上海大学
SHANGHAI UNIVERSITY

非线性规划

主要内容

- 基本概念
- 凸函数
- 线性搜索算法
- 无约束极值问题
- 约束极值问题

基本概念

非线性规划 { 无约束优化　$\min\limits_{x\in E^n} f(x)$

约束优化

约束优化问题数学模型：

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

$$s.t. \quad g_i(x)\geq 0, \quad i=1,2,\cdots,l$$

$$h_j(x)=0, \quad j=1,2,\cdots,m.$$

基本概念

例:某公司经营两种设备。第一种设备每件售价为30元，第二种设备每件售价为450元。且知，售出第一、二种设备分别需时为每件约0.5小时和（2+0.25x_2）小时，其中x_2为第二种设备售出数量。公司的总营业时间为800小时。

求：公司为获取最大营业额（销售额）的最优营业计划

[解]设公司应经营销售第一、二种设备数额分别为x_1件和x_2件，则有

$$\max: f(X) = 30x_1 + 450x_2$$

$$s.t.\ 0.5x_1 + 2x_2 + 0.25x_2^2 \leq 800$$

$$x_1 \geq 0,\ x_2 \geq 0$$

范数

若函数 $\|\cdot\|$：$\Re^n \to \Re$满足下面条件：

（1）正定性：$\forall x \in \Re^n, \| x \| \geq 0,$并且$\| x \| = 0 \Leftrightarrow x = 0;$

（2）三角不等式：$\forall x, y \in \Re^n, \| x + y \| \leq \| x \| + \| y \|;$

(3) 正齐次性：$\forall \alpha \in \Re, \forall x \in \Re^n, \| \alpha x \| = | \alpha | \| x \|.$

则称 $\|\cdot\|$ 为$\Re^n$上的范数.

常用范数：

$$\| x \|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \qquad \| x \|_1 = \sum_{i=1}^{n} | x_i | \qquad \| x \|_\infty = \max_{1 \leq i \leq n} | x_i |$$

集合

x^0的$\varepsilon-$邻域：$N_\varepsilon=\{x\,|\,\|x-x^0\|<\varepsilon,\varepsilon>0\}$

内点：设$x^0\in S\subset\Re^n$，若存在$\varepsilon>0$，使得$N_\varepsilon(x^0)\subset S$，则称$x^0$为$S$的一个内点。

补集：集合S的补集定义为$S^C=\{x\,|\,x\notin S,x\in\Re^n\}$

开集：若对$\forall x\in S$，x为内点，则称S为开集。

闭集：若集合S的补集S^C为开集，则称S为闭集。

有界集：若存在正数$M>0$，使得$\forall x\in S,\|x\|\le M$成立，则称$S$为有界集。

紧集：有界闭集称为紧集

可行性

定义

满足约束条件的点$x \in \mathbb{R}^n$称为可行点。所有可行点的集合称为可行域。

$$X = \left\{ x \in \mathbb{R}^n \;\middle|\; \begin{array}{ll} c_i(x) = 0, & i = 1, \cdots, m' \\ c_i(x) \geq 0, & i = m'+1, \cdots, m \end{array} \right\}.$$

- 考虑可行点$\bar{x}$和不等式约束$c_i(x) \geq 0$:
  - 如果有$c_i(\bar{x}) = 0$, 就称不等式约束$c_i(x) \geq 0$在点$\bar{x}$是有效约束或起作用约束(active constraint), 并称可行点$\bar{x}$位于约束 $c_i(x) \geq 0$的边界
  - 如果有$c_i(\bar{x}) > 0$,就称不等式约束$c_i(x) \geq 0$在点$\bar{x}$是无效约束或不起作用约束(inactive constraint), 并称$\bar{x}$是约束$c_i(x) \geq 0$的内点。
- 在任意可行点，所有的等式约束都被认为是有效约束。

可行性

在可行点$\bar{x}$，所有有效约束的全体被称为该可行点的有效集(active set)，其指标集记为

$$\mathcal{A}_{\bar{x}} = \{i \mid i = 1, 2, \cdots, m', m'+1, \cdots, p, \quad c_i(\bar{x}) = 0\}.$$

在可行点$\bar{x}$如果没有一个不等式约束是有效的，就称$\bar{x}$是可行域的内点，不是内点的可行点就是可行域的边界点。

可行性

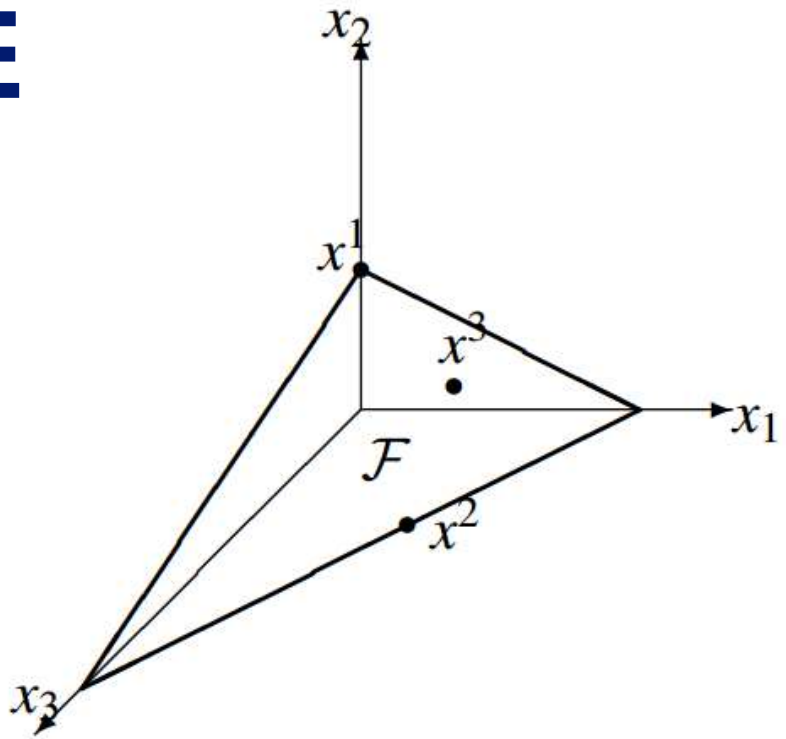
考虑可行域X:

$$c_1(x) = 2x_1 + 3x_2 + x_3 - 6 = 0,$$

$$x_1 \geq 0,$$

$$x_2 \geq 0,$$

$$x_3 \geq 0.$$



- 对于可行点x^1，约束$x_1 \geq 0$和$x_3 \geq 0$是有效约束，而$x_2 \geq 0$ 是无效约束。
- 对于可行点x^2,则刚好相反，约束$x_2 \geq 0$ 是有效约束，而$x_1 \geq 0$和$x_3 \geq 0$是无效约束。
- 对于可行点x^3，三个不等式约束都是无效约束。
- 图中可行域的边界由粗线表示

可行性

[例]求解下述非线性规划

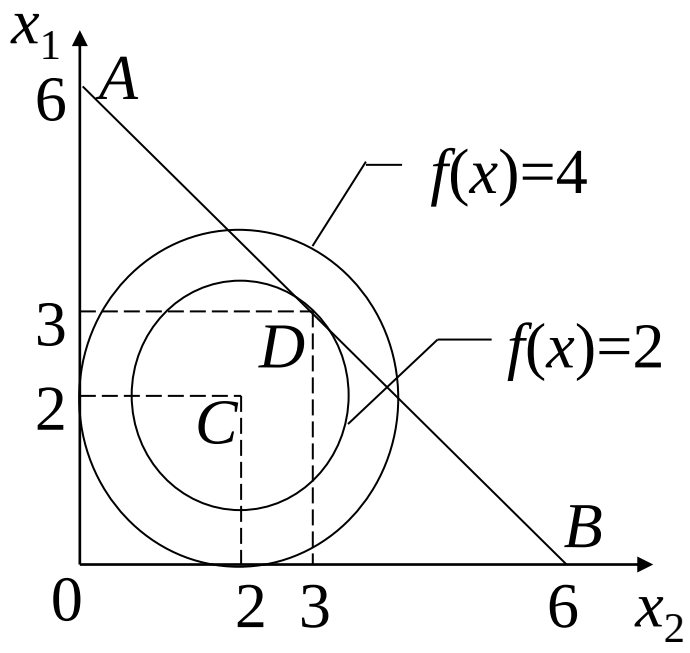
$$\min\ f(x)=(x_1-2)^2+(x_2-2)^2$$

$$h(x)=x_1+x_2-6=0$$

显然，与直线AB相切的点必为最优解。

图中的D点即为最优点，此时目标函数值为：

$f(x^*)=2$，$x_1^*=x^*_2=3$

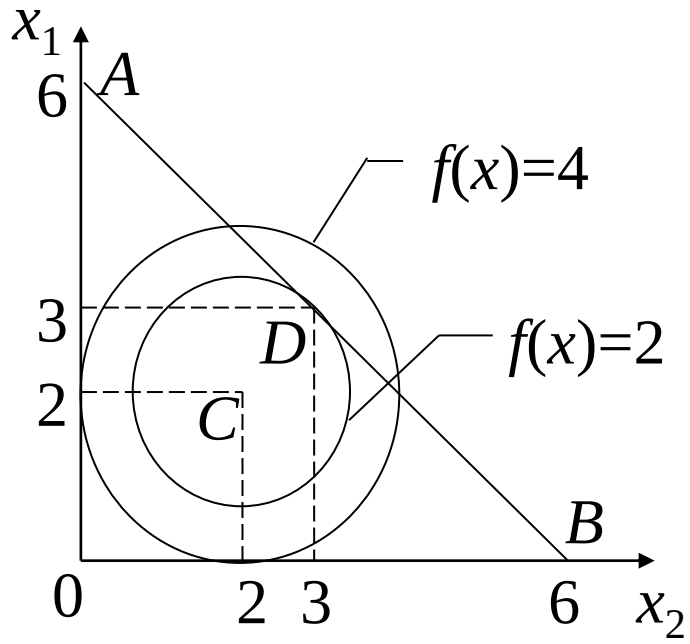


可行性

[例]非线性规划为

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

$h(x)=x_1+x_2-6\leq 0$



最优解为$x_1^*=x_2^*=2$，$f(x^*)=0$，该点落在可行域内部，其边界约束失去作用。

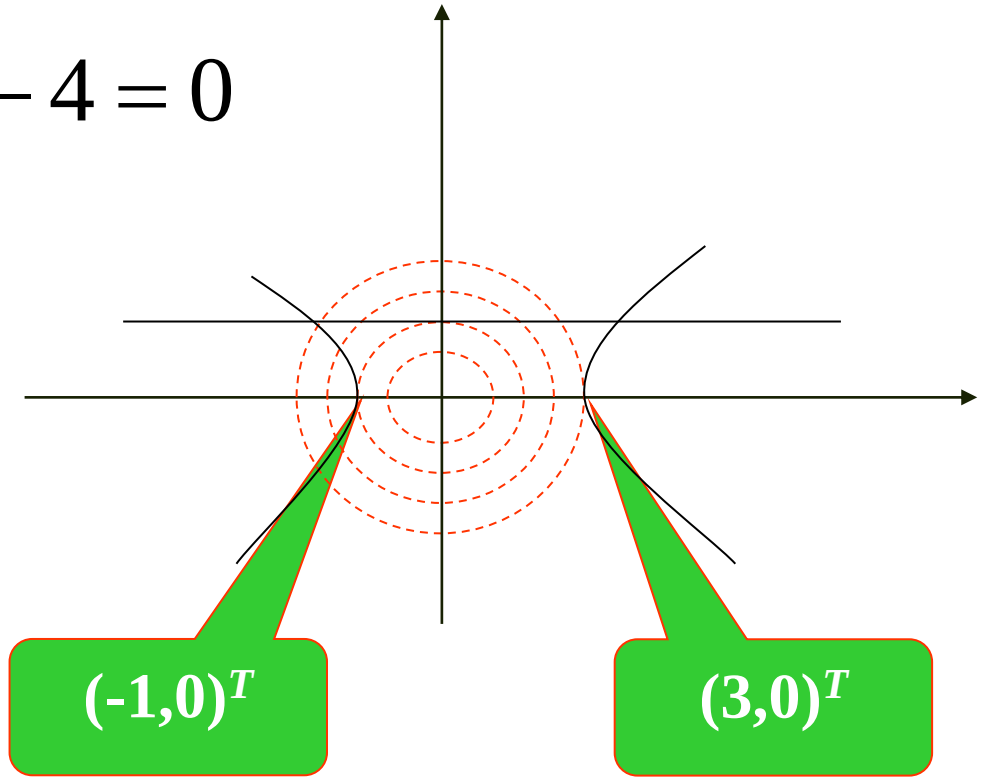
可行性

求下列约束问题的解：

$$\min\ x_1^2 + x_2^2$$

$$s.t.\quad (x_1 - 1)^2 - x_2^2 - 4 = 0$$

$$x_2 - 1 \le 0$$



最优解

全局最优解

一个可行点$x^* \in X$ 称为问题(1.1)-(1.3)的全局(或总体)最优解(极小点)，如果有

$$f(x^*) \leq f(x), \quad \forall x \in X \tag{1.11}$$

成立。 如果上述不等式对所有不同于x^*的可行点x严格成立， 即

$$f(x^*) < f(x), \quad \forall x \in X, \ x \neq x^*, \tag{1.12}$$

则x^*称为严格全局(或总体)最优解。

最优解

局部最优解

对于可行点x^*, 如果存在一个邻域

$$\mathcal{N}(x^*) = \{x \mid \|x - x^*\| \leq \delta\},$$

使得

$$f(x^*) \leq f(x), \ \forall x \in \mathcal{N}(x^*) \cap X \tag{1.13}$$

成立，则称x^*为优化问题(1.1)-(1.3)的局部最优解， 其中$\delta > 0$是一个小的正数。如果不等式(1.13)对所有$x \in \mathcal{N}(x^*) \cap X$, $x \neq x^*$严格成立，则称x^*为严格局部最优解(极小点).

上述范数$\|\cdot\|$可以是任意向量范数，但一般常用ℓ_2范数

$$\|x\|_2 = \left[\sum_{i=1}^{n} x_i^2\right]^{\frac{1}{2}}.$$

凸集

"To be, or not to be, that is the question"

---- 《Hamlet》 William Shakespeare 1599~1602

"凸或者不凸，那才是问题的本质"

---- 数学规划分会 徐宗本院士 2010

凸集

定义： 设x,y为欧式空间R^n中相异的两个点，则点集

$$P=\{\lambda x+(1-\lambda)y|\lambda\in R\}$$

称为通过x和y的直线.

定义： 设$S\subseteq R^n$, **若对**$\forall x^{(1)}, x^{(2)}\in S$**及**$\forall\lambda\in[0,1]$, **都有**

$$\lambda x^{(1)}+(1-\lambda)x^{(2)}\in S$$

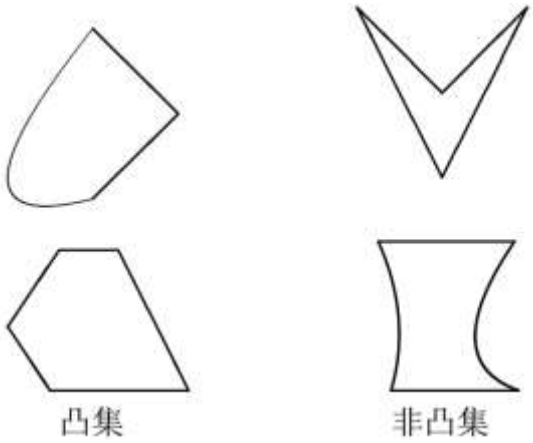
则称S**为凸集.**

设$x^{(1)}, x^{(2)}, \cdots, x^{(k)}\in S$, **称**

$$\lambda_1 x^{(1)}+\lambda_2 x^{(2)}+\ldots+\lambda_k x^{(k)}$$

(其中$\lambda_1+\lambda_2+\cdots+\lambda_k=1$)

为$x^{(1)}, x^{(2)}, \cdots, x^{(k)}$**的凸组合.**



特殊的凸集合

- $H=\{x|p^{T}x=a\}$ 超平面(hyper-plane)
- $H=\{x|p^{T}x\leqslant a\}$ （闭）半空间
- $L=\{x|x=x^{(0)}+\lambda d,\lambda\geq 0\}$ 射线

凸集的性质

设S_1和S_2为R^n中的两个凸集，β是实数，则

(1) $\beta S_1 = \{\beta x \mid x \in S_1\}$

(2) $S_1 \cap S_2$

(3) $S_1 + S_2 = \{x^{(1)} + x^{(2)} \mid x^{(1)} \in S_1,\ x^{(2)} \in S_2\}$

(4) $S_1 - S_2 = \{x^{(1)} - x^{(2)} \mid x^{(1)} \in S_1,\ x^{(2)} \in S_2\}$

都是凸集。

凸锥和多面体

定义：给定集合$C \subset R^n$，若对C中每一个x，及任意的$\lambda \geq 0$，都有$\lambda x \in C$，则称C为锥；若C为凸集，则称C为凸锥。

定义：有限个半空间的交$\{x \mid Ax \leq b\}$称为多面集。

- 非空有界的多面集称为多面体（**polyhedron**）
- 称多面集 $\{x \mid Ax \leq \mathbf{0}\}$ 为多面锥

凸集分离定理

定义1.4.6

设S_1, S_2为R^n中的两个非空集合，$H = \{x | p^T x = \alpha\}$为超平面，如果对每个$x \in S_1$，都有$p^T x \geq \alpha$，对每个$x \in S_2$，都有$p^T x \leq \alpha$（或情形恰好相反），则称超平面$H$分离集合$S_1$和$S_2$.

