优化理论与最优控制

数学准备

梯度与Hessian矩阵

函数 $f(x)$ $(x \in R^n)$， $x=(x_1, x_2, \ldots x_n)^T$在某一点$x^k$处的n 个一阶偏导数构成的n 维向量

$$\nabla f(x^k) = (\frac{\partial f(x^k)}{\partial x_1}, \frac{\partial f(x^k)}{\partial x_2}, \cdots \frac{\partial f(x^k)}{\partial x_n})^T$$

$$H(x^k) = \nabla^2 f(x^k) = \begin{bmatrix} \frac{\partial^2 f(x^k)}{\partial x_1^2} & \frac{\partial^2 f(x^k)}{\partial x_1 x_2} & \cdots & \frac{\partial^2 f(x^k)}{\partial x_1 x_n} \\ \frac{\partial^2 f(x^k)}{\partial x_2 x_1} & \frac{\partial^2 f(x^k)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x^k)}{\partial x_2 x_n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2 f(x^k)}{\partial x_n x_1} & \frac{\partial^2 f(x^k)}{\partial x_n x_2} & \cdots & \frac{\partial^2 f(x^k)}{\partial x_n^2} \end{bmatrix}$$

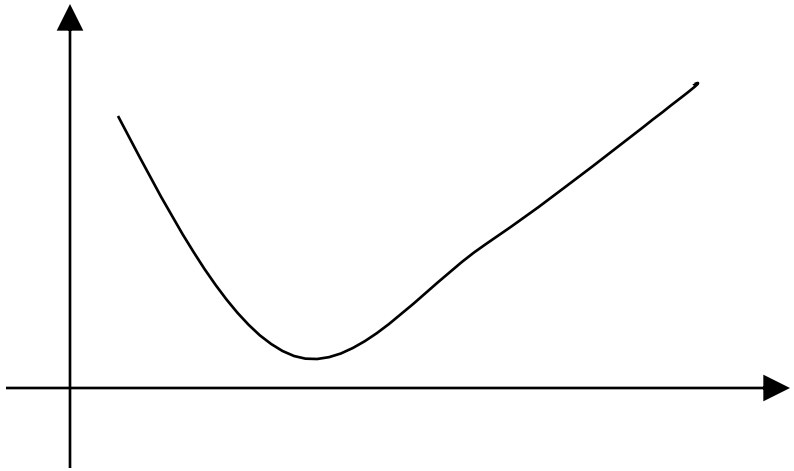
求解 $f(x)=2x_1^2+5x_2^2+x_3^2+2x_2x_3+2x_3x_1-6x_2+3$

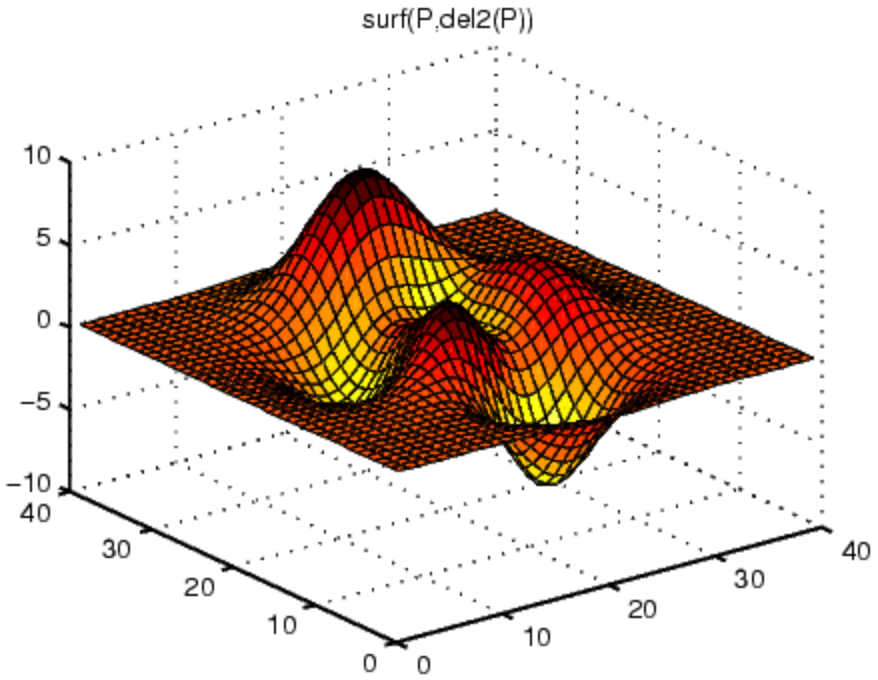
在 $x_1=1$， $x_2=1$， $x_3=-2$时的一阶偏导数和Hessian矩阵

$$\nabla f(x^k)=(\frac{\partial f(x^k)}{\partial x_1},\frac{\partial f(x^k)}{\partial x_2},\cdots\frac{\partial f(x^k)}{\partial x_n})^T$$

$$H(x^k)=\nabla^2 f(x^k)=\begin{bmatrix}\frac{\partial^2 f(x^k)}{\partial x_1^2} & \frac{\partial^2 f(x^k)}{\partial x_1 x_2} & \cdots & \frac{\partial^2 f(x^k)}{\partial x_1 x_n}\\ \frac{\partial^2 f(x^k)}{\partial x_2 x_1} & \frac{\partial^2 f(x^k)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x^k)}{\partial x_2 x_n}\\ \cdots & \cdots & \cdots & \cdots\\ \frac{\partial^2 f(x^k)}{\partial x_n x_1} & \frac{\partial^2 f(x^k)}{\partial x_n x_2} & \cdots & \frac{\partial^2 f(x^k)}{\partial x_n^2}\end{bmatrix}$$

