高维统计笔记

作者：Crab

时间：2025 年 1 月 3 日

说明：本笔记根据王成、刘卫东老师翻译的 Martin J. Wainwright《High-Dimensional Statistics: A Non-Asymptotic Viewpoint》整理而成

ElegantLaTeX Program

I don’t have any secret. I’m just passionate about what I do. ——Samuelson

目　　录

第 1 章 简介

1.1 经典理论和高维理论

1.2 高维会产生什么问题

内容提要

- ❑ 线性判别分析
- ❑ 非参数回归
- ❑ 协方差估计

1.3 高维中什么能帮助我们

我们期望数据具有一定的低维结构，这使得我们能够简化高维问题。

内容提要

- ❑ 向量的稀疏性
- ❑ 回归形式的结构
- ❑ 协方差矩阵中的结构

1.4 什么是非渐进的观点

1.5 全书概述

- 工具和方法：第 2-5 章、第 12 章、第 14-15 章
- 模型和估计：第 6-11 章、第 13 章

第 2 章 基本尾部概率界和集中不等式

2.1 经典的界

2.1.1 从马尔可夫不等式到 Chernoff 界

- **马尔可夫不等式**：对任意一个均值有限的非负随机变量 X，有

$$\mathbb{P}[X \geqslant t] \leqslant \frac{\mathbb{E}[X]}{t}, \quad \forall t > 0. \tag{2.1}$$

- **Chebyshev 不等式**：如果这个随机变量 X 还有有限的方差，则有

$$\mathbb{P}[|X-\mu| \geqslant t] \leqslant \frac{\operatorname{Var}(X)}{t^2}, \quad \forall t > 0. \tag{2.2}$$

- **Chernoff 界**：若随机变量 X 在 0 的领域 $[0, b]$ 内有矩母函数，则有

$$\mathbb{P}\left[(X-\mu) \geqslant t\right] \leqslant \inf_{\lambda \in [0,b]} \frac{\mathbb{E}\left[e^{\lambda(X-\mu)}\right]}{e^{\lambda t}} \tag{2.3}$$

注 一般情况下，马尔可夫不等式和 Chebyshev 不等式所给的界是最优的。

2.1.2 次高斯随机变量和 Hoeffding 界

例 2.1 (次高斯尾部界)

对于随机变量 $X \sim \mathcal{N}(\mu, \sigma^2)$，有上偏差不等式：

$$\mathbb{P}[X \geqslant \mu + t] \leqslant \exp\left(-\frac{t^2}{2\sigma^2}\right), \quad \forall t \geqslant 0. \tag{2.4}$$

注 事实上，根据对 Mills 比的探究，这个界是除了多项式修正项之外最优的。

定义 2.1 (次高斯随机变量)

设随机变量 X 的均值 $\mu = \mathbb{E}[X]$，若存在 $\sigma > 0$，使得

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leqslant \exp\left\{\frac{\sigma^2\lambda^2}{2}\right\}, \quad \forall \lambda \in \mathbb{R}, \tag{2.5}$$

则称这个随机变量 X 是次高斯 (sub-Gaussian) 的，常数 σ 称为次高斯参数。

将上偏差不等式和下偏差不等式结合，可以得到对任意次高斯随机变量 X 的**集中不等式**：

$$\mathbb{P}[|X-\mu| \geqslant t] \leqslant 2\exp\left(-\frac{t^2}{2\sigma^2}\right), \quad \forall t \in \mathbb{R}. \tag{2.6}$$

例 2.2 (Rademacher 随机变量)

一个 **Rademacher** 随机变量 ε 是指等概率地取 $\{-1, 1\}$ 的随机变量，它是一个参数为

$\sigma=1$ 的次高斯随机变量。利用指数函数的级数展开，有

$$\begin{aligned}\mathbb{E}[e^{\lambda\varepsilon}]&=\frac{1}{2}\{e^{-\lambda}+e^{\lambda}\}=\frac{1}{2}\left\{\sum_{k=0}^{\infty}\frac{(-\lambda)^k}{k!}+\sum_{k=0}^{\infty}\frac{(\lambda)^k}{k!}\right\}\\&=\sum_{k=0}^{\infty}\frac{\lambda^{2k}}{(2k)!}\leqslant 1+\sum_{k=1}^{\infty}\frac{\lambda^{2k}}{2^k k!}=e^{\lambda^2/2}.\end{aligned}$$

例 2.3 (有界随机变量的次高斯性)

设 X 是均值为 0，支撑集为某个区间 $[a,b]$ 的随机变量。利用对称化技巧，可以证明 X 是一个次高斯随机变量，其中参数 σ 至多为 $b-a$。(实际上可以提升为 $\dfrac{b-a}{2}$)

$$\begin{aligned}\mathbb{E}_X[e^{\lambda X}]&=\mathbb{E}_X[e^{\lambda(X-\mathbb{E}_{X'}[X'])}]\leqslant\mathbb{E}_{X,X'}[e^{\lambda(X-X')}]\\&=\mathbb{E}_{X,X'}[\mathbb{E}_\varepsilon[e^{\lambda\varepsilon(X-X')}]]\overset{\text{(i)}}{\leqslant}\mathbb{E}_{X,X'}[e^{\frac{\lambda^2(X-X')^2}{2}}]\leqslant e^{\frac{\lambda^2(b-a)^2}{2}}.\end{aligned}$$