定理1.4.2（闭凸集上的投影）：

设S为R^n的非空闭凸集，$y \notin S$，则存在唯一的$\bar{x} \in S$，使得

$$\| y - \bar{x} \| = \inf_{x \in S} \| y - x \| > 0.$$

$\bar{x}$是这一最小距离点 $\Leftrightarrow (y - \bar{x})^T (\bar{x} - x) \geq 0, \quad \forall x \in S.$

存在性

凸集分离定理

证明：令 $\inf_{x\in S} \| y-x \| = r > 0$

$\Rightarrow \exists$序列$\{x^{(k)}\}, x^{(k)} \in S$, 使得 $\| y-x^{(k)} \| \to r$。

先证$\{x^{(k)}\}$为Cauchy序列。

$$\| x^{(k)} - x^{(m)} \|^2$$

$$= 2\| x^{(k)} - y \|^2 + 2\| x^{(m)} - y \|^2 - 4\left\| \frac{x^{(k)} + x^{(m)}}{2} - y \right\|^2$$

$$\le 2\| x^{(k)} - y \|^2 + 2\| x^{(m)} - y \|^2 - 4r^2$$

$$\to 0 \quad (\text{当} m \to \infty, k \to \infty)$$

$\therefore \{x^{(k)}\}$为Cauchy序列，$\Rightarrow \{x^{(k)}\}$极限存在，设为$\bar{x}$,

$\because S$为闭集，$\therefore \ \bar{x} \in S.$

凸集分离定理

唯一性

证明：假设存在$\hat{x} \in S$，使得$\| y-\bar{x} \|=\|y-\hat{x}\|=r$

$\because S$为凸集，$\bar{x}, \hat{x} \in S$，　$\therefore$　$\dfrac{\bar{x}+\hat{x}}{2} \in S.$

$$\therefore \quad r \le \left\|y-\frac{\bar{x}+\hat{x}}{2}\right\| \le \frac{1}{2}\| y-\bar{x} \| + \frac{1}{2}\|y-\hat{x}\| = r$$

$$\Rightarrow \quad \left\|y-\frac{\bar{x}+\hat{x}}{2}\right\| = \frac{1}{2}\| y-\bar{x} \| + \frac{1}{2}\|y-\hat{x}\|$$

$$\Rightarrow \quad y-\bar{x}=\lambda(y-\hat{x})$$

$$\Rightarrow \quad \| y-\bar{x} \| = |\lambda| \|y-\hat{x}\|$$

$$\Rightarrow \quad \lambda = 1 \quad \Rightarrow \quad \bar{x}=\hat{x}.$$

凸集分离定理

等价性： $\bar{x}$是这一最小距离点 $\Leftrightarrow (y-\bar{x})^T(\bar{x}-x)\geq 0,\ \forall x\in S.$

证明： “$\Leftarrow$”假设$(y-\bar{x})^T(\bar{x}-x)\geq 0$，

则对任意的$x\in S$，有

$$\|y-x\|^2=\|y-\bar{x}+\bar{x}-x\|^2$$

$$=\|y-\bar{x}\|^2+\|\bar{x}-x\|^2+2(y-\bar{x})^T(\bar{x}-x)$$

$$\geq\|y-\bar{x}\|^2$$

$\therefore$ $\bar{x}$是最小距离点。

凸集分离定理

等价性： $\bar{x}$是这一最小距离点 $\Leftrightarrow (y-\bar{x})^T(\bar{x}-x)\geq 0,\ \forall x\in S.$

“$\Rightarrow$”假设$\bar{x}$是最小距离点,则对$\forall x\in S$，有

$$\|y-x\|^2\geq\|y-\bar{x}\|^2.$$

$\because S$是凸集，$\therefore \forall\lambda\in(0,1)$,有$\bar{x}+\lambda(x-\bar{x})\in S.$

$$\therefore\quad \|y-(\bar{x}+\lambda(x-\bar{x}))\|^2\geq\|y-\bar{x}\|^2,$$

$$\because \|y-(\bar{x}+\lambda(x-\bar{x}))\|^2=\|y-\bar{x}\|^2+\lambda^2\|x-\bar{x}\|^2-2\lambda(y-\bar{x})^T(x-\bar{x})$$

$$\therefore\quad \lambda^2\|x-\bar{x}\|^2-2\lambda(y-\bar{x})^T(x-\bar{x})\geq 0$$

$$\Rightarrow\quad \lambda\|x-\bar{x}\|^2-2(y-\bar{x})^T(x-\bar{x})\geq 0$$

令$\lambda\to 0$,得 $\quad -2(y-\bar{x})^T(x-\bar{x})\geq 0$

$$\therefore\quad (y-\bar{x})^T(\bar{x}-x)\geq 0.$$

定理1.4.3（点与凸集可强分离）：

设S是R^n的非空闭凸集，$y \notin S$，则存在非零向量p及数$\varepsilon > 0$，使得对$\forall x \in S$，有$p^T y \geq \varepsilon + p^T x.$

证明：∵ S为闭凸集，$y \notin S$，由定理1.4.2，$\exists \bar{x} \in S$，使

$$\|y - \bar{x}\| = \inf_{x \in S} \|y - x\| > 0$$

令$p = y - \bar{x} \neq 0, \quad \varepsilon = p^T(y - \bar{x}) = \|y - \bar{x}\|^2 > 0$

$$\because \quad p^T(y - x) = p^T(y - \bar{x} + \bar{x} - x)$$

$$= p^T(y - \bar{x}) + p^T(\bar{x} - x)$$

$$= \varepsilon + (y - \bar{x})^T(\bar{x} - x) \geq \varepsilon$$

$$\therefore \quad p^T y \geq \varepsilon + p^T x.$$

定理1.4.4: 设S是R^n的非空凸集，$y \in \partial S$(S的边界)，则存在非零向量p，使得对$\forall x \in clS$(S的闭包，由S的内点和边界点组成)，有$p^T y \geq p^T x$.

证明：$\because S$是凸集，$\therefore clS$是闭凸集。

$\because y \in \partial S$，则存在序列$\{y^{(k)}\} \notin clS$，使得$y^{(k)} \to y$.

对每个点$y^{(k)}$，由定理1. 4.3，存在单位向量$p^{(k)}$，

使得对每个$x \in clS$，有$\left(p^{(k)}\right)^T y^{(k)} > \left(p^{(k)}\right)^T x$.

$\because$ 序列$\{p^{(k)}\}$有界(单位向量)，$\therefore$ 存在收敛的子序列$\{p^{(k_j)}\}$，

其极限为单位向量p.

$\because \left(p^{(k_j)}\right)^T y^{(k)} > \left(p^{(k_j)}\right)^T x$对每个$x \in clS$成立，

$\therefore$ 令$k_j \to \infty$，得到$p^T y \geq p^T x, \quad \forall x \in clS$.

推 论: 设S是R^n的非空凸集，$y \notin S$，则存在非零向量p，使得对$\forall x \in clS$(S的闭包，由S的内点和边界点组成)，有$p^T(x-y) \leq 0$.

(1) $y \notin clS$ ----定理1.4.3

(2) $y \in \partial S \setminus S$ ----定理1.4.4

定理1.4.5（两个凸集可分离）：

设S_1和S_2是R^n的两个非空凸集，$S_1 \cap S_2 = \varnothing$，则存在非零向量$p$，使得

$$\inf\{\, p^T x \mid x \in S_1\} \geq \sup\{\, p^T x \mid x \in S_2\}.$$

(或　$p^T y \geq p^T x$　对$\forall y \in S_1, \forall x \in S_2$成立)

证明：　设$S = S_2 - S_1 = \{x^{(2)} - x^{(1)} \mid x^{(1)} \in S_1, x^{(2)} \in S_2\}$

$\because$　S_1, S_2是非空凸集，

$\therefore$　S是凸集且$S \neq \varnothing$.

$\because$　$S_1 \cap S_2 = \varnothing$,　$\therefore$　$0 \notin S$

$\Rightarrow$　存在$p \neq 0$,对$\forall x \in S$,有$p^T(x-0) \leq 0$

$\Rightarrow$　$p^T x^{(2)} \leq p^T x^{(1)}$

凸函数

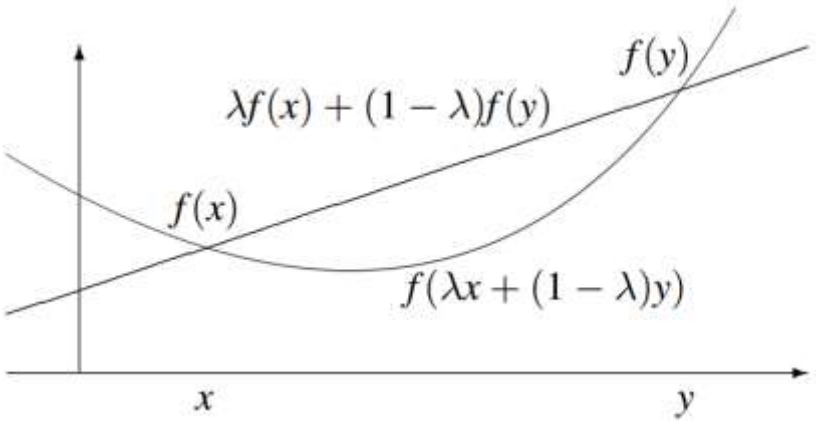
定义 1.2

设函数$f(x)$在凸集D上有定义，如果对任意$x, y \in D$和任意 $\lambda \in [0,1]$有

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y), \tag{1.15}$$

则称$f(x)$是凸集D上的凸函数。如果上述不等式对$x \neq y$与任意$\lambda \in (0,1)$严格成立，则称f是凸集D上的严格凸函数。

凸函数的定义表明：如果$f(x)$是凸集D上的凸函数，则对于凸集D上的任意两点 x, y,连结点$(x,f(x))$与点$(y,f(y))$之间的直线段位于函数图形(曲线或曲面)的上方。



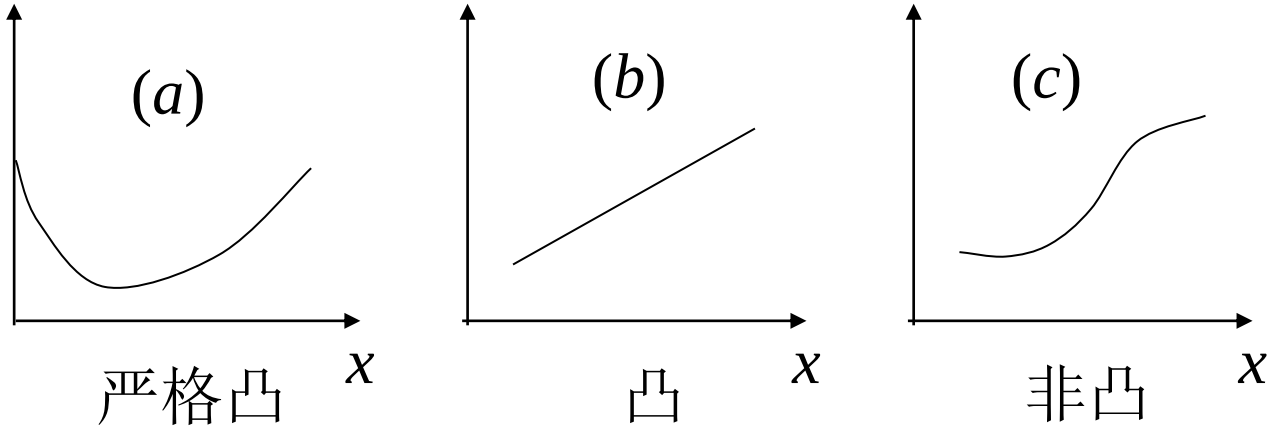
凹函数

一个函数$f(x)$称为是凸集D上的(严格)凹函数，如果$-f(x)$是凸集D上的(严格)凸函数。具体地说，如果对任意$x, y \in D$和任意$\lambda \in [0,1]$有

$$f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda)f(y), \tag{1.16}$$

则称$f(x)$是凸集D上的凹函数。如果上述不等式对$x \neq y$与任意$\lambda \in (0,1)$严格成立，则称f是凸集D上的严格凹函数。

- 例：当$x \geq 0$，$f(x) = x^{\frac{1}{2}}$是凹函数。
- 例：线性函数$f(x) = ax + b$既是凸函数，也是凹函数。



凸函数基本性质

(1) 设$f_1(x)$，$f_2(x)$是凸集S上的凸函数，则函数$f_1(x) + f_2(x)$在S上也是凸函数。

(2) 设$f(x)$是凸集S上的凸函数，则对任意的$a \geq 0$，函数$af(x)$是凸的。

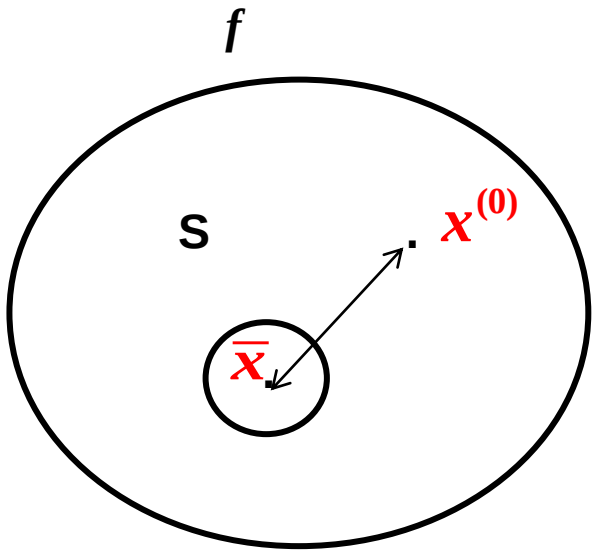
推广：设$f_1(x)$，$f_2(x)$, …, $f_k(x)$是凸集S上的凸函数，$a_i \geq 0$, 则$a_1 f_1(x)+a_2 f_2(x)+ \ldots + a_k f_k(x)$也是凸集$S$上的凸函数.

(3) 设$f(x)$是凸集S上的凸函数，对每一个实数c，则集合

$$S_c = \{x \mid x \in S，\ f(x) \leq c\}$$是凸集。

凸函数的根本重要性

(4)设S是R^n中的非空凸集，f是定义在S上的凸函数，则f在S上的局部极小点是整体极小点，且极小点的集合是凸集.



证明：　设$\bar{x}$是f在S中的局部极小点，即存在$\bar{x}$的$\varepsilon > 0$邻域$N_\varepsilon(\bar{x})$使得对$\forall x \in S \cap N_\varepsilon(\bar{x})$，有$f(x) \geq f(\bar{x})$。

若$\bar{x}$不是整体极小点，则$\exists x^{(0)} \in S$使$f(\bar{x}) > f(x^{(0)})$，

∵　S是凸集，　∴　对$\forall \lambda \in (0,\ 1)$有$\lambda x^{(0)} + (1-\lambda)\bar{x} \in S$，

∵　f是S上的凸函数，

∴　$f(\lambda x^{(0)} + (1-\lambda)\bar{x}) \leq \lambda f(x^{(0)}) + (1-\lambda) f(\bar{x})$

$< \lambda f(\bar{x}) + (1-\lambda) f(\bar{x}) = f(\bar{x})$

当λ取得充分小时，可使$\lambda x^{(0)} + (1-\lambda)\bar{x} \in S \cap N_\varepsilon(\bar{x})$，

这与$\bar{x}$为局部极小点矛盾，

∴　$\bar{x}$是f在S上的整体极小点。

f在S上的极小值也是它在S上的最小值。

极小点集合为$\Gamma_\alpha = \{x | x \in S, f(x) \leq \alpha\}$，

则由性质（3），Γ_α为凸集。

凸函数的判别

二次型的正定性

定义： 给定二次型$f(X)=X^TAX$，若对$\forall X\neq 0$，都有$f(X)=X^TAX>0$成立，则称$f(X)$为正定二次型，A为正定矩阵。

定理： 对于n阶实对称矩阵A，下列命题等价：

（1）X^TAX是正定二次型（或A是正定矩阵）；

（2）A的n个顺序主子式都大于零；

（3）A的n个特征值都大于零；

（4）存在可逆矩阵P，使得$A=P^TP$.

