北京交通大學

2020-2021学年第一学期全校研究生公共课

最优化方法 I

Optimization Methods I

主讲教师：陈丙振

chenbingzhen6026@163.com

助教: 曲文涛 18118018@bjtu.edu.cn

成绩评定

成绩评定：

总成绩=作业(30%)+小论文(20%)+期末成绩(50%)

中文参考书

[1] 袁亚湘, 孙文瑜. 最优化理论与方法[M]. 科学出版社, 1997.

[2] 陈宝林. 最优化理论与算法[M]. 清华大学出版社, 2005.

[3] 李航. 统计学习方法[M]. 清华大学出版社, 2012.

[4] 周志华. 机器学习[M]. 清华大学出版社, 2016.

[5] 王宜举, 修乃华. 非线性最优化理论与方法（第三版）[M].
科学出版社，2019.

英文参考书

- Dimitri P. **Bertsekas**, Angelia Nedic, and Asuman E. Ozdaglar, **Convex Analysis and Optimization**, Athena Scientific, Belmont; Massachusetts, 2003 (http://web.mit.edu/dimitrib/www/home.html)
- Dimitri P. **Bertsekas**, **Convex Optimization Theory**, MIT, 2009
- Jorge **Nocedal,** Stephen Wright, **Numerical Optimization**, Springer, 2006 (http://users.iems.northwestern.edu/~nocedal/)
- Stephen **Boyd**, Lieven Vandenberghe, **Convex Optimization**, Cambridge University Press, 2004 (https://stanford.edu/~boyd/)
- R. T. **Rockafellar** and R. J-B. Wets, **Variational Analysis**, Springer, 1997
- R. T. **Rockafellar**, **Convex Analysis**, Princeton University Press, 1970

第一讲

- 最优化方法简史
- 无约束最优化理论

□ 最优化方法简史

发展简介

- 运筹学在国外

英国称为 *Operational Research*

美国称为 *Operations Research*

- ✓ 起源于二战期间的军事问题，如雷达的设置、运输船队的护航舰队的规模、反潜作战中深水炸弹的深度、飞机出击队型、军事物资的存储等。
- ✓ 二战后运筹学应用于经济管理领域（LP、计算机）

1948年英国首先成立运筹学会；1952年美国成立。

1952年，Morse 和 Kimball出版《运筹学方法》

1959年成立国际运筹学联合会

运筹学在国内

- 1956年成立运筹学小组
- 1958年提出运输问题的图上作业法
- 1962年提出中国邮路问题
- 1964年华罗庚推广统筹方法
- 我国于1982年加入国际运筹学联合会，并于1999年8月组织了第15届大会

最优化问题

Q： 什么是最优化?

A： 根据国际数学优化学会定义,最优化是指在一定约束条件下极大化或极小化某一目标函数的问题，其变量可能是连续或离散或随机的.

A： 通俗解释，在所有可能中挑选最好的.

最优化问题无处不在！

- 早在18世纪，著名数学家欧拉就曾说：宇宙万物无不与最小化或最大化的原理有关系.
- 可以说，最优化的原理渗入到社会发展的各个方面，甚至在我们的日常生活里也有各种各样的最优化问题.

最优化问题无处不在！

- 经济金融： 最大利润、最小风险
- 交通运输： 列车运行图、物流
- 信息科学： 数据挖掘、图像处理
- 生命科学： DNA 序列、蛋白质折叠
- 工程力学： 最大载重、结构最优
- 军事国防： 摆兵布阵、后勤保障

最优化的起源

- 中国古代优化思想--田忌赛马(公元前340年)