单峰与多峰函数





最优解

最优解的概念

设 $f(x)$ 为定义在 n 维空间 R^n 上的某一领域 N 上的 n 元实函数，

其中 $x=(x_1,\ x_2,\ \ldots x_n)^T$

§ 局部最优解

对于 $x^* \in N$ 如果存在 $\varepsilon>0, \|x-x^*\|<\varepsilon$ 均满足 $f(x)\geq f(x^*)$

称 x^* 为 $f(x)$ 在 N上的局部最优解

§ 全局最优解

对于 $x^* \in N$，而对于所有 $x \in N$ 均满足 $f(x)\geq f(x^*)$

称 x^* 为 $f(x)$ 在 N上的全局最优解

算法分类 （以极小化无约束优化问题为例）

古典微分法

$$\min\ (\max)\quad f(x)$$

$$\text{s. t.}\quad g(x) \geq 0$$

$$h(x) = 0$$

$$\min\quad f(x)$$

最优化问题从数学意义上讲实际上是在无约束或者约束条件下目标函数的极值问题

极值点

设定义在 n 维空间R^n中的函数f（x）具有连续一阶导数， 并且

必要条件

对于极值点x^*，

$$\nabla f(\bar{x}^*) = (\frac{\partial f(\bar{x}^*)}{\partial x_1}, \frac{\partial f(\bar{x}^*)}{\partial x_2}, \cdots \frac{\partial f(\bar{x}^*)}{\partial x_n})^T = \bar{0}$$

设定义在 n 维空间R^n中的函数f（x）具有连续二阶导数

充要条件

对于极值点x^*，

$$\nabla f(\bar{x}^*) = (\frac{\partial f(\bar{x}^*)}{\partial x_1}, \frac{\partial f(\bar{x}^*)}{\partial x_2}, \cdots \frac{\partial f(\bar{x}^*)}{\partial x_n})^T = \bar{0}$$

x^*的Hessian矩阵正定

求解 f（x）=2 $x_1^2 + 5\,x_2^2 + x_3^2 + 2\,x_2\,x_3 + 2\,x_3\,x_1 - 6x_2 + 3$ 的极值点

古典微分法的不足

$f(x)$不具有解析性

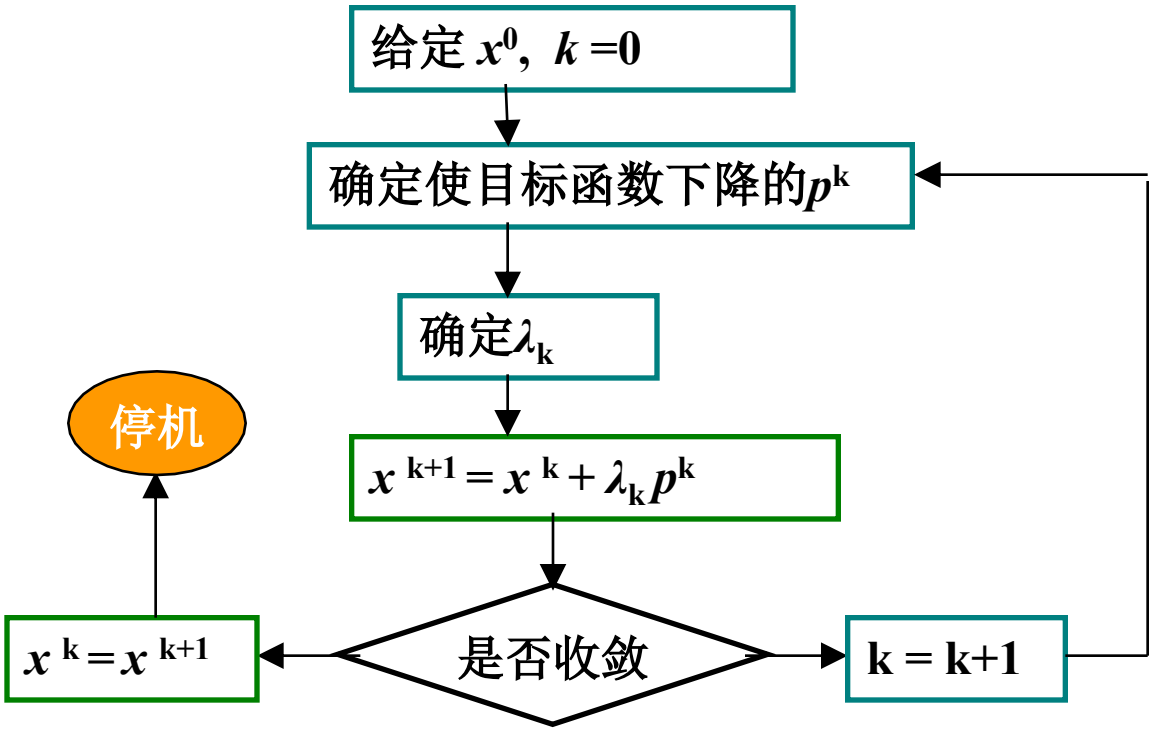
求解　　$\nabla f(\bar{x}^*) = \bar{0}$　　非常困难