凸函数的判别

二次型的半正定性

定义：对实二次型$f(X)=X^TAX$，若$\forall X\neq 0$，都有$f(X)=X^TAX\geq 0$成立，则称$f(X)$为半正定二次型，A为半正定矩阵。

定理： 对于n阶实对称矩阵A，下列命题等价：

（1）X^TAX是半正定二次型（或A是半正定矩阵）；

（2）A的所有主子式（行数与列数取成相同的子式）都大于等于零，而且至少有一个等于零；

（3）A的n个特征值都大于等于零，而且至少有一个等于零。

凸函数的判别

梯度： $\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \cdots, \frac{\partial f(x)}{\partial x_n} \right)^T$

Hessian(Hesse)矩阵：

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

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

$$\nabla f(x) = Ax + b$$

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

凸函数的判别

方向导数

设$x^0 \in E^n$, $p \in E^n$, $p \neq 0$，则函数$f(x)$在点x^0关于方向p的方向导数定义为：

$$\frac{\partial f(x^0)}{\partial p} = \lim_{t \to 0^+} \frac{f(x^0 + tp) - f(x^0)}{t}.$$

我们用$Df(x^0; p)$表示f在点x^0关于方向p的方向导数。

方向导数通常用下面的公式计算：

$$Df(x^0; p) = \nabla f(x^0)^T p.$$

凸函数的判别

Taylor展开

$$f(x)=\frac{f(x_0)}{0!}+\frac{f'(x_0)}{1!}(x-x_0)+\frac{f''(x_0)}{2!}(x-x_0)^2+\ldots+\frac{f^{(n)}(x_0)}{n!}(x-x_0)^n+R_n(x)$$

拉格朗日余项（其中$\theta\in(0,1)$）：

$$R_n(x)=f^{(n+1)}[x_0+\theta(x-x_0)]\frac{(x-x_0)^{n+1}}{(n+1)!}$$

$$f(x+p)=f(x)+\nabla f(x)^T p+\|p\|\alpha(x,p),$$

$$f(x+p)=f(x)+\nabla f(x)^T p+\frac{1}{2}p^T\nabla^2 f(x)p+\|p\|^2\alpha(x,p)$$

其中当$p\to 0^+$时，$\alpha(x,p)\to 0$

凸函数的判别

凸函数的充要条件

定理（一阶充要条件）

设S是E^n中非空开凸集，$f(x)$是定义在S上的可微函数，则$f(x)$为凸函数的充要条件是对任意两点$x^{(1)}, x^{(2)} \in S$，有

$$f(x^{(2)}) \geq f(x^{(1)}) + \nabla f(x^{(1)})^T (x^{(2)} - x^{(1)});$$

$f(x)$为严格凸函数的充要条件是对任意互不相同两点$x^{(1)}, x^{(2)} \in S$，有

$$f(x^{(2)}) > f(x^{(1)}) + \nabla f(x^{(1)})^T (x^{(2)} - x^{(1)}).$$

凸函数的判别

证 明

“$\Rightarrow$”设f是S上的凸函数，则对$\forall x^{(1)}, x^{(2)} \in S$及$\lambda \in (0,\ 1)$，有

$$f(\lambda x^{(2)} + (1-\lambda)x^{(1)}) \leq \lambda f(x^{(2)}) + (1-\lambda) f(x^{(1)})$$

即 $$\frac{f(x^{(1)} + \lambda(x^{(2)} - x^{(1)})) - f(x^{(1)})}{\lambda} \leq f(x^{(2)}) - f(x^{(1)})$$

令$\lambda \to 0^+$，由f的可微性，得f在点$x^{(1)}$关于方向$x^{(2)} - x^{(1)}$的方向导数

$$Df(x^{(1)}; x^{(2)} - x^{(1)}) = \nabla f(x^{(1)})^T (x^{(2)} - x^{(1)}) \leq f(x^{(2)}) - f(x^{(1)}),$$

$$\Rightarrow f(x^{(2)}) \geq f(x^{(1)}) + \nabla f(x^{(1)})^T (x^{(2)} - x^{(1)}).$$

“$\Rightarrow$”当f是S上的严格凸函数时，对$\forall x^{(1)}, x^{(2)} \in S$及$x^{(1)} \neq x^{(2)}$，

取$\lambda = \frac{1}{2}$，则$y = \frac{1}{2}x^{(1)} + \frac{1}{2}x^{(2)} \in S$且

$$f(y) = f\left(\frac{1}{2}x^{(1)} + \frac{1}{2}x^{(2)}\right) < \frac{1}{2}f(x^{(1)}) + \frac{1}{2}f(x^{(2)}).$$

$\because$ f是凸函数，

$$\therefore \quad f(y) \geq f(x^{(1)}) + \nabla f(x^{(1)})^T (y - x^{(1)}).$$

$$\therefore \quad \frac{1}{2}f(x^{(1)}) + \frac{1}{2}f(x^{(2)}) > f(x^{(1)}) + \nabla f(x^{(1)})^T (y - x^{(1)})$$

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

$$\therefore \quad f(x^{(2)}) > f(x^{(1)}) + \nabla f(x^{(1)})^T (x^{(2)} - x^{(1)}).$$

“⇐” 设对$\forall x^{(1)}, x^{(2)} \in S$，有

$f(x^{(2)}) \geq f(x^{(1)}) + \nabla f(x^{(1)})(x^{(2)} - x^{(1)})$。

$\forall \lambda \in (0,1)$，令$y = \lambda x^{(1)} + (1-\lambda)x^{(2)}$，则$y \in S$。

由假设，对$x^{(1)}, y \in S$及$x^{(2)}, y \in S$有

$$f(x^{(1)}) \geq f(y) + \nabla f(y)^T (x^{(1)} - y)$$

$$f(x^{(2)}) \geq f(y) + \nabla f(y)^T (x^{(2)} - y)$$

$$\therefore \quad \lambda f(x^{(1)}) + (1-\lambda) f(x^{(2)})$$

$$\geq f(y) + \nabla f(y)^T (\lambda x^{(1)} + (1-\lambda)x^{(2)} - y)$$

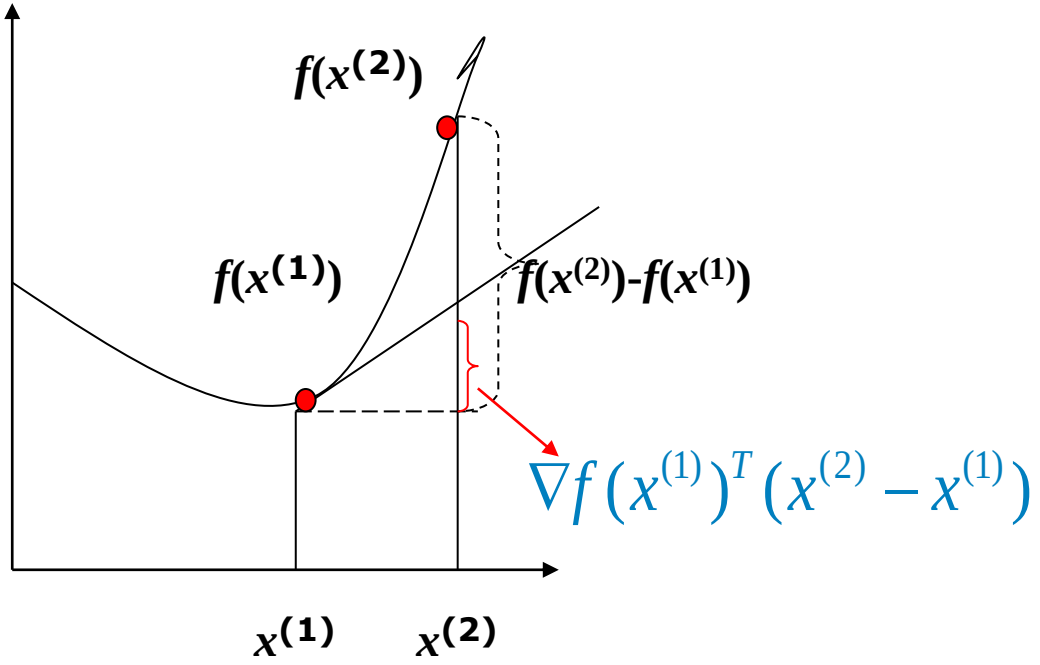
$$= f(y) = f(\lambda x^{(1)} + (1-\lambda)x^{(2)})$$

$\therefore$ f是凸函数。

凸函数的判别

几何意义

$f(x)$是凸函数
当且仅当任意点处的切线增量不超过函数的增量。



推论：设S是E^n中的凸集，$\bar{x} \in S, f(x)$是定义在E^n上的凸函数，且在点$\bar{x}$可微，则对任意$x \in S$，有

$$f(x) \geq f(\bar{x}) + \nabla f(\bar{x})^T(x - \bar{x})$$

凸函数的判别

定理(二阶充要条件)：设S是E^n中非空开凸集，$f(x)$是定义在S上的二次可微函数，则$f(x)$为凸函数的充要条件是对任意$x \in S$，$f(x)$在x处的$Hessian$矩阵$\nabla^2 f(x)$是半正定的。

凸函数的判别

证明： "$\Rightarrow$"设f是S上的凸函数, 对任意$\bar{x} \in S$

$\because S$是开集，则对$\forall x \in E^n$，$\exists \delta > 0$使当$\lambda \in (0, \delta)$，有$\bar{x} + \lambda x \in S$。

$$\therefore f(\bar{x} + \lambda x) \geq f(\bar{x}) + \lambda \nabla f(\bar{x})^T x \qquad (1)$$

$\because f$在点$\bar{x} \in S$二次可微，

$$\therefore f(\bar{x} + \lambda x) = f(\bar{x}) + \lambda \nabla f(\bar{x})^T x + \frac{1}{2}\lambda^2 x^T \nabla^2 f(\bar{x}) x + \lambda^2 \| x \|^2 \alpha(\bar{x}, \lambda x) \qquad (2)$$

其中$\lim_{\lambda \to 0} \alpha(\bar{x}, \lambda x) = 0$。令$a = \alpha(\bar{x}, \lambda x)$.

由(1),(2)得，$\frac{1}{2}\lambda^2 x^T \nabla^2 f(\bar{x}) x + \lambda^2 \| x \|^2 a \geq 0$。

两边除以λ^2后，令$a \to 0$，得，$x^T \nabla^2 f(\bar{x}) x \geq 0$。

凸函数的判别

"$\Leftarrow$"设$\nabla^2 f(x)$在任意点$x \in S$半正定，对$\forall x, \bar{x} \in S$，

由带Lagrange余项的二阶*Taylar*展开式，得

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

其中$\xi = \lambda \bar{x} + (1 - \lambda)x$，$\lambda \in (0,1)$

因为S是凸集，所以$\xi \in S$，又$\nabla^2 f(x)$半正定，

$$\therefore \frac{1}{2}(x - \bar{x})^T \nabla^2 f(\xi)(x - \bar{x}) \geq 0$$

$$\Rightarrow f(x) \geq f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x})$$

$\therefore f$是凸函数。

凸函数的判别

定理：设S是E^n中非空开凸集，$f(x)$是定义在S上的二次可微函数，如果对任意$x \in S$，有f在x处的*Hessian*矩阵$\nabla^2 f(x)$正定，则$f(x)$为严格凸函数。

对二次函数$f(x)=\frac{1}{2}x^T Ax+bx+c$，

若A正定，则$f(x)$为严格凸函数；

若A半正定，则$f(x)$为凸函数。

凸函数的判别

Ex：判断下列函数是否为凸函数.

(1) $f(x_1, x_2) = x_1^2 - 2x_1x_2 + x_2^2 + x_1 + x_2$

(2) $f(x_1, x_2) = x_1 e^{-(x_1+x_2)}$

(3) $f(x_1, x_2, x_3) = x_1x_2 + 2x_1^2 + x_2^2 + 2x_3^2 - 6x_1x_3.$

全局极小点的判别

定理：设$f(x)$是定义在凸集S上的可微凸函数，若$\exists x^* \in S$，使对$\forall x \in S$，都有

$$\nabla f(x^*)^T (x - x^*) \geq 0,$$

则x^*是$f(x)$在凸集S上的全局极小点。

证明：对$\forall x \in S$，因为$f(x)$是凸函数，所以有

$$f(x) \geq f(x^*) + \nabla f(x^*)^T (x - x^*)$$

$$\because \nabla f(x^*)^T (x - x^*) \geq 0$$

$$\therefore f(x) \geq f(x^*)$$

$\Rightarrow$ x^*为全局极小点。

凸规划

- 凸规划：求凸函数在凸集上的极小点。

$$\begin{aligned} \min \quad & f(x) \\ s.t. \quad & g_i(x) \geq 0, i = 1, \cdots, m \\ & h_j(x) = 0, \ j = 1, \cdots, l \end{aligned}$$

若$f(x)$是凸函数，$g_i(x)(i = 1, \cdots, m)$是凹函数，$h_j(x)(j = 1, \cdots, l)$是线性函数，则原问题为凸规划。

凸规划

定理 1.2

考虑非空可行域

$$X = \{x \mid c_i(x) \geq 0, \quad i = 1, 2, \cdots, m\}. \tag{1.17}$$

如果每一个约束函数$c_i(x)$是凹函数，则可行域X是凸集。

注 1.2

如果可行域是

$$X = \{x \mid c_i(x) \leq 0, \quad i = 1, 2, \cdots, m\}, \tag{1.18}$$

则当每一个约束函数$c_i(x)$是凸函数时，可行域X是凸集。

无约束优化问题的最优性条件

$$\begin{aligned} \min \quad & f(x) \\ s.t. \quad & x \in E^n \end{aligned}$$

定义： 对 $\min\limits_{x \in E^n} f(x)$，设 $\bar{x} \in E^n$ 是任给一点，$d \neq 0$，若存在 $\delta > 0$，使得对任意的 $\lambda \in (0, \delta)$，有 $f(\bar{x} + \lambda d) < f(\bar{x})$，则称 d 为 $f(x)$ 在点 $\bar{x}$ 处的下降方向。

无约束优化问题的最优性条件

引理：设函数$f(x)$在点$\bar{x}$可微，若存在$d \neq \mathbf{0}$使$\nabla f(\bar{x})^T d < \mathbf{0}$，则存在$\delta > \mathbf{0}$，使对$\forall \lambda \in (\mathbf{0}, \delta)$，有$f(\bar{x} + \lambda d) < f(\bar{x})$。

证明：对$\lambda > 0$, $d \neq 0$由泰勒展开公式

$$f(\bar{x} + \lambda d) = f(\bar{x}) + \lambda \nabla f(\bar{x})^T d + \| \lambda d \| \alpha(\bar{x}, \lambda d)$$

$$= f(\bar{x}) + \lambda [\nabla f(\bar{x})^T d + \| d \| \alpha(\bar{x}, \lambda d)]$$

其中当$\lambda \to 0^+$时，$\alpha(\bar{x}, \lambda d) \to 0$

故有 $\lim\limits_{\lambda \to 0^+} \dfrac{f(\bar{x} + \lambda d) - f(\bar{x})}{\lambda} = \nabla f(\bar{x})^T d < 0$

$\therefore$ 存在$\delta > 0$，使得当$\lambda \in (0, \delta)$时，有

$f(\bar{x} + \lambda d) < f(\bar{x})$.

无约束优化问题的最优性条件

必要条件

定理1:(一阶必要条件)设函数$f(x)$在点$\bar{x}$处可微，若$\bar{x}$是局部极小点，则$\nabla f(\bar{x})=0.$

证明：设$\nabla f(\bar{x})\neq 0$，取$d=-\nabla f(\bar{x})\neq 0,$

则有$\nabla f(\bar{x})^T d=\nabla f(\bar{x})^T(-\nabla f(\bar{x}))$

$$=-\|\nabla f(\bar{x})\|^2<0$$

由引理，存在$\delta>0,$使当$\lambda\in(0,\delta)$时，有

$$f(\bar{x}+\lambda d)<f(\bar{x})$$

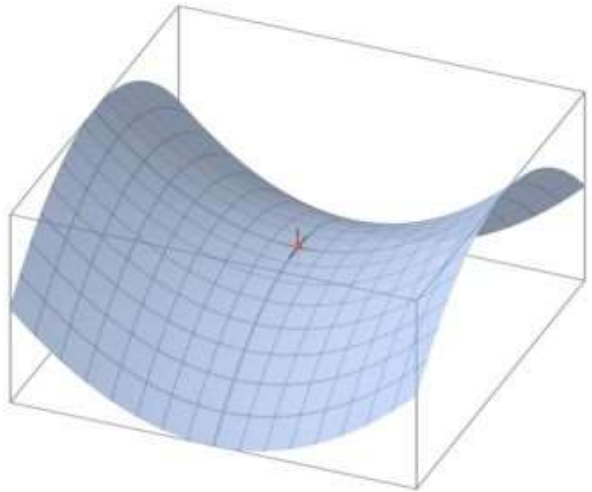
与$\bar{x}$是局部极小点矛盾。

在点为(0,0)的海塞矩阵是:

无约束优化问题的最优性条件

定义：若$f(x)$在点x^*可微，并且$\nabla f(x^*)=0$，则x^*称为$f(x)$的一个驻点(或者平稳点，stationary point)。既不是极小点，也不是极大点的驻点称为鞍点(saddle point)。

例：对于二次函数$f(x)=x_1^2-x_2^2$, $x^*=(0,0)^T$是它的驻点，但是该点既不是极小点，也不是极大点，所以x^*是$f(x)$的鞍点。



无约束优化问题的最优性条件

定理2:(二阶必要条件)设$f(x)$在$\bar{x}$处二阶可微，若$\bar{x}$是局部极小点，则$\nabla f(\bar{x})=0$，且*Hessian*矩阵$\nabla^2 f(\bar{x})$是半正定的。

证明：由定理1，$\nabla f(\bar{x})=0.$

设d是任意一个n 维非零向量，$\because$ $f(x)$在$\bar{x}$处二阶可微,且$\nabla f(\bar{x})=0,$

$$\therefore\ f(\bar{x}+\lambda d)=f(\bar{x})+\lambda\nabla f(\bar{x})^T d+\frac{1}{2}\lambda^2 d^T\nabla^2 f(\bar{x})d+\lambda^2\|d\|^2\alpha(\bar{x},\lambda d)$$

$$\Rightarrow\quad \frac{f(\bar{x}+\lambda d)-f(\bar{x})}{\lambda^2}=\frac{1}{2}d^T\nabla^2 f(\bar{x})d+\|d\|^2\alpha(\bar{x},\lambda d)$$

其中当$\lambda\to 0$时,$\alpha(\bar{x},\lambda d)\to 0$

$\because$ $\bar{x}$是局部极小点,当$|\lambda|$充分小时,必有$f(\bar{x}+\lambda d)\geq f(\bar{x})$

$\therefore$ 当$\lambda\to 0$时,有$d^T\nabla^2 f(\bar{x})d\geq 0$，即$\nabla^2 f(\bar{x})$为半正定的.

