最优化方法 Optimization

第二讲

算法概述

主要内容

- 算法概念
- 算法收敛准则

  全局收敛， 局部收敛，收敛速度

  算法二次终止性
- 算法复杂性
- 一维搜索

优化问题的方法概述

- 解析法
- 图解法与实验法
- 形式转化法
- 智能算法
- 数值迭代算法（重点）

解析法

解析法是借助微分学、变分学等数学工具通过逻辑推理与分析运算给出最优解解析式的方法。

优点：精确、简洁、直观、适用于理论分析

缺点：一般来说缺乏求解最优化问题实用性。

解析法

例： 无约束优化问题

$$\min_{x\in\mathbb{R}^n} f(x)=\frac{1}{2}x^T Ax+b^T x+c$$

其中 A 是一个 n 阶对称正定矩阵 。

解析法：由最优性条件 $0=\nabla f\left(x^*\right)=Ax^*+b$

得到唯一最优解 $x^*=-A^{-1}b$

计算量大：矩阵求逆运算量 $O(n^3)$

稳定性差：当 A 趋近于奇异矩阵(最小特征值接近0)

数据的误差将对最优解产生极大影响。

图解法与实验法

解析法与实验法 属于“手工作坊”类的方法。

优点： 操作简单、通俗易懂

缺点： 一般效率低、变量维数小。

例：求解线性规划

$$\max Z = f(x) = 6x_1 + 4x_2$$

$$s.t.\quad 2x_1 + x_2 \le 10,$$

$$x_1 + x_2 \le 8,$$

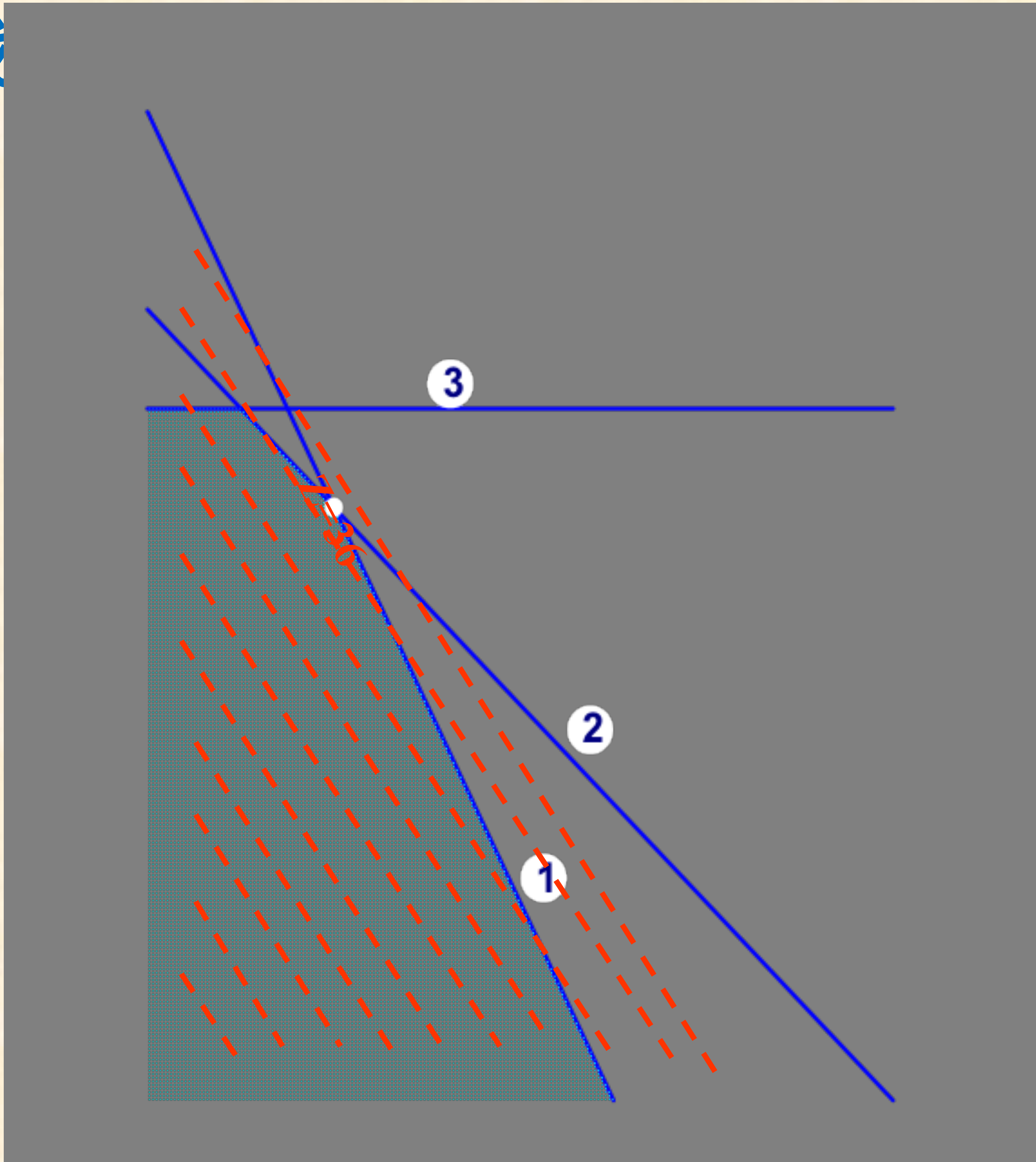
$$x_2 \le 7,$$

$$x_1, x_2 \ge 0.$$

最优解：$x_1 = 2$

$x_2 = 6$

$Z = 36$



形式转化法

形式转化法 是指利用非线性最优化问题的结构性质或者最优性条件将其转化为有别于原问题的另一类数学问题，然后对后者套用现有的方法求解。

优点： 利用问题结构信息与优化理论简化问题

缺点： 未提供具体算法，且转化需要条件。

形式转化法

例如：

约束优化问题

$$\min \ f(x)$$
$$s.t.\quad g_i(x)\ge 0 \ i=1,2,\cdots,m$$
$$h_j(x)=0 \ j=1,2,\cdots,l$$

罚函数 →

无约束优化问题

$$\min \ f(x)+\sigma p(x)$$

非连续优化问题

$$\text{e.g.,}\ \min \ x^T Vx$$
$$\text{s.t.}\quad r^T x\ge \mu,\ e^T x=1,$$
$$x\ge 0, \|\mathrm{x}\|_0 \le s$$

松弛与正则化 →

连续优化问题

$$\text{e.g.,}\ \min \ x^T Vx+\lambda e^T x$$
$$\text{s.t.}\quad r^T x\ge \mu,\ e^T x=1,$$
$$x\ge 0$$

智能算法

智能算法 是人们受自然界规律启迪并根据其原理来模拟某些自然现象而建立的一种随机搜索算法。

应用广泛的智能算法有： 遗传算法、蚁群算法、模拟退火算法、神经网络算法等。

数值迭代算法

数值迭代算法 主要利用问题的函数值信息或者（一阶与\或二阶）导数信息等按照一定的规则从当前迭代点产生下一个更好的迭代点，直到不能改进为止。这类算法得到的是数值解，为近似解。

