组合优化与凸优化(combinatorial optimization and convex optimization)

# 课堂讨论

刘绍辉

计算机科学与技术学院 哈尔滨工业大学

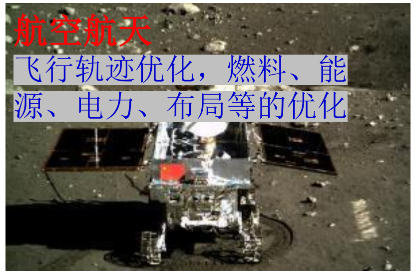
shliu@hit.edu.cn

2025年春季

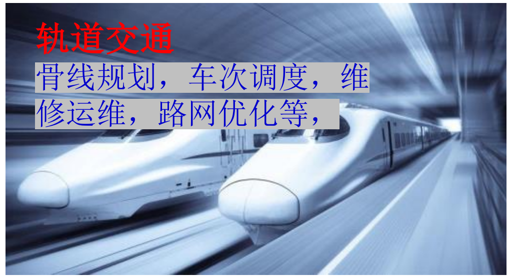
# 简介

- https://optimization-online.org/： 类似arxiv.org的网站， 专门发表跟优化相关的电子论文,2000年
- https://www.coin-or.org/： COmputational Infrastructure for Operations Research, Open Source for the operations research community, 2000年.
- http://Neos-server.org : NEOS Server: SOTA solvers for Numerical Optimization, a free internet-based for solving numerical optimization problems.由威斯康星大学麦迪逊分校托管的求解器运行在由HTCondor软件支持的分布式高性能机器上;远程求解器在亚利桑那州立大学、奥地利克拉根福大学和葡萄牙米尼奥大学的机器上运行。

# 优化求解应用广泛-涉及国民经济的各个行业










**问题建模+优化求解是解决实际应用问题的关键**


调度

规划

• • •

决策

# 优化求解器

- 科学技术
  - 科学-发现规律，总结规律
  - 技术-利用规律，解决问题
- 在国防、能源、制造、交通、金融、通信、计算机等各个领域出现了很多实际问题，对其建模后，很多都形式化为运筹优化问题，并指导各个领域和行业的精益化发展
- 实际问题通过规律和实际约束对其进行数学建模，然后对这些数学问题进行求解，这就是专门解决优化问题的求解器(Solver)
- 优化求解器主要以工业软件的形式为各个领域提供问题求解工具，能够将实际问题中的大规模复杂问题进行优化求解！

# 商用求解器vs开源求解器

- 数据驱动

数据获取 → 分析内在规律 → 建模与求解

- 数学规划/优化理论
  - 问题建模与求解的核心所在
  - 求解器：编程实现求解方法，给出最优解
  - 决策部门根据最优解及其敏感性分析作出最优决策！
- 商用和开源
  - Gurobi, IBM CPLEX, FICO XPRESS，Matlab，Excel (Frontline Solvers)
  - SCIP（德国），MOSEK（丹麦），GLPK（俄罗斯）
  - CVX-求解器建模平台（斯坦福大学：Stephen Boyd, Yinyu Ye）

华为：OPTV
阿里：MDOPT
杉数：COPT
中科院：CMIP
上财：LEAVES

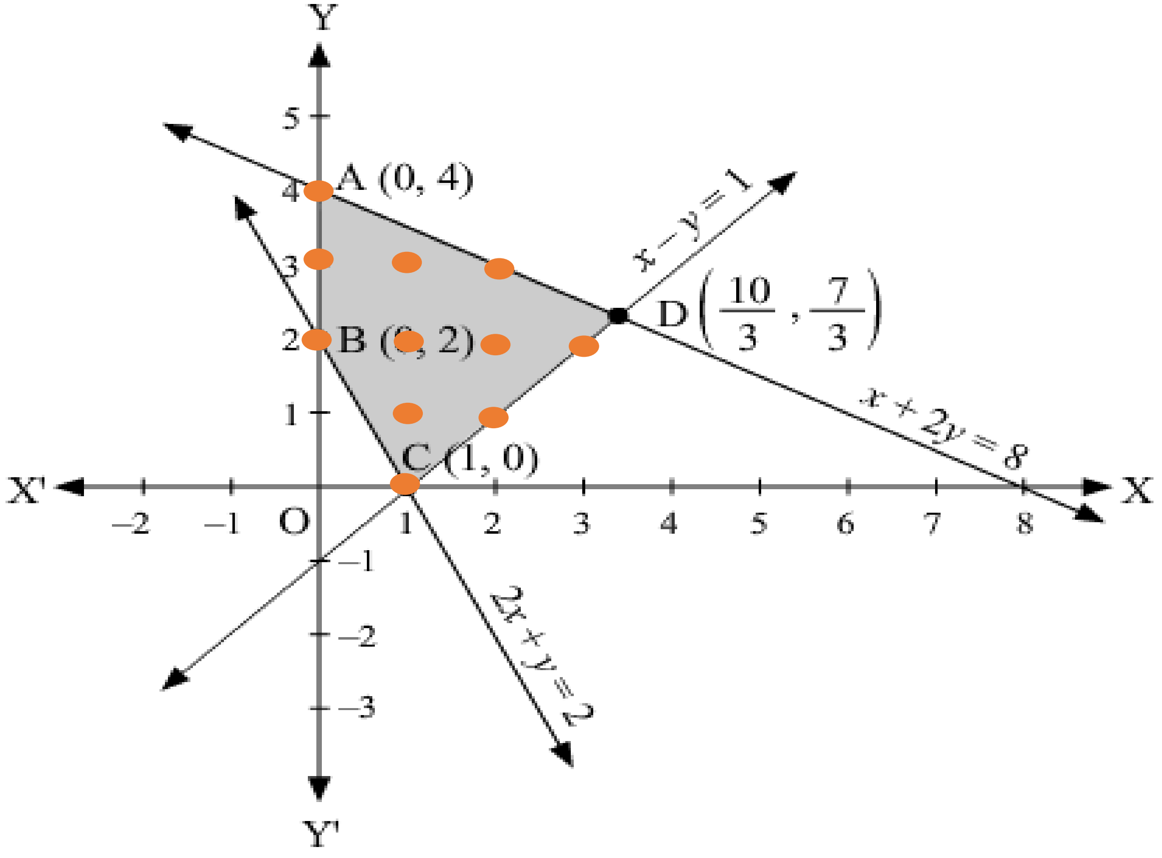
# 数学规划的分类

- 连续优化问题vs离散优化问题
  - COP：约束集包含无穷多个元素，并具有连续性特征，利用微积分和凸函数性质来分析
  - DOP：约束集是有限集，调度、路径规划、匹配问题、整数规划，利用组合数学和离散数学进行分析，同时也利用COP相关的方法
- 无约束优化（Unconstrained）vs约束优化
- 凸优化vs非凸优化：局部最优和全局最优，如何更有效！
- 线性规划（Linear Programming）vs非线性规划（Nonlinear Programming）
- 网络优化问题：同时具有连续优化和离散优化问题特点

# 线性规划LP-IP

LP
- 可行域内找最优解
- 规模大：数百万变量和约束
- 稀疏性：稀疏矩阵计算
- 数值：受计算机精度影响

IP
- 可行域是离散点
- 经典的NP完全问题
- 问题结构复杂多样
- 调用LP求解



# 非线性规划（NLP）

- 最小二乘LS型
  - $\min f(x) = \frac{1}{2}||g(x)||^2 = \frac{1}{2}\sum_{i=1}^{m}||g_i(x)||^2, s.t. x \in R^n$，这里函数$g$连续可微，包含$m$个组成函数$g_1, \cdots, g_m, g_i: R^n \to R^{r_i}$
  - 曲线拟合：假设要估计一个数学模型的$n$个参数，使得其能够很好地拟合一个基于测量的物理系统，假设其有如下形式：$z = h(x, y)$,其中$h$表示模型的已知函数，$x \in R^n$是一个未知参数向量，$z \in R^r$是模型输出，$y \in R^p$是模型输入，测得物理系统的$m$个输入-输出数据对$(y_1, z_1), \cdots, (y_m, z_m)$，想要寻找在最小平方误差和意义下能够最佳匹配该数据的参数向量$x$:
  - $\frac{1}{2}\sum_{i=1}^{m}||z_i - h(x, y_i)||$
  - 动态系统辨识：单输入单输出动力系统模型一般通过一个线性方程来建立输入序列$\{y_k\}$和输出序列$\{z_k\}$之间的联系：$\sum_{j=0}^{n}\alpha_j z_{k-j} = \sum_{j=0}^{n}\beta_j y_{k-j}$，给定来自实际系统的一组输入输出序列$y_1, z_1, \cdots, y_m, z_m$，期望得到一组参数$\{\alpha_j, \beta_j | j = 0,1, \cdots, n\}$,使其能够最小化：$\sum_{k=n}^{m}\left(\sum_{j=0}^{n}\alpha_j z_{k-j} - \sum_{j=0}^{n}\beta_j y_{k-j}\right)^2$

# 非线性规划（NLP）

- 最小二乘LS型
  - 神经网络MLP：第$k$层由$n_k$个激活单元组成，其中每一个单元都是$\phi: R \to R$上的单输入单输出映射，第$(k+1)$层的第j个激活单元的输出以$x_{k+1}^j$表示，其输入为第$k$层的输出向量$x_k = (x_k^1, \cdots, x_k^{n_k})$的线性函数，则
  - $x_{k+1}^j = \phi\left(u_k^{0j} + \sum_{s=1}^{n_k} x_k^s u_k^{sj}\right), j = 1, \cdots, n_{k+1}$，$(*)$其中$u_k^{sj}$（权重）是需要确定的。假设多层感知机有$N$层，令向量$u$表示所有的系数：$u = \left\{u_k^{sj} \middle| k = 0, \cdots, N-1, s = 0, \cdots, n_k, j = 1, \cdots, n_{k+1}\right\}$,则对给定的系数向量$u$,第1层的输入向量$x_0$通过$(*)$会生成唯一的第$N$层想来给你$x_N$，因此，可以将多层感知机看作参数为$u$的映射$h$，它将一个输入向量$x_0$转换成输出向量$x_N = h(u, x_0)$,同样，假设物理系统有$m$个输入输出对$(y_1, z_1), \cdots, (y_m, z_m)$，则适当选择参数$u$就能使MLP的映射与物理系统映射相匹配，一种常见的方法就是选取合适的参数$u$使得如下平方误差和最小：

$$\frac{1}{2}\sum_{i=1}^{m} ||z_i - h(u, y_i)||^2 ,$$

  - 寻找这种最优权重$u$的过程就是训练网络！

一般如何求解？

高斯牛顿法

# 非线性规划（NLP）

- 线性归回是最小二乘的一种特殊情况
- 逻辑回归：
  - $f: X \to Y, Y$是一个n维的二值向量，$f(x) = logit(\beta^T x), logit(\alpha) = \frac{1}{1+\exp(-\alpha)}$
  - 一般通过最大化观察数据的似然性来求解$\beta$,从而形成如下的优化问题：$\hat{\beta} = argmax_\beta \sum_{i=1}^{n}[y_i\beta^T x_i - \log(1 + \exp(-y_i\beta^T x_i))]$
- SVM:
  - 与逻辑回归类似，SVM试图找到可线性分类两个类的函数，这时元素$Y$的取值为1或-1.从而SVM形式化为：$\hat{\beta} = argmin_\beta \frac{1}{2}||\beta||_2^2, s.t.\ y_i(\beta^T x_i) \geq 1, \forall i = 1, \cdots, n$

# 非线性规划（NLP）

- 正则化/惩罚
  - 几乎可用于所有的机器学习问题，大部分情况下是为了简化所学的函数，通常通过强迫参数变“小”（数量或秩），并设定很多值为0，正则化也可以结合先验知识到问题的形式化表达当中
  - 岭回归（Ridge Regression）,增加$l_2$惩罚，目标函数为：
    - $\hat{\beta} = argmin_\beta ||X\beta - Y||_2^2 + \lambda||\beta||_2$,此时若$\lambda$越大则$\beta$的0值越多

有解析解closed form

  - Lasso回归，增加$l_1$惩罚，目标阐述为：
    - $\hat{\beta} = argmin_\beta ||X\beta - Y||_2^2 + \lambda||\beta||_1$

对标准的一阶方法，没有封闭解

# 非线性规划（NLP）

- 最小二乘LS，线性规划LP，以及凸二次规划（QP）的统一和推广
- 其中凸规划（Convex Programming）是非线性规划的子集，近二十年来，凸规划已成为与LS、LP和QP一样无处不在的技术。
- 与成熟的LS，LP和QP只需要对计算有基本的理解即可不一样，一般的CP的广泛使用要求具有较高的专业知识水平，用户必须理解凸分析的基础知识，这既是合理的，也是不可避免的;但事实上，还需要更深入的理解。
- 用户必须找到一种方法将他的问题转化为众多有限的标准形式之一;或者，如果做不到这一点，开发一个自定义求解器。对于那些专注于特定应用而不是底层数学的面向应用的用户来说，这些要求构成了使用凸规划的巨大障碍，特别是如果还不能确定结果是否会比其他方法更好的话。

# 优化求解器

- LP，MIP，NLP在实际应用中的占比
  - LP：15%
  - MIP：79%
  - NLP：7%
- 线性规划LP：约5万行代码
- 整数规划IP：100~200万行
- 非线性各个模块：10万行以下

# 第一章、绪论

- 各行各业都或多或少的应用运筹学或最优化的思想

**问题一：你觉得在哪些方面已经应用了这种思想？请从宏观和微观方面，举例说明。**

**问题二：结合你自己的研究方向？举例说明。**

machine learning = representation + optimization + evaluation

P. Domingos, an AAAI Fellow and a Professor of University of Washington

# 第一章、绪论

- 线性系统$\boldsymbol{Ax = b, A \in R^{m\times n}, m < n}$
  - 无穷多解
  - 无解：为此假设$A$是满秩矩阵
- 实际工程考虑
  - 图像超分辨率，输入为模糊下采样后的图像$\boldsymbol{b}$,矩阵$\boldsymbol{A}$表示降质运算，目标是从输入$\boldsymbol{b}$重构原始图像$\boldsymbol{x}$
  - $\boldsymbol{min_x J(x)\ s.t. b = Ax}$
  - $\boldsymbol{J(x)}$用来评估解的可取性，例如超分中可以是平滑性、分段平滑性等
  - 常见的$\boldsymbol{J(x)}$为$\boldsymbol{||x||_2^2}$,此时定义拉格朗日函数为：
  - $\boldsymbol{L(x) = ||x||_2^2 + \lambda^T(Ax - b)}$,求导数为0即可得：$\boldsymbol{\hat{x}_{opt} = -\frac{1}{2}A^T\lambda}$
  - 代入约束$\boldsymbol{Ax = b \Rightarrow A\hat{x}_{opt} = -\frac{1}{2}AA^T\lambda = b \Rightarrow \lambda = -2\left(AA^T\right)^{-1}b}$
  - 最后得$\boldsymbol{\hat{x}_{opt} = A^T\left(AA^T\right)^{-1}b = A^+b}$, 其中$\boldsymbol{A^+}$表示伪逆
  - 问题：如果$\boldsymbol{J(x) = ||Bx||_2^2}$（注意：$\boldsymbol{B^TB}$可逆），最优解=?

$$\boldsymbol{\hat{x}_{opt} = \left(B^TB\right)^{-1}A^T\left(A\left(B^TB\right)^{-1}A^T\right)^{-1}b}$$

# 第一章、绪论

- 为何$\boldsymbol{l_2}$使用广泛?
  - 具有唯一的解析形式的解，简单
  - 在反问题，信号表达等很多地方常见
  - 那是否$\boldsymbol{l_2}$就是最好的选择呢？为什么?
- 如果不是，如何给出解?
  - $\boldsymbol{l_p}$范数$(\boldsymbol{p} \geq \mathbf{1})$：$||\boldsymbol{x}||_{\boldsymbol{p}}^{\boldsymbol{p}} = \sum_i |\boldsymbol{x}_i|^{\boldsymbol{p}}$,其中特殊情况为$\boldsymbol{l}_\infty = \boldsymbol{max}_i(|\boldsymbol{x}_i|), \boldsymbol{l}_\mathbf{1} = \sum_i |\boldsymbol{x}_i|$
  - 而其中$\boldsymbol{l_1}$在一定条件下往往倾向于稀疏解，并且$\boldsymbol{l_p}(\boldsymbol{p} \geq \mathbf{1})$是凸的，因此很常用
  - 并且，此时该问题与线性规划问题等价!
  - $\boldsymbol{l_0}$范数：不是真正的范数，且$\boldsymbol{l_0}$非凸，求解困难 为什么？

# 第一章、绪论

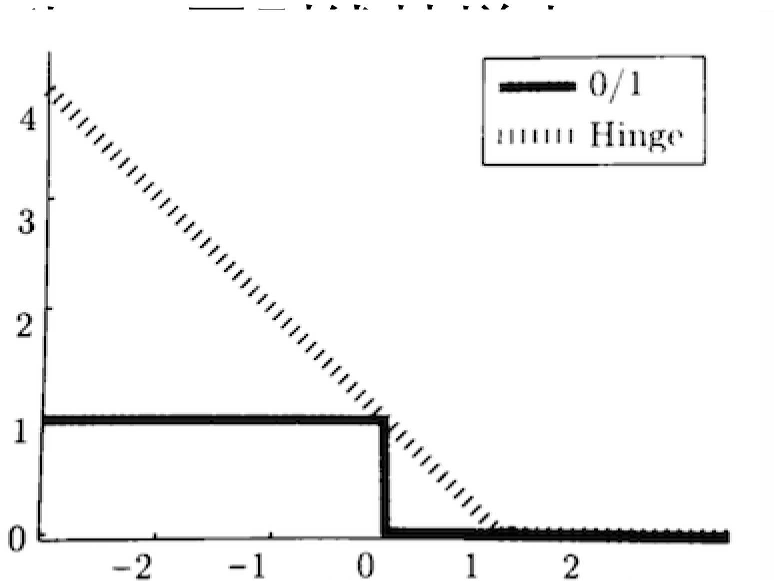
- 压缩感知（CS：Compressive sensing)
  - 求解方程$\boldsymbol{Ax}=\boldsymbol{b},\boldsymbol{x}\in\boldsymbol{R}^{\boldsymbol{n}},\boldsymbol{b}\in\boldsymbol{R}^{\boldsymbol{m}},\boldsymbol{A}\in\boldsymbol{R}^{\boldsymbol{m}\times\boldsymbol{n}},\boldsymbol{m}\ll\boldsymbol{n}$
  - 如接收到信号$\boldsymbol{b}$,已知矩阵$\boldsymbol{A}$，重构向量$\boldsymbol{x}$（无穷多），先验：稀疏性，即$\boldsymbol{x}$中有很多0元素
  - m=128;n=256;A=randn(m,n);u=sprandn(n,1,0.1);b=A*u;
  - 理论上，$\boldsymbol{u}$是如下$\boldsymbol{l_0}$范数问题的最优解：
  - $\boldsymbol{min_{x\in R^n}||x||_0,s.t.Ax=b}$
  - 这是一个NP难的，如何求解？
  - 定义$\boldsymbol{l_1}$范数：$\boldsymbol{||x||_1=\sum_{i=1}^{n}|x_i|}$,则在一定条件下，$\boldsymbol{l_0}$可等价转换为$\boldsymbol{l_1}$问题：$\boldsymbol{min_{x\in R^n}||x||_1,s.t.Ax=b}$
  - 能否使用$\boldsymbol{l_2}$范数来替代$\boldsymbol{l_0}$范数呢？

# 第一章、绪论

- SVM
  - 最大间隔算法与拉格朗日对偶
  - Hinge loss
- 二分类问题：标签$\mathbf{y} = \pm\mathbf{1}$,预测值$\hat{\boldsymbol{y}} \in \boldsymbol{R}$
  - $\hat{\mathbf{y}} > +\mathbf{1}/\hat{\boldsymbol{y}} < -\mathbf{1}$，此时分类正确，损失为0
  - $\hat{\mathbf{y}} \in (-\mathbf{1}, \mathbf{1})$，分类不确定，损失不为0
  - $\hat{\mathbf{y}} = \mathbf{0}$,损失最大
- 对输出$\mathbf{y}$,则损失可写为
  - $\boldsymbol{l}(\boldsymbol{y}) = \boldsymbol{max}\{\mathbf{0}, \mathbf{1} - \boldsymbol{y} \cdot \hat{\boldsymbol{y}}\}$
  - 这就是Hinge loss在二分类问题的变体，可看做是双向Hinge loss.

# 第一章、绪论

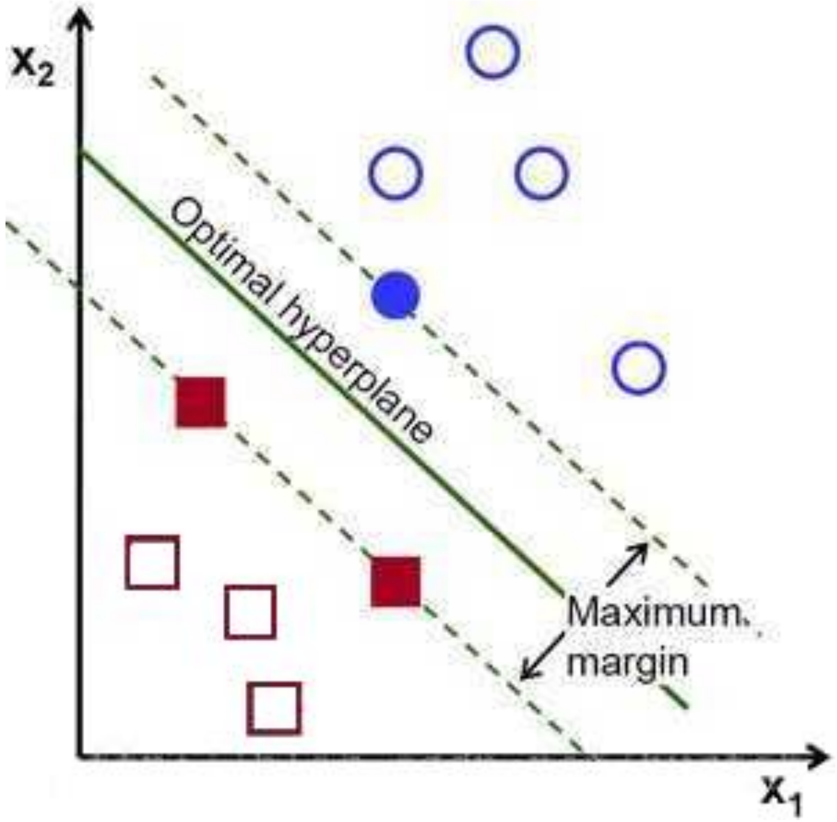
- 单向hinge loss
  - $\boldsymbol{y} = \pm\mathbf{1}, \boldsymbol{y} \geq \mathbf{1}$,los




- SVM呢？

# 第一章、绪论

- 线性可分SVM的预测值$\hat{\mathbf{y}} = \boldsymbol{wx} + \boldsymbol{b} \in \boldsymbol{R}$，如果分离超平面在如图，且支持向量与分割平面之间的距离为1
  - 每个$\mathbf{y} = \mathbf{1}$的样本$\hat{\mathbf{y}} \geq \mathbf{1}$，$\mathbf{y} = -\mathbf{1}$的样本$\hat{\mathbf{y}} \leq -\mathbf{1}$，样本集上损失为0
  - 如果分类错误，则Hinge loss大于0。Hinge loss约束超平面作出调整。如果超平面距离支持向量的距离小于1，则Hinge loss大于0，且就算分离超平面满足最大间隔，Hinge loss仍大于0



# 第一章、绪论

- 预测值$|\hat{y}|$越大，表明样本点越远离分离超平面，分类越容易。选择优化设计时，没必要关注离超平面很远的点，因此可通过对距离分离超平面的距离选择一个阈值，来去除这些样本！
  - $l(y) = max\{0, 1 - y \cdot \hat{y}\}$，其中的1就是阈值，可看做是超参数，对预测值超过阈值的样本无需考虑
  - 考虑多分类
    - $l(y) = \max\{0, 1 + \max_{\hat{y} \neq y} \mathbf{w}_{\hat{y}}\mathbf{x} - \mathbf{w}_{\mathbf{y}}\mathbf{x}\}$
    - $l(y) = \sum_{t \neq y} \max\{0, 1 + \mathbf{w}_{\hat{y}}\mathbf{x} - \mathbf{w}_{\mathbf{y}}\mathbf{x}\}$
- 损失函数知道，如何求解？
  - $\frac{\partial l}{\partial w_i} = \begin{cases} -y \cdot x_i & y \cdot \hat{y} < 1 \\ 0 & \text{其它} \end{cases}$，有什么问题？

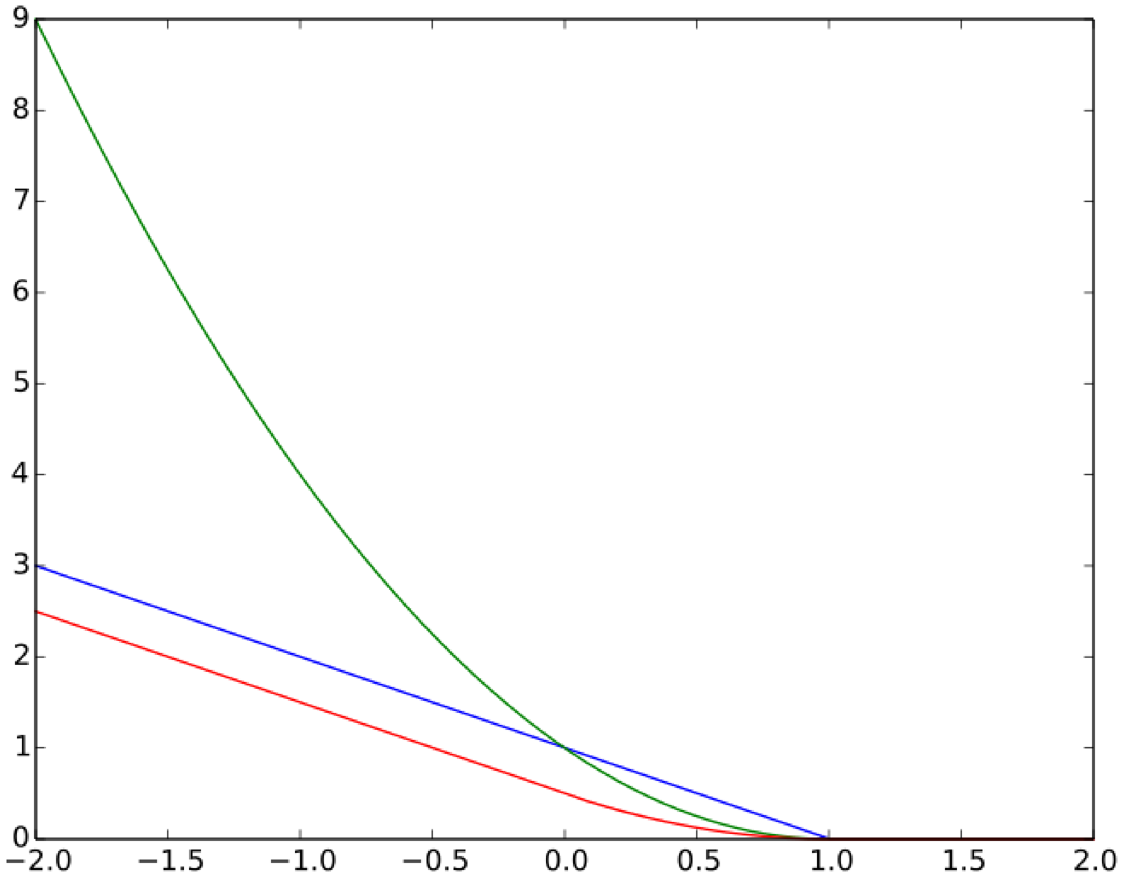
# 第一章、绪论

- 损失函数在$\mathbf{y}\cdot\hat{\boldsymbol{y}} = \mathbf{1}$处不可导
  - 各种光滑变体
    - 分段光滑[2]: $\boldsymbol{l}(\hat{\boldsymbol{y}}) = \begin{cases} \frac{1}{2} - y\cdot\hat{y} & y\cdot\hat{y} \le 0 \\ \frac{1}{2}(1 - y\cdot\hat{y})^2 & 0 < y\cdot\hat{y} \le 1 \\ 0 & 1 \le y\cdot\hat{y} \end{cases}$
    - 平方光滑： $\boldsymbol{l}_\gamma(\hat{\boldsymbol{y}}) = \begin{cases} \frac{1}{2\gamma} max\{0, 1 - y\cdot\hat{y})^2 & y\cdot\hat{y} \ge 1-\gamma \\ 1 - \frac{\gamma}{2} - y\cdot\hat{y} & \text{其它} \end{cases}$
    - 其中更改的Huberloss是$\gamma = 2$是平方光滑的特例$L(y, \hat{y}) = 4l_2(\hat{y})$

1.Zhang, Tong. Solving large scale linear prediction problems using stochastic gradient descent algorithms (PDF). ICML. 2004.
2.Rennie, Jason D. M.; Srebro, Nathan. Loss Functions for Preference Levels: Regression with Discrete Ordered Labels (PDF). Proc. IJCAI. 2005

# 第一章、绪论

- “普通变体”（蓝色），平方变体（绿色），以及 Rennie 和 Srebro 提出的分段



# 第一章、绪论

- 典型问题
- $\mathbf{min}_{\mathbf{w}\in\mathbb{R}^n} \frac{1}{\mathbf{m}} \sum_{i=1}^{m} l(p(x_i;w),y_i) + \lambda R(w)$
  - $l(p,y) = \frac{1}{2}(p-y)^2; l(p,y) = log(1+e^{-py}); l(p,y) = max\{0,1-py\}$
  - $p(x;w) = w^T x - b; p(x;W) = \phi(W_n\phi(W_{n-1}\cdots\phi(W_1 x)\cdots))$
  - $R(w) = \frac{1}{2}||w||^2; R(w) = ||w||_1$
- SVM:hinge loss, 线性分类函数，$\mathbf{l}_2$正则化
- 正则化逻辑斯谛回归：logistic损失，线性回归函数，$\mathbf{l}_2$正则化
- 多层感知机：平方损失，前向反馈函数，R(W)=0
- LASSO问题：平方损失，线性回归函数，$\mathbf{l}_1$正则化