通过Hessian矩阵值判定是否为最优值非常困难

迭代法

$$\min \quad f(x) \qquad x \in R^n$$

最优化方法通常采用迭代方法求最优解 $x^{k+1} = x^k + \lambda_k p^k$



迭代法的收敛速度和终止准则

不存在

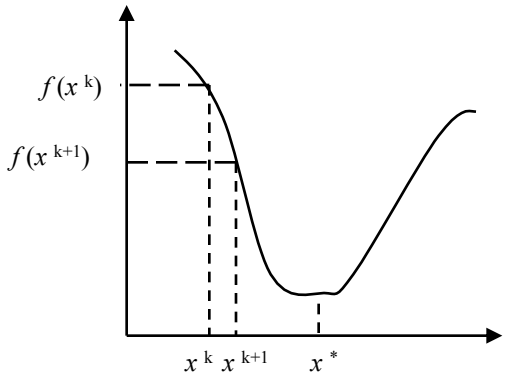
设某种算法产生的点列 $\{x^k\}$ 收敛于优化问题的最优解 x^*，且满足

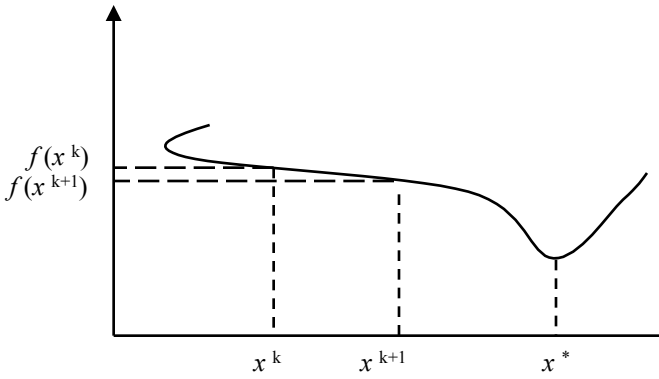
$$\lim_{k\to\infty}\frac{\|x^{k+1}-x^*\|}{\|x^k-x^*\|^P}=c$$

