



统计估计

Statistical Estimation · Chapter 7

梁经纬 Jingwei Liang

Institute of Natural Sciences
School of Mathematical Sciences

Lecture 07: 统计估计

2026 年秋季学期

饮水思源 · 爱国荣校

本讲内容

参数分布估计 §7.1
最大似然、最大后验、常见噪声模型与协方差估计



非参数分布估计 §7.2
概率单纯形、矩约束、最大熵、KL 散度与风险界



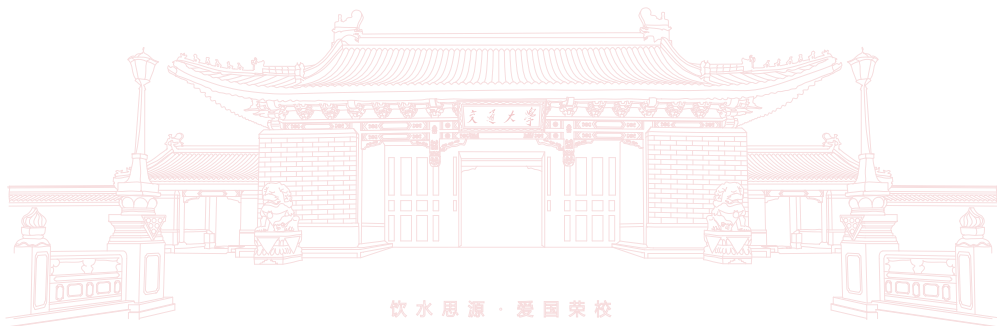
检测器与假设检验 §7.3
检测矩阵、ROC、随机化检测器与稳健设计



概率界与实验设计 §§7.4–7.5
Chebyshev/Chernoff 界、信息矩阵与 A/D/E 最优设计

01

前序知识与统计建模



饮水思源 · 爱国荣校

与前序课程的连接

Lectures 01B–01C, 02

凸函数、log-sum-exp、范数、正则化与回归



Lecture 04

共轭函数、对偶变量、entropy 与 support function



Lectures 03, 05–06

Newton/KKT 系统与 interior-point solvers

本讲重新解释这些对象

损失函数 = $-\log$ 似然;

正则项 = $-\log$ 先验;

约束集 = 已有统计信息;

对偶变量 = 矩或风险约束的边际代价。

优化模型由概率假设决定; 求解仍调用前面建立的凸性与数值方法。

从数据到估计量



参数估计

分布族的形式已知，有限维参数 θ 未知，例如均值、回归系数或协方差矩阵。

非参数估计

直接把离散分布 $p \in \mathbb{R}^n$ 作为变量，用矩、概率或熵刻画已有信息。

何时得到凸问题

最大化形式

若对每个观测 y , 对数似然

$$\ell(\theta) = \log p_{\theta}(y)$$

关于 θ 为凹函数, 并且先验约束集 C 为凸集, 则

$$\underset{\theta \in C}{\text{maximize}} \ell(\theta)$$

是凸优化问题。

关键: “统计估计可由凸优化求解”依赖参数化方式; 换变量常把非凸表达改成凸表达。

最小化形式

等价地定义负对数似然

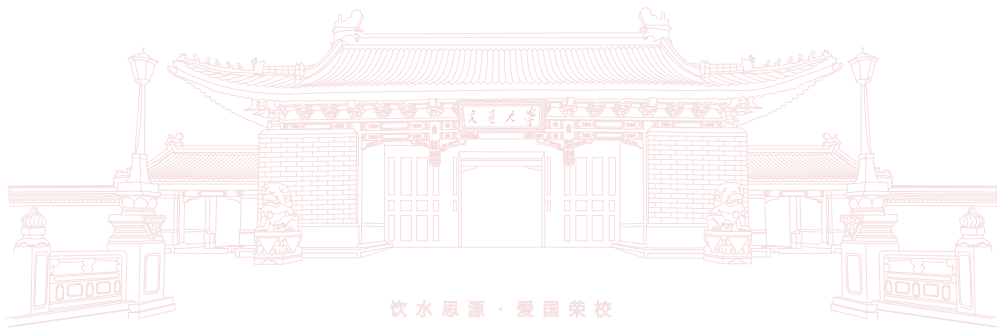
$$L(\theta; y) = -\log p_{\theta}(y).$$

若 L 为凸函数, 则多个独立样本对应

$$\min_{\theta \in C} \sum_{i=1}^N L(\theta; y_i).$$

02

参数分布估计



饮水思源 · 爱国荣校

最大似然估计

观测 $y \in \mathbb{R}^m$ 来自参数为 $x \in \mathbb{R}^n$ 的分布，密度记为 $p_x(y)$ 。最大似然估计 (maximum likelihood, ML) 为

$$\hat{x}_{\text{ml}} \in \arg \max_{x \in C} p_x(y) = \arg \max_{x \in C} \ell(x), \quad \ell(x) = \log p_x(y).$$

变量与参数

优化变量是分布参数 x ；已观测的样本 y 在优化问题中是固定数据。

先验约束

C 可表达物理范围、符号、守恒或结构信息；也可令 $p_x(y) = 0$ 于 $x \notin C$ 。

问题 若两个参数产生相同的最大似然值，估计是否唯一？还需要哪些结构信息？

Boyd & Vandenberghe, *Convex Optimization*, §7.1.1, equation (7.1)

线性测量与 IID 噪声

$$y_i = \mathbf{a}_i^\top \mathbf{x} + v_i, \quad i = 1, \dots, m, \quad v_i \stackrel{\text{iid}}{\sim} p.$$

独立性把联合似然分解为乘积，因此

$$p_{\mathbf{x}}(\mathbf{y}) = \prod_{i=1}^m p(y_i - \mathbf{a}_i^\top \mathbf{x}), \quad \ell(\mathbf{x}) = \sum_{i=1}^m \log p(y_i - \mathbf{a}_i^\top \mathbf{x}).$$

优化形式

令 $A = [\mathbf{a}_1^\top; \dots; \mathbf{a}_m^\top] \in \mathbb{R}^{m \times n}$ ，残差 $r_i = y_i - \mathbf{a}_i^\top \mathbf{x}$ ：

$$\min_{\mathbf{x}} \sum_{i=1}^m \phi(r_i), \quad \phi(t) = -\log p(t).$$

若 p 对数凹，则 ϕ 为凸函数。

$\mathbf{x} \in \mathbb{R}^n$ ：待估参数



$A\mathbf{x} \in \mathbb{R}^m$ ：模型预测



$\mathbf{y} - A\mathbf{x} \in \mathbb{R}^m$ ：残差

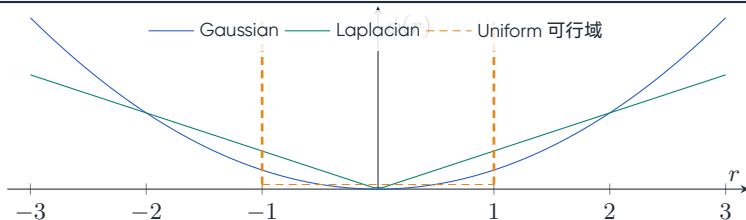


$\sum_i -\log p(r_i)$ ：总损失

Boyd & Vandenberghe, *Convex Optimization*, §7.1.1, equation (7.2)

噪声模型决定损失函数

噪声密度 $p(r)$	$-\log p(r)$ (忽略常数)	ML 模型	残差特征
Gaussian $\propto e^{-r^2/(2\sigma^2)}$	$r^2/(2\sigma^2)$	$\min \ Ax - y\ _2^2$	强调大残差
Laplacian $\propto e^{- r /a}$	$ r /a$	$\min \ Ax - y\ _1$	对离群点更稳健
Uniform on $[-a, a]$	0 或 $+\infty$	$\ Ax - y\ _\infty \leq a$	所有残差有硬界



Boyd & Vandenberghe, *Convex Optimization*, Example 7.1

应用例 1: 传感器偏置估计

四次标定读数为 $y = (1.0, 1.2, 0.8, 4.0)$, 模型 $y_i = x + v_i$ 。比较 Gaussian 与 Laplacian 假设:

$$\hat{x}_2 = \arg \min_x \sum_i (y_i - x)^2 = 1.75, \quad \hat{x}_1 \in \arg \min_x \sum_i |y_i - x| = [1.0, 1.2].$$

平方损失

样本均值受读数 4.0 明显拉动; 这与 Gaussian 噪声的轻尾假设一致。

绝对值损失

任一中位数均最优; 单个大离群点不会产生平方级影响。



自编数值例; 模型依据 §71.1

损失与密度的双向解释

给定凸罚函数 ϕ ，若归一化常数有限，则可构造

$$p(z) = \frac{e^{-\phi(z)}}{\int_{-\infty}^{\infty} e^{-\phi(u)} du}.$$

因此任何可归一化的可分离罚函数逼近

$$\min_{\mathbf{x}} \sum_{i=1}^m \phi(b_i - \mathbf{a}_i^\top \mathbf{x})$$

都可解释为某个 IID 噪声模型下的 ML 估计。

增长快的 ϕ

密度尾部衰减快，模型认为大残差极不可能，因此强烈惩罚离群点。

建模含义：选择损失函数等价于声明残差的概率结构。

增长慢的 ϕ

密度具有更重的尾部，模型允许少量大残差，估计通常更稳健。

Boyd & Vandenberghe, *Convex Optimization*, §7.1.1, “ML interpretation of penalty function approximation”