# 第一章、绪论

- 典型问题
  - $\boldsymbol{min}_{X\in\mathbb{R}^{m\times n}}||X||_*, s.t. X_{ij}=D_{ij}, \forall(i,j)\in\Omega$，其中$\Omega$表示观测值的位置
- 低秩表示问题
  - $\mathbf{min}_{\mathrm{Z}\in\mathbb{R}^{m\times n}, E\in\mathbb{R}^{m\times n}}||\mathrm{Z}||_* + \lambda||E||_1, s.t. D = DZ + E$
- 为减少计算代价和存储空间，低秩矩阵可分解为两个小得多的矩阵乘积：$\mathbf{X}=\mathbf{UV^T}$,此时上述矩阵补全问题可形式化为如下的非凸问题
  - $\mathbf{min}_{\mathrm{U}\in\mathbb{R}^{m\times r}, V\in\mathbb{R}^{n\times r}}\frac{1}{2}\sum_{(i,j)\in\Omega}||U_iV_j^T - D_{ij}||_F^2 + \frac{\lambda}{2}(||U||_F^2 + ||V||_F^2)$

# 第一章、绪论

- 优化分类
- $\min \boldsymbol{f_0(x)}, \boldsymbol{s.t.\, f_j(x)} \leq (\geq, =)\mathbf{0}, \boldsymbol{j} = \mathbf{1}, \cdots \boldsymbol{m}, \boldsymbol{x} \in \boldsymbol{Q}$
- **可行集** $\mathcal{F} = \{\boldsymbol{x} \in Q | f_j(x) \leq 0, j = 1, \cdots, m\}$
- 约束问题：$\mathcal{F} \subset \mathbb{R}^{\boldsymbol{n}}$;无约束问题：$\mathcal{F} = \mathbb{R}^{\boldsymbol{n}}$;光滑问题：$\boldsymbol{f_j}(\cdot)$都可微；非光滑问题：$\boldsymbol{f_k}(\cdot)$不可微；线性约束问题：$f_j(x) = \sum_{i=1}^{n} a_j^{(i)} x^{(i)} + b_j \equiv < a_j, x > + b_j, j = 1, \cdots, m$都是仿射的，如果$f_0(x)$也是仿射的，则实线性优化问题；如果目标函数$f_0(x)$是二次的，则为二次优化问题；如果所有约束都是二次函数，则为二次约束的二次优化问题

# 第一章、绪论

- 优化问题也可分为可行的，和不可行的
- 如果$\mathcal{F} \neq \emptyset$,则称问题是可行的
- 如果存在某个$\boldsymbol{x} \in \boldsymbol{Q}$，使得所有不等式约束和等式严格成立，则称为严格可行的
- 一般说来，优化问题应该是不可解的（？）
  - 的确，从我们的现实生活经验来看，很难相信存在一种能够解决世界上所有问题的万能工具。
- 根据函数的性质称为零阶，一阶和二阶
  - 零阶：只利用函数值
  - 一阶：利用函数值和导数值
  - 二阶：利用函数值、导数值和Hessian矩阵值

# 第一章、绪论

- 优化(Optimization)无所不在，包罗万象，但在国内的应用跟国外还有一些差距

**问题一：在宏观方面，你觉的哪些方面能够采用这些技术来提高效率？**

**问题二：优化技术应用的难点在哪？**

# 第一章、绪论

- 优化(Optimization)无所不在，难不难解？

**问题一：线性优化容易，非线性优化难？**

**问题二：凸优化容易，非凸优化难？**

# 第一章、绪论

- 优化或运筹学是解决实际应用的问题，这显然需要建模

**问题一：建模的基本流程是怎么样的?**

**问题二：数学建模的难点在哪?**

**问题三：你听说过幸存者定律，墨菲定律（Murphy's Law）？**

Anything that can go wrong will go wrong

# 第一章、绪论

- 最优化(Optimization) 是数学当中比较流行的词汇

**问题一：你听说过次梯度，梯度投影，拉格朗日对偶，强对偶，弱对偶等名词吗？**

**问题二：听说过秩1校正、秩2校正吗？**

# 第一章、绪论

- 最优化(Optimization) 是数学当中比较流行的词汇

**李普希兹连续，梯度李普希兹连续，变量拆分，块坐标下降（block coordinate descent）,Q-收敛速度，超线性、线性、次线性、二次收敛，R-收敛速度，优化算法复杂度（$N(\epsilon) = O\left(\frac{1}{\epsilon^2}\right)$）**

# 第一章、绪论

- 典型的优化问题
  - $\mathcal{P}: minimize\ f(x), s.t. x \in D$
  - 实际上就是研究函数的性质
- 极限：$\lim_{x \to a} f(x)$表示函数$f$随变量$x$趋近于$a$时，函数所趋近的值!
- 连续性：函数$f(x)$在定义域内一点$x'$连续，表明存在序列$x_1, x_2, \cdots \to x'$,则值$f(x_1), f(x_2), \cdots$趋近于$f(x')$.即：$\lim_{(i \to \infty)} f(x_i) = f\left(\lim_{i \to \infty} x_i\right)$，在定义域内任一殿都连续，则函数$f$称为连续函数
- 可微性：函数$f: \mathbb{R}^n \to \mathbb{R}$在定义域内点上是可微的，则存在向量$\nabla f(x)$满足：

$$\lim_{z \in dom f, z \neq x, z \to x} \frac{||f(z) - f(x) - Df(x)(z-x)||_2}{||z-x||_2} = 0, \nabla f(x)\text{称为函数的导数}$$

- 光滑性：如果函数$f$的导数在其定义域内的每一点都是连续的，则称函数是光滑的。该定义可以递归应用到高级光滑性，例如函数是$n$阶光滑的，则表明$f^{(n)}(x)$在定义域上是连续的

# 第一章 绪论

- 李普希兹性（Lipschitz）:函数$f$是常数$L$李普希兹连续的，表明其满足：$||f(x)-f(y)|| \le L||x-y||, \forall x,y \in dom f$
  - 满足李普希兹性质的函数不仅连续，而且不会快速改变其取值
  - 这与光滑性也不无关系，但函数可以是李普希兹连续但不是光滑的!
  - $p$阶lipschitz连续：即对$\forall \boldsymbol{x},\boldsymbol{y} \in \mathrm{dom} f$,都有$||\boldsymbol{\nabla^p f(x)} - \boldsymbol{\nabla^p f(y)}|| \le \boldsymbol{L}||\boldsymbol{x}-\boldsymbol{y}||$
  - 讨论算法收敛性的时候往往会往目标函数上加上各种条件，其中Lipschitz是常见的一种

# 第一章、绪论

- 作为最基本的线性规划

**问题一：听说过丹齐格，敏感性分析吗？**

**问题二：为什么要学线性规划？**

# 第一章、绪论

- 作为最优化迭代的基本形式

$\boldsymbol{x_{k+1} = x_k + \alpha p_k}$,其中$\boldsymbol{\alpha}$为步长，$\boldsymbol{p_k}$为方向，

**问题一：你知道几种步长和方向的求法?**

**问题二：你知道线搜索和信任域方法的区别吗?**

# 第一章、绪论

- 优化算法中一般涉及泰勒逼近：

$$f(x+p)=f(x)+\nabla f(x+tp)^{T}p, t\in(0,1)$$

$$f(x+p)=f(x)+\nabla f(x)^{T}p+\frac{1}{2}p^{T}\nabla^{2}f(x+tp)p,\ t\in(0,1)$$

**问题一：连续可微性为什么很重要？**

**问题二：迭代求解中，如果保证目标函数值下降，是否可以求得极小值？**

# 第一章、绪论

- 优化算法中一般需要求极小值:

$$f(x+p)=f(x)+\nabla f(x+tp)^T p, t\in(0,1)$$

$$f(x+p)=f(x)+\nabla f(x)^T p+\frac{1}{2}p^T\nabla^2 f(x+tp)p,\ t\in(0,1)$$

**问题一：连续可微性为什么很重要?**

**问题二：迭代求解中，如果保证目标函数值下降，是否可以求得极小值?**

# 第二章 线性规划

- 集合的概念
- 开集，闭集，紧集
- 每个有界并单调非递减或非递增序列都收敛。每个单调非递减序列$\{\boldsymbol{x_k}\}$要么有界，要么$\boldsymbol{x_k} \to \infty$
- 集合$\boldsymbol{X} \subset \boldsymbol{R^n}$,$\boldsymbol{x}$为集合$\boldsymbol{X}$的封闭点(closure)，如果存在序列$\{\boldsymbol{x_k}\} \subset \boldsymbol{X}$收敛于$\boldsymbol{x}$。集合$\boldsymbol{X}$的闭包：$\boldsymbol{cl}(\boldsymbol{X})$为所有闭点的集合。
- 闭集：闭包与其本身相等
- 开集：闭集的补集
- 紧集：闭且有界

# 第二章 线性规划

- 设$\boldsymbol{L} \subset \boldsymbol{R}^n$为子空间，其余子空间或正交补记为：

$$\boldsymbol{L}^{\perp} = \{\boldsymbol{x} \in \boldsymbol{R}^n | \boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{0}, \forall \boldsymbol{y} \in \boldsymbol{L}\}$$

- 显然，任意$\boldsymbol{z} \in \boldsymbol{R}^n$,存在唯一分解$\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{y}$,使得$\boldsymbol{x} \in \boldsymbol{L}$,$\boldsymbol{y} \in \boldsymbol{L}^{\perp}$.称$\boldsymbol{x}, \boldsymbol{y}$分别为$\boldsymbol{z}$在子空间$\boldsymbol{L}$，$\boldsymbol{L}^{\perp}$上的投影，这时有：$||\boldsymbol{z}||^2 = ||\boldsymbol{x}||^2 + ||\boldsymbol{y}||^2$,这就是投影定理
- 问题1.这个定理内容是否理解？
- 问题2.你是否听说过投影定理（Projection Theorem）

# 第二章 线性规划

- 设$\boldsymbol{L}$为$\boldsymbol{R^n}$为子空间，那么，$\forall \boldsymbol{z} \in \boldsymbol{R^n}$,∃唯一$\boldsymbol{x} \in \boldsymbol{L}, \boldsymbol{y} \in \boldsymbol{L}^{\perp}$,使得$\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{y}$,且$\boldsymbol{x}$为问题：

$$\begin{cases} \boldsymbol{min}||\boldsymbol{z} - \boldsymbol{u}|| \\ \boldsymbol{s.t.} \, \boldsymbol{u} \in \boldsymbol{L} \end{cases}$$

的唯一解，而问题的最优值为$||\boldsymbol{y}||$

- 问题1.这个定理跟之前的投影定理是否有不同？
- 问题2.子空间$\boldsymbol{L}$是否是凸的？
- 问题3.参看教材17页的定理1-2.

# 第二章 线性规划

- 你知道对线性规划影响最大的三种算法是什么算法吗？
  - 单纯形法（simplex algorithms）
  - 内点法（interior point algorithm)
  - 椭球法（ellipsoid method）
- 迄今未发现单纯形法的任何变种具有多项式运行时间
- 椭球法已证明对LP问题具有多项式时间算法，但其效率低下不适合实际使用
- 但内点法（最坏情况具有指数运行时间），和单纯形法远比椭球法更有效，因而实际广泛使用

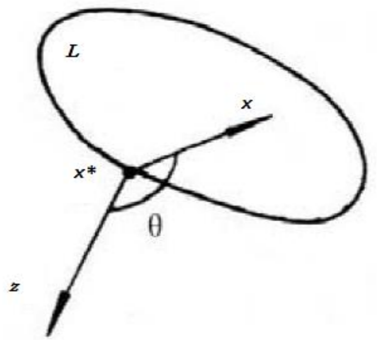
# 第二章 线性规划

- 设$\boldsymbol{L}$为$\boldsymbol{R^n}$的非空闭凸子集，那么，$\forall \boldsymbol{z} \in \boldsymbol{R^n}$,∃唯一的向量$\boldsymbol{x^*} \in \boldsymbol{L}$,且$\boldsymbol{x^*}$为问题：

$$\begin{cases} \boldsymbol{min||z-x||} \\ \boldsymbol{s.t. x \in L} \end{cases}$$

的唯一解，称其为向量$\boldsymbol{z}$在集合$\boldsymbol{L}$上的投影。并且，向量$\boldsymbol{x^*}$是$\boldsymbol{z}$在$\boldsymbol{L}$上的投影当前仅当（if and only if）：

$\boldsymbol{(z-x^*)'(x-x^*) \leq 0, \forall x \in L}$ ---------(*)



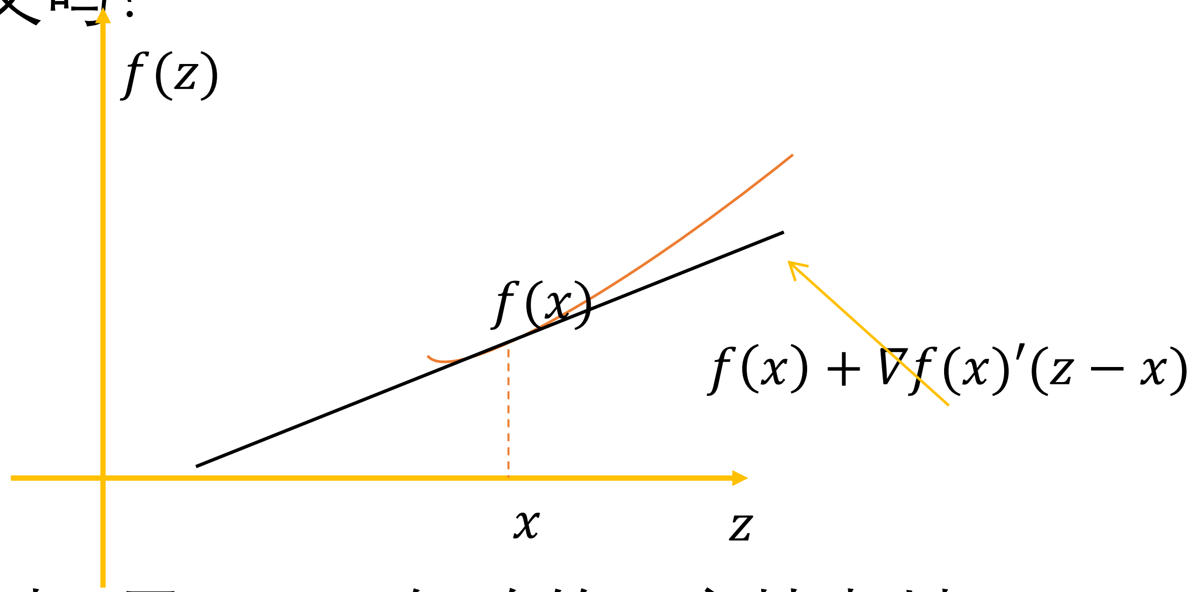

- 问题1.这种描述之前听说过吗？
- 问题2.是否会证明？
  - 最小化$\boldsymbol{||z-x||}$等价于最小化凸可微函数：$\boldsymbol{f(x) = \frac{1}{2}||z-x||^2}$,则$\boldsymbol{x^*}$在集合$\boldsymbol{L}$上最小化$\boldsymbol{f}$等价于$\boldsymbol{\nabla f(x^*)'(x-x^*) \geq 0, \forall x \in L}$,二这正好是上述(*)的内容!
- 问题3.唯一性如何证明？
  - 在集合$\boldsymbol{L}$上最小化$\boldsymbol{f}$等价于在紧集$\boldsymbol{L \cap \{||z-x|| \leq ||z-w||\}, \forall w \in L}$.由Weierstrass定理（连续函数$\boldsymbol{f: R^n \to R}$在$\boldsymbol{R^n}$的任何紧子集上都有极小点），存在一个最小化向量。若存在两个最小化向量，则必定相等!

# 第二章 线性规划

- 凸函数的操作:
- $\boldsymbol{f}: \boldsymbol{R}^{\boldsymbol{m}} \rightarrow (-\infty, \infty], \boldsymbol{A}$为$\boldsymbol{m} \times \boldsymbol{n}$的矩阵，令$\boldsymbol{F}: \boldsymbol{R}^{\boldsymbol{n}} \rightarrow (-\infty, \infty], \boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{A}\boldsymbol{x}), \boldsymbol{x} \in \boldsymbol{R}^{\boldsymbol{n}}$
- 问题1.如果$\boldsymbol{f}$是凸函数，则$\boldsymbol{F}$是否是凸函数？如果$\boldsymbol{f}$是闭的，则$\boldsymbol{F}$也是闭的。如何证明?
- 问题2.函数的和是否是与这些函数的凸性有关呢?
  - $\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{\gamma}_{\mathbf{1}}\boldsymbol{f}_{\mathbf{1}}(\boldsymbol{x}) + \cdots + \boldsymbol{\gamma}_{\boldsymbol{m}}\boldsymbol{f}_{\boldsymbol{m}}(\boldsymbol{x}), \boldsymbol{x} \in \boldsymbol{R}^{\boldsymbol{n}}, \boldsymbol{\gamma}_{\boldsymbol{i}} > \mathbf{0}$
- 问题3.多个函数的上确界呢?
  - $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{sup}_{\boldsymbol{i} \in \boldsymbol{I}}\boldsymbol{f}_{\boldsymbol{i}}(\boldsymbol{x}), \boldsymbol{I}$为任意索引集，$\boldsymbol{f}_{\boldsymbol{i}}: \boldsymbol{R}^{\boldsymbol{n}} \rightarrow (-\infty, \infty]$
  - $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{max}_{\boldsymbol{i} \in \boldsymbol{I}}\boldsymbol{f}_{\boldsymbol{i}}(\boldsymbol{x})$

# 第二章 线性规划

- 凸函数的几何意义等价于
  - $\boldsymbol{f(z) \geq f(x) + \nabla f(x)'(z-x), \forall x, z, \in C}$,C为非空凸子集
  - 严格凸的时候，则等号没有
- 问题1.理解凸函数的几何意义吗？
- 问题2.如何证明上述结论？




反例：
$f(x)=x_1^2-x_2^2, C=\{(x_1,0)|x_1\in R\}, f$是凸的，但$\nabla^2 f(x)$并不是正定的；又如：$f(x)=x^4$,严格凸，但$\nabla^2 f(x)$并不一定$>0$，例如在0点就为0.

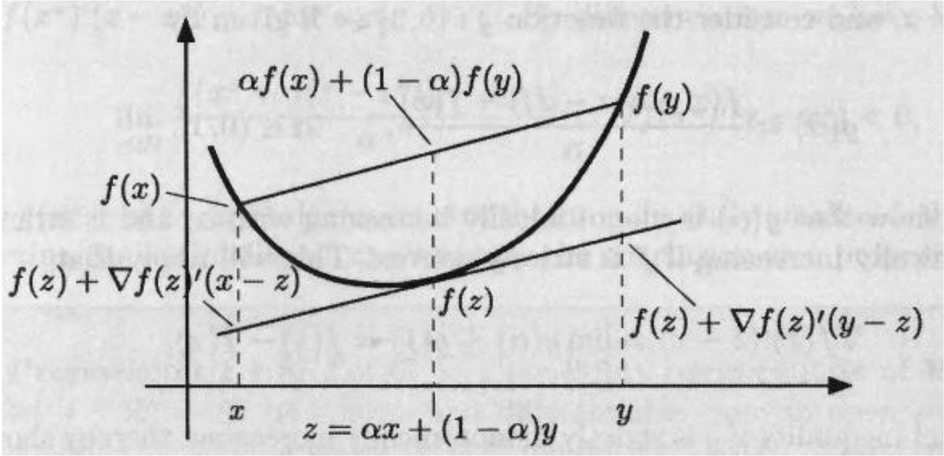
- 问题3.如果函数二次可微，则可用Hessian矩阵的正定性来判断：如果为凸$\boxed{\leftrightarrow}$ $\boldsymbol{\nabla^2 f(x) \geq 0}$;若$\boldsymbol{\nabla^2 f(x) > 0 \Rightarrow f}$为严格凸。
  问：若$\boldsymbol{f}$严格凸，则$\boldsymbol{\nabla^2 f(x) > 0}$是否成立？

# 第二章 线性规划

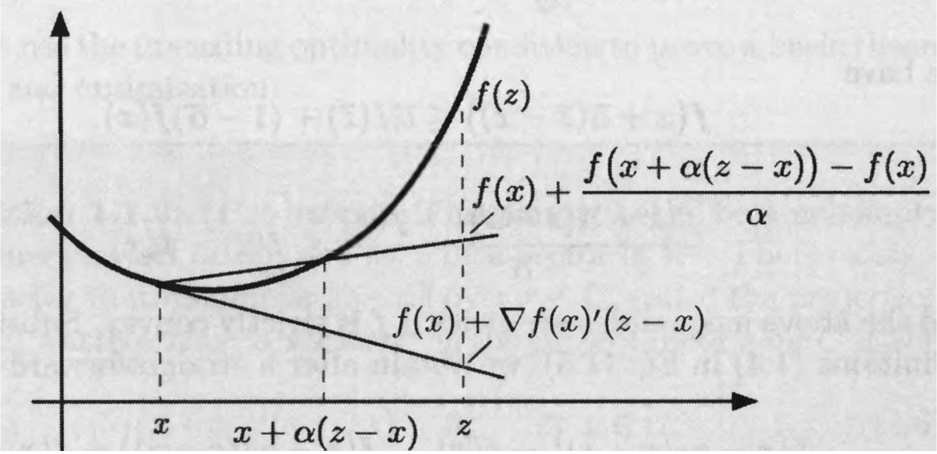
- 凸函数组合的几何意义
  - $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{C}, \boldsymbol{\alpha} \in [\mathbf{0}, \mathbf{1}]$,设$\boldsymbol{z} = \boldsymbol{\alpha x} + (\mathbf{1} - \boldsymbol{\alpha})\boldsymbol{y}$,则根据上页的不等式有:
  - $\boldsymbol{f}(\boldsymbol{x}) \geq \boldsymbol{f}(\boldsymbol{z}) + \nabla \boldsymbol{f}(\boldsymbol{z})'(\boldsymbol{x} - \boldsymbol{z}),----(\mathbf{1})$
  - $\boldsymbol{f}(\boldsymbol{y}) \geq \boldsymbol{f}(\boldsymbol{z}) + \nabla \boldsymbol{f}(\boldsymbol{z})'(\boldsymbol{y} - \boldsymbol{z}),----(\mathbf{2})$
  - $(1) * \boldsymbol{\alpha} + (\mathbf{2}) * (\mathbf{1} - \boldsymbol{\alpha}) \Rightarrow \boldsymbol{\alpha f}(\boldsymbol{x}) + (\mathbf{1} - \boldsymbol{\alpha})\boldsymbol{f}(\boldsymbol{y}) \geq \boldsymbol{f}(\boldsymbol{z}) + \nabla \boldsymbol{f}(\boldsymbol{z})'[\boldsymbol{\alpha}(\boldsymbol{x} - \boldsymbol{z}) + (\mathbf{1} - \boldsymbol{\alpha})(\boldsymbol{y} - \boldsymbol{z})] = \boldsymbol{f}(\boldsymbol{z}) + \nabla \boldsymbol{f}(\boldsymbol{z})'[\boldsymbol{\alpha x} + (\mathbf{1} - \boldsymbol{\alpha})\boldsymbol{y} - \boldsymbol{z}] = \boldsymbol{f}(\boldsymbol{z})$
- 问题1.上述定义的几何意义怎么表达?能画出图形吗?
- 问题2.当$\boldsymbol{\alpha}$变化时候,$\boldsymbol{f}(\boldsymbol{x}) + \frac{\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{\alpha}(\boldsymbol{z}-\boldsymbol{x}))-\boldsymbol{f}(\boldsymbol{x})}{\boldsymbol{\alpha}}$有何意义?(参见下页图)

# 第二章 线性规划

- 问题1：解释下图的意义




- 问题2：




这表示什么意思？

考虑$0<\alpha_1<\alpha_2<1$, 令$\bar{\alpha}=\frac{\alpha_1}{\alpha_2}$,$\bar{z}=x+\alpha_2(z-x)$, 则有：

$f\big(x+\bar{\alpha}(\bar{z}-x)\big)\le\bar{\alpha}f(\bar{z})+(1-\bar{\alpha})f(x)$即：

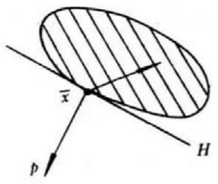
$\frac{f(x)+\bar{\alpha}(\bar{z}-x))-f(x)}{\bar{\alpha}}\le f(\bar{z})-f(x)$, 将$\bar{z}$代入：

$\frac{f(x+\alpha_1(z-x))-f(x)}{\alpha_1}\le\frac{f(x+\alpha_2(z-x))-f(x)}{\alpha_2}$,

令$g(\alpha)=\frac{f(x+\alpha(z-x))-f(x)}{\alpha}$, $\alpha\in(0,1]$, 则其单调递增！

# 第二章 线性规划

- 分离定理
- 两个凸集可以用超平面分开！直观上很显然！但由此可以推导出Farkas定理.后续 ！




- 定义：对于非空集合$C \subset R^n$,$\bar{x} \in \partial C$表示边界上的点，若有：
  - $C \subset H^+ = \{x \in C | p^T(x - \bar{x}) \geq 0$或
  - $C \subset H^- = \{x \in C | p^T(x - \bar{x}) \leq 0$
  - 则称超平面$H = \{x | p^T(x - \bar{x}) = 0\}$是$C$在$\bar{x}$处的支撑超平面。若$C$不包含在$H$中，则$H$称为$C$在$\bar{x}$的正常支撑超平面。
- 从上述凸集和性质，可以看出，凸集在每个边界点都有一个切平面支撑。
- 两个凸集$C_1, C_2$,若$p^T x \geq \alpha, \forall x \in C_1, p^T x \leq \alpha, \forall x \in C_2$，则称超平面$H = \{x | p^T x = \alpha\}$分离$C_1, C_2$.若都是不等号，则称为严格分离。
- 分离定理：两个不相交凸集存在分离超平面，即存在非量向量$p \in R^n$,使得：

$$p^T x_1 \leq p^T x_2, \forall x_1 \in cl(C_1), \forall x_2 \in cl(C_2)$$

- 并可经一步推广到强分离定理：

$$inf\{p^T x | x \in C_1\} \geq \epsilon + sup\{p^T x | x \in C_2\}, p \neq 0, \epsilon > 0$$

# 第二章 线性规划

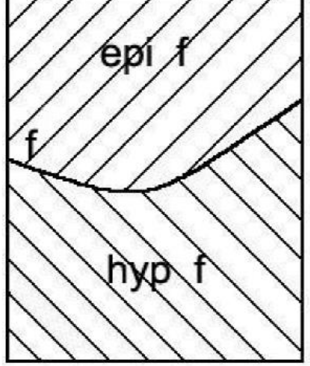
- 典型的凸函数，凸函数由于具有很好的性质，在实际应用中颇为广泛
- 问题1.函数$\boldsymbol{f(x)=x^p}$,若$\boldsymbol{x>0,0<p<1}$,函数$\boldsymbol{f(x)}$的凸凹性如何？若$\boldsymbol{x>0,p>1}$呢？
- 问题2.尝试证明：Minkowshi不等式：$\boldsymbol{||x+y||_p\le||x||_p+||y||_p}$,$\boldsymbol{p\ge 1}$
- 问题3.令$\boldsymbol{M_\alpha(x)=\left(\frac{x_1^\alpha+x_2^\alpha+\cdots+x_n^\alpha}{n}\right)^{\frac{1}{\alpha}}}$,$\boldsymbol{\alpha\ge 0}$,则$\boldsymbol{M_\alpha(x)}$是$\boldsymbol{\alpha}$的减函数。
- 问题4.$\boldsymbol{f(x)=lnx,x\in(0,+\infty)}$是(0,+∞)上的凸函数还是凹函数？$\boldsymbol{f(x)=xlnx(0<x<+\infty)}$呢？

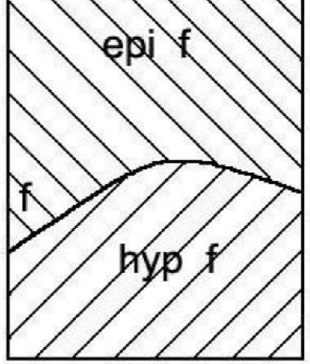
# 第二章 线性规划

- 很多著名不等式都可以采用类似的办法
  - Young不等式：$x,y>0,p,q>1,\frac{1}{p}+\frac{1}{q}=1,x^{\frac{1}{p}}y^{\frac{1}{q}}\le\frac{x}{p}+\frac{y}{q}$
  - Holder不等式：$\alpha>1,\beta>1,\frac{1}{\alpha}+\frac{1}{\beta}=1,a_i>0,b_i>0,i=1,2,\cdots,n$,则

    有：$\sum_{i=1}^{n}a_ib_i\le(\sum_{i=1}^{n}a_i^{\alpha})^{\frac{1}{\alpha}}\left(\sum_{i=1}^{n}b_i^{\beta}\right)^{\frac{1}{\beta}}$,且$a_i^{\alpha},b_i^{\beta}$成正比例时等号成立。

# 第二章 线性规划




- 水平集(level set)
  - 函数$\boldsymbol{f}(\boldsymbol{x})$在集合$\boldsymbol{S}$上关于实数$\boldsymbol{\alpha}$的水平集定义为:
    $$\boldsymbol{L}(\boldsymbol{f},\boldsymbol{\alpha})=\{\boldsymbol{x}\in\boldsymbol{S}|\boldsymbol{f}(\boldsymbol{x})\le\boldsymbol{\alpha}\}$$
  - 这里$\boldsymbol{S}\subset\boldsymbol{R}^{\boldsymbol{n}},\boldsymbol{f}:\boldsymbol{R}^{\boldsymbol{n}}\to\boldsymbol{R}$.并且函数$\boldsymbol{f}(\boldsymbol{x})$是凸集$\boldsymbol{S}\subset\boldsymbol{R}^{\boldsymbol{n}}$上的凸函数时，水平集也是凸集
- 设$\boldsymbol{f}(\boldsymbol{x})$是凸集合$\boldsymbol{S}\subset\boldsymbol{R}^{\boldsymbol{n}}$上的凸函数，则对$\forall\boldsymbol{\alpha}\in\boldsymbol{R}^{\mathbf{1}}$,水平集$\boldsymbol{L}(\boldsymbol{f},\boldsymbol{\alpha})$是凸集。
- 函数$\boldsymbol{f}$的上图或上镜图(Epigraph):
  $$\boldsymbol{epi}(\boldsymbol{f})=\left\{\begin{pmatrix}\boldsymbol{x}\\\boldsymbol{\alpha}\end{pmatrix}|\boldsymbol{f}(\boldsymbol{x})\le\boldsymbol{\alpha},\boldsymbol{x}\in\boldsymbol{S},\boldsymbol{\alpha}\in\boldsymbol{R}^{\mathbf{1}}\right\}$$
- 问题1.函数$\boldsymbol{f}(\boldsymbol{x})$在凸集$\boldsymbol{S}$上是凸函数，等价于$\boldsymbol{f}$的上镜图$\boldsymbol{epi}(\boldsymbol{f})$是$\boldsymbol{R}^{\boldsymbol{n}+\mathbf{1}}$上的凸集，是否正确？(尝试证明)