则称该算法为 p 阶收敛

$p=1$	线性收敛
$1<p<2$	超线性收敛
$p=2$	二阶收敛

终止准则 $\|x^{k+1}-x^k\|\le\varepsilon$ $|f(x^{k+1})-f(x^k)|<\varepsilon$





$$\frac{|f(x^{k+1})-f(x^k)|}{|f(x^k)|}<\varepsilon$$

$$\frac{\|x^{k+1}-x^k\|}{\|x^k\|}\le\varepsilon$$

一维极小化方法

$$\min \quad f(x)$$

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

进退法

0.618法

插值法

不精确一维搜索法

最速下降法

牛顿法

单变量问题最优化

$$\min \quad f(x)$$

消去法

黄金分割法

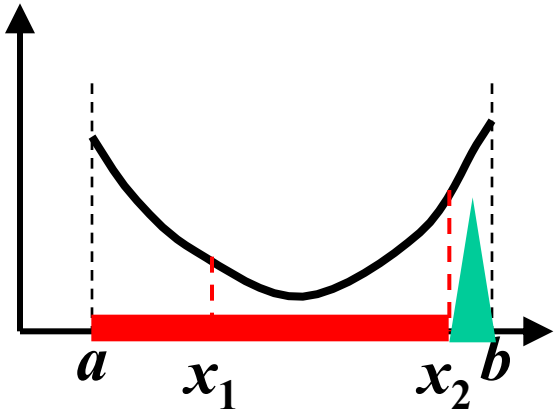
插值法

牛顿法

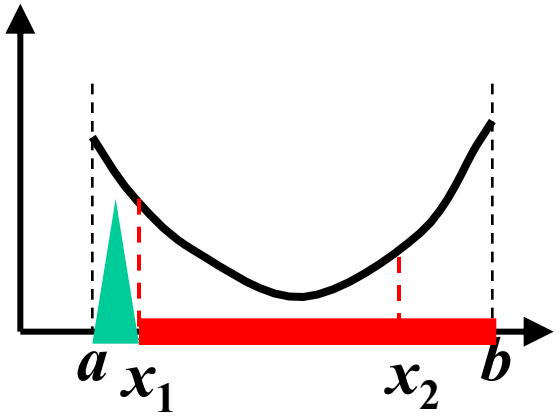
消去法的基本思路

设凸函数 $f(x)$在区间 $[a\ \ b]$内存在极小点x^*。

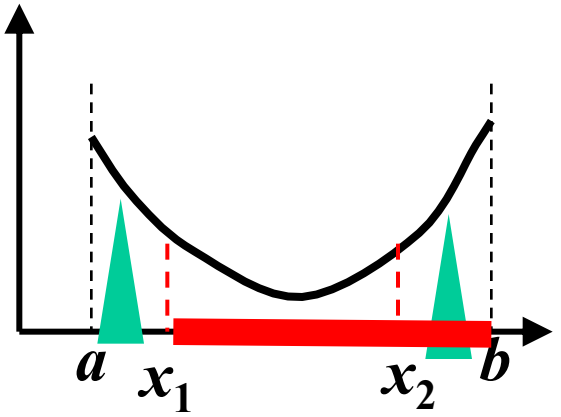
在 $[a\ \ b]$内按一定规则对称地取两个内部点 x_1和x_2，且 $x_1<x_2$，计算 $f(x_1)$和$f(x_2)$，可能有三种情况：