无约束优化问题的最优性条件

注意：设函数$f(x)$在点$\bar{x}$的邻域内二次可微，若梯度$\nabla f(\bar{x})=0$,且*Hessian*矩阵$\nabla^2 f(x)$在该邻域内半正定，则$\bar{x}$不一定是局部极小点，可能是鞍点。

例：$f(x)=x^4-y^4$,

梯度$\nabla f(x,y)=(4x^3,-4y^3)^T$,有$\nabla f(0,0)=(0,0)^T$.

*Hessian*矩阵$\nabla^2 f(x)=\begin{bmatrix}12x^2 & 0\\ 0 & -12y^2\end{bmatrix}$

在点$(0,0)$邻域内半正定，但在该点x轴方向增加，在y轴方向减小。则点$(0,0)$是鞍点。

无约束优化问题的最优性条件

定理3（二阶充分条件）：设函数$f(x)$在点$\bar{x}$处二次可微，若梯度$\nabla f(\bar{x})=0$,且$Hessian$矩阵$\nabla^2 f(\bar{x})$正定，则$\bar{x}$是严格局部极小点。

证明：对任意的$d\in E^n, d\neq 0, \because f(x)$在$\bar{x}$处二阶可微且$\nabla f(\bar{x})=0$,

$$\therefore f(\bar{x}+\lambda d)=f(\bar{x})+\frac{1}{2}\lambda^2 d^T\nabla^2 f(\bar{x})d+\lambda^2\|d\|^2\alpha(\bar{x},\lambda d)$$

其中$\lim_{\lambda\to 0}\alpha(\bar{x},\lambda d)=0.$

$\because d^T\nabla^2 f(\bar{x})d>0, \therefore$ 存在$\delta>0$,使得当$\lambda\in(0,\delta)$时,有

$$\frac{1}{2}\lambda^2 d^T\nabla^2 f(\bar{x})d+\lambda^2\|d\|^2\alpha(\bar{x},\lambda d)>0$$

$$\Rightarrow \quad f(\bar{x}+\lambda d)>f(\bar{x})$$

由d的任意性,知$\bar{x}$是严格局部极小点.

无约束优化问题的最优性条件

推论：对于正定二次函数$f(x)=\frac{1}{2}x^TAx+b^Tx+c$

(A对称正定)，有唯一极小点$x^*=-A^{-1}b$.

证明：$\because\ f(x)=\frac{1}{2}x^TAx+b^Tx+c$

$\therefore\quad \nabla f(x)=Ax+b,\quad \nabla^2 f(x)=A$

令$\nabla f(x)=0,\quad \because A$正定，$\therefore Ax+b=0$有唯一解$x^*=-A^{-1}b$.

显然$\nabla f(x^*)=0$且$\nabla^2 f(x^*)=A$正定,

$\therefore x^*$为唯一极小点.

无约束优化问题的最优性条件

定理4：设$f(x)$是定义在E^n上的可微凸函数，$\bar{x} \in E^n$，则$\bar{x}$为整体极小点的充要条件是$\nabla f(\bar{x}) = 0.$

证明：只证充分性。

设$\nabla f(\bar{x}) = 0.$

$\because$ $f(x)$是E^n上的可微凸函数，

$\therefore$ 对任意的$x \in E^n$，有

$$f(x) \geq f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}) = f(\bar{x})$$

$\therefore$ $\bar{x}$为整体极小点。

性质：凸规划的局部极小点就是整体极小点，且极小点的集合为凸集。

无约束优化问题的最优性条件

例：$\min\ f(x) = 5x_1^2 - 6x_1x_2 + 5x_2^2$

解：$\dfrac{\partial f}{\partial x_1} = 10x_1 - 6x_2, \dfrac{\partial f}{\partial x_2} = -6x_1 + 10x_2$

令$\nabla f(x) = 0$,得$x^* = (0,\ 0)^T$,

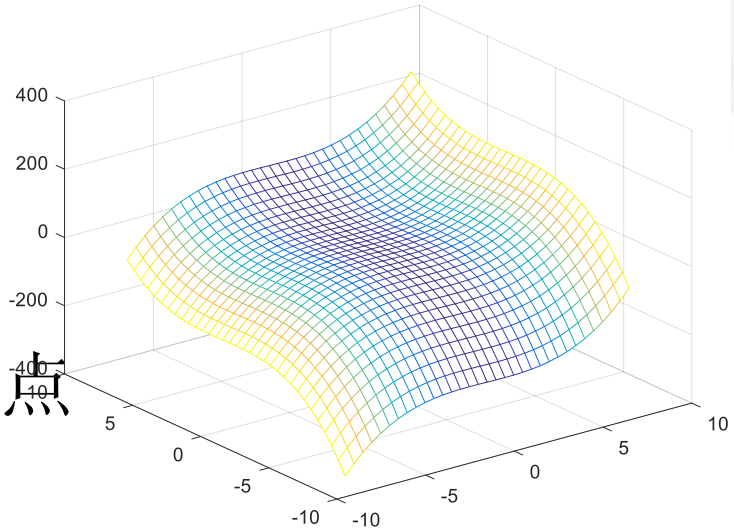
$\because \nabla^2 f(x) = \begin{pmatrix} 10 & -6 \\ -6 & 10 \end{pmatrix}$正定且$f(x)$为凸函数

$\therefore x^*$为整体极小点。

无约束优化问题的最优性条件

例：求$f(x)=\frac{1}{3}x_1^3+\frac{1}{3}x_2^3-x_2^2-x_1$的局部极小点。

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



令$\nabla f(x)=0$,得$\begin{cases}x_1^2-1=0\\ x_2^2-2x_2=0\end{cases}$，得驻点

$x^{(1)}=\begin{pmatrix}1\\0\end{pmatrix},\ x^{(2)}=\begin{pmatrix}1\\2\end{pmatrix},\ x^{(3)}=\begin{pmatrix}-1\\0\end{pmatrix},\ x^{(4)}=\begin{pmatrix}-1\\2\end{pmatrix}$

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

下降迭代算法

迭代： 从一点$x^{(k)}$出发，按照某种规则A，求出后继点$x^{(k+1)}$，用$k+1$代替k，重复以上过程，得到一个解的序列$\{x^{(k)}\}$，若该序列有极限点x^*，即

$$\lim_{k\to\infty}\left\|x^{(k)}-x^*\right\|=0$$

则称它收敛于x^*。

下降： 在每次迭代中，后继点处的函数值要有所减少。

下降迭代算法

下降迭代算法的步骤：

1.选定某一初始点$x^{(0)}$，置$k=0$。

2.确定搜索方向$d^{(k)}$。

3.从$x^{(k)}$出发，沿方向$d^{(k)}$求步长λ_k，以产生下一个迭代点$x^{(k+1)}$。

4.检查$x^{(k+1)}$是否为极小点或近似极小点，若是，则停止迭代；否则，令$k:=k+1$，返回2。

选取搜索方向是最关键的一步，各种算法的区别，主要在于确定搜索方向的方法不同。

下降迭代算法

确定步长λ_k的主要方法

1.令它等于某一常数。

2.可接受点算法，即只要能使目标函数值下降，可任意选取步长λ_k。

3.基于沿搜索方向使目标函数值下降最多，即沿射线

$$x = x^{(k)} + \lambda d^{(k)}$$

求目标函数$f(x)$的极小

$$f(x^{(k)} + \lambda_k d^{(k)}) = \min f(x^{(k)} + \lambda d^{(k)}).$$

由于这项工作是求以λ 为变量的一元函数的极小点，故常称这一过程为（最优）一维搜索，这样确定的步长为最佳步长。

下降迭代算法

定理：

设目标函数$f(x)$具有一阶偏导数，$x^{(k+1)}$由下列规则产生：

$$\begin{cases} f(x^{(k)}+\lambda_k d^k)=\min\limits_{\lambda} f(x^{(k)}+\lambda d^k) \\ x^{(k+1)}=x^{(k)}+\lambda_k d^k \end{cases}$$

则有$\nabla f(x^{(k+1)})^T d^k=0$。

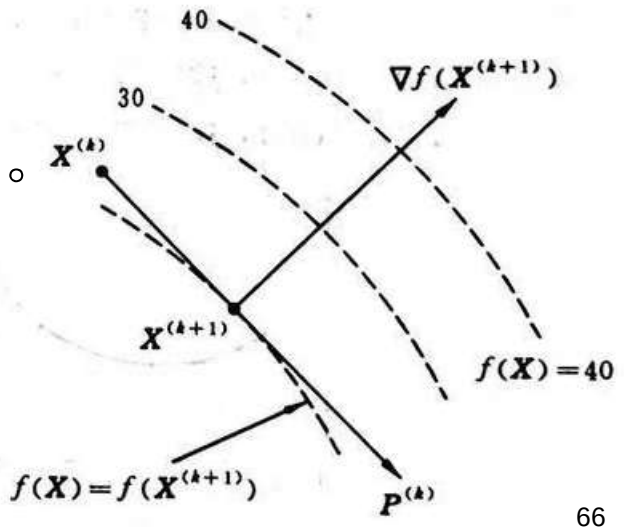
证明： 记$\varphi(\lambda)=f(x^{(k)}+\lambda d^k)$

若λ_k是最优步长，则λ_k为$\varphi(\lambda)$的驻点。

$\therefore \quad \varphi'(\lambda_k)=0$

而$\varphi'(\lambda_k)=\nabla f(x^{(k)}+\lambda_k d^k)^T d^k$

$\therefore \quad \nabla f(x^{(k)}+\lambda_k d^k)^T d^k=0$



下降迭代算法

实用收敛准则

1. $\left\|x^{(k+1)}-x^{(k)}\right\|<\varepsilon$或者$\dfrac{\left\|x^{(k+1)}-x^{(k)}\right\|}{\left\|x^{(k)}\right\|}<\varepsilon.$

2. $f(x^{(k)})-f(x^{(k+1)})<\varepsilon$或者$\dfrac{f(x^{(k)})-f(x^{(k+1)})}{\left|f(x^{(k)})\right|}<\varepsilon.$

3. $\left\|\nabla f(x^{(k)})\right\|<\varepsilon$(无约束最优化中).

收敛速率

定义：设序列$\{\gamma^{(k)}\}$收敛于γ^*，定义满足

$$\lim_{k\to+\infty}\frac{\left\|\gamma^{(k+1)}-\gamma^*\right\|}{\left\|\gamma^{(k)}-\gamma^*\right\|^p}=\beta<\infty$$

的非负数p的上确界为序列$\{\gamma^{(k)}\}$的收敛级。

若序列的收敛级为p，则称序列是p级收敛的。

若$p=1$且$0<\beta<1$，则称序列是以收敛比β线性收敛的。

若$p>1$,或者$p=1$且$\beta=0$，则称序列是超线性收敛的。

收敛级p越大，序列收敛得越快；当收敛级p相同时，收敛比 β 越小，序列收敛得越快。

收敛速率

例： $\{a^k\}\quad 0<a<1$

$$\because \lim_{k\to\infty} a^k = 0,\quad 又 \lim_{k\to\infty}\frac{a^{k+1}}{a^k} = a<1, 且 \lim_{k\to\infty}\frac{a^{k+1}}{(a^k)^r} = \infty(当 r>1 时),$$

$$\therefore \{a^k\} 以收敛比 a 线性收敛于 0。$$

例： $\{a^{2^k}\}\quad 0<|a|<1$

$$\because \lim_{k\to\infty} a^{2^k} = 0,\quad 又 \lim_{k\to\infty}\frac{a^{2^{k+1}}}{\left(a^{2^k}\right)^2} = \lim_{k\to\infty}\frac{a^{2^{k+1}}}{a^{2^{k+1}}} = 1, 且 \lim_{k\to\infty}\frac{a^{2^{k+1}}}{\left(a^{2^k}\right)^r} = \infty(当 r>2 时),$$

$$\therefore \{a^{2^k}\} 是 2 级收敛的。$$

收敛速率

例：$\left\{\left(\frac{1}{k}\right)^k\right\}$

$$\because \lim_{k\to\infty}\left(\frac{1}{k}\right)^k = 0, \quad 又 \lim_{k\to\infty}\frac{\left(\frac{1}{k+1}\right)^{k+1}}{\left(\frac{1}{k}\right)^k} = \lim_{k\to\infty}\left(\frac{k}{k+1}\right)^k \times \frac{1}{k+1} = 0$$

$$\lim_{k\to\infty}\frac{\left(\frac{1}{k+1}\right)^{k+1}}{\left(\left(\frac{1}{k}\right)^k\right)^p} = \lim_{k\to\infty}\left(\frac{k}{k+1}\right)^{k+1} \times \frac{k^{(p-1)k}}{k} = \infty (p>1)$$

$\therefore \left\{\left(\frac{1}{k}\right)^k\right\}$ 是超线性收敛的。

一维搜索

线性搜索方法的迭代格式为

$$x_{k+1} = x_k + \alpha_k d_k, \tag{3.1}$$

其中，x_k，d_k是n维向量，α_k是一个数。 关键是构造搜索方向d_k和步长因子α_k. 设

$$\varphi(\alpha) = f(x_k + \alpha d_k) \tag{3.2}$$

是α的单变量函数。从x_k出发，沿搜索方向d_k，确定步长因子α_k， 使

$$\varphi(\alpha_k) < \varphi(0)$$

的问题就是关于α的线性搜索问题。上式意味着$f(x_k + \alpha_k d_k) < f(x_k)$.

一维搜索

理想的方法是使目标函数沿方向d_k达到极小，即使得

$$f(x_k + \alpha_k d_k) = \min_{\alpha>0} f(x_k + \alpha d_k), \tag{3.3}$$

或者选取$\alpha_k > 0$使得

$$\alpha_k = \min\{\alpha > 0 \mid \nabla f(x_k + \alpha d_k)^T d_k = 0\}. \tag{3.4}$$

满足(3.3) 或(3.4) 的线性搜索称为精确线性搜索，所得到的α_k称为精确步长因子。 一般地，精确线性搜索不但需要的计算量很大，而且在实际上也不必要。 因此人们提出了既花费较少的计算量，又能达到足够下降的不精确线性搜索方法。

一维搜索

精确与非精确线搜索

$$(LS)\quad \min_{\alpha\geq 0} f(x_k+\alpha d_k)$$

- 如果求得的α_k，使得

$$f\left(x_k+\alpha_k d_k\right)=\min_{\alpha} f\left(x_k+\alpha d_k\right)$$

则称该一维搜索为精确线搜索(Exact Line Search).

- 如果存在α_k，使得

$$f\left(x_k+\alpha_k d_k\right)<f\left(x_k\right)$$

则称该一维搜索为非精确线搜索(Inexact Line Search).