齐威王		田忌	齐威王		田忌
A	>	B	A	>	F
C	>	D	C	<	B
E	>	F	E	<	D
3	:	0	1	:	2

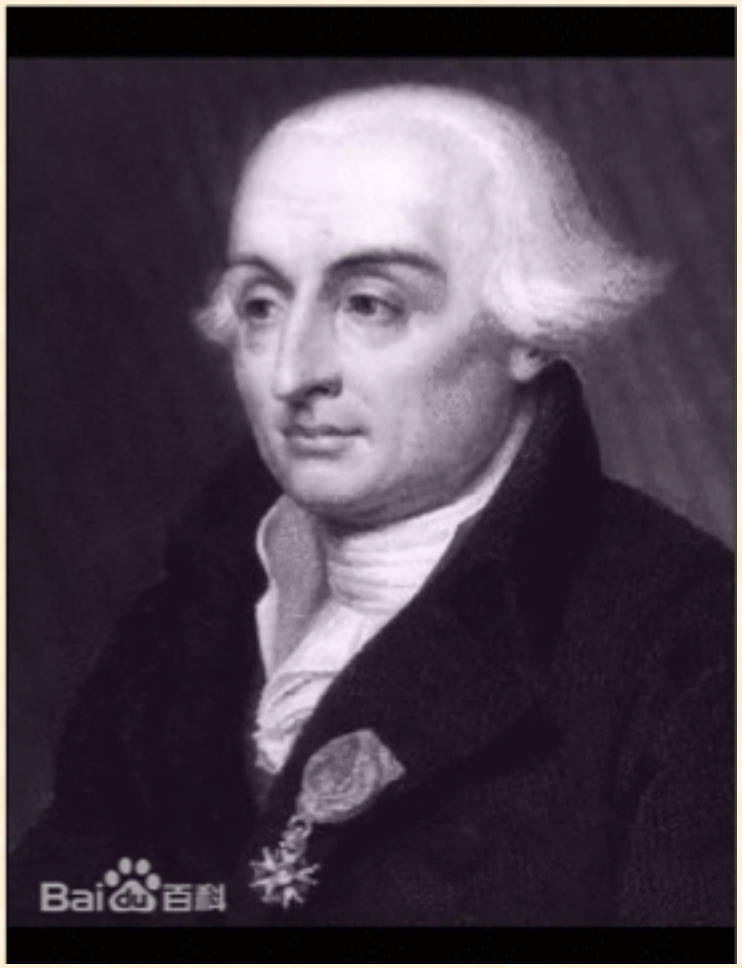
反败为胜！

最优化的起源

- 追溯到18世纪L.Euler，J.L.Lagrange等对与力学相关的极值问题或者变分问题统一处理方法的研究。

莱昂哈德·欧拉（Leonhard Euler）

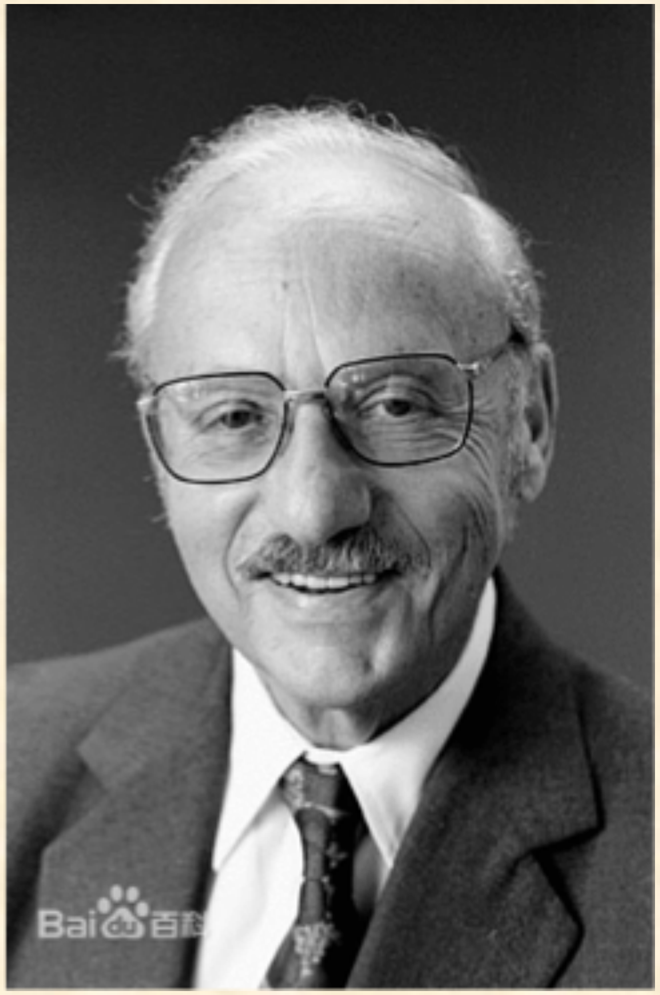
1707.4.15～1783.9.18



1736.1.25～1813.4.10

最优化的起源

- 线性规划与单纯形法—George Dantzig 1947



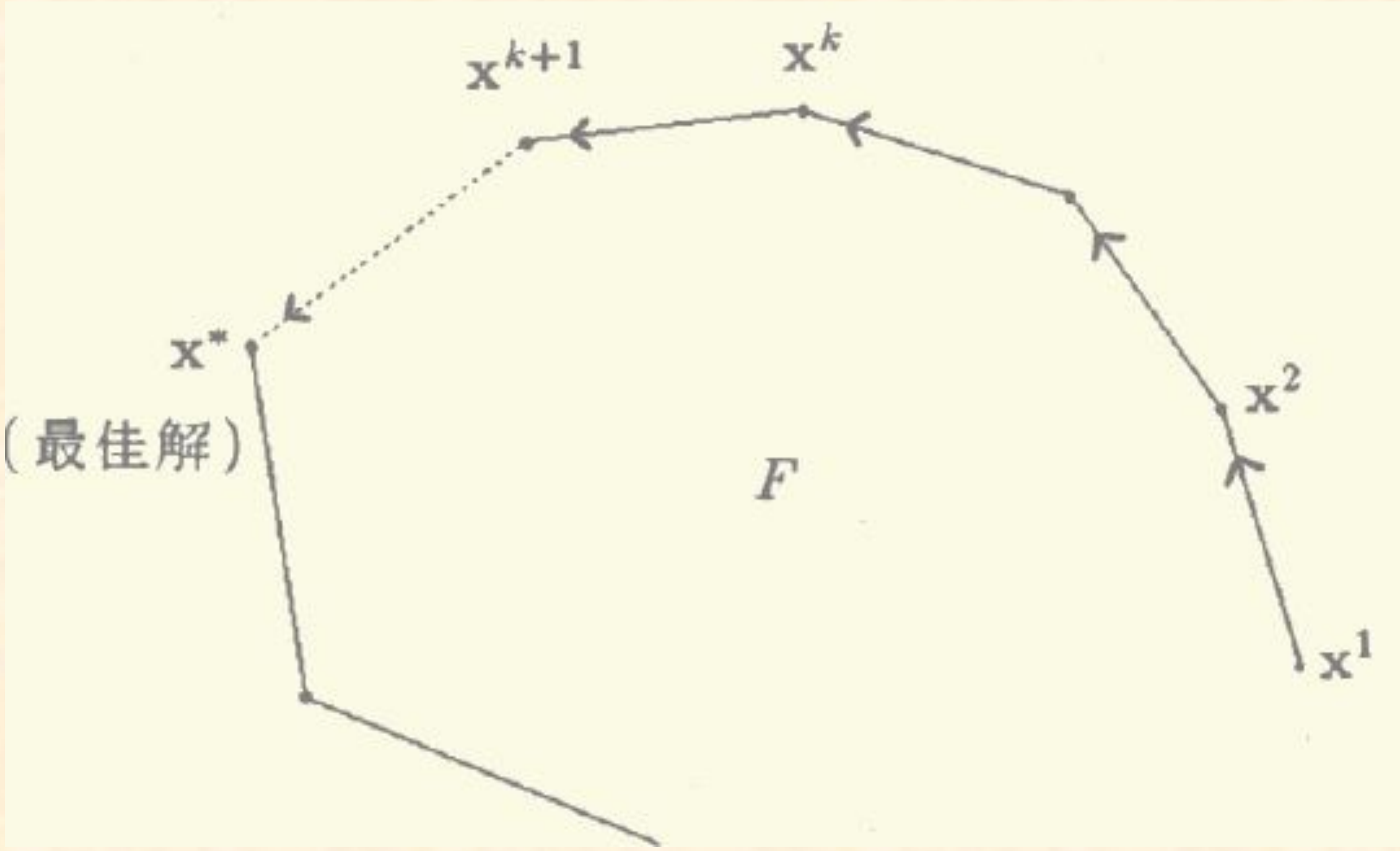
1914.11.8～2005.5.13

- 线性规划之父
- 师从著名统计学家J. Neyman

- 一个人的潜能是难以预料的，成功的障碍往往来自于心理上的畏难情绪；一定要相信自己，保持积极的态度。

北京交通大學

□ 线性规划与单纯形法—George Dantzig 1947



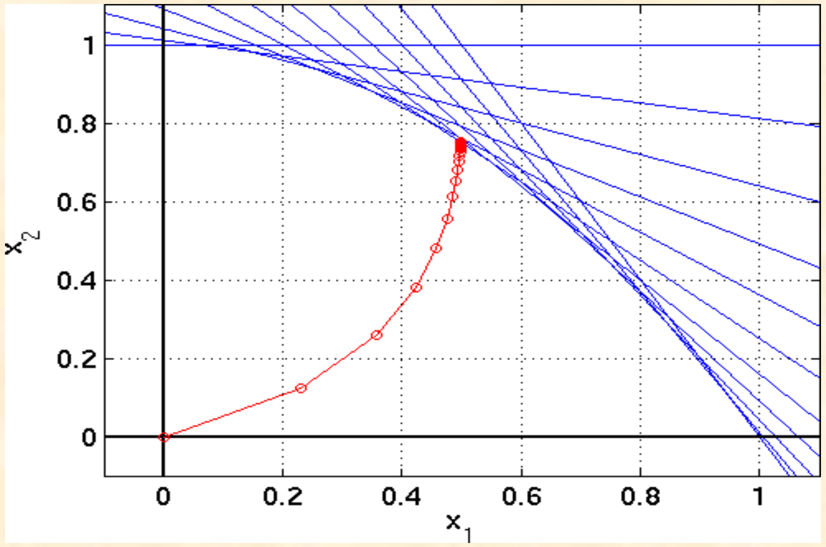
最优化的发展

- Dantzig，Fulkerson和Johnson在1950年研究旅行商问题时提出了线性整数规划问题。
- 随后，Gomory提出的割平面方法则奠定了现代整数规划算法的基础。
- 1951年Kuhn和Tucker提出了约束最优化问题必要条件，后称为Karush-Kuhn-Tucker (KKT)条件，标志着现代非线性规划理论研究的开端。