分类：模式搜索法与梯度法

模式搜索法

模式搜索法： 根据函数值变化规律探测函数的下降方向并沿着该方向寻求更优的点。

优点： 简单、直观、无需计算目标函数的梯度

缺点： 适用于变量较少、约束简单的情形

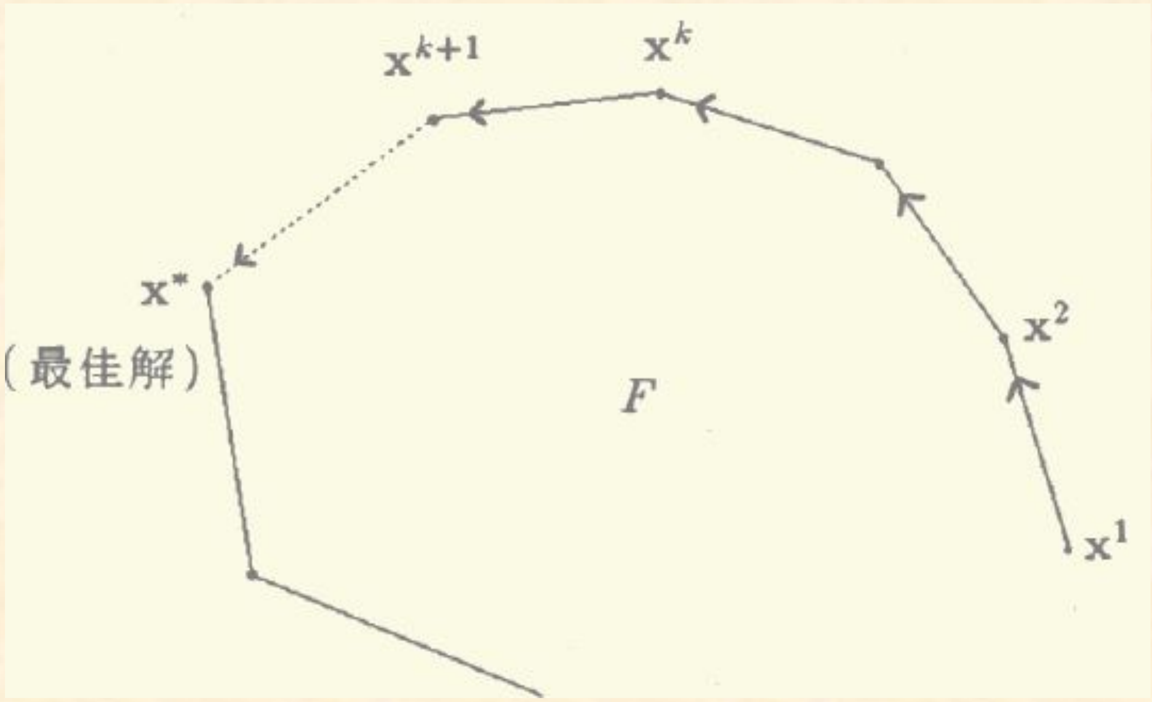
常见算法： 坐标轮转法、Hooke-Jeeves法、Powell共轭方向法、单纯形法

单纯形法求解线性规划

单纯形表：

	x_B	x_N	右端
x_B	I_m	$B^{-1}N$	$B^{-1}b$
f	0	$c_BB^{-1}N-c_N$	$c_BB^{-1}b$

搜索轨迹：



迭代算法

基本思想： 利用函数值及其微分信息等按照一定的搜索策略进行迭代。

优点： 收敛速度快且更容易建立其理论性质

缺点： 要求函数连续性且其梯度比较容易计算

两种搜索策略： 线搜索方法、信赖域方法

- 算法概念
- 算法收敛准则

  全局收敛， 局部收敛，收敛速度

  算法二次终止性
- 算法复杂性
- 一维搜索

迭代算法
算法概念

一．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。

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。

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

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

二. 算法映射

定义： 给定集合$X \subset E^n$，记其幂集（即所有子集构成的集合）为2^X，称集值映射$A: X \to 2^X$为一个算法映射(algorithm mapping).

例： 考虑标准形式的线性规划

$$(LP) \quad \min\{cx \mid Ax = b, x \geq 0\},$$

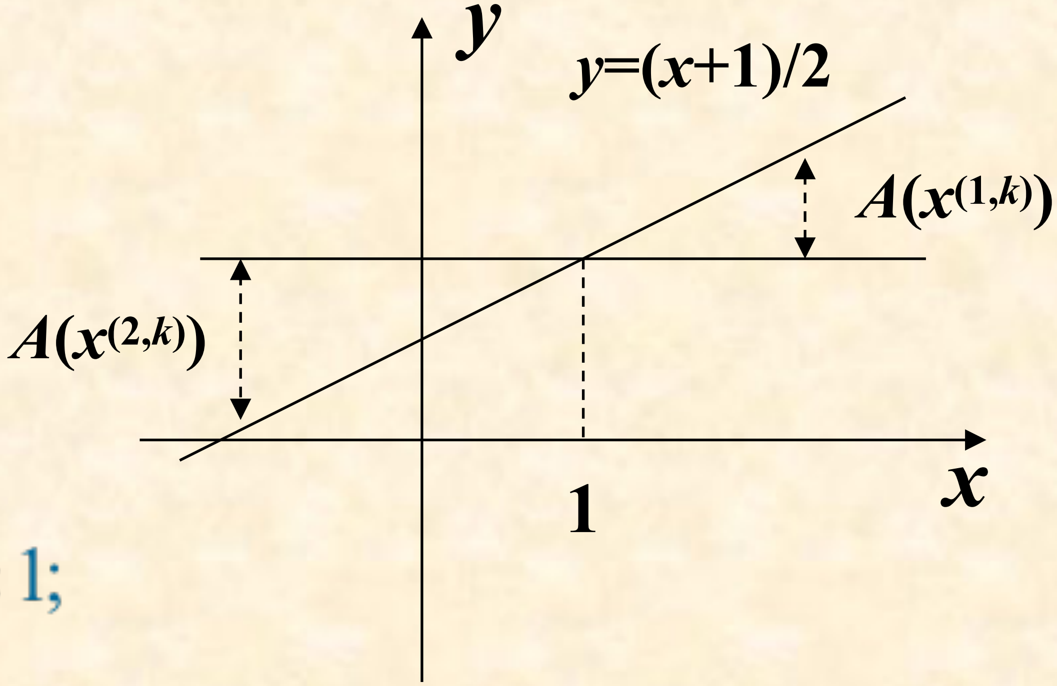
令$X = \{x \in \mathbb{R}^n \mid x$为$LP$的基本可行解$\}$，若定义算法映射

$A(x) = \{y \in \mathbb{R}^n \mid y$为$LP$的基本可行解，并且$y$和$x$的基矩阵是相邻的$\}$，

那么对于任意一个基本可行解$x^{(0)} \in X$，迭代格式$x^{(k+1)} \in A(x^k)$就生成一个相邻的基本可行解序列。

例： 考虑下列非线性规划：

$$\begin{cases} \min\ x^2 \\ s.t.\quad x \ge 1. \end{cases}$$



定义算法映射：

$$A(x)=\begin{cases} \left[1, \dfrac{1}{2}(x+1)\right] & x \ge 1; \\ \left[\dfrac{1}{2}(x+1), 1\right] & x < 1. \end{cases}$$

利用算法A可以产生不同的点列：

$$\left\{3, 2, \frac{3}{2}, \frac{5}{4}, \cdots\right\},\quad \left\{3, \frac{3}{2}, \frac{9}{8}, \frac{33}{32}, \cdots\right\},\quad \left\{3, \frac{5}{3}, \frac{7}{6}, \frac{25}{24}, \cdots\right\}$$

解集合：将满足某些最优性条件的点集定义为**解集合**.