一维搜索

- 精确线搜索

  试探法: 黄金分割法、Fibonacci法、二分法

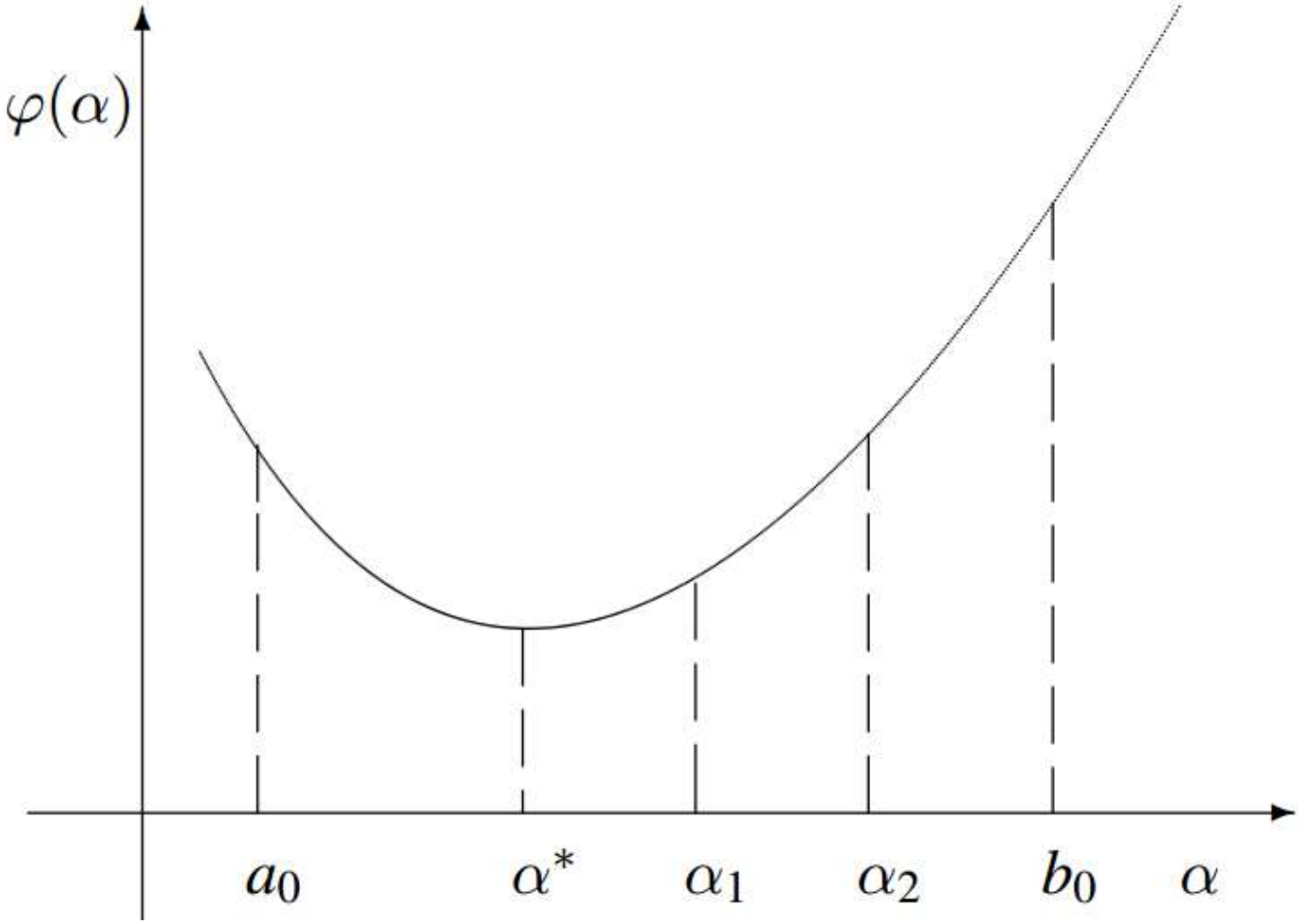
  函数逼近法: Newton法、割线法、抛物线法、三次插值法

- 非精确线搜索

  Armijo步长规则、Goldstein步长规则、Wolfe步长规则

一维搜索

- 一般用于单峰函数的子区间



一维搜索

优化问题的模型： $\min\limits_{x\in R^n} f(x)$ 其中 f 至少一阶连续可微

符号说明： $g(x)=\nabla f(x),\quad g_k=\nabla f(x_k)$

最优步长：设 d_k 为目标函数在 x_k 处的下降方向，称

$$\alpha_k=\arg\min_{\alpha\geq 0} f(x_k+\alpha d_k)$$

为精确步长，又称最优步长。

最优步长的正交性： $d_k^{\ T}\ \nabla f(x_k+\alpha_k d_k)=0$

（一阶最优性条件+复合函数求导的链式法则）

多元复合函数求导法则

记$\varphi(\lambda)=f(x+\lambda d)$

$$u=x+\lambda d=(x_1+\lambda d_1,x_2+\lambda d_2,\cdots,x_n+\lambda d_n)^T=(u_1,u_2,,u_n)^T$$

$$\varphi'(\lambda)=\frac{\partial f(u)}{\partial u_1}\frac{\partial u_1}{\partial \lambda}+\cdots+\frac{\partial f(u)}{\partial u_n}\frac{\partial u_n}{\partial \lambda}=\frac{\partial f(u)}{\partial u_1}d_1+\cdots+\frac{\partial f(u)}{\partial u_n}d_n$$

$$=\nabla f(u)^T d=\nabla f(x+\lambda d)^T d$$

设$x^0\in E^n$, $p\in E^n$, $p\neq 0$，则函数$f(x)$在点x^0关于方向p的方向导数定义为：

$$\frac{\partial f(x^0)}{\partial p}=\lim_{t\to 0^+}\frac{f(x^0+tp)-f(x^0)}{t}=\nabla f(x^0)^T p$$

精确线搜索

精确线搜索方法的基本框架

步1. 取初始点x_0及精确参数$\varepsilon \geq 0$，令$k=0$.

步2. 若$\|g_k\| \leq \varepsilon$,算法终止，否则进入下一步.

步3. 计算 x_k点处的下降方向 d_k ，使得$d_k{}^T g_k < 0$.

步4. 计算最优步长 $\alpha_k = \arg\min\left\{ f\left(x_k + \alpha d_k\right) \mid \alpha \geq 0 \right\}$

步5. 令 $x_{k+1} = x_k + \alpha_k d_k, k = k+1,$ 转步2.

精确线搜索

精确线搜索的特点

最优步长：$$\alpha_k = \arg\min_{\alpha \geq 0} f(x_k + \alpha d_k)$$

- 目标函数在当前迭代点处的下降量达到最大（理想化的搜索策略）
- 便于算法理论分析（如二次终止性、收敛速率等）
- 求解一元极小化问题的计算量
- 搜索方向上的目标函数最大下降量vs全局意义下靠近最优解的程度

精确线搜索

在实际计算时……

- $\alpha_k = \arg\min_{\alpha \geq 0} f(x_k + \alpha d_k)$

$\alpha_k > 0: \quad f(x_k + \alpha_k d_k) < f(x_k)$

非精确线搜索

How much?

- 线搜索

邻域内搜索

信赖域方法

精确线搜索—0.618法

0.618法又称为黄金分割法

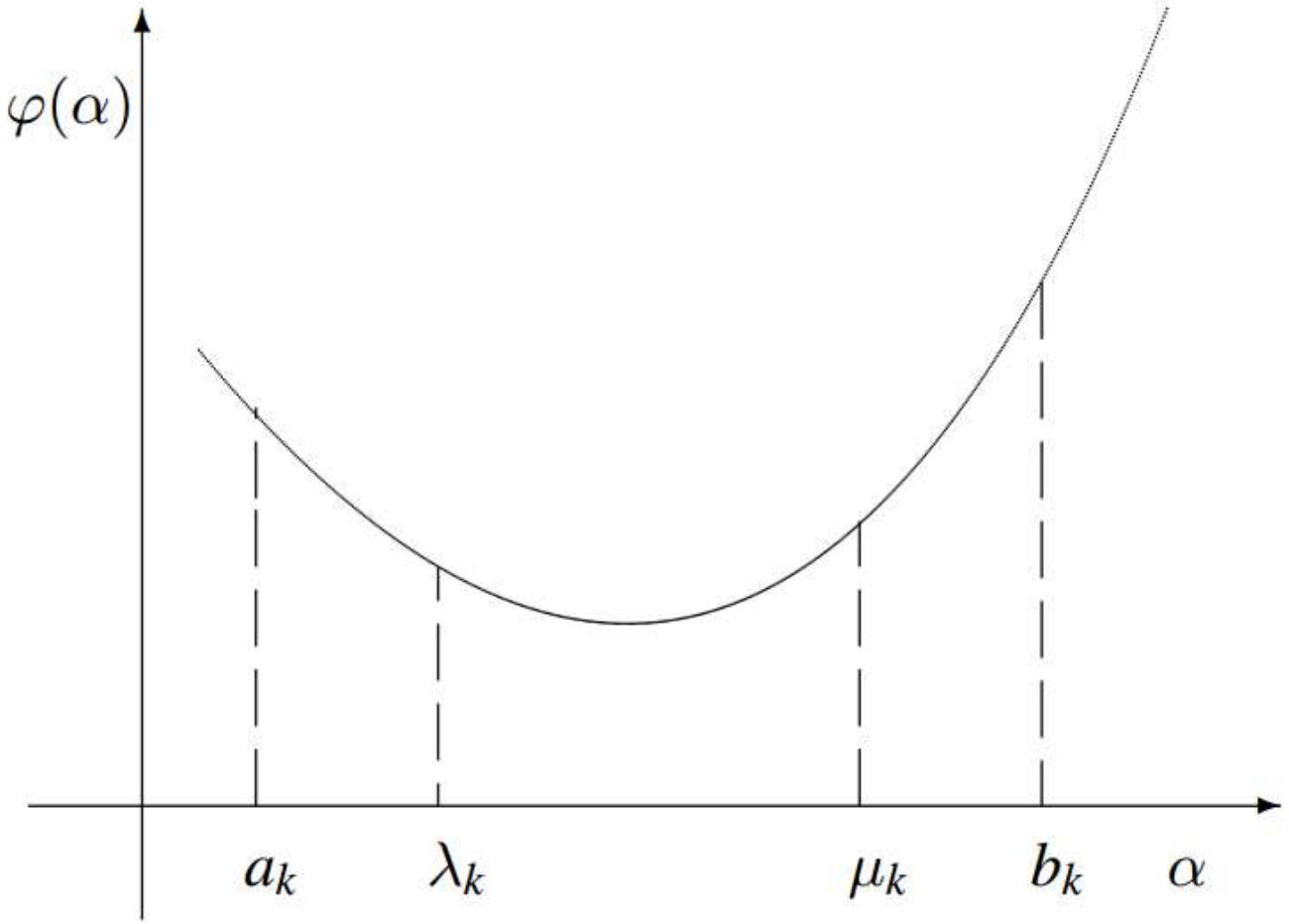
设包含极小点α^*的初始搜索区间为$[a_0, b_0]$, 设

$$\varphi(\alpha) = f(x_k + \alpha d_k)$$

在$[a_0, b_0]$上是凸函数。 0.618法的基本思想是在搜索区间$[a_k, b_k]$上选取两个对称点λ_k, μ_k且$\lambda_k < \mu_k$, 通过比较这两点处的函数值$\varphi(\lambda_k)$和$\varphi(\mu_k)$的大小来决定删除左半区间$[a_k, \lambda_k)$, 还是删除右半区间$(\mu_k, b_k]$. 删除后的新区间长度是原区间长度的0.618倍。新区间包含原区间中两个对称点中的一点， 我们只要再选一个对称点，并利用这两个新对称点处的函数值继续比较。重复这个过程, 最后确定出极小点α^*.

精确线搜索—0.618法

0.618法又称为黄金分割法



精确线搜索—0.618法

设区间$[a_0, b_0]$经k次缩短后变为$[a_k, b_k]$. 在区间$[a_k, b_k]$上选取两个试探点λ_k和μ_k, 要求 满足下列条件:

$$\frac{\mu_k - a_k}{b_k - a_k} = \frac{b_k - \lambda_k}{b_k - a_k} = \tau, \tag{3.6}$$

这里τ 是每次迭代的一个区间缩短率(常数)。这样,

$$\mu_k - a_k = b_k - \lambda_k = \tau(b_k - a_k). \tag{3.7}$$

也就是新的区间长度

$$b_{k+1} - a_{k+1} = \tau(b_k - a_k). \tag{3.8}$$

由(3.7)得到

$$\lambda_k = b_k - \tau(b_k - a_k) = a_k + (1 - \tau)(b_k - a_k), \tag{3.9}$$

精确线搜索—0.618法

$$\mu_k = a_k + \tau(b_k - a_k). \tag{3.10}$$

计算$\varphi(\lambda_k)$和$\varphi(\mu_k)$. 如果$\varphi(\lambda_k) \le \varphi(\mu_k)$, 则删掉右半 区间$(\mu_k, b_k]$, 保留$[a_k, \mu_k]$, 从而新的搜索区间为

$$[a_{k+1}, b_{k+1}] = [a_k, \mu_k]. \tag{3.11}$$

为进一步缩短区间， 需在$[a_{k+1}, b_{k+1}]$上取试探点λ_{k+1}, μ_{k+1}. 由(3.10),

$$\begin{aligned} \mu_{k+1} &= a_{k+1} + \tau(b_{k+1} - a_{k+1}) \\ &= a_k + \tau(\mu_k - a_k) \\ &= a_k + \tau(a_k + \tau(b_k - a_k) - a_k) \\ &= a_k + \tau^2(b_k - a_k). \end{aligned} \tag{3.12}$$

若令

$$\tau^2 = 1 - \tau, \tag{3.13}$$

精确线搜索—0.618法

则由 (3.13)和(3.9) 得到

$$\mu_{k+1} = a_k + (1-\tau)(b_k - a_k) = \lambda_k. \tag{3.14}$$

这样，新的试探点μ_{k+1}不需要重新计算，只要取λ_k就行了，从而在每次迭代中(第一次迭代除外) 只需选取一个试探点即可。

类似地，在$\varphi(\lambda_k) > \varphi(\mu_k)$的情形，新的试探点$\lambda_{k+1} = \mu_k$, 它也不需要重新 计算。在这种情形，我们删去左半区间$[a_k, \lambda_k)$, 保留$[\lambda_k, b_k]$, 这时新的搜索区间为

$$[a_{k+1}, b_{k+1}] = [\lambda_k, b_k]. \tag{3.15}$$

令

$$\lambda_{k+1} = \mu_k, \tag{3.16}$$

$$\mu_{k+1} = a_{k+1} + \tau(b_{k+1} - a_{k+1}). \tag{3.17}$$

然后再比较$\varphi(\lambda_{k+1})$和$\varphi(\mu_{k+1})$. 重复上述过程，直到$b_{k+1} - a_{k+1} \le \varepsilon$.

精确线搜索—0.618法

解方程(3.13)立得

$$\tau = \frac{-1 \pm \sqrt{5}}{2}.$$

由于$\tau > 0$, 故取

$$\tau = \frac{\sqrt{5}-1}{2} \approx 0.618. \tag{3.18}$$

这样(3.9)和(3.10)可分别写成

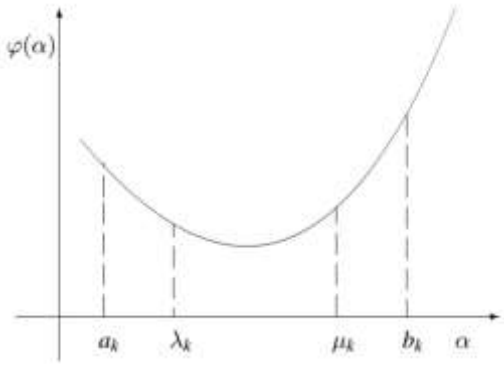
$$\lambda_k = a_k + 0.382(b_k - a_k), \tag{3.19}$$

$$\mu_k = a_k + 0.618(b_k - a_k). \tag{3.20}$$

精确线搜索—0.618法

步 1. 选取初始数据。确定初始搜索区间$[a_0, b_0]$和精度要求$\varepsilon > 0$. 计算最初两个试探点λ_0, μ_0,

$$\begin{aligned} \lambda_0 &= a_0 + 0.382(b_0 - a_0) \\ \mu_0 &= a_0 + 0.618(b_0 - a_0) \end{aligned}$$



计算$\varphi(\lambda_0)$和$\varphi(\mu_0)$，令$k = 0$.

步 2. 比较目标函数值. 若$\varphi(\lambda_k) > \varphi(\mu_k)$, 转步3; 否则转步4.

步 3. 若$b_k - \lambda_k \leq \varepsilon$, 则停止计算，输出$\mu_k$; 否则，令

$$a_{k+1} := \lambda_k,\ b_{k+1} := b_k,\ \lambda_{k+1} := \mu_k,$$

$$\varphi(\lambda_{k+1}) := \varphi(\mu_k),\ \mu_{k+1} := a_{k+1} + 0.618(b_{k+1} - a_{k+1}).$$

计算$\varphi(\mu_{k+1})$, 转步5.

精确线搜索—0.618法

步 4. 若$\mu_k - a_k \leq \varepsilon$, 则停止计算，输出$\lambda_k$; 否则，令

$$a_{k+1} := a_k,\ b_{k+1} := \mu_k,\ \mu_{k+1} := \lambda_k,$$

$$\varphi(\mu_{k+1}) := \varphi(\lambda_k),\ \lambda_{k+1} := a_{k+1} + 0.382(b_{k+1} - a_{k+1}).$$

计算$\varphi(\lambda_{k+1})$, 转步5.

步 5. $k := k + 1$, 转步2.

精确线搜索—0.618法

应用上述0.618法于单变量函数极小问题$\min f(x) = e^x - 5x$, 给出$\varepsilon = 0.01$, $a_0 = 0.0, b_0 = 2.0$. 计算结果如下：

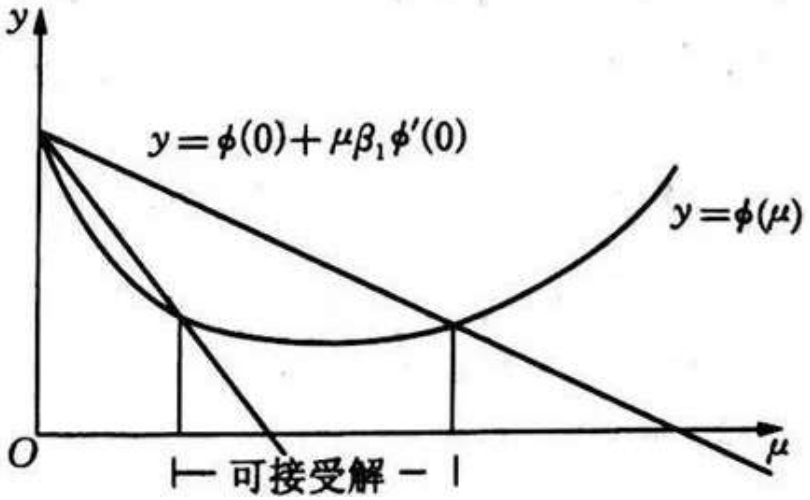
k	a_k	b_k	$\lvert b_k - a_k \rvert$
0	0	2	2
1	0.7640	2	1.2360
2	1.2362	2	0.7638
3	1.2362	1.7082	0.4721
4	1.4165	1.7082	0.2917
5	1.5279	1.7082	0.1803
6	1.5279	1.6393	0.1114
7	1.5705	1.6393	0.0689
8	1.5968	1.6393	0.0426
9	1.5968	1.6231	0.0263
10	1.5968	1.6130	0.0163
11	1.6030	1.6130	0.0100

非精确线搜索- Goldstein

非精确一维搜索的基本思想是求 μ，使得 $\phi(\mu)<\phi(0)$，但不希望 μ 值过大，因为 μ 值过大会引起点列 $\{x^k\}$ 产生大幅度的摆动；也不希望 μ 值过小，因为 μ 值过小会使得点列 $\{x^k\}$ 在未达到 x^* 之前进展缓慢.