最优化的发展

- **1970年，Victor Klee & George Minty给出实例证明了单纯形方法不是多项式时间的，而是指数级** $O(2^n)$

- 椭球法---L.G.Khachian 1979 $O(n^6L^2)$
- 内点法---Karmarkar 1984 $O(n^{3.5}L^2)$
- Nesterov和Nemirovski等推广到凸优化问题(90年代）.



最优化发展与现状

- 向量优化
- 互补与均衡问题
- 组合优化
- 随机优化
- 半定规划
- 鲁棒优化
- 稀疏优化
- 统计优化
- 张量与多项式优化
- 非光滑优化

最优化问题分类

- CONSTRAINED AND UNCONSTRAINED OPTIMIZATION
- GLOBAL AND LOCAL OPTIMIZATION
- STOCHASTIC AND DETERMINISTIC OPTIMIZATION
- CONTINUOUS VERSUS DISCRETE OPTIMIZATION
- ……

模型与分类

一般模型:

目标函数

$$\min \ f(x)$$

$$s.t.\ x \in \Omega$$

决策变量

可行域or决策集

无约束优化

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

约束优化

约束函数

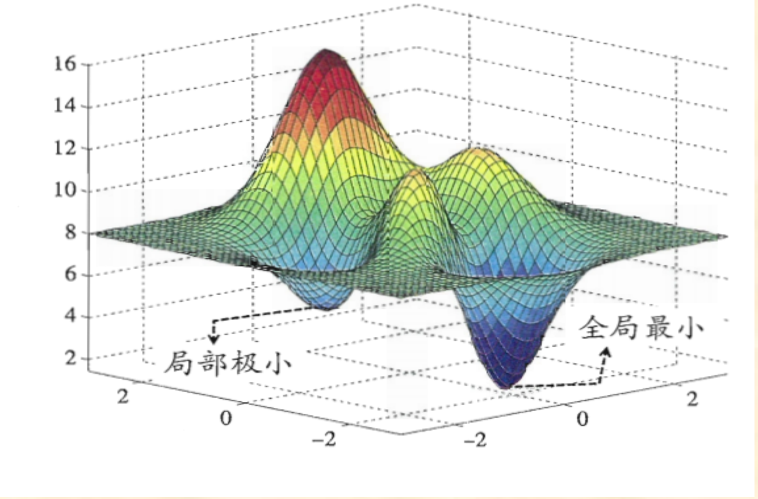