# 第二章 线性规划

- 设$\boldsymbol{f(x)}$定义在凸集$\boldsymbol{S \subseteq R^n}$上， $\boldsymbol{x, y \in S, x \neq y}$.令$\boldsymbol{\phi(t) = f(tx + (1-t)y), t \in [0, 1]}$.则$\boldsymbol{f(x)}$是$\boldsymbol{S}$上的凸函数的充要条件为：对$\forall \boldsymbol{x, y \in S, x \neq y}$,一元函数$\boldsymbol{\phi(t)}$是$\boldsymbol{[0, 1]}$上的凸函数。
- 问题1.上述论断是否成立，为什么？

若：$C=\{(x_1,x_2)|x_1^2+(x_2-1)^2\le 1\}$
问：$cone(C)=?$

# 第二章 线性规划

- 非空集合$\boldsymbol{C}\subset\boldsymbol{R}^n$的元素的非负组合为：

$\boldsymbol{x_i}\in\boldsymbol{C},\boldsymbol{\alpha_i}\ge\boldsymbol{0},\sum_{i=1}^{m}\boldsymbol{\alpha_i x_i}$,若所有$\boldsymbol{\alpha_i}$为正的，则称为正组合。

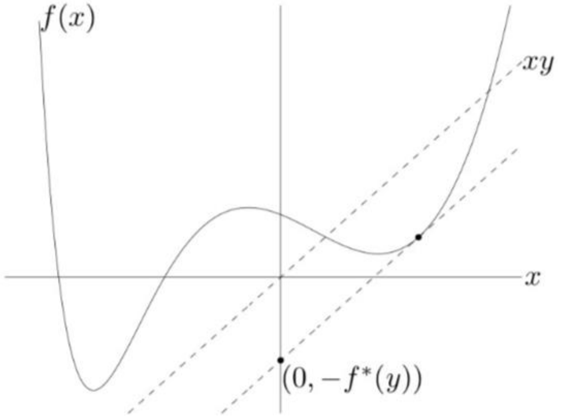
- 定义：由集合$\boldsymbol{C}$所形成的锥定义为：集合$\boldsymbol{C}$的元素所形成的所有非负组合集合，表示为$\boldsymbol{cone(C)}$.注意锥是凸的！【所有包含集合$\boldsymbol{C}$的凸集的交称为集合$\boldsymbol{C}$的凸包:convex hull】
- 定理：非空子集$\boldsymbol{C}\subset\boldsymbol{R}^n$，$\boldsymbol{cone(X)}$中的任意非零向量都可以表示为$\boldsymbol{C}$中线性无关向量的正组合.并且锥中的任意向量都可以表示为$\boldsymbol{C}$中不超过$\boldsymbol{n+1}$个向量的凸组合！
- 问题1.如何证明这条定理？
  - $\boldsymbol{x\neq 0,\in cone(C)}$,令$\boldsymbol{m}$是满足：$\boldsymbol{x=\sum_{i=1}^{m}\alpha_i x_i,\alpha_i>0,x_i\in C,i=1,\cdots,m}$的最小正整数。反证法！若向量$\boldsymbol{x_i}$线性相关，则存在$\boldsymbol{\lambda_i},\sum_{i=1}^{m}\boldsymbol{\lambda_i x_i}$线性相关，且至少一个$\boldsymbol{\lambda_i>0}$.则线性组合：$\boldsymbol{\sum_{i=1}^{m}(\alpha_i-\bar{\gamma}\lambda_i)x_i}$,其中$\boldsymbol{\bar{\gamma}=min_i\left\{\frac{\alpha_i}{|\lambda_i|}\right\}}$,提供了一个$\boldsymbol{m-1}$个元素的正组合，从而矛盾。注：$\boldsymbol{\alpha_i-\gamma\lambda_i>0}\Rightarrow\begin{cases}\boldsymbol{\gamma<\frac{\alpha_i}{\lambda_i},\lambda_i>0}\\\boldsymbol{\gamma>\frac{\alpha_i}{\lambda_i},\lambda_i<0}\end{cases}\Rightarrow$

    取其交集!
  - 将上述定理应用到子空间：$\boldsymbol{Y=\{(y,1)|y\in C\}}$上即可！

定理的作用是什么？**能将任意非凸集凸化为凸集，而通过凸化凸函数的上镜图，则可以将非凸函数转化为凸函数！**

# 第二章 线性规划

- 都说凸集、凸函数有很好的性质，但实际上大部分集合和函数都不是凸的？
- 问题1.如何将非凸集合或函数，转换为凸集或凸函数呢？
- 问题2.听说过共轭函数吗？什么叫共轭函数？Conjugate function: Fenchel conjugate, 如果$\boldsymbol{f}$可微，又称为Legendre 变换
  - 令$\boldsymbol{f}:\boldsymbol{A}\to\boldsymbol{R}$是定义在子集$\boldsymbol{A}\subset\boldsymbol{R}^{\boldsymbol{n}}$上的一个函数。其共轭函数$\boldsymbol{f}^*$:$\boldsymbol{R}^{\boldsymbol{n}}\to\boldsymbol{R}$定义为：
  - $\boldsymbol{f}^*(\boldsymbol{y})=\boldsymbol{sup}_{\boldsymbol{x}\in\boldsymbol{A}}\left(\boldsymbol{y}^{\boldsymbol{T}}\boldsymbol{x}-\boldsymbol{f}(\boldsymbol{x})\right),\boldsymbol{y}\in\boldsymbol{R}^{\boldsymbol{n}}$
  - 例子：假设$\boldsymbol{f}=\boldsymbol{ax}+\boldsymbol{b},\boldsymbol{x},\boldsymbol{a},\boldsymbol{b}\in\boldsymbol{R}$,则$\boldsymbol{f}^*(\boldsymbol{y})=\begin{cases}-\boldsymbol{b},\boldsymbol{y}=\boldsymbol{a}\\+\infty,\boldsymbol{otherwise}\end{cases}$



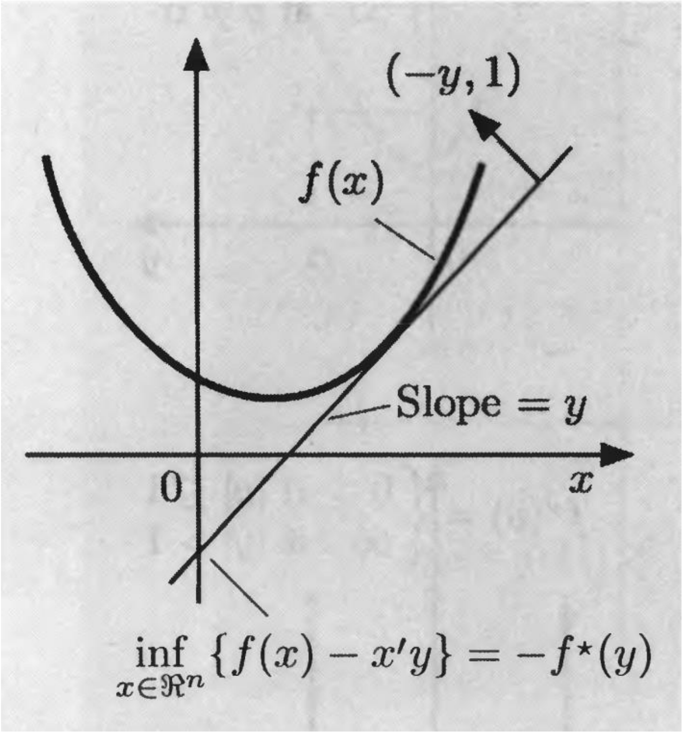
# 第二章 线性规划

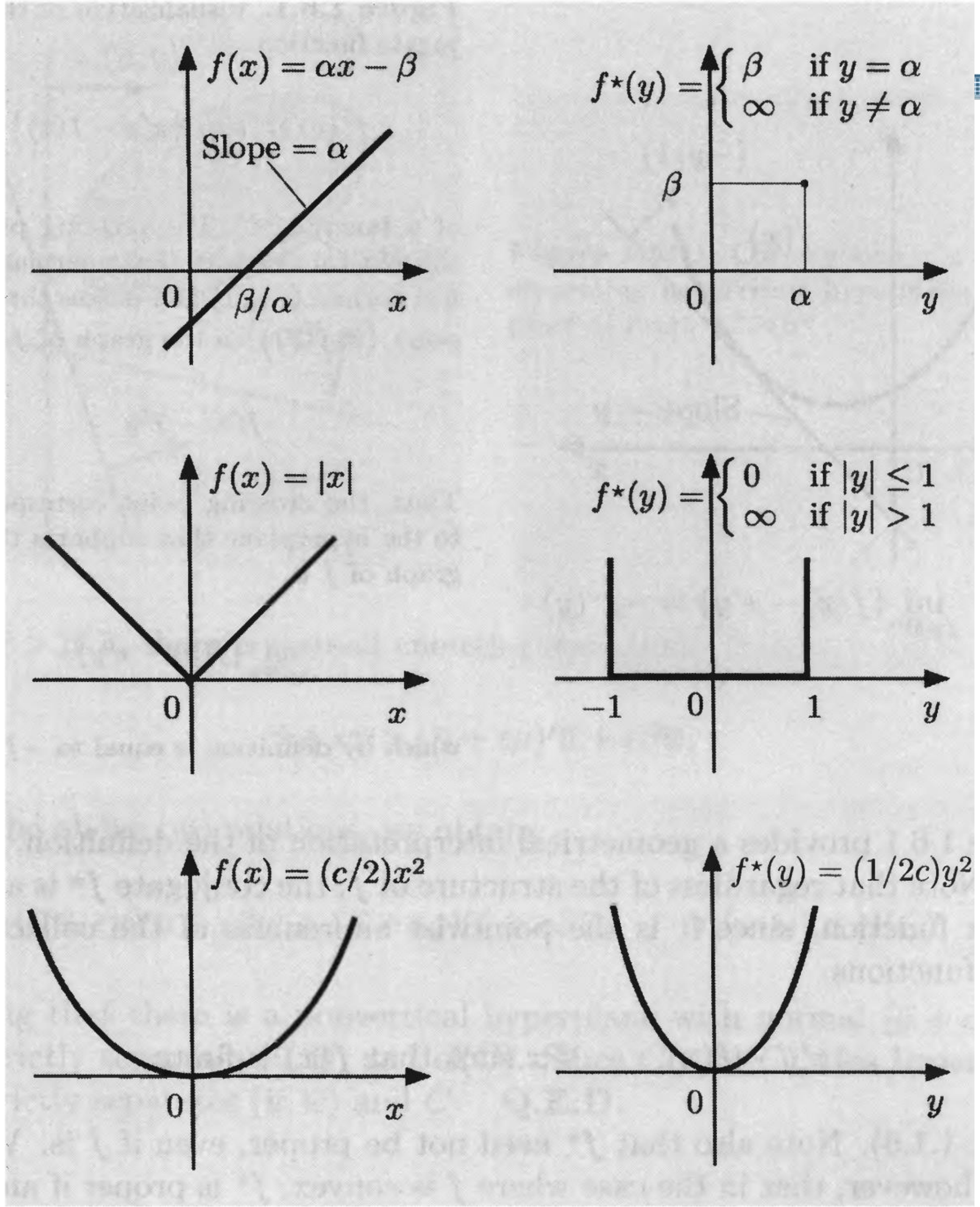
- 共轭函数的计算
  - $\boldsymbol{f(x) = -logx, x \in R, x > 0}$
  - $\boldsymbol{f(x) = e^{x}, x \in R}$
  - $\boldsymbol{f(x) = xlogx, x \geq 0}$
  - $\boldsymbol{f(x) = \frac{1}{2}x^{T}Ax, A > 0}$, 对称， $\boldsymbol{x \in R^{n}}$
  - $\boldsymbol{f(X) = logdet(X^{-1}), X}$对称正定矩阵
  - $\boldsymbol{f(x) = ||x||, dom(f) = R^{n}}$, $||\cdot||$为任意范数
- 问题1.如何计算共轭函数?
- 问题2.从几何上如何理解共轭函数?

$$f^{*}(y) = -\log(-y) - 1, y \in R, y < 0$$
$$f^{*}(y) = ylogy - y, y \geq 0, 0log0 = 0$$
$$f^{*}(y) = e^{y-1}, y \in R$$
$$f^{*}(y) = -\frac{1}{2}y^{T}A^{-1}y$$
$$f(Y) = logdet\left((-Y)^{-1}\right) - n$$

# 第二章 线性规划

- 几何直观解释



# 第二章 线性规划

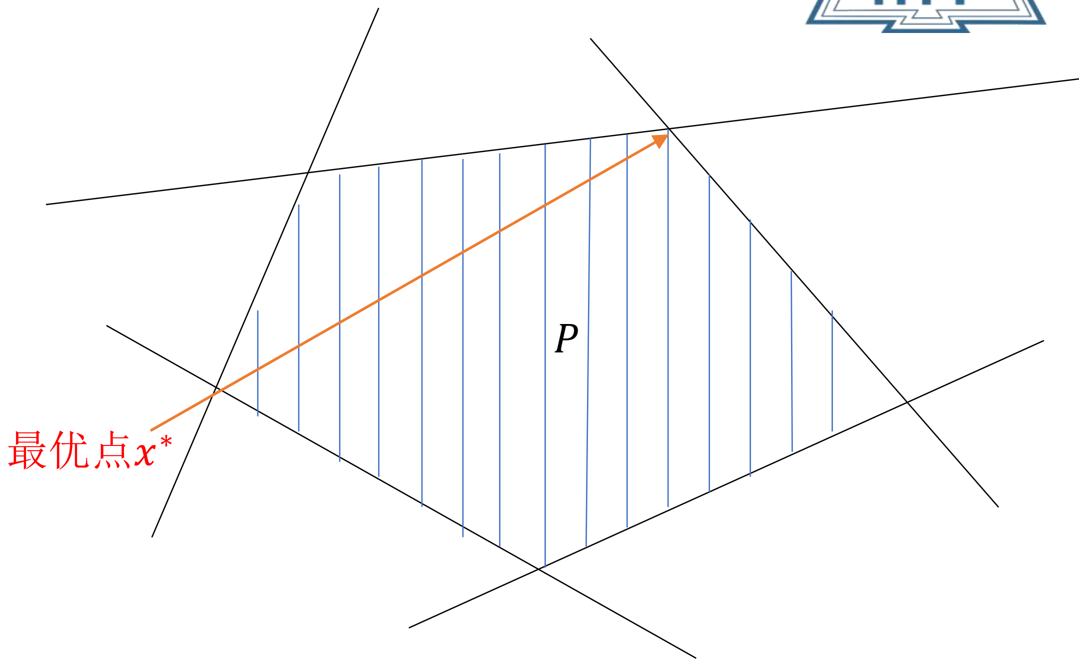
- 线性规划和整数规划有什么联系？
  - $\boldsymbol{A} \in \mathbb{Z}^{\boldsymbol{m}\times\boldsymbol{n}}$, 且$\boldsymbol{b} \in \mathbb{Z}^{\boldsymbol{m}}, \boldsymbol{c} \in \mathbb{Z}^{\boldsymbol{n}}$， 求$\boldsymbol{x} \in \mathbb{Z}^{\boldsymbol{n}}$, 使得$\boldsymbol{Ax} \le \boldsymbol{b}, \boldsymbol{max\ cx}$
- 几乎所有的组合优化问题都可以形式化表达为整数规划问题
  - 可行解表示为$\{x: Ax \le b, x \in \mathbb{Z}^n\}$，集合$\boldsymbol{P} \coloneqq \{\boldsymbol{x} \in \mathbb{R}^{\boldsymbol{n}}: \boldsymbol{Ax} \le \boldsymbol{b}\}$为多面体，令$\boldsymbol{P_I} \coloneqq \{\boldsymbol{x} \in \mathbb{R}^{\boldsymbol{n}}: \boldsymbol{Ax} \le \boldsymbol{b}\}$为P中整数向量的凸包(convex hull),称为P的整数凸包
- 命题：令$P = \{x: Ax \le b\}$是整数凸包非空的有理多面体，令$c$为某一向量（不一定是有理数）。则$\max\{cx: x \in P\}$有界$\Leftrightarrow \max\{cx: x \in P_I\}$是有界的。
- 割平面法求解：$P \supset P' \supset P_I$,期望$\max\{cx: x \in P'\}$在整数向量上达到

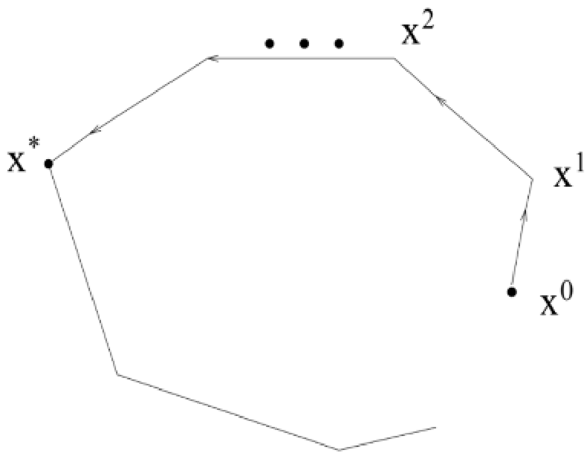
# 第二章 线性规划内点法

- 线性意味着退化，如果用一阶或二阶方法，一阶导数为常数，二阶为0，因此一般的最优化理论的微分法无法直接使用！
- 从前面求解可以看出，这也是一个组合优化问题，就是从初始点到最优点的路径遍历，但顶点数目大的时候，有 $\frac{n!}{m!(n-m)!}$ 种可能，穷举也不可行！
- 单纯形法则限制在顶点上，利用变量的连续性质来从一个顶点到另一个顶点去逐步优化目标函数值，在大约50年中一直作为求解线性规划的有效方法！

# 第二章 线性规划内点法

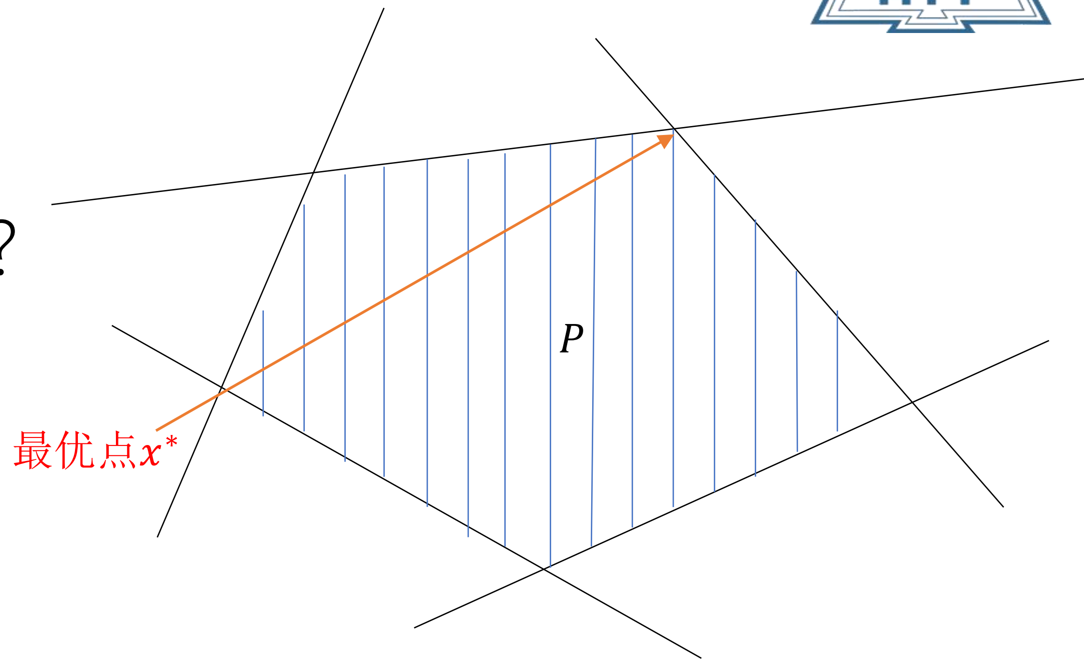
- 解LP问题
  - 如何去找到最优点$x^*$?
- 单纯形法（Simplex method）
  - Step1. 从初始点$x^0$出发;
  - Step2.判断当前点$x^k$是否是最优点
  - 是，则停止,$x^k$即是最优点$x^*$;否，则继续下一步
  - Step3.按照准则，找下一个领域更优顶点$x^{k+1}$,
  - 转Step2.

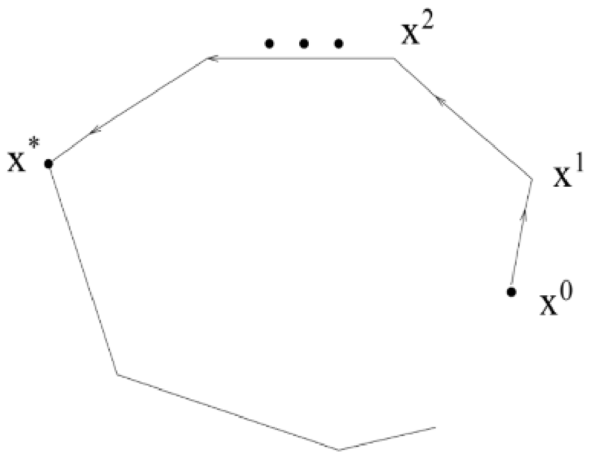


# 第二章 线性规划内点法

- 单纯形法（Simplex method）是否是最好的？
- 是否是唯一的方法？
  - 一般来说，需要访问顶点的数据大约是：
  - $0.7159m^{0.9522}n^{0.3109}$个（大约按照m为线性，n为次线性）
  - 最差的情况，Klee,Minty(1971)，需要访问$2^n-1$个顶点（指数复杂性）
  - 大规模问题因此需要很长的时间来运行！
- 内点法的思想
  - 停留在可行域内部，检查更多的移动方向，缩短遍历的路径
    - 例如，通过增加每次迭代的复杂性，但减少迭代次数等



# 第二章 线性规划椭球法

- 1972年，Klee-Minty提出的一类LP问题，每个采用单纯形求解的迭代次数都是指数次迭代
- 如：$maximize\ \sum_{j=1}^{n} 10^{n-j} x_j$ ,
  - $s.t.\ 2\sum_{j=1}^{i-1} 10^{i-j} x_j + x_i \leq 100^{i-1}, i = 1, \cdots, n$
  - $x_j \geq 0, j = 1, \cdots, n$
- 对于$n = 3$，其标准形式为：
- $maximize\ 100x_1 + 10x_2 + x_3$
- S.t. $\begin{cases} x_1 \leq 1 \\ 20x_1 + x_2 \leq 100 \\ 200x_1 + 20x_2 + x_3 \leq 10000 \\ x_1 \geq 0, x_2 \geq 0, x_3 \geq 0 \end{cases}$
- 一般来说，这类问题都需要顶点数-1次迭代。$n = 50, 2^{50} - 1 \approx 10^{15}$,每秒百万次计算能力也需要大约33年，但已经证明解具有固定折扣率的MDP问题是多项式的。

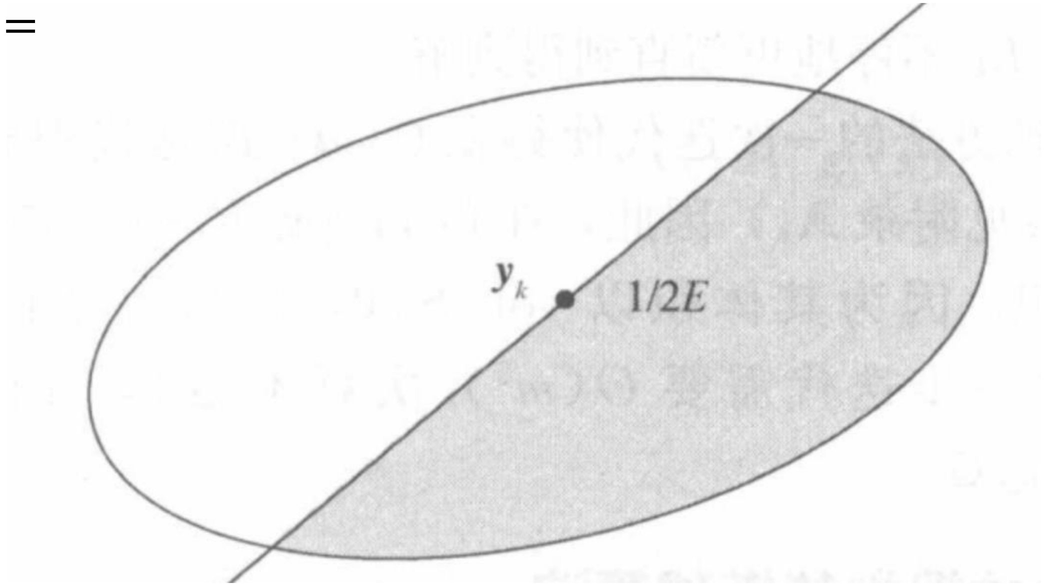
# 第二章 线性规划椭球法

- **椭球算法**：本质就是在越来越小的椭球上包围感兴趣的区域。哈奇扬的巨大贡献在于证明了在某种假设下，椭球算法构成线性规划的一个多项式时间算法
- 旨在找到由线性不等式组$\Omega = \{y \in R^m : y^T a_j \le c_j, j = 1, \cdots, n\}$定义的多面体集合$\Omega$的一个点。求解$\Omega$的一个点可认为等价于求解一个线性规划问题。
  - 假设1：$\exists y_0 \in R^m, R > 0$,使得以$y_0$为中心，以$R$为半径的闭球$S(y_0, R)$，即$\{y \in R^m | |y - y_0| \le R\} \supseteq \Omega$
  - 假设2：若$\Omega$非空，$\exists r > 0$,使得$\Omega \supseteq S(y^*, r)$。这表明若$\Omega$非空，则存在非空内部，并且其提及至少为$vol\big(S(0, r)\big)$
- 定义$R^m$中的椭球是形为：$E = \{y \in R^m : (y - z)^T Q (y - z) \le 1\}$的集合，这里$z \in R^m$为中心，$Q$为$m \times m$的正定矩阵，椭球记为$E(z, Q)$.
  - 以原点为中心的单位球$S(0,1)$表示$Q = I$（单位矩阵）的特殊椭球。
  - 一般椭球的轴为矩阵$Q$的特征向量，其轴长度记为$\lambda_1^{-\frac{1}{2}}, \cdots, \lambda_m^{-\frac{1}{2}}$,其中$\lambda_i$为矩阵的特征值。椭球体积为$vol(E) = vol\big(S(0,1)\big)\Pi_{i=1}^m \lambda_i^{-\frac{1}{2}} = vol\big(S(0,1)\big)\det\big(Q^{-\frac{1}{2}}\big)$

# 第二章 线性规划椭球法

- **椭球算法**：割平面以及包含新的椭球
- 椭球算法定义一系列以$y_k$为中心并且$Q = B_k^{-1}$的椭球$E_k$，$B_k$是对称正定矩阵。
- 算法的每步迭代中都有$\Omega \subset E_k$。因此验证$y_k \in \Omega$是否属于$\Omega$是可行的。若是，则已经找到$\Omega$中所要求的元素。若不是，则说明至少一个约束条件不成立。假设$a_j^T y_k > c_j$，则$\Omega \subset \frac{1}{2} E_k = \{y \in E_k | a_j^T y \le a_j^T y_k\}$,这个集合是半个椭球，通过从中心将椭球一分为二。
  - 下一个椭球$E_{k+1}$为包含$\frac{1}{2}E_k$的最小的椭球，构造如下：令$\tau = \frac{1}{m+1}, \delta = \frac{m^2}{m^2-1}, \sigma = 2\tau,$
  - 设$y_{k+1} = y_k - \frac{\tau}{\left(a_j^T B_k a_j\right)^{\frac{1}{2}}} B_k a_j$；$B_{k+1} = \delta\left(B_k - \sigma \frac{B_k a_j a_j^T B_k}{a_j^T B_k a_j}\right)$
  - **定理**：如上定义的椭球是包含$\frac{1}{2}E_k$的最小椭球，而且

$$\frac{vol(E_{k+1})}{vol(E_k)} = \left(\frac{m^2}{m^2-1}\right)^{\frac{m-1}{2}} \frac{m}{m+1} < \exp\left(-\frac{1}{2(m+1)}\right) < 1$$



# 第二章 线性规划椭球法

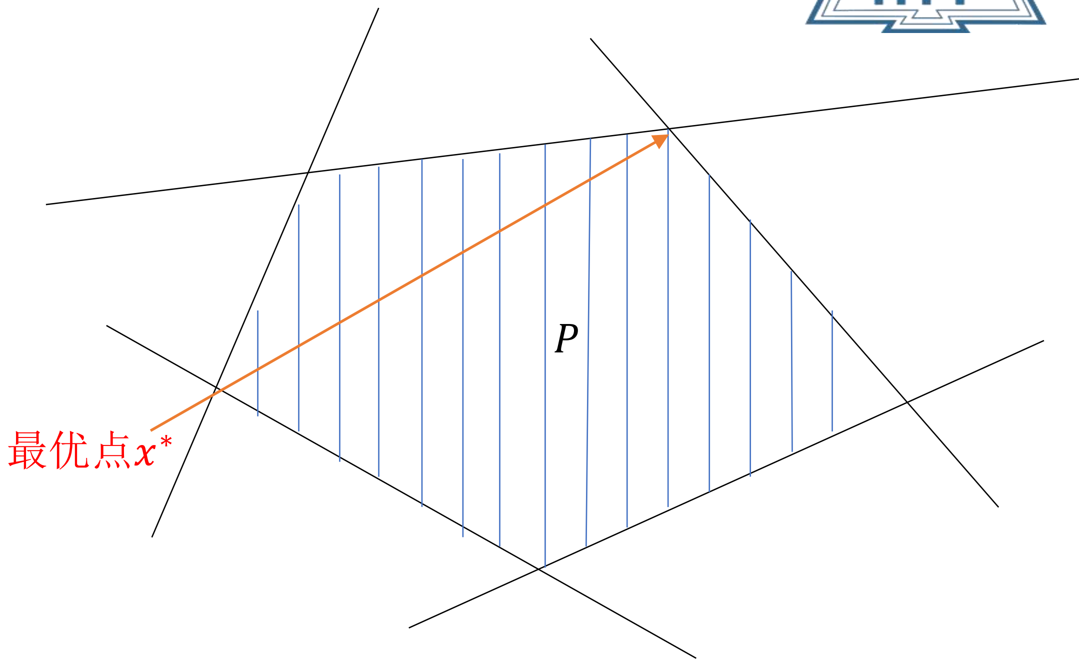
- **证明**：$\frac{vol(E_{k+1})}{vol(E_k)} = \frac{\det\left(B_{k+1}^{\frac{1}{2}}\right)}{\det\left(B_k^{\frac{1}{2}}\right)}$,通过坐标变换设$B_k = I$,则$B_{k+1}$有$m-1$个等于$\delta = \frac{m^2}{m^2-1}$的特征值以及一个等于$\delta - 2\delta\tau = \frac{m^2}{m^2-1}\left(1-\frac{2}{m+1}\right) = \left(\frac{m}{m+1}\right)^2$的特征值。体积的减少量为它们的平方根的乘积，于是可得定理中的等式。
- 然后应用$(1+x)^p \le e^{xp}$,可得$\left(\frac{m^2}{m^2-1}\right)^{\frac{m-1}{2}}\frac{m}{m+1} = \left(1+\frac{1}{m^2-1}\right)^{\frac{m-1}{2}}\left(1-\frac{1}{m+1}\right) < \exp\left(\frac{1}{2(m+1)} - \frac{1}{m+1}\right) = \exp\left(-\frac{1}{2(m+1)}\right)$
- **收敛性**：初始选择满足假设1的$y_0, R$。接下来设$B_0 = R^2 I, E_0 \supset \Omega$,不断更新$E_k$直到得到解。显然，椭球方法一次迭代使得在$O(m)$次迭代中椭球的体积减少为初始值的一半。因此在$O\left(m^2 \log\left(\frac{R}{r}\right)\right)$次迭代后体积减少为小于半径为$r$的球的体积，因为其体积以$vol(S(0,1))r^m$为下界，且初始体积为$vol(S(0,1))R^m$。通常一步迭代需要$O(m^2)$的算术运算。因此，整个过程需要$O\left(m^4 \log\left(\frac{R}{r}\right)\right)$次算术运算。