其中步骤 (i) 是先固定 (X,X') 取条件期望得到。

注 例 2.3 中所用的是对称化技巧的一个简单例子：先引入一个和 X 独立同分布的 X'，然后用一个 Rademacher 随机变量把问题对称化。

命题 2.1 (Hoeffding 界)

设次高斯随机变量 X_i 相互独立，对应的均值和次高斯参数分别为 μ_i,σ_i，则对 $\forall t\geqslant 0$，有

$$\mathbb{P}\left[\sum_{i=1}^{n}(X_i-\mu_i)\geqslant t\right]\leqslant\exp\left(-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2}\right).\tag{2.7}$$

其中用到如下性质：

性质 设 X_1,X_2 是独立的次高斯随机变量，对应的参数分别为 σ_1,σ_2，则 X_1+X_2 是参数为 $\sqrt{\sigma_1^2+\sigma_2^2}$ 的次高斯随机变量。

对次高斯随机变量的以下三种不同形式的刻画是等价的：

1. 通过计算矩母函数或得到矩母函数的界直接验证次高斯性；
2. 任意的次高斯随机变量在某种意义下会被一个正态随机变量控制；
3. 次高斯性可以通过控制随机变量的矩来得到。

定理 2.2 (次高斯随机变量定义的等价性)

对任意均值为 0 的随机变量 X，下面的性质是等价的：

(I) 存在常数 $\sigma\geqslant 0$ 使得

$$\mathbb{E}[e^{\lambda X}]\leqslant\exp\left(\frac{\lambda^2\sigma^2}{2}\right),\quad\forall\lambda\in\mathbb{R}.$$

(II) 存在常数 $c\geqslant 0$ 和正态随机变量 $Z\sim\mathcal{N}(0,\tau^2)$ 使得

$$\mathbb{P}[|X|\geqslant s]\leqslant c\mathbb{P}[|Z|\geqslant s],\quad\forall s\geqslant 0.$$

(III) 存在常数 $\theta \geqslant 0$ 使得

$$\mathbb{E}[X^{2k}] \leqslant \frac{(2k)!}{2^k k!}\theta^{2k}, \quad \forall k = 1, 2, \cdots .$$

(IV) 存在常数 $\sigma \geqslant 0$ 使得

$$\mathbb{E}[e^{\frac{\lambda X^2}{2\sigma^2}}] \leqslant \frac{1}{\sqrt{1-\lambda}}, \quad \forall \lambda \in [0, 1).$$

2.1.3 次指数随机变量和 Bernstein 界

次高斯的定义相对比较严格，考虑一些更为宽泛的情形。

定义 2.2 (次指数随机变量)

对一个均值为 $\mu = \mathbb{E}[X]$ 的随机变量 X，如果存在非负参数对 (ν, α) 满足

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leqslant \exp\left(\frac{\nu^2\lambda^2}{2}\right), \quad \forall |\lambda| < \frac{1}{\alpha}. \tag{2.8}$$

则称该随机变量是次指数 (sub-exponential) 的。

次高斯随机变量都是次指数的，但次指数随机变量不一定是次高斯的。

类似次高斯性，次指数随机变量也有相应的偏差和集中不等式。当 t 充分小的时候，这些界本质上是次高斯的（指数部分为 t^2 阶）；而当 t 较大的时候，这个界的指数部分会与 t 呈线性关系。

命题 2.3 (次指数尾部不等式)

设 X 是参数为 (ν, α) 的次指数随机变量，则

$$\mathbb{P}[X - \mu \geqslant t] \leqslant \begin{cases} e^{-\frac{t^2}{2\nu^2}}, & 0 \leqslant t \leqslant \frac{\nu^2}{\alpha}, \\ e^{-\frac{t}{2\alpha}}, & t > \frac{\nu^2}{\alpha}. \end{cases} \tag{2.9}$$

次指数性可以通过计算矩母函数或得到矩母函数的界来验证，但在许多情形下，这种直接计算的方法是不可行的，其中一种替代方法是控制 X 的多项式形式的矩。

给定均值为 $\mu = \mathbb{E}[X]$，方差为 $\sigma^2 = \mathbb{E}[X^2] - \mu^2$ 的随机变量 X，称参数为 b 的 **Bernstein 条件**成立，如果

$$|\mathbb{E}[(X-\mu)^k]| \leqslant \frac{1}{2}k!\sigma^2 b^{k-2} \quad k = 2, 3, 4, \cdots . \tag{2.10}$$

Bernstein 条件的一个充分条件是 X 有界。(?)