$$\begin{aligned} \min \quad & f(x) \\ s.t. \quad & c_i(x) \geq 0, \quad i \in I \\ & c_j(x) = 0, \quad j \in \varepsilon \end{aligned} \qquad (1)$$

北京交通大學

最优解概念

定义 1：设$f(x)$为目标函数，Ω为可行域，$x^0 \in \Omega$，若对$\forall x \in \Omega$, 有$f(x) \geq f(x^0)$，则x^0称为极小化问题$\min \{f(x), x \in \Omega\}$ 的（全局）最优解。

定义 2：设$f(x)$为目标函数，Ω为可行域，若存在x^0的ε邻域$N_\varepsilon(x^0) = \{x| \parallel x - x^0 \parallel < \varepsilon, \varepsilon > 0\}$,使得对$\forall x \in \Omega \cap N_\varepsilon(x^0)$, 有$f(x) \geq f(x^0)$，则$x^0$称为极小化问题$\min \{f(x), x \in \Omega\}$的局部最优解。若此时不等式严格成立，则称$x$为严格局部最优解。



最优化主要研究内容

理论:

- 一阶/二阶最优性条件
- 对偶理论
- 鞍点问题
- 灵敏度分析
- 复杂性

算法:

- 最速下降方法
- 共轭梯度方法
- 牛顿方法
- 拟牛顿方法
- 投影方法
- 罚函数方法

应用:

- 建立模型
- 理论分析
- 编程计算
- 解决实际问题

What will you learn?