预先指定两个参数 β_1，β_2（精度要求），满足 $0<\beta_1<\beta_2<1$，用下面两个不等式来限定步长 μ，即

$$\phi(\mu)\leqslant\phi(0)+\mu\beta_1\phi'(0);\ \phi(\mu)\geqslant\phi(0)+\mu\beta_2\phi'(0).$$



最速下降法

考虑求解无约束优化问题

$$\min_{x\in\mathbb{R}^n} f(x) \tag{4.1}$$

最速下降法是以负梯度方向作为下降方向的极小化算法，又称梯度法，是1874年法国数学家Cauchy提出的。最速下降法是无约束最优化中最简单的方法。

设目标函数$f(x)$在x_k附近连续可微，且$g_k \triangleq \nabla f(x_k) \neq 0$.

将$f(x)$在x_k处Taylor展开，

$$f(x) = f(x_k) + g_k^T(x - x_k) + o(\| x - x_k \|). \tag{4.2}$$

记$x - x_k = \alpha d_k$，则上式写为

$$f(x_k + \alpha d_k) = f(x_k) + \alpha g_k^T d_k + o(\| \alpha d_k \|). \tag{4.3}$$

进一步，我们有

$$f(x_k) - f(x_k + \alpha d_k) = -\alpha g_k^T d_k + o(\| \alpha d_k \|) \tag{4.4}$$

$$= \alpha |g_k^T d_k| + o(\| \alpha d_k \|). \tag{4.5}$$

最速下降法

显然，若d_k满足$g_k^T d_k < 0$，则d_k是下降方向，它使得$f(x_k + \alpha d_k) < f(x_k)$.当$\alpha$取定后，$|g_k^T d_k|$的值越大, 函数$f(x)$在$x_k$处下降量越大。由Cauchy-Schwartz不等式

$$| d_k^T g_k | \leq \| d_k \| \| g_k \| \tag{4.6}$$

可知，当$d_k = -g_k$时，$d_k^T g_k < 0$且$|d_k^T g_k|$达到最大，从而$d_k = -g_k$是最速下降方向。以$d_k = -g_k$为下降方向的方法叫最速下降法。

最速下降法的迭代格式为

$$x_{k+1} = x_k - \alpha_k g_k, \tag{4.7}$$

其中步长因子α_k由线性搜索策略确定。

最速下降法

算法 4.1

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

步 2. 计算$d_k = -g_k$；如果$\| g_k \| \le \varepsilon$，停止。

步 3. 由线性搜索求步长因子α_k.

步 4. 计算$x_{k+1} = x_k + \alpha_k d_k$.

步 5. 如果$\|x_{k+1} - x_k\| \le \varepsilon$, 停止。

步 6. $k := k + 1$，转步2.

定理 4.2

设函数$f(x)$二次连续可微，且$\| \nabla^2 f(x) \| \le M$， 其中$M$是某个正常数。对任何给定的初始点$x_0$，最速下降算法*4.1*或有限终止，或$\lim_{k\to\infty} f(x_k) = -\infty$，或$\lim_{k\to\infty} g_k = 0$. 进一步，收敛速度满足

$$\|x_{k+1} - x^*\| \le (1 - \frac{\lambda_n^2}{\lambda_1^2})\|x_k - x^*\|, \tag{4.8}$$

其中，λ_1 和 λ_n 分别为$\nabla^2 f(x^*)$ 的最大和最小特征值。

最速下降法

例：求$\min\ f(x)=(x_1-1)^2+(x_2-1)^2$，取$x^{(1)}=(0,0)^T,\varepsilon=1e-5.$

解：$$\nabla f(x)=\left(2(x_1-1),2(x_2-1)\right)^T$$

第一次迭代

$$d^{(1)}=-(-2,-2)^T=(2,2)^T,\left\|d^{(1)}\right\|=2\sqrt{2}>\varepsilon$$

$\therefore$从$x^{(1)}$出发，沿方向$d^{(1)}$进行一维搜索，求步长λ_1.

$$\min_{\lambda\geq 0}\varphi(\lambda)=\min_{\lambda\geq 0}\ f(x^{(1)}+\lambda d^{(1)})$$

$$\because\ x^{(1)}+\lambda d^{(1)}=(2\lambda,2\lambda)^T,\therefore\varphi(\lambda)=(2\lambda-1)^2+(2\lambda-1)^2.$$

令$\varphi'(\lambda)=4(2\lambda-1)+4(2\lambda-1)=0$，得$\lambda_1=\dfrac{1}{2}$

最速下降法

第二次迭代

令$x^{(2)} = x^{(1)} + \lambda_1 d^{(1)} = (1,1)^T.$

$$\nabla f(x) = \left(2(x_1 - 1), 2(x_2 - 1)\right)^T$$

$$d^{(2)} = (0,0)^T$$

$$\because \left\| d^{(2)} \right\| = 0 < \varepsilon, \therefore (1,1)^T \text{为最优解。}$$

最速下降法

考虑下面的例子：

$$\min_{x\in\mathbb{R}^2} f(x) = (x_1 - 2)^4 + (x_1 - 2x_2)^2.$$

显然，

$$\nabla f(x) = \begin{bmatrix} 4(x_1 - 2)^3 + 2(x_1 - 2x_2) \\ -4(x_1 - 2x_2) \end{bmatrix}.$$

任意选择初始点$x^{(0)} = [0, 3]^T$, 则$f(x^{(0)}) = 52$. 于是，

$$d_0 = -\nabla f(x^{(0)}) = -\begin{bmatrix} 4(0 - 2)^3 + 2(0 - 2\cdot 3) \\ -4(0 - 2\cdot 3) \end{bmatrix} = \begin{bmatrix} 44 \\ -24 \end{bmatrix}.$$

沿方向d_0 作精确线性搜索，得$\alpha_0 = 0.062$. 于是，

$$x^{(1)} = x^{(0)} + \alpha_0 d_0 = \begin{bmatrix} 0 \\ 3 \end{bmatrix} + 0.062 \begin{bmatrix} 44 \\ -24 \end{bmatrix} = \begin{bmatrix} 2.70 \\ 1.51 \end{bmatrix}.$$

最速下降法

继续下去， 直到$\|\nabla f_k\| \leq \varepsilon$ 或$\|x_{k+1} - x_k\| \leq \varepsilon$为止。 若取$\varepsilon = 0.05$, 则在得到$x^{(5)}$后算法终止， 计算结果如下：

k	x_k	d_k	α_k	$\|x_{k+1} - x_k\|$
0	[0.00, 3.00]	[44.00, -24.00]	0.062	3.08
1	[2.70, 1.51]	[-0.73, -1.28]	0.24	0.36
2	[2.52, 1.20]	[-0.80, 0.48]	0.11	0.10
3	[2.43, 1.25]	[-0.18, -0.28]	0.31	0.11
4	[2.37, 1.16]	[-0.30, 0.20]	0.12	0.04
5	[2.33, 1.18]			

最速下降法

考虑二次函数

$$\min f(x) = \frac{1}{2}x^T Gx$$

的梯度法的显式最优步长因子，其中G是对称正定矩阵。

由于

$$\nabla f(x_k) = Gx_k,$$

故梯度法为

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

这样，

$$\begin{aligned} f(x_{k+1}) &= \frac{1}{2}(x_k - \alpha G x_k)^T G(x_k - \alpha G x_k) \\ &= \frac{1}{2}[x_k^T G x_k - 2\alpha x_k^T G^2 x_k + \alpha^2 x_k^T G^3 x_k], \\ \frac{\partial f(x_{k+1})}{\partial \alpha} &= \frac{1}{2}[-2x_k^T G^2 x_k + 2\alpha x_k^T G^3 x_k]. \end{aligned}$$

最速下降法

令 $\frac{\partial f(x_{k+1})}{\partial \alpha} = 0$, 得

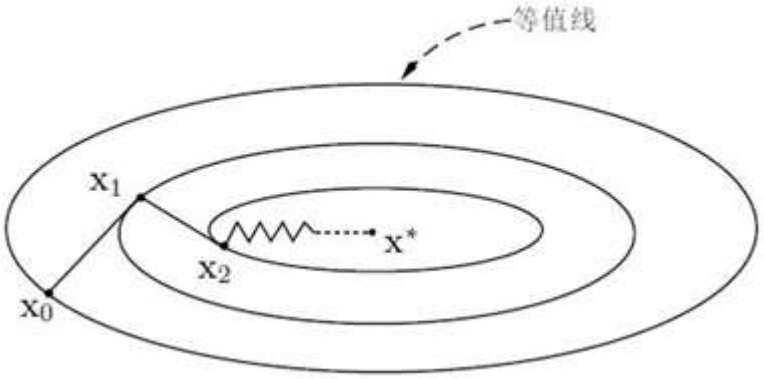
$$\alpha_k = \frac{x_k^T G^2 x_k}{x_k^T G^3 x_k}. \tag{4.10}$$

于是，显式最速下降法为

$$x_{k+1} = x_k - \frac{x_k^T G^2 x_k}{x_k^T G^3 x_k} G x_k. \tag{4.11}$$

最速下降法

最速下降法具有算法和程序设计简单，计算工作量小，存储量小，对初始点没有特别要求等优点。但是，最速下降方向仅是函数的局部性质，对整体求解过程而言，这个方法下降非常缓慢。数值试验表明，当目标函数的等值线 接近于一个圆（球）时，最速下降法下降较快，而当目标函数的等值线是一个扁长的椭球时，最速下降法开始几步下降较快，后来就出现锯齿现象，下降很缓慢。在大量的工程设计和计算中，最速下降法得到了广泛的应用。为了利用 最速下降法的优点，克服其缺陷，若干基于最速下降法的新方法出现了。



牛顿法

考虑求解无约束优化问题

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

牛顿法的基本思想是利用目标函数$f(x)$在迭代点x_k处的二次Taylor展开作为模型函数， 并用这个二次模型函数的极小点序列去逼近目标函数的极小点。

牛顿法

设$x^{(k)}$是$f(x)$的极小点x^*的第k次近似，将$f(x)$在$x^{(k)}$点作二阶*Taylor*展开，得

$$f(x) \approx \varphi(x) = f(x^{(k)}) + \nabla f(x^{(k)})^T (x - x^{(k)})$$

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

为求$\varphi(x)$的极小点，令$\nabla \varphi(x) = \nabla f(x^{(k)}) + \nabla^2 f(x^{(k)})(x - x^{(k)}) = 0$

设$\nabla^2 f(x^{(k)})$可逆，则得

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

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

牛顿法

定理： 设$f(x)$为二次可微函数，$x \in E^n$，$\bar{x}$满足$\nabla f(\bar{x}) = 0$，且$\nabla^2 f(\bar{x})^{-1}$存在。又设$x^{(1)}$充分接近$\bar{x}$，使得存在$k_1, k_2 > 0$，满足$k_1 k_2 < 1$，且对每一个

$$x \in X = \left\{ x \middle| \|x - \bar{x}\| \le \|x^{(1)} - \bar{x}\| \right\}$$

$\left\|\nabla^2 f(x)^{-1}\right\| \le k_1$和$\dfrac{\left\|\nabla f(\bar{x}) - \nabla f(x) - \nabla^2 f(x)(\bar{x} - x)\right\|}{\|\bar{x} - x\|} \le k_2$

成立，则牛顿法产生的序列收敛于$\bar{x}$。

对于正定二次函数，牛顿法一步即可达到最优解。对于一般非二次函数，牛顿法并不能保证经过有限次迭代求得最优解，但如果初始点x_0充分靠近极小点，牛顿法的收敛速度一般是快的。

令$x \in X$，且$x \neq \bar{x}$，又令$y = x - \nabla^2 f(x)^{-1} \nabla f(x)$.

因为$\nabla f(\bar{x}) = 0$

$$\therefore y - \bar{x} = x - \nabla^2 f(x)^{-1} \nabla f(x) - \bar{x}$$

$$= (x - \bar{x}) - \nabla^2 f(x)^{-1} [\nabla f(x) - \nabla f(\bar{x})]$$

$$= \nabla^2 f(x)^{-1} [\nabla f(\bar{x}) - \nabla f(x) - \nabla^2 f(x)(\bar{x} - x)]$$

$$\therefore \|y - \bar{x}\| \leq \left\|\nabla^2 f(x)^{-1}\right\| \left\|\nabla f(\bar{x}) - \nabla f(x) - \nabla^2 f(x)(\bar{x} - x)\right\|$$

$$\leq k_1 k_2 \|\bar{x} - x\| < \|x - \bar{x}\|$$

因迭代产生的序列$\{x^{(k)}\} \subset X$。易知

X为紧集，因此迭代产生的序列含于紧集中。

所以迭代产生的序列$\{x^{(k)}\}$必收敛于$\bar{x}$。

牛顿法

牛顿法计算步骤：

1.给定初点$x^{(0)} \in E^n$，允许误差$\varepsilon > 0$，置$k = 0$。

2.若$\left\|\nabla f(x^{(k)})\right\| < \varepsilon$，则停止计算；否则，转3。

3.计算

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

置$k := k+1$，返回2。

牛顿法

例：求 $\min\ f(x) = x_1^2 + 25x_2^2$

解：取 $x^{(0)} = (2,\ 2)^T$，则

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

$$\nabla^2 f(x^{(0)}) = \begin{pmatrix} 2 & 0 \\ 0 & 50 \end{pmatrix} \qquad \nabla^2 f(x^{(0)})^{-1} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{50} \end{pmatrix}$$

$$x^{(1)} = x^{(0)} - \nabla^2 f(x^{(0)})^{-1} \nabla f(x^{(0)})$$

$$= \begin{pmatrix} 2 \\ 2 \end{pmatrix} - \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{50} \end{pmatrix} \begin{pmatrix} 4 \\ 100 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

牛顿法

例:求 $\min\ f(x)=4x_1^2+x_2^2-x_1^2x_2$

分别取初始点为$x_A=(1,1)^T, x_B=(3,4)^T,$

$x_C=(2,0)^T$,精度要求$\varepsilon=10^{-3}$.

解: $f(x)=4x_1^2+x_2^2-x_1^2x_2$

$$\nabla f(x)=\left(8x_1-2x_1x_2, 2x_2-x_1^2\right)^T$$

$$\nabla^2 f(x)=\begin{pmatrix} 8-2x_2 & -2x_1 \\ -2x_1 & 2 \end{pmatrix}.$$

牛顿法

(1) 取$x^{(1)} = x_A = (1,1)^T$, 则

k	$x^{(k)}$	$f\left(x^{(k)}\right)$	$\nabla f\left(x^{(k)}\right)$	$\left\|\nabla f\left(x^{(k)}\right)\right\|$	$\nabla^2 f\left(x^{(k)}\right)$
1	1.0000 1.0000	4.000	6.0000 1.0000	6.0828	6.0000 −2.0000 −2.000 2.0000
2	−0.7500 −1.2500	4.5156	−7.8750 −3.0625	8.4495	10.500 1.5000 1.5000 2.0000
3	−0.1550 −0.1650	0.1273	−1.2911 −0.3540	1.3388	8.3300 0.3100 0.3100 2.0000
4	−0.0057 −0.0111	0.0003	−0.0459 −0.0223	0.0511	8.0222 0.0115 0.0115 2.0000
5	−0.0000 −0.0000	0.0000	−0.0001 −0.0000	0.0001	8.0000 0.0000 0.0000 2.0000

牛顿法

(2) 取$x^{(1)} = x_B = (3,4)^T$, 则

k	$x^{(k)}$	$f\left(x^{(k)}\right)$	$\nabla f\left(x^{(k)}\right)$	$\left\|\nabla f\left(x^{(k)}\right)\right\|$	$\nabla^2 f\left(x^{(k)}\right)$
1	3.0000 4.0000	16.000	0.0000 −1.0000	1.0000	0.0000 −6.0000 −6.000 2.0000
2	2.8333 4.0000	16.0000	0.0000 −0.2078	0.0278	0.0000 −5.6667 −5.6667 2.0000
3	2.8284 4.0000	16.0000	0.0000 0.0000	0.0000	0.0000 −5.6569 −5.6569 2.0000

牛顿法

(3) 取$x^{(1)} = x_C = (2,0)^T$, 得到

$$\nabla^2 f\left(x^{(1)}\right) = \begin{pmatrix} 8 & -4 \\ -4 & 2 \end{pmatrix}$$

由于*Hessian*矩阵不可逆，无法进行下一步。

用Newton法求解无约束问题会出现以下情形：

（1）收敛到极小点

（2）收敛到鞍点

（3）Hessian矩阵不可逆，无法迭代下去

牛顿法

优点:（1）Newton法产生的点列$\{x^{(k)}\}$若收敛，则收敛速度快---具有至少二阶收敛速率。

(2) Newton法具有二次终止性—对于严格凸二次规划

证明:设A为对称,正定矩阵,且

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

令$\nabla f(x)=Ax+b=0$,得$x^*=-A^{-1}b.$

若用$Newton$迭代公式,从任一点$x^{(0)}$出发,得

$$x^{(1)}=x^{(0)}-\nabla^2 f(x^{(0)})^{-1}\nabla f(x^{(0)})$$

$$=x^{(0)}-A^{-1}(Ax^{(0)}+b)=-A^{-1}b=x^*.$$

牛顿法

例:用牛顿方法求 $\min\ f(x)=\sqrt{1+x^2}$.

