Practical Optimization Algorithms
实用优化算法

徐 翔

数学科学学院
浙江大学

Feb 28, 2022

第一讲：引论

General Information

- Prerequisites(前置课程):
  - Some knowledge of linear algebra 数学分析;
  - The standard sequence of calculus 高等代数;
  - Some knowledge of numerical linear algebra (数值线性代数)
- Reference Books(参考书):
  - 袁亚湘 孙文瑜 《最优化理论与方法》 科学出版社 北京 1997, 2007.
  - W-Y. Sun and Y-X. Yuan. *Optimization Theory and Methods: Nonlinear Programming*. Springer, New York, USA., 2006.
  - J. Nocedal and S. Wright. *Numerical Optimization* Second Edition. Springer, New York, USA., 2006.
- Additional Online Recourse(线上资源):
  - http://neos-server.org/neos/
  - https://neos-server.org/neos/solvers/index.html
- Teaching Assistant 助教:

Introduction: 简介

What is Optimization?

Optimization is the minimization or maximization of a function subject to constraints on its variables. ("在某些约束条件下求目标函数最小或最大")

- It traces the roots to the calculus of variations (变分法) and the work of Euler and Lagrange.
- It is often called mathematical programming(数学规划), a somewhat confusing term coined in the 1940s, before the word "programming" became inextricably linked with computer software.

NOTATION（数学记号）

We will use the following notation in this course:

- 未知向量： x is the vector of variables, also called unknowns or parameters;
- 目标函数： $f(x)$ is the objective function, a (scalar) function of x that we want to maximize or minimize;
- 约束条件： c_i are constraint functions, which are scalar functions of x that define certain equations and inequations that the unknown vector x must satisfy.

Mathematics Formulation 数学形式

Using the notations, the optimization problem can be written as the following Standardized Formulation(标准形式):

$$
\begin{aligned}
\min \quad & f(x) && && (2.1a)\\
s.t. \quad & c_i(x) = 0, && i \in \mathcal{E} = \{1, \cdots, m_e\} && (2.1b)\\
& c_i(x) \geq 0, && i \in \mathcal{I} = \{m_e + 1, \cdots, m\} && (2.1c)
\end{aligned}
$$

Here "s.t." means "subject to", $\mathcal{E}$ and $\mathcal{I}$ are the sets of indices for equality(等式约束集) and inequality constraints(不等式约束集), respectively.

EXAMPLE: A TRANSPORTATION PROBLEM （运输问题）

A chemical company has 2 factories F_1 and F_2 and a dozen retail outlets $R_1, R_2, \cdots, R_{12}$.

- Each factory F_i can produce a_i tons of a certain chemical product each week; a_i is called the *capacity* of the plant.
- Each retail outlet R_j has a known weekly demand of b_j tons of the product.
- The cost of shipping one tone of the product from factory F_i to retail outlet R_j is c_{ij}

The problem is to determine how much of the product to ship from each factory to each outlet so as to satisfy all the requirements and minimize cost.

EXAMPLE: A TRANSPORTATION PROBLEM （运输问题）

Denote the number of tons of the product shipped from factory F_i to retail outlet R_j as x_{ij} . We can write the problem as

$$
\begin{aligned}
\min \quad & \sum_{i=1}^{2}\sum_{j=1}^{12} c_{ij}x_{ij} \\
s.t. \quad & \sum_{j=1}^{12} x_{ij} \le a_i, \qquad i = 1, 2, \\
& \sum_{i=1}^{2} x_{ij} \ge b_j, \quad j = 1, \cdots, 12, \\
& x_{ij} \ge 0, \qquad i = 1, 2, j = 1, \cdots, 12.
\end{aligned}
$$

Example: A Transportation Problem （运输问题）

Suppose there are p factories instead of 2, i.e. $F_1, \cdots, F_p$, and each factory has q outlets $R_1, R_2, \cdots, R_q$, p and q are positive integer. Then the model can be formulated as

$$\min \quad \sum_{i=1}^{p} \sum_{j=1}^{q} c_{ij} x_{ij} \tag{2.2a}$$

$$s.t. \quad \sum_{j=1}^{q} x_{ij} \leq a_i, \qquad i = 1, \cdots, p, \tag{2.2b}$$

$$\sum_{i=1}^{p} x_{ij} \geq b_j, \quad j = 1, \cdots, q, \tag{2.2c}$$

$$x_{ij} \geq 0, \qquad i = 1, \cdots, 2, j = 1, \cdots, q. \tag{2.2d}$$

Example: A Transportation Problem （运输问题）

Suppose there are volume discounts for shipping the product. For example the cost could be represented by $c_{ij}\sqrt{\delta + x_{ij}}$, where $\delta > 0$ is a small subscription fee. Then the model can be formulaes as

$$\min \quad \sum_{i=1}^{p}\sum_{j=1}^{q} c_{ij}\sqrt{\delta + x_{ij}} \tag{2.3a}$$

$$s.t. \quad \sum_{j=1}^{q} x_{ij} \le a_i, \qquad i = 1, \cdots, p, \tag{2.3b}$$

$$\sum_{i=1}^{p} x_{ij} \ge b_j, \quad j = 1, \cdots, q, \tag{2.3c}$$

$$x_{ij} \ge 0, \qquad i = 1, \cdots, 2, j = 1, \cdots, q. \tag{2.3d}$$

EXAMPLE: A TRANSPORTATION PROBLEM （运输问题）

- The problem (2.2) is known as linear programming(线性规划). Because objective and constraints are all linear function(线性函数).
- For problem (2.3), since the objective is described as anonlinear function(非线性函数). We call the problem as a nonlinear programming(非线性规划).