- **Models --- the art: How we choose to represent real problems 【线性规划、非线性规划】**

- **Theory --- the science: What we know about different classes of models; e.g. necessary and sufficient conditions for optimality 【无约束优化、约束优化最优性条件】**

- **Algorithms --- the tools: How we apply the theory to robustly and efficiently solve powerful models 【单纯形法、无约束优化方法（最速下降法、牛顿法、拟牛顿法、共轭梯度法）约束优化方法（罚函数法、ALM、ADMM等）】**

最优化的简单实例

北京交通大學

工厂生产问题

已知：

	木门	木窗	总工时
木工	4小时	3小时	120小时/日
油漆工	2小时	1小时	50小时/日
利润	56	30	

问：每日安排生产多少扇木门多少木窗，才能使得总利润最大?

解：设该车间每日安排 生产木门x_1扇，木窗x_2

max $z=56\,x_1+30\,x_2$

s.t. $4\,x_1+3\,x_2\le 120$

$2\,x_1+x_2\le 50$

$x_1,\ x_2\ge 0$

最优化问题—线性规划

求解 →

最优解：(15,20)—最优决策

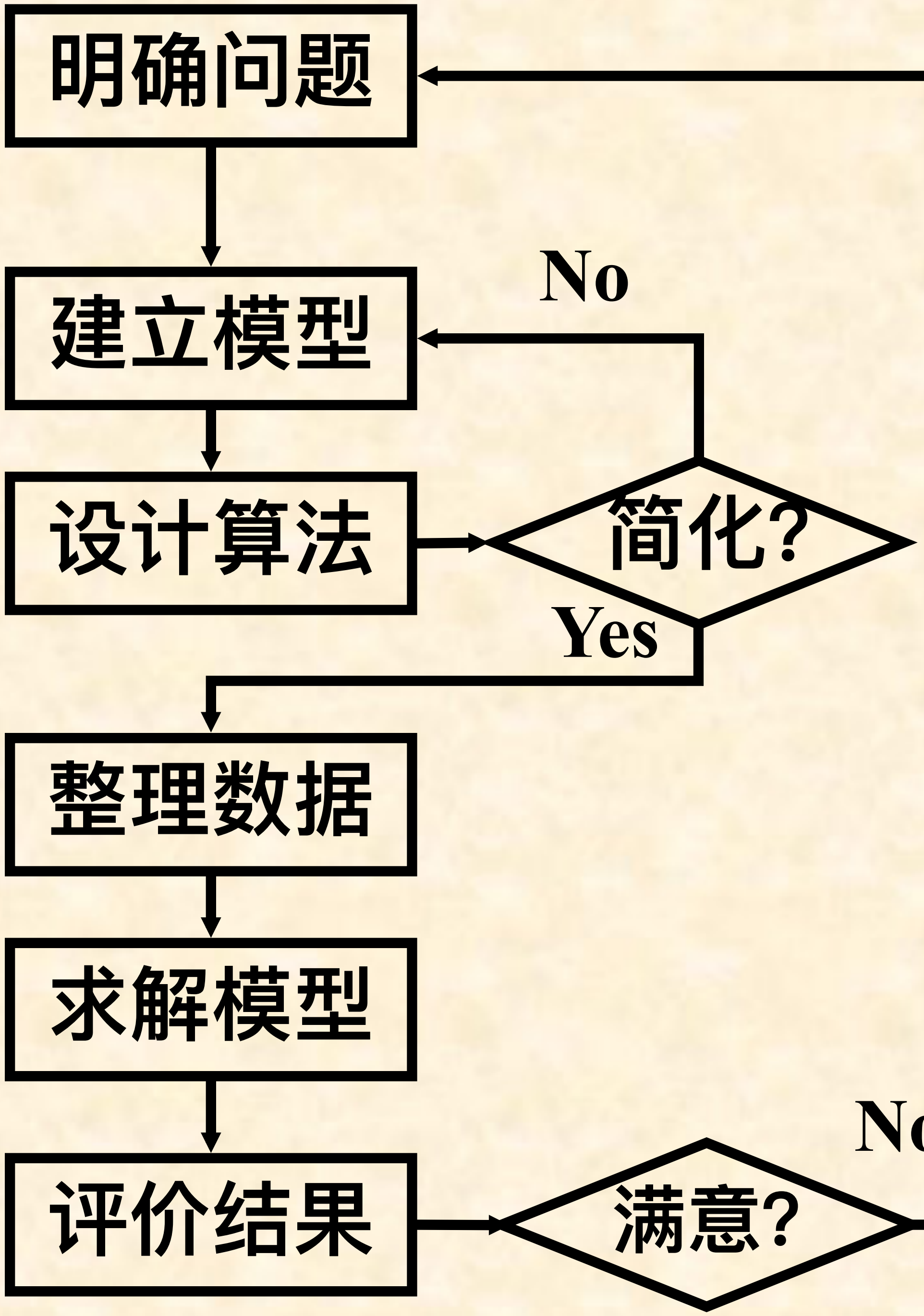
运输问题

把某种货物从m个工厂运到n个商店去，其中每个工厂的库存量为$a_1,a_2,\ldots,a_m$,各商店的需求量为$b_1,b_2,\ldots,b_n$,从工厂i到商店j的运费（每单位货物）为c_{ij},确定从工厂i到商店j的运输量$x_{ij}(i=1,\ldots,m,j=1,\ldots,n)$,使在满足供求的条件下，总的运费最小。

$$
\begin{cases}
\min \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}x_{ij} \\
s.t. \quad \sum_{i=1}^{m} x_{ij} = b_j \quad j=1,2,\cdots,n \\
\sum_{j=1}^{n} x_{ij} \le a_{ij} \quad i=1,2,\cdots,m \\
x_{ij} \ge 0 \quad i=1,2,\cdots,m \ \ j=1,2,\cdots,n
\end{cases}
$$

最优化处理实际问题的工作步骤

- 明确问题
- 建立模型
- 设计算法
- 整理数据
- 求解模型
- 评价结果



□ 无约束最优化理论

泰勒展开式 (Taylor Expansion)

- 一阶泰勒展开式：给定$f: \mathbb{R}^n \to \mathbb{R}$且$f \in C^1$，以及一点$\bar{x} \in \mathbb{R}^n$，则$f$在点$\bar{x}$的一阶泰勒展开式为