# 第二章 线性规划椭球法

- 一般形式的线性规划椭球算法
- $LP$: $\begin{cases} \max c^T x \\ s.t. Ax \leq b, x \geq 0 \end{cases}$ 及其对偶$DLP$: $\begin{cases} \min y^T b \\ s.t. y^T A \geq c^T, y \geq 0 \end{cases}$
- 注意到求不等式: $\begin{cases} -c^T x + b^T y \leq 0 \\ \mathrm{Ax} \leq b \\ -A^T y \leq -c \\ \mathrm{x}, y \geq 0 \end{cases}$ 的可行解，可得上述两个问题的解，$x, y$为变量。因此，求解线性规划的所有算术运算的总和以$O\left((m+n)^4 \log\left(\frac{R}{r}\right)\right)$为界。

# 第二章 线性规划内点法

- Interior-Point法
  - Step1.从一个初始内部可行解开始;
  - Step2.如果当前解是最优解，停止！否则下一步;
  - Step3.移动到更好的内点，转到step2
    - 更好的方向?
    - 恰当的步长?
- Projective scaling 方法
- Affine scaling 方法
  - Primal affine scaling 算法
  - Dual affine scaling算法
  - Primal-Dual 算法
- Potential reduction方法
- Path-Following 方法




内点法要避免可行集的边缘，只是最终收敛到某个顶点！

# 第二章 线性规划内点法

- **内点法**不关注边界，从可行域内部迭代最后收敛到边界上的顶点。
- **分析中心（analytic center）**用来描述集合的内部，远离边界的区域。然后引入特殊的函数，称为壁垒或势（barrier/potential）。
- 考虑一系列不等式定义的$R^n$的子集$\mathcal{X}$中的集合$S=\{x\in\mathcal{X}|g_j(x)\geq 0, j=1,2,\cdots,m\}$，并且假设函数$g_j$是连续的。$S$有非空内域$S_{int}=\{x\in\mathcal{X}|g_j(x)>0,\forall j\}$。定义集合上$S_{int}$上的势函数(potential function)为：

$$\psi(x)=-\sum_{j=1}^{m}\log\left(g_j(x)\right)$$

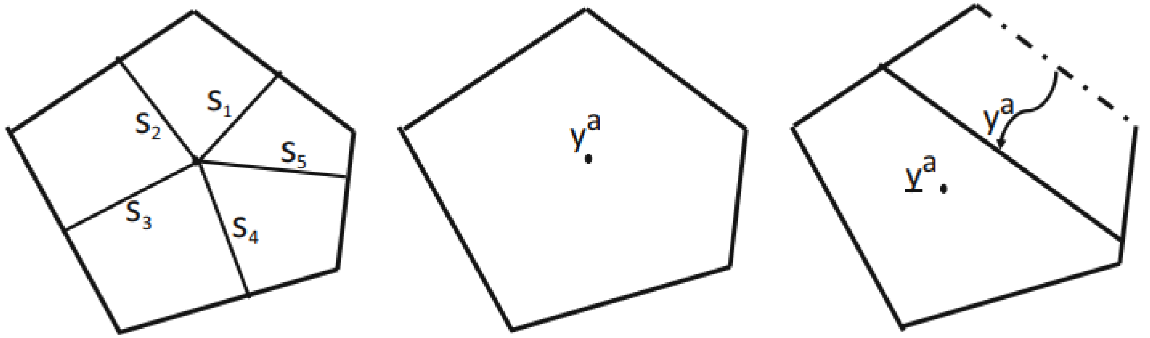
- 集合的分析中心是最小化势的向量或向量集；即满足：
- $min\psi(x)=\min\left\{-\sum_{j=1}^{m}\log\left(g_j(x)\right)|x\in\mathcal{X},g_j(x)>0,\forall j\right\}$
- 例：考虑由$x_i\geq 0,(1-x_i)^d.\geq 0, i=1,2,\cdots,n$定义的集合$S$。当$d=1$，这里$S=[0,1]^n$,即$R^n$中的立方体，通过求导可以确定分析中心为:对所有$i,x_i=\frac{1}{2}$。因此，分析中心等同于单位立方体的中心。
- 分析中心依赖于集合的定义，并且重复增加不等式约束也会改变分析中心。如果奇数$d>1$，解$x_i=\frac{1}{d+1}$,对于比较大的$d$,该点靠近单位立方体的内部角点。

# 第二章 线性规划内点法

- 考虑通过与由$n(>m)$个线性不等式表示的$R^m$中的有界多面体相关的分析中心，即$\Omega=\{y\in R^m|c^T-y^TA\geq 0\}, A\in R^{m\times n}, c\in R^n, Rank(A)=m, \Omega_{int}=\{y\in R^m|c^T-y^TA>0\}$
- 该集合的势函数为：$\psi_\Omega(y)\equiv -\sum_{j=1}^{n}\log\left(c_j-y^Ta_j\right)=-\sum_{j=1}^{n}logs_j$
- 这里$s\equiv c-A^Ty$是一个松弛向量。因此势函数是松弛变量的对数和的相反数。
- $\Omega$的分析中心是$\Omega$的最小化势函数的内点。记这个点为$y^a$,并且$s^a=c-A^Ty^a$。点对$(y^a,s^a)$是唯一定义的，因为势函数在有界凸集$\Omega$上是严格凸的。
- 令$\psi(y)$关于每个$y_i$的导数为0，可得：$\sum_{j=1}^{n}\frac{a_{ij}}{c_j-y^Ta_j}=0,\forall i$,
- 可进一步写为：$\sum_{j=1}^{n}\frac{a_{ij}}{s_j}=0,\forall i$
- 现在对每个$j$定义$x_j=\frac{1}{s_j}$。若定义元素相乘：$x\circ s\equiv(x_1s_1,x_2s_2,\cdots,x_ns_n)^T$,这样分析中心由条件：

$$\begin{cases} x\circ s=1 \\ \mathrm{Ax}=0 \\ \mathrm{A}^Ty+s=c \end{cases}$$ 来定义。

# 第二章 线性规划内点法




- 当内域为空或有等式时，分析中心也可以定义如下，例如

$$\Omega = \{y \in R^m | c^T - y^T A \geq 0, By = b\}$$

- 此时，选择分析中心为在线性曲面$\{y|By = b\}$上最大化松弛变量$s = c - A^T y$的乘积的点。因此，这时，$\Omega$的内域为松弛变量的正象限的内域。内域的定义只依赖于松弛变量的区域。即使$\Omega$中只有一个点，对某些$y$来说，满足$s = c - A^T y$，其中当$s > 0$时$By = b$，我们仍然可以说$\Omega_{int}$是非空的。
- 割平面和分析体（analytic volume）的化简
  - 定义$V^a(A, c) \coloneqq \Pi_{j=1}^n s_j^a = \Pi_{j=1}^n (c_j - a_j^T y^a)$为多面体$\Omega(A, c)$的分析体，如果一个约束超平面，例如第一个，需要去平移，将$c_1 - a_1^T y \geq 0$变为$a_1^T y^a - a_1^T y \geq 0$;也就是第一个约束超平面平行向切割中心$y^a$移动；如最右边的图所示。然后得到一个由($A$，$\underline{c}$)给定的多面体，这里$\underline{c}_1$=$a_1^T y^a (< c_1)$,且$c_j$=$\underline{c}_j$, $\forall j = 2, \cdots, n$.则分析体$V^a\left(A, \underline{c}\right)$与$V^a(A, c)$相比如何？
- 定理：$\frac{V^a\left(A, \underline{c}\right)}{\mathrm{V}^a(\mathrm{A,c})} \leq \mathrm{e}^{-1} \leq \frac{1}{2.718}$.想象平移超平面表示目标超平面并持续平移到最优解，分析中心也将支付向下移动并且其轨迹行成一条路径。

# 第二章 线性规划内点法

- LP问题的内点法基于非线性规划的分析技术和方法。内点法的分析一般基于这个问题定义的函数的微分。一般非线性规划的对偶通常利用拉格朗日乘子来表示。非线性规划的计算机算法本质上就是迭代，通过搜索算法来刻画。先建立方向，然后搜索最优步长，梯度法和牛顿法都属于这一类。
- 非线性方法改进了LP的求解，而且LP的内点法得到推广为非线性规划的求解提供了新方法。
- 考虑$LP: \min c^T x, s.t. Ax = b, x \geq 0$，记其可行域为$\mathcal{F}_p$,假设其内域记为$\mathcal{F}_{p\,int} = \{x | Ax = b, x > 0\}$非空且问题的最优解集有界。
- 对$\mu \geq 0$，定义障碍问题$BP: \min c^T x - \mu \sum_{j=1}^{n} log x_j, s.t. Ax = b, x > 0$
  - 显然，当$\mu = 0$时，这个问题即原始问题。当$\mu \to \infty$，问题的解趋于可行域的分析中心（可行域有界），因为障碍值超出$c^T x$。当$\mu \to 0$时，由BP问题的解可以定义一条路径$x(\mu)$，这条路径就称为原始中心路径(primal central path)。当$\mu \to 0$时，这条路径收敛于最优面$\{x | c^T x = z^*, Ax = b, x \geq 0\}$的分析中心，这里$z^*$是LP问题的最优值。

# 第二章 线性规划内点法

- 内点法的核心思想：因此求解LP问题的策略之一就是求解当$\mu$越来越小时的BP问题，直到接近LP问题的解。
- 对任意$\mu > 0$，在LP问题假设下，唯一有界解的充要条件通过引入对应于线性等式约束的拉格朗日乘子向量$y$以形成拉格朗日函数而得到：

$$c^T x - \mu \sum_{j=1}^{n} log x_j - y^T (Ax - b)$$

- 对$x_j$求导并设为0，可得条件：

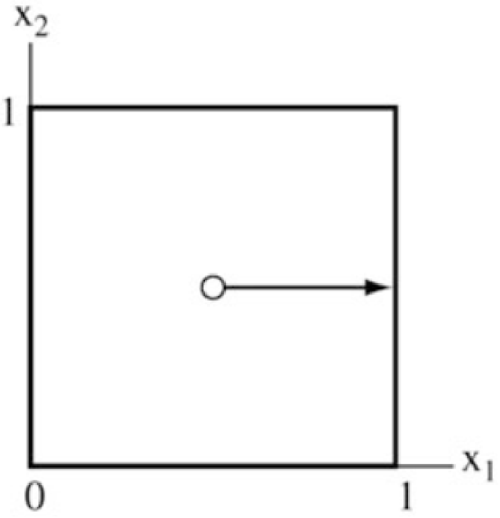
$$c_j - \frac{\mu}{x_j} - y^T a_j = 0, \forall j,$$

- 即$\mu X^{-1}\mathbf{1} + A^T y = c$,其中$a_j$表示$A$的第$j$列，$\mathbf{1}$表示所有分量都是1的向量，$X$表示对角线元素为$x > 0$的分量的对角矩阵。设$s_j = \mu / x_j$，所有条件可改写为：
- $\begin{cases} x \circ s = \mu \mathbf{1} \\ \text{Ax} = b \\ \text{A}^T y + s = c \end{cases}$，注意，$y$是一个对偶可行解并且$c - A^T y > 0$

# 第二章 线性规划内点法

- 例（正方形问题的原始问题）在单位正方形 $S=[0,1]^2$ 内最大化 $x_1$ 。形式化表达为：

$$\min -x_1, \quad s.t. \begin{cases} x_1+x_3=1 \\ x_2+x_4=1 \\ x_1\ge 0, x_2\ge 0, x_3\ge 0, x_4\ge 0 \end{cases}$$




- 这里$x_3,x_4$为松弛变量。对于$x(\mu)$的最优性条件包括原始的两个线性方程：
- $y_1+s_1=-1, y_2+s_2=0, y_1+s_3=0, y_2+s_3=0$,其中$s_j=\frac{\mu}{x_j}, j=1$

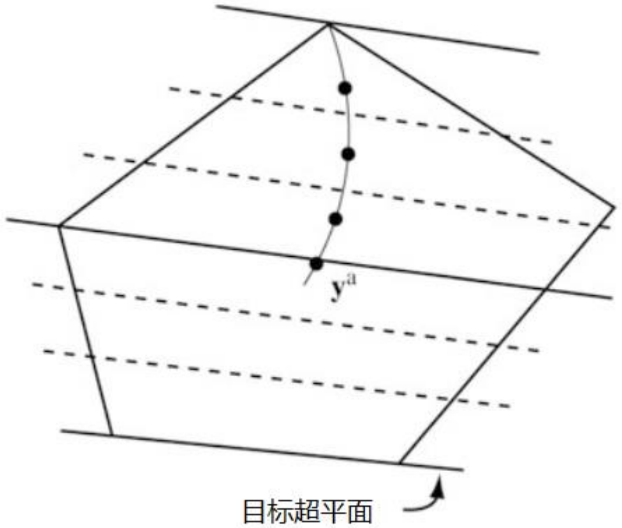
$x_1(\mu)=\frac{1-2\mu\pm\sqrt{1+4\mu^2}}{2}, x_2(\mu)=\frac{1}{2}$.从这可以看出，当$\mu\to 0, x_+\to(1,\frac{1}{2})$,+表示“+”解。注意，这个解不是立方体的角点。而是在最优面（$\{x: x_1=1,0\le x_2\le 1\}$）的分析中心。而当$\mu\to\infty$,x→$(\frac{1}{2},\frac{1}{2})$.因此，这种情况下，其中心路径是从方形区域的分析中心$(\frac{1}{2},\frac{1}{2})$到最优面的分析中心$(1,\frac{1}{2})$的一条直线。

# 第二章 线性规划内点法

- 对偶中心路径（Dual Central Path）
- 对偶问题：$(LD)\max y^Tb\,,s.t.\,y^TA+s^T=c^T,s\geq 0$
- 其障碍问题：$(BD)\max y^Tb+\mu\sum_{j=1}^{n}logs_j\,,s.t.\,y^TA+s^T=c^T,s>0$
- 假设对偶可行集$\mathcal{F}_d$有内点,集合$\overset{o}{\mathcal{F}_d}=\{(y,s):y^TA+s^T=c^T,s>0\}$非空，$(LD)$问题的最优解有界。则当$\mu\to 0$时，由$(BD)$问题的解定义了一条路径$\big(y(\mu),s(\mu)\big)$,该路径称为对偶中心路径。
- 为了求出充要条件，我们引入x作为拉格朗日乘子，形成拉格朗日对偶形式：
- $L(y,\mu,x)=y^Tb+\mu\sum_{j=1}^{n}logs_j-(y^TA+s^T-c^T)x$
- 令其对$y_i$的偏导数为0，得：
- $b_i-a^i$x=0,$\forall i,a^i$表示矩阵$A$的第$i$行，令其对$s_j$的偏导数为0，则得：
- $\frac{\mu}{s_j}-x_j=0$,或者，$1-x_js_j=0,\forall j$
- 结合这些条件得到与原中心路径$\begin{cases}x\circ s=\mu\mathbf{1}\\ \mathrm{Ax}=b\\ \mathrm{A}^Ty+s=c\end{cases}$等价的最优性条件。注意x确实是原可行解并且x> 0

# 第二章 线性规划内点法

- 再看看对偶中心路径的几何解释，考虑对偶水平集：
- $\Omega(z) = \{y: c^T - y^T A \geq 0, y^T b \geq z\}, \forall\, z < z^*, z^*$是（LD）问题的最优值。则$\Omega(z)$的分析中心$\big(y(z), s(z)\big)$与当$z \to z^*$时生成的对偶中心路径一致。如图所示的对偶几何的可行区域。水平集$\Omega(z)$对各种z值都有显示。这些水平集的分析中心对应对偶中心路径。



# 第二章 线性规划内点法

- 例（方形对偶）考虑前一个例子的对偶问题。
- $\max y_1 + y_2 \, , s.t. \begin{cases} y_1 \le -1 \\ y_1 \le 0 \\ y_2 \le 0 \end{cases}$ ,对偶障碍问题的解很容易从原障碍问题得到：$y_1(\mu) = -1 - \frac{\mu}{x_1(\mu)}, y_2(\mu) = -2\mu$
- 当$\mu \to 0$,有$y_1 \to -1, y_2 \to 0$,这是对偶LP问题的唯一解。然而，随着$\mu \to \infty, y$无界，因为这种情况下对偶可行集本身无界。

# 第二章 线性规划内点法

- 原对偶中心路径（Primal-Dual Central Path）
- 假设原问题LP的可行区域有内点，其最优解有界。则对偶问题也有内点。原对偶中心路径定义为满足下列条件的向量$(x(\mu)>0, y(\mu), s(\mu)>0)$集合$\begin{cases} x\circ s=\mu\mathbf{1} \\ \mathrm{Ax}=b \\ \mathrm{A}^T y+s=c \end{cases}$，$\forall 0\le\mu\le\infty$.
- 因此中心路径不用显示参考优化问题来定义。简单根据等式和不等式条件来定义即可。该条件跟前面介绍的条件等同，原对偶中心路径可以分裂成投影到相关空间的两个成分：
- 命题：假设原问题和对偶问题的可行集包含内点。则原对偶中心路径$(x(\mu), y(\mu), s(\mu))$对于所有的$\mu, 0\le\mu<\infty$都存在。进一步，$x(\mu)$是原中心路径，$\big(y(\mu), s(\mu)\big)$是对偶中心路径。而且$x(\mu)$和$\big(y(\mu), s(\mu)\big)$当$\mu\to 0$时，分别收敛到原问题最优解的分析中心和对偶解面的分析中心。

# 第二章 线性规划内点法

- 对偶间隙（Duality Gap）
- 令原对偶中心路径为$(x(\mu), y(\mu), s(\mu))$，则之前的条件$\begin{cases} x \circ s = \mu\mathbf{1} \\ \mathrm{A}x = b \\ \mathrm{A}^T y + s = c \end{cases}$可得：
- $c^T x - y^T b = y^T A x + s^T x - y^T b = s^T x = n\mu$
- 值$c^T x - y^T b = s^T x$是原目标函数值与对偶目标函数值的差。这个值总是非负，也就是弱对偶定理，并称之为对偶间隙，与非线性优化中的对偶间隙一致。在原对偶中心路径上的任一点，其对偶间隙都为$n\mu$，很清楚当$\mu \to 0$对偶间隙趋近于0， $x(\mu)$和$\big(y(\mu), s(\mu)\big)$当$\mu \to 0$时，分别收敛到原问题和对偶问题的最优解。
- 对偶间隙提供了一种对于接近最优值情况的度量。对于任意原可行解x，因为$c^T x \geq z^*$，因此$c^T x$提供了一个上界，这里$z^*$是对偶问题的最优值
- 类似，对于任意对偶可行对$(y, s)$，$y^T b$给出了一个下界，因为$y^T b \leq z^*$。其差值，也就是对偶间隙$g = c^T x - y^T b$为$z^*$提供了有界范围，因为$z^* \geq c^T x - g$。因此，在一个可行点x处，对偶可行点$(y, s)$也存在，x的质量可以通过$c^T x - z^* \leq g$来进行度量。

# 第二章 线性规划内点法

- 求解策略
- 中心路径的各种定义直接表明了线性规划求解的对应策略。一般有三种方法:
  - 原障碍或路径跟随方法
  - 原对偶路径跟随方法
  - 原对偶势缩减方法

| | 原可行性 | 对偶可行 | 零对偶间隙 |
|---|---|---|---|
| 原单纯形 | ✓ | | ✓ |
| 对偶单纯形 | | ✓ | ✓ |
| 原障碍函数 | ✓ | | |
| 原-对偶路径跟随 | ✓ | ✓ | |
| 原对偶势缩减 | ✓ | ✓ | |

- 原单纯形不断改进原可行解，保持零对偶间隙（互补松弛性条件）；对偶单纯形则不断改进对偶可行解，维持零对偶间隙（互补性条件）并不断移向原可行解。原障碍法不断改进原始可行解，并向对偶可行互补方向发展。原-对偶内点法不断改进原始和对偶可行解对，并向互补方向发展

# 第二章 线性规划内点法

- 使用牛顿法来求解这些最小化问题，对于目前的策略，通过考虑中心路径问题$\begin{cases} x \circ s = \mu\mathbf{1} \\ \mathrm{A}x = b \\ \mathrm{A}^T y + s = c \end{cases}$来用牛顿法求解固定$\mu$问题：
- $\min c^T x - \mu \sum_{j=1}^{n} log x_j\, , s.t.\, Ax = b, x > 0$ （**）
- 从一个给定点$x \in \overset{0}{\mathcal{F}_p}$,牛顿法在从中心路径问题的线性化版本确定的方向$d_x, d_y, d_s$上移动到一个相邻点$x^+ \in \overset{0}{\mathcal{F}_p}$,
- $\begin{cases} \mu X^{-2} d_x + d_s = \mu x^{-1}\mathbf{1} - \boldsymbol{c} \\ \mathrm{A} d_{\mathrm{x}} = 0 \\ -\mathrm{A}^T d_y - d_s = 0 \end{cases}$,注意X是对角矩阵，其对角线的值x> 0.新的点在方向$d_x$上来进行更新：$x^+ = x + d_x$.
- 注意，如对某个$s = c - A^T y, x \circ s = \mu\mathbf{1}$,则$d \overset{\text{def}}{=} (d_x, d_y, d_s) = 0$,因为当前点满足$Ax = b$，因此对$\mu$已经是中心路径解。如果某些成分小于$\mu$,则$d$将倾向于增加解来提高该成分，反之亦然。
- 该过程可以重复，知道解足够接近障碍问题的解，也就是前面的充要条件($\mu X^{-1}1 + A^T y = c$)满足为止。

# 第二章 线性规划内点法

- 为了求解$\begin{cases} \mu X^{-2}d_x + d_s = \mu x^{-1}\mathbf{1} - \boldsymbol{c} \\ \mathrm{A}d_\mathrm{x} = 0 \\ -\mathrm{A}^T d_y - d_s = 0 \end{cases}$,两边乘上$X^2$,并乘上$A$，利用$Ad_x = 0$,则得：
- $AX^2 d_s = \mu AX\mathbf{1} - AX^2 c$,使用$d_s = -A^T d_y$,则有：
- $(AX^2A^T)d_y = -\mu AX\mathbf{1} + AX^2 c$,从而解得$d_y$。$d_s$可从第三个方程求得，最后得$d_x$，每个牛顿步的计算量为$O(nm^2 + m^3)$

# 第二章 线性规划内点法

- 原障碍函数方法
- 直接求解就是使用障碍函数构造并求解问题:
- $\min c^T x - \mu \sum_{j=1}^{n} log x_j, s.t. Ax = b, x > 0$ （**）
- 这里$\mu$是一个非常小的值。实际上，如果期望将对偶间隙减少的$\epsilon$则唯一必要的条件就是求解当 $\mu = \epsilon/n$ 时的问题。但当 $\mu$ 很小时，上面的问题可能高度病态，必要条件变得几乎奇异，因而就很难直接去求解小的$\mu$值。
- 因此，总体策略就是从适当大的$\mu$(如$\mu = 100$)开始，来近似求解该问题。对应的解是一个近似在原始中心路径上的点，但它很可能离极限对应的点（$\mu \to 0$）相当远。但作为一个比$\mu = 100$更小的$\mu$值问题作为一个初始点还是比较合适的。每次可以按照$\mu_{k+1} = \gamma\mu_k$的方式减少$\mu$，$\gamma$是小于1的正缩减因子。
- 由$\mu_k = \gamma^k \mu_0 \Rightarrow \frac{\mu_k}{\mu_0} < \epsilon \Rightarrow k = \frac{log\epsilon}{log\gamma}$，得出需要$\frac{log\epsilon}{log\gamma}$步来达到目标。

# 第三章非线性规划的数学模型

- 线性规划转换
  - 基追踪问题
  - $\boldsymbol{min_{x\in\mathbb{R}^n}||x||_1, s.t. Ax = b}$
    - 法1.引入新变量$\boldsymbol{z_i}$, $\boldsymbol{min_{(z\in\mathbb{R}^n)}\sum_{i=1}^n z_i, s.t. Ax = b, -z_i \le x_i \le z_i, i = 1, 2, \cdots, n}$
    - 法2.引入$\boldsymbol{u, v \in \mathbb{R}^n, u \ge 0, v \ge 0}$,使得$\boldsymbol{x = u - v}$,则$\boldsymbol{min_{u,v\in R^n}\sum_{i=1}^n(u + v), s.t. Au - Av = b, u, v \ge 0}$
  - 数据拟合
    - 最小二乘，最小$\boldsymbol{l_1}$范数，最小$\boldsymbol{l_\infty}$范数模型
    - $\boldsymbol{min_{x\in\mathbb{R}^n}||Ax - b||_1, min_{x\in\mathbb{R}^n}||Ax - b||_\infty}$
    - 引入变量$\boldsymbol{y = Ax - b}$,则转换为$\boldsymbol{min_{x,y\in\mathbb{R}^n}||y||_1, s.t. y = Ax - b}$
    - 引入$\boldsymbol{t = ||Ax - b||_\infty}$，等价转换为$\boldsymbol{min_{x\in R^n, t\in R}t, s.t. ||Ax - b||_\infty \le t \Leftrightarrow min_{x\in R^n, t\in R}t, s.t. -tE \le Ax - b \le tE.}$

# 第三章 非线性规划的数学模型

- 问题1.我们都说非线性规划，那你认为最简单的非线性规划是什么？
- 问题2.为什么二次规划很重要？$\boldsymbol{f(x)=\frac{1}{2}x^TAx+x^Tb+c}$,有什么实际物理意义吗？这里矩阵$\boldsymbol{A}$对称！
  - 实际上，这可以看作是一种能量的表达形式，大家回忆，在计算机中是不是很多情况都涉及能量？都是怎么来表示能量的？
  - 物理和工程中的许多问题都可以表达为能量函数的最小化！力学的基本原理就是保证能量最小化（熵增，稳态对应能量最小）
  - 最简单的能量函数就是二次函数！$\boldsymbol{f(x)=\frac{1}{2}x^TAx\pm x^Tb}$
- 问题3.在什么样的条件下，$\boldsymbol{f(x)}$有全局最小值，并唯一呢？

# 第三章 非线性规划的数学模型

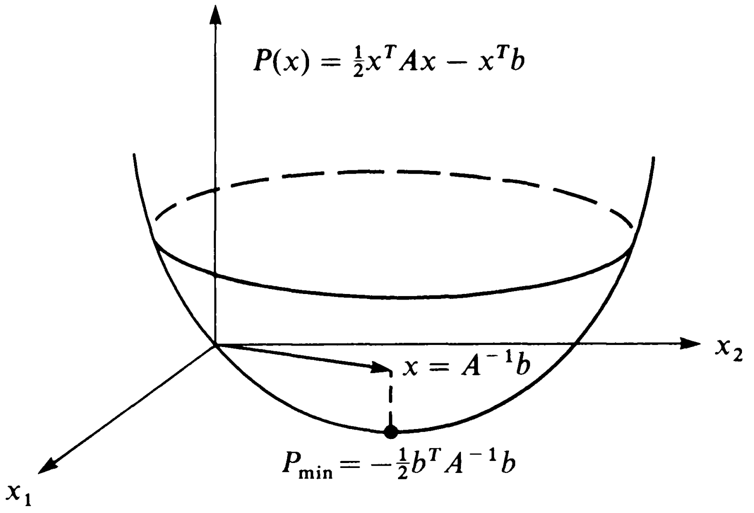
- 函数$\boldsymbol{f(x)=\frac{1}{2}x^TAx-x^Tb+c}$当矩阵$\boldsymbol{A}$是对称正定的时候，函数$\boldsymbol{f(x)}$有唯一全局极小点.
- 结论：二次函数$\boldsymbol{f(x)=\frac{1}{2}x^TAx-x^Tb+c}$中，如果$\boldsymbol{A}$是对称正定的，则$\boldsymbol{f(x)}$的全局唯一极小点是线性方程$\boldsymbol{Ax=b}$的解，$\boldsymbol{f}$的极小值为：$\boldsymbol{f\left(A^{-1}b\right)=-\frac{1}{2}b^TA^{-1}b+c}$.
- 问题1.如何证明上述结论?
  - $\boldsymbol{A}$对称正定，可逆，因此其特征值均为正，令$\boldsymbol{x=A^{-1}b}$,对$\forall \boldsymbol{y}\in \boldsymbol{R^n}$计算$\boldsymbol{f(y)-f(x)=\frac{1}{2}y^TAy-y^Tb-\frac{1}{2}x^TAx+x^Tb=\frac{1}{2}y^TAy-y^TAx+\frac{1}{2}x^TAx=\frac{1}{2}(y-x)^TA(y-x)\geq 0}$

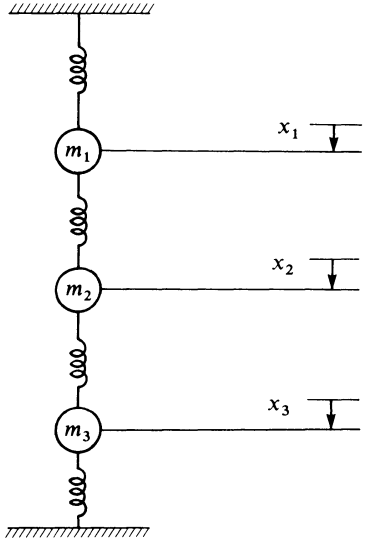
# 第三章 非线性规划的数学模型

- 上述二次函数最小值的求解等价于求解线性方程组$\boldsymbol{Ax}=\boldsymbol{b}$.这表明也可以将$\boldsymbol{Ax}=\boldsymbol{b}$表述为二次函数求极值（对应泛函求极值，采用变分法求解）
- 问题 1.假设二次函数为$\boldsymbol{Q(x_1,x_2)=\frac{1}{2}(x_1^2+x_2^2), s.t.\ 2x_1-y_1=5}$,函数$\boldsymbol{Q}$的最小值点是什么？
  - $\boldsymbol{(2,-1)}$