Poisson 计数模型

若 $Y \sim \text{Poisson}(\mu)$, 则

$$\mathbf{P}(Y = k) = \frac{e^{-\mu} \mu^k}{k!}, \quad \mu > 0.$$

令均值由解释变量 $\mathbf{u}_i$ 决定: $\mu_i = \mathbf{a}^\top \mathbf{u}_i + b$ 。对数据 $(\mathbf{u}_i, y_i)$, 忽略常数后的对数似然为

$$\ell(\mathbf{a}, b) = \sum_{i=1}^m [y_i \log(\mathbf{a}^\top \mathbf{u}_i + b) - (\mathbf{a}^\top \mathbf{u}_i + b)],$$

定义域要求 $\mathbf{a}^\top \mathbf{u}_i + b > 0$ 。每一项关于 $(\mathbf{a}, b)$ 为凹函数。

典型数据

光子到达数、事故数、呼叫次数、故障次数等非负整数计数。

与最小二乘的区别

Poisson 方差等于均值, 噪声尺度随信号变化;
恒方差 Gaussian 模型并不匹配。

Boyd & Vandenberghe, *Convex Optimization*, §7.1.1, “Counting problems with Poisson distribution”

应用例 2：事故频次与交通流量

假设某路段三天的标准化车流量为 $u = (0, 1, 2)$ ，事故数为 $y = (1, 2, 4)$ ，取 $\mu_i = au_i + b$ 。最大似然问题为

$$\max_{a,b} \sum_{i=1}^3 [y_i \log(au_i + b) - (au_i + b)] \quad \text{s.t.} \quad au_i + b > 0.$$

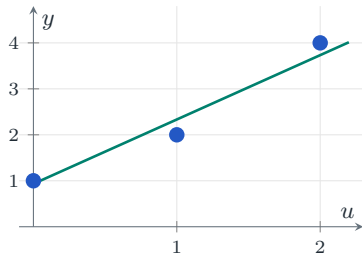
一阶条件写为

$$\sum_i \left(\frac{y_i}{\mu_i} - 1 \right) = 0, \quad \sum_i u_i \left(\frac{y_i}{\mu_i} - 1 \right) = 0.$$

解得

$$\hat{a} = \frac{7}{5}, \quad \hat{b} = \frac{14}{15}, \quad \hat{\mu} = \left(\frac{14}{15}, \frac{7}{3}, \frac{56}{15} \right).$$

两个方程分别对应截距与斜率方向的似然平衡。



自编数值例；模型依据 §7.1.1

逻辑回归

二元响应 $Y \in \{0, 1\}$ 的成功概率由

$$p(\mathbf{u}) = \mathbf{P}(Y = 1 \mid \mathbf{u}) = \frac{e^{\mathbf{a}^\top \mathbf{u} + b}}{1 + e^{\mathbf{a}^\top \mathbf{u} + b}}$$

给出。对样本 $(\mathbf{u}_i, y_i)$ ，负对数似然为

$$L(\mathbf{a}, b) = \sum_{i=1}^m \left[\log(1 + e^{\mathbf{a}^\top \mathbf{u}_i + b}) - y_i(\mathbf{a}^\top \mathbf{u}_i + b) \right].$$

$\log\text{-sum-exp}$ 为凸函数，因此 L 关于 $(\mathbf{a}, b)$ 凸。

参数含义

$\mathbf{a}$ 控制各特征对 $\log\text{-odds}$ 的影响， b 控制基准概率。

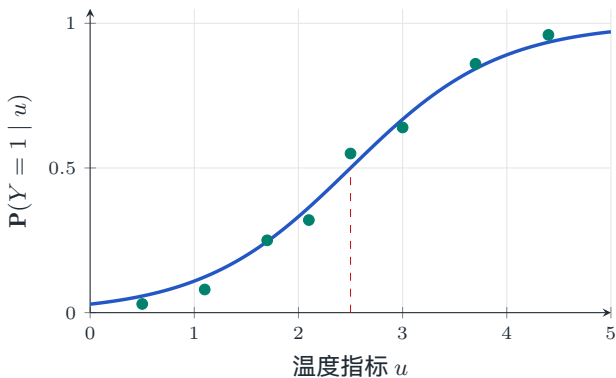
预测与估计

优化阶段估计概率模型；阈值化 $p(\mathbf{u})$ 后才得到分类规则。

Boyd & Vandenberghe, *Convex Optimization*, §7.1.1, equation (7.3)

应用例 3：设备故障概率

令 u 为归一化温度，取 $a = 1.4, b = -3.5$ ，则 $p(u) = \sigma(1.4u - 3.5)$ 。



解释

$u = 2.5$ 时 $p = 0.5$ ；温度每增加一单位，故障 odds 乘以 $e^{1.4}$ 。

建模检查

若样本完全可分，未正则化的 ML 参数可能无有限最优解；可加入 ℓ_2 先验或约束。

自编数值例；逻辑模型依据 §7.1.1

Gaussian 协方差估计

零均值 Gaussian 样本 $\mathbf{y}_1, \dots, \mathbf{y}_N \in \mathbb{R}^n$ 的协方差为 $R \succ 0$, 样本协方差为

$$Y = \frac{1}{N} \sum_{k=1}^N \mathbf{y}_k \mathbf{y}_k^\top.$$

以 R 为变量的对数似然并非全局凹; 改用信息矩阵 $S = R^{-1}$:

$$\max_{S \in \mathcal{S}} \log \det S - \text{tr}(SY), \quad S \succ 0.$$

目标对 S 为凹函数。无额外约束且 $Y \succ 0$ 时, 一阶条件 $S^{-1} = Y$ 给出

$$\hat{R}_{\text{ml}} = Y.$$

参数化的作用

同一统计模型在 R 与 $S = R^{-1}$ 下具有不同的凸性; 选择变量是建模的一部分。

结构化协方差

关于协方差 R 的先验可转化为信息矩阵 S 的凸约束:

$$L \preceq R \preceq U \iff U^{-1} \preceq S \preceq L^{-1}.$$

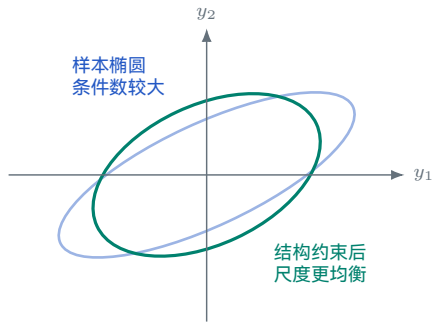
条件数界可写为存在 $u > 0$ 使

$$uI \preceq S \preceq \kappa_{\max} uI.$$

线性组合方差界

$$\mathbf{E}(\mathbf{c}_i^\top \mathbf{y})^2 = \mathbf{c}_i^\top S^{-1} \mathbf{c}_i \leq \alpha_i$$

也是关于 S 的凸约束。



Boyd & Vandenberghe, *Convex Optimization*, §7.1.1, covariance estimation for Gaussian variables

应用例 4: 两通道协方差正则化

样本协方差 $Y = \text{diag}(100, 1)$ 的条件数为 100。若要求 $\kappa(R) \leq 10$, 可解

$$\max_{S \succ 0, u > 0} \log \det S - \text{tr}(SY) \quad \text{s.t.} \quad uI \preceq S \preceq 10uI.$$

由于 $R = S^{-1}$ 与 S 条件数相同, 该模型避免估计椭圆过度扁长。

未约束 ML

$\hat{R} = Y$, 两个方向的标准差比为 10; 有限样本下可能非常不稳定。

结构化 ML

条件数约束牺牲部分似然, 换取更稳定的逆协方差与后续判别计算。

自编数值例; 结构约束依据 §7.1.1

最大后验估计

最大后验 (maximum a posteriori, MAP) 把参数 x 也视为随机变量。Bayes 公式给出

$$p(\boldsymbol{x} \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{x})p_{\boldsymbol{x}}(\boldsymbol{x}).$$

因此

$$\hat{\boldsymbol{x}}_{\text{map}} \in \arg \max_{\boldsymbol{x}} [\log p(\boldsymbol{y} \mid \boldsymbol{x}) + \log p_{\boldsymbol{x}}(\boldsymbol{x})].$$

似然项

说明数据在给定参数下出现的可能性；对应数据拟合项。

若似然和先验均对数凹，则 MAP 是凸优化问题。

先验项

说明参数在观测前的可信程度；负对数先验对应正则项。

Boyd & Vandenberghe, *Convex Optimization*, §7.1.2, equation (7.6)

线性测量下的 MAP

在线性模型 $y_i = \mathbf{a}_i^\top \mathbf{x} + v_i$ 中，若噪声密度为 p_v 、先验密度为 p_x ，则

$$\hat{\mathbf{x}}_{\text{map}} \in \arg \max_{\mathbf{x}} \left[\log p_x(\mathbf{x}) + \sum_{i=1}^m \log p_v(y_i - \mathbf{a}_i^\top \mathbf{x}) \right].$$

Gaussian 噪声与 $\mathbf{x} \sim \mathcal{N}(\bar{\mathbf{x}}, \Sigma)$ 给出

$$\min_{\mathbf{x}} \frac{1}{2\sigma^2} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}})^\top \Sigma^{-1}(\mathbf{x} - \bar{\mathbf{x}}).$$

均匀噪声 $v_i \in [-a, a]$ 与 Gaussian 先验给出 QP

$$\min_{\mathbf{x}} (\mathbf{x} - \bar{\mathbf{x}})^\top \Sigma^{-1}(\mathbf{x} - \bar{\mathbf{x}}) \quad \text{s.t.} \quad \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_\infty \leq a.$$

Boyd & Vandenberghe, *Convex Optimization*, §7.1.2, equation (7.7)