$f(x_1) < f(x_2)$



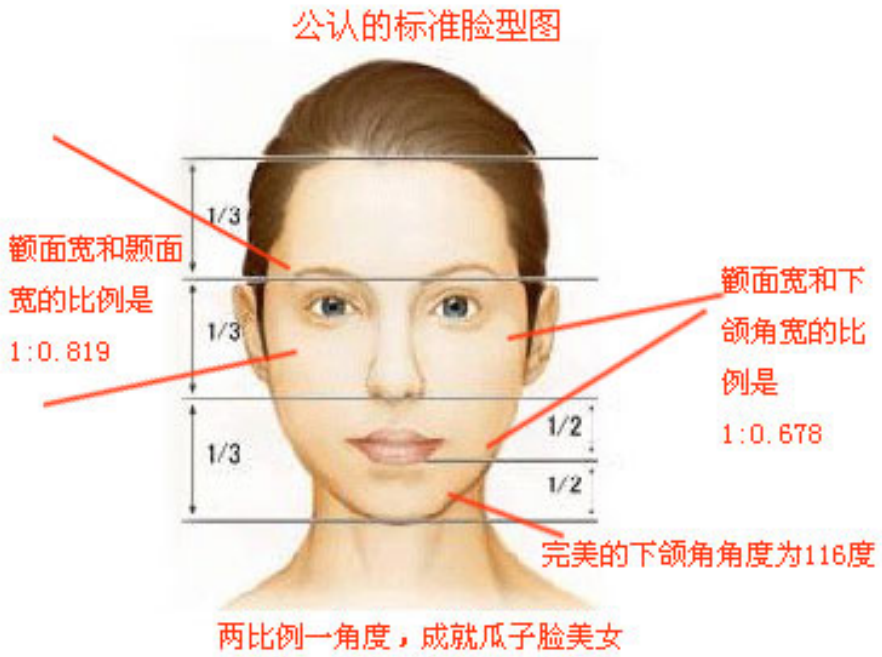
$f(x_1) > f(x_2)$



$f(x_1) = f(x_2)$

区间短缩的过程 ⟶ 区间缩短率？ $\dfrac{\text{红}}{[a\ \ b]}$

黄金分割法

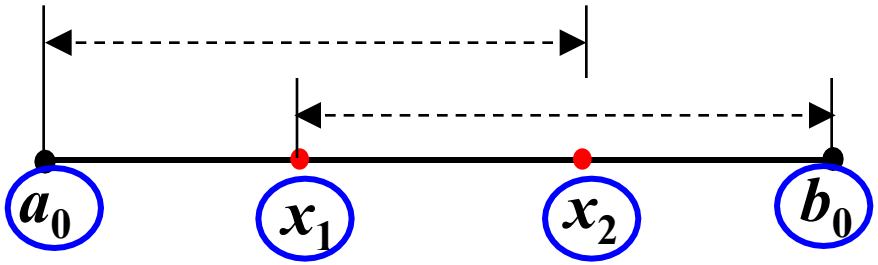


人体结构中有14个“黄金点”（物体短段与长段之比值为 0.618）

12个“黄金矩形”（宽与长比值为 0.618的长方形）

2个“黄金指数”（两物体间的比例关系为 0.618）

黄金分割法



$$\frac{|f(x_{k+1})-f(x_k)|}{|f(x_k)|}<\varepsilon_1$$

$$\frac{\|x_{k+1}-x_k\|}{\|x_k\|}\le\varepsilon_2$$

计算$f(x_1)$和$f(x_2)$并比较

若$f(x_1)<f(x_2)$ 则新区间$[a_1\ b_1]=[a_0\ x_2]$

若$f(x_1)>f(x_2)$ 则新区间$[a_1\ b_1]=[x_1\ b_0]$

精度判别；重复步骤

计算流程图

例 用（Golden section method）法求解下列问题

$$\min \quad f(x) = 2x^2 - x - 1$$

初始区间 $[a_1, b_1]$ =[-1，1]，区间长度要求精度 $\varepsilon \le 0.16$

（终止准则简化）

k	a_k	b_k	x'_k	x_k	$f(x'_k)$	$f(x_k)$
1	-1	1	-0.236	0.236	-0.653	-1.125
2	-0.236	1	0.236	0.528	-1.125	-0.970
3	-0.236	0.528	0.056	0.236	-1.050	-1.125
4	0.056	0.528	0.236	0.348	-1.125	-1.106
5	0.056	0.348	0.168	0.236	-1.112	-1.125
6	0.168	0.348	0.236	0.279	-1.125	-1.123
7	0.168	0.279				

二次多项式近似法

基本思想是在搜索区间中不断用二次多项式 P(*x*) 来近似目标函救 *f*(*x*)，并逐步用插值多项式的极小点来逼近一维搜索问题的极小点

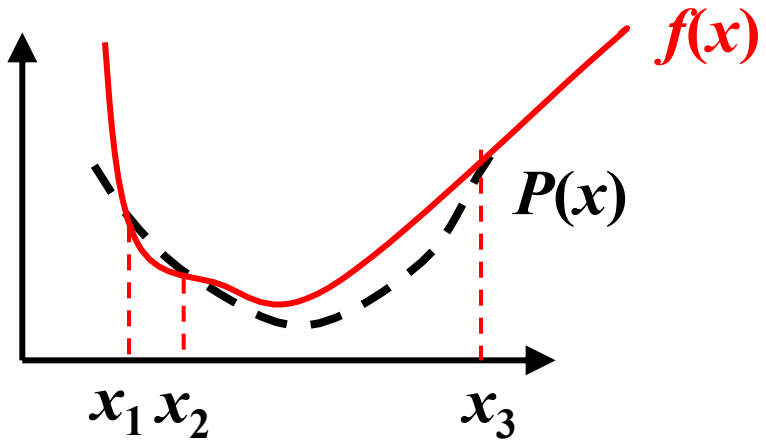
？多项式函数具有连续可微的解析性质，可以采用微分法求极小点的解析格式

$$P(x)=\beta_0+\beta_1 x+\beta_2 x^2$$

？β_0 β_1 β_2 怎么求解

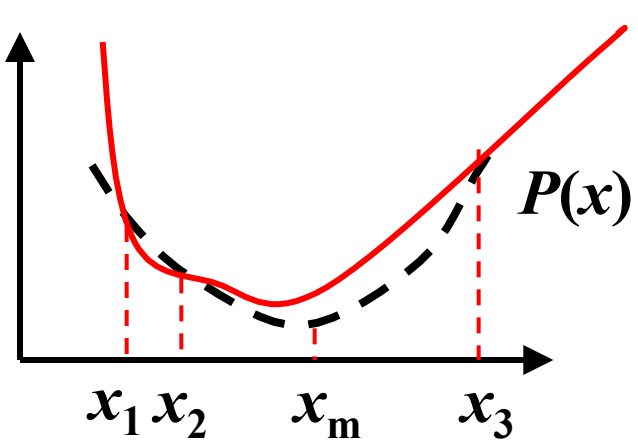
$$P(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