满足 Bernstein 条件的随机变量是次指数的。利用指数函数的幂级数展开可以得到，

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leqslant \exp\left(\frac{\lambda^2\sigma^2/2}{1-b|\lambda|}\right) \leqslant \exp\left(\frac{\lambda^2(\sqrt{2}\sigma)^2}{2}\right), \quad \forall |\lambda| < \frac{1}{2b}. \tag{2.11}$$

即 X 是参数为 $(\sqrt{2}\sigma, 2b)$ 的次指数随机变量。

由 Bernstein 条件，可以得到常数项更紧的尾部概率界。

命题 2.4 (Bernstein 型界)
对任意满足 Bernstein 条件 (2.10) 的随机变量 X，有

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leqslant \exp\left(\frac{\lambda^2\sigma^2/2}{1-b|\lambda|}\right), \quad \forall|\lambda| < \frac{1}{b}. \tag{2.12}$$

以及集中不等式

$$\mathbb{P}[|X-\mu| \geqslant t] \leqslant 2\exp\left(-\frac{t^2}{2(\sigma^2+bt)}\right), \quad \forall t \geqslant 0. \tag{2.13}$$

证明 运用矩母函数的界，在 Chernoff 界中取 $\lambda = \dfrac{t}{bt+\sigma^2} \in \left[0, \dfrac{1}{b}\right)$ 即证。 □

性质 由定义易得，和次高斯性一样，独立次指数随机变量的和同样保持次指数性。

设 $\{X_k\}_{k=1}^n$ 是独立的次指数随机变量，对应的均值为 μ_k，参数为 (ν_k, α_k)，则 $\sum\limits_{k=1}^n (X_k-\mu_k)$ 是参数为 $(\sqrt{\sum\limits_{k=1}^n \nu_k^2}, \max\limits_{1\leqslant k\leqslant n} \alpha_k)$ 的次指数随机变量。

对次指数随机变量，同样有很多等价的描述方式。

定理 2.5 (次指数随机变量的等价描述)
对一个均值为 0 的随机变量 X，下面的几个定义等价：

(I) 存在非负参数 (ν, α) 使得

$$\mathbb{E}[e^{\lambda X}] \leqslant \exp\left(\frac{\nu^2\lambda^2}{2}\right), \quad \forall|\lambda| < \frac{1}{\alpha}.$$

(II) 存在一个正常数 $c_0 > 0$，使得对任意 $|\lambda| \leqslant c_0$，都有 $\mathbb{E}[e^{\lambda X}] < \infty$。

(III) 存在常数 $c_1, c_2 > 0$，使得

$$\mathbb{P}[|X| \geqslant t] \leqslant c_1 e^{-c_2 t}, \quad \forall t > 0.$$

(IV) 量 $\gamma := \sup\limits_{k\geqslant 2}\left[\dfrac{\mathbb{E}[X^k]}{k!}\right]^{\frac{1}{k}} < \infty$。

2.1.4 一些单边结果

命题 2.6 (单侧 Bernstein 不等式)
如果 $X \leqslant b$, a.s. ，那么

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \leqslant \exp\left(\frac{\frac{\lambda^2}{2}\mathbb{E}[X^2]}{1-\frac{b\lambda}{3}}\right), \quad \forall\lambda \in [0, 3/b). \tag{2.14}$$

相对应的，给定 n 个满足条件 $X_i \leqslant b$, a.s. 的独立随机变量，有

$$\mathbb{P}\left[\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \geqslant n\delta\right] \leqslant \exp\left(-\frac{n\delta^2}{2\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i^2] + \frac{b\delta}{3}\right)}\right). \tag{2.15}$$

如果一个随机变量 X 有下界，考虑 $-X$ 可以得到它的下尾部不等式。特别地，对独立非

负随机变量 $Y_i \geqslant 0$，有

$$\mathbb{P}\left[\sum_{i=1}^{n}(Y_i - \mathbb{E}[Y_i]) \leqslant -n\delta\right] \leqslant \exp\left(-\frac{n\delta^2}{\frac{2}{n}\sum_{i=1}^{n}\mathbb{E}[Y_i^2]}\right). \tag{2.16}$$

证明 对 $e^{\lambda X}$ 泰勒展开并取期望

$$\mathbb{E}[e^{\lambda X}] = 1 + \lambda\mathbb{E}[X] + \frac{1}{2}\lambda^2\mathbb{E}[X^2 h(\lambda X)],$$

其中

$$h(u) := 2\frac{e^u - u - 1}{u^2} = 2\sum_{k=2}^{\infty}\frac{u^{k-2}}{k!}.$$

由 $h(u)$ 的单调性可得

$$h(\lambda x) \leqslant h(\lambda b) \leqslant \sum_{k=2}^{\infty}\left(\frac{\lambda b}{3}\right)^{k-2} = \frac{1}{1-\frac{\lambda b}{3}}, \quad \lambda \in [0, 3/b).$$

故

$$\begin{aligned}
\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] &\leqslant e^{-\lambda\mathbb{E}[X]}\left(1 + \lambda\mathbb{E}[X] + \frac{1}{2}\lambda^2\mathbb{E}[X^2]h(\lambda b)\right) \\
&\leqslant \exp\left\{\frac{\lambda^2\mathbb{E}[X^2]}{2}h(\lambda b)\right\} \\
&\leqslant \exp\left\{\frac{\frac{\lambda^2}{2}\mathbb{E}[X^2]}{1-\frac{b\lambda}{3}}\right\}, \quad \forall\lambda \in [0, 3/b).
\end{aligned}$$