无噪声测量与最小罚函数

若测量精确满足 $Ax = y$, MAP 估计为

$$\max_x \log p_x(x) \quad \text{s.t.} \quad Ax = y.$$

若 x_i 独立同分布, $p_x(x) = \prod_i p(x_i)$, 则

$$\min_x \sum_{i=1}^n \phi(x_i) \quad \text{s.t.} \quad Ax = y, \quad \phi(t) = -\log p(t).$$

这正是 Lecture 02 的最小罚函数问题。

Gaussian 先验

$\phi(t) \propto t^2$, 得到最小二范数解。

Laplacian 先验

$\phi(t) \propto |t|$, 得到 basis pursuit, 并偏好稀疏解。

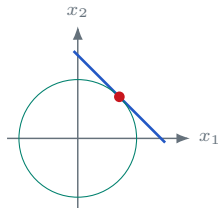
Boyd & Vandenberghe, *Convex Optimization*, §7.1.2, “MAP with perfect linear measurements”

应用例 5: 欠定测量中的稀疏 MAP

由一个测量 $x_1 + x_2 = 1$ 恢复 $\mathbf{x} \in \mathbb{R}^2$ 。

Gaussian 先验对应

$$\min x_1^2 + x_2^2 \quad \text{s.t.} \quad x_1 + x_2 = 1,$$

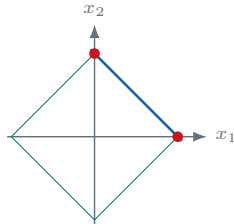


唯一解为 $(1/2, 1/2)$ 。

Laplacian 先验对应

$$\min |x_1| + |x_2| \quad \text{s.t.} \quad x_1 + x_2 = 1.$$

在非负象限整条线段均最优；加入微小权重差即

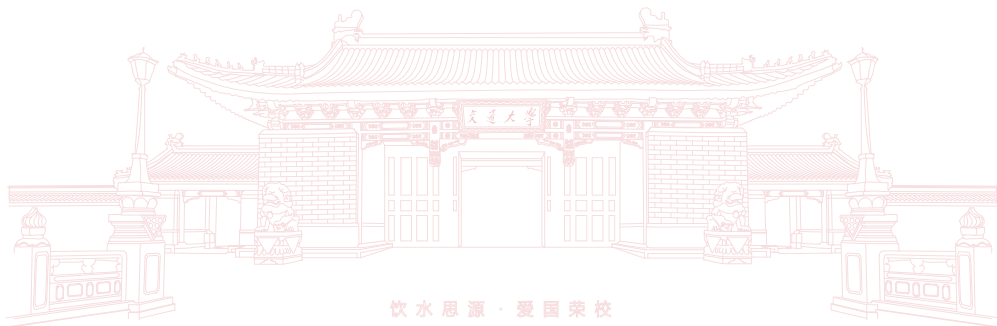


可选择端点稀疏解。

自编二维例；MAP 解释依据 §7.1.2

03

非参数分布估计

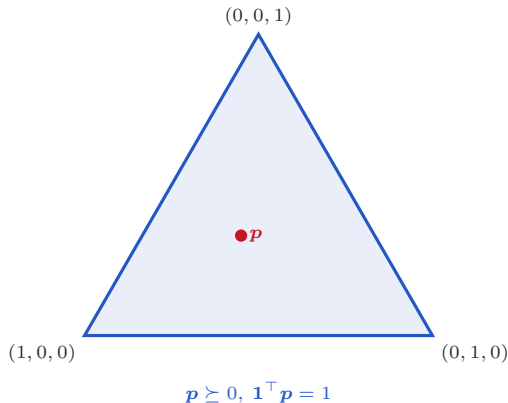


概率分布作为优化变量

随机变量 X 只取有限个值 $\{\alpha_1, \dots, \alpha_n\}$, 令

$$p_i = \mathbf{P}(X = \alpha_i), \quad \mathbf{p} \succeq 0, \quad \mathbf{1}^\top \mathbf{p} = 1.$$

所有可能分布恰好构成概率单纯形。



非参数的含义

不预先指定 Gaussian、Poisson 等分布族；直接估计每个支持点上的概率质量。

凸性基础

矩、事件概率和样本对数似然都可写成 $\mathbf{p}$ 的线性或凸函数。

Boyd & Vandenberghe, *Convex Optimization*, §7.2

矩与事件概率是线性的

对任意函数 f 与集合 C ,

$$\mathbf{E}f(X) = \sum_{i=1}^n p_i f(\alpha_i), \quad \mathbf{P}(X \in C) = \sum_{\alpha_i \in C} p_i.$$

因此已有信息可直接写成线性等式或不等式。例如

$$\sum_i \alpha_i p_i = \bar{\alpha}, \quad \sum_i \alpha_i^2 p_i = \beta, \quad \sum_{\alpha_i \geq 0} p_i \leq 0.3.$$

统计语言

已知均值、二阶矩和事件概率上界。

优化语言

概率单纯形与若干超平面的交集，仍是凸集。

非线性先验信息

方差

$$\text{var}(X) = \sum_i \alpha_i^2 p_i - \left(\sum_i \alpha_i p_i \right)^2$$

关于 p 为凹函数，因此下界

$$\text{var}(X) \geq \sigma_{\min}^2$$

定义凸可行集。

条件概率为线性分式函数

$$\mathbf{P}(X \in A \mid X \in B) = \frac{\mathbf{c}^\top \mathbf{p}}{\mathbf{d}^\top \mathbf{p}}.$$

在 $\mathbf{d}^\top \mathbf{p} > 0$ 时，区间约束等价于

$$l \mathbf{d}^\top \mathbf{p} \leq \mathbf{c}^\top \mathbf{p} \leq u \mathbf{d}^\top \mathbf{p}.$$

建模检查

“下界作用于凹函数”与“上界作用于凸函数”都产生凸约束；方向写反通常破坏凸性。

熵与 KL 散度

离散熵与 Kullback–Leibler 散度为

$$H(\mathbf{p}) = - \sum_{i=1}^n p_i \log p_i, \quad D_{\text{KL}}(\mathbf{p} \parallel \mathbf{q}) = \sum_{i=1}^n p_i \log \frac{p_i}{q_i}.$$

熵下界

H 为凹函数，所以 $H(\mathbf{p}) \geq h_{\min}$ 是凸约束。熵越大，概率质量越分散。

约定 $0 \log 0 = 0$ ；若 $p_i > 0, q_i = 0$ ，则 $D_{\text{KL}}(\mathbf{p} \parallel \mathbf{q}) = +\infty$ 。

KL 邻域

D_{KL} 对 $(\mathbf{p}, \mathbf{q})$ 联合凸；给定参考分布 $\mathbf{q}$ 时， $D_{\text{KL}}(\mathbf{p} \parallel \mathbf{q}) \leq \rho$ 是凸约束。

Boyd & Vandenberghe, *Convex Optimization*, §7.2; §3.1.5, Example 3.19

概率与期望的最坏情形界

设先验信息定义凸集合 $\mathcal{P}$ 。对函数 f ，所有相容分布下的精确下界由

$$\begin{array}{ll} \underset{\mathbf{p}}{\text{minimize}} & \sum_i f(\alpha_i) p_i \\ \text{subject to} & \mathbf{p} \in \mathcal{P} \end{array}$$

给出；最大化则给出上界。事件概率只需取 $f = \mathbf{1}_C$ 。

不是估计一个分布

目标是在所有相容分布中寻找使关注量最小或最大的分布。

鲁棒统计解释

最优值是由现有信息保证的最紧界；最优分布是达到该界的对抗性分布。

Boyd & Vandenberghe, *Convex Optimization*, §7.2, “Bounding probabilities and expected values”

离散分布的 ML 估计

观察 N 个独立样本，支持点 α_i 出现 k_i 次， $\sum_i k_i = N$ 。对数似然为

$$\ell(\mathbf{p}) = \sum_{i=1}^n k_i \log p_i.$$

带先验信息的 ML 问题为

$$\max_{\mathbf{p} \in \mathcal{P}} \sum_i k_i \log p_i.$$

仅有单纯形约束

拉格朗日条件给出 $\hat{p}_i = k_i/N$ ，即经验频率。

当 $k_i > 0$ 时最优解必须满足 $p_i > 0$ ，否则对数似然为 $-\infty$ 。

加入矩或形状约束

估计不再等于经验频率，但仍可由凹最大化稳定求解。

Boyd & Vandenberghe, *Convex Optimization*, §7.2, “Maximum likelihood estimation”

最大熵与最小 KL

在相容集合 $\mathcal{P}$ 中，可用两种准则选择单一分布：

$$\text{最大熵: } \min_{p \in \mathcal{P}} \sum_i p_i \log p_i,$$

$$\text{最小 KL: } \min_{p \in \mathcal{P}} \sum_i p_i \log(p_i/q_i).$$

最大熵

不引入约束之外的集中结构；在只有归一化约束时得到均匀分布。

最小 KL

在满足新信息的前提下尽量接近参考分布 q ；若 $q_i = 1/n$ ，退化为最大熵。

Boyd & Vandenberghe, *Convex Optimization*, §7.2, “Maximum entropy” and “Minimum Kullback–Leibler divergence”

例 7.2: 矩约束下的最大熵

在 $[-1, 1]$ 上取 100 个等距支持点 α_i , 寻找满足

$$\begin{aligned} \mathbf{E}X &\in [-0.1, 0.1], & \mathbf{E}X^2 &\in [0.5, 0.6], \\ \mathbf{E}(3X^3 - 2X) &\in [-0.3, -0.2], & \mathbf{P}(X < 0) &\in [0.3, 0.4] \end{aligned}$$