CLASSIFICATION (分类)

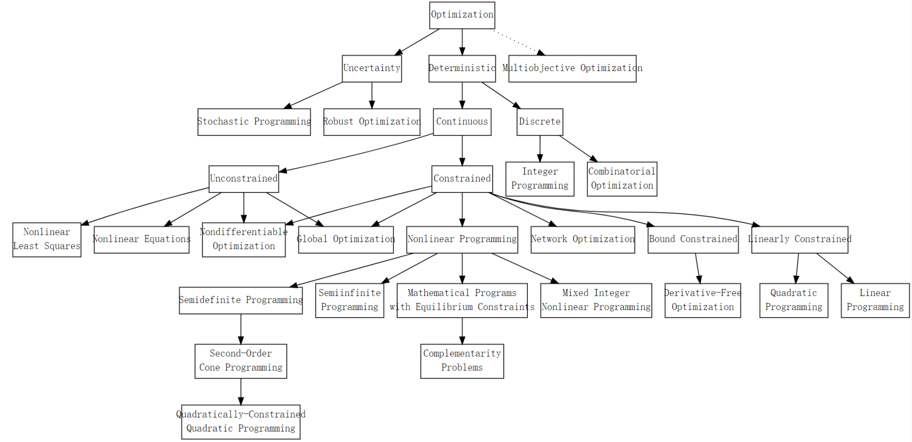
Problems with the general form (2.1) can be classified according to

- the nature of the objective function and constraints（目标函数与约束函数性质）: linear, nonlinear, convex;
- the number of variables（未知量个数）: large or small;
- the smoothness of functions（函数光滑性）: differentiable or non-differentiable;
- the aim of optimization;
- and so on.

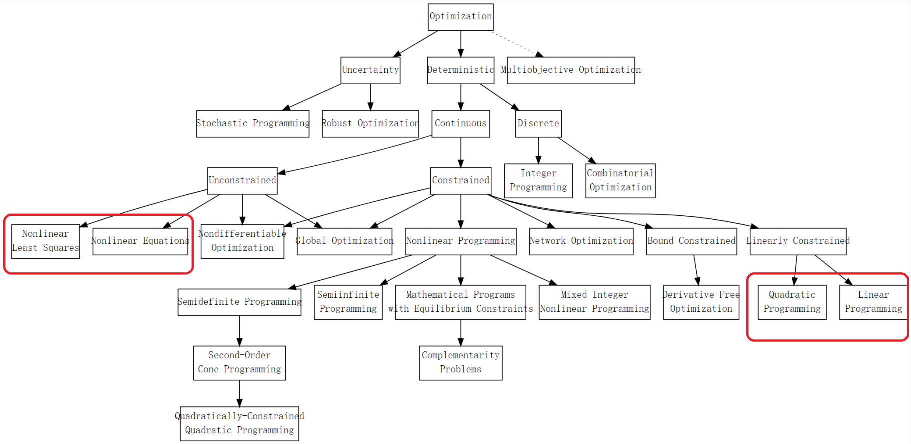
Classification (分类)

- Linear Programming vs Nonlinear Programming;
- Continuous vs Discrete Optimization;
- Smooth vs Nonsmooth optimization;
- Global vs Local Optimization;
- Stochastic vs Deterministic Optimization;
- Constrained vs Unconstrained Optimization;
- If $\mathcal{E} = \mathcal{I} = \emptyset$, (1.1-1.3) is called unconstrained optimization;
- If $\mathcal{E} \neq \emptyset$ or $\mathcal{I} \neq \emptyset$, (2.1) is called constrained optimization;
- Convex vs Nonconvex Optimization.

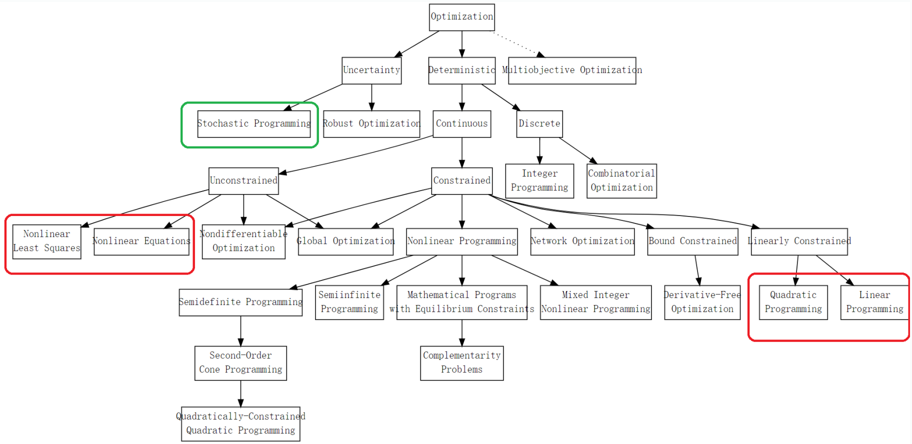
Classification (分类)



Classification (分类)



Classification (分类)



GOAL (目标)

Give a comprehensive description of the most powerful, state-of-the-art techniques for finding the local optimal solution of the continuous optimization problems (给出找到连续优化问题的局部最优解的最新方法)

- Concentrate on presenting the motivating ideas (给出想法)
- Try to make the technical details easier to follow (细节简化)
- Try to keep formal mathematical requirements to a minimum (数学形式简单)
- Proofs for the theoretical properties (理论证明)

ALGORITHM (算法)

An algorithm is a procedure or formula for solving a problem. (求解问题的一个过程)

- In mathematics and computer science, an algorithm is a step-by-step procedure for calculations. More precisely, it is an effective method expressed as a finite list of well-defined instructions for calculating a function.
- In computational mathematics, an iterative method (迭代方法) is a mathematical procedure that generates a sequence of improving approximate solutions for a class of problems. (产生一个逼近解序列)
- A specific implementation of an iterative method, including the termination criteria(停机准则), is an algorithm of the iterative method.
- An iterative method is called convergent (收敛) if the corresponding sequence converges for given initial approximations.

OPTIMIZATION ALGORITHMS (优化算法)

Generally speaking, optimization algorithms are iterative.

- They begin with an initial guess of the variable x and generate a sequence of improved estimates (called "iterates") until they terminate, hopefully at a solution.
- The strategy (策略) used move from one iterate to the next distinguishes one algorithm from another.
- Most strategies make use of the values of the objective function f, the constrained function c_i, and possibly the first and second derivatives of these functions (涉及目标函数值、约束条件及其导数).
- Some algorithms accumulate information gathered at previous iterations, while others use only local information obtained at the current point. (有些会利用前几步信息，有些仅利用当前局部信息)

The Mathematical theory of optimization is used both to characterize optimal points and to provide the basis for most algorithms. (如何刻画最优点并找到它，即找到最优点满足的方程)

Optimization Algorithms (优化算法)

Good algorithms should possess the following properties:

- **Robustness 稳健**. They should perform well on a wide variety of problems in their class, for all reasonable values of the starting point.
- **Efficiency 效率**. They should not require excessive computer time or storage.
- **Accuracy 精度**. They should be able identify a solution with precision, without being overly sensitive to errors in the data or to the arithmetic rounding errors that occur when the algorithm is implemented on a computer.