# 第三章 非线性规划的数学模型

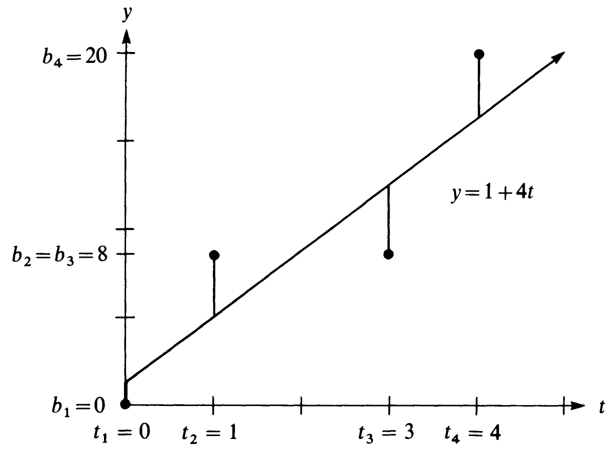
- 二次型的几何意义
  - 最小势能：四个弹簧+3个质块，三个质块的位移 $\boldsymbol{x}: x_1, x_2, x_3$ 应该是怎么样的？弹簧的延伸长度$\boldsymbol{e}$，弹簧上的作用力$\boldsymbol{y}$，以及节点上的外力$\boldsymbol{f}$.
    - $\boldsymbol{Ax = e, Ce = y, A^T y = f \Rightarrow A^T C A x = f}$,且$\boldsymbol{K = A^T C A}$是对称正定的，物理中称为刚性矩阵(stiffness matrix)
    - 弹簧的势能可以表示为：$\boldsymbol{\frac{1}{2} x^T A^T C A x}$,质块的势能可以表示为：$\boldsymbol{-x^T f}$,关于位移$\boldsymbol{x}$的系统的整体势能



# 第三章 非线性规划的数学模型

- 最小二乘解$\boldsymbol{Ax=b}$
  - $\boldsymbol{m}$个方程，$\boldsymbol{n}$个未知数，$\boldsymbol{m>n}$.超定方程!
  - 通过右端项$\boldsymbol{b}$可得：矩阵$\boldsymbol{A}$的列空间，该列空间在$\boldsymbol{m}$维子空间当中（因为所有列都只有$\boldsymbol{m}$个分量）
  - 问题1.$\boldsymbol{Ax=b}$什么时候有解？
  - 若$\begin{cases} C+t_1D=b_1 \\ C+t_2D=b_2 \\ C+t_3D=b_3 \\ C+t_4D=b_4 \\ C+t_5D=b_5 \end{cases} \Rightarrow \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix}$



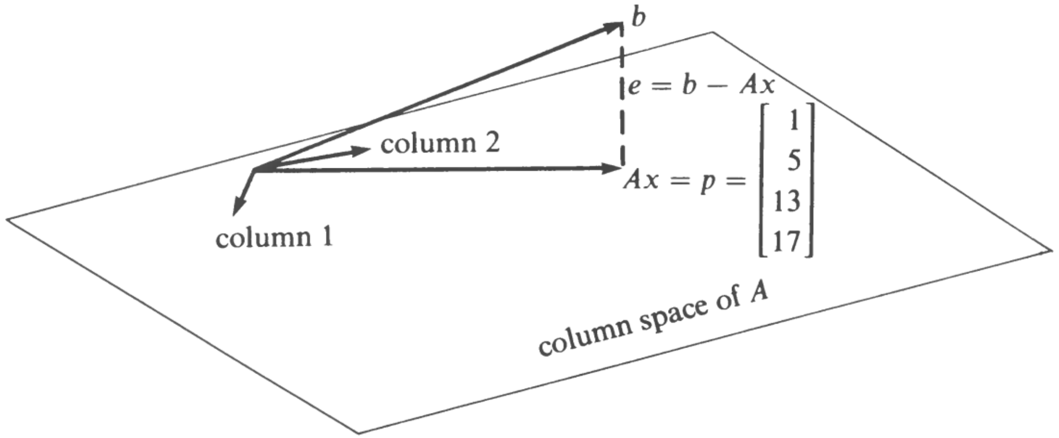

  - 问题2.这个方程的几何意义是什么?什么时候有解？

# 第三章 非线性规划的数学模型

- error误差$\boldsymbol{e}=\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}$
  - 最好的直线就是使误差$\boldsymbol{e}$尽可能小的直线
  - $\boldsymbol{min}||\boldsymbol{Ax}-\boldsymbol{b}||^{2}=(Ax-b)^{T}(Ax-b)$
  - $||\boldsymbol{e}||^{2}=\boldsymbol{e}^{T}\boldsymbol{e}=\boldsymbol{e}_{1}^{2}+\boldsymbol{e}_{2}^{2}+\cdots+\boldsymbol{e}_{n}^{2}$
  - 问题1.哪个向量$\boldsymbol{x}$**能最小化**$(\boldsymbol{Ax}-\boldsymbol{b})^{T}(\boldsymbol{Ax}-\boldsymbol{b})=\boldsymbol{x}^{T}\boldsymbol{A}^{T}\boldsymbol{Ax}-\boldsymbol{x}^{T}\boldsymbol{A}^{T}\boldsymbol{b}-\boldsymbol{b}^{T}\boldsymbol{Ax}+\boldsymbol{b}^{T}\boldsymbol{b}$?
    - 最后可化简为$\boldsymbol{P}=\frac{1}{2}\boldsymbol{x}^{T}\boldsymbol{A}^{T}\boldsymbol{Ax}-\boldsymbol{x}^{T}\boldsymbol{A}^{T}\boldsymbol{b}$,这跟之前的问题没有本质区别，只是$\boldsymbol{A}\rightarrow\boldsymbol{A}^{T}\boldsymbol{A},\boldsymbol{b}\rightarrow\boldsymbol{A}^{T}\boldsymbol{b}\Rightarrow\boldsymbol{Ax}=\boldsymbol{b}\rightarrow\boldsymbol{A}^{T}\boldsymbol{Ax}=\boldsymbol{A}^{T}\boldsymbol{b}$
    - 显然：$\boldsymbol{min}_{\boldsymbol{x}}||\boldsymbol{Ax}-\boldsymbol{b}||^{2}\Rightarrow\boldsymbol{A}^{T}\boldsymbol{Ax}=\boldsymbol{A}^{T}\boldsymbol{b}$,其解为：$\boldsymbol{A}^{T}\boldsymbol{Ax}=\boldsymbol{A}^{T}\boldsymbol{b}$
  - 向量$\boldsymbol{x}=\left(\boldsymbol{A}^{T}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{T}\boldsymbol{b}$是方程$\boldsymbol{Ax}=\boldsymbol{b}$的最小二乘解
    - 误差$\boldsymbol{e}=\boldsymbol{b}-\boldsymbol{Ax}$不为0，但与$\boldsymbol{A}$的每列向量的内积为0.$\boldsymbol{A}^{T}\boldsymbol{e}=\boldsymbol{0}$,或$\boldsymbol{A}^{T}\boldsymbol{Ax}=\boldsymbol{A}^{T}\boldsymbol{b}$,表明误差$\boldsymbol{e}$垂直于$\boldsymbol{A}$的列空间，而$\boldsymbol{b}$分解为

# 第三章 非线性规划的数学模型

- 而$\boldsymbol{b}$分解为:
- $\boldsymbol{b} = \boldsymbol{Ax} + (\boldsymbol{b} - \boldsymbol{Ax}) =$ 列空间中最近的点 $+$ 误差$\boldsymbol{error}$




- 问题1.最近的点的几何意义?
  - $\boldsymbol{p}$称为$\boldsymbol{b}$在列空间上的投影!
  - 从代数意义上：如果$\mathbf{Ax} = \mathbf{b}$无解，则$\boldsymbol{A}^{\boldsymbol{T}}(\boldsymbol{Ax}) = \boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{b}$再解方程!
  - 这就是线性回归（Linear Regression）
  - 预条件（preconditioning）

# 第三章 非线性规划的数学模型

- 回归问题的表示
  - 给定不同时刻$\boldsymbol{t_1,\cdots,t_m}$的测量值$\boldsymbol{b_1,\cdots,b_m}$,最小化误差$||\boldsymbol{e}||^2$的直线$\boldsymbol{y=C+Dt}$由方程:
  - $\boldsymbol{A^TAx=A^Tb}$ 或者$\begin{bmatrix}\boldsymbol{m} & \boldsymbol{\sum t_i}\\ \boldsymbol{\sum t_i} & \boldsymbol{\sum t_i^2}\end{bmatrix}\begin{bmatrix}\boldsymbol{C}\\ \boldsymbol{D}\end{bmatrix}=\begin{bmatrix}\boldsymbol{\sum b_i}\\ \boldsymbol{\sum t_ib_i}\end{bmatrix}$
  - 并且最佳拟合直线方程就是:
  - $\boldsymbol{y=\bar{b}+D(t-\bar{t}),D=\frac{\sum(t_i-\bar{t})b_i}{\sum(t_i-\bar{t})^2}}$,这条直线通过点中心点:$\boldsymbol{(\bar{t},\bar{b})}$
- 问题1.如果矩阵$\boldsymbol{A}$的列不独立，这时$\boldsymbol{A^TA}$就不是正定的，不是正定的怎么办？$\boldsymbol{A^TAx=A^Tb}$
  - 伪逆(pseudoinverse).实际应用中一般列都是独立的

# 第三章 非线性规划的数学模型

- Q收敛速率
  - $(\boldsymbol{\alpha},\boldsymbol{\beta})$有关，在优化中可用其刻画迭代序列的收敛速度。
- 问题1.假设有两个序列$\{\boldsymbol{x}_1^k\},\{\boldsymbol{x}_2^k\}$,其$\boldsymbol{Q}$阶和$\boldsymbol{Q}$因子分别为$(\boldsymbol{\alpha}_1,\boldsymbol{\beta}_1),(\boldsymbol{\alpha}_2,\boldsymbol{\beta}_2)$,若$\boldsymbol{\alpha}_1>\boldsymbol{\alpha}_2$，那个序列收敛快？；若$\boldsymbol{\alpha}_1=\boldsymbol{\alpha}_2,\boldsymbol{\beta}_1<\boldsymbol{\beta}_2$,这时候那个序列的收敛速度快？
- 一般主要就是Q-线性，Q-超线性，以及Q-二次收敛，通常算法的收敛速度为超线性或二次收敛，则称为具有快速收敛速度
- 问题2.拟牛顿方法的收敛速度属于哪类？牛顿法的收敛速度属于哪类？

# 第三章 非线性规划的数学模型

- R收敛（Root-convergence rate）速率
  - 比Q收敛要弱,令$\{\boldsymbol{x}^k\} \subset \boldsymbol{R}^n$为任意收敛到$\boldsymbol{x}^*$的序列。令$\boldsymbol{R_p} = \begin{cases} \boldsymbol{limsup_{k\to\infty}||x^k - x^*||^{\frac{1}{k}}}, & \boldsymbol{if\ p = 1} \\ \boldsymbol{limsup_{k\to\infty}||x^k - x^*||^{\frac{1}{p^k}}}, & \boldsymbol{if\ p > 1} \end{cases}$
  - 如果$\boldsymbol{R_1 = 0}$，则$\{\boldsymbol{x^k}\}$称为$\boldsymbol{R}$−超线性收敛到$\boldsymbol{x}^*$.
  - 如果$\boldsymbol{0 < R_1 < 1}$，则$\{\boldsymbol{x^k}\}$称为$\boldsymbol{R}$−线性收敛到$\boldsymbol{x}^*$.
  - 如果$\boldsymbol{R_1 = 1}$，则$\{\boldsymbol{x^k}\}$称为$\boldsymbol{R}$−次线性收敛到$\boldsymbol{x}^*$(sublinearly)
  - $\boldsymbol{R_2 = 0}$,则称为超平方收敛，$\boldsymbol{0 < R_2 < 1}$,R-平方收敛，$\boldsymbol{R_2 \geq 1}$，R-次平方收敛
- 问题1.如果存在非负标量序列$\{\boldsymbol{q^k}\}$,使得$\boldsymbol{||x^k - x^*|| \leq q_k}, \forall \boldsymbol{k}$,并且$\{\boldsymbol{q^k}\}$Q-线性收敛到0，则序列$\{\boldsymbol{x^k}\}$也是$\boldsymbol{R}$−线性收敛的，是否正确？实际上，在这种情况下，根据$\{\boldsymbol{q^k}\}$的Q收敛特性对应R-收敛特性！
- 问题1.$\boldsymbol{R}$ − 收敛速率依赖于R-阶 $\boldsymbol{p}$和$\boldsymbol{R}$ − 因子$\boldsymbol{R_p}$。$\boldsymbol{p}$越大，则收敛速度越大？若$\boldsymbol{R}$ − 阶$\boldsymbol{p}$相同，则$\boldsymbol{R}$ − 因子$\boldsymbol{R_p}$越小收敛速度越快？

# 第三章 非线性规划的数学模型

- 令$\mathcal{X} \subset \boldsymbol{R}^n$,$\boldsymbol{min f(x)}$, $\boldsymbol{x} \in \mathcal{X}$,迭代法的基本方式
  - $\boldsymbol{x^{k+1} = x^k + \lambda_k p_k}$
- 如果$\boldsymbol{x}^*$为凸规划问题的最优解，则表明从该点出发的任何方向都不是可行下降方向
  - $\boldsymbol{S_d(x) = \{d \in R^n | d^T \nabla f(x) < 0\}}$ =下降方向集合
  - $\boldsymbol{S_f(x) = \{d \in R^n | d = x' - x, \forall x' \in \mathcal{X}\}}$=可行方向集合
- 上述最优性条件表示为:
  - $\boldsymbol{x}^* \in \mathcal{X} \Leftrightarrow \boldsymbol{S_f(x^*)} \cap \boldsymbol{S_d(x^*)} = \emptyset$

变分不等式

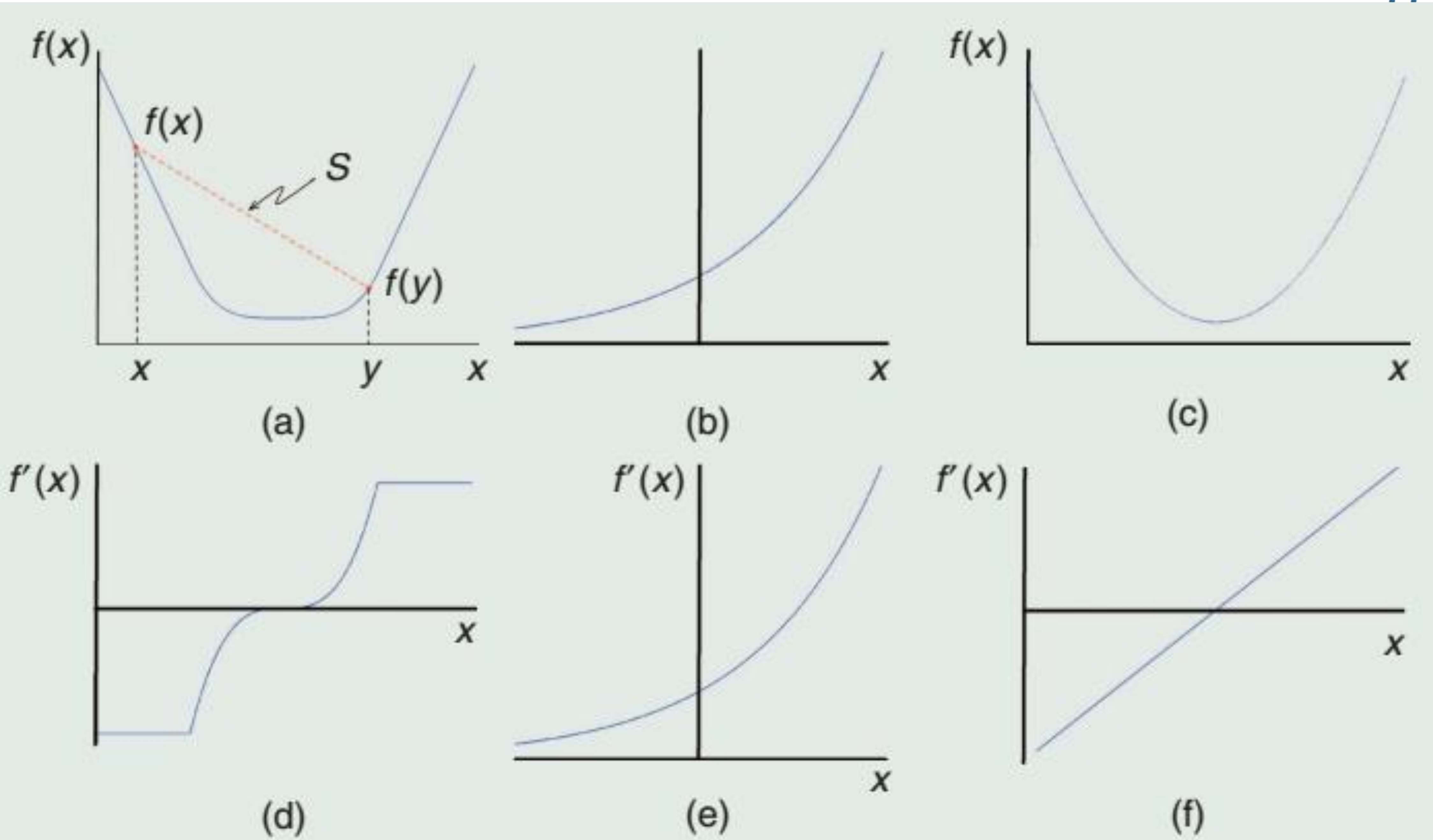
- 实际上也等价于：$\boldsymbol{(x' - x)^{\mathrm{T}} \nabla f(x) \geq 0}, \forall \boldsymbol{x'} \in \mathcal{X}$
- 由于$\boldsymbol{f}$的凸性$\Leftrightarrow$ $\boldsymbol{(x - y)^T (\nabla f(x) - \nabla f(y)) \geq 0}$,也称凸函数的梯度算子为单调算子!

# 第三章 非线性规划的数学模型

- 变分不等式问题（Variational Inequality Problem）
  - 给定闭凸集$F\colon \mathcal{X} \subseteq \mathbb{R}^n$和映射$F\colon \mathcal{X} \to \mathbb{R}^n$，变分问题表示为$VI(\mathcal{X}, F)$:找到向量$x^* \in \mathcal{X}$使得：
  - $(y - x^*)^T F(x^*) \geq 0, \forall y \in \mathcal{X}$
  - 即找到变分问题的一个解$x^*$。
- 注：
  - 如果某个函数$f$,其梯度$\nabla f \to F$,则显然有：$(y - x^*)^T \nabla f(x^*) \geq 0, \forall y \in \mathcal{X}$,即满足$x^*$为函数$f$的极小值点！
  - 因此，在一般意义下，变分问题的解等价于寻找连续可微凸函数的极值
  - 如果函数$F$不能表示成某些函数的梯度形式，则变分问题和最优化问题不同，但VI问题包含最优化问题
  - 但如果变分问题函数$F$的Jacobian矩阵是对称的，则$F$可表示为一个函数$f$的梯度形式，例如$F(x) = Ax + b, A$是$n \times n$的对称方阵，则$f(x) = \frac{1}{2}(x^T Ax + b^T x)$即满足$F = \nabla f$

# 第三章 非线性规划的数学模型

- 若$\mathcal{X} = \mathbb{R}^n$:方程组求解，此时$VI(\mathbb{R}^n, F)$等价于找到一个$x^* \in \mathbb{R}^n$使得$F(x^*) = 0$,因为唯一一个向量$F(x^*)$与所有$\mathbb{R}^n$中的向量成非钝角的向量就是0向量!
- 若$\mathcal{X} = \mathbb{R}^n_+$:非线性互补问题（Nonlinear complementarity Problem-NCP）,找向量$x^* \in \mathcal{X}$,使得：$0 \leq x^* \perp F(x^*) \geq 0$,等价于$x_i^* F_i(x^*) = 0, \forall i = 1,2,\cdots,n$
- 变分不等式解的存在性和唯一性如何?
  - 可以从最优化问题的解所必须满足的条件出发去思考?


f(x)
f(x)
S
f(y)
x
y
x
(a)
x
(b)
f(x)
x
(c)
f′(x)
x
(d)
f′(x)
x
(e)
f′(x)
x
(f)

# 第三章 非线性规划的数学模型

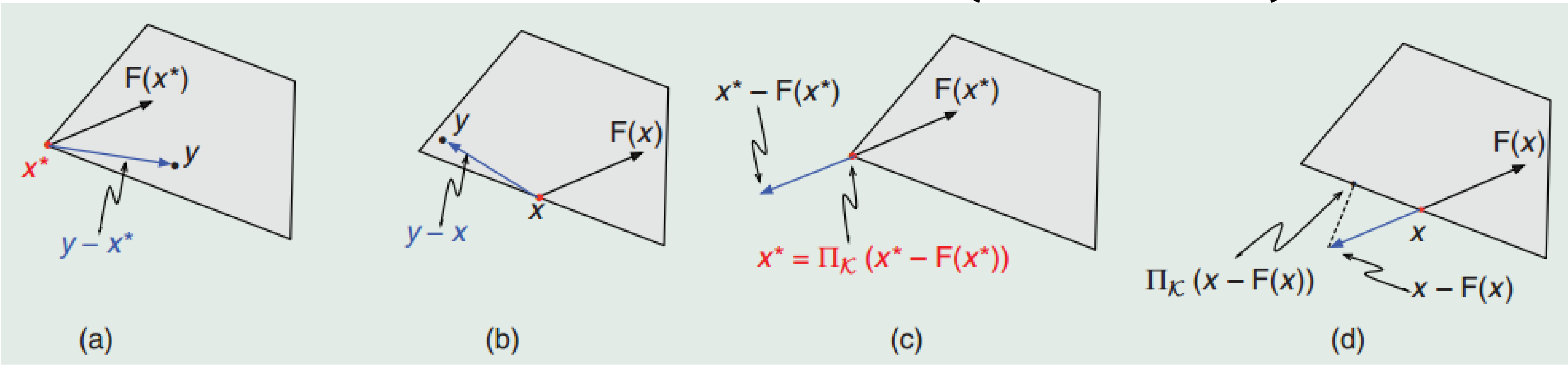
- 强单调→严格单调→单调，反之不然
- 这种单调性和函数$F$的Jacobian矩阵的半正定性之间存在联系！
- 对于仿射函数,$F(x) = Ax + b$,$A_{n\times n}$为矩阵，$b$为n维向量，以下结果成立：
- $F(x) = Ax + b$是单调的，当且仅当$A$是半正定的。而对正定矩阵而言，严格单调和强单调等价。如果向量函数$F$是某个标量函数$f$的梯度$\nabla f$,上述的单调性与函数$f$的凸性相关
  - **$f$是凸的$\Leftrightarrow$ $\nabla f$单调**
  - **$f$是严格凸的$\Leftrightarrow$ $\nabla f$严格单调**
  - **$f$是强凸的$\Leftrightarrow$ $\nabla f$强单调**

# 第三章 非线性规划的数学模型

- 基于上述结论，在不要求集合$\mathcal{X}$的紧性情况下，变分问题$VI(\mathcal{X},F)$的解具有如下的性质（注意$F$在集合$\mathcal{X}$上的连续性）
  - **若$\boldsymbol{F}$在$\boldsymbol{\mathcal{X}}$上单调，则$\boldsymbol{VI(\mathcal{X},F)}$的解集是闭且凸的**
  - **若$\boldsymbol{F}$在$\boldsymbol{\mathcal{X}}$上严格单调，则$\boldsymbol{VI(\mathcal{X},F)}$存在最多一个解**
  - **若$\boldsymbol{F}$在$\boldsymbol{\mathcal{X}}$上强单调，则$\boldsymbol{VI(\mathcal{X},F)}$存在唯一解**
- 其中第2条表明，严格单调并不能保证其解存在，例如$F(x)=e^x$严格单调，但$VI(\mathbb{R},e^x)$无解
- 同时，上述结果可以引导出凸优化解的存在性和唯一性
  - 例如，如果如果$f$强凸，则$\min_x f(x), s.t. x\in\mathcal{X}$有唯一解，这与下述论述等价：如果$\nabla f$强单调，$VI(\mathcal{X},\nabla f)$存在唯一解

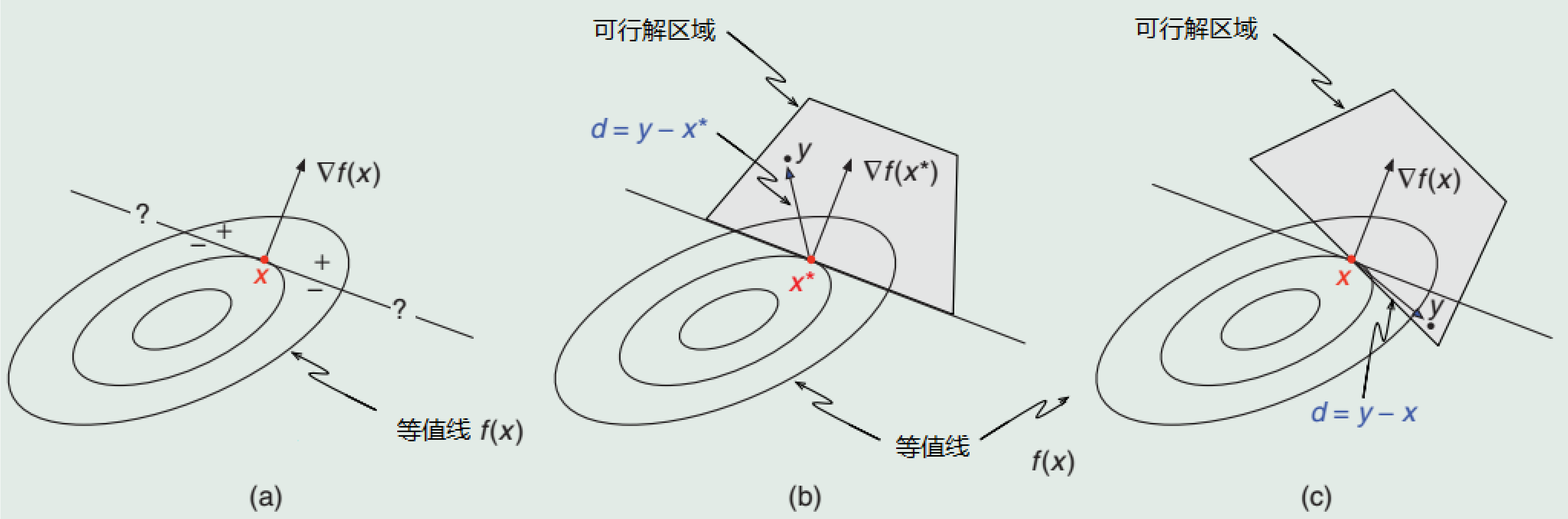
# 第三章 非线性规划的数学模型

- 投影定理：向量$x_0$在闭凸集$\mathcal{X}$的欧几里德投影表示为$\Pi_{\mathcal{X}}(x_0)$,表示$\mathcal{X}$中的与$x_0$的欧几里德范数意义上最近的唯一向量，采用优化的观点表示为：$\min_{y} ||y - x_0||^2, s.t. y \in \mathcal{X}$的解，目标函数是强凸的，因此其解存在且唯一。
- $VI(\mathcal{X}, F)$可以形式化为一个不动点问题：
- $x^*$是变分问题$VI(\mathcal{X}, F)$的解$\Leftrightarrow x^* = \Pi_{\mathcal{X}}\left(x^* - F(x^*)\right)$

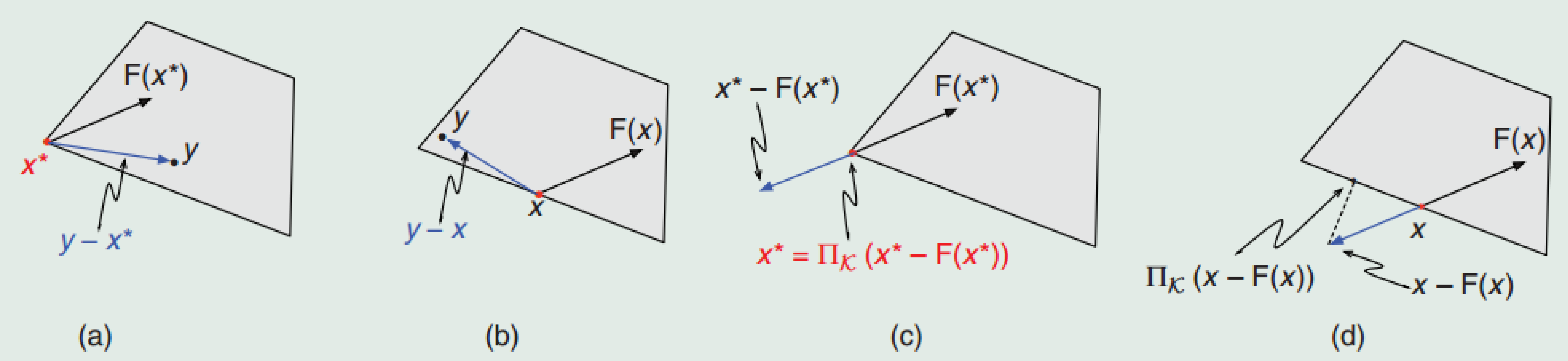


# 第三章 非线性规划的数学模型

- 传统使用优化方法来求解的问题，博弈模型也越来越广泛的用来解决这类问题，尤其是当参与者之间的交互是不可忽略的情形下，单纯的中心化方法不再适合!
- 非合作博弈（Noncooperative games）
  - 纳什均衡（Nash Equilibrium problem: NEPs）和广义纳什均衡（GNEPs）
  - 前者只在目标函数级别上进行交互，后者会考虑对手的行为
- NEPs
  - 假设$Q$个玩家分别控制变量$x_i \in \mathbb{R}^{n_i}$,其中用$x \triangleq (x_1, x_2, \cdots, x_Q)$，用$x_{-i} \triangleq (x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_Q)$表示除第$i$个玩家外所有其他玩家对应的变量。给定其他玩家的策略$x_{-i}$，玩家$i$的目标是选择一个$x_i \in Q_i$来最小化其支付函数$f_i(x_i, x_{-i})$，即
  - $\min_{x_i} f_i(x_i, x_{-i}), s.t. x_i \in Q_i$
  - 大概来看，一个NEP就是一个耦合的优化问题集合。假设$f_i$连续可微，单作为$x_i$的函数，是凸的集合$Q_i \subseteq R^{n_i}$是闭凸集。如果$x_i \in Q_i$对所有玩家$i$都成立，则点$x$称为可行点。一个纯粗略NE或简单一个NEP的解是可行点$x^*$:
  - $f_i(Ix_i^*, x_{-i}^*) \leq f_i(x_i, x_{-i}^*), \forall x_i \in Q_i$,对所有每一个玩家$i = 1,2, \cdots, Q$成立。
  - 一个NE是具有以下性质的一个可行策略档$x^*$：如果所有玩家都遵循这个策略，则没有单个玩家能够从$x_i^*$的单边偏离中获益。