常用的解集合：

$$S^* = \{\bar{x} \mid \nabla f(\bar{x}) = 0\}$$ 【无约束优化一阶必要条件（第1章）】

$$S^* = \{\bar{x} \mid \bar{x}\text{为}KKT\text{点}\}$$ 【约束优化一阶最优性条件（第7章）】

$$S^* = \{\bar{x} \mid \bar{x} \in S, f(\bar{x}) \le a\},$$

其中a是某个可接受的目标函数值。

值得注意的是：相邻两次迭代点的距离比较小，或者相邻两次迭代点对应的目标函数值的差比较小并不能作为解集合的一种，因此，以这类停机准则的算法并不能称之为严格意义上的算法收敛性，尽管很多工程问题的算法以此为停机准则。

全局收敛与局部收敛

收敛性： 设 S^* 为最优化问题的某种解集合，A 为某种算法。给定一个集合 Y，若对于任意的初始点 $x^{(0)} \in Y$，算法 A 所产生的迭代序列 $\{x^{(k)}\}$ 中任一收敛子列的极限都属于 S^*，则称算法 A 在 Y 上**收敛**。

全局收敛性： 若集合 Y 是任意选取的（该集合不必限定在解集合的很小邻域内）则称相应的收敛性为**全局收敛性**（global convergence）；

局部收敛性： 若集合 Y 只能取接近解集合的点集，则称相应的收敛性为**局部收敛性**（local convergence）.

收敛速率与二次终止性

□ Q–收敛速率：(Quotient)

定义：设点列$\{x_k\}$收敛到x^*，$p\in[0,+\infty), r\geq 1$,定义满足

$$\limsup_{k\to+\infty}\frac{\|x_{k+1}-x^*\|}{\|x_k-x^*\|^r}\leq p<\infty$$

的点列$\{x_k\}$Q-r 阶收敛到x^*。

- ⋆ 若$r=1$且 $0<p<1$，则称点列$\{x_k\}Q-$线性收敛到x^*
- ⋆ 若$r=1$且 $p=0$，则称点列$\{x_k\}Q-$超线性收敛到x^*
- ⋆ 若$r>1$，则点列$\{x_k\}$必定超线性收敛到x^*
- ⋆ 若r=2，则称点列$\{x_k\}$Q-2阶收敛到x^*

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

北京交通大學

收敛速率与二次终止性

□ R–收敛速率：(Root)

定义： 设点列$\{x_k\}$收敛到x^*，若存在$\kappa > 0, q \in (0,1)$使

$$\|x_k - x^*\| \le \kappa q^k$$

则称点列$\{x_k\}$R-线性收敛到x^*；若存在$\kappa > 0$和收敛到0的正数列$\{q_k\}$使

$$\|x_k - x^*\| \le \kappa \prod_{i=1}^{k} q_i$$

则称点列$\{x_k\}$R-超线性收敛到x^*.

Q-(超)线性收敛 ⟶ R-(超)线性收敛

简单分析：

$$
\begin{aligned}
\|x_k - x*\| &\le p\|x_{k-1} - x^*\| \\
&\le p^2\|x_{k-2} - x^*\| \\
&\ \vdots \\
&\le p^k\|x_0 - x^*\| \\
&=:\ \kappa\, q^k
\end{aligned}
$$

目标函数值的收敛速率

1. Q-收敛速率

2. R-收敛速率

3. O($\frac{1}{k^r}$)-收敛速率: $f(x_k)-f^* \le O\left(\frac{1}{k^r}\right)$

最优值

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).

例：　$\{a^k\}$　$0<a<1$

$$\because \lim_{k\to\infty} a^k = 0$$

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

$\therefore \{a^k\}$以收敛比a线性收敛于0。

例：　$\{a^{2^k}\}$　$0<|a|<1$

$$\because \lim_{k\to\infty} a^{2^k} = 0$$

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

$\therefore \{a^{2^k}\}$是2级收敛的。

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

$$\because \lim_{k\to\infty}\left(\frac{1}{k}\right)^k = 0$$

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

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

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

算法的二次终止性

定义：若某个算法对任意的严格凸二次函数，从任意的初始点出发，都能经有限步迭代达到其极小值点，则称该算法具有二次终止性。

用二次终止性作为判断算法优劣的原因：

- 严格凸二次函数具有某些较好的性质，因此一个好的算法应能够在有限步内达到其极小点。
- 对于一般的目标函数，若在其极小点处Hesse矩阵正定，

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

因此可以猜想，对正定二次函数好的算法，对于一般目标函数也应具有较好的性质。

算法复杂性

算法复杂性 描述算法的存储要求和运行时间要求，分为算法的空间复杂性和算法的时间复杂性。

- 利用算法需要的初等运算次数表示算法的时间复杂性。

求解实例I的算法的基本计算总次数$C(I)$是实例输入长度$d(I)$的一个函数，该函数被另一个函数$C(I)$控制，即存在一个函数$g(x)$和一个常数a，使得

$$C(I) \le a g(d(I))$$

多项式时间算法与指数时间算法

- 假设问题和解决该问题的一个算法已经给定，若给定该问题的一个实例I，存在多项式函数$g(\boldsymbol{x})$，使得

$$C(I) \le ag(d(I))$$

成立，则称该算法对实例I是多项式时间算法.
若存在$g(\boldsymbol{x})$为多项式函数且对该问题任意一个实例I，都有上式成立，则称该算法为解决该问题的多项式时间算法。

- 当$g(\boldsymbol{x})$为指数函数时，称相应的算法为指数时间算法。

多项式时间算法的优点：

- 随着问题输入规模的增加，算法的计算量（即算法复杂性）呈多项式增长.
- 一个多项式时间算法利用另一个多项式时间算法作为其“子程序”，构造一个新的复合型算法，则新算法仍是多项式时间算法。

经典结论： 单纯形方法是指数时间算法

【CPLEX】https://www.ibm.com/analytics/cplex-optimizer

椭球法、内点法是多项式时间算法

【Gurobi】http://www.gurobi.com/

单纯形算法的复杂性

$$
\begin{aligned}
\min \quad & -\sum_{i=1}^{n} 10^{n-i} x_i \\
s.t. \quad & x_i + 2\sum_{j<i} 10^{i-j} x_j \le 10^{2i-2}, i = 2,\cdots,n \\
& x_i \ge 0, \quad i = 1,2,\cdots,n
\end{aligned}
$$