分析： 显然0 是唯一最优解，

$$f'(x)=\frac{x}{\sqrt{1+x^2}},\ f''(x)=\frac{1}{\left(1+x^2\right)^{3/2}}$$

牛顿迭代过程为：

$$x_{k+1}=x_k-\frac{f'(x_k)}{f''(x_k)}=x_k-x_k(1+x_k{}^2)=-x_k{}^3$$

显然：

若初始点x_0 满足$|x_0|<1$,则迭代点快速收敛到最优解；

而当$|x_0|\geq 1$时， 目标函数值快速增加， 算法不收敛。

牛顿法

缺点:

（1）可能会出现在某步迭代时，目标函数值上升.

（2）当初始点远离极小点时，牛顿法产生的点列可能不收敛，或者收敛到鞍点，或者Hessian矩阵不可逆，无法计算.

（3）需要计算Hessian矩阵，计算量大.

阻尼牛顿法

步骤：

1.给定初点$x^{(1)} \in E^n$, 允许误差$\varepsilon > 0$, 置$k = 1$。

2.计算$\nabla f(x^{(k)})$, $\nabla^2 f(x^{(k)})^{-1}$。

3.若$\left\|\nabla f(x^{(k)})\right\| < \varepsilon$, 则停止迭代；否则，令

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

4.从$x^{(k)}$出发，沿方向$d^{(k)}$作一维搜索：

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

令$x^{(k+1)} = x^{(k)} + \lambda_k d^{(k)}$

5.置$k := k+1$, 转2。

修正牛顿法

1.给定初点$x^{(1)} \in E^n$,允许误差$\varepsilon > 0$,置$k = 1$。

2.计算$\nabla f(x^{(k)})$，$G_k = \nabla^2 f(x^{(k)})$。若$\left\|\nabla f(x^{(k)})\right\| \le \varepsilon$,

则停止计算，得点$x^{(k)}$；否则转3。

3.置$B_k = G_k + \varepsilon_k I$,其中$\varepsilon_k$是一个非负数，选取$\varepsilon_k$，使得$B_k$是对称正定矩阵，计算修正牛顿方向$d^{(k)} = -B_k{}^{-1}\nabla f(x^{(k)})$.

4.从$x^{(k)}$出发，沿方向$d^{(k)}$作一维搜索：

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

令$x^{(k+1)} = x^{(k)} + \lambda_k d^{(k)}$

5.置$k := k + 1$，转2。

随机梯度下降（SGD）

Consider minimizing an average of functions

$$\min_x \frac{1}{m}\sum_{i=1}^m f_i(x)$$

As $\nabla \sum_{i=1}^m f_i(x) = \sum_{i=1}^m \nabla f_i(x)$, gradient descent would repeat:

$$x^{(k)} = x^{(k-1)} - t_k \cdot \frac{1}{m}\sum_{i=1}^m \nabla f_i(x^{(k-1)}), \quad k = 1, 2, 3, \ldots$$

In comparison, stochastic gradient descent or SGD (or incremental gradient descent) repeats:

$$x^{(k)} = x^{(k-1)} - t_k \cdot \nabla f_{i_k}(x^{(k-1)}), \quad k = 1, 2, 3, \ldots$$

where $i_k \in \{1, \ldots, m\}$ is some chosen index at iteration k

随机梯度下降（SGD）

Two rules for choosing index i_k at iteration k:

- Randomized rule: choose $i_k \in \{1, \ldots, m\}$ uniformly at random
- Cyclic rule: choose $i_k = 1, 2, \ldots, m, 1, 2, \ldots, m, \ldots$

Randomized rule is more common in practice. For randomized rule, note that

$$\mathbb{E}[\nabla f_{i_k}(x)] = \nabla f(x)$$

so we can view SGD as using an unbiased estimate of the gradient at each step

Main appeal of SGD:

- Iteration cost is independent of m (number of functions)
- Can also be a big savings in terms of memory useage

随机梯度下降（SGD）

Given $(x_i, y_i) \in \mathbb{R}^p \times \{0, 1\}$, $i = 1, \ldots, n$, recall logistic regression:

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \underbrace{\Big(- y_i x_i^T \beta + \log(1 + \exp(x_i^T \beta)) \Big)}_{f_i(\beta)}$$

Gradient computation $\nabla f(\beta) = \frac{1}{n} \sum_{i=1}^{n} \big(y_i - p_i(\beta)\big) x_i$ is doable when n is moderate, but not when n is huge

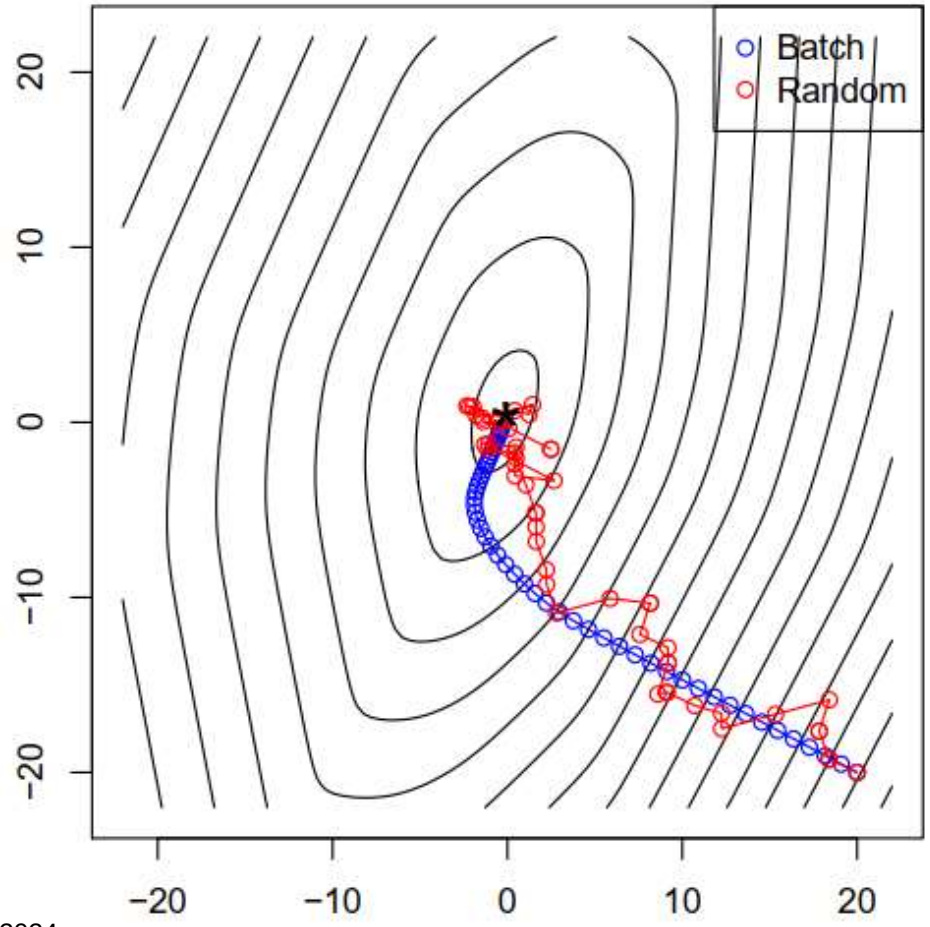
Full gradient (also called batch) versus stochastic gradient:

- One batch update costs $O(np)$
- One stochastic update costs $O(p)$

Clearly, e.g., 10K stochastic steps are much more affordable

随机梯度下降（SGD）

Small example with $n = 10$, $p = 2$ to show the "classic picture" for batch versus stochastic methods:



Blue: batch steps, $O(np)$
Red: stochastic steps, $O(p)$

Rule of thumb for stochastic methods:

- generally thrive far from optimum
- generally struggle close to optimum

Ryan Tibshirani

随机梯度下降（SGD）

Standard in SGD is to use diminishing step sizes, e.g., $t_k = 1/k$

Why not fixed step sizes? Here's some intuition. Suppose we take cyclic rule for simplicity. Set $t_k = t$ for m updates in a row, we get:

$$x^{(k+m)} = x^{(k)} - t \sum_{i=1}^{m} \nabla f_i(x^{(k+i-1)})$$

Meanwhile, full gradient with step size mt would give:

$$x^{(k+1)} = x^{(k)} - t \sum_{i=1}^{m} \nabla f_i(x^{(k)})$$

The difference here: $t \sum_{i=1}^{m} [\nabla f_i(x^{(k+i-1)}) - \nabla f_i(x^{(k)})]$, and if we hold t constant, this difference will not generally be going to zero

随机梯度下降（SGD）

Back to logistic regression, let's now consider a regularized version:

$$\min_{\beta} \frac{1}{n}\sum_{i=1}^{n}\Big(-y_i x_i^T\beta + \log(1+e^{x_i^T\beta})\Big) + \frac{\lambda}{2}\|\beta\|_2^2$$

Write the criterion as

$$f(\beta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\beta), \quad f_i(\beta) = -y_i x_i^T\beta + \log(1+e^{x_i^T\beta}) + \frac{\lambda}{2}\|\beta\|_2^2$$

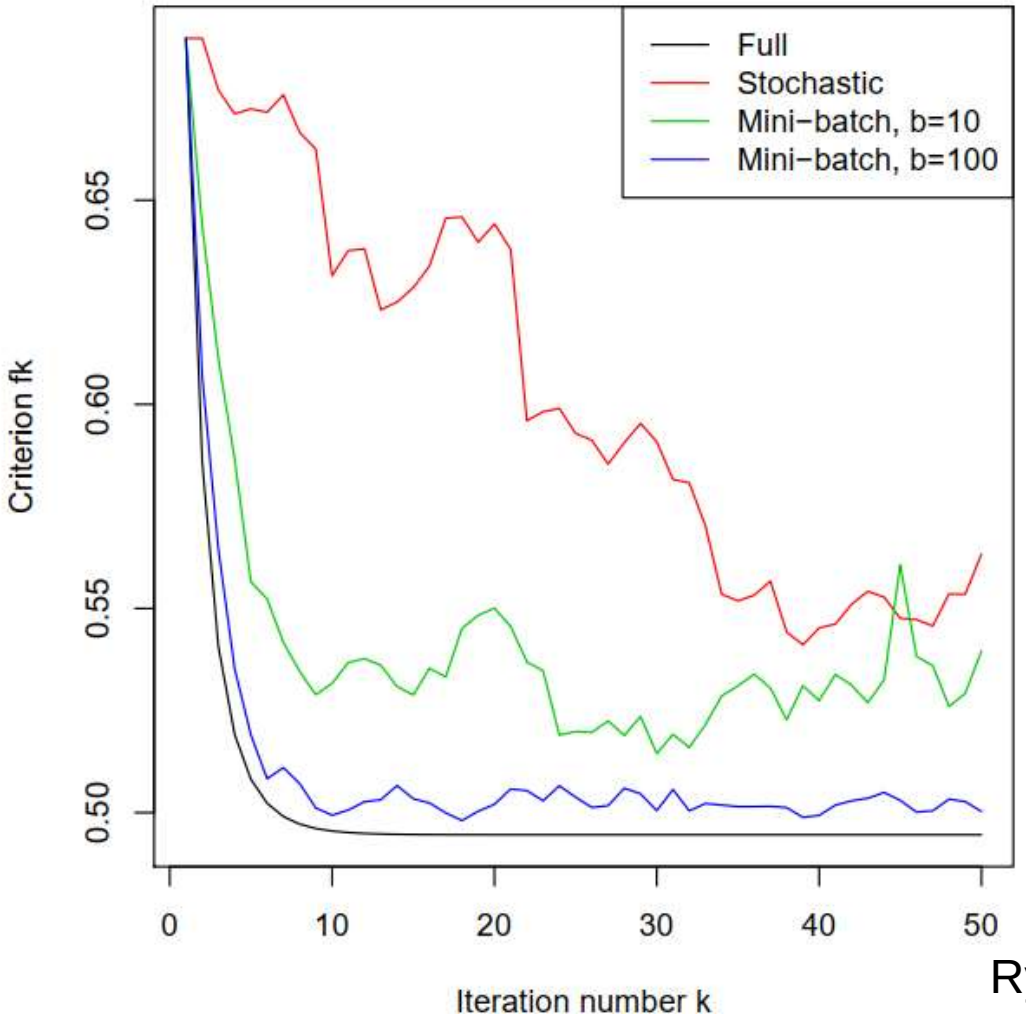
Full gradient computation is $\nabla f(\beta) = \frac{1}{n}\sum_{i=1}^{n}\big(y_i - p_i(\beta)\big)x_i + \lambda\beta$.

Comparison between methods:

- One batch update costs $O(np)$
- One mini-batch update costs $O(bp)$
- One stochastic update costs $O(p)$

随机梯度下降（SGD）

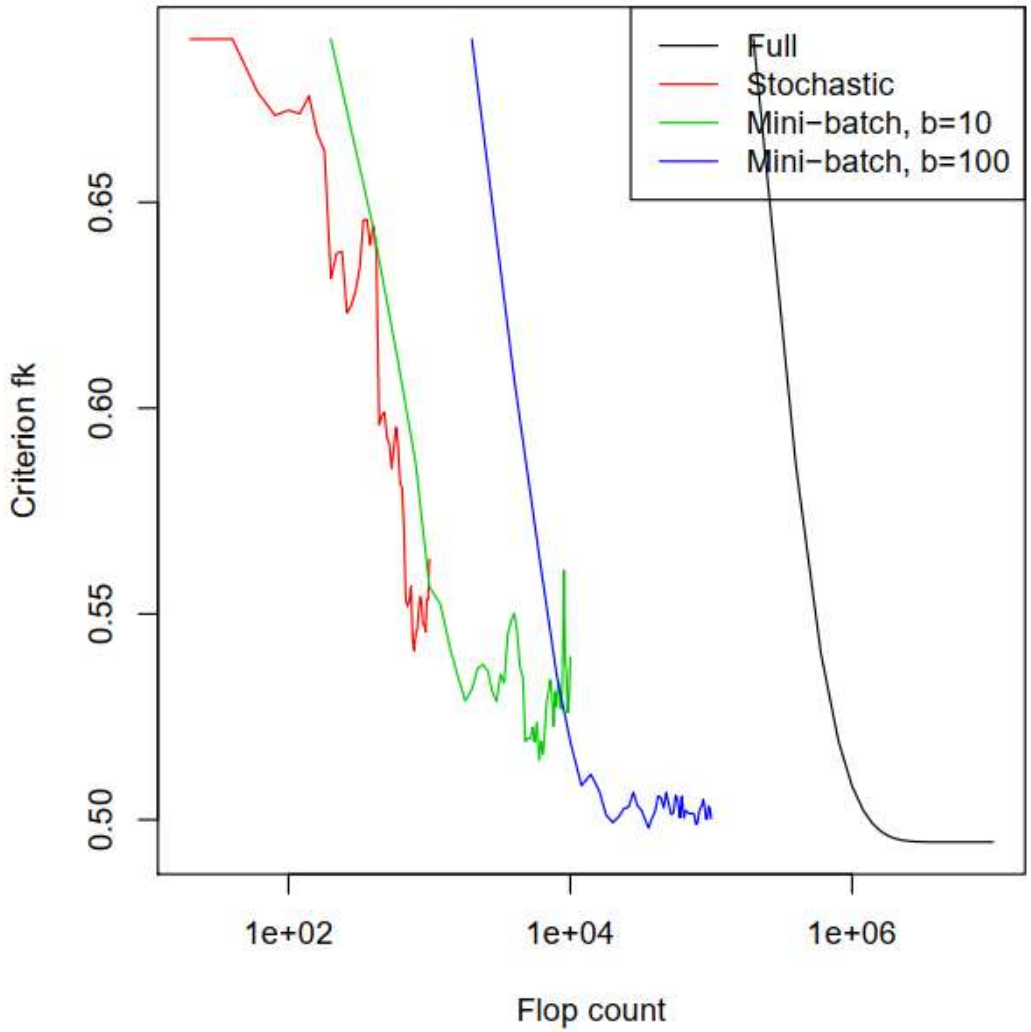
Example with $n = 10,000$, $p = 20$, all methods use fixed step sizes:



Ryan Tibshirani

随机梯度下降（SGD）

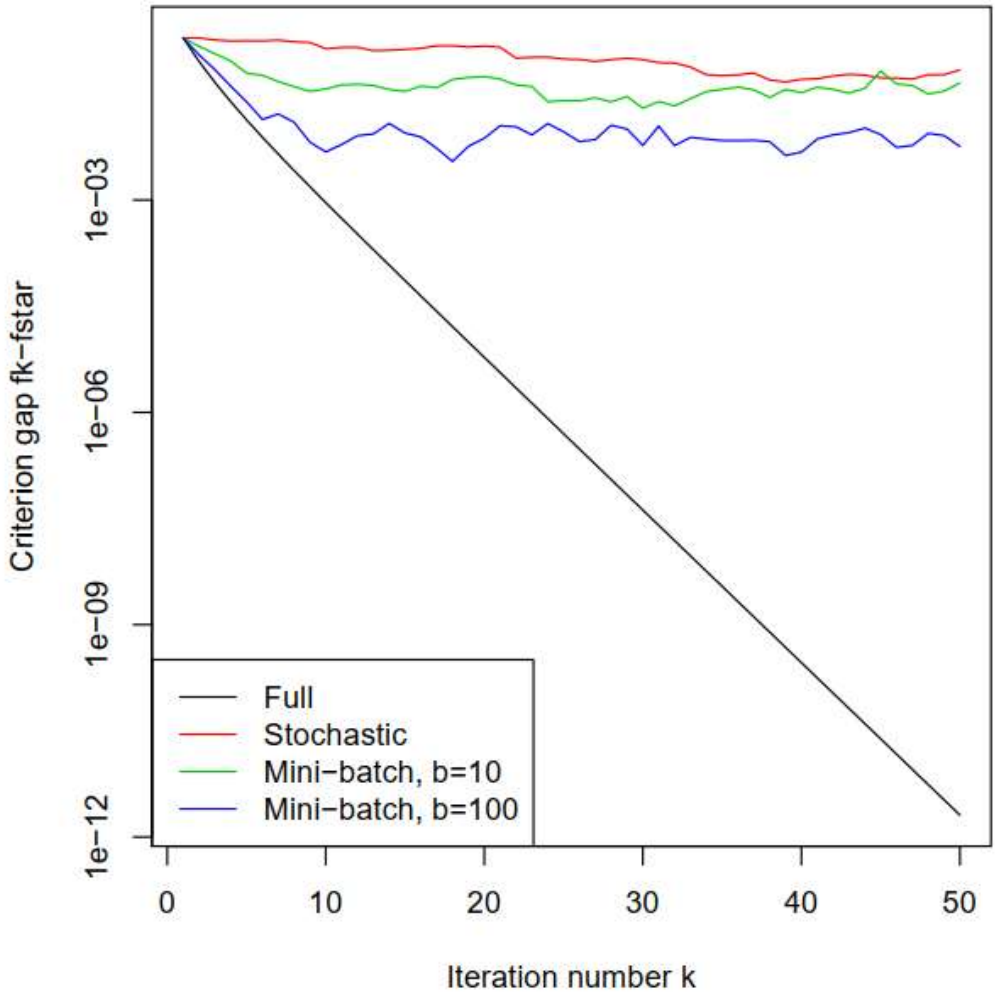
What's happening? Now let's parametrize by flops:



Ryan Tibshirani

随机梯度下降（SGD）

Finally, looking at suboptimality gap (on log scale):



Ryan Tibshirani

随机梯度下降（SGD）

Short story:

- SGD can be super effective in terms of iteration cost, memory
- But SGD is slow to converge, can't adapt to strong convexity
- And mini-batches seem to be a wash in terms of flops (though they can still be useful in practice)

Is this the end of the story for SGD?