∇f(x)
?
+
−
x
等值线 f(x)
(a)
可行解区域
d = y − x*
y
∇f(x*)
x*
等值线
f(x)
(b)
可行解区域
∇f(x)
x
y
d = y − x
(c)

F(x*)
x*
y
y − x*
(a)
y
F(x)
x
y − x
(b)
x* − F(x*)
F(x*)
x* = Π_K (x* − F(x*))
(c)
F(x)
x
Π_K (x − F(x))
x − F(x)
(d)

# 第三章 非线性规划的数学模型

- 变分表达：设$\mathcal{X} \subset \boldsymbol{R}^n$是闭凸集，$\boldsymbol{\theta}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{x})$都是$\boldsymbol{R}^n \to \boldsymbol{R}$上的凸函数，如果$\boldsymbol{\phi}(\boldsymbol{x})$可微并且最优化问题：$\boldsymbol{min}\{\boldsymbol{\theta}(\boldsymbol{x}) + \boldsymbol{\phi}(\boldsymbol{x}) | \boldsymbol{x} \in \mathcal{X}\}$有解，则$\tilde{\boldsymbol{x}} \in \boldsymbol{argmin}\{\boldsymbol{\theta}(\boldsymbol{x}) + \boldsymbol{\phi}(\boldsymbol{x}) | \boldsymbol{x} \in \mathcal{X}\}$的充要条件是：
$$\tilde{\boldsymbol{x}} \in \mathcal{X}, \boldsymbol{\theta}(\boldsymbol{x}) - \boldsymbol{\theta}(\tilde{\boldsymbol{x}}) + (\boldsymbol{x} - \tilde{\boldsymbol{x}})^T \nabla \boldsymbol{\phi}(\tilde{\boldsymbol{x}}) \geq \boldsymbol{0}, \forall \boldsymbol{x} \in \mathcal{X}$$
- 问题1.如何证明？
  - 必要性: $\boldsymbol{x}^*$存在，令$\boldsymbol{x}_{\boldsymbol{\alpha}} = (\boldsymbol{1} - \boldsymbol{\alpha})\boldsymbol{x}^* + \boldsymbol{\alpha x}, \forall \boldsymbol{\alpha} \in (\boldsymbol{0}, \boldsymbol{1}]$,则有,$\frac{\boldsymbol{\theta}(\boldsymbol{x}_{\alpha}) - \boldsymbol{\theta}(\boldsymbol{x}^*)}{\boldsymbol{\alpha}} + \frac{\boldsymbol{\phi}(\boldsymbol{x}_{\alpha}) - \boldsymbol{\phi}(\boldsymbol{x}^*)}{\boldsymbol{\alpha}} \geq \boldsymbol{0}$
  - $\boldsymbol{\theta}(\boldsymbol{x})$是凸函数，$\boldsymbol{\theta}(\boldsymbol{x}_{\boldsymbol{\alpha}}) \leq (\boldsymbol{1} - \boldsymbol{\alpha})\boldsymbol{\theta}(\boldsymbol{x}^*) + \boldsymbol{\alpha \theta}(\boldsymbol{x})$,从而：$\boldsymbol{\theta}(\boldsymbol{x}) - \boldsymbol{\theta}(\boldsymbol{x}^*) \geq \frac{\boldsymbol{\theta}(\boldsymbol{x}_{\alpha}) - \boldsymbol{\theta}(\boldsymbol{x}^*)}{\boldsymbol{\alpha}}, \forall \boldsymbol{\alpha} \in (\boldsymbol{0}, \boldsymbol{1}]$,因此
  - $\boldsymbol{\theta}(\boldsymbol{x}) - \boldsymbol{\theta}(\boldsymbol{x}^*) + \frac{\boldsymbol{\phi}(\boldsymbol{x}_{\alpha}) - \boldsymbol{\phi}(\boldsymbol{x}^*)}{\boldsymbol{\alpha}} \geq \boldsymbol{0}$，由于$\boldsymbol{\phi}(\boldsymbol{x}_{\boldsymbol{\alpha}}) = \boldsymbol{\phi}\big(\boldsymbol{x}^* + \boldsymbol{\alpha}(\boldsymbol{x} - \boldsymbol{x}^*)\big)$,若令$\boldsymbol{\alpha} \to \boldsymbol{0}_+$,立即可得$\boldsymbol{\theta}(\boldsymbol{x}) - \boldsymbol{\theta}(\tilde{\boldsymbol{x}}) + (\boldsymbol{x} - \tilde{\boldsymbol{x}})^T \nabla \boldsymbol{\phi}(\tilde{\boldsymbol{x}}) \geq \boldsymbol{0}, \forall \boldsymbol{x} \in \mathcal{X}$
  - 充分性：$\boldsymbol{\phi}(\boldsymbol{x}_{\boldsymbol{\alpha}}) \leq (\boldsymbol{1} - \boldsymbol{\alpha})\boldsymbol{\phi}(\boldsymbol{x}^*) + \boldsymbol{\alpha \phi}(\boldsymbol{x}) \Rightarrow \boldsymbol{\phi}(\boldsymbol{x}_{\boldsymbol{\alpha}}) - \boldsymbol{\phi}(\boldsymbol{x}^*) \leq \boldsymbol{\alpha}\big(\boldsymbol{\phi}(\boldsymbol{x}) - \boldsymbol{\phi}(\boldsymbol{x}^*)\big) \Rightarrow \boldsymbol{\phi}(\boldsymbol{x}) - \boldsymbol{\phi}(\boldsymbol{x}^*) \geq \frac{\boldsymbol{\phi}(\boldsymbol{x}_{\alpha}) - \boldsymbol{\phi}(\boldsymbol{x}^*)}{\boldsymbol{\alpha}} = \frac{\boldsymbol{\phi}\big(\boldsymbol{x}^* + \boldsymbol{\alpha}(\boldsymbol{x} - \boldsymbol{x}^*)\big) - \boldsymbol{\phi}(\boldsymbol{x}^*)}{\boldsymbol{\alpha}}, \forall \boldsymbol{\alpha} \in (\boldsymbol{0}, \boldsymbol{1}]$,两边令$\boldsymbol{\alpha} \to \boldsymbol{0}_+ \Rightarrow \boldsymbol{\phi}(\boldsymbol{x}) - \boldsymbol{\phi}(\boldsymbol{x}^*) \geq \nabla \boldsymbol{\phi}(\boldsymbol{x}^*)^T (\boldsymbol{x} - \boldsymbol{x}^*) \Rightarrow \boldsymbol{\theta}(\boldsymbol{x}) - \boldsymbol{\theta}(\boldsymbol{x}^*) + \boldsymbol{\phi}(\boldsymbol{x}) - \boldsymbol{\phi}(\boldsymbol{x}^*) \geq \boldsymbol{0}$

# 第三章 非线性规划的数学模型

- 具有线性等式约束的凸优化问题求解
  - $\boldsymbol{min}\{\boldsymbol{\theta}(\boldsymbol{x})|\boldsymbol{Ax}=\boldsymbol{b},\boldsymbol{x}\in\mathcal{X}\}$,其中函数$\boldsymbol{\theta}(\boldsymbol{x})$为凸函数,$\boldsymbol{A}_{\boldsymbol{m}\times\boldsymbol{n}}$
- 问题1.该问题的拉格朗日乘子法的形式应该是什么?
  - $\boldsymbol{L}(\boldsymbol{x},\boldsymbol{\lambda})=\boldsymbol{\theta}(\boldsymbol{x})-\boldsymbol{\lambda}^{\boldsymbol{T}}(\boldsymbol{Ax}-\boldsymbol{b})$
  - 点$(\boldsymbol{x}^*,\boldsymbol{\lambda}^*)$称为拉格朗日函数的鞍点，如果满足下式:

$$\boldsymbol{L}_{\boldsymbol{\lambda}\in\boldsymbol{R}^{\boldsymbol{m}}}(\boldsymbol{x}^*,\boldsymbol{\lambda})\le\boldsymbol{L}(\boldsymbol{x}^*,\boldsymbol{\lambda}^*)\le\boldsymbol{L}_{\boldsymbol{x}\in\mathcal{X}(\boldsymbol{x},\boldsymbol{\lambda}^*)}$$

- 该条件等价于如下条件:
- $\begin{cases}\boldsymbol{x}^*\in\mathcal{X}, & \boldsymbol{\theta}(\boldsymbol{x})-\boldsymbol{\theta}(\boldsymbol{x}^*)+(\boldsymbol{x}-\boldsymbol{x}^*)^{\boldsymbol{T}}\left(-\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{\lambda}^*\right)\ge\boldsymbol{0},\forall\boldsymbol{x}\in\mathcal{X}\\ \boldsymbol{\lambda}^*\in\boldsymbol{R}^{\boldsymbol{m}}, & (\boldsymbol{\lambda}-\boldsymbol{\lambda}^*)^{\boldsymbol{T}}(\boldsymbol{Ax}^*-\boldsymbol{b})\ge\boldsymbol{0},\forall\boldsymbol{\lambda}\in\boldsymbol{R}^{\boldsymbol{m}}\end{cases}$
- 若令$\boldsymbol{w}=\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{\lambda}\end{pmatrix}$,$\boldsymbol{F}(\boldsymbol{w})=\begin{pmatrix}-\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{\lambda}\\ \boldsymbol{Ax}-\boldsymbol{b}\end{pmatrix}$,$\boldsymbol{w}\in\mathcal{X}\times\boldsymbol{R}^{\boldsymbol{m}}$

# 第三章 非线性规划的数学模型

- 此时最优性条件可表达为变分不等式:
- $\boldsymbol{w}^* \in \mathcal{X} \times \boldsymbol{R}^m, \boldsymbol{\theta}(\boldsymbol{x}) - \boldsymbol{\theta}(\boldsymbol{x}^*) + (\boldsymbol{w} - \boldsymbol{w}^*)^T \boldsymbol{F}(\boldsymbol{w}^*) \geq \boldsymbol{0}, \forall \boldsymbol{w} \in \mathcal{X} \times \boldsymbol{R}^m$
- 注意算子$\boldsymbol{F}$的单调性，因为
  - $(\boldsymbol{w} - \widetilde{\boldsymbol{w}})^T \left(\boldsymbol{F}(\boldsymbol{w}) - \boldsymbol{F}(\widetilde{\boldsymbol{w}})\right) \geq \boldsymbol{0}$,实际上
  - $(\boldsymbol{w} - \widetilde{\boldsymbol{w}})^T \left(\boldsymbol{F}(\boldsymbol{w}) - \boldsymbol{F}(\widetilde{\boldsymbol{w}})\right) = 0$

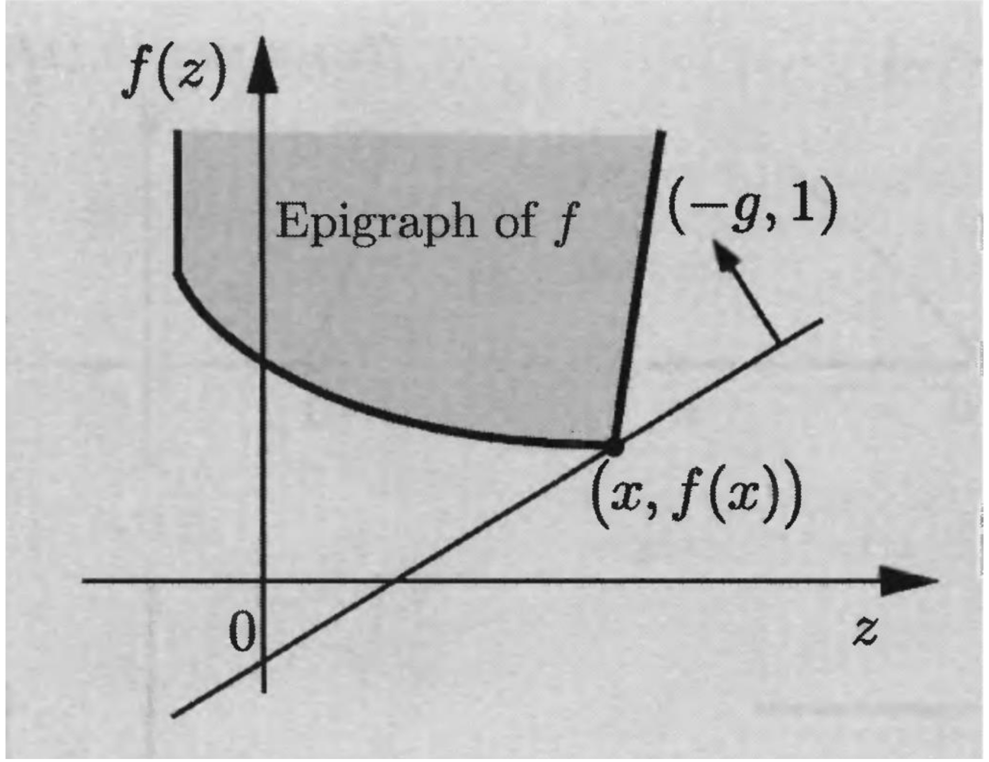
# 第三章 非线性规划的数学模型

- 次梯度(subgradient)
  - 梯度的替代项，不可微的时候用次梯度
  - $\boldsymbol{f}(\boldsymbol{z}) \geq \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{g}^{\boldsymbol{T}}(\boldsymbol{z} - \boldsymbol{x})$

$\boldsymbol{g}$是函数$\boldsymbol{f}$在点$\boldsymbol{x}$的次梯度等价于

$$f(z) - g^T z \geq f(x) - g^T x$$

在$n+1$维，过点$(x, f(x))$，法向量为$(-g, 1)$的超平面支撑函数$f$的上镜图(epigraph)



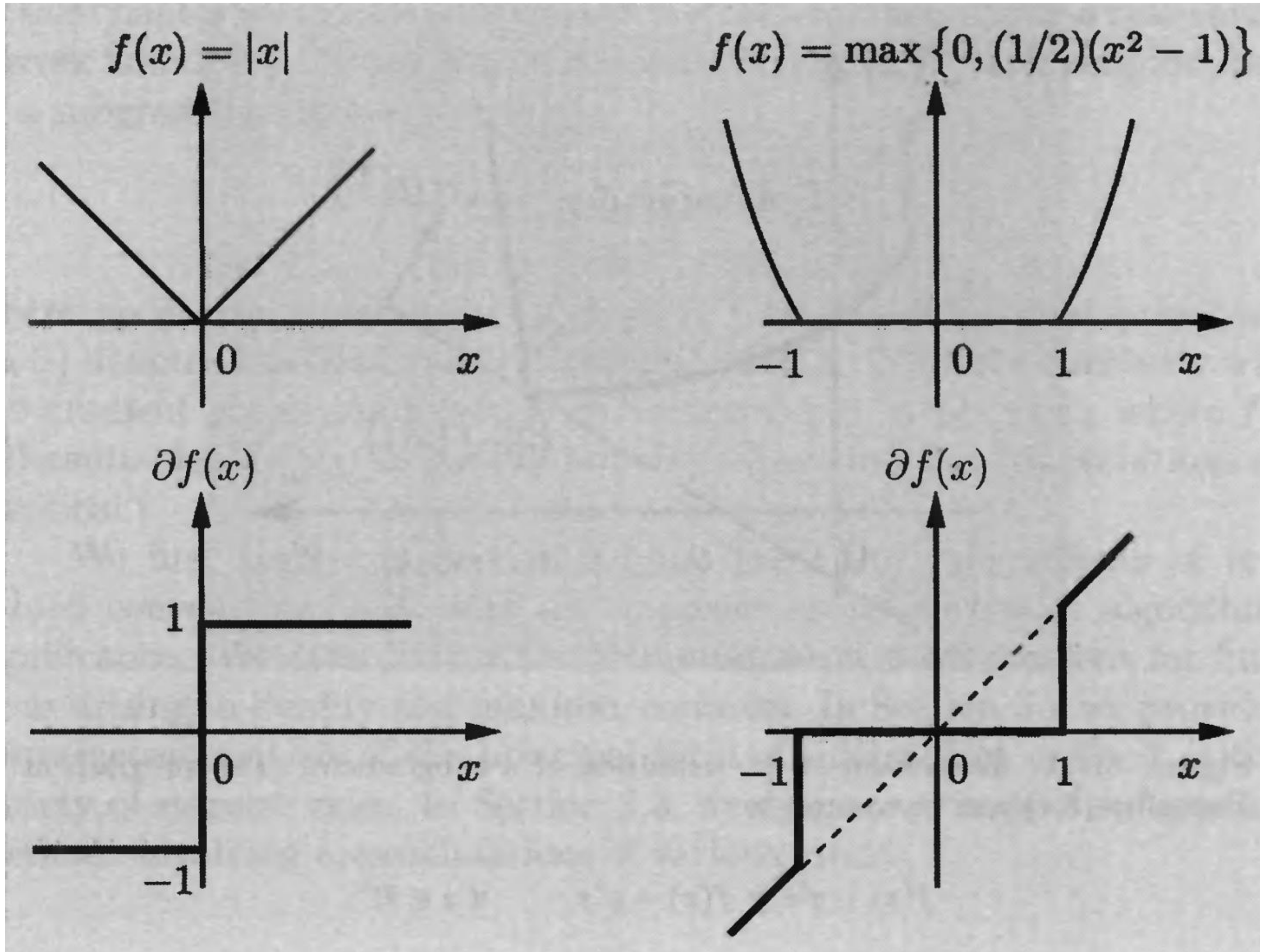
# 第三章 非线性规划的数学模型

- 次梯度的计算

  向量$\boldsymbol{x}$是凸函数$\boldsymbol{f}$在的极小值点等价于存在一个次梯度$g$，使得

  $g^T(z-x) \geq 0, \forall z \in \mathcal{X}$

  在$\boldsymbol{x}$点处的所有次梯度的集合称为该点的次微分$\boldsymbol{\partial f(x)}$



# 第三章 非线性规划的数学模型

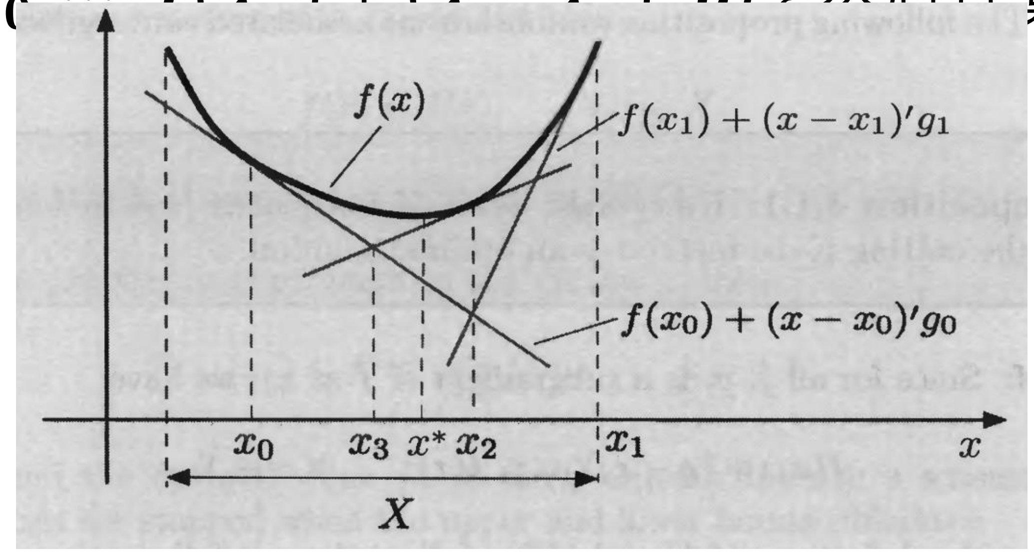
- 临近点算法PPA（Proximal Point Algorithms）
  - 本章介绍了优化算法的一般迭代格式$\boldsymbol{x}^{k+1} = \boldsymbol{x}^{k} + \boldsymbol{\lambda}_{k}\boldsymbol{p}_{k}$
  - 多次强调在步长和方向上的变化!
- 问题1.你觉得除了步长和方向上有变化外，还能有哪些变化或扩展?
  - $\boldsymbol{x}^{k+1} = \boldsymbol{P}_{\mathcal{X}}(\boldsymbol{x}^{k} - \boldsymbol{\lambda}_{k}\boldsymbol{p}_{k})$,其中$\boldsymbol{p}_{k}$可以是任何可行的方向，梯度方向，次梯度方向，$\boldsymbol{\lambda}_{k}$是正的步长，$\boldsymbol{P}_{\mathcal{X}}$表示在可行集合上的欧式投影
  - 方向可以是次梯度，此时对应次梯度方法
  - 若将次梯度放松，$\boldsymbol{f}(\boldsymbol{z}) \geq \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{g}^{T}(\boldsymbol{z} - \boldsymbol{x}) - \boldsymbol{\epsilon}, \boldsymbol{\epsilon} > \boldsymbol{0}$,则有$\boldsymbol{\epsilon}$-次梯度方法

# 第三章 非线性规划的数学模型

- 多边形近似算法（Polyhedral Approximation Algorithms）
  - 在迭代算法中每次生成的$\boldsymbol{x}^{\boldsymbol{k}+\mathbf{1}}$满足：
  - $\boldsymbol{x}^{\boldsymbol{k}+\mathbf{1}} \in \boldsymbol{argmin}_{\boldsymbol{x}\in\mathcal{X}_{\boldsymbol{k}}}\boldsymbol{F}_{\boldsymbol{k}}(\boldsymbol{x})$
  - 其中$\boldsymbol{F}_{\boldsymbol{k}}$是近似$\boldsymbol{f}$的一个多面体函数，$\mathcal{X}_{\boldsymbol{k}}$是近似$\mathcal{X}$的一个多面体集合
- 从而将原来问题变为逼近问题，多面体结构比原问题易于求解，从而解决原问题
  - 外部线性化方法-割平面法
  - 凸函数的求解中可以采用迭代点列的割平面（支撑）所围成的区域来近似

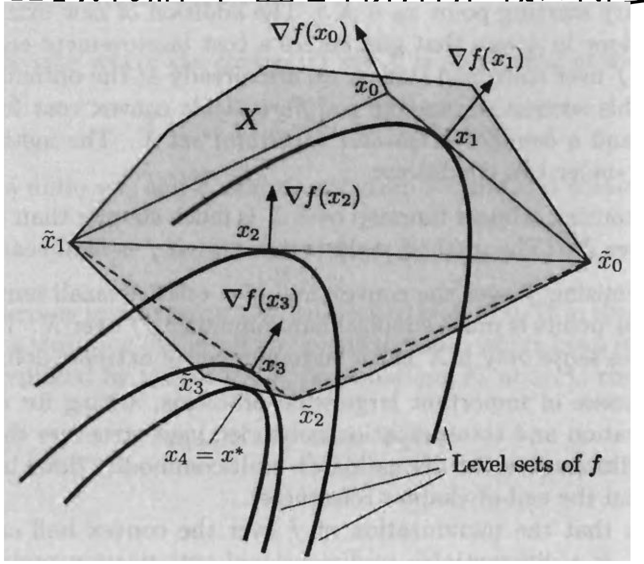
# 第三章 非线性规划的数学模型

- 外部线性化方法
  - 凸函数$\boldsymbol{f}: \boldsymbol{R}^n \to \boldsymbol{R}$,初始点$\boldsymbol{x}^0$,次梯度$\boldsymbol{g}_0 \in \partial \boldsymbol{f}(\boldsymbol{x}_0)$,其迭代格式表示为:
  - $\boldsymbol{min}\ \boldsymbol{F}_k(\boldsymbol{x}), \boldsymbol{s.t.}\ \boldsymbol{x} \in \mathcal{X}$
  - $\boldsymbol{f}$采用多面体近似，由已有$\boldsymbol{x}^0, \boldsymbol{x}^1, \cdots, \boldsymbol{x}^k$来进行近似，且相关的次梯度为$\boldsymbol{g}_i \in \partial \boldsymbol{f}(\boldsymbol{x}_i)$，例如
  - $\boldsymbol{F}_k(\boldsymbol{x}) = \boldsymbol{max}\{\boldsymbol{f}(\boldsymbol{x}^0) + (\boldsymbol{x} - \boldsymbol{x}^0)^T \boldsymbol{g}_0, \cdots, \boldsymbol{f}(\boldsymbol{x}^k) + (\boldsymbol{x} - \boldsymbol{x}^k)^T \boldsymbol{g}_k\}$然后计算
  - $\boldsymbol{x}^{k+1} \in \boldsymbol{argmin}_{x \in \mathcal{X}} \boldsymbol{F}_k(\boldsymbol{x})$



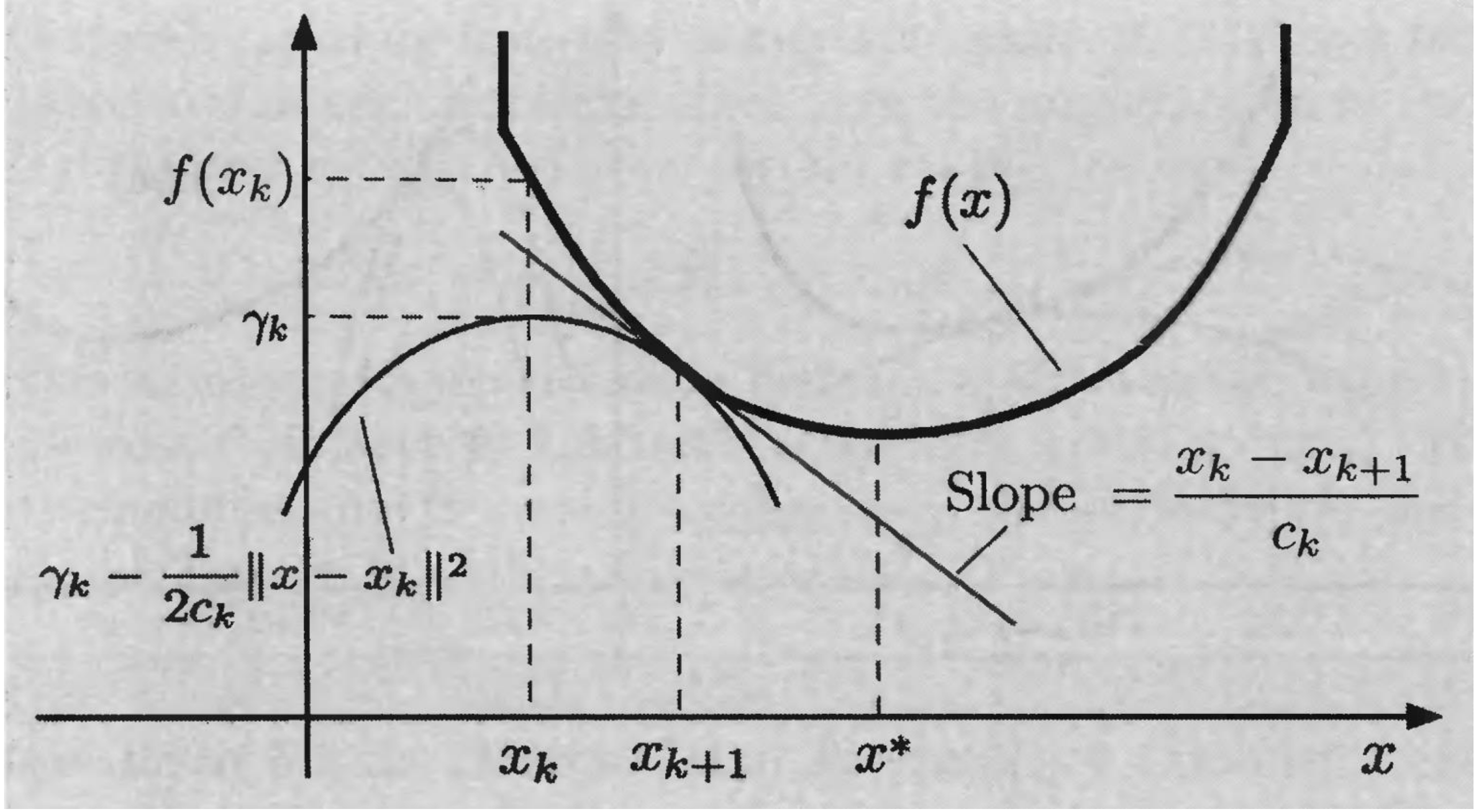
# 第三章 非线性规划的数学模型

- 最小化闭凸集上的凸函数$\boldsymbol{f}$
  - 通过有限集合$\boldsymbol{X_k} \subset \mathcal{X}$的凸包来逼近可行域$\mathcal{X}$,其中$\boldsymbol{X_k}$包括$\mathcal{X}$的极点和任一个初始点$\boldsymbol{x^0} \in \mathcal{X}$
- 基本步骤
  - 令初始点$\boldsymbol{x^0}$构成初始集合$\boldsymbol{X_0} = \{\boldsymbol{x^0}\}$,求解如下问题生成$\boldsymbol{\tilde{x}^k}$作为$\mathcal{X}$的极点:
  - $\boldsymbol{min\ \nabla f\left(x^k\right)^T\left(x - x^k\right), s.t.\, x} \in \mathcal{X}$
  - 然后将$\boldsymbol{\tilde{x}^k}$加入到集合$\boldsymbol{X_k}$中， $\boldsymbol{X_{k+1}} = \left\{\boldsymbol{\tilde{x}^k}\right\} \cup \boldsymbol{X_k}$,求解下面的优化问题得到$\boldsymbol{x^{k+1}}$:
  - $\boldsymbol{min\ f(x), s.t.\, x} \in \boldsymbol{conv(X_{k+1})}$



# 第三章 非线性规划的数学模型

- 添加正则项求解凸规划
  - $\boldsymbol{x^{k+1}} \in \boldsymbol{argmin_{x \in R^n}} \left\{ \boldsymbol{f(x)} + \frac{\mathbf{1}}{\mathbf{2}\boldsymbol{c_k}} ||\boldsymbol{x} - \boldsymbol{x^k}||^{\mathbf{2}} \right\}$,其中初始点任意，$\boldsymbol{c_k} > \mathbf{0}$为标量。