上例用单纯形算法需要2^n-1次迭代

- 输入规模(input size)：表示一个实例所需要的字符串长度。
- 一般的，使用 $1+\lceil \log_2 r \rceil$ 位二进制就可以表示任意整数*r*。
- 线性规划的输入规模为：

$$L=\lceil 1+\log_2 m \rceil+\lceil 1+\log_2 n \rceil+\sum_{j=1}^{n}\{\lceil 1+\log_2 |c_j| \rceil\}$$

$$+\sum_{i=1}^{m}\sum_{j=1}^{n}\{\lceil 1+\log_2 |a_{ij}| \rceil\}+\sum_{i=1}^{m}\{\lceil 1+\log_2 |b_i| \rceil\}$$

$$\leq mn+m+n+2+\log_2 |P| \, (P\text{为}A,b,c\text{中所有非零数的乘积})$$

■ 数学模型： $\min\limits_{\alpha \geq 0} f(x_k + \alpha d_k)$

其中 d_k 为当前的搜索方向，一般为目标函数的下降方向.

函数的下降方向:

设 $x, d \in \mathbb{R}^n$,对函数$f: \mathbb{R}^n \to \mathbb{R}$,若存在$\delta$>0,使得对于任意的 $t \in (0, \delta]$,都有$f(x + td) < f(x)$,则称d为函数f 在x 点处的下降方向.

$d^T \nabla f(x) < 0$ $\Longrightarrow$ d 为f 在x 点处的下降方向

[Hint: 一阶泰勒展开]

■ 线搜索的目标：

在当前给定的搜索方向d_k上选取一个合适的迭代步长.

精确与非精确线搜索

- **精确线搜索**：如果求得的α_k，使得

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

则称该一维搜索为精确线搜索(Exact Line Search).

- **非精确线搜索**：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).

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

2.1 精确线搜索

精确线搜索

- 优化问题的模型：$\min_{x \in \mathbb{R}^n} f(x)$ 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

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

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）

精确线搜索方法的基本框架

算法2.1.1

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.

精确线搜索下降算法的收敛性

定理2.1.1 设目标函数$f:\mathbb{R}^n\rightarrow\mathbb{R}$二阶连续可微有下界，搜索方向$d_k$与$-g_k$的夹角满足$\theta_k\le\frac{\pi}{2}-\theta$,其中$\theta\in\left(0,\frac{\pi}{2}\right]$. 若算法2.2.1产生无穷迭代点列且满足$\left\|\nabla^2 f(x_k+\alpha d_k)\right\|\le M,\forall\alpha>0$,其中$M>0$为常数，则 $\lim_{k\to\infty} g_k=0$

收敛到稳定点

- 存在常数c,使得对于任意x,有$f(x)\ge$c
- 此时的d_k为下降方向$\left(d_k{}^T(-\nabla f(x_k))=d_k{}^T(-g_k)=\|d_k\|\|-g_k\|\cos(\theta_k)>0\right)$
- 理论分析中可以取$\boldsymbol{\varepsilon}=\mathbf{0}$,产生无穷点列$\{\boldsymbol{x}_k\}$,而在数值计算中,取精度参数$\boldsymbol{\varepsilon}=\mathbf{10^{-3}},\mathbf{10^{-4}},\cdots$
- $\|\nabla^2 f(x_k+\alpha d_k)\|$为矩阵的谱范数,即矩阵的最大奇异值.对于任意矩阵$\boldsymbol{A}$,有 $\|\boldsymbol{Ax}\|\le\|\boldsymbol{A}\|\ \|\boldsymbol{x}\|$

精确线搜索下降算法的收敛性

定理2. 1. 2 设目标函数$f:\mathbb{R}^n \to \mathbb{R}$连续可微有下界，梯度函数$\nabla f$在包含水平集$L(x_0)$的某邻域内一致连续. 对算法2.2.1,设搜索方向d_k与$-g_k$的夹角满足$\theta_k \le \frac{\pi}{2}-\theta$,其中$\theta \in (0, \frac{\pi}{2}]$. 若算法2.2.1不是有限步终止，则 $\lim_{k\to\infty} g_k = 0.$

- 水平集 $L(x_0) \coloneqq \{x \in \mathbb{R}^n | f(x) \le f(x_0)\}$
- 梯度函数∇f在邻域N内一致连续:对任意给定的$\varepsilon > 0$,存在$\delta > 0$使得对任意$x, y \in$ N满足$\|x-y\| \le \delta$，有$\|\nabla f(x) - \nabla f(y)\| \le \varepsilon$成立.

北京交通大學

精确线搜索下降算法的收敛性

定理2.1.3 设目标函数$f:\mathbb{R}^n\to\mathbb{R}$连续可微，算法2.2.1 产生无穷迭代点列. 若存在收敛子列$\{x_k\}_{k\in N_0}$，设极限为x^*，使得$\lim\limits_{k\in N_0,k\to\infty} d_k=d^*$,则$g(x^*)^T d^*=0$.进一步，若目标函数$f$二阶连续可微，则$\left(d^*\right)^T\nabla^2 f\left(x^*\right)d^*\geq 0.$

- 该定理中对搜索方向与目标函数的梯度不做其它要求，因而得到的结论比前两个定理弱。

北京交通大學

精确线搜索下降算法的收敛性

定理2.1.4 设搜索方向 d_k 满足 $\cos(d_k, g_k) \geq \mu > 0$.若算法2.2.1产生的点列 $\{x_k\}$ 收敛到 $f(x)$ 的极小值点 x^*, $f(x)$ 在 x^* 点附近二阶连续可微且存在 $\delta > 0$ 以及 $M > m > 0$,使得 $\forall x \in N(x^*, \delta)$, $y \in R^n$, 有 $m\|y\|^2 \leq y^T \nabla^2 f(x) y \leq M\|y\|^2$,则 $\{x_k\}$ R-线性收敛到 x^*.

精确线搜索的特点

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

- 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

在实际计算时……

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

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

非精确线搜索

How much?