由 Chernoff 界得

$$\mathbb{P}\left[\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \geqslant n\delta\right] \leqslant \exp\left\{-\lambda n\delta + \frac{\frac{\lambda^2}{2}\sum_{i=1}^{n}\mathbb{E}[X_i^2]}{1-\frac{b\lambda}{3}}\right\}, \quad \forall\lambda \in [0, 3/b).$$

取

$$\lambda = \frac{n\delta}{\sum_{i=1}^{n}\mathbb{E}[X_i^2] + \frac{n\delta b}{3}} \in [0, 3/b)$$

即证。 □

2.2 基于鞅的方法

推导随机变量的更一般函数的不等式界，解决这类问题的一种经典技巧是鞅的分解。

2.2.1 背景

令 $\{X_k\}_{k=1}^{n}$ 为一列独立随机变量，对于函数 $f: \mathbb{R}^d \to \mathbb{R}^d$，$f(X) = f(X_1, \cdots, X_n)$，假设我们的目标是研究函数 f 与其均值之间偏差的概率界。考虑随机变量序列 $Y_0 = \mathbb{E}[f(X)]$，$Y_n = f(X)$，以及

$$Y_k = \mathbb{E}[f(X) \mid X_1, \cdots, X_k], \quad k = 1, \cdots, n-1. \tag{2.17}$$

假定上述所有条件期望均存在，基于下面的伸缩分解：

$$f(X)-\mathbb{E}[f(X)]=Y_n-Y_0=\sum_{k=1}^{n}(Y_k-Y_{k-1})=:\sum_{k=1}^{n}D_k.$$

序列 $\{Y_k\}_{k=1}^n$ 是一个鞅序列，称为 **Doob 鞅**，而 $\{D_k\}_{k=1}^n$ 则是一个鞅差序列。

定义 2.3 (鞅)

令 $\{\mathcal{F}_k\}_{k=1}^\infty$ 为一列非降的 σ 域，即对 $\forall k\geqslant 1$ 有 $\mathcal{F}_k\subset\mathcal{F}_{k+1}$，这样的序列称为域流。给定一个适应于域流 $\{\mathcal{F}_k\}_{k=1}^\infty$ 的随机变量序列 $\{Y_k\}_{k=1}^\infty$，如果对 $\forall k\geqslant 1$，有

$$\mathbb{E}[|Y_k|]<\infty \quad 和 \quad \mathbb{E}[Y_{k+1}\mid\mathcal{F}_k]=Y_k, \tag{2.18}$$

那么称 $\{(Y_k,\mathcal{F}_k)\}_{k=1}^\infty$ 是一个鞅。

例 2.4 (Doob 构造)

假设函数 f 绝对可积，即 $\mathbb{E}[|f(X)|]<\infty$，那么前面讨论的 Doob 构造确实是一个鞅。记 $X_1^k=(X_1,X_2,\cdots,X_k)$，由 Jenson 不等式及条件期望的性质，

$$\mathbb{E}[|Y_k|]=\mathbb{E}[|\mathbb{E}[f(X)\mid X_1^k]|]\leqslant\mathbb{E}[|f(X)|]<\infty,$$

$$\mathbb{E}[Y_{k+1}\mid X_1^k]=\mathbb{E}[\mathbb{E}[f(X)\mid X_1^{k+1}]\mid X_1^k]=\mathbb{E}[f(X)\mid X_1^k]=Y_k.$$

一个密切相关的概念是**鞅差序列**，表示一个适应序列 $\{(D_k,\mathcal{F}_k)\}_{k=1}^\infty$ 对 $\forall k\geqslant 1$，满足

$$\mathbb{E}[|D_k|]<\infty \quad 和 \quad \mathbb{E}[D_{k+1}\mid\mathcal{F}_k]=0. \tag{2.19}$$

2.2.2 鞅差序列的集中度界

首先通过在鞅差序列上加上次指数条件，得到一个一般的鞅差序列的 Bernstein 型界。

定理 2.7 (鞅差序列的 Bernstein 界)

令 $\{(D_k,\mathcal{F}_k)\}_{k=1}^\infty$ 为一个鞅差序列，并假设对任意的 $|\lambda|<1/\alpha_k$，有 $\mathbb{E}[e^{\lambda D_k}\mid\mathcal{F}_{k-1}]\leqslant e^{\lambda^2\nu_k^2/2}$, a.s.，那么下面的结论成立：

(a) $\sum_{k=1}^n D_k$ 是参数为 $\left(\sqrt{\sum_{k=1}^n\nu_k^2},\alpha_*\right)$ 的次指数随机变量，其中 $\alpha_*=\max_{1leqk\leqslant n}\alpha_k$。

(b) 有集中不等式

$$\mathbb{P}\left[\left|\sum_{k=1}^n D_k\right|\geqslant t\right]\leqslant\begin{cases}2\exp\left(-\dfrac{t^2}{2\sum_{k=1}^n\nu_k^2}\right), & 0\leqslant t\leqslant\dfrac{\sum_{k=1}^n\nu_k^2}{\alpha_*},\\[2ex] 2\exp\left(-\dfrac{t}{2\alpha_*}\right), & t>\dfrac{\sum_{k=1}^n\nu_k^2}{\alpha_*}.\end{cases} \tag{2.20}$$