These goals may conflict . Trade off (权衡) between convergence rate(收敛速度) and storage requirements(存储量), and between robustness (稳定性)and speed(速度), and so on, are central issues in numerical optimization.

Mathematical Foundation: 数学基础

NORM (范数)

定义

映射$\|\cdot\|: R^n \rightarrow R$称为$R^n$上的半范数，当且仅当它具有如下性质：

- $\|x\| \geq 0$, for all $x \in R^n$,
- $\alpha x = |\alpha| \|x\|$, for all $\alpha \in R$, $x \in R^n$,
- $\|x+y\| \leq \|x\| + \|y\|$, for all $x, y \in R^n$.
  此外，除了以上三个性质之外，如果映射还满足
- $\|x\| = 0 \leftrightarrows x = 0.$
  则称$\|\cdot\|$为R^n上的范数。

Norm (范数)

常见的范数

设$x = (x_1, \cdots, x_n)^T \in R^n$

$$\|x\|_\infty = \max_i |x_i|, \ (l_\infty \text{ 范数}), \qquad \|x\|_1 = \sum_{i=1}^n |x_i|, \ (l_1 \text{ 范数})$$

$$\|x\|_2 = \left(\sum_{i=1}^n x_i^2\right)^{\frac{1}{2}}, \ (l_2 \text{ 范数})$$

这些都是l_p范数的特例。一般说来，对于$1 \leq < \infty$，l_p向量范数定义为

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}}, \ l_p\text{范数}$$

Norm (范数)

设$A \in R^{n\times n}$, 其诱导的矩阵范数定义为

$$\|A\| = \max_{x\neq 0}\left\{\frac{\|Ax\|}{\|x\|}\right\}$$

其中$\|x\|$是某一个向量范数。特别地，l_1范数可以诱导矩阵的列和范数

$$\|A\|_1 = \max_j\{\|a_{\cdot j}\|_1\} = \max_j \sum_{i=1}^n |a_{ij}|$$

l_∞诱导矩阵行和范数

$$\|A\|_\infty = \max_i \|a_{i\cdot}\|_1 = \max_i \sum_{j=1}^n |a_{ij}|$$

l_2诱导矩阵谱范数

$$\|A\|_2 = (\lambda_{A^TA})^{1/2}, \text{ 这里 } \lambda_{A^TA} \text{ 代表 } A^TA\text{的最大特征值}$$

Norm (范数)

很显然，

$$\|A^{-1}\| = \frac{1}{\min_{x\neq 0} \frac{\|Ax\|}{\|x\|}}.$$

此外，经常采用的矩阵范数还有Frobenius范数，定义为

$$\|A\|_F = \left(\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|^2\right)^{\frac{1}{2}} = [tr(A^TA)]^{1/2}$$

加权的Frobenius范数和加权的l_2范数可以如下定义

$$\|A\|_{M,F} = \|MAM\|_F, \quad \|A\|_{M,2} = \|MAM\|_2$$

其中M为对称正定矩阵。矩阵范数满足相容性条件：$\|AB\| \leq \|A\|\|B\|$ 诱导$p-$范数和Frobenius范数是满足相容性条件，并且还满足

$$\|AB\|_F \leq \min\{\|A\|_2\|B\|_F, \|A\|_F\|B\|_2\}$$

Norm (范数)

椭球向量范数定义

设$x \in R^n$, $A \in R^{n\times n}$是对称正定矩阵，则为

$$\|x\|_A = (x^T Ax)^{1/2}.$$

正交变换下不变的矩阵范数

设U为正交矩阵，如果 $\|UA\| = \|A\|$, 则称$\|\cdot\|$为正交不变矩阵范数。
显然谱范数和Frobenius 范数是正交不变范数

NORM (范数)

范数等价定义

设$\|\cdot\|_\alpha$和$\|\cdot\|_\beta$是R^n上任意两个范数，如果存在$\mu_1, \mu_2 > 0$, 使得

$$\mu_1\|x\|_\alpha \le \|x\|_\beta \le \mu_2\|x\|_\alpha, \text{对于任意的 } x \in R^n$$

, 则称范数$\|x\|_\alpha$ 和范数 $\|x\|_\beta$ 是等价的。

常见等价范数

$$\begin{aligned}
&\|x\|_2 \le \|x\|_1 \le \sqrt{n}\|x\|_2, \qquad && \|x\|_\infty \le \|x\|_2 \le \sqrt{n}\|x\|_\infty \\
&\|x\|_\infty \le \|x\|_1 \le n\|n\|_\infty, && \|x\|_\infty \le \|x\|_2 \le \|x\|_1, \\
&\sqrt{\lambda}\|x\|_2 \le \|x\|_A \le \sqrt{\Lambda}\|x\|_2
\end{aligned}$$

其中λ 和 Λ分别为A的最小和最大特征值。

Norm (范数)

定义：依范数收敛

设向量$\{x_k\}$序列，如果

$$\lim_{k\to\infty} \|x_k - x^*\| = 0$$

，则称序列$\{x_k\}$依范数收敛到x^*.

定义：Cauchy序列

在R^n中，序列$\{x_k\}$满足

$$\lim_{m,l\to\infty} \|x_m - x_l\| = 0$$

在R^n中，序列$\{x_k\}$收敛，当且仅当$\{x_k\}$是Cauchy序列。

Inequalities (不等式)

- Cauchy-Schwarz 不等式

$$|x^T y| \leq \|x\| \|y\|, \text{ 当且仅当 } x\text{和}y\text{线性相关时等式成立}$$

- 设A为正定矩阵，则

$$|x^T A y| \leq \|x\|_A \|y\|_A, \text{ 当且仅当 } x\text{和}y\text{线性相关时等式成立}$$

- 设A为正定矩阵，则

$$|x^T y| \leq \|x\|_A \|y\|_{A^{-1}}, \text{ 当且仅当 } x\text{和}A^{-1}y\text{线性相关时等式成立}$$

- Young不等式：假设 $p, q > 1$,且$\frac{1}{p} + \frac{1}{q} = 1$, 如果$x, y \in R$, 则

$$xy \leq \frac{x^p}{p} + \frac{y^q}{q}, \text{ 当且仅当 } x^p = y^q\text{时等式成立}$$

Inequalities (不等式)

- Holder不等式:

$$|x^T y| \leq \|x\|_p \|y\|_q \leq \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} \left(\sum_{i=1}^{n} |x_i|^q\right)^{1/q}$$

其中$p, q > 1$且满足$\frac{1}{p} + \frac{1}{q} = 1$.