- 线搜索

邻域内搜索

信赖域方法

2.2 非精确线搜索

- Armijo 步长规则
- Goldstein 步长规则
- Wolfe步长规则

Armijo步长规则

Armijo（1966）

设$\beta>0$, $\gamma\in(0,1)$, $\sigma\in(0,1)$. 取步长$\alpha_k=\beta\gamma^{m_k}$,其中$m_k$是满足下式的最小非负整数:

$$f(x_k+\beta\gamma^m d_k)\le f(x_k)+\underline{\sigma\beta\gamma^m\nabla f(x_k)^T d_k} \qquad (2.2.1)$$

下降量

当$\alpha_k<\beta$时

$$f(x_k+\beta\gamma^{m_k}d_k)\le f(x_k)+\sigma\beta\gamma^{m_k}g_k^T d_k$$

$$f(x_k+\beta\gamma^{m_k-1}d_k)> f(x_k)+\sigma\beta\gamma^{m_k-1}g_k^T d_k$$

Armijo步长

Armijo步长的存在性

$$\begin{cases} f(x_k+\beta\gamma^{m_k}d_k)=f(x_k)+\beta\gamma^{m_k}g_k{}^T d_k+o(\beta\gamma^{m_k}) \\ g_k{}^T d_k<0 \end{cases}$$

Armijo步长的两种改进 (为获取更大步长)

- 当 $\alpha_k=\beta$ 时，Calamai & More （1987）建议将 α_k 放大 $\frac{1}{\gamma}$ 倍，直至不能满足(2.2.1)式。
- Grippo 等（1986）建议将 (2.2.1)松弛为: (非单调线搜索)

$$f(x_k+\beta\gamma^m d_k)\le \max_{0\le j\le m(k)}\left\{f\left(x_{k-j}\right)\right\}+\sigma\beta\gamma^m g_k{}^T d_k$$

其中： $m(0)=0, 0\le m(k)\le \min\left\{m(k-1)+1, M\right\}$

Goldstein步长规则

Goldstein（1965）

设 $\sigma \in (0, \frac{1}{2})$. 取步长 α_k 满足下式:

下降量

$$
\begin{aligned}
f(x_k + \alpha d_k) &\le f(x_k) + \underline{\sigma\alpha\nabla f(x_k)^T d_k}, \\
f(x_k + \alpha d_k) &> f(x_k) + (1-\sigma)\alpha\nabla f(x_k)^T d_k.
\end{aligned} \tag{2.2.2}
$$

由于 α>0充分小时, 第二式必不成立, 故该规则在保证目标函数下降的前提下, 使下一迭代点尽可能远离当前迭代点.

Goldstein步长的存在性

定理2. 2. 1: 若$\varphi_k(\alpha)=f(x_k+\alpha d_k)$关于$\alpha>0$有下界，则必存在$\alpha_k>0$满足*Goldstein*步长规则.

证明：由于$g_k^T d_k<0$，故当$\alpha\to+\infty$时，

$\phi_1(\alpha)=f(x_k)+\sigma\alpha g_k^T d_k \to -\infty.$

当α>0充分小时，$\phi_1(\alpha)$在$\varphi_k(\alpha)$的上方.

由于$\varphi_k(\alpha)$关于$\alpha>0$有下界，故

$\phi_1(\alpha)$在$\varphi_k(\alpha)$在α的正半轴有交点

（见图2. 2. 1）． 类似地，

$\phi_2(\alpha)=f(x_k)+(1-\sigma)\alpha g_k^T d_k$

和$\varphi_k(\alpha)$在α的正半轴也有交点. 取$\varphi_k(\alpha$

在$\phi_1(\alpha)$与$\phi_2(\alpha)$在的α的正半轴上最

小交点$\bar{\alpha}_1$，$\bar{\alpha}_2$． 因0<$\sigma<\frac{1}{2}$， 故$\bar{\alpha}_1>\bar{\alpha}_2$. 则对

于任意$\alpha\in(\bar{\alpha}_2,\bar{\alpha}_1)$， *Goldstein*步长规则成立.

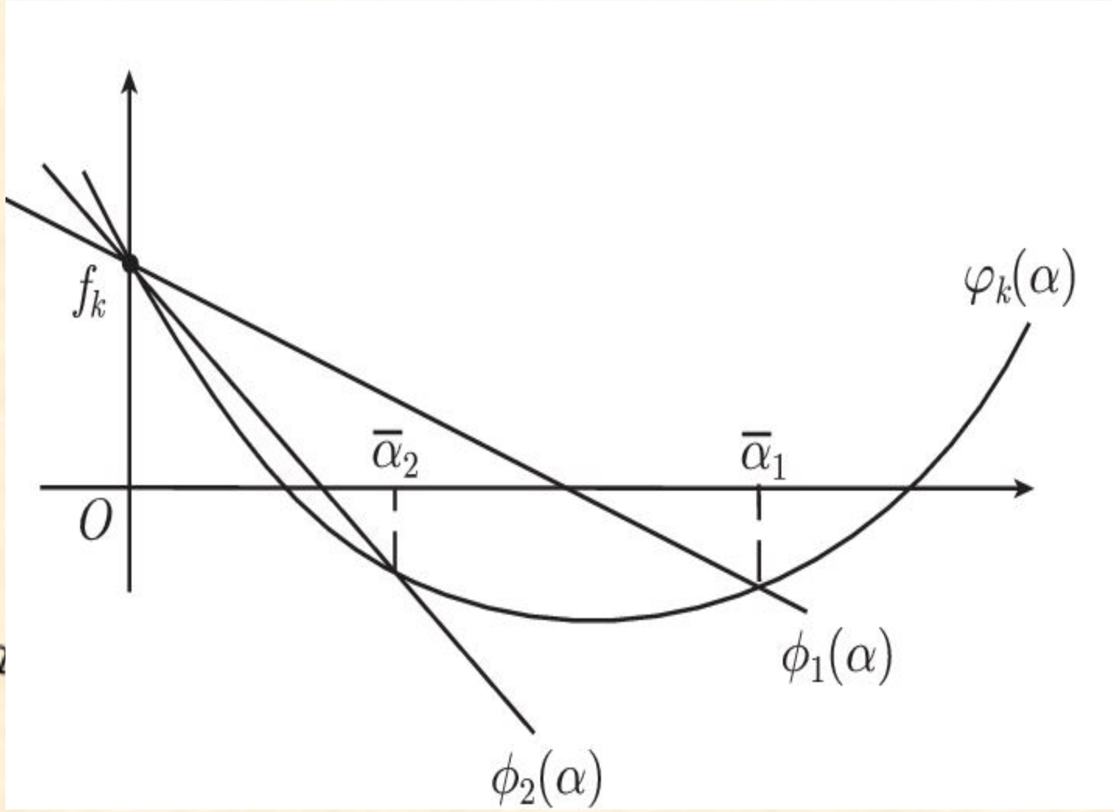


图2.2.1 Goldstein 步长规则

Wolfe步长规则

Wolfe（1968）

设$0<\sigma_1<\sigma_2<1$. 取步长α_k满足下式:

$$f(x_k+\alpha d_k)\le f(x_k)+\underline{\sigma_1\alpha g_k^T d_k},$$

$$\nabla f(x_k+\alpha d_k)^T d_k\ge \sigma_2 g_k^T d_k.$$

下降量

该规则使函数$f(x_k+\alpha d_k)$的陡度在α_k点比在$\alpha=0$点有所减缓, 从而使下一迭代点尽可能远离当前迭代点.