证明 在 $\mathcal{F}_{k-1}$ 下求条件期望

$$\mathbb{E}\left[e^{\lambda\sum_{k=1}^n D_k}\right]=\mathbb{E}\left[e^{\lambda\sum_{k=1}^{n-1}D_k}\mathbb{E}\left[e^{\lambda D_n}\mid\mathcal{F}_{n-1}\right]\right]\leqslant\mathbb{E}\left[e^{\lambda\sum_{k=1}^{n-1}D_k}\right]e^{\lambda^2\nu_n^2/2}, \tag{2.21}$$

可以验证次指数性，再由命题 2.3 即得集中不等式。 □

推论 2.8 (Azuma-Hoeffding)

令 $\{(D_k,\mathcal{F}_k)\}_{k=1}^{\infty}$ 为一个鞅差序列，满足对所有的 $k=1,\cdots,n$，存在常数 $\{(a_k,b_k)\}_{k=1}^{n}$，使得 $D_k\in[a_k,b_k]$, a.s.，那么，对 $\forall t\geqslant 0$，

$$\mathbb{P}\left[\left|\sum_{k=1}^{n}D_k\right|\geqslant t\right]\leqslant 2\exp\left(-\frac{2t^2}{\sum_{k=1}^{n}(b_k-a_k)^2}\right). \tag{2.22}$$

推论 2.8 的一个重要应用是研究满足有界差性质的函数。给定向量 $\boldsymbol{x},\boldsymbol{x}'\in\mathbb{R}^n$ 和一个指标 $k\in\{1,2,\cdots,n\}$，定义新的向量 $\boldsymbol{x}^{\backslash k}\in\mathbb{R}^n$

$$x_j^{\backslash k}:=\begin{cases}x_j & \text{如果} j\neq k,\\ x_k' & \text{如果} j=k.\end{cases} \tag{2.23}$$

我们称 $f:\mathbb{R}^n\to\mathbb{R}^n$ 满足参数为 $(L_1,\cdots,L_n)$ 的**有界差不等式**，如果对 $\forall k=1,2,\cdots,n$，有

$$|f(\boldsymbol{x})-f(\boldsymbol{x}^{\backslash k})|\leqslant L_k,\quad \forall\boldsymbol{x},\boldsymbol{x}'\in\mathbb{R}^n. \tag{2.24}$$

推论 2.9 (有界差不等式)

假设 f 满足参数为 $(L_1,\cdots,L_n)$ 的有界差性质 (2.24)，且随机向量 $\mathbf{X}=(X_1,\cdots,X_n)$ 各分量独立，则

$$\mathbb{P}[|f(X)-\mathbb{E}[f(X)]|\geqslant t]\leqslant 2\exp\left(-\frac{2t^2}{\sum_{k=1}^{n}L_k^2}\right),\quad \forall t\geqslant 0. \tag{2.25}$$

证明 考虑 Doob 分解对应的鞅差序列

$$D_k=\mathbb{E}[f(X)\mid X_1,\cdots,X_k]-\mathbb{E}[f(X)\mid X_1,\cdots,X_{k-1}].$$

记

$$A_k:=\inf_x\mathbb{E}[f(X)\mid X_1,\ldots,X_{k-1},x]-\mathbb{E}[f(X)\mid X_1,\ldots,X_{k-1}],$$
$$B_k:=\sup_x\mathbb{E}[f(X)\mid X_1,\ldots,X_{k-1},x]-\mathbb{E}[f(X)\mid X_1,\ldots,X_{k-1}].$$

证明 $A_k\leqslant D_k\leqslant B_k$, a.s. 且 $B_k-A_k\leqslant L_k$，再由推论 2.8 Azuma-Hoeffding 不等式即证。 □

例 2.5 (有界差经典 Hoeffding 界)

考虑函数 $f(x_1,\cdots,x_n)=\sum_{i=1}^{n}(x_i-\mu_i)$，其中 $\mu_i=\mathbb{E}[X_i]$。对 $\forall k\in\{1,\cdots,n\}$，有 $|f(\boldsymbol{x})-f(\boldsymbol{x}^{\backslash k})|=|(x_k-\mu_k)-(x_k'-\mu_k)|=|x_k-x_k'|\leqslant b-a$，从而可以得到独立随机变量的经典 Hoeffding 不等式界

$$\mathbb{P}\left[\left|\sum_{i=1}^{n}(X_i-\mu_i)\right|\geqslant t\right]\leqslant 2\exp\left(-\frac{2t^2}{n(b-a)^2}\right).$$

例 2.6 (U 统计量)

令 $g:\mathbb{R}^2\to\mathbb{R}$ 为一个关于两个自变量对称的函数，给定一列独立同分布的随机变量 $X_k(k\geqslant 1)$，统计量

$$U := \frac{1}{\binom{n}{2}}\sum_{j<k} g(X_j, X_k) \tag{2.26}$$

被称为 U 统计量。如果 g 有界（即 $\|g\|_\infty \leqslant b$），那么就有

$$\mathbb{P}[|U-\mathbb{E}[U]|\geqslant t]\leqslant 2\exp\left(-\frac{nt^2}{8b^2}\right).$$