For a while, the answer was believed to be yes. Slow convergence for strongly convex functions was believed inevitable, as Nemirovski and others established matching lower bounds ... but this was for a more general stochastic problem, where $f(x) = \int F(x, \xi)\, dP(\xi)$

New wave of "variance reduction" work shows we can modify SGD to converge much faster for finite sums (more later?)

Ryan Tibshirani

随机梯度下降（SGD）

SGD has really taken off in large-scale machine learning

- In many ML problems we don't care about optimizing to high accuracy, it doesn't pay off in terms of statistical performance
- Thus (in contrast to what classic theory says) fixed step sizes are commonly used in ML applications
- One trick is to experiment with step sizes using small fraction of training before running SGD on full data set[4]
- Momentum/acceleration, averaging, adaptive step sizes are all popular variants in practice
- SGD is especially popular in large-scale, continous, nonconvex optimization, but it is still not particular well-understood there (a big open issue is that of implicit regularization)

Ryan Tibshirani

随机梯度下降（SGD）

Suppose p is large and we wanted to fit (say) a logistic regression model to data $(x_i, y_i) \in \mathbb{R}^p \times \{0, 1\}$, $i = 1, \ldots, n$

We could solve (say) ℓ_2 regularized logistic regression:

$$\min_\beta \frac{1}{n} \sum_{i=1}^n \Big(- y_i x_i^T \beta + \log(1 + e^{x_i^T \beta}) \Big) \quad \text{subject to} \quad \|\beta\|_2 \leq t$$

We could also run gradient descent on the unregularized problem:

$$\min_\beta \frac{1}{n} \sum_{i=1}^n \Big(- y_i x_i^T \beta + \log(1 + e^{x_i^T \beta}) \Big)$$

and stop early, i.e., terminate gradient descent well-short of the global minimum

随机梯度下降（SGD）

Consider the following, for a very small constant step size ϵ:

- Start at $\beta^{(0)} = 0$, solution to regularized problem at $t = 0$
- Perform gradient descent on unregularized criterion

$$\beta^{(k)} = \beta^{(k-1)} - \epsilon \cdot \frac{1}{n} \sum_{i=1}^{n} (y_i - p_i(\beta^{(k-1)}))x_i, \quad k = 1, 2, 3, \ldots$$

  (we could equally well consider SGD)

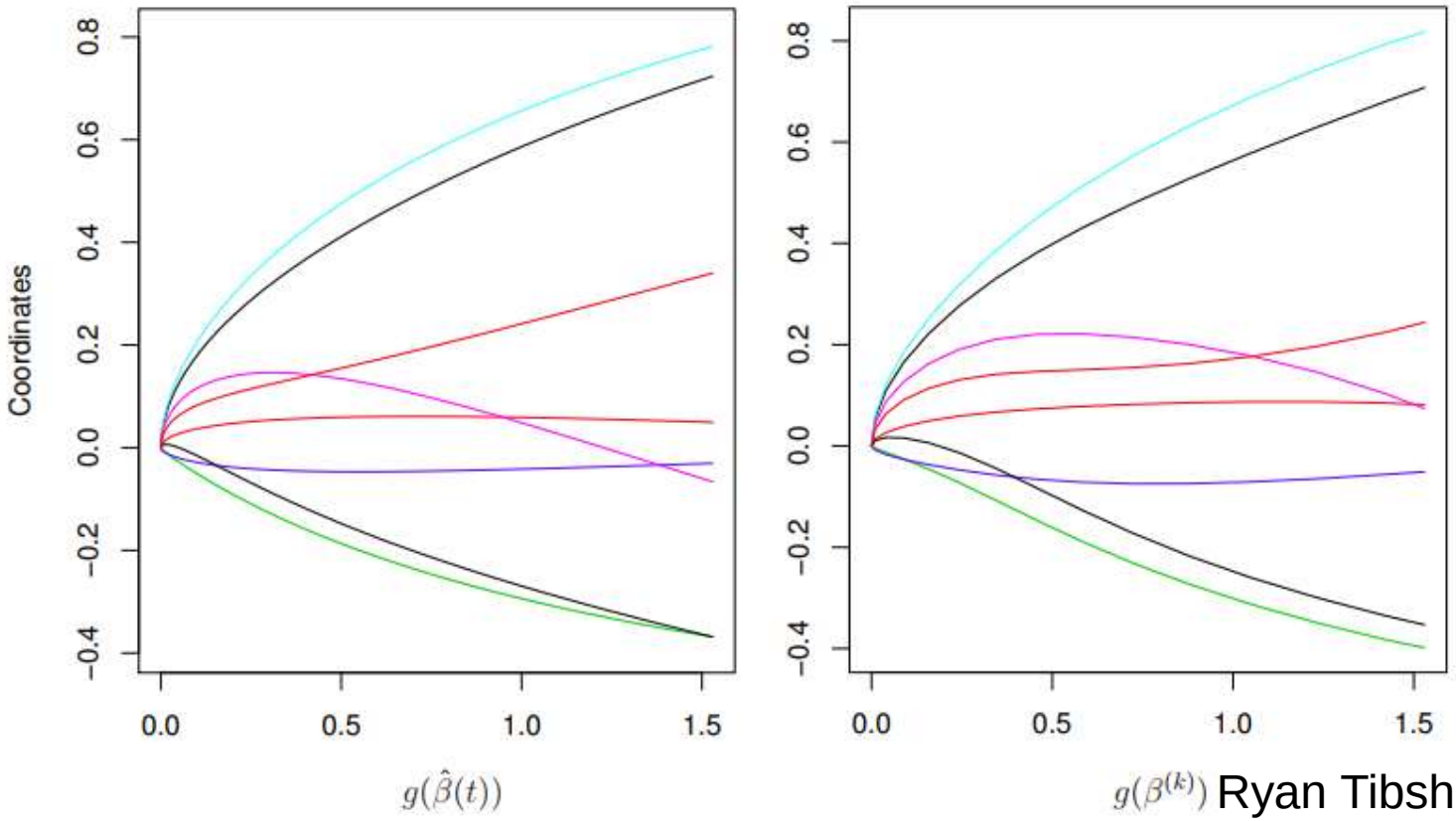
- Treat $\beta^{(k)}$ as an approximate solution to regularized problem with $t = \|\beta^{(k)}\|_2$

This is known as early stopping for gradient descent. Why do this? It's both more convenient and potentially much more efficient than using explicit regularization

随机梯度下降（SGD）

When we plot gradient descent iterates ... it resembles the solution path of the ℓ_2 regularized problem for varying t!

Logistic example with $p = 8$, solution path and grad descent path:



Ryan Tibshirani

随机梯度下降（SGD）

The intuitive connection comes from the steepest descent view of gradient descent. Let $\|\cdot\|$ and $\|\cdot\|_*$ be dual norms (e.g., ℓ_p and ℓ_q norms with $1/p + 1/q = 1$)

Steepest descent updates are $x^+ = x + t \cdot \Delta x$, where

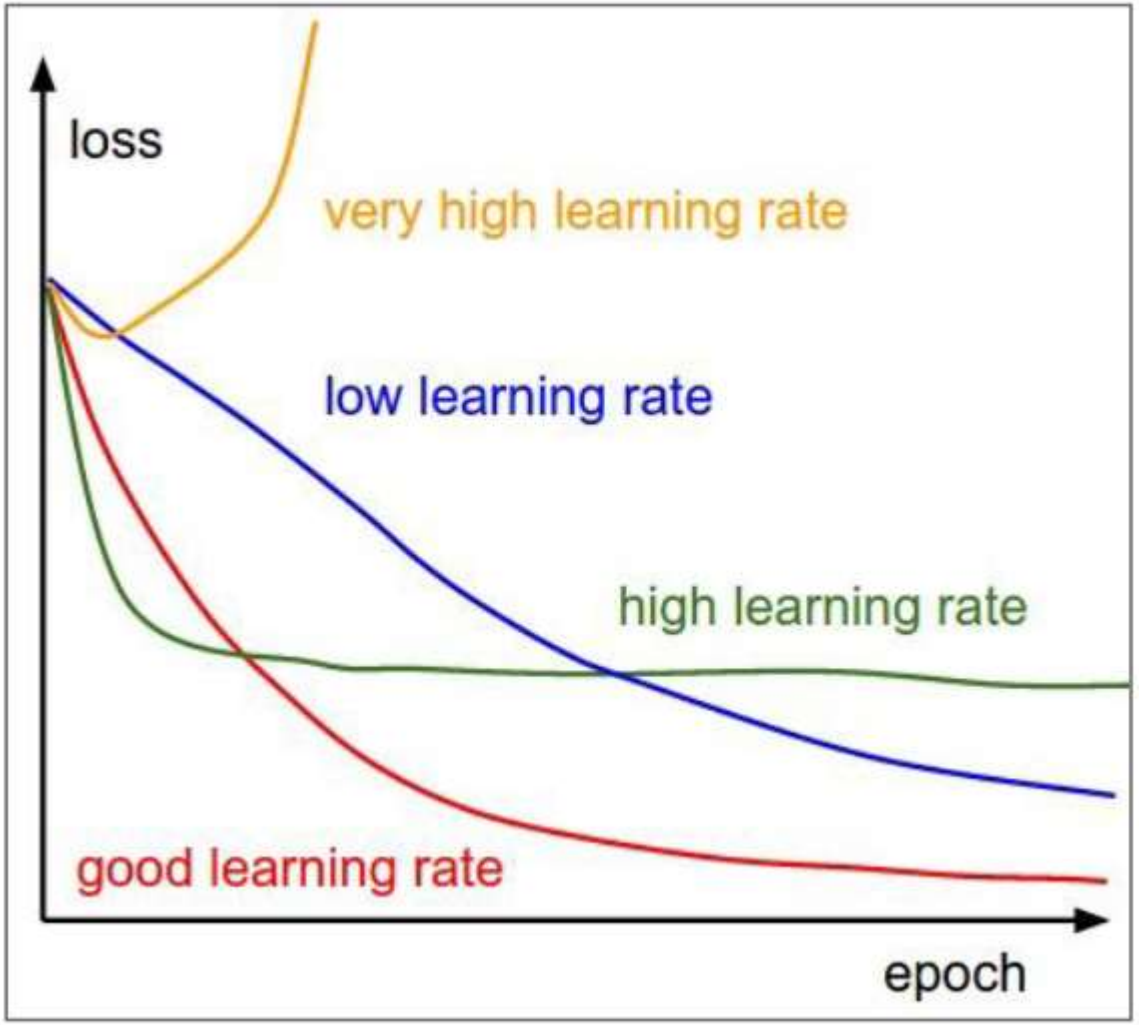
$$\Delta x = \|\nabla f(x)\|_* \cdot u$$

$$u = \underset{\|v\| \leq 1}{\operatorname{argmin}} \ \nabla f(x)^T v$$

If $p = 2$, then $\Delta x = -\nabla f(x)$, and so this is just gradient descent (check this!)

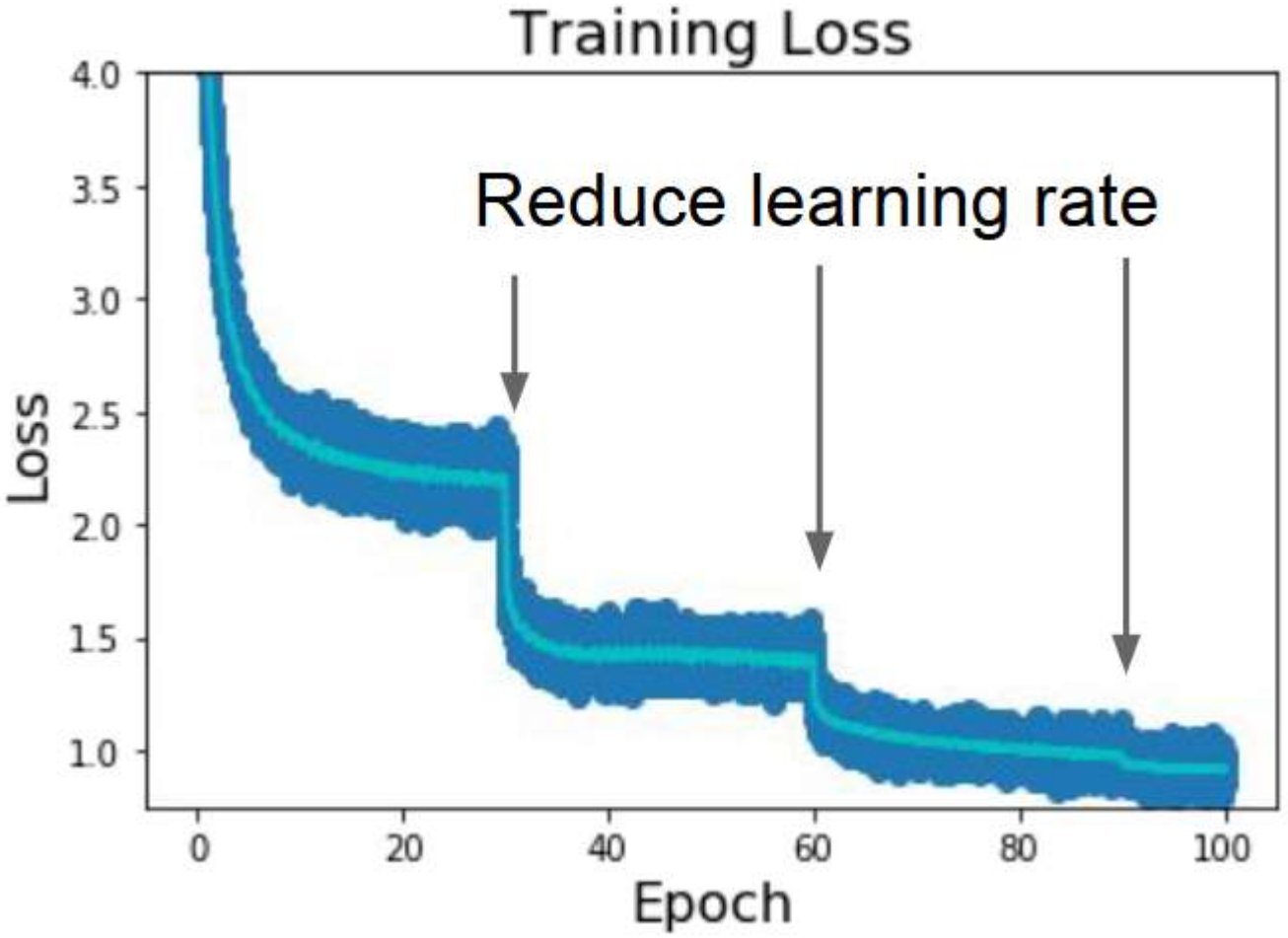
Thus at each iteration, gradient descent moves in a direction that balances decreasing f with increasing the ℓ_2 norm, same as in the regularized problem

SGD in Deep Learning



Fei-fei Li

SGD in Deep Learning



Fei-fei Li

作 业

运筹学教程，p.218

① 6.1

② 6.10(2)

③ 6.12

④ 6.16

⑤ 6.19

谢　谢！