的最大熵分布。

所有约束对 p 线性, 因此问题为带线性约束的可分离凸优化。教材解满足

$$\begin{aligned} \mathbf{E}X &= 0.056, & \mathbf{E}X^2 &= 0.5, \\ \mathbf{E}(3X^3 - 2X) &= -0.2, & \mathbf{P}(X < 0) &= 0.4. \end{aligned}$$

活跃约束

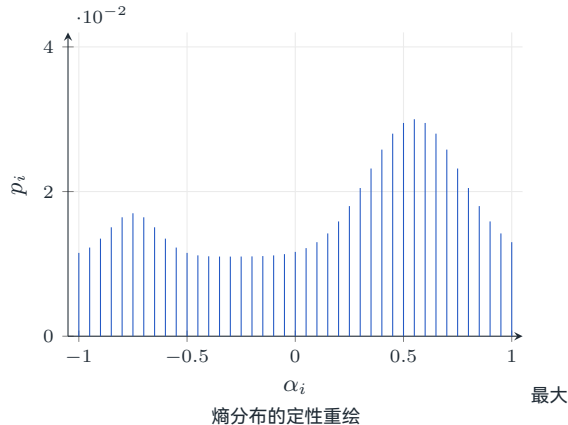
后面三个约束在教材最优解处取边界值, 说明它们真正限制了可达到的熵。

几何

目标的等值面在概率单纯形与约束多面体的交集中寻找切点。

Boyd & Vandenberghe, *Convex Optimization*, Example 7.2, equation (7.8)

例 7.2: 分布形状与 CDF 区间



对每个阈值 α_j , 分别解两个 LP:

$$\min / \max_{p \in \mathcal{P}} F(\alpha_j) = \sum_{i \leq j} p_i.$$

得到 CDF 的逐点上下界。最大熵 CDF 位于界之间, 但它并不等同于任一风险方向的最坏情形分布。

计算代价

若有 n 个阈值, 需要 $2n$ 个结构相同的 LP, 可并行求解或利用参数延续。

Boyd & Vandenberghe, *Convex Optimization*, Example 7.2, Figures 7.2–7.3 (定性重绘)

应用例 6: 骰子的最大熵分布

设骰子点数 $X \in \{1, \dots, 6\}$, 只知道 $\mathbb{E}X = 4$ 。最大熵问题为

$$\min_{\mathbf{p} \succeq 0} \sum_{i=1}^6 p_i \log p_i \quad \text{s.t.} \quad \mathbf{1}^\top \mathbf{p} = 1, \quad \sum_{i=1}^6 i p_i = 4.$$

KKT 条件给出指数族形式

$$p_i = \frac{e^{\lambda_i}}{\sum_{j=1}^6 e^{\lambda_j}},$$

其中 λ 由均值约束确定。

结构结论

一个线性矩约束产生一个指数权重；多个矩约束产生相应充分统计量的线性组合。

数值求解

原问题可直接解，也可用一维求根确定 λ ；二者给出同一分布。

自编例；最大熵模型依据 §7.2

例 7.3: 已知边缘分布的风险界

投资收益 $X \in \{\alpha_i\}_{i=1}^n$ 、 $Y \in \{\beta_j\}_{j=1}^m$ ，联合概率 p_{ij} 未知，但边缘分布已知：

$$\sum_j p_{ij} = r_i, \quad \sum_i p_{ij} = q_j, \quad p_{ij} \geq 0.$$

总收益低于 γ 的最坏概率由运输型 LP 给出：

$$\max_{p_{ij}} \sum_{\alpha_i + \beta_j < \gamma} p_{ij} \quad \text{s.t. 已知边缘分布约束.}$$

为什么需要优化

边缘分布不决定相关性；不同联合分布可能产生完全不同的组合尾部风险。

与最优运输的联系

p_{ij} 是把一个边缘分布耦合到另一个边缘分布的运输计划，目标对特定区域计费。

Boyd & Vandenberghe, *Convex Optimization*, Example 7.3, equation (79)

应用例 7: 相关性未知时的损失概率

$X, Y \in \{-1, +1\}$, 且各自取 -1 与 $+1$ 的概率均为 $1/2$ 。关注事件 $X + Y < 0$, 即 $(X, Y) = (-1, -1)$ 。

	$Y = -1$	$Y = +1$	row sum
$X = -1$	p_{11}	p_{12}	$1/2$
$X = +1$	p_{21}	p_{22}	$1/2$
column sum	$1/2$	$1/2$	

由边缘约束可知 $0 \leq p_{11} \leq 1/2$, 所以

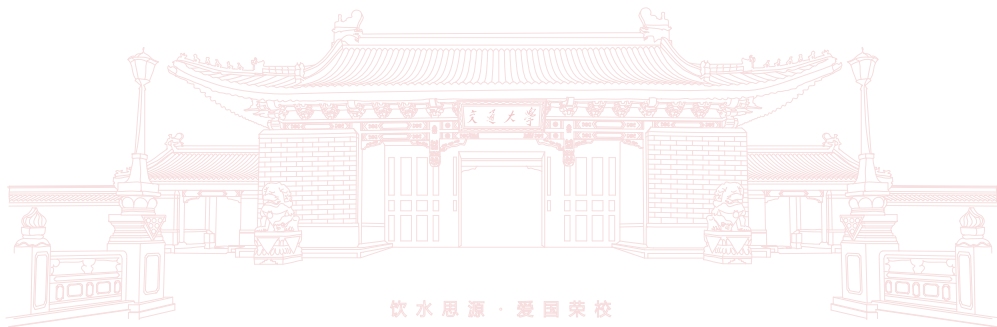
$$0 \leq \mathbf{P}(X + Y < 0) \leq \frac{1}{2}.$$

解释

独立性会给出 $1/4$, 但它不是由边缘分布推出的; LP 给出不额外假设相关结构时的精确范围。

04

检测器与假设检验



饮水思源 · 爱国荣校

检测器设计问题

未知假设 $\theta \in \{1, \dots, m\}$, 离散观测 $X \in \{1, \dots, n\}$ 。已知

$$P_{kj} = \mathbf{P}(X = k \mid \theta = j), \quad P \in \mathbb{R}^{n \times m},$$

即 P 的第 j 列是第 j 个假设下的观测分布。



目标

设计从观测 X 到判断 $\hat{\theta}$ 的映射, 使误判概率满足给定准则。

确定性与随机化检测器

检测器由 $T \in \mathbb{R}^{m \times n}$ 表示:

$$T_{ik} = \mathbf{P}(\hat{\theta} = i \mid X = k), \quad \mathbf{t}_k \succeq 0, \quad \mathbf{1}^\top \mathbf{t}_k = 1.$$

其中 $\mathbf{t}_k$ 是 T 的第 k 列。

确定性检测器

每一列 $\mathbf{t}_k$ 是某个单位向量; 观测 k 唯一对应一个判断。

随机化检测器

允许同一观测以给定概率作出不同判断; $\mathbf{t}_k$ 位于概率单纯形内部。

$$X = k \mapsto \hat{\theta} = i(k)$$

所有随机化检测器构成 n 个概率单纯形的笛卡尔积, 是凸集。

$$X = k \mapsto \hat{\theta} \sim \mathbf{t}_k$$

检测概率矩阵

联合两阶段条件概率，得到

$$D_{ij} = \mathbf{P}(\hat{\theta} = i \mid \theta = j) = \sum_{k=1}^n T_{ik} P_{kj}, \quad D = TP.$$

矩阵维度

$$\underbrace{T}_{m \times n} \underbrace{P}_{n \times m} = \underbrace{D}_{m \times m}.$$

D 的第 j 列描述真实假设为 j 时各判断的分布。

D 的每列和为 1；检测器设计是关于 T 的线性多目标问题。

正确与错误

D_{jj} 是正确检测概率； D_{ij} ($i \neq j$) 是将 j 错判为 i 的概率。

$$P_{e,j} = 1 - D_{jj}.$$

检测性能的线性约束与代价

检测矩阵 $D = TP$ 的每个元素都对设计变量 T 线性，因此常见的性能要求直接给出 LP:

$$D_{jj} \geq L_j, \quad D_{ij} \leq U_{ij} (i \neq j), \quad \min_T \max_j (1 - D_{jj}).$$

若假设对应有数值的参数 $\theta_1, \dots, \theta_m$ ，则真实为 θ_i 时

$$\text{bias}_i = \sum_{j=1}^m (\theta_j - \theta_i) D_{ji},$$

$$\text{MSE}_i = \sum_{j=1}^m (\theta_j - \theta_i)^2 D_{ji},$$

$$\mathbf{P}(|\hat{\theta} - \theta_i| > \delta \mid \theta_i) = \sum_{|\theta_j - \theta_i| > \delta} D_{ji}.$$

结论

偏差、均方误差、大误差概率以及正确检测率均是 D 的仿射函数；对它们加上界、取加权和或 minimax 仍得到 LP。

Boyd & Vandenberghe, *Convex Optimization*, §7.3.3, limits, bias, and mean-square error

多准则检测器设计

所有非对角元素 D_{ij} 都希望尽量小，可写成向量多目标问题

$$\text{minimize } \{D_{ij} \mid i \neq j\} \quad \text{s.t.} \quad D = TP, t_k \succeq 0, \mathbf{1}^\top t_k = 1.$$