在函数 $f(x)$ 的寻优区间内任意确定x_1，x_2，x_3三点。

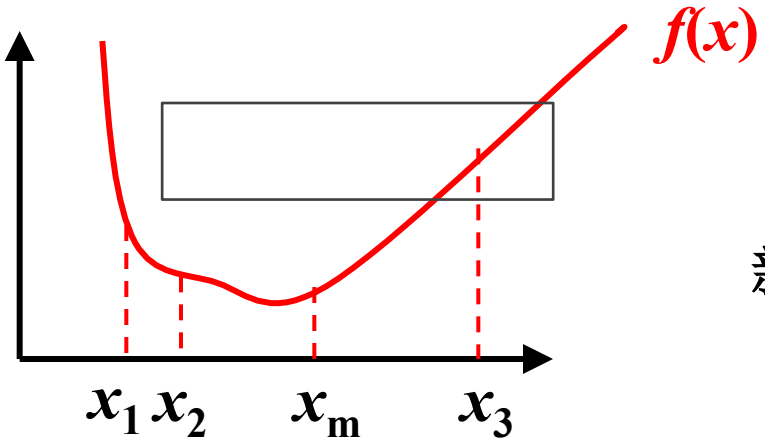


求 β_0 β_1 β_2 → 求 $P(x)$ 的极小值点以及对应的x_m

提示： (x_1, f_1)，(x_2, f_2)，(x_3, f_3)已知



比较$f(x_2)$和$f(x_m)$，取较小的为下一个中间点，其左右两边的点保持，构成新的三点

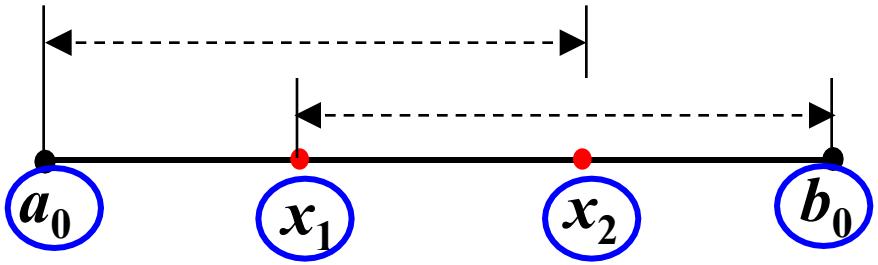


新区间，继续二次多项式，直至收敛

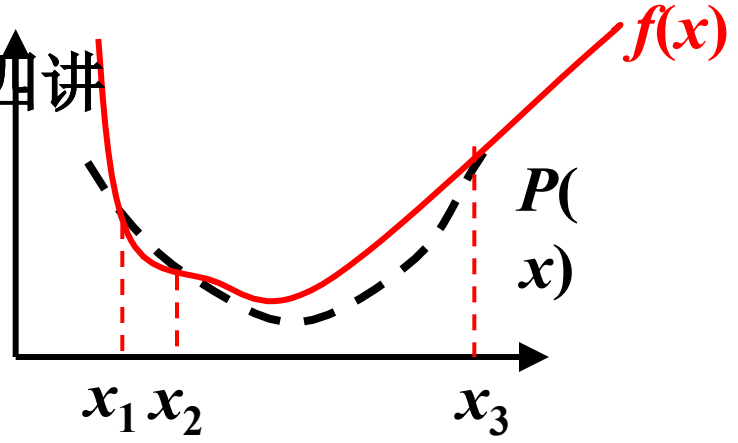
外推内插法

上述两种方法难点？

黄金分割法



二次多项式近似法



极小点存在的区间　　经验？数值搜索？

牛顿法

设凸函数 $f(x)$的在区间 $[a\ \ b]$内存在极小点x^*，且二阶连续可微。那么$f(x)$的二阶泰勒展开为：

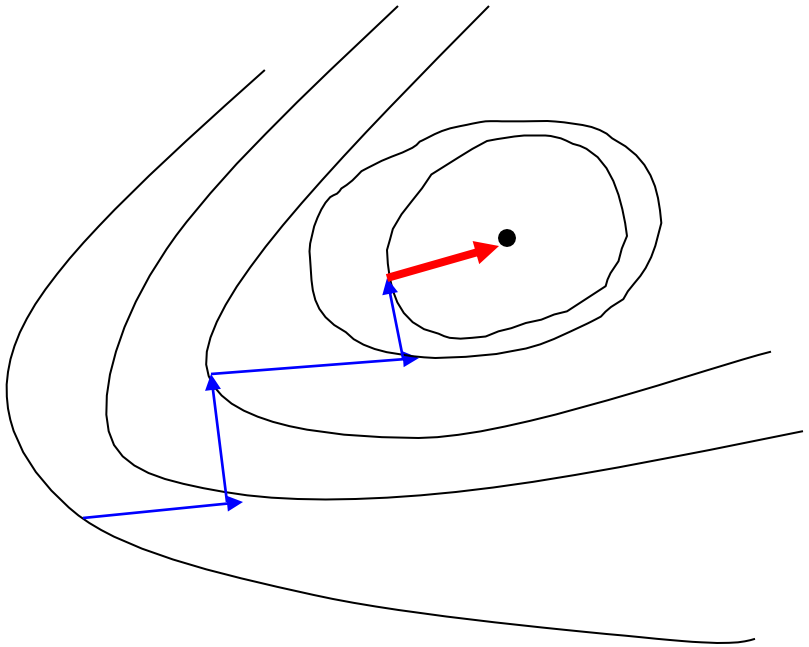
$$L(x) = f(x^*) + f'(x^*)(x - x^*) + 0.5 f''(x^*)(x - x^*)^2$$