强Wolfe步长

Wolfe步长的加强版

$$\nabla f(x_k + \alpha d_k)^T d_k \geq \sigma_2 g_k^T d_k$$

$$\downarrow$$

$$\left|\nabla f(x_k + \alpha d_k)^T d_k\right| \leq \sigma_2 \left|g_k^T d_k\right|$$

- 当 $\sigma_2=0$ 时，即 $g_{k+1}^T d_k=0$ （正交性），即为最优步长
- Wolfe步长没有考虑 $g_{k+1}^T d_k > 0$ 的情形，因而可能导致新的迭代点进入离稳定点较远的区域；而强Wolfe步长规则可以避免这一情形的发生。

(强)Wolfe步长的存在性

定理2. 2. 2: 若$\varphi_k(\alpha)=f(x_k+\alpha d_k)$关于$\alpha>0$有下界，则必存在$\alpha_k>0$满足（强）Wolfe 步长规则.

证明　由于$\varphi_k(\alpha)=f(x_k+\alpha d_k)$关于$\alpha>0$有下界和$0<\sigma_1<1$，故射线$\phi(\alpha)=f(x_k)+\sigma_1\alpha g_k^T d_k$与曲线$\varphi_k(\alpha)$在$\alpha$的正半轴上有交点. 记最小的交点为$\alpha_k'$，则

$$f(x_k+\alpha_k' d_k)=f(x_k)+\sigma_1\alpha_k' g_k^T d_k. \tag{2.2.5}$$

显然，对任意的$\alpha\in(0,\alpha_k')$，(2.2.3)成立. 对(2.2.5)利用中值定理，存在$\alpha_k''\in(0,\alpha_k')$，使

$$f(x_k+\alpha_k' d_k)-f(x_k)=\alpha_k'\nabla f(x_k+\alpha_k'' d_k)^T d_k.$$

结合(2.2.5)，并利用$0<\sigma_1<\sigma_2<1$及$g_k^T d_k<0$得

$$\nabla f(x_k+\alpha_k'' d_k)^T d_k=\sigma_1 g_k^T d_k>\sigma_2 g_k^T d_k. \tag{2.2.6}$$

三种非精确步长的异同

相同之处：

- 定义式中均要求目标函数有一个满意的下降量且控制步长不能太小。

区别之处：

- Goldstein步长最早，但已很少用；其第二式将最优步长排除
- Armijo步长最为常见，但不适用于共轭梯度法和拟牛顿方法
- Wolfe步长特别适用于共轭梯度与拟牛顿方法，且包含最优步长的情形
- Armijo & Goldstein步长可以通过进退试探法求，而Wolfe步长可借助多项式插值法求得。

非精确线搜索的下降算法基本框架

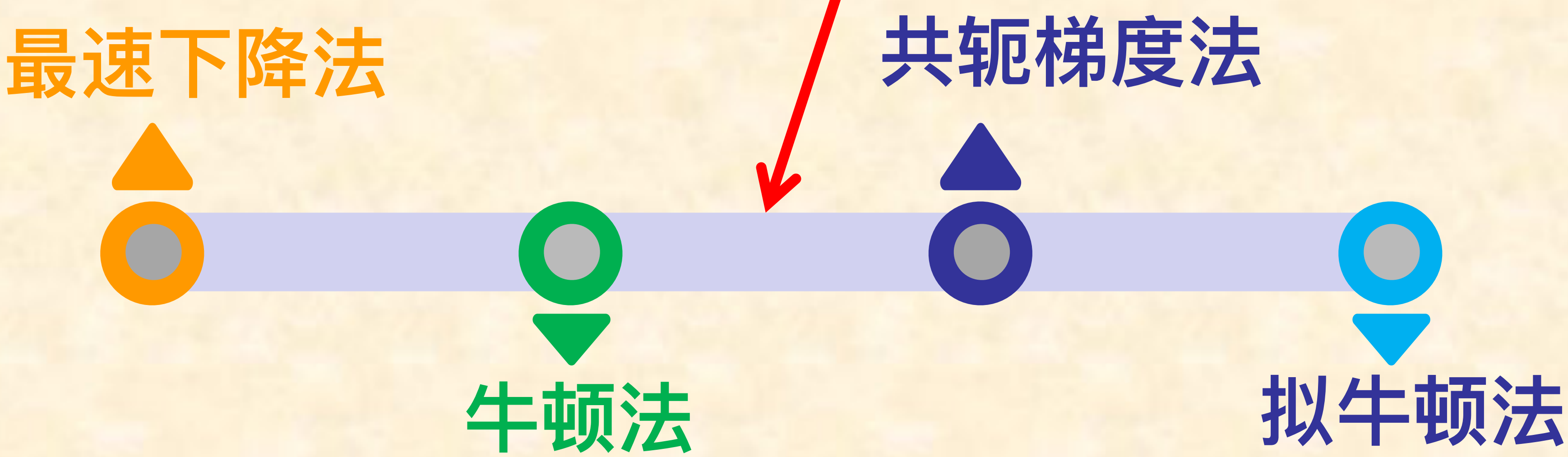
步1. 取初始点x_0及精确参数$\varepsilon \geq 0$，令$k=0$.

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. 按Armijo步长规则或者Wolfe步长规则选取步长 a_k

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



非精确线搜索方法收敛性

Armijo步长规则下的最速下降算法收敛性

定理2.2.3. 设目标函数 f 在$\mathbb{R}^n$上连续可微，若Armijo步长规则下的最速下降算法产生无穷迭代点列$\{x_k\}$,则其任一聚点为目标函数的稳定点。

非精确线搜索方法收敛性

Wolfe步长规则下的下降算法收敛性

定理2.2.4. 设目标函数 $f:\mathbb{R}^n\to\mathbb{R}$连续可微有下界且 $\nabla f(x)$在水平集$L(x_0)$上Lipschitz连续，则Wolfe步长规则下的下降算法产生的点列$\{x_k\}$满足

$$\sum_{k=1}^{\infty}\|g_k\|^2\cos^2(d_k,-g_k)<\infty.$$

非精确线搜索方法收敛性

Wolfe步长规则下的下降算法收敛性

定理2.2.5. 设目标函数 $f:\mathbb{R}^n \to \mathbb{R}$连续可微有下界且 $\nabla f(x)$在水平集$L(x_0)$上一致连续，记 $s_k = \alpha_k d_k$，则Wolfe步长规则下的下降算法产生的点列$\{x_k\}$满足

$$\lim_{k\to\infty} \frac{g_k^{\mathrm{T}} s_k}{\|s_k\|} = 0.$$

最速下降法

无约束优化问题模型

$$\min \quad f(x)$$
$$s.t. \quad x \in E^n$$

$f(x)$具有一阶连续偏导数

基本思想

每次沿负梯度方向进行搜索，即搜索方向为

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

最速下降方向

考虑函数$f(x)$在$x^{(k)}$的近似二次表达

$$f(x)\approx\varphi(x)=f\left(x^{(k)}\right)+\left\langle\nabla f\left(x^{(k)}\right),x-x^{(k)}\right\rangle+\frac{1}{2\lambda}\left\|x-x^{(k)}\right\|_2^2,a>0$$

其中$f\left(x^{(k)}\right)+\left\langle\nabla f\left(x^{(k)}\right),x-x^{(k)}\right\rangle$是$f(x)$在$x=x^{(k)}$处的一阶线性近似，

而$\frac{1}{2\lambda}\left\|x-x^{(k)}\right\|_2^2$是加权邻近项。

显然$\varphi(x)$是二次强凸函数，易求$\min\ \varphi(x)$有唯一的最优解且满足$\nabla\varphi(x)=0$ **,即:**

$$\nabla f\left(x^{(k)}\right)+\frac{1}{\lambda}\left(x-x^{(k)}\right)=0.$$

解得

$$x=x^{(k)}-\lambda\nabla f\left(x^{(k)}\right)$$

带精确线搜索的最速下降法算法步骤：

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

2.计算搜索方向$d^{(k)} = -\nabla f(x^{(k)})$;

3.若$\|d^{(k)}\| \le \varepsilon$，则停止计算；否则，从$x^{(k)}$出发，沿$d^{(k)}$进行一维搜索，求$\lambda_k$，使

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

4.令$x^{(k+1)} = x^{(k)} + \lambda_k d^{(k)}$，置$k := k+1$，返回2.