常见标量化包括：

$$\text{minimax:} \quad \min_T \max_j (1 - D_{jj}),$$

$$\text{Bayes:} \quad \min_T \sum_j q_j (1 - D_{jj}),$$

$$\text{加权误判:} \quad \min_T \sum_{i \neq j} W_{ij} D_{ij}.$$

凸性

$D = TP$ 对 T 线性，列单纯形约束线性；上述三种模型均可写为 LP。

加权和解析决策规则

令 $\mathbf{c}_k$ 为 $WP^\top$ 的第 k 列，则加权目标可分解为

$$\sum_{i,j} W_{ij} D_{ij} = \sum_{k=1}^n \mathbf{c}_k^\top \mathbf{t}_k.$$

对每个观测 k 独立求解

$$\min_{\mathbf{t}_k \succeq 0, \mathbf{1}^\top \mathbf{t}_k = 1} \mathbf{c}_k^\top \mathbf{t}_k.$$

线性函数在单纯形顶点取到最优值，因此可选择确定性规则

$$\hat{\theta}(k) \in \arg \min_j (WP^\top)_{jk}.$$

结论的边界

每个固定权重矩阵都有确定性最优解；但非光滑的 minimax 准则可能在 Pareto 边之间选择随机化检测器。

MAP 与 ML 检测器

若假设先验为 $q_j = \mathbf{P}(\theta = j)$, Bayes 平均误差最小化给出

$$\hat{\theta}(k) \in \arg \max_j P_{kj} q_j.$$

这就是 MAP 检测器, 因为 $P_{kj} q_j$ 与后验概率 $\mathbf{P}(\theta = j \mid X = k)$ 成正比。

非均匀先验

即使某假设对观测 k 的似然略低, 较大的先验概率仍可能使其后验最大。

均匀先验

$q_j = 1/m$ 时, 规则化为

$$\hat{\theta}(k) \in \arg \max_j P_{kj},$$

即 ML 检测器。

二元假设检验

设假设 1 为正常、假设 2 为事件发生。检测矩阵写成

$$D = \begin{bmatrix} 1 - P_{\text{fp}} & P_{\text{fn}} \\ P_{\text{fp}} & 1 - P_{\text{fn}} \end{bmatrix}.$$

假阳性 (false positive)

真实为正常却判断为事件：代价可能是误报警、无效检修或额外检查。

ROC 的 Pareto 边界描述 P_{fp} 与 P_{fn} 之间可达到的最优权衡。

假阴性 (false negative)

真实有事件却判断为正常：代价可能是漏检故障、疾病或安全风险。

似然比阈值规则

令观测在两假设下的分布分别为 p_k 与 q_k 。加权误差最小化给出

$$\hat{\theta}(k) = \begin{cases} 1, & W_{21}p_k > W_{12}q_k, \\ 2, & W_{21}p_k \leq W_{12}q_k. \end{cases}$$

若 $q_k > 0$ ，等价于比较似然比

$$\frac{p_k}{q_k} \geq \frac{W_{12}}{W_{21}}.$$

改变阈值会依次改变各观测的判决，从而沿 ROC 的顶点移动。

顶点之间的线段由相邻规则随机化得到；这就是 Neyman–Pearson 结构。

Boyd & Vandenberghe, *Convex Optimization*, §7.3.5, Neyman–Pearson lemma

例 7.4：四种观测的二元检测

教材给出

$$P = \begin{bmatrix} .70 & .10 \\ .20 & .10 \\ .05 & .70 \\ .05 & .10 \end{bmatrix} = [p \ q].$$

似然比 p_k/q_k 依次为 7, 2, 1/14, 1/2。随着阈值降低，观测按 1, 2, 4, 3 的顺序被归入假设 1。

三个中间确定性规则

$$\{1, 2, 4\} \rightarrow (P_{fp}, P_{fn}) = (.05, .30),$$

$$\{1, 2\} \rightarrow (.10, .20),$$

$$\{1\} \rightarrow (.30, .10).$$

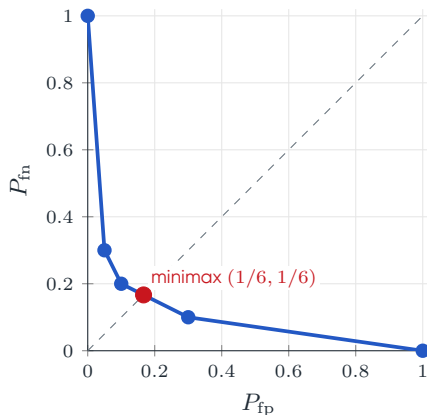
集合表示判断为假设 1 的观测。

端点

总判断为假设 1 得 (0, 1)；总判断为假设 2 得 (1, 0)。

五个点的下凸折线构成 ROC 的最优边界。

例 7.4: ROC 与随机化 minimax



在规则 (.10, .20) 与 (.30, .10) 之间随机化, 可到达对角线上的

$$P_{fp} = P_{fn} = \frac{1}{6}.$$

每个确定性规则至少有一个错误概率大于 $1/6$, 所以 minimax 最优检测器必须随机化。

Boyd & Vandenberghe, *Convex Optimization*, Example 7.4, Figure 7.4 (重绘)

应用例 8: 医学筛查阈值

某评分在健康与患病人群中的离散分布为

$$p = (.50, .30, .15, .05), \quad q = (.05, .15, .30, .50),$$

四档评分由低到高。将高于阈值者判为阳性。

阈值后阳性档位	P_{fp}	P_{fn}
{4}	.05	.50
{3, 4}	.20	.20
{2, 3, 4}	.50	.05

决策含义

漏诊代价较高时选择更低阈值；误检代价较高时选择更高阈值。阈值不是由分类准确率单独决定的。

若要求 $P_{fn} \leq .1$ ，第三个规则可行且在上述候选中使 P_{fp} 最小。

自编数值例；二元检测结构依据 §7.3.5

稳健检测器

当观测分布矩阵不确定，只知 $P \in \mathcal{P}$ ，定义最坏情形检测概率

$$D_{ij}^{\text{wc}} = \sup_{P \in \mathcal{P}} (TP)_{ij} \ (i \neq j), \quad D_{ii}^{\text{wc}} = \inf_{P \in \mathcal{P}} (TP)_{ii}.$$

相应最坏误差为 $P_{e,i}^{\text{wc}} = 1 - D_{ii}^{\text{wc}}$ 。稳健 minimax 设计等价于

$$\max_T \min_i \inf_{P \in \mathcal{P}} (TP)_{ii} \quad \text{s.t.} \quad \mathbf{t}_k \succeq 0, \mathbf{1}^\top \mathbf{t}_k = 1.$$

注意

D^{wc} 的不同元素可能由不同的 P 取得，因此其列和一般不等于 1；它不是某一个实际分布下的检测矩阵。

有限情景下的稳健设计

若不确定集合有限, $\mathcal{P} = \{P^{(1)}, \dots, P^{(K)}\}$, 引入保证水平 γ :

$$\begin{aligned} & \underset{T, \gamma}{\text{maximize}} && \gamma \\ & \text{subject to} && (TP^{(s)})_{ii} \geq \gamma, \quad i = 1, \dots, m, s = 1, \dots, K, \\ & && \mathbf{t}_k \succeq 0, \quad \mathbf{1}^\top \mathbf{t}_k = 1, \quad k = 1, \dots, n. \end{aligned}$$

这是 LP; 它同时保证每个假设、每个情景下的正确检测概率不低于 γ 。

情景扩充

增加一个可能分布只增加一组线性约束, 最优保证水平不会上升。

凸包不变性

把有限集合替换为其凸包不会改变最坏值, 因为 $(TP)_{ii}$ 关于 P 线性。

多面体不确定集与 LP 对偶

假设第 i 个观测分布满足

$$\mathcal{P}_i = \{\mathbf{p} \mid A_i \mathbf{p} = \mathbf{b}_i, \mathbf{1}^\top \mathbf{p} = 1, \mathbf{p} \succeq 0\}.$$

对检测器第 i 行 $\tilde{\mathbf{t}}_i^\top$, 最坏正确率的内层问题为

$$\inf_{\mathbf{p} \in \mathcal{P}_i} \tilde{\mathbf{t}}_i^\top \mathbf{p} = \sup_{\nu_i, \mu_i} \{\nu_i^\top \mathbf{b}_i + \mu_i \mid A_i^\top \nu_i + \mu_i \mathbf{1} \preceq \tilde{\mathbf{t}}_i\}.$$

把右侧对偶变量与 T, γ 一起优化, 得到单层 LP:

$$\nu_i^\top \mathbf{b}_i + \mu_i \geq \gamma, \quad A_i^\top \nu_i + \mu_i \mathbf{1} \preceq \tilde{\mathbf{t}}_i.$$

作用

对偶消除了“对所有相容分布”的无穷约束, 把稳健检测器直接变成有限维 LP。

Boyd & Vandenberghe, *Convex Optimization*, §7.3.6, robust minimax detector for polyhedral $\mathcal{P}$

例 7.5: 只知道均值与二阶矩

观测 $X \in \{a_1, \dots, a_n\}$ 。两个假设分别只给出

$$\mathbf{E}X = \alpha_1, \mathbf{E}X^2 = \alpha_2, \quad \mathbf{E}X = \beta_1, \mathbf{E}X^2 = \beta_2.$$

令 $\tilde{t}$ 为判断假设 1 的概率, 则两类最坏正确率都是矩约束 LP; 例如

$$\inf_{\mathbf{p} \succeq 0} \left\{ \tilde{t}^\top \mathbf{p} \mid \sum_i a_i p_i = \alpha_1, \sum_i a_i^2 p_i = \alpha_2, \mathbf{1}^\top \mathbf{p} = 1 \right\}.$$