- Minkowski不等式：

$$\|x + y\|_p \leq \|x\|_p + \|y\|_p$$

其中$p \geq 1$.

Inversion of Matrix 矩阵求逆

von Neumann Lemma

设$E \in R^{n\times n}$, I是单位矩阵，$\|\cdot\|$是满足$\|I\| = 1$的相容矩阵范数。

- 如果$\|E\| < 1$, 则$(I - E)$是非奇异矩阵，且

$$(I - E)^{-1} = \sum_{k=0}^{\infty} E^k, \quad \|(I - E)^{-1}\| \le \frac{1}{1 - \|E\|}$$

- 如果A非奇异，$\|A^{-1}(B - A)\| < 1$, 则B也非奇异，且

$$B^{-1} = \sum_{k=0}^{\infty} (I - A^{-1}B)^k A^{-1}, \quad \|B^{-1}\| \le \frac{\|A^{-1}\|}{1 - \|A^{-1}(B - A)\|}.$$

表明当B充分靠近一个可逆矩阵A时，B也可逆。

Inversion of Matrix 矩阵求逆

等价形式定理

设$A, B \in R^{n\times n}$, A可逆，$\|A^{-1}\| \le \alpha$, 如果$\|A-B\| \le \beta$, $\alpha\beta < 1$, 则B可逆，且

$$\|B^{-1}\| \le \frac{\alpha}{1-\alpha\beta}$$

RANK ONE UPDATE 秩一校正

Sherman-Morrison定理

设$A \in R^{n\times n}$是非奇异矩阵，$u, v \in R^n$是任意向量，若$1 + v^T A^{-1} u \neq 0$, 则A的秩一校正矩阵$A + uv^T$非奇异，且其逆矩阵可以表示为

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1} u v^T A^{-1}}{1 + v^T A^{-1} u}$$

S-M-Woodburg定理

设$A \in R^{n\times n}$是非奇异矩阵，$U, V \in R^{n\times m}$，若$I + v^* A^{-1} U$可逆, 则$A + UV^*$非奇异，且其逆矩阵可以表示为

$$(A + UV^*)^{-1} = A^{-1} - A^{-1} U (I + V^* A^{-1} U)^{-1} V^* A^{-1}$$

低秩校正行列式和特征值

- 对于秩一校正，我们有$\det(I+uv^T)=1+u^Tv$.
- 对于秩二校正，我们有

$$\det(I+u_1u_2^T+u_3u_4^T)=(1+u_1^Tu_2)(1+u_3^Tu_4)-(u_1^Tu_4)(u_2^Tu_3).$$

- 秩一校正的Frobenius范数为

$$\|A+uv^T\|_F^2=\|A\|_F^2+2v^TA^Tu+\|u\|^2\|v\|^2.$$

- 设$P=I-\frac{uv^T}{\|u\|\|v\|}$，则$P$有$n-1$个特征值为1.考虑$P^TP$的最大特征值，可知

$$\|P\|_2=\frac{v^Tu}{\|u\|\|u\|}$$

低秩校正行列式和特征值

联锁特征值定理

设A是$n \times n$ 的对称矩阵，其特征值为$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$，又设$\bar{A} = A + \sigma u u^T$，其特征值为$\bar{\lambda}_1 \geq \bar{\lambda}_2 \geq \cdots \geq \bar{\lambda}_n$. 那么

1. 若$\sigma > 0$，则
$$\bar{\lambda}_1 \geq \lambda_1 \geq \bar{\lambda}_2 \geq \lambda_2 \geq \cdots \geq \bar{\lambda}_n \geq \lambda_n.$$
2. 若$\sigma < 0$，则
$$\lambda_1 \geq \bar{\lambda}_1 \geq \lambda_2 \geq \bar{\lambda}_2 \geq \cdots \geq \lambda_n \geq \bar{\lambda}_n.$$

函数和微分：梯度与方向导数

- 连续函数 f称为连续可微，如果$\frac{\partial f}{\partial x_i}$ 存在且连续, 记作$f \in C^1(D)$。定义梯度

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

- 连续函数 f称为二次连续可微，如果$\frac{\partial^2 f}{\partial x_i \partial x_j}$ 存在且连续, 记作$f \in C^2(D)$。定义Hessian

$$[\nabla^2 f]_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j},\ 1 \leq i, j \leq n.$$

- 沿着方向$d \in R^n$的导数：

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

函数和微分：Tayor展开

- 一阶Taylor展开：对于任何$x, x+d, y \in D$, f在D上连续可微，则有

$$f(x+d) = f(x) + \int_0^1 \nabla f(x+td)^T d dt = f(x) + \int_x^{x+d} \nabla f(\xi) d\xi$$

因而，我们有

$$\begin{aligned} & f(x+d) = f(x) + \nabla f(\xi)^T d,\ \xi \in (x, x+d), \\ \text{或}\quad & f(y) = f(x) + \nabla f(x+t(y-x))^T (y-x),\ t \in (0,1) \\ \text{或}\quad & f(y) = f(x) + \nabla f(x)^T (y-x) + o(\|y-x\|). \end{aligned}$$

- 二阶Taylor展开：对于任何$x, x+d, y \in D$, f在D上二次连续可微，则有

$$\begin{aligned} & f(x+d) = f(x) + \nabla f(x)^T d + \frac{1}{2} d^T \nabla^2 f(\xi) d;\ \xi \in (x, x+d) \\ \text{或}\quad & f(x+d) = f(x) + \nabla f(x)^T d + \frac{1}{2} d^T \nabla^2 f(x) d + o(\|d\|^2) \end{aligned}$$