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。

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

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

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

第二次迭代

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

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

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

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

收敛性分析

最速下降法的收敛性

设$f(x)$是连续可微实函数，解集合$\Omega=\{\bar{x}\mid\nabla f(\bar{x})=0\}$，最速下降算法产生的序列$\left\{x^{(k)}\right\}$含于某个紧集，则序列$\left\{x^{(k)}\right\}$的每个聚点$\hat{x}\in\Omega$。

二次函数情形

$$f(x)=\frac{1}{2}x^TQx-b^Tx$$

其中$Q_{n\times n}$是正定矩阵，其特征值为$A=a_1\geq\cdots\geq a_n=a>0$

最速下降法表示为

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

其中$g_k=\nabla f\left(x^{(k)}\right)=Qx^{(k)}-b.$

$$\because f\left(x^{(k)}-\lambda g_k\right)=\frac{1}{2}\left(x^{(k)}-\lambda g_k\right)^T Q\left(x^{(k)}-\lambda g_k\right)-b^T\left(x^{(k)}-\lambda g_k\right)$$

$\therefore f\left(x^{(k)}-\lambda g_k\right)$在$\lambda_k=\dfrac{g_k{}^T g_k}{g_k{}^T Q g_k}$处达到极小。

$$\Rightarrow\quad x^{(k+1)}=x^{(k)}-\frac{g_k{}^T g_k}{g_k{}^T Q g_k}g_k$$

定理（二次规划收敛性）

对于强凸二次规划问题 $f(x)=\frac{1}{2}x^TQx-b^Tx$,其中Q是对称正定矩阵。最优步长下的最速下降法线性收敛，且产生的点列$\{x^{(k)}\}$收敛到唯一的最优解x^*，且满足

$$f\left(x^{(k+1)}\right)-f^*\le\left(\frac{A-a}{A+a}\right)^2\left[f\left(x^{(k)}\right)-f^*\right] \qquad (*)$$

等价于

$$\left\|x^{(k+1)}-x^*\right\|_Q^2\le\left(\frac{A-a}{A+a}\right)^2\left\|x^{(k)}-x^*\right\|_Q^2. \qquad (**)$$

其中a和A分别为Q的最小和最大特征值，$\|\cdot\|_Q^2$定义为

$$\|x\|_Q^2=\langle x,Qx\rangle$$

证明：首先，证明(*)式与(**)式等价。对于二次函数

$$f(x)=\frac{1}{2}x^TQx-b^Tx$$

有梯度$\nabla f(x)=Qx-b$. 最优解x^*满足$\nabla f(x^*)=0$，即$Qx^*=b$.

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

$$\Longleftrightarrow f(x)=f(x^*)+\frac{1}{2}\|x-x^*\|_Q^2 \Longleftrightarrow f(x)-f(x^*)=\frac{1}{2}\|x-x^*\|_Q^2$$

这证明了(*)式与(**)式等价.

下求最优步长.由$\lambda_k=\underset{\lambda\geq 0}{argmin}\ f(x^{(k)}-\lambda\nabla f(x^{(k)}))$

得λ_k满足$\langle\nabla f(x^{(k)}-\lambda_k\nabla f(x^{(k)})),\nabla f(x^{(k)})\rangle=0$

即$\langle Qx^{(k)}-b-\lambda_kQ(Qx^{(k)}-b),Qx^{(k)}-b\rangle=0$

从而$\lambda_k=\dfrac{\|Qx^{(k)}-b\|_2^2}{\langle Q(Qx^{(k)}-b),Qx^{(k)}-b\rangle}=\dfrac{\|Q(x^{(k)}-x^*)\|_2^2}{\|Q(x^{(k)}-x^*)\|_Q^2}$

下证（**）式成立.

$$\|x^{(k+1)}-x^*\|_Q^2=\|x^{(k)}-\lambda_k\nabla f(x^{(k)})-x^*\|_Q^2=\|x^{(k)}-x^*-\lambda_k(Qx^{(k)}-b)\|_Q^2$$

$$=\|x^{(k)}-x^*-\lambda_k Q(x^{(k)}-x^*)\|_Q^2=\|(I-\lambda_k Q)(x^{(k)}-x^*)\|_Q^2$$

$$=(x^{(k)}-x^*)^T(I-\lambda_k Q)Q(I-\lambda_k Q)(x^{(k)}-x^*)$$

$$=\langle x^{(k)}-x^*,Q(x^{(k)}-x^*)\rangle-2\lambda_k\langle x^{(k)}-x^*,Q^2(x^{(k)}-x^*)\rangle$$

$$+\lambda_k^2\langle x^{(k)}-x^*,Q^3(x^{(k)}-x^*)\rangle$$

$$=\|x^{(k)}-x^*\|_Q^2-2\frac{\|Q(x^{(k)}-x^*)\|_2^2}{\|Q(x^{(k)}-x^*)\|_Q^2}\|Q(x^{(k)}-x^*)\|_2^2$$

$$+\frac{\|Q(x^{(k)}-x^*)\|_2^4}{\|Q(x^{(k)}-x^*)\|_Q^4}\|Q(x^{(k)}-x^*)\|_Q^2$$

$$=\|x^{(k)}-x^*\|_Q^2-\frac{\|Q(x^{(k)}-x^*)\|_2^4}{\|Q(x^{(k)}-x^*)\|_Q^2}$$

$$=\left(1-\frac{\|Q(x^{(k)}-x^*)\|_2^4}{\|Q(x^{(k)}-x^*)\|_Q^2\|Q(x^{(k)}-x^*)\|_{Q^{-1}}^2}\right)\|x^{(k)}-x^*\|_Q^2$$

其中$\|x^{(k)}-x^*\|_Q^2=\|Q(x^{(k)}-x^*)\|_{Q^{-1}}^2$.

根据Kantorovich不等式，对于任意x有

$$\frac{(x^T x)^2}{(x^T Qx)(x^T Q^{-1}x)} \geq \frac{4Aa}{(A+a)^2},$$ (其中a和A分别为正定矩阵Q的最小和最大特征值)

从而，有

$$\|x^{(k+1)} - x^*\|_Q^2 \leq \left(1 - \frac{4Aa}{(A+a)^2}\right)\|x^{(k)} - x^*\|_Q^2 = \left(\frac{A-a}{A+a}\right)^2 \|x^{(k)} - x^*\|_Q^2$$

显然$0 < \dfrac{A-a}{A+a} < 1$,从而$x^{(k+1)} - x^*$收敛至0,即$\{x^{(k)}\}$收敛到唯一最优解x^*.