# 第四章 一维搜索与无约束最优化

- 黑箱模型
  - 假设计算资源无限，约束集$\mathcal{X}$已知，目标函数$\boldsymbol{f}: \mathcal{X} \to \mathbb{R}$未知，但可以通过oracle查询
    - 零阶的oracle接受$\boldsymbol{x} \in \mathcal{X}$作为输入，输出函数$\boldsymbol{f}$在点$\boldsymbol{x}$处的函数值
    - 一阶的oracle接受$\boldsymbol{x} \in \mathcal{X}$作为输入，输出函数$\boldsymbol{f}$在点$\boldsymbol{x}$处的次梯度
  - 凸优化的oracle复杂性：必须经过多少对oracle的查询才足以找到凸函数的$\boldsymbol{\epsilon}$ −近似极小值
    - 能够导出一个完整的凸优化理论，获得匹配各种有趣凸函数子类的oracle复杂度上下界
    - 模型本身并不限制计算资源，允许对约束集的任何操作，但会注意算法的计算复杂性（即算法需要执行基本操作的数量）
    - 如果约束集合$\mathcal{X}$是未知的，并且只能通过分离oracle得到：给定$\boldsymbol{x} \in \mathbb{R}, \Rightarrow \boldsymbol{x} \in \mathcal{X}$或者$\boldsymbol{x} \notin \mathcal{X}$,那么它输出$\boldsymbol{x}$和$\mathcal{X}$之间的分离超平面
  - 开发维度无关的oracle复杂性算法是可能的，对高维优化问题非常有意义
  - 在黑箱模型中开发的算法对oracle输出中的噪声具有鲁棒性，这对于随机优化特别有意义，并且与机器学习应用紧密相关
- 结构性优化，试图考虑约束集和目标函数的全局结构，如内点法

# 第四章 一维搜索与无约束最优化

- 通用迭代算法的复杂性

**输入**：初始点$x_0$和精度$\epsilon > 0$
**初始化**：令$k = 0, \psi_{-1} = \emptyset$,这里$k$是迭代计数，$\psi_k$是累积的信息集
**主循环**：
1. 在点$x_k$处调用Oracle $\mathcal{O}$
2. 更新信息集：$\psi_k = \psi_{k-1} \cup \left(x_k, \mathcal{O}(x_k)\right)$
3. 将方法$\mathcal{M}$的规则应用于$\psi_k$，生成一个新点$x_{k+1}$
4. 检验停止准则$\mathcal{T}_\epsilon$:如果满足停止准则，则输出$\bar{x}$；否则置$k := k + 1$,转到第1步

引入两种度量准则来衡量算法$\mathcal{M}$求解优化问题$\mathcal{P}$的复杂度
**解析复杂度（Analytical Complexity)**：为使问题$\mathcal{P}$达到精度$\epsilon$，需要调用Oracle的次数
**算术复杂度（Arithmetical Complexity)**：为使问题$\mathcal{P}$达到精度$\epsilon$，需要的算术运算总量（包括Oracle的调用计算量和算法$\mathcal{M}$的计算量）。

# 第四章 一维搜索与无约束最优化

- 了解优化方法的复杂性吗？
- 问题$\mathcal{P}$： $\mathbf{min}_{\mathrm{x}\in B_n}\mathbf{f}(\mathbf{x}),\mathbf{B_n}=\{\mathbf{x}\in\mathbb{R}^{\boldsymbol{n}}|\mathbf{0}\le\boldsymbol{x_i}\le\mathbf{1},\boldsymbol{i}=\mathbf{1}\cdots\boldsymbol{n}\}$
  - 假设距离为$\boldsymbol{l}_\infty := ||\boldsymbol{x}||_\infty=\boldsymbol{max}_{\mathbf{1}\le\boldsymbol{i}\le\boldsymbol{n}}|\boldsymbol{x_i}|$
  - 目标函数$\boldsymbol{f}(\cdot):\mathbb{R}^{\boldsymbol{n}}\to\mathbb{R}$是在$\mathbf{B_n}$上Lipschitz连续的：$|\mathbf{f}(\mathbf{x})-\mathbf{f}(\mathbf{y})|\le\mathbf{L}||\mathbf{x}-\mathbf{y}||_{(\infty)},\forall\mathbf{x},\mathbf{y},\in\boldsymbol{B_n}$,L为Lipschitz常数
- 假设采用均匀网格方法求解$\mathcal{P}$， 其输入参数为整数*p≥1*
  - $\mathbf{x}_{\boldsymbol{\alpha}}=\left(\frac{\mathbf{2i_1-1}}{\mathbf{2p}},\frac{\mathbf{2i_2-1}}{\mathbf{2p}},\cdots,\frac{\mathbf{2}\boldsymbol{i_n}\mathbf{-1}}{\mathbf{2}\boldsymbol{p}}\right)^{\boldsymbol{T}},\boldsymbol{\alpha}\equiv(\boldsymbol{i_1},\cdots,\boldsymbol{i_n})\in\{\mathbf{1},\cdots,\boldsymbol{p}\}^{\boldsymbol{n}}$
  - 在所有点$\mathbf{x}_{\boldsymbol{\alpha}}$上求具有最小目标函数值的点$\bar{\mathbf{x}}$
  - 方法输出为$(\bar{\mathbf{x}},\mathbf{f}(\bar{\mathbf{x}}))$

# 第四章 一维搜索与无约束最优化

该算法在$\boldsymbol{B_n}$内形成测试点的均匀网格，在网格上计算目标的最优值，并将此值作为问题$\mathcal{P}$的近似解。属于零阶迭代算法，现在看看其效率估计。

- 定理：若$\boldsymbol{f^*}$为全局最优解，则$\boldsymbol{f(\bar{x}) - f^* \leq \frac{L}{2p}}$
  - 对多索引$\boldsymbol{\alpha} \equiv (\boldsymbol{i_1}, \cdots, \boldsymbol{i_n})$，定义$\boldsymbol{X_\alpha} = \left\{\boldsymbol{x} \in \boldsymbol{R^n}: ||\boldsymbol{x} - \boldsymbol{x_\alpha}||_\infty \leq \frac{1}{2p}\right\}$,显然$\bigcup_{\alpha\in\{1,\cdots,p\}^n} \boldsymbol{X_\alpha} = \boldsymbol{B_n}$,由$\boldsymbol{x^*}$是全局解，存在多索引$\boldsymbol{\alpha^*}$使得$\boldsymbol{x^*} \in \boldsymbol{X_{\alpha^*}}$。注意到$||\boldsymbol{x^*} - \boldsymbol{x_{\alpha^*}}||_\infty \leq \frac{1}{2p}$,从而得证。
- 推论：假设原问题变为：求$\bar{\mathbf{x}} \in \boldsymbol{B_n}: \boldsymbol{f(\bar{x}) - f^* \leq \epsilon}$，则有：上述问题的解析复杂性最多为$\left(\left\lfloor\frac{L}{2\epsilon}\right\rfloor + 1\right)^n$.令$\mathbf{p} = \left\lfloor\frac{L}{2\epsilon}\right\rfloor + \mathbf{1}$, 则$\mathbf{p} \geq \frac{L}{2\epsilon} \Rightarrow \mathbf{f(\bar{x}) - f^* \leq \frac{L}{2p} \leq \epsilon}$.
- 确定了问题类的复杂度上界。存在的问题1.证明粗糙，实际性能可能会更好；2.不能确定是否算法就是解决问题的合理方法，可能存在更好的。下界?
- 定理：对于$\boldsymbol{\epsilon} < \frac{1}{2}\boldsymbol{L}$,问题的解析复杂度至少为$\left\lfloor\frac{L}{2\epsilon}\right\rfloor^n$次调用oracle.（令$\mathbf{p} = \left\lfloor\frac{L}{2\epsilon}\right\rfloor (\geq \mathbf{1})$)
- 对于上述均匀网格法的性能，将其效率估计值与下界进行比较$\left(\left\lfloor\frac{L}{2\epsilon}\right\rfloor + \mathbf{1}\right)^n \Leftrightarrow \left\lfloor\frac{L}{2\epsilon}\right\rfloor^n$，如果$\boldsymbol{\epsilon} \leq \boldsymbol{O}\left(\frac{L}{n}\right)$，则除了一个绝对常数乘子的意义下，下界和上界是一致的。这意味着，对于这个精度，算法对该问题类来说是最优的

# 第四章 一维搜索与无约束最优化

- 考虑上述问题参数为：$\boldsymbol{L=2,n=10,\epsilon=0.01}$.问题规模非常小，且只要求适中的精度1%。
  - 该问题的复杂度下界是Oracle的$\left\lfloor\frac{\boldsymbol{L}}{\boldsymbol{2\epsilon}}\right\rfloor^{\boldsymbol{n}}$次调用，对于这个例子，看看具体值
  - 下界：Oracle 的$\boldsymbol{10^{20}}$次调用
  - Oracle的复杂度：至少$\boldsymbol{n}$次算术运算
  - 总体复杂度：$\boldsymbol{10^{21}}$次算术运算
  - 处理器性能：每秒$\boldsymbol{10^{6}}$次算术运算
  - 总时间：$\boldsymbol{10^{15}}$秒
  - 一年：不超过$\boldsymbol{3.2\times10^{7}}$秒
  - 需要：31250000年，即使处理器达到$\boldsymbol{10^{8}}$,n=11时仍然成立！
  - 与组合优化中的NP难问题的复杂度 比较，结果也令人沮丧，为找到精确解，最难的组合问题只需要$\boldsymbol{2^{n}}$次算术运算！

# 第四章 一维搜索与无约束最优化

- 令$\boldsymbol{Q}$是$\boldsymbol{R}^n$中的一个子集，用$\boldsymbol{C}_L^{k,p}(\boldsymbol{Q})$表示满足下面性质的函数类
  - 任意$\boldsymbol{f} \in \boldsymbol{C}_L^{k,p}(\boldsymbol{Q})$在$\boldsymbol{Q}$上是$\boldsymbol{k}$次连续可微的
  - 其$\boldsymbol{p}$阶导数在$\boldsymbol{Q}$上关于常数$\boldsymbol{L}$是李普希兹连续的，即对$\forall \boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{Q}$,都有$||\nabla^p \boldsymbol{f}(\boldsymbol{x}) - \nabla^p \boldsymbol{f}(\boldsymbol{y})|| \le \boldsymbol{L}||\boldsymbol{x} - \boldsymbol{y}||$
  - 显然，$\boldsymbol{p} \le \boldsymbol{k}$。如果$\boldsymbol{q} \ge \boldsymbol{k}$，则$\boldsymbol{C}_L^{q,p}(\boldsymbol{Q}) \subseteq \boldsymbol{C}_L^{k,p}(\boldsymbol{Q})$
  - 如果$\boldsymbol{f}_1 \in \boldsymbol{C}_{L_1}^{k,p}(\boldsymbol{Q}), \boldsymbol{f}_2 \in \boldsymbol{C}_{L_2}^{k,p}(\boldsymbol{Q}), \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \boldsymbol{R}$,则对$\boldsymbol{L}_3 = |\boldsymbol{\alpha}_1|\boldsymbol{L}_1 + |\boldsymbol{\alpha}_2|\boldsymbol{L}_2$,我们有$\boldsymbol{\alpha}_1\boldsymbol{f}_1 + \boldsymbol{\alpha}_2\boldsymbol{f}_2 \in \boldsymbol{C}_{L_3}^{k,p}(\boldsymbol{Q})$
- **定理**：函数$\boldsymbol{f}(\cdot) \in \boldsymbol{C}_L^{2,1}(\boldsymbol{R}^n) \subset \boldsymbol{C}_L^{1,1}(\boldsymbol{R}^n)$,当且仅当$\forall \boldsymbol{x} \in \boldsymbol{R}^n$,我们有$||\nabla^2 \boldsymbol{f}(\boldsymbol{x})|| \le \boldsymbol{L}$
  - 证明：注意$\forall \boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{R}^n, \nabla \boldsymbol{f}(\boldsymbol{y}) = \nabla \boldsymbol{f}(\boldsymbol{x}) + \int_0^1 \nabla^2 \boldsymbol{f}\big(\boldsymbol{x} + \boldsymbol{\tau}(\boldsymbol{y} - \boldsymbol{x})\big)(\boldsymbol{y} - \boldsymbol{x})d\boldsymbol{\tau}$得证⇒。
  - $\Leftarrow \forall \boldsymbol{s} \in \boldsymbol{R}^n, \boldsymbol{\alpha} > 0$,有$||\left(\int_0^1 \nabla^2 \boldsymbol{f}(\boldsymbol{x} + \boldsymbol{\tau s})d\boldsymbol{\tau}\right) \cdot \boldsymbol{s}|| = ||\nabla \boldsymbol{f}(\boldsymbol{x} + \boldsymbol{\alpha s}) - \nabla \boldsymbol{f}(\boldsymbol{x})|| \le \boldsymbol{\alpha L}||\boldsymbol{s}||$,用$\boldsymbol{\alpha}$除这个等式，同时令$\boldsymbol{\alpha} \downarrow 0$,即得证。

# 第四章 一维搜索与无约束最优化

- 定理：令$\boldsymbol{f} \in \boldsymbol{C}_L^{1,1}(\boldsymbol{R}^n)$,则$\forall \boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{R}^n$，我们有
  - $|\boldsymbol{f}(\boldsymbol{y}) - \boldsymbol{f}(\boldsymbol{x}) - <\nabla \boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}>| \le \frac{L}{2}||\boldsymbol{y} - \boldsymbol{x}||^2$
- 由此，可知，对于任意$\boldsymbol{f} \in \boldsymbol{C}_l^{1,1}(\boldsymbol{R}^n)$,取定点$\boldsymbol{x_0} \in \boldsymbol{R}^n$，则可构造两个二次函数
  - $\boldsymbol{\phi_1}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x_0}) + <\nabla \boldsymbol{f}(\boldsymbol{x_0}), \boldsymbol{x} - \boldsymbol{x_0}> - \frac{\boldsymbol{L}}{2}||\boldsymbol{x} - \boldsymbol{x_0}||^2$
  - $\boldsymbol{\phi_2}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x_0}) + <\nabla \boldsymbol{f}(\boldsymbol{x_0}), \boldsymbol{x} - \boldsymbol{x_0}> + \frac{\boldsymbol{L}}{2}||\boldsymbol{x} - \boldsymbol{x_0}||^2$
- 则函数$\boldsymbol{f}$的图像位于$\boldsymbol{\phi_1}$和$\boldsymbol{\phi_2}$之间，即
  - $\boldsymbol{\phi_1}(\boldsymbol{x}) \le \boldsymbol{f}(\boldsymbol{x}) \le \boldsymbol{\phi_2}(\boldsymbol{x}), \forall \boldsymbol{x} \in \boldsymbol{R}^n$

# 第四章 一维搜索与无约束最优化

- 定理：令$f \in C_L^{2,2}(R^n), \forall x, y \in R^n$,我们有
  - $||\nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y-x)|| \leq \frac{L}{2}||y-x||^2$
  - $|f(y) - f(x) - < \nabla f(x), y-x > - \frac{1}{2} < \nabla^2 f(x)(y-x), y-x > | \leq \frac{L}{6}||y-x||^3$
  - 证明略
- 推论：令$f \in C_L^{2,2}(R^n)$和$x, y \in R^n$,满足$||y-x|| = r$,则有
- $\nabla^2 f(x) - LrI_n \leq \nabla^2 f(y) \leq \nabla^2 f(x) + LrI_n$
- 证明：令$G = \nabla^2 f(y) - \nabla^2 f(x)$,因为$f \in C_L^{2,2}(R^n)$,从而$||G|| < Lr$,因此矩阵$G$的特征值$|\lambda_i| \leq Lr, i = 1, \cdots, n$, 因此$-LrI_n \leq G \leq LrI_n$

- 梯度法及其收敛性分析
  - $x_0 \in R^n$
  - $x_{k+1} = x_k - h_k \nabla f(x_k), k = 0, 1, \cdots$,步长$h_k > 0$
- $h_k$的选取有很多变形
  - $\{\mathrm{h_k}\}_{k=0}^{\infty}$:如$h_k = h > 0, h_k = \frac{h}{\sqrt{k+1}}$
  - 全松弛（精确步长）：$h_k = argmin_{h\geq 0} f(x_k - h\nabla f(x_k))$
  - Armijo规则：对$h > 0$,确定$x_{k+1} = x_k - h\nabla f(x_k)$,满足
    - $\alpha < \nabla f(x_k), x_k - x_{k+1} > \leq f(x_k) - f(x_{k+1})$
    - $\beta < \nabla f(x_k), x_k - x_{k+1} > \geq f(x_k) - f(x_{k+1})$
    - 其中,$0 < \alpha < \beta < 1$是一些固定参数.
- 看第三种策略，确定$x \in R^n, \nabla \mathrm{f}(\mathrm{x}) \neq 0$,则只需研究单变量函数$\phi(h) = f\left(x - h\nabla f(x)\right), h \geq 0$

# 第四章 一维搜索与无约束最优化-梯度法收敛性分析

- 看第三种策略，确定$\boldsymbol{x} \in \boldsymbol{R}^n, \nabla \mathbf{f}(\mathbf{x}) \neq \mathbf{0}$,则只需研究单变量函数$\boldsymbol{\phi}(\boldsymbol{h}) = \boldsymbol{f}\left(\boldsymbol{x} - \boldsymbol{h}\nabla \boldsymbol{f}(\boldsymbol{x})\right), \boldsymbol{h} \geq \mathbf{0}$
  - 则Armijo规则表明可接受的步长值对应于函数$\boldsymbol{\phi}$的图像的特定部分，该部分介于两个线性函数的图像之间
    - $\boldsymbol{\phi}_1(\boldsymbol{h}) = \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{\alpha}\boldsymbol{h}||\nabla \boldsymbol{f}(\boldsymbol{x})||^2, \boldsymbol{\phi}_2 = \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{\beta}\boldsymbol{h}||\nabla \boldsymbol{f}(\boldsymbol{x})||^2$
  - 注意$\boldsymbol{\phi}(\mathbf{0}) = \boldsymbol{\phi}_1(\mathbf{0}) = \boldsymbol{\phi}_2(\mathbf{0}), \boldsymbol{\phi}'(\mathbf{0}) < \boldsymbol{\phi}_2'(\mathbf{0}) < \boldsymbol{\phi}_1'(\mathbf{0}) < \mathbf{0}$,因此除非$\boldsymbol{\phi}$没有下界，否则可接受的步长总是存在。
- 考虑问题$\boldsymbol{min}_{\boldsymbol{R}^n}\boldsymbol{f}(\mathbf{x})$，满足$\boldsymbol{f} \in \boldsymbol{C}_L^{1,1}(\boldsymbol{R}^n)$,并假设函数$\boldsymbol{f}$在$\boldsymbol{R}^n$有下界
  - 考虑$\boldsymbol{y} = \boldsymbol{x} - \boldsymbol{h}\nabla \boldsymbol{f}(\boldsymbol{x})$,此时$\boldsymbol{f}(\boldsymbol{y}) \leq \boldsymbol{f}(\boldsymbol{x}) + < \nabla \boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} > + \frac{\boldsymbol{L}}{2}||\boldsymbol{y} - \boldsymbol{x}||^2 = \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{h}||\nabla \boldsymbol{f}(\boldsymbol{x})||^2 + \frac{\boldsymbol{h}^2}{2}\boldsymbol{L}||\nabla \boldsymbol{f}(\boldsymbol{x})||^2 = \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{h}\left(\mathbf{1} - \frac{\boldsymbol{h}}{2}\boldsymbol{L}\right)||\nabla \boldsymbol{f}(\boldsymbol{x})||^2$
  - 为了获得减少量的最优上界，必须解$\boldsymbol{\Delta}(\boldsymbol{h}) = -\boldsymbol{h}\left(\mathbf{1} - \frac{\boldsymbol{h}}{2}\boldsymbol{L}\right) \rightarrow \boldsymbol{min}_{\boldsymbol{h}}$,计算其导数，得最优步长必满足方程：$\boldsymbol{\Delta}'(\boldsymbol{h}) = \boldsymbol{h}\boldsymbol{L} - \mathbf{1} = \mathbf{0}$.因为$\boldsymbol{\Delta}''(\boldsymbol{h}) = \boldsymbol{L} > \mathbf{0}$,因此$\boldsymbol{h}^* = \frac{\mathbf{1}}{\boldsymbol{L}}$就是$\boldsymbol{\Delta}(\boldsymbol{h})$的极小点。表明一步至少按$\boldsymbol{f}(\boldsymbol{y}) \leq \boldsymbol{f}(\boldsymbol{x}) - \frac{\mathbf{1}}{\mathbf{2}\boldsymbol{L}}||\nabla \boldsymbol{f}(\boldsymbol{x})||^2$来降低目标函数值。

# 第四章 一维搜索与无约束最优化-梯度法收敛性分析

- 令$x_{k+1}=x_k-h_k\nabla f(x_k)$,则对于定步长策略，$h_k=h$,我们有：
  - $f(x_k)-f(x_{k+1})\geq h\left(1-\frac{1}{2}Lh\right)||\nabla f(x_k)||^2$
- 因此，如果选择$h_k=\frac{2\alpha}{L}$,满足$\alpha\in(0,1)$，则
  - $f(x_k)-f(x_{k+1})\geq\frac{2}{L}\alpha(1-\alpha)||\nabla f(x_k)||^2$
  - 当然最优选择为$h_k=\frac{1}{L}$
- 对于全松弛策略，我们有
  - $f(x_k)-f(x_{k+1})\geq\frac{1}{2L}||\nabla f(x_k)||^2$
- 最大的减少量不会比步长为$h_k=\frac{1}{L}$的情形差
- 最后，对于Armijo规则，有
  - $f(x_k)-f(x_{k+1})\leq\beta<\nabla f(x_k),x_k-x_{k+1}>=\beta h_k||\nabla f(x_k)||^2$
- 根据下降法一步梯度推导：
  - $f(x_k)-f(x_{k+1})\geq h_k\left(1-\frac{1}{2}Lh_k\right)||\nabla f(x_k)||^2$
- 因此$h_k\geq\frac{2}{L}(1-\beta)$,根据Armijo第一条规则，有

# 第四章 一维搜索与无约束最优化-梯度法收敛性分析

- $f(x_k) - f(x_{k+1}) \geq \alpha < \nabla f(x_k), x_k - x_{k+1} >= \alpha h_k ||\nabla f(x_k)||^2$
- 结合$h_k \geq \frac{2}{L}(1-\beta)$,则可以得到
  - $f(x_k) - f(x_{k+1}) \geq \frac{2}{L}\alpha(1-\beta)||\nabla f(x_k)||^2$
- 从而证明了，所有条件下都满足
  - $f(x_k) - f(x_{k+1}) \geq \frac{\omega}{L}||\nabla f(x_k)||^2$,其中$\omega$是一个正常数
- 梯度法的性能如何?
  - 从$f(x_k) - f(x_{k+1}) \geq \frac{\omega}{L}||\nabla f(x_k)||^2$能得到什么?
  - $\frac{\omega}{L}\sum_{k=0}^{N}||\nabla f(x_k)||^2 \leq f(x_0) - f(x_{N+1}) \leq f(x_0) - f^*$
  - $f^*$为目标函数值的下界，上式有界，从而$\lim_{k\to\infty}||\nabla f(x_k)|| \to 0$

# 第四章 一维搜索与无约束最优化-梯度法收敛性分析

- 收敛率如何?
  - 定义$g_N^* = min_{0\le k\le N}||\nabla f(x_k)||$,根据$\frac{\omega}{L}\sum_{k=0}^{N}||\nabla f(x_k)||^2 \le f(x_0) - f(x_{N+1}) \le f(x_0) - f^*$，可得
  - $g_N^* \le \frac{1}{\sqrt{N+1}}\left[\frac{1}{\omega}L(f(x_0) - f^*)\right]^{\frac{1}{2}}$
  - 这描述了序列$\{g_N^*\}$收敛到0的速率
- 一般的非线性优化中，只想找到接近优化问题的局部极小点，但这个目标，有时候对梯度法也实现不了。
- 例：考虑下列函数：$f(x) = f(x_1, x_2) = \frac{1}{2}x_1^2 + \frac{1}{4}x_2^4 - \frac{1}{2}x_2^2$,其梯度为$\nabla f(x) = \left(x_1, x_2^3 - x_2\right)^T$,其稳定点分别为$x^{(1)} = (0,0)^T, x^{(2)} = (0,-1)^T, x^{(3)} = (0,1)^T$.计算其Hessian矩阵
- $\nabla^2 f(x) = \begin{bmatrix} 1 & 0 \\ 0 & 3x_2^2 - 1 \end{bmatrix}$,容易判断三个稳定点的极值情况.其中$x^{(1)}$为稳定点，但非极值点，$f\left(x^{(1)}\right) = 0$,对任意$\epsilon > 0, f\left(x^{(1)} + \epsilon e_2\right) = \frac{\epsilon^4}{4} - \frac{\epsilon^2}{2} < 0$
- 此外，以$x_0 = (1,0)$为梯度法的初始点，则其迭代路径产生的序列收敛到$x^{(1)} = (0,0)^T$。因此，对于一阶无约束极小化方法，若没有额外限制，不能保证全局收敛到一个局部极小点，只能靠近稳定点。

# 第四章 一维搜索与无约束最优化-梯度法收敛性分析

- 研究如下问题类：
  - 模型：$\mathbf{1}$.无约束极小化，2.$\boldsymbol{f} \in \boldsymbol{C}_{\boldsymbol{L}}^{\mathbf{1,1}}(\boldsymbol{R}^{\boldsymbol{n}})$,3.$\boldsymbol{f}^*$是$\boldsymbol{f}(\cdot)$的下界
  - Oracle：一阶黑箱
  - $\boldsymbol{\epsilon}$ −最优解：$\boldsymbol{f}(\bar{\boldsymbol{x}}) \leq \boldsymbol{f}(\boldsymbol{x}_0), ||\nabla \boldsymbol{f}(\bar{\boldsymbol{x}})|| \leq \boldsymbol{\epsilon}$

- 注意到$\boldsymbol{g}_N^* \leq \frac{1}{\sqrt{N+1}}\left[\frac{1}{\omega}L(f(x_0)-f^*)\right]^{\frac{1}{2}}$用于得到迭代次数的上界，这对找到梯度范数小的点很必要。为此,令$\boldsymbol{g}_N^* \leq \frac{1}{\sqrt{N+1}}\left[\frac{1}{\omega}L(f(x_0)-f^*)\right]^{\frac{1}{2}} \leq \boldsymbol{\epsilon} \Rightarrow$如果$\boldsymbol{N}+\mathbf{1} \geq \frac{L}{\omega\epsilon^2}(\boldsymbol{f}(\boldsymbol{x}_0)-\boldsymbol{f}^*)$,则必然有$\boldsymbol{g}_N^* \leq \boldsymbol{\epsilon}$，因此我们用$\boldsymbol{g}_N^* \leq \frac{1}{\sqrt{N+1}}\left[\frac{1}{\omega}L(f(x_0)-f^*)\right]^{\frac{1}{2}}$来作为该问题类的复杂度上界，注意这个上界比之前用均匀网格法的上界$\left(\left\lfloor\frac{L}{2\epsilon}\right\rfloor+1\right)^n$更好，与$\boldsymbol{n}$无关！但其复杂度下界还未知!