$$f(x) = f(\bar{x}) + \nabla f(\bar{x})^T(x-\bar{x}) + o(\|x-\bar{x}\|),$$

其中$o(\|x-\bar{x}\|)$表示当$\|x-\bar{x}\| \to 0$时，它是$\|x-\bar{x}\|$的高阶无穷小量**(例如：$\frac{1}{2}(x-\bar{x})^T\nabla^2 f(\bar{x})(x-\bar{x})$就是$\|x-\bar{x}\|$的高阶无穷小量).**

- 函数f在点$\bar{x}$处的最佳线性近似：

$$l(x) = f(\bar{x}) + \nabla f(\bar{x})^T(x-\bar{x})$$ **【最速下降法的基础】**

泰勒展开式 (Taylor Expansion)

- 二阶泰勒展开式：给定$f:\mathbb{R}^n \to \mathbb{R}$且$f \in C^2$，以及一点$\bar{x} \in \mathbb{R}^n$，则$f$在点$\bar{x}$的二阶泰勒展开式为

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

其中$o(\|x-\bar{x}\|^2)$表示当$\|x-\bar{x}\| \to 0$时，它是$\|x-\bar{x}\|^2$的高阶无穷小量。

- 函数f在点$\bar{x}$处的最佳二次近似：

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

【牛顿法的基础】

二阶连续可微函数

(Twice continuously differentiable function)

- **定义：** 设$f(x)$是$\mathbb{R}^n$上的一个连续可微函数，如果对每一点$x \in \mathbb{R}^n$，对所有$i, j = 1,2,\dots n$，二阶偏导数$\frac{\partial^2 f(x)}{\partial x_i \partial x_j}$存在且连续，则称$f$在$\mathbb{R}^n$上二阶连续可微，记作$f \in C^2$.

例题3： $\forall x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2$,**判断如下二元函数是否二阶连续可微**

(1) $f(x) = x_1^2 + x_2^2$ ✔

(2) $f(x) = \left(\max\{x_1, 0\}\right)^2 + \left(\max\{x_2, 0\}\right)^2$ ✘

连续可微 ⟹ 函数的梯度(Gradient)

- 连续可微函数f在点$x \in \mathbb{R}^n$处的梯度为n维列向量：

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \ldots, \frac{\partial f(x)}{\partial x_n}\right)^T$$

例题4： $\forall x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2$, 计算如下二元函数的梯度.