这个尾部不等式保证了 U 是 $\mathbb{E}[U]$ 的相合估计。

证明 记 $U=f(\boldsymbol{x})=f(x_1,\cdots,x_n)$，对任意给定的 k，有

$$|f(\boldsymbol{x})-f(\boldsymbol{x}^{\backslash k})|\leqslant\frac{1}{\binom{n}{2}}\sum_{j\neq k}|g(x_j,x_k)-g(x_j,x_k')|\leqslant\frac{(n-1)(2b)}{\binom{n}{2}}=\frac{4b}{n},$$

即 f 满足参数为 $L_k=\dfrac{4b}{n}$ 的有界差性质，由推论 2.9 有界差不等式即证。 □

例 2.7 (Rademacher 复杂度)

令 $\{\varepsilon_k\}_{k=1}^n$ 为一个独立同分布的 Rademacher 随机变量序列。给定一个向量集合 $\mathcal{A}\subset\mathbb{R}^n$，定义随机变量

$$Z(\mathcal{A}) := \sup_{a\in\mathcal{A}}\left[\sum_{k=1}^n a_k\varepsilon_k\right]=\sup_{a\in\mathcal{A}}[\langle\boldsymbol{a},\boldsymbol{\varepsilon}\rangle]. \tag{2.27}$$

这个随机变量 Z 从某种意义上度量了 $\mathcal{A}$ 的大小，而其期望 $\mathcal{R}(\mathcal{A}) := \mathbb{E}[Z(\mathcal{A})]$ 称为集合 $\mathcal{A}$ 的 **Rademacher** 复杂度。

由有界差不等式可以得到，$Z(\mathcal{A})$ 是次高斯的，其参数至多为 $\sqrt{\sum_{k=1}^n \sup_{a\in\mathcal{A}} a_k^2}$。

证明 给定 $\mathcal{A}$，把 $Z(\mathcal{A})$ 看成是 $\boldsymbol{\varepsilon}$ 的函数 $f(\boldsymbol{\varepsilon})=f(\varepsilon_1,\cdots,\varepsilon_n)$，由于

$$\langle\boldsymbol{a},\boldsymbol{\varepsilon}\rangle-f(\boldsymbol{\varepsilon}^{\backslash k})\leqslant\langle a,\boldsymbol{\varepsilon}-\boldsymbol{\varepsilon}^{\backslash k}\rangle=a_k(\varepsilon_k-\varepsilon_k')\leqslant 2|a_k|,$$

两边同时在 $\mathcal{A}$ 上取最大值，则得到不等式

$$f(\boldsymbol{\varepsilon})-f(\boldsymbol{\varepsilon}^{\backslash k})\leqslant 2\sup_{a\in\mathcal{A}}|a_k|.$$

从而 f 满足有界差性质，由推论 2.9 得

$$\mathbb{P}\left[|Z(\mathcal{A})-\mathbb{E}[Z(\mathcal{A})]|\geqslant t\right]\leqslant 2\exp\left(\frac{t^2}{2\sum_{k=1}^n\sup_{a\in\mathcal{A}}a_k^2}\right)$$

□

2.3 高斯随机变量的 Lipschitz 函数

对于经典的高斯随机变量的 Lipschitz 函数的集中不等式，有一个非常有用的性质：集中度与维数无关。

称一个函数 $f:\mathbb{R}^n\to\mathbb{R}^n$ 在欧几里得范数 $\|\cdot\|_2$ 意义下是 L**-Lipschitz**的，如果

$$|f(\boldsymbol{x})-f(\boldsymbol{y})|\leqslant L\|\boldsymbol{x}-\boldsymbol{y}\|_2,\quad \forall \boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n. \tag{2.28}$$

定理 2.10

令 $(X_1,\cdots,X_n)$ 是由独立同分布的标准正态随机变量生成的向量，函数 $f:\mathbb{R}^n\to\mathbb{R}$ 是在欧几里得范数意义下的 L-Lipschitz 函数。那么随机变量 $f(\boldsymbol{X})-\mathbb{E}[f(\boldsymbol{X})]$ 是参数至多为 L 的次高斯随机变量，且

$$\mathbb{P}[|f(\boldsymbol{X})-\mathbb{E}[f(\boldsymbol{X})]|\geqslant t]\leqslant 2\exp\left(-\frac{t^2}{2L^2}\right),\quad \forall t\geqslant 0. \tag{2.29}$$

注 这个定理保证了对一个标准正态随机向量的任意 L-Lipschitz 函数，无论其维数如何，都是次高斯的，且参数最大为 L，其集中度的表现与一元的方差为 L^2 的正态随机变量相似。

证明 利用下面的引理可以证明一个常数弱化后的版本。

引理 2.11

若函数 $f:\mathbb{R}^n\to\mathbb{R}$ 可微，则对任意凸函数 $\phi:\mathbb{R}\to\mathbb{R}$，有

$$\mathbb{E}[\phi(f(X)-\mathbb{E}[f(X)])]\leqslant\mathbb{E}\left[\phi\left(\frac{\pi}{2}\langle\nabla f(X),Y\rangle\right)\right], \tag{2.30}$$