由(**)式可知，最速下降法的收敛速度与矩阵Q的条件数$r=A/a$ 有关。r越接近1，算法收敛速度越快；特别地，当r=1时，即Q的特征值都相等时，算法只需要一次迭代即可求出最优解。

例：考虑二次函数$f(x)=\frac{1}{2}x^TQx-b^Tx$,其中

$$Q=\begin{pmatrix} 0.78 & -0.02 & -0.12 & -0.14 \\ -0.02 & 0.86 & -0.04 & 0.06 \\ -0.12 & -0.04 & 0.72 & -0.08 \\ -0.14 & 0.06 & -0.08 & 0.74 \end{pmatrix}$$

$b=\left(0.76,0.08,1.12,0.68\right)^T$

其最小特征值$a=0.52$,最大特征值$A=0.94$

$$\left(\frac{A-a}{A+a}\right)^2=0.081$$

$\Rightarrow$每次迭代将使目标函数的误差减至十分之一以上.

等价于每次迭代将增加大约一位数字的精确度.

k	$f\left(x^{(k)}\right)$
0	0
1	-2.1563635
2	-2.1744062
3	-2.1746440
4	-2.1746585
5	-2.1746595
6	-2.1746595

定理（非二次情形）：

设$f(x)$存在连续二阶偏导数，$\bar{x}$是局部极小点，$Hesse$矩阵$\nabla^2 f(\bar{x})$的最小特征值$a>0$，最大特征值为A，算法产生的序列$\{x^{(k)}\}$收敛于点$\bar{x}$，则目标函数值的序列$\{f(x^{(k)})\}$以不大于

$$\left(\frac{A-a}{A+a}\right)^2$$

的收敛比线性的收敛于$f(\bar{x})$。

令$r=\dfrac{A}{a}$,则

条件数

$$\left(\frac{A-a}{A+a}\right)^2=\left(\frac{r-1}{r+1}\right)^2<1.$$

结论：

最优步长规则下的最速下降法，在相邻两次迭代中搜索方向互相正交.

锯齿现象

证明：令

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

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

为求出从$x^{(k)}$出发沿方向$d^{(k)}$的极小点，令

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

得 $$-\nabla f(x^{(k+1)})^T \nabla f(x^{(k)})=0$$

即方向$d^{(k+1)}=-\nabla f(x^{(k+1)})$与$d^{(k)}=-\nabla f(x^{(k)})$正交。

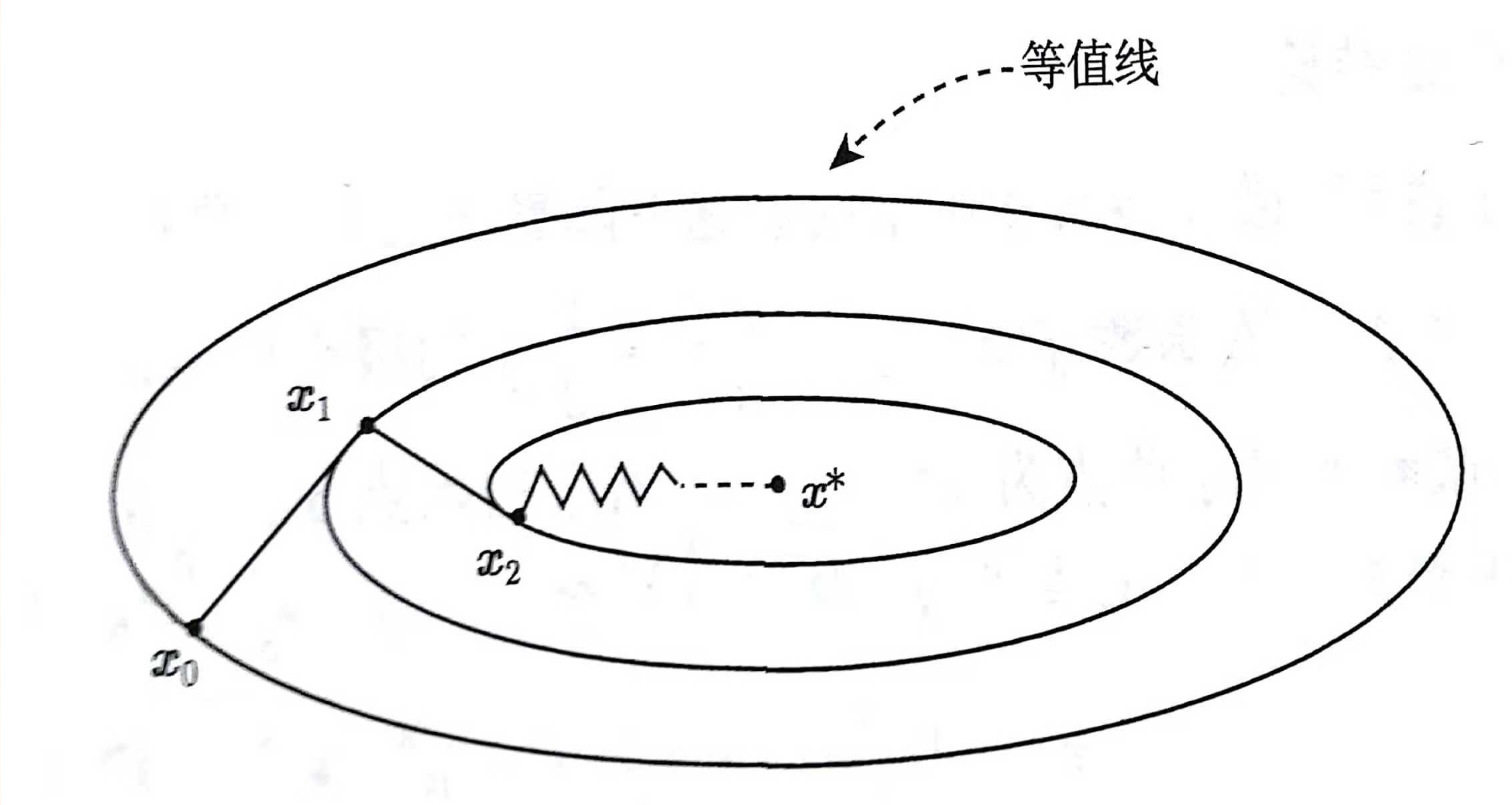


图 最速下降法的迭代过程

最速下降法的评价

最速下降方向具有计算简单，存储量小的优点。

从局部看，最速下降方向确是函数值下降最快的方向，选择这样的方向进行搜索是有利的；

从全局看，在远离极小值点处，每次迭代能够使目标函数有较大的下降，但越接近极小值点，由于锯齿现象的影响，算法收敛速率显著减慢。