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

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}\right)^T = \begin{pmatrix} 2x_1 \\ 2x_2 \end{pmatrix}$$

(2) $f(x) = (\max\{x_1, 0\})^2 + (\max\{x_2, 0\})^2$

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}\right)^T = \begin{pmatrix} 2\max\{x_1, 0\} \\ 2\max\{x_2, 0\} \end{pmatrix}$$

二阶连续可微 $\Longrightarrow$ 函数的Hesse矩阵

- 二阶连续可微函数f在点$x \in \mathbb{R}^n$处的Hesse 矩阵为$n \times n$矩阵$\nabla^2 f(x)$，其表达式为

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

$$\frac{\partial^2 f(x)}{\partial x_i \partial x_j} = \frac{\partial^2 f(x)}{\partial x_j \partial x_i}$$

$$\Downarrow$$

$\nabla^2 f(x)$是对称矩阵

例题5：$\forall x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2$, 计算如下二元函数 $f(x) = x_1^2 + x_2^2$ 的Hesse矩阵

$$\nabla f(x) = \begin{pmatrix} 2x_1 \\ 2x_2 \end{pmatrix} \Rightarrow \nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} = 2I$$

单位矩阵

注意到:

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

$$\nabla^2 f(x)=\begin{bmatrix}\frac{\partial}{\partial x_1}\left(\frac{\partial f(x)}{\partial x_1}\right) & \frac{\partial}{\partial x_2}\left(\frac{\partial f(x)}{\partial x_1}\right) & \cdots & \frac{\partial}{\partial x_n}\left(\frac{\partial f(x)}{\partial x_1}\right)\\ \frac{\partial}{\partial x_1}\left(\frac{\partial f(x)}{\partial x_2}\right) & \frac{\partial}{\partial x_2}\left(\frac{\partial f(x)}{\partial x_2}\right) & \cdots & \frac{\partial}{\partial x_n}\left(\frac{\partial f(x)}{\partial x_2}\right)\\ & \vdots & & \\ \frac{\partial}{\partial x_1}\left(\frac{\partial f(x)}{\partial x_n}\right) & \frac{\partial}{\partial x_2}\left(\frac{\partial f(x)}{\partial x_n}\right) & \cdots & \frac{\partial}{\partial x_n}\left(\frac{\partial f(x)}{\partial x_n}\right)\end{bmatrix} \Rightarrow \begin{pmatrix}(\nabla(\boldsymbol{\nabla f(x)})_1)^T\\ (\nabla(\boldsymbol{\nabla f(x)})_2)^T\\ \vdots\\ (\nabla(\boldsymbol{\nabla f(x)})_n)^T\end{pmatrix}$$

由向量值函数$\boldsymbol{\nabla f(x)}: \mathbb{R}^n \to \mathbb{R}^n$的一阶导数信息产生的矩阵

无约束优化最优性条件

- 一阶必要条件
- 二阶必要条件
- 二阶充分条件
- 凸优化的一阶充要条件

无约束优化问题

数学模型：

$$\min_{x\in R^n} f(x) \qquad (1)$$

这里我们假设目标函数是连续可微的.

函数的下降方向：

设 $\bar{x}\in R^n$ 是任一给定向量， $d\neq 0$,若存在 $\delta>0$，对任意 $\lambda\in(0,\delta)$，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。

局部最优解的一阶必要条件

定理1：(一阶必要条件)若x^*是无约束优化问题(1)的局部最优解，则$\nabla f(x^*)=0.$

证明：反证法假设$\nabla f(x^*)\neq 0$，取$d=-\nabla f(x^*)\neq 0$,则对充分小的$\alpha>0$,由Taylor展开式得

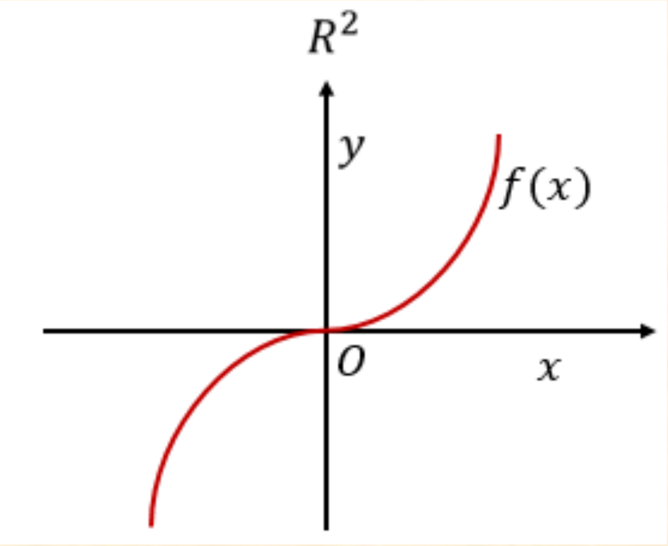
$$
\begin{aligned}
f(x^*+\alpha d) &= f(x^*)+\alpha\nabla f(x^*)^T d+o(\alpha)\\
&= f(x^*)-\alpha\|\nabla f(\bar{x})\|^2+o(\alpha)\\
&< f(x^*).
\end{aligned}
$$

这与x^*是局部最优解矛盾. 定理得证.

一阶条件的非充分性

定义：对无约束优化(1)，若$\nabla f(x^*)=0$，则x^*称为$f(x)$的一个**稳定点**(stationary point). 既不是极小值点，也不是极大值点的驻点称为**鞍点**(saddle point).