基本思想是用 $L(x)$ 的极小点近似 $f(x)$ 的极小点 $L'(x) = 0$

$$x_{k+1} = x_k - f'(x_k)/f''(x_k)$$ 收敛于 $$|f'(x_{k+1})| < \varepsilon$$

以梯度为基础的优化方法

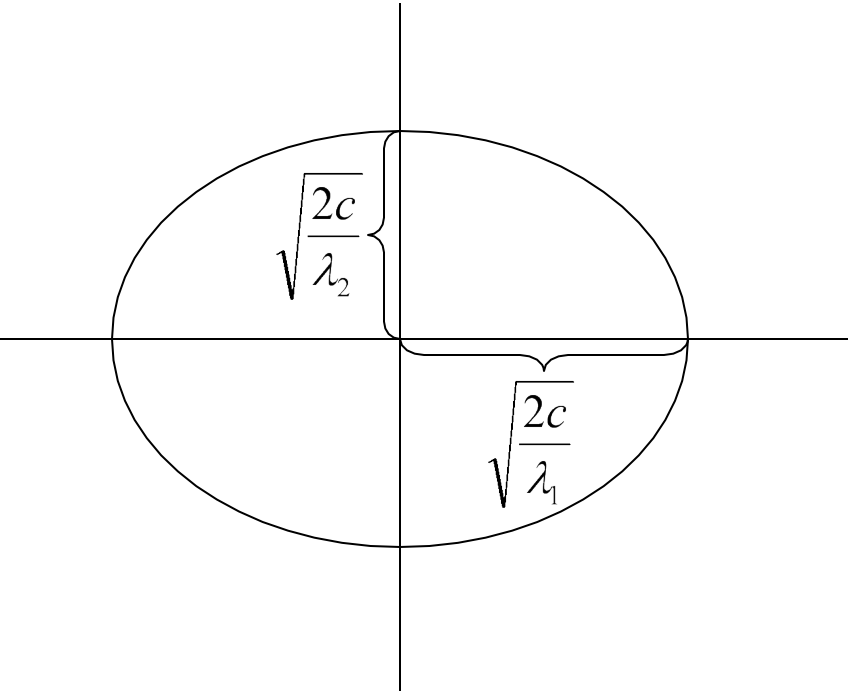
最速下降法



最速下降法的迭代点在向极小点靠近的过程中走的是曲折的路：后一个搜索方向P_{k+1}与前一次搜索方向 P_k 总是互相垂直的。称为锯齿现象。

在远离极小点的地方每次迭代可能使目标函数值较多的下降。可是在接近极小点的地方，由于锯齿现象使每次迭代行进的距离缩短，因而收敛速度不快。这正是最速下降法的缺点。

最速下降法收敛特性的几何意义：



考虑二元二次函数：

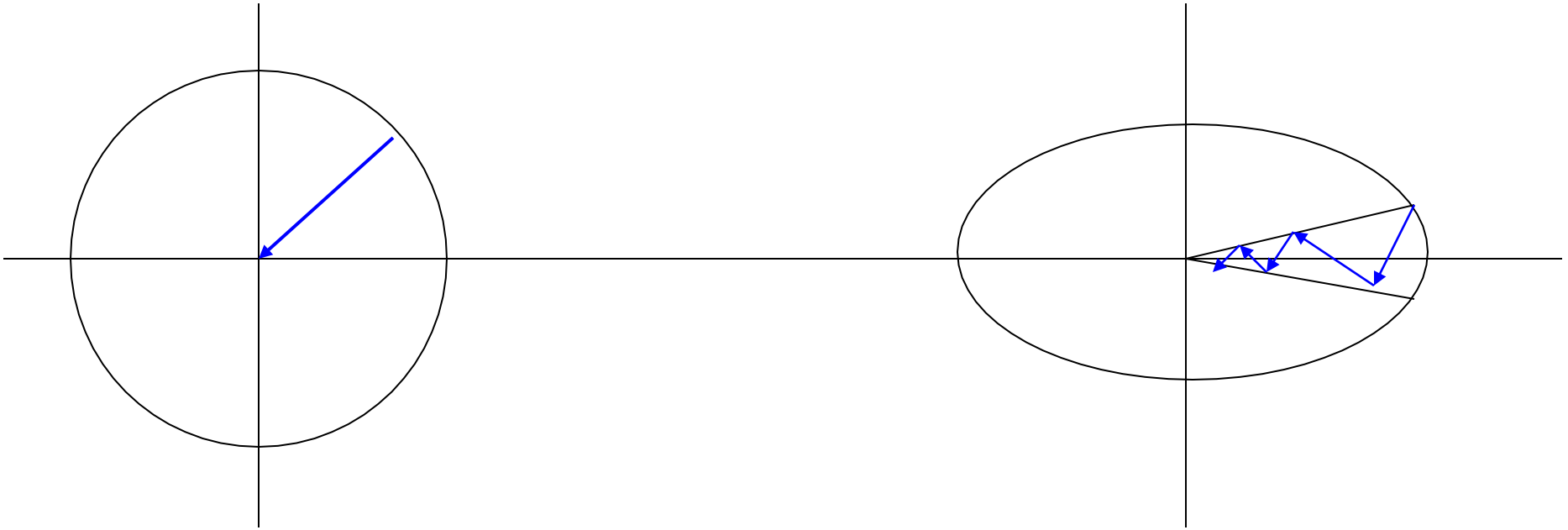
$$f(x)=\frac{1}{2}x^{T}Qx=\frac{1}{2}\left(\lambda_1 x_1^2+\lambda_2 x_2^2\right)$$