其对偶寻找一条二次函数 $z_0 + z_1 a + z_2 a^2$, 在所有支持点上不超过检测概率 $\tilde{t}_i$ 。

几何解释

矩信息不确定时, 稳健检测器与“用低次多项式从下方逼近决策函数”联系起来。

应用例 9：两种工况下的稳健告警

振动等级 $X \in \{0, 1, 2\}$ 。正常工况的候选分布为 $p^{(1)} = (.70, .25, .05)$ 、 $p^{(2)} = (.60, .35, .05)$ ；故障工况为 $q^{(1)} = (.10, .30, .60)$ 、 $q^{(2)} = (.05, .35, .60)$ 。

阈值规则“ $X = 2$ 时告警”在四个情景中的

$$P_{\text{fp}} = .05, .05, \quad P_{\text{fn}} = .40, .40.$$

规则“ $X \geq 1$ 时告警”给出

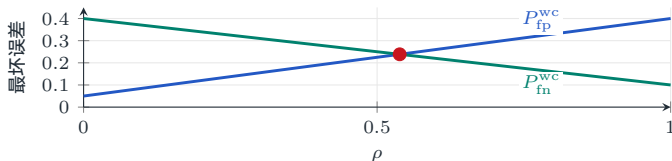
$$P_{\text{fp}} = .30, .40, \quad P_{\text{fn}} = .10, .05.$$

稳健选择

以 ρ 的概率使用“ $X \geq 1$ ”规则，则两类最坏误差为

$$.05 + .35\rho, \quad .40 - .30\rho.$$

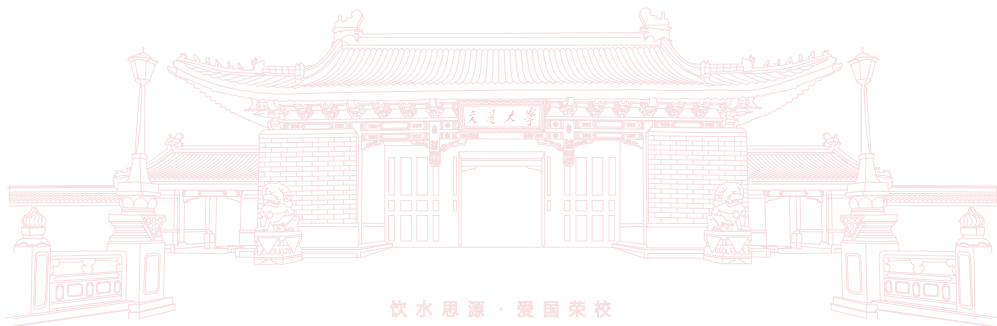
取 $\rho = 7/13$ 使二者相等，稳健 minimax 值为 $31/130 \approx .2385$ 。



自编有限情景例；模型依据 §7.3.6

05

Chebyshev 与 Chernoff 概率界



从经典概率界到优化

只知道部分统计信息时，事件概率通常不能精确计算，但可以求对所有相容分布都成立的界。

Markov 界

若 $X \geq 0$ 、 $\mathbf{E}X = \mu$ ，则

$$\mathbf{P}(X \geq 1) \leq \mu.$$

用线性函数 $f(z) = z$ 从上方覆盖事件指示函数。

两者的共同结构是：在一类可计算期望的函数中，寻找事件指示函数的最佳上界。

Chebyshev 界

若 $\mathbf{E}X = \mu$ 、 $\mathbf{E}(X - \mu)^2 = \sigma^2$ ，则

$$\mathbf{P}(|X - \mu| \geq 1) \leq \sigma^2.$$

用二次函数从上方覆盖指示函数。

广义 Chebyshev 界

随机向量 X 支持于 S , 关注事件 $C \subseteq S$ 。已知

$$\mathbf{E}f_i(X) = a_i, \quad i = 0, \dots, n, \quad f_0 \equiv 1, a_0 = 1.$$

寻找线性组合 $f(z) = \sum_{i=0}^n x_i f_i(z)$, 使 $f \geq \mathbf{1}_C$:

$$\begin{aligned} & \underset{x}{\text{minimize}} && \sum_{i=0}^n a_i x_i \\ & \text{subject to} && \sum_{i=0}^n x_i f_i(z) \geq 1, \quad z \in C, \\ & && \sum_{i=0}^n x_i f_i(z) \geq 0, \quad z \in S \setminus C. \end{aligned}$$

则 $\mathbf{P}(X \in C) \leq \mathbf{E}f(X) = a^\top x$ 。

问题类型

目标线性、每个固定 z 给出一个线性不等式, 因此这是半无限 LP; 适当结构可把它化为 SDP。

Boyd & Vandenberghe, *Convex Optimization*, §7.4.1, equation (7.17)

Markov 界的优化推导

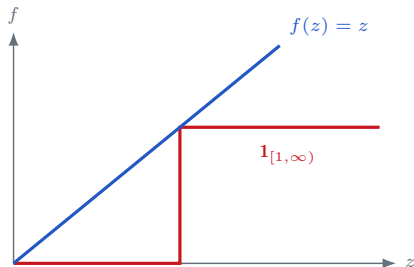
取 $S = \mathbb{R}_+$ 、 $C = [1, \infty)$ 、 $f_0(z) = 1$ 、 $f_1(z) = z$ ，且 $\mathbf{E}X = \mu \leq 1$ 。覆盖条件化为

$$x_0 \geq 0, \quad x_1 \geq 0, \quad x_0 + x_1 \geq 1.$$

于是

$$\min_{x_0, x_1} x_0 + \mu x_1 \quad \text{s.t.} \quad x_0, x_1 \geq 0, x_0 + x_1 \geq 1.$$

最优解为 $(x_0, x_1) = (0, 1)$ ，得到 $\mathbf{P}(X \geq 1) \leq \mu$ 。



最优上界函数在 $z = 1$ 处接触指示函数；期望将点态上界转为概率上界。

Boyd & Vandenberghe, *Convex Optimization*, §74.1, Markov bound example

已知一、二阶矩的 SDP

若 $X \in \mathbb{R}^m$, 已知 $\mathbf{E}X = \mathbf{a}$ 、 $\mathbf{E}XX^\top = \Sigma$, 选二次上界

$$f(z) = z^\top Qz + 2\mathbf{q}^\top z + r, \quad \mathbf{E}f(X) = \text{tr}(\Sigma Q) + 2\mathbf{q}^\top \mathbf{a} + r.$$

全局非负等价于

$$\begin{bmatrix} Q & \mathbf{q} \\ \mathbf{q}^\top & r \end{bmatrix} \succeq 0.$$

若坏事件为开多面体 $\mathcal{P} = \{z \mid \mathbf{a}_i^\top z < b_i\}$ 的补集, 则由半空间形式的 S-procedure, 覆盖条件等价于存在 $\tau_i \geq 0$ 使

$$\begin{bmatrix} Q & \mathbf{q} \\ \mathbf{q}^\top & r - 1 \end{bmatrix} \succeq \tau_i \begin{bmatrix} 0 & \mathbf{a}_i/2 \\ \mathbf{a}_i^\top/2 & -b_i \end{bmatrix}, \quad \tau_i \geq 0.$$

最小化 $\mathbf{E}f(X)$ 即得到 SDP (7.19)。

Chebyshev 界的对偶含义

原始无穷维问题是在所有满足矩约束的概率测度中最大化坏事件概率：

$$\begin{array}{ll}\text{maximize} & \int_C \pi(dz) \\ \text{subject to} & \int_S f_i(z) \pi(dz) = a_i, \quad i = 0, \dots, n, \\ & \pi \geq 0.\end{array}$$

广义 Chebyshev 问题寻找覆盖 1_C 的函数，是它的形式对偶。

原问题

寻找使坏事件概率最大的分布。

在教材的一、二阶矩多面体情形中，强对偶还可构造达到界的离散分布，说明界是锐的。

对偶问题

寻找可由已知矩计算期望的最紧上界函数。

Boyd & Vandenberghe, *Convex Optimization*, §74.1, Remarks 7.1–7.2

Chernoff 界

对任意 $\lambda \geq 0$, 事件 $X \geq u$ 上有 $e^{\lambda(X-u)} \geq 1$, 故

$$\mathbf{P}(X \geq u) \leq \inf_{\lambda \geq 0} \mathbf{E} e^{\lambda(X-u)}.$$

取对数得到一维凸问题

$$\log \mathbf{P}(X \geq u) \leq \inf_{\lambda \geq 0} \{-\lambda u + K(\lambda)\}, \quad K(\lambda) = \log \mathbf{E} e^{\lambda X}.$$

K 是累积量生成函数 (cumulant generating function), 总为凸函数。

Gaussian 例

若 $X \sim \mathcal{N}(0, 1)$, 则 $K(\lambda) = \lambda^2/2$; $u \geq 0$ 时最优 $\lambda = u$, 得到 $\mathbf{P}(X \geq u) \leq e^{-u^2/2}$ 。

多维 Chernoff 界

对 $C \subseteq \mathbb{R}^m$, 取指数上界函数 $f(z) = e^{\lambda^\top z + \mu}$ 。若 $-\lambda^\top z \leq \mu$ 对所有 $z \in C$ 成立, 则 $f \geq \mathbf{1}_C$, 从而

$$\log \mathbf{P}(X \in C) \leq \inf_{\lambda} [S_C(-\lambda) + K_X(\lambda)],$$

其中

$$S_C(y) = \sup_{z \in C} y^\top z, \quad K_X(\lambda) = \log \mathbf{E} e^{\lambda^\top X}.$$