其中 $X,Y\sim\mathcal{N}(0,\boldsymbol{I}_n)$ 是独立的标准多元正态向量。

由此可以推出 $f(X)-\mathbb{E}[f(X)]$ 是参数为 $\dfrac{\pi L}{2}$ 的次高斯随机变量，取 Hoeffding 界即证。□

注 在上述定理的证明过程中，标准正态分布的各种性质起到了关键作用。但事实上，类似的结果对其他非正态分布同样成立，如球面上的均匀分布和任意严格对数凹的分布。但是如果没有关于函数 f 的结构性质，上述不依赖于维数集中度的结论对任意的次高斯分布未必成立。

例 2.8 (χ^2 集中度)

对任意给定的独立同分布的标准正态随机变量序列 $\{Z_k\}_{k=1}^n$，随机变量 $Y:=\sum\limits_{k=1}^n Z_k^2$ 服从自由度为 n 的 χ^2 分布。注意 Z_k^2 是次指数的且相互独立，可以得到 Y 的尾部不等式。事实上，基于正态分布的 Lipschitz 函数的集中不等式，可以给出另一种推导方式。

令 $V=\sqrt{\dfrac{Y}{n}}$，由于欧几里得范数是 1-Lipschitz 函数，对其应用定理 2.10。

例 2.9 (次序统计量)

给定随机向量 $(X_1,\cdots,X_n)$ 和 $(Y_1,\cdots,Y_n)$，有 $|X_{(k)}-Y_{(k)}|\leqslant\|\boldsymbol{X}-\boldsymbol{Y}\|_2$，故每个次序

统计量都是 1-Lipschitz 函数。因此，当 $\boldsymbol{X}$ 是一个正态随机向量时，由定理 2.10 有

$$\mathbb{P}\left[\left|X_{(k)}-\mathbb{E}[X_{(k)}]\right|\geqslant\delta\right]\leqslant 2e^{-\frac{\delta^2}{2}},\quad \forall\delta\geqslant 0.$$

例 2.10 (高斯复杂度)

令 $\{W_k\}_{k=1}^n$ 为独立同分布的标准正态随机变量序列。给定一个向量集合 $\mathcal{A}\subset\mathbb{R}^n$，定义随机变量

$$Z(\mathcal{A}):=\sup_{a\in\mathcal{A}}\left[\sum_{k=1}^n a_k W_k\right]=\sup_{a\in\mathcal{A}}[\langle\boldsymbol{a},\boldsymbol{W}\rangle]. \tag{2.31}$$

和 Rademacher 复杂度一样，随机变量 Z 也是刻画 $\mathcal{A}$ 大小的一种方式。

把 Z 看成 $\boldsymbol{W}$ 的函数 $f(\boldsymbol{W})=f(W_1,\cdots,W_n)$，则 f 是参数为 $\sup_{\boldsymbol{a}\in\mathcal{A}}\|\boldsymbol{a}\|_2$ 的 Lipschitz 函数。

证明 对 $\forall\boldsymbol{w},\boldsymbol{w}'\in\mathbb{R}^n$，令 $\boldsymbol{a}^*\in\mathcal{A}$ 为使 $f(\boldsymbol{w})$ 达到最大值的向量，则

$$f(\boldsymbol{w})-f(\boldsymbol{w}')\leqslant\langle\boldsymbol{a}^*,\boldsymbol{w}-\boldsymbol{w}'\rangle\leqslant\|\boldsymbol{a}^*\|_2\|\boldsymbol{w}-\boldsymbol{w}'\|_2\leqslant D(\mathcal{A})\|\boldsymbol{w}-\boldsymbol{w}'\|_2,$$

其中 $D(\mathcal{A})=\sup_{\boldsymbol{a}\in\mathcal{A}}\|\boldsymbol{a}\|_2$，故 f 是参数为 $D(\mathcal{A})$ 的 Lipschitz 函数，从而有

$$\mathbb{P}[|Z-\mathbb{E}[Z]|\geqslant\delta]\leqslant 2\exp\left(-\frac{\delta^2}{2D^2(\mathcal{A})}\right). \tag{2.32}$$

□

设 $\boldsymbol{Q}\in\mathbb{R}^{n\times n}$ 为一个对称矩阵，则矩阵 $\boldsymbol{Q}$ 的 ℓ_2 **算子范数**为

$$|||\boldsymbol{Q}|||_2:=\sup_{\|u\|_2=1}\|\boldsymbol{Q}u\|_2. \tag{2.33}$$

矩阵 $\boldsymbol{Q}$ 的 **Frobenius 范数**为

$$|||\boldsymbol{Q}|||_{\mathrm{F}}:=\sqrt{\sum_{i=1}^n\sum_{j=1}^n Q_{ij}^2}. \tag{2.34}$$