函数与微分：向量值函数

- 连续可微: 连续函数$F: R^n \to R^m$在$x \in R^n$称为连续可微，如果其每个分量函数f_i在x处连续可微。
- 梯度： F在x处的导数$F'(x)$叫做F在x处的Jacobi矩阵，它的转置叫做F在x处的梯度

$$F'(x) = J(x) = \nabla F(x)^T$$

$$\text{其中} \quad [F'(x)]_{ij} = [J(x)]_{ij} = \frac{\partial f_i}{\partial x_j}, \ 1 \le i \le m; 1 \le j \le n.$$

- Lipschitz连续: $G: R^n \to R^{m\times n}$ 在$x \in D \subset R^n$称为 Lipschitz连续，如果$\forall v \in D$,

$$\|G(v) - G(x)\| \le \gamma \|v - x\|$$

其中γ称为Lipschitz常数.
- 一阶Taylor展开定理： $F: R^n \to R^m$连续可微， F'Lipschitz连续，则对于任意的d, 有

$$|F(x+d) - F(x) - F'(x)d| \le \frac{\gamma}{2}\|d\|^2.$$

函数与微分：向量值函数

定理

设$F: R^n \to R^m$是连续可微函数，则对于任意的x, u, v，有

$$\|F(u) - F(v) - F'(x)(u - v)\| \le \left[\sup_{0\le t\le 1} \|F'(v + t(u - v)) - F'(x)\right]\| \|u - v\|.$$

如果F'是Lipschitz连续的，则有

$$\|F(u) - F(v) - F'(x)(u - v)\| \le \gamma\sigma(u, v)\|u - v\|,$$

$$\|F(u) - F(v) - F'(x)(u - v)\| \le \gamma\frac{\|u - x\| + \|v - x\|}{2}\|u - v\|,$$

其中 $\sigma(u, v) = \max\{\|u - x\|, \|v - x\|\}.$

如果$F'(x)$还是可逆的，则存在$\varepsilon > 0$, $0 \le \alpha \le \beta$, 使得对于任意的u, v，当$\max\{\|u - v\|, \|v - x\|\} \le \varepsilon$时，有

$$\alpha\|u - v\| \le \|F(u) - F(v)\| \le \beta\|u - v\|.$$

函数与微分：有限差分

设$F: R^n \to R^m$, 其Jacobi矩阵(梯度转置)的第(i,j)个分量可以用有限差分近似

$$a_{ij} = \frac{f_i(x+e_j) - f_i(x)}{h}, \quad \text{其中}f_i\text{是}F\text{的第}i\text{个分量}, e_j\text{是第}j\text{个单位向量}, h > 0.$$

$$A_{\cdot j} = \frac{F(x+he_j) - F(x)}{h}, \quad \text{表示}A\text{的第}j\text{列}.$$

定理：一阶逼近

如果$F: R^n \to R^m$满足连续可微, F'是Lipschitz连续，所采用的范数$\|\cdot\|$满足$\|e_j\| = 1$，则有

$$\|A_{\cdot j} - J(x)_{\cdot j}\| \le \frac{\gamma}{2}|h|.$$

如果采用l_1范数，则有

$$\|A - J(x)\|_1 \le \frac{\gamma}{2}|h|.$$

函数与微分：有限差分

定理：二阶逼近

如果$f: R^n \to R$满足连续可微, ∇fLipschitz连续，所采用的范数$\|\cdot\|$满足$\|e_j\| = 1$，

$$a_j = \frac{f(x + he_j) - f(x - he_j)}{2h}$$

则有

$$\|a_j - [\nabla f(x)]_j\| \leq \frac{\gamma}{6} h^2.$$

如果采用l_∞范数，则有

$$\|a - \nabla f(x)\|_\infty \leq \frac{\gamma}{6} h^2.$$

函数与微分：有限差分

定理：二阶导数逼近

如果$f: R^n \to R$满足连续可微, ∇fLipschitz连续，所采用的范数$\|\cdot\|$满足$\|e_j\| = 1$，

$$a_{ij} = \frac{f(x+he_i+he_j) - f(x+he_i) - f(x+he_j) + f(x)}{2h^2}$$

则有

$$\|a_{ij} - [\nabla^2 f(x)]_{ij}\| \le \frac{\gamma}{4}h.$$

如果采用l_1, l_2或Frobenius范数，则有

$$\|A - \nabla^2 f(x)\| \le \frac{\gamma}{4}hn.$$

CONVEXITY：凸集与凸函数

定义：凸集

设集合$S \subset R^n$, 称S为凸集，如果对任意的$x_1, x_2 \in S$

$$\alpha x_1 + (1-\alpha)x_2 \in S, \quad \forall \alpha \in [0,1].$$

- 超平面$H = \{x | p^T x = \alpha\}$是个凸集，其中$0 \neq p \in R^n$是法向量, $\alpha \in R$.
- 闭半空间$H^- = \{x | p^T x \leq \alpha\}$是凸集. 开半空间$\dot{H}^+ = \{x | p^T x > \alpha\}$是凸集.
- 射线$S = \{x | x + 0 + \lambda d, \lambda \geq 0, d \neq 0\}$是凸集.
- A是$m \times n$矩阵, $b \in R^n$, $\{x | Ax = b\}$是凸集.
- 有限个闭半空间的交集称为多面集, 多面集是闭凸集.
- 设S_1, S_2是R^n中的两个凸集，则
  1. $S_1 \cap S_2$是凸集
  2. $S_1 \pm S_2 = \{x_1 \pm x_2 | x_1 \in S_1, x_2 \in S_2\}$是凸集.

CONVEXITY： 凸集与凸函数

- 凸包：设$S \subset R^n$, 包含子集S的所有凸集的交集称为S的凸包, 记作conv(S).

$$\text{conv}(S) = \left\{ x \middle| x = \sum_{i=1}^{m} \alpha_i x_i, x_i \in S, \sum_{i=1}^{m} \alpha_i = 1, \alpha_i \geq 0. \right\}$$

- 锥： R^n的一个子集K称为锥, 如果它关于正的数乘运算是封闭的. 即当$x \in K$, $\lambda > 0$, $\lambda x \in K$.
- 凸锥：如果一个锥是凸集。 凸锥的充要条件是它关于加法和正的数乘运算是封闭的.
- 凸集的极值点: 设$S \subset R^n$是非空凸集, $x \in S$, 若$x = \theta x_1 + (1-\theta)x_2$, $x_1, x_2 \in S$能推出 $x = x_1 = x_2$, 则称x为凸集S的极值点. (即: x不在S中任何线段的内部.)
- 凸集的极值方向:

Convexity : 凸集与凸函数

- 凸集的方向: 设$S\subset R^n$为闭凸集, d为非零向量, 如果对每一个$x\in S$, $x+\lambda d\in S$, $\forall\lambda\geq 0$, 则称d为S的方向.
- 不同方向: 设d_1, d_2为S的两个方向，如果$d_1\neq\alpha d_2$, $\forall\alpha>0$, 则称d_1和d_2是S的两个不同方向.
- 极值方向： 如果S的方向d不能表示称该集合的两个不同方向的正的线性组合, 即如果$d=\lambda_1 d_1+\lambda_2 d_2$, $\lambda_1,\lambda_2>0$, 必然得到$d_1=\alpha d_2$, $\alpha>0$, 则称d为S的极值方向.
- 考虑多面集$S=x|Ax=b,x\geq 0$, $A\in R^{m\times n}$, $rank(A)=m$, $b\in R^m$. 不失一般性设$A=[B,N]$, x_B,x_N分别对应B和N的分量, 即 $Bx_B+Nx_N=b,\ \ x_B\geq 0,x_N\geq 0.$
  1. x是多面集S的极值点充要条件为 $x=\begin{bmatrix} x_B \\ x_N \end{bmatrix}=\begin{bmatrix} B^{-1}b \\ 0 \end{bmatrix}$.
  2. d是S的一个方向的充要条件为 $Ad=0,d\geq 0$.
  3. $\bar{d}$是S的一个极值方向的充要条件是

$$B^{-1}a_j\leq 0,\ \text{对某个}a_j\text{是}N\text{的列},\quad \text{并且}\bar{d}=\alpha d=\alpha\begin{pmatrix} B^{-1}a_j \\ e_j \end{pmatrix}$$

Convexity : 凸集与凸函数

- 凸函数：设$S \subset R^n$是凸集, $\alpha \in (0,1)$, 如果对任意的$x_1, x_2 \in S$, 有

$$f(\alpha x_1 + (1-\alpha)x_2) \le \alpha f(x_1) + (1-\alpha) f(x_2)$$

则称函数f是S上的凸函数.

- 如果当$x_1 \neq x_2$时，上式中不等式严格成立，则称f是严格凸函数.
- 如果存在某个$c > 0$, 使得对任意的$x_1, x_2 \in S$, 有

$$\alpha f(x_1) + (1-\alpha) f(x_2) \ge f(\alpha x_1 + (1-\alpha)x_2) + c\alpha \|x_1 - x_2\|^2,$$

则称f在S上是一致凸的.

- 如果$-f$是S上的凸(严格凸)函数, 则称f为S上的凹(严格凹)函数.
- 设f_i是定义在凸集S上的凸函数, $\alpha_i \ge 0$, 则$\sum_{i=1}^m \alpha_i f_i$也是定义在$S$上的凸函数.

Convexity： 凸集与凸函数

- 设$S \subset R^n$是非空开凸集, f是定义在S上的连续可微函数，则f为凸函数的充要条件

$$f(y) \geq f(x) + \nabla f(x)^T (y-x), \ \ \forall x, y \in S.$$

- f为严格凸函数的充要条件

$$f(y) > f(x) + \nabla f(x)^T (y-x), \ \ \forall x, y \in S, \ x \neq y.$$

- 如果f是定义在S上的二次可微函数，则f为凸函数的充要条件是在每一点的Hessian矩阵半正定.
- 如果f是定义在S上的二次可微函数，则f为严格凸函数的充要条件是在每一点的Hessian矩阵正定.
- f是定义在S上的凸函数, $\alpha \in R$, 则$L_\alpha = \{x | f(x) \leq \alpha, x \in S\}$是凸集.
- f在S上二次连续可微，存在$\alpha > 0$, 使得

$$u^T \nabla^2 f(x) u \geq \alpha \|u\|^2, \ \ \forall x \in L(x_0), \ u \in R^n,$$

则$L(x_0) = \{x \in S | f(x) \leq f(x_0)\}$是有界闭凸集.

Convexity： 凸集的支撑与分离

- 点到凸集的距离： 设S是非空闭凸集, $y \notin S$, 则存在唯一的点$\bar{x} \in S$, 它与y的距离最短. 进一步，$\bar{x}$与y距离最短的充要条件是$(x-\bar{x})^T(\bar{x}-y) \geq 0,\ \forall x \in S$.
- 点与凸集分离: 设$S \subset R^n$是非空闭凸集, $y \notin S$, 则存在向量$p \neq 0$和实数α，使得
$$p^T y > \alpha, p^T x \leq \alpha,\ \forall x \in S.$$
即存在超平面$H = \{x|p^T x = \alpha\}$严格分离y和S.
- Farkas 定理: 设$A \in R^{m\times n}$, $c \in R^n$. 则下列方程组有且仅有一组解:
$$Ax \leq 0, c^T x > 0, \quad \text{对某个} x \in R^n,$$
$$A^T y = c, y \geq 0, \quad \text{对某个} y \in R^m.$$

CONVEXITY：凸集的支撑与分离

- 支撑超平面定义：设$S \subset R^n$是非空凸集, $p \in R^n$, $\bar{x} \in \partial S$. 若有

  $S \subset H^+ = \{x \in S | p^T(x - \bar{x}) \geq 0\}$, 或 $S \subset H^- = \{x \in S | p^T(x - \bar{x}) \leq 0\}$

  则称$H = \{x | p^T(x - \bar{x}) = 0\}$是S在$\bar{x}$处的支撑超平面.