这个函数的等值线为：常数$c>0$)。

$$f(x_1,x_2)=c$$

$$\frac{x_1^2}{\left(\sqrt{\frac{2c}{\lambda_1}}\right)^2}+\frac{x_2^2}{\left(\sqrt{\frac{2c}{\lambda_2}}\right)^2}=1$$ ，这是以$\sqrt{\frac{2c}{\lambda_1}}$ &$\sqrt{\frac{2c}{\lambda_2}}$ 为半轴的椭圆

两个特征值的相对大小决定了最速下降法的收敛特性



(a)　　(b)

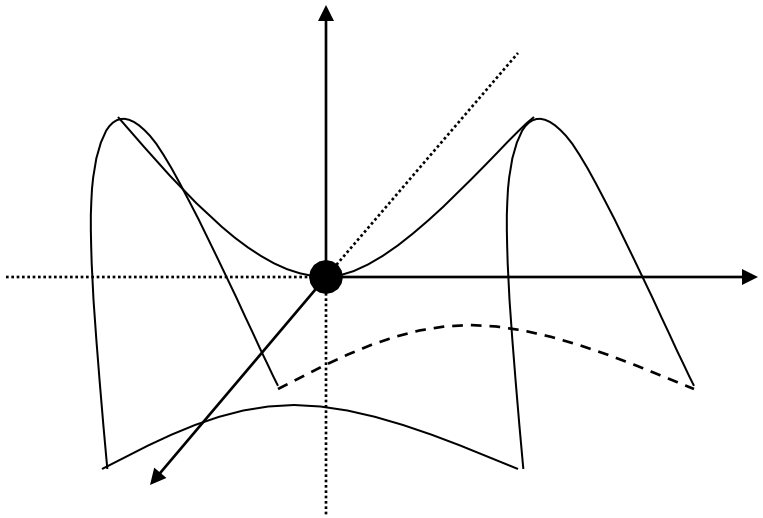
当$\lambda_1 = \lambda_2$时，等值线变为圆，其几何意义是，最速下降法施用于等值线为圆的目标函数时，只须迭代一次就得到了极小点。

当$\lambda_1 < \lambda_2$时，等值线为椭圆。对于一般的初始点，将出现锯齿现象。

Newton法

- 如果目标函数的梯度和Hessian矩阵的表达式容易求到，并且对极小点可能给出较好的估值，那么还是使用Newton法为宜。
- 对于表达式很复杂的目标函数，其Hessian矩阵很难或不可能求出，这正是Newton法的局限性。
- 在通常情况下，极小点相应于正定的Hessian矩阵，极大点相应于负定的Hessian矩阵，而鞍点相应于一个不定矩阵，即*特征值有正有负*的矩阵。

$$z = y^2 - x^2$$



共轭梯度法

(一) 共轭方向

共轭方向的一般定义：对于对称正定矩阵A，称满足

$$[S^{(i)}]^{\mathrm{T}} A S^{(j)}=0 \quad (i\neq j;\ \ i, j=1, 2, \ldots, n)$$

$$[S^{(i)}]^{\mathrm{T}} A S^{(i)}\neq 0 \quad (i=1, 2, \ldots, n)$$

的非零向量$S^{(1)}, S^{(2)}, \ldots, S^{(n)}$为关于$A$两两共轭。当$A$=I（单位矩阵）时，$[S^{(i)}]^{\mathrm{T}} S^{(j)}=0$, 此时为通常的**正交条件**。所以**共轭方向是正交方向的推广**。

1. 二维正定二次函数依次沿两个共轭方向进行一维搜索，则两步就可以达到极小点。这一结论可以推广到n维二次函数，其结论是：若$f(X)$为n维正定二次函数，则依次沿n个共轭方向进行一维搜索，n步就可以收敛到极小点。

2. 由于高次函数可以按泰勒公式展开到二次项，所以当迭代计算已接近极点时，按二次收敛速度收敛于最优点。

共轭梯度法包含：

F-R(Fletcher-Reeves)共轭梯度法,

D-M(Dixon-Myers)共轭梯度法,

P-R-P(Polak-Pibiere-Polyok)共轭梯度法.

对于正定二次函数，这三种形式是等价的，但对于一般的目标函数，它们产生的方向不同，**其中F-R最常用，但也有理论表明P-R-P法略优。**

DFP法

DFP的基本思想：不需要计算二阶导数，又能很好地逼近，那么计算会变得收敛速度很快，而且又不繁琐。

?

$$H_{K+1} = H_K + \frac{\Delta\chi_k(\Delta\chi_k)^t}{(\Delta\chi_k)^t \Delta g_k} - \frac{H_k \Delta g_k (\Delta g_k^t) H_K}{\Delta g_k^t H_k \Delta g_k}$$

DFP方法具有数值　不稳定性，有时产生数值上奇异Hessian阵，BFGS方法就是它的修正。

Levenberg – Marqurdt 算法

$$\min_{x \in r^n} f(x) = \frac{1}{2} r(x)^t r(x) = \frac{1}{2} \sum_{i=1}^{m} [r_i(x)]^2 , m \geq n$$

$$p_k(u) = -(A_k{}^T A_k + uI)^{-1} A_k{}^T f_k$$