集合信息

支持函数 S_C 描述事件集合在方向 $-\lambda$ 上的最远位置。

两项均凸, 因此求最优方向是凸优化问题。

分布信息

累积量生成函数 K_X 描述随机向量沿方向 λ 的指数矩。

Gaussian 落入多面体的界

若 $X \sim \mathcal{N}(0, I)$ 、 $C = \{x \mid Ax \preceq b\} \neq \emptyset$ ，则 $K_X(\lambda) = \|\lambda\|_2^2/2$ 。利用 S_C 的 LP 对偶，Chernoff 界由 QP

$$\begin{array}{ll} \underset{u, \lambda}{\text{minimize}} & b^\top u + \frac{1}{2} \|\lambda\|_2^2 \\ \text{subject to} & u \succeq 0, \quad A^\top u + \lambda = 0 \end{array}$$

给出。该 QP 又与投影问题对偶，最终得到

$$\mathbf{P}(X \in C) \leq \exp\left(-\frac{1}{2} \text{dist}(0, C)^2\right).$$

几何意义

集合离 Gaussian 中心越远，概率按距离平方的指数速度衰减；投影点决定最有效的分离方向。

Boyd & Vandenberghe, *Convex Optimization*, §7.4.2, equations (7.21)–(7.22)

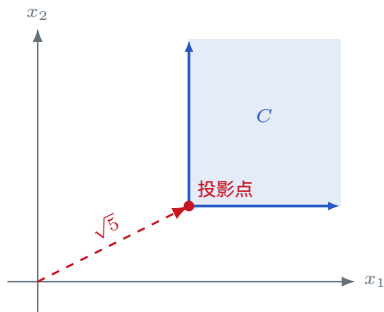
应用例 10: 二维告警区域

设 $X \sim \mathcal{N}(0, I_2)$, 告警区域为

$$C = \{x \in \mathbb{R}^2 \mid x_1 \geq 2, x_2 \geq 1\}.$$

原点到 C 的投影是 $(2, 1)$, 距离为 $\sqrt{5}$, 故

$$\mathbf{P}(X \in C) \leq e^{-5/2} \approx 0.0821.$$



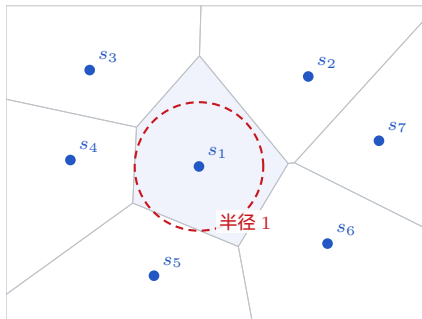
实际概率为 $\mathbf{P}(X_1 \geq 2)\mathbf{P}(X_2 \geq 1)$, 可由独立性精确计算; Chernoff 界只利用 Gaussian 累积量和集合几何, 适用于更复杂多面体。

§7.4.3: 信号星座与 Voronoi 区域

发送信号 $s_k \in \mathbb{R}^n$ ，接收 $x = s_k + v$ 。最小距离检测器选择距离 x 最近的信号。正确检测 s_k 等价于

$$2(s_j - s_k)^\top (s_k + v) < \|s_j\|_2^2 - \|s_k\|_2^2, \quad j \neq k.$$

这些线性不等式定义 s_k 的 Voronoi 区域 V_k 。



Chebyshev

只假设 $\mathbb{E}v = 0$ 、 $\mathbb{E}vv^\top = \sigma^2 I$ ，用 SDP 给出对所有相容噪声分布有效的正确率下界。

Chernoff

进一步假设 Gaussian 噪声，利用到错误 Voronoi 区域的距离给出指数型误差上界。

Boyd & Vandenberghe, *Convex Optimization*, §7.4.3, Figures 7.5–7.8 (定性重绘)

两类概率界的区别

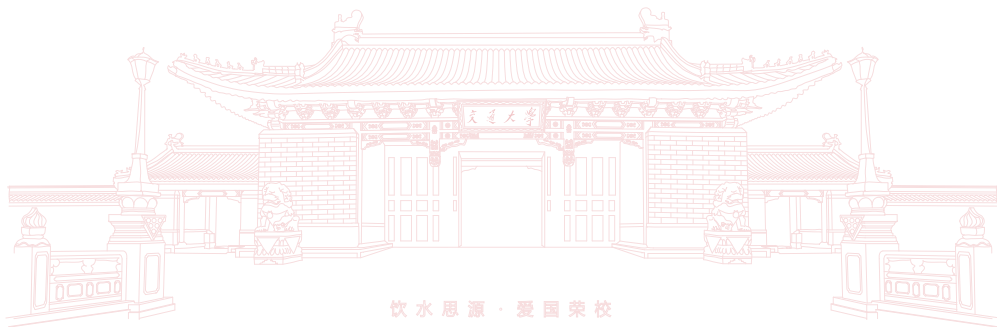
	Chebyshev 型	Chernoff 型
已知信息	有限个矩或期望	完整的累积量生成函数
上界函数	f_i 的线性组合, 常为多项式	指数仿射函数 $e^{\lambda^\top z + \mu}$
典型优化	半无限 LP、SDP	光滑凸问题、QP
鲁棒性	对所有满足矩条件的分布有效	依赖特定分布族或指数矩
紧度	信息少时较保守, 但可为锐界	分布假设正确时通常更紧

选择原则

界的可靠性由假设决定: 矩信息可信而分布未知时用 Chebyshev; 分布或累积量模型可信时用 Chernoff。

06

实验设计



饮水思源 · 爱国荣校

选择哪些测量

线性测量模型

$$y_i = \mathbf{a}_i^\top \mathbf{x} + w_i, \quad w_i \sim \mathcal{N}(0, 1), \quad i = 1, \dots, m,$$

中, $\mathbf{a}_i$ 从候选向量 $v_1, \dots, v_p \in \mathbb{R}^n$ 中选择。最小二乘估计误差协方差为

$$E = \left(\sum_{i=1}^m \mathbf{a}_i \mathbf{a}_i^\top \right)^{-1}.$$

设计变量

不是待估参数 x , 而是每类实验做多少次; 设计发生在采集数据之前。

评价对象

E 越小, 任意线性函数 $q^\top x$ 的估计方差 $q^\top E q$ 越小, 置信椭圆也越小。

信息矩阵与实验次数

设第 j 类实验做 m_j 次, $m_j \in \mathbb{Z}_+$ 、 $\sum_j m_j = m$, 则

$$E = \left(\sum_{j=1}^p m_j v_j v_j^\top \right)^{-1}.$$

定义信息矩阵

$$M = \sum_{j=1}^p m_j v_j v_j^\top,$$

则 $E = M^{-1}$ 。每次实验以秩一矩阵 $v_j v_j^\top$ 增加信息。

若 $M \succeq \widetilde{M} \succ 0$, 则 $M^{-1} \preceq \widetilde{M}^{-1}$: 更多方向性信息缩小误差协方差。

若候选向量未张成 $\mathbb{R}^n$, M 必然奇异; 存在某个参数方向无法由任何实验辨识。

Boyd & Vandenberghe, *Convex Optimization*, §7.5, equation (7.23)

连续松弛

令 $\lambda_j = m_j/m$ 为第 j 类实验的比例，则

$$E(\lambda) = \frac{1}{m} \left(\sum_{j=1}^p \lambda_j v_j v_j^\top \right)^{-1}, \quad \lambda \succeq 0, \quad \mathbf{1}^\top \lambda = 1.$$

忽略“ $m\lambda_j$ 必须为整数”得到凸的连续实验设计问题。

下界

松弛扩大可行集，因此其最优值给整数设计的理想下界。

λ 也可解释为随机选择实验类型的概率分布。

回到整数设计

取 $m_j \approx m\lambda_j$ 并调整总和；当 m 相对 p 较大时，舍入误差通常较小。

D、E、A 三种最优性准则

令 $M(\lambda) = \sum_j \lambda_j v_j v_j^\top \succ 0$, 忽略常数 $1/m$ 。

准则	优化目标	几何含义	统计含义	三
D-opt	$\min -\log \det M$	最小化置信椭球体积	最大化总体信息体积	
E-opt	$\max \lambda_{\min}(M)$	最小化最长半轴	最小化最坏方向方差	
A-opt	$\min \operatorname{tr}(M^{-1})$	控制各轴平方和	最小化均方参数误差	

者都在概率单纯形 $\lambda \succeq 0, \mathbf{1}^\top \lambda = 1$ 上构成凸优化, 但会选择不同的候选实验和权重。

E-opt 与 A-opt 的 SDP 形式

E-opt 可直接写为

$$\begin{aligned} & \underset{\lambda, t}{\text{maximize}} && t \\ & \text{subject to} && \sum_j \lambda_j v_j v_j^\top \succeq tI, \\ & && \lambda \succeq 0, \quad \mathbf{1}^\top \lambda = 1. \end{aligned}$$

A-opt 用 Schur 补表示 $u_k \geq e_k^\top M^{-1} e_k$:

$$\min_{\lambda, u} \mathbf{1}^\top u \quad \text{s.t.} \quad \begin{bmatrix} M(\lambda) & e_k \\ e_k^\top & u_k \end{bmatrix} \succeq 0, k = 1, \dots, n, \quad \lambda \succeq 0, \mathbf{1}^\top \lambda = 1.$$

维度

$M(\lambda) \in \mathbb{S}^n$; 每个 Schur 补 LMI 的尺寸为 $(n+1) \times (n+1)$ 。

对偶椭球与支持实验

D-opt 的对偶可写为

$$\max_{W \succ 0} \log \det W + n \log n \quad \text{s.t.} \quad v_i^\top W v_i \leq 1, i = 1, \dots, p.$$