- 凸集在边界每个点都存在一个支撑超平面： 设$S \subset R^n$是非空凸集, $\bar{x} \in \partial S$. 那么$\bar{x}$处存在一个超平面支撑S, 即存在非零向量p，使得

$$p^T(x - \bar{x}) \leq 0, \ \forall x \in \bar{S}.$$

- 凸集外任意一点都存在一个支撑超平面: 设$S \subset R^n$是非空凸集, $\bar{x} \notin S$. 那么$\bar{x}$处存在一个超平面支撑S, 即存在非零向量p，使得

$$p^T(x - \bar{x}) \leq 0, \ \forall x \in \bar{S}.$$

Convexity：凸集的支撑与分离

- 两个凸集分离定义: 设$S_1, S_2 \subset R^n$是非空凸集，若

$$p^T x \geq \alpha,\ \forall x \in S_1 \text{ 和 } p^T x \leq \alpha,\ \forall x \in S_2,$$

则称超平面$H = \{x | p^T x = \alpha\}$分离$S_1$和$S_2$.
严格分离定义: $\quad p^T x > \alpha,\ \forall x \in S_1$ 和 $p^T x < \alpha,\ \forall x \in S_2.$
强分离定义: $\quad p^T x \geq \alpha + \varepsilon,\ \varepsilon > 0, \forall x \in S_1$ 和 $p^T x \leq \alpha,\ \forall x \in S_2.$

- 凸集分离定理: S_1, S_2是非空凸集, $S_1 \cap S_2 = \emptyset$, 则存在超平面分离S_1和S_2, 即存在非零向量p, 使得

$$p^T x_1 \leq p^T x_2,\ \forall x_1 \in \bar{S}_1,\ \forall x_2 \in \bar{S}_2.$$

- 凸集强分离定理:S_1, S_2是非空闭凸集, 且S_1有界, $S_1 \cap S_2 = \emptyset$, 则存在超平面强分离S_1和S_2, 即存在非零向量p和$\varepsilon > 0$, 使得

$$\inf\{p^T x | x \in S_1\} \geq \varepsilon + \sup\{p^T | x \in S_2\}.$$

Unconstrained Optimization
无约束优化问题的最优性条件

MATHEMATICAL FORMULATION (数学形式)

In unconstrained optimization, we minimize an objective function that depends on real variables, with no restriction at all on the values of these variables.
The mathematical formulation is

$$\min_x f(x) \tag{4.4}$$

where
$x \in \mathcal{R}^n$ is a real vector with $n \geq 1$ components
$f : \mathcal{R}^n \to \mathcal{R}$ is a continuous function.

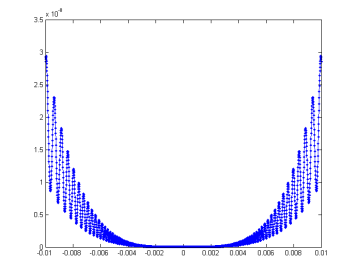
Solution Definition (解的定义)

- A point x^* is a global minimizer(全局解) if $f(x^*) \leq f(x)$ for all x;
- A point x^* is a (weak) local minimizer (局部解) if there is a neighborhood $\mathcal{N}$ of x^* such that $f(x^*) \leq f(x)$ for all $x \in \mathcal{N}$;
- A point x^* is a strict local minimizer (严格局部解) (also called a strong local minimizer) if there is a neighborhood $\mathcal{N}$ of x^* such that $f(x^*) < f(x)$ for all $x \in \mathcal{N}$ with $x \neq x^*$;
- A point x^* is an isolated local minimizer(孤立局部解) if there is a neighborhood $\mathcal{N}$ of x^* such that x^* is the only local minimizer in $\mathcal{N}$;

A COUNTER EXAMPLE(反例)

- All isolated local minimizer are strict.
- Strict minimizer are not always isolated.
- For example, consider the function

$$f(x) = \begin{cases} x^4 \cos(\frac{1}{x}) + 2x^4 & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$



- Obviously, $x = 0$ is a strict local minimizer.
- However, there are strict local minimizers at many nearby points $x_j = \frac{2}{4j\pi+3\pi}$, and we can label these points so that $x_j \to 0$ as $j \to \infty$.

RECOGNIZING A LOCAL MINIMUM (局部极小值)

- 如何判断一个点是否是局部极小值?
- 从定义出发，检查邻域内所有点.
- 当f是连续可微函数，存在更实际有效的方法定位局部极小值.
- 特别地, 如果 f 二阶连续可微, 我们可以通过$\nabla f(x^*)$和 $\nabla^2 f(x^*)$直接判断x^*是局部极小值 (或严格局部极小值).
- 数学工具是TAYLOR 定理.

RECOGNIZING A LOCAL MINIMUM

Theorem (Taylor's Theorem)

Suppose that $f : \mathcal{R}^n \to \mathcal{R}$ *is a continuously differentiable and that* $p \in \mathcal{R}^n$. *Then we have that*

$$f(x+p) = f(x) + \nabla f(x+tp)^T p \tag{4.5}$$

for some $t \in (0,1)$. *Moreover, if* f *is twice continuously differentiable, we have that*

$$\nabla f(x+p) = \nabla f(x) + \int_0^1 \nabla^2 f(x+tp) p dt \tag{4.6}$$

and that

$$f(x+p) = f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x+tp) p, \tag{4.7}$$

for some $t \in (0,1)$.

Necessary Conditions: 必要条件

Theorem (First-Order Necessary Conditions)

If x^* is a local minimizer and f is a continuously differentiable in an open neighborhood of x^*, then $\nabla f(x^*) = 0$.

一阶必要条件.

- 设x^*是局部极小值, 考虑序列$x_k = x^* - \alpha_k \nabla f(x^*)$.
- 利用泰勒展开可以得到，对于充分大的k, 有

$$0 \leq f(x_k) - f(x^*) = -\alpha_k \nabla f(\eta_k)^T \nabla f(x^*), \text{其中}\eta_k\text{是}x_k\text{和}x^*\text{凸组合}.$$

- 两边同除以α_k, 并取极限，可以得到

$$0 \leq -\|\nabla f(x^*)\|^2$$

显然只有$\nabla f(x^*) = 0$才成立.

Necessary Conditions: 必要条件

Theorem (Second-Order Necessary Conditions)

If x^* is a local minimizer and $\nabla^2 f$ exists and is continuous in an open neighborhood of x^*, then $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \geq 0$ which means positive semidefinite.

Proof and Remark

- 我们只需证明第二项.设$x_k = x^* + \alpha_k d$, d任意. 由于f二阶连续可微, 根据泰勒展开和一阶必要条件, 有

$$0 \leq f(x_k) - f(x^*) = \frac{1}{2}\alpha_k d^T \nabla^2 f(\eta_k) d, \quad \text{其中}\eta_k\text{是}x_k\text{和}x^*\text{凸组合}.$$

- 两边同除以α_k, 并取极限, 得到

$$d^T \nabla^2 f(x^*) d \geq 0, \quad \forall d \in R^n.$$

- 称$\nabla f(x^*) = 0$的x^* 为 stationary point(稳定点). 局部极小值都是稳定点

SUFFICIENT CONDITIONS: 充分条件

Theorem (Theorem (Second-Order Sufficient Conditions))

Suppose that $\nabla^2 f(x)$ is continuous in an open neighborhood of x^* and that $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite. Then x^* is a strict local minimizer of f .