# 第四章 一维搜索与无约束最优化-梯度法收敛性分析

- 下面研究梯度法的局部收敛怎么描述!考虑无约束极小化问题
- $\boldsymbol{min_{x\in R^n} f(x)}$,满足如下假设
  - $\boldsymbol{f \in C_M^{2,2}(R^n)}$
  - $\boldsymbol{f}$有一个局部极小值点$\boldsymbol{x^* \in R^n}$,该点处的Hessian矩阵正定
  - 该点处的Hessian矩阵，知道其上下界$\boldsymbol{0 < \mu \le L < \infty}$,即
$$\boldsymbol{\mu I_n \le \nabla^2 f(x^*) \le L I_n}$$
  - 初始点$\boldsymbol{x_0}$足够接近$\boldsymbol{x^*}$
- 研究迭代过程：$\boldsymbol{x_{k+1} = x_k - h_k \nabla f(x_k)}$,因$\boldsymbol{\nabla f(x^*) = 0}$，因此
- $\boldsymbol{\nabla f(x_k) = \nabla f(x_k) - \nabla f(x^*) = \int_0^1 \nabla^2 f\big(x^* + \tau(x_k - x^*)\big)(x_k - x^*)d\tau = G_k(x_k - x^*)}$,其中$\boldsymbol{G_k = \int_0^1 \nabla^2 f\big(x^* + \tau(x_k - x^*)\big)d\tau}$.因此
- $\boldsymbol{x_{k+1} - x^* = x_k - x^* - h_k G_k(x_k - x^*) = (I_n - h_k G_k)(x_k - x^*)}$
- 此时跟前面提到 的收缩映射有关。设序列$\boldsymbol{\{a_k\}}$定义为
- $\boldsymbol{a_0 \in R^n, a_{k+1} = A_k a_k}$,其中$\boldsymbol{A_k}$是$\boldsymbol{n \times n}$矩阵，$\boldsymbol{\forall k \ge 0, q \in (0,1)}$，有$\boldsymbol{||A_k|| \le 1 - q}$,这样估计序列$\boldsymbol{\{a_k\}}$的收敛速度

# 第四章 一维搜索与无约束最优化-梯度法收敛性分析

- 估计序列$\{a_k\}$的收敛到0的速度
- $||a_{k+1}|| \le (1-q)||a_k|| \le (1-q)^{k+1}||a_0|| \to 0$
- 现在估计$||I_n - h_k G_k||$。令$r_k = ||x_k - x^*||$,根据之前函数类的推论有：$\nabla^2 f(x^*) - \tau M r_k I_n \le \nabla^2 f\left(x^* + \tau(x_k - x^*)\right) \le \nabla^2 f(x^*) + \tau M r_k I_n$,结合条件3
- $\left(\mu - \frac{r_k}{2}M\right) I_n \le G_k \le \left(L + \frac{r_k}{2}M\right) I_n$，从而有
- $\left(1 - h_k\left(L + \frac{r_k}{2}M\right)\right) I_n \le I_n - h_k G_k \le (1 - h_k\left(\mu - \frac{r_k}{2}M\right)) I_n$,得
- $||I_n - h_k G_k|| \le max\{a_k(h_k), b_k(h_k)\}$,
- 其中$a_k(h) = 1 - h\left(\mu - \frac{r_k}{2}M\right), b_k(h) = h\left(L + \frac{r_k}{2}M\right) - 1$
- 注意到$a_k(0) = 1, b_k(0) = -1$,因此如果$0 < r_k < \bar{r} = \frac{2\mu}{M}$,则$a_k(\cdot)$是一个严格递减函数，对足够小的$h_k$可确保$||I_n - h_k G_k|| < 1$,此时有$r_{k+1} < r_k$

# 第四章 一维搜索与无约束最优化-梯度法收敛性分析

- 步长选择策略，如$h_k=\frac{1}{L}$，极小化$||I_n-h_kG_k||\le max\{a_k(h_k),b_k(h_k)\}\to \min_h$的右端项,假设$r_0<\bar{r}$,我们利用最优策略序列得到序列$\{x_k\}$,可以保证$r_{k+1}<r_k<\bar{r}$.进一步最优步长可从方程
- $a_k(h)=b_k(h)\Leftrightarrow 1-h\left(\mu-\frac{r_k}{2}M\right)=h\left(L+\frac{r_k}{2}M\right)-1$得到
- $h_k^*=\frac{2}{L+\mu}$,注意与M无关，此时
- $r_{k+1}\le\frac{(L-\mu)r_k}{L+\mu}+\frac{Mr_k^2}{L+\mu}$
- 现在估计迭代过程的收敛速度，令$q=\frac{2\mu}{L+\mu}$, $a_k=\frac{M}{L+\mu}r_k(<q)$，则
- $a_{k+1}\le(1-q)a_k+a_k^2=a_k\big(1+(a_k-q)\big)=\frac{a_k(1-(a_k-q)^2)}{1-(a_k-q)}\le\frac{a_k}{1+q-a_k}$，因此$\frac{1}{a_{k+1}}\ge\frac{1+q}{a_k}-1$

# 第四章 一维搜索与无约束最优化-梯度法收敛性分析

- 或者$\frac{q}{a_{k+1}}-1\ge\frac{q(1+q)}{a_k}-q-q=(1+q)\left(\frac{q}{a_k}-1\right)$
- 所以，$\frac{q}{a_k}-1\ge(1+q)^k\left(\frac{q}{a_0}-q\right)=(1+q)^k\left(\frac{2\mu}{L+\mu}\cdot\frac{L+\mu}{r_0M}-1\right)=(1+q)^k\left(\frac{\bar{r}}{r_0}-1\right)$
- 从而有$a_k\le\frac{qr_0}{r_0+(1+q)^k(\bar{r}-r_0)}\le\frac{qr_0}{\bar{r}-r_0}\left(\frac{1}{1+q}\right)^k$
- 定理：设函数$f(\cdot)$满足我们的假设，且初始点$x_0$足够接近一个严格局部极小点$x^*$,即

$$r_0=||x_0-x^*||<\bar{r}=\frac{2\mu}{M}$$

则步长为$h_k^*=\frac{2}{L+\mu}$的梯度法收敛如下：

- $||x_k-x^*||\le\frac{\bar{r}r_0}{\bar{r}-r_0}\left(1-\frac{2\mu}{L+3\mu}\right)^k$,这种收敛速度称为线性收敛

- 牛顿法
  - 最开始用于单变量函数求根.令$\boldsymbol{\phi}(\cdot): \boldsymbol{R} \to \boldsymbol{R}$,考虑$\boldsymbol{\phi}(\boldsymbol{t}^*) = \mathbf{0}$,其原理由线性近似得到。假设知道距$\boldsymbol{t}^*$足够近的$\boldsymbol{t} \in \boldsymbol{R}$,注意到$\boldsymbol{\phi}(\boldsymbol{t} + \boldsymbol{\Delta t}) = \boldsymbol{\phi}(\boldsymbol{t}) + \boldsymbol{\phi}'(\boldsymbol{t})\boldsymbol{\Delta t} + \boldsymbol{o}(|\Delta t|)$,因此方程$\boldsymbol{\phi}(\boldsymbol{t} + \boldsymbol{\Delta t}) = \mathbf{0}$的解可用右端来近似为：$\boldsymbol{\phi}(\boldsymbol{t}) + \boldsymbol{\phi}'(\boldsymbol{t})\boldsymbol{\Delta t}$=0,在某些条件下，我们希望增量$\boldsymbol{\Delta t}$是最优增量$\boldsymbol{\Delta t}^* = \boldsymbol{t}^* - \boldsymbol{t}$的一个好的近似.将其转化为算法得到：$\boldsymbol{t}_{k+1} = \boldsymbol{t}_k - \frac{\boldsymbol{\phi}(\boldsymbol{t}_k)}{\boldsymbol{\phi}'(\boldsymbol{t}_k)}$
  - 该算法可自然推广到解非线性方程组的问题：$\boldsymbol{F}(\boldsymbol{x}) = \mathbf{0}, \boldsymbol{x} \in \boldsymbol{R}^n, \boldsymbol{F}(\cdot): \boldsymbol{R}^n \to \boldsymbol{R}^n$。此时定义一个增量$\boldsymbol{\Delta x}$是下面线性方程组的解：$\boldsymbol{F}(\boldsymbol{x}) + \boldsymbol{F}'(\boldsymbol{x})\boldsymbol{\Delta x} = \mathbf{0}$(牛顿系统）。如果Jacobian矩阵$\boldsymbol{F}'(\boldsymbol{x})$非退化，我们可以计算增量$\boldsymbol{\Delta x} = -[\boldsymbol{F}'(\boldsymbol{x})]^{-1}\boldsymbol{F}(\boldsymbol{x})$,相应的迭代方法如下：
  - $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - [\boldsymbol{F}'(\boldsymbol{x}_k)]^{-1}\boldsymbol{F}(\boldsymbol{x}_k)$
  - 最优，由无约束优化的必要条件求解$\boldsymbol{\nabla f}(\boldsymbol{x}) = \mathbf{0}$来代替求解$\boldsymbol{minf}(\boldsymbol{x})$(非退化情况），为解必要条件，使用标准的解非线性方程组的牛顿法，这时，牛顿系统为：$\boldsymbol{\nabla f}(\boldsymbol{x}) + \boldsymbol{\nabla}^2\boldsymbol{f}(\boldsymbol{x})\boldsymbol{\Delta x} = \mathbf{0}$
  - 因此，对优化问题，牛顿法写为：

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \left[\boldsymbol{\nabla}^2\boldsymbol{f}(\boldsymbol{x}_k)\right]^{-1}\boldsymbol{\nabla f}(\boldsymbol{x}_k)$$

  - 注意，也可以用二阶逼近来推导!

如何用二阶逼近
来推导牛顿法？

- 牛顿法在严格局部极小点的一个邻域内，其收敛速度非常快
  - 约束1：若$\nabla^2 f(x_k)$退化，则牛顿法失败
  - 约束2：邻域内，否则可能不收敛
- 例：使用牛顿法求解$\phi(t)=\frac{t}{\sqrt{1+t^2}}$的一个根。显然，$t^*=0$,但$\phi'(t)=\frac{1}{(1+t^2)^{\frac{3}{2}}}$,因此牛顿法的步骤如下：
  - $t_{k+1}=t_k-\frac{\phi(t_k)}{\phi'(t_k)}=t_k-\frac{t_k}{\sqrt{1+t_k^2}}\cdot(1+t_k^2)^{\frac{3}{2}}=-t_k^3$
  - 若$|t_0|<1$，则算法收敛且收敛速度很快
  - 点$t_0=\pm1$为算法的震荡点
  - $|t_0|>1$，则算法发散
- 为避免可能的发散，在实际中可以使用阻尼牛顿法：

$$x_{k+1}=x_k-h_k\left[\nabla^2 f(x_k)\right]^{-1}\nabla f(x_k),$$ 其中$h_k$为步长参数

# 第五章 无约束最优化方法及其收敛性分析-牛顿法

- 注意，一般$\boldsymbol{h_k}$使用梯度法中的步长策略，但最后阶段，选择$\boldsymbol{h_k = 1}$比较合理。
- 下面推导牛顿法的局部收敛速度
- 考虑问题$\boldsymbol{min_{x \in R^n} f(x)}$
- 满足如下假设：
  - 1.$\boldsymbol{f} \in \boldsymbol{C_M^{2,2}(R^n)}$
  - 2.函数$\boldsymbol{f}$存在一个局部极小点$\boldsymbol{x}^*$，该点处Hessian矩阵正定：$\boldsymbol{\nabla^2 f(x^*) \succcurlyeq \mu I_n, \mu > 0}$
  - 3.初始点$\boldsymbol{x_0}$足够接近$\boldsymbol{x}^*$
- $\boldsymbol{x_{k+1} = x_k - \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k)}$的收敛速度如何？

# 第五章 无约束最优化方法及其收敛性分析-牛顿法

**推论：令$f \in C_L^{2,2}(R^n)$和$x, y \in R^n$,满足$||y-x|| = r$,则有**
**$\nabla^2 f(x) - LrI_n \le \nabla^2 f(y) \le \nabla^2 f(x) + LrI_n$**

- 与梯度法类似
- $x_{k+1} - x^* = x_k - x^* - [\nabla^2 f(x_k)]^{-1} \nabla f(x_k) = x_k - x^* - [\nabla^2 f(x_k)]^{-1} \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*))(x_k - x^*) d\tau = [\nabla^2 f(x_k)]^{-1} G_k (x_k - x^*)$
- 其中$G_k = \int_0^1 \left(\nabla^2 f(x_k) - \nabla^2 f\left(x^* + \tau(x_k - x^*)\right)\right) d\tau$
- 令$r_k = ||x_k - x^*||$,则
- $||G_k|| = ||\int_0^1 (\nabla^2 f(x_k) - \nabla^2 f(x^* + \tau(x_k - x^*))) d\tau|| \le \int_0^1 ||(\nabla^2 f(x_k) - \nabla^2 f(x^* + \tau(x_k - x^*)))|| d\tau \le \int_0^1 M(1-\tau) r_k d\tau = \frac{r_k}{2} M$
- 因此，由之前P71页推论我们有：$\nabla^2 f(x_k) \succcurlyeq \nabla^2 f(x^*) - Mr_k I_n \succcurlyeq (\mu - Mr_k) I_n$
- 因此，若$r_k < \frac{\mu}{M}$，则$\nabla^2 f(x_k)$正定，且有$[\nabla^2 f(x_k)]^{-1} \le (\mu - Mr_k)^{-1}$
- 所以，对于足够小的$r_k (r_k \le \frac{2\mu}{3M})$,有$r_{k+1} \le \frac{Mr_k^2}{2(\mu - Mr_k)} (\le r_k)$

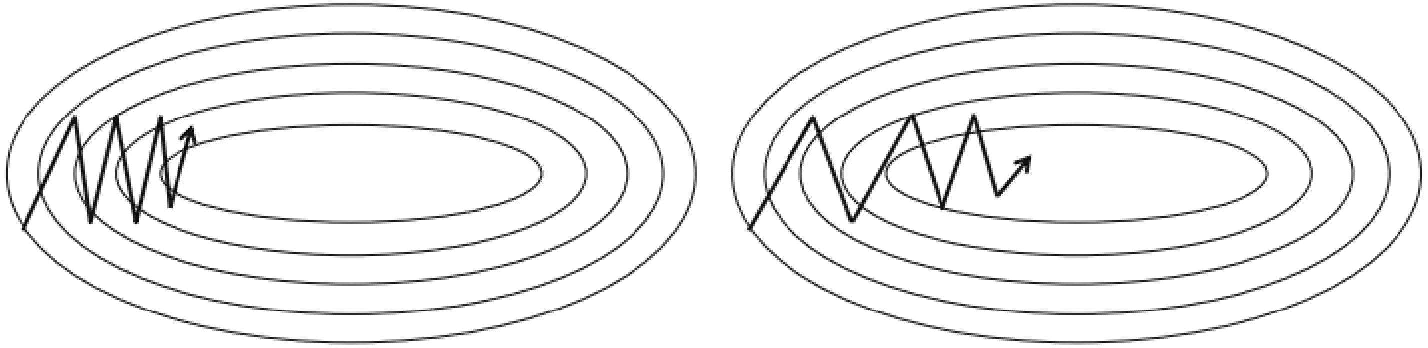
二次收敛

- **定理(牛顿法收敛性)**：令函数$\boldsymbol{f}(\cdot)$满足上述假设，假设初始点$\boldsymbol{x_0}$足够接近$\boldsymbol{x}^*$,即$||\boldsymbol{x_0}-\boldsymbol{x}^*|| \le \bar{\boldsymbol{r}}=\frac{2\boldsymbol{\mu}}{3\boldsymbol{M}}$
- 则对任意$\boldsymbol{k}$有$||\boldsymbol{x_k}-\boldsymbol{x}^*|| \le \bar{\boldsymbol{r}}$,牛顿法二次收敛，即：
- $||\boldsymbol{x_{k+1}}-\boldsymbol{x}^*|| \le \frac{\boldsymbol{M}||\boldsymbol{x_k}-\boldsymbol{x}^*||^2}{2(\boldsymbol{\mu}-\boldsymbol{M}||\boldsymbol{x_k}-\boldsymbol{x}^*||)}$
- 与梯度法的局部收敛率相比较，牛顿法显然收敛更快。
  - ！ 牛顿法的二次收敛区域与梯度法的线性收敛区域几乎相同
  - 表明：标准推荐极小化过程的初始阶段使用梯度法来接近局部极小点，然后再利用牛顿法快速收敛到最优点！

- 通过梯度法和牛顿法的收敛率，其与复杂度的界之间的对应关系，可知这些问题类的解析复杂度的上界是收敛率的反函数
  - 次线性速率。该速率由迭代计算器的幂函数来表示。例如，假设对于某算法可以证明其收敛率为$\boldsymbol{r_k \le \frac{c}{\sqrt{k}}}$，此时对于相应的问题类，该算法的复杂度上界是$\left(\frac{c}{\epsilon}\right)^2$ $(\because r_k \le \frac{c}{\sqrt{k}} < \epsilon \Rightarrow k > \left(\frac{c}{\epsilon}\right)^2)$
  - 次线性速率是比较慢的，就复杂度而言，最优值中每得到一位新的正确数字都需要经历与以前的总的工作量相当的迭代次数。注意常数$\boldsymbol{c}$对相应的复杂度的上界影响很大
  - 线性速度。该速率是根据迭代计数器的指数函数给出的。例如，$\boldsymbol{r_k \le c(1-q)^k \le ce^{-qk}, 0 < q \le 1}$,注意其相应的复杂度上界是$\boldsymbol{\frac{1}{q}\left(\ln c + \ln\frac{1}{\epsilon}\right)}$$(\because ce^{-qk} < \epsilon \Rightarrow e^{-qk} < \frac{\epsilon}{c} \Rightarrow e^{qk} > \frac{c}{\epsilon} \Rightarrow qk > \ln\left(\frac{c}{\epsilon}\right) \Rightarrow k > \frac{1}{q}(lnc + \ln\left(\frac{1}{\epsilon}\right))$)
  - 这个速率很快：最优值中每得到一位新的正确数字需要经历大约固定的迭代次数。此外，复杂度估计对于常数$\boldsymbol{c}$的依赖性非常弱
  - 二次收敛速率。该速率对迭代计数器都是双指数依赖的。例如，$\boldsymbol{r_{k+1} \le cr_k^2}$,相应复杂度估计依赖于所需精度的双对数：$\boldsymbol{\ln\ln\frac{1}{\epsilon}}$
  - 这个收敛率非常快：每次迭代都会双倍增加最优值的正确数字。常数$\boldsymbol{c}$仅对二次收敛的开始时刻很重要（$\boldsymbol{cr_k < 1}$）。例如，在$\boldsymbol{cr_k \le \frac{1}{2}}$之后，我们可以保证一个较大的收敛速率$\boldsymbol{r_{k+1} \le \frac{1}{2} r_k}$,不再依赖于常数$\boldsymbol{c}$。

# 第五章 有约束最优化方法及其收敛性分析-牛顿法

- 非线性优化中的一阶方法
  - 在梯度下降法中，选取$\boldsymbol{h_k} = \frac{1}{L}$,一般其收敛速度为$\mathcal{O}\left(\frac{1}{k}\right)$,若目标函数为$\boldsymbol{\mu}$强凸函数(满足：$\boldsymbol{f(\alpha x + (1-\alpha)y) \le \alpha f(x) + (1-\alpha)f(y) - \frac{\mu\alpha(1-\alpha)}{2}||x-y||^2}$),则其收敛速率为$\mathcal{O}\left(\left(1-\frac{\mu}{L}\right)^k\right)$
  - 加速梯度法可将收敛速率提至$\mathcal{O}\left(\frac{1}{k^2}\right)$和$\mathcal{O}\left(\left(1-\sqrt{\frac{\mu}{L}}\right)^k\right)$
  - 加速梯度法$\boldsymbol{x_0 = x_{-1} \in R^n}$
    - $\boldsymbol{y_k = x_k + \beta_k(x_k - x_{k-1}), x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k), k = 0, 1, \cdots,}$
    - 例如令$\boldsymbol{\beta_k = \frac{k-1}{k+2}}$
  - 



- 二阶加速算法

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \le L_2\|x-y\|_2,\ \forall x, y \in \mathbb{R}^n$$

| | Lower Bound | Upper Bound |
|---|---|---|
| $p=2$ | $\Omega\left(\left(\frac{L_2\|x_0-x^*\|^3}{\epsilon}\right)^{\frac{2}{11}}\right)$ [Agarwal and Hazan (2018)] | $O\left(\left(\frac{L_2\|x_0-x^*\|^3}{\epsilon}\right)^{\frac{1}{3}}\right)$ [Nesterov (2008)] |
| | $\Omega\left(\left(\frac{L_2\|x_0-x^*\|^3}{\epsilon}\right)^{\frac{2}{7}}\right)$ [Arjevani et al. (2018)] [Nesterov (2018)] | $\tilde{O}\left(\left(\frac{L_2\|x_0-x^*\|^3}{\epsilon}\right)^{\frac{2}{7}}\right)$ [Monteiro, Svaiter (2013)] |

Key idea, second order approximation:

$$f(x) \approx f(x^i) + (x-x^i)^\top \nabla f(x^i) + \frac{1}{2}(x-x_i)^\top \nabla^2 f(x^i)(x-x^i)$$

| | 下界 | 上界 |
|---|---|---|
| $p=1$ | $\Omega\left(\left(\frac{L_1\|\|x_0-x^*\|\|^2}{\epsilon}\right)^{\frac{1}{2}}\right)$ [Nemirovski,Yudin(1983)] | $\mathcal{O}\left(\left(\frac{L_1\|\|x_0-x^*\|\|^2}{\epsilon}\right)^{\frac{1}{2}}\right)$ [Nesterov(1983)] |
| $p=2$ | $\Omega\left(\left(\frac{L_2\|\|x_0-x^*\|\|^3}{\epsilon}\right)^{\frac{2}{7}}\right)$ [Arjevani et.al (2018)] | $\mathcal{O}\left(\left(\frac{L_2\|\|x_0-x^*\|\|^3}{\epsilon}\right)^{\frac{2}{7}}\right)$ [Monteiro,Svaiter(2013)] |
| $p\geq 3$ | $\Omega\left(\left(\frac{L_p\|\|x_0-x^*\|\|^{p+1}}{\epsilon}\right)^{\frac{2}{3p+1}}\right)$ [Arjevani et al(2018)] [Nesterov[2018]] | $\mathcal{O}\left(\left(\frac{L_p\|\|x_0-x^*\|\|^{p+1}}{\epsilon}\right)^{\frac{1}{p+1}}\right)$ [Baes(2009)] [Nesterov [2018]] |
| $p\geq 3$ | | $\mathcal{O}\left(\left(\frac{L_p\|\|x_0-x^*\|\|^{p+1}}{\epsilon}\right)^{\frac{2}{3p+1}}\right)$ |

Bo Jiang, Haoyue Wang, Shuzhong Zhang, Mathematics of OperationsResearch, published online, 2021 (posted on arXiv:1812.06557, 2018)

# 自适应三次正则化方法

- 优化问题求解基本步骤是通过对目标函数取一次\二次泰勒展开式估计得到一个近似模型,这样构成原问题的子问题
  - 在实际计算中目标函数的精确矩阵用近似矩阵取代，通常采取校正公式更新目标函数的精确Hessian矩阵
- 或者，假设目标函数的二阶导函数是Lipschitz连续的，Lipschitz常数为$L$，利用二阶导函数的Lipschitz连续性得到一个带三次项的近似模型，这个带三次项的近似模型构成原问题的子问题，且具有更好的精确性和自适应性
- 非线性无约束优化问题的求解
  - 拟牛顿法
  - 信赖域方法
  - 共轭梯度法

# 自适应三次正则化方法

- 信赖域方法，和线搜索方法通常联合使用来解决非线性无约束优化问题
- 自适应性的三次正则化方法是通过对目标函数取一个三次过高估计，作为一种正则技术来计算从上一步迭代到下一步迭代的步长
- 目标函数$f$二次连续可微，其Hessian矩阵满足常数为$L$的Lipschitz连续性假设，迭代中从$x^{(k+1)} \to x^{(k)}, x^{(k+1)} = x^{(k)} + s^{(k)}$,将其在$x^{(k)}$展开，则有：
- $f\left(x^{(k)}+s^{(k)}\right)=f\left(x^{(k)}\right)+{s^{(k)}}^T\nabla f\left(x^{(k)}\right)+\frac{1}{2}{s^{(k)}}^T\nabla^2 f\left(x^{(k)}\right)s^{(k)}+\int_0^1(1-\tau){s^{(k)}}^T\left[\nabla^2 f\left(x^{(k)}+\tau s^{(k)}\right)-\nabla^2 f\left(x^{(k)}\right)\right]s^{(k)}d\tau \le f\left(x^{(k)}\right)+{s^{(k)}}^T\nabla f\left(x^{(k)}\right)+\frac{1}{2}{s^{(k)}}^T\nabla^2 f\left(x^{(k)}\right)s^{(k)}+\int_0^1(1-\tau)\tau L||s^{(k)}||^3 d\tau = \boldsymbol{f\left(x^{(k)}\right)+{s^{(k)}}^T\nabla f\left(x^{(k)}\right)+\frac{1}{2}{s^{(k)}}^T\nabla^2 f\left(x^{(k)}\right)s^{(k)}+\frac{1}{6}L||s^{(k)}||^2 = m_k^c\left(s^{(k)}\right)}, \forall s^{(k)} \in R^n$
- 因此只要存在$s^{(k)} \in R^n$满足下式：

$$m_k^c\left(s^{(k)}\right) < m_k^c(0) = f\left(x^{(k)}\right),$$

- 则目标函数$f$在$x^{(k)}+s^{(k)}$处的三次正则过高估计为$m_k^c\left(s^{(k)}\right)$，从而可以通过求$m_k^c\left(s^{(k)}\right)$的一个局部最优解$s^{(k)}$来确定目标函数$f$在$x^{(k)}$点处下降迭代点$x^{(k+1)} = x^{(k)} + s^{(k)}$,所以有下式成立
- $f\left(x^{(k+1)}\right) = f\left(x^{(k)}+s^{(k)}\right) \le f\left(x^{(k)}\right)$

# 自适应三次正则化方法

- 实际上$\int_0^1 (1-\tau){s^{(k)}}^T\left[\nabla^2 f\left(x^{(k)}+\tau s^{(k)}\right)-\nabla^2 f\left(x^{(k)}\right)\right]s^{(k)}d\tau = \int_0^1 (1-\tau){s^{(k)}}^T\nabla^2 f\left(x^{(k)}+\tau s^{(k)}\right)s^{(k)}d\tau - \int_0^1 (1-\tau){s^{(k)}}^T\nabla^2 f\left(x^{(k)}\right)s^{(k)}d\tau = \int_0^1 (1-\tau){s^{(k)}}^T d\nabla f\left(x^{(k)}+\tau s^{(k)}\right) - \frac{1}{2}{s^{(k)}}^T\nabla^2 f\left(x^{(k)}\right)s^{(k)} = -{s^{(k)}}^T\nabla f\left(x^{(k)}\right)+\int_0^1 \nabla f\left(x^{(k)}+\tau s^{(k)}\right)d\tau - \frac{1}{2}{s^{(k)}}^T\nabla^2 f\left(x^{(k)}\right)s^{(k)} = -{s^{(k)}}^T\nabla f\left(x^{(k)}\right)+f\left(x^{(k)}+s^{(k)}\right)-f\left(x^{(k)}\right)-\frac{1}{2}{s^{(k)}}^T\nabla^2 f\left(x^{(k)}\right)s^{(k)}$
- 这里$\tau\in(0,1)$

# 自适应三次正则化方法

- Griewank用一个与$k$有关的参数$\sigma_k > 0$,代替前面三次估计模型$m_k^c(s^{(k)})$中的Lipschitz常数$\frac{1}{2}L$,得到一个新的模型$m_k^G(s)$，记为
- $m_k^G(s) = f\left(x^{(k)}\right) + s^T \nabla f\left(x^{(k)}\right) + \frac{1}{2} s^T \nabla^2 f\left(x^{(k)}\right) s + \frac{1}{3}\sigma_k ||s||^3$
- 显然$m_k^G(s)$是牛顿二次近似的一种正则化形式，称之为目标函数$f(x)$在$x^{(k)}$处的自适应三次正则估计
- Coralia Cartis, N.I.M.Gould, Ph.L.Tonint2009年推广Griewank的结果：目标函数$f(x)$一阶可微，不要求其Hessian矩阵局部或全局连续，在迭代中利用拟牛顿校正公式计算对称矩阵来近似Hessian矩阵$\nabla^2 f\left(x^{(k)}\right) \approx B_k$，记$g_k = \nabla f(x^{(k)})$,则可得如下新的自适应三次正则估计模型：
- $m_k(s) = f(x_k) + s^T g_k + \frac{1}{2} s^T B_k s + \frac{1}{3}\sigma_k ||s||^3$

# 自适应三次正则化方法

- 利用自适应性的三次正则化估计去求解非线性无约束优化问题，提出了自适应性的三次正则化算法并在一定的条件下证明了自适应性三次正则化算法的超线性收敛和全局收敛，同时他们还进行了基于自适应性三次正则化算法的大规模数值实验，数值结果表明该算法比信赖域算法收敛速度要快速且更稳定，因此自适应性的三次正则化算法是一种比较好的求解非线性无约束优化的算法
- 证明了自适应性三次正则化算法的迭代步数的上界和梯度计算的复杂性

Cartis, C., Gould, N.I.M., Toint, Ph.L. Adaptive cubic overestimation methods for un-constrained optimization. Part I: Motivation,convergence and numerical results [J]. Math.Program.,2010; Ser. A 127(2),245-295.

Cartis, C., Gould, N.I.M.,Toint,Ph.L.Adaptive cubic regularisation methods for uncon-strained optimization.Part II: worst-case function- and derivative-evaluation complexity [J].Mathematical Programming. 2011; 130(1),295-319.

# 拟牛顿方法

- 美国物理学家Davidon于20世纪50年代中期,第一次提出用拟牛顿算法求解非线性无约束优化问题.
- 后来Powell和Fletche证明了由Davidon提出的拟牛顿算法比其他的无约束优化算法更有效.
- 在接下来的20多年时间里，拟牛顿算法成了非线性无约束优化问题算法的研究热点，
- 经过长期对拟牛顿算法的研究,人们得到一些比较经典的结论，如20世纪60年代末Broyden, Fletcher, Goldfrab和Shanno分别从不同的角度提出了BFGS校正公式，
- BFGS校正公式的拟牛顿算法成为当今求解非线性无约束优化最有效的拟牛顿算法.
- 还有其他一些比较有影响力的成果,如对称秩-1校正公式和Broyden族校正公式的拟牛顿算法
- 像最速下降法一样,拟牛顿方法在每次迭代计算只要求知道目标函数的梯度,而且不要求目标函数的二阶导数,所以拟牛顿方法比牛顿方法在某些问题上是更有效的.在当今求解优化问题的实际算法中，人们通常比较喜欢用拟牛顿方法去求解大规模的非线性无约束优化问题和非线性约束优化问题.下面来介绍关于一些拟牛顿校正公式的由来，首先来了解目标函数$f(x)$在当前迭代点$x_k$处的二次近似模型如下

$$m_k(s) = f(x_k) + s^T g_k + \frac{1}{2} s^T G_k s$$

# 拟牛顿方法

- 关于目标函数$f(x)$做以下三个说明
  - 目标函数$f(x)$在$R^n$二阶可微
  - 目标函数$f(x)$在$x_k$处的梯度记为$g_k$
  - 目标函数$f(x)$在$x_k$处的Hessian矩阵记为$G_k$
- 我们用$m_k(s)$表示目标函数$f(x)$在$x_k$处的二阶泰勒展开式，当$s$的范数$||s||$足够小，则可以认为目标函数$f(x)$在$x_k+s$处的函数值$f(x_k+s)$近似等于$m_k(s)$. 因此目标函数$f(x)$在$x_k$的二次近似模型如下

$$f(x_k+s)\approx f(x_k)+s^T g_k+\frac{1}{2}s^T G_k s$$

- 所以目标函数$f(x)$在$x_{k+1}$处的二次近似模型为

$$f(x)\approx f_{k+1}+g_{k+1}^T(x-x_{k+1})+\frac{1}{2}(x-x_{k+1})^T G_{k+1}(x-x_{k+1})$$

# 拟牛顿方法

- 对上面的逼近式求导数为0，可得

$$g(x) \approx g_{k+1} + G_{k+1}(x - x_{k+1})$$

- 令$x = x_k$,则有：

$$g_k \approx g_{k+1} + G_{k+1}(x_k - x_{k+1})$$

- 与以前类似，令$s_k = x_{k+1} - x_k, y_k = g_{k+1} - g_k$,则有

$$G_{k+1}s_k \approx y_k$$

- 但二阶Hessian矩阵计算量大，采用对称矩阵$B_{k+1}$来近似$G_{k+1}$

$$B_{k+1}s_k = y_k$$

拟牛顿方程

- 从而可以达到两个拟牛顿校正公式

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k)(\mathrm{y_k} - \mathrm{B_k s_k})^{\mathrm{T}}}{(\mathrm{y_k} - \mathrm{B_k s_k})^{\mathrm{T}}\mathrm{s_k}},$$

对称秩1校正

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k},$$

对称秩2校正

# 自适应性的拟牛顿校正公式

- 自适应性的拟牛顿方程
  - 在自适应性的三次正则拟牛顿方法中,我们用一个近似矩阵$B_k$取代目标函数的Hessian矩阵$G_k$，为了确保算法更有效, $B_k$需要满足以下3个条件
    - $B_k$应该充分的接近$G_k$,即$B_k \approx G_k$,这样可以使$x_k$处的线搜索方向近似于牛顿方向,以至于算法有一个好的收敛性
    - 对于任意正整数$k$, $B_k$是一个对称正定矩阵,这样可以保证$x_k$处的线搜索方向是目标函数的下降方向
    - 对于$B_k$的校正公式:$B_{k+1} = B_k + E_k$,其中$E_k$矩阵的秩为1或2.
  - 根据目标函数自适应性的三次正则化估计模型,我们可以得到目标函数在$x_{k+1}$处的自适应性的三次正则化近似模型为:
  - $f(x) \approx f(x_{k+1}) + s^T g_{k+1} + \frac{1}{2} s^T G_{k+1} s + \frac{1}{3}\sigma_{k+1}||s||^3, s = x - x_{k+1}$
  - 对式两边对$x$求导得:$g_k \approx g_{k+1} + G_{k+1}(x - x_{k+1}) + \sigma_{k+1}||x - x_{k+1}||(x - x_{k+1})$
  - 式中，令$x = x_k, s_k = x_{k+1} - x_k, y_k = g_{k+1} - g_k$，则化为:
  - $g_{k+1} - g_k \approx G_{k+1}s_k + \sigma_k||s_k||s_k$
  - 这等价于$y_k \approx (G_{k+1} + \sigma_k||s_k||I)s_k$

谢谢！