W^* 定义包含全部候选点的最小体积中心椭球

$$\mathcal{E} = \{x \mid x^\top W^* x \leq 1\}.$$

互补松弛给出

$$\lambda_i^* (1 - v_i^\top W^* v_i) = 0.$$

正权重实验必须位于对偶椭球边界；椭球内部候选点可有零权重。

E-opt 与 A-opt 也有相似椭球解释，只是对 W 采用不同的尺寸准则。

E-opt 与 A-opt 的对偶

两种设计也可由包含候选测量向量的中心椭圆刻画：

E-opt 对偶

$$\begin{aligned} & \underset{W \succeq 0}{\text{maximize}} && \text{tr } W \\ & \text{subject to} && v_i^\top W v_i \leq 1, \quad i = 1, \dots, p. \end{aligned}$$

它选择控制最坏估计方向的椭圆尺度。

A-opt 对偶

$$\begin{aligned} & \underset{W \succeq 0}{\text{maximize}} && (\text{tr } W^{1/2})^2 \\ & \text{subject to} && v_i^\top W v_i \leq 1, \quad i = 1, \dots, p. \end{aligned}$$

它选择控制各参数方向平均误差的椭圆尺度。

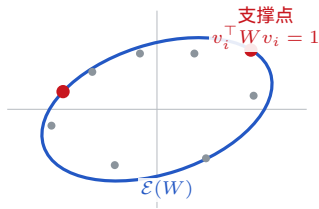
共同的几何约束

令

$$\mathcal{E}(W) = \{z \mid z^\top W z \leq 1\}.$$

约束 $v_i^\top W v_i \leq 1$ 要求所有候选向量均位于 $\mathcal{E}(W)$ 内；两种准则只改变对椭圆轴长的评价方式。

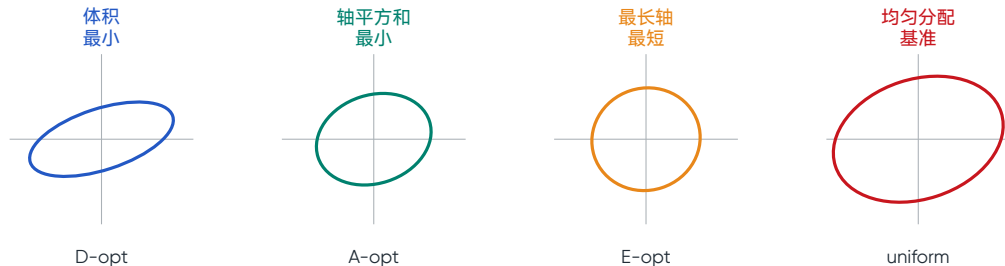
两者仍满足 $\lambda_i^*(1 - v_i^\top W^* v_i) = 0$ ，因此正权重只落在支撑点上。



Boyd & Vandenberghe, *Convex Optimization*, §7.5.2, equations (7.30)–(7.31)

不同准则的置信椭球

同一候选集合下，D/A/E-opt 与均匀设计产生不同的误差协方差；下图为教材 Figure 7.12 的结构性示意。



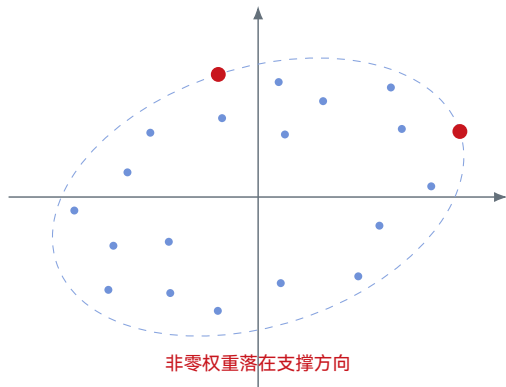
比较方式

三种最优准则回答不同问题，椭球之间一般不存在 PSD 意义下的全序；选择准则应对应实际误差代价。

Boyd & Vandenberghe, *Convex Optimization*, §7.5.2, Figure 7.12 (结构性重绘)

教材例：候选测量与稀疏设计

教材在 $\mathbb{R}^2$ 中给出 20 个候选测量向量。三种最优设计只对少数向量赋予非零权重：D-opt 使用 2 个，E-opt 使用 2 个，A-opt 使用 3 个。



稀疏性的来源

目标是矩阵准则，但对偶椭球只需由少数边界点支撑；因此连续设计常自动集中于少量实验。

Boyd & Vandenberghe, *Convex Optimization*, §7.5.2, Figures 7.9–7.12 (定性重绘)

应用例 11: 二维传感器方向选择

候选测量方向为

$$v_1 = (1, 0), \quad v_2 = (0, 1), \quad v_3 = \frac{1}{\sqrt{2}}(1, 1).$$

若只使用 v_1, v_2 , 取权重 $(\lambda_1, \lambda_2) = (t, 1 - t)$, 则

$$M = \text{diag}(t, 1 - t), \quad -\log \det M = -\log[t(1 - t)].$$

D-opt 在 $t = 1/2$ 取得最优, 给出各向同性信息矩阵 $M = I/2$ 。

只测一个方向

$t = 0$ 或 1 时 M 奇异, 另一个参数方向完全不可辨识。

加入斜向测量

v_3 可在方向受限或成本不同时改善设计; 最优权重由完整凸问题决定。

资源约束与多输出实验

资源限制

若第 j 类实验成本为 c_j ，总预算为 B ，则加入

$$mc^\top \lambda \leq B.$$

多个资源对应多条线性不等式，可同时表示时间、能耗与样本数限制。

联合测量还可单独作为一种候选实验，并赋予体现批量折扣的成本。

每次得到多个测量

若实验 j 产生矩阵 $V_j = [u_{j1} \cdots u_{jk_j}]$ ，其信息贡献为

$$V_j V_j^\top = \sum_{\ell=1}^{k_j} u_{j\ell} u_{j\ell}^\top.$$

原有设计公式保持不变。

应用例 12: 预算约束下的实验组合

三类实验的信息贡献为

$$v_1 v_1^\top = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad v_2 v_2^\top = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad v_3 v_3^\top = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix},$$

其中 $v_3 = (1, 1)$ 同时测量两个参数。成本 $c = (1, 1, 1.6)$, 平均成本上限为 1.2。由对称性可取

$$\lambda_3 = z, \quad \lambda_1 = \lambda_2 = \frac{1-z}{2}, \quad c^\top \lambda = 1 + 0.6z \leq 1.2,$$

而信息矩阵的行列式为

$$\det M(z) = \frac{1 + 2z - 3z^2}{4}.$$

它在可行区间 $0 \leq z \leq 1/3$ 上递增, 故 $\lambda^* = (1/3, 1/3, 1/3)$, 且预算约束取等号。

模型解释

联合实验虽更昂贵, 但它提供的额外信息使 D-opt 解主动用尽预算; 权重由边际信息增益与成本共同决定。

连续设计如何落地

连续解 λ^* 需要转换为总次数 m 的整数方案。

1. 先取 $\tilde{m}_j = \lfloor m\lambda_j^* \rfloor$;
2. 将剩余次数分给小数部分最大的候选;
3. 检查信息矩阵是否满秩, 并计算实际 A/D/E 指标;
4. 若预算或成组约束被破坏, 在邻域内做小规模整数搜索。

例

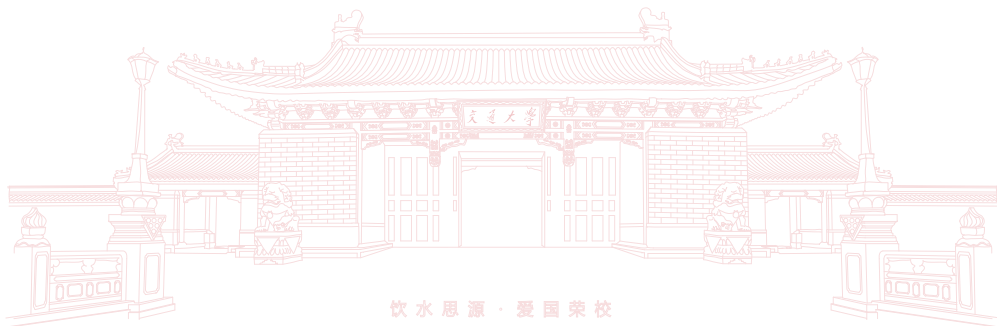
$m = 10$ 、 $\lambda^* = (.46, .34, .20)$ 给出 $(5, 3, 2)$, 总数恰为 10。

验证

舍入后的方案必须重新评价; 连续最优值只是整数设计的参照, 不是最终性能保证。

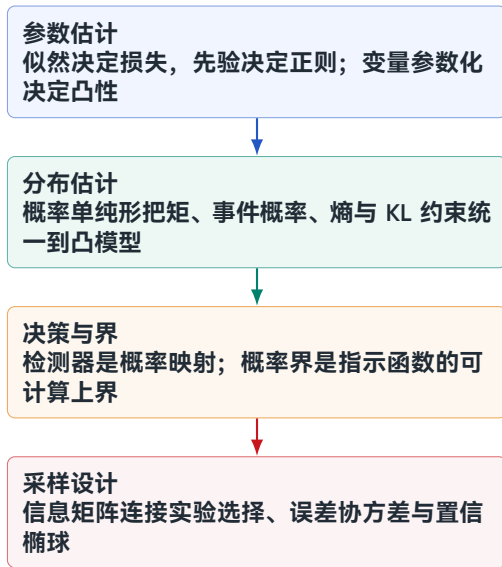
07

总结



饮水思源 · 爱国荣校

第 7 章的结构



下一讲：几何问题

协方差与信息矩阵将转化为 Gram geometry;
置信椭圆与实验设计将转化为 extremal ellipsoid problems。