例题8：对于二次函数$f(x)=x^3$, $x^*=0$是它的稳定点，但是该点既不是极小值点，也不是极大值点，所以x^*是$f(x)$的鞍点.



局部最优解的二阶必要条件

定理3:(二阶必要条件)若$\bar{x}$是无约束优化(1)的局部最优解，则$\nabla f(\bar{x})=0$，且$Hesse$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，

Q $f(x)$在$\bar{x}$处二阶可微,且$\nabla f(\bar{x})=0$,

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

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

其中当$\lambda\to 0$时,$o(\lambda^2\|d\|^2)/\lambda^2\to 0$

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

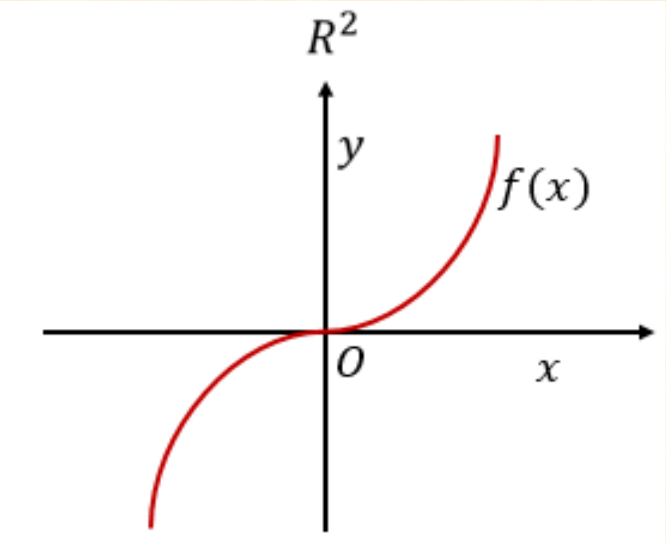
$\therefore$ 当$\lambda\to 0$时,有$d^T\nabla^2 f(\bar{x})d\geq 0$,即$\nabla^2 f(\bar{x})$为半正定的.

二阶必要条件的非充分性

例题： 对于二次函数$f(x)=x^3$, 在$x=0$ 处,

$$\nabla f\left(x\right)=0, \nabla^2 f\left(x\right)=O,$$

从而二阶必要性条件满足，然而0不是局部最优解。



二阶充分条件

定理4：设函数$f(x)$在点$\bar{x}$处二阶连续可微，若梯度$\nabla f(\bar{x})=0$，且$Hesse$矩阵$\nabla^2 f(\bar{x})$正定，则$\bar{x}$是(1)的严格局部最优解。

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

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

$$\text{其中}\lim_{\lambda\to 0} o(\lambda^2||d||^2)/\lambda^2=0.$$

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

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

由d的任意性,知$\bar{x}$是严格局部最优解. □

二阶充分条件

定理3′： 设函数$f(x)$在$\bar{x}$的邻域内二阶可微，若梯度$\nabla f(\bar{x}) = 0$，且Hesse矩阵$\nabla^2 f(x)$在该邻域内半正定，则$\bar{x}$是局部最优解。特别地，对于邻域内的任意点$x \neq \bar{x}$,若$\nabla^2 f(x)$是正定矩阵，则$\bar{x}$是优化问题（1）的严格的局部最优解。

例题： $\min\ f(x)=\frac{1}{3}x_1^3+\frac{1}{3}x_2^3-x_2^2-x_1$

解：由$\frac{\partial f}{\partial x_1}=x_1^2-1,\quad \frac{\partial f}{\partial x_2}=x_2^2-2x_2$，令$\nabla f(x)=0$，

得$\begin{cases}x_1^2-1=0\\ x_2^2-2x_2=0\end{cases}$，从而稳定点有：

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

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

显然，只有在$x^{(2)}$处，相应的$\nabla^2 f(x^{(2)})$正定，

其它三个点处均不是半正定矩阵，因此$x^{(2)}$是一个局部极小值点.

Homework

1.设函数 $f: R^n \to R$连续可微，$d \in R^n$为该函数在点$x \in R^n$处的下降方向. 试建立函数$\phi(\alpha) = f(x+\alpha d)$在$\alpha \geq 0$上的最小值点的必要条件.

2.设$a_1, a_2, \ldots, a_m \in R^n$. 求下面问题的最优解

$$\min_{x \in R^n} \sum_{i=1}^{m} \| a_i - x \|^2 .$$

注意：请大家将作业于下周六上课之前交给助教【提交方式：电子文档标明姓名、学号、学院）发送到邮箱　18118018 @bjtu.edu.cn】