j + \alpha_{j0}\vec{p}_0 + \alpha_{j1}\vec{p}_1 + \cdots + \alpha_{j\,j-1}\vec{p}_{j-1}\right)$$

$$= 0, \ \ \forall 0 \le j < k$$

$$\nabla^T f(X_k)\nabla f(X_k)$$

$$= \nabla^T f(X_k)\left(-\vec{p}_k + \alpha_{k0}\vec{p}_0 + \alpha_{k1}\vec{p}_1 + \cdots + \alpha_{k\,k-1}\vec{p}_{k-1}\right)$$

$$= -\nabla^T f(X_k)\vec{p}_k$$

$$\nabla^T f(X_k)\nabla f(X_j)=0,\ \ \forall 0\le j<k$$

$$\nabla^T f(X_k)\nabla f(X_k)=-\nabla^T f(X_k)\vec{p}_k$$

$$\Rightarrow$$

$$\alpha_{kj}=\frac{\nabla^T f(X_k)\left(\nabla f\left(X_{j+1}\right)-\nabla f\left(X_j\right)\right)}{-\vec{p}_j^T\nabla f(X_j)}$$

$$=\frac{\nabla^T f(X_k)\nabla f\left(X_{j+1}\right)-\nabla^T f(X_k)\nabla f\left(X_j\right)}{-\vec{p}_j^T\nabla f(X_j)}$$

$$\alpha_{kj}=\begin{cases}0 & \text{if}\quad j<k-1\\ \dfrac{\nabla^T f(X_k)\nabla f\left(X_k\right)}{\nabla^T f(X_{k-1})\nabla f(X_{k-1})} & \text{if}\quad j=k-1\end{cases}$$

$$\alpha_{kk-1} = \frac{\nabla^T f(X_k)\left(\nabla f\left(X_k\right) - \nabla f\left(X_{k-1}\right)\right)}{-\vec{p}_{k-1}^T \nabla f(X_{k-1})}$$

$$= \frac{\nabla^T f(X_k) \nabla f\left(X_k\right)}{\nabla^T f(X_{k-1}) \nabla f(X_{k-1})}$$

$$-\vec{p}_{k-1}^T \nabla f(X_{k-1})$$

$$= 0 - \vec{p}_{k-1}^T \nabla f(X_{k-1})$$

$$= \vec{p}_{k-1}^T \nabla f(X_k) - \vec{p}_{k-1}^T \nabla f(X_{k-1})$$

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

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

$$= \frac{\nabla^T f(X_k)\left(\nabla f\left(X_k\right) - \nabla f\left(X_{k-1}\right)\right)}{\vec{p}_{k-1}^T\left(\nabla f\left(X_k\right) - \nabla f\left(X_{k-1}\right)\right)}$$

要点：三种共轭梯度法

共轭梯度法（Fletcher-Reeves）

1）任取 $X_0 \in R^n$，令 $k=0$

2）如果 $\|\nabla f(X_k)\| \le \varepsilon$，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 $D_k = -\nabla f(X_k) + \alpha_{k-1} D_{k-1}$，其中

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

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

共轭梯度法（Polak-Ribiere / Polyak）

1）任取 $X_0 \in R^n$，令 $k=0$

2）如果 $\|\nabla f(X_k)\| \le \varepsilon$ ，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 $D_k = -\nabla f(X_k) + \alpha_{k-1} D_{k-1}$ ，其中

$$\alpha_{k-1} = \frac{\nabla^T f(X_k)\left(\nabla f(X_k) - \nabla f(X_{k-1})\right)}{\|\nabla f(X_{k-1})\|^2}$$

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

共轭梯度法（Beale-Sorenson / Hestenes-Stiefel）

1）任取 $X_0 \in R^n$，令 $k=0$

2）如果 $\left\|\nabla f(X_k)\right\| \le \varepsilon$ ，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 $D_k = -\nabla f(X_k) + \alpha_{k-1} D_{k-1}$ ，其中

$$\alpha_{k-1} = \frac{\nabla^T f(X_k)\left(\nabla f(X_k) - \nabla f(X_{k-1})\right)}{D_{k-1}^T\left(\nabla f(X_k) - \nabla f(X_{k-1})\right)}$$

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

关于共轭梯度法的结论

1）共轭梯度法是下降算法

2）对于正定二次目标函数

$$f(X)=\frac{1}{2}X^{T}AX+B^{T}X+C$$

如果从相同的初始点出发，三种共轭梯度法前进的轨迹完全相同，即每一步一维搜索得到的点均相同，并且，经过 n 次精确的一维搜索后一定找到最优解，即 $X_n=-A^{-1}B$

意义：对一般非线性函数在最优解附近快速收敛

要点：几种算法的性能比较

三种基于梯度的搜索方向的比较

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