Proof

- 假设x^*满足$\nabla f(x^*) = 0$, $\nabla^2 f(x^*) > 0$. 由Taylor展开, 对于任意方向$d \in R^n$,

$$f(x^* + \varepsilon d) = f(x^*) + \varepsilon \nabla f(x^*)^T d \frac{1}{2} \varepsilon^2 d^T \nabla^2 f(x^* + \theta \varepsilon d) d.$$

- 由于$\nabla^2 f$连续且在x^*是正定的, 那么在x^*领域内, $\nabla^2 f(x)$都是正定的, 从而$d^T \nabla^2 f(x^* + \theta \varepsilon d) d > 0$, 这样

$$f(x^* + \varepsilon d) > f(x^*),$$

即 x^*是严格局部极小值点.

Sufficient Conditions

Remark

- 二阶充分条件比二阶必要条件更强，可以保证局部极小值点是严格的.
- 二阶充分条件并不是必要的, 点x^*可以是严格局部极小值点，但不满足充分条件.
- *A simple example:*

$$f(x) = x^4 \tag{4.8}$$

$x^* = 0$ 是严格局部极小值点, 但是在该点的*Hessian*矩阵是零. *(*因此并不是正定*)*.

- 如果目标函数是凸函数, 那么局部极小值点x^*一定是全局极小值点. 如果目标函数还是可微的, 那么任何稳定点就是全局极小值点.
- 以上这些结果提供了非约束优化的数学基础:

寻找x^*, 使得$\nabla f(x^*) = 0$.

Structure of Optimization Algorithms
最优化算法结构

Overview of Algorithms: 算法概论

最优化方法通常采用迭代方法求解，其基本思想是

- 给定一个初始点$x_0 \in R^n$
  - 通常如果有一些关于最优解的先验信息，可以给出一个很好的x_0
  - 或者有一种系统选择初值的方法，或者任意选.
- 按照某种迭代规则产生一个点列x_k
- 如果点列$\{x_k\}$是有限的, 其最后一个点就是最优化模型问题的解.
  如果点列$\{x_k\}$是无限的, 它有极限点，并且为最优化模型问题的解.
- 按照某种终止条件, 停止迭代.
  - 要么无法再改进结果,
  - 要么已经接近精确解到足够精度要求.

Two Strategies: 两种策略

- 如何设计迭代规则, 即决定怎么从一个点x_k移动到下一个点x_{k+1}?
- 通常需要用到f在x_k处的信息, 甚至更早点x_{k-1}, x_{k-2}等处的信息
- 通常计算这些信息计算量会比较大, 因此我们希望尽可能少用调用f
- 如果找不到单调递减的算法, 我们可以要求f可以每隔m次迭代再下降, 即$f(x_k) \leq f(x_{k-m})$.
- 总的说来，一般有两种策略设计迭代格式，分别是线搜索(Line Search)和信赖域(Trust Region).

Convergence: 收敛速度

收敛速度是衡量最优化方法有效性的一个重要指标.

- 一种是Q-α阶收敛速度.
  设$\{x_k\}$在某种范数意义下收敛，即$\lim_{k\to\infty}\|x_k - x^*\| = 0$. 如果存在实数$\alpha > 0$以及与$k$无关的常数$q > 0$，使得

$$\lim_{k\to\infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^\alpha} = q$$

- 当$\alpha = 1, q > 0$, Q-线性收敛.
- 当$1 < \alpha < 2, q > 0$或$\alpha = 1, q = 0$时，Q-超线性收敛.
- 当$\alpha = 2$时，Q-二阶收敛.

- 另一种叫R-收敛速度(根收敛速度). 设

$$R_p = \begin{cases} \lim\sup\limits_{k\to\infty} \|x_k - x^*\|^{1/k}, & p = 1, \\ \lim\sup\limits_{k\to\infty} \|x_k - x^*\|^{1/p^k}, & p > 1. \end{cases}$$

- 如果$0 < R_1 < 1$，称为R-线性收敛
- 如果$R_1 = 1$，称为R-次线性收敛
- 如果$R_1 = 0$，称为R-超线性收敛
- 如果$0 < R_2 < 1$，称为R-平方收敛
- 如果$R_2 \geq 1$，称为R-次平方收敛
- 如果$R_2 = 0$，称为R-超平方收敛

Convergence: 收敛速度

Theorem (超线性收敛特征)

如果序列$\{x_k\}$ Q-超线性收敛到x^*, 那么

$$\lim_{k\to\infty}\frac{\|x_{k+1}-x_k\|}{\|x_k-x^*\|}=1.$$

反之一般不成立.

构造反例证明反之不成立. 定义序列$\{x_k\}$为

$$x_{2i-1}=(i!)^{-1}, x_{2i}=2x_{2i-1}$$

显然 $x^*=0$, 但是

$$\frac{\|x_{k+1}-x_k\|}{\|x_k-x^*\|}=\begin{cases}1, & k=2i-1,\\ 1-\frac{1}{i+1}, & k=2i.\end{cases}$$

这样$\{x_k\}$并不是超线性收敛到x^*.

Proof 对于$k>0$,

$$\frac{\|x_{k+1}-x^*\|}{\|x_k-x^*\|}\geq\left|\frac{\|x_{k+1}-x_k\|}{\|x_k-x^*\|}-\frac{\|x_k-x_*\|}{\|x_k-x^*\|}\right|$$

根据超线性收敛的定义，可知左侧极限为0, 因此可得到结论.

- 这个定理表明 $\|x_{k+1}-x_k\|$ 可以来代替 $\|x_k-x^*\|$ 给出终止判断.
- 实际使用 $|f(x_k)-f(x^*)|\leq\varepsilon$ 或 $|x_k-x^*|\leq\varepsilon$ 是不实用的.
- 有时$|x_{k+1}-x_k|<\varepsilon$, 而$|f(x_{k+1})-f(x_k)|$仍然比较大, 或反之亦然.
- Himmeblau提出的相对误差小.

Thanks For Your Attention