2.4 附录 A：次高斯随机变量的等价性

2.5 附录 B：次指数随机变量的等价性

第 3 章 测度集中度

3.1 基于熵技巧的集中度

3.1.1 熵及其相关性质

给定一个凸函数 $\phi : \mathbb{R} \to \mathbb{R}$，只要 $\mathbb{E}[X]$ 和 $\mathbb{E}[\phi(X)]$ 都存在，就可以定义概率空间上的一个泛函

$$\mathbb{H}_\phi(X) := \mathbb{E}[\phi(X)] - \phi(\mathbb{E}[X]), \tag{3.1}$$

其中 $X \sim \mathbb{P}$，称为随机变量 X 的 ϕ **熵**。由 ϕ 的凸性可知 $\mathbb{H}_\phi(X)$ 总是非负的，它是一个度量随机性变化大小的量。

实际上，有一些常见的度量随机性变化大小的量其实是 ϕ 熵的一种特殊形式：

- 取 $\phi(u) = u^2$，$\mathbb{H}_\phi(X) = \mathrm{Var}(X)$，这时的熵对应的是随机变量的方差；
- 取 $\phi(u) = -\log u$，$\mathbb{H}_\phi(X) = \log \mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right]$，这时的熵对应的是中心化的矩母函数。

在本章中，考虑的 $\phi : [0, +\infty) \to \mathbb{R}$ 定义为

$$\phi(u) := u \log u, \quad \forall u > 0, \quad \text{以及} \quad \phi(0) := 0. \tag{3.2}$$

此时，对任意非负随机变量 $Z \geqslant 0$，定义 ϕ 熵为

$$\mathbb{H}(Z) = \mathbb{E}[Z \log Z] - \mathbb{E}[Z] \log \mathbb{E}[Z]. \tag{3.3}$$

对于随机变量 $Z := e^{\lambda X}$，熵可以通过矩母函数 $\varphi_X(\lambda) = \mathbb{E}\left[e^{\lambda X}\right]$ 及其一阶导数表示为

$$\mathbb{H}(e^{\lambda X}) = \lambda \varphi_X'(\lambda) - \varphi_X(\lambda) \log \varphi_X(\lambda). \tag{3.4}$$

例 3.1 (正态随机变量的熵)

对于一元正态随机变量 $X \sim \mathcal{N}(0, \sigma^2)$，有 $\varphi_X(\lambda) = e^{\lambda^2\sigma^2/2}$，$\varphi_X'(\lambda) = \lambda\sigma^2\varphi_X(\lambda)$，因此

$$\mathbb{H}(e^{\lambda X}) = \lambda^2\sigma^2\varphi_X(\lambda) - \frac{1}{2}\lambda^2\sigma^2\varphi_X(\lambda) = \frac{1}{2}\lambda^2\sigma^2\varphi_X(\lambda) \tag{3.5}$$

3.1.2 Herbst 方法及其延伸

直观上来说，熵是一个度量随机变量波动性的量，因此熵的控制可以转化为尾部概率界的控制。考虑一类特定的随机变量，假设存在常数 $\sigma > 0$ 使得 $e^{\lambda X}$ 的熵满足上界

$$\mathbb{H}(e^{\lambda X}) \leqslant \frac{1}{2}\sigma^2\lambda^2\varphi_X(\lambda). \tag{3.6}$$

注 任意正态随机变量 $X \sim \mathcal{N}(0, \sigma^2)$ 对 $\forall \lambda \in \mathbb{R}$ 满足上述条件且取等号，任意有界随机变量也满足上述条件。

命题 3.1 (Herbst 方法)

假设熵 $\mathbb{H}(e^{\lambda X})$ 对 $\forall\lambda\in I$ 满足不等式 (3.6)，这里 I 可以是区间 $[0,+\infty)$ 或 $\mathbb{R}$，则 X 满足

$$\log\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}]\leqslant\frac{1}{2}\lambda^2\sigma^2,\quad \forall\lambda\in I. \tag{3.7}$$

注 当 $I=\mathbb{R}$ 时，等价于中心化后的随机变量 $X-\mathbb{E}[X]$ 是参数为 σ 的次高斯随机变量。当 $I+[0,+\infty)$ 时，对应的是次高斯变量的单侧尾部概率界。

证明 由熵的矩母函数表达式，条件等价于矩母函数 $\varphi=\varphi_X$ 满足

$$\lambda\varphi'(\lambda)-\varphi(\lambda)\log\varphi(\lambda)\leqslant\frac{1}{2}\sigma^2\lambda^2\varphi(\lambda),\quad \forall\lambda\geqslant 0.$$

定义函数 $G(\lambda)=\frac{1}{\lambda}\log\varphi(\lambda),\varphi(0)=\mathbb{E}[X]$，则上式等价于 $G'(\lambda)\leqslant\frac{1}{2}\sigma^2$，进而有

$$G(\lambda)-\mathbb{E}[X]\leqslant\frac{1}{2}\lambda\sigma^2,$$

化简即证。 □

命题 3.2 (Bernstein 熵的界)

假设存在正常数 b 和 σ 使得熵 $\mathbb{H}(e^{\lambda X})$ 满足

$$\mathbb{H}(e^{\lambda X})\leqslant\lambda^2\{b\varphi_x'(\lambda)+\varphi_x(\lambda)(\sigma^2-b\mathbb{E}[X])\},\quad \forall\lambda\in[0,1/b). \tag{3.8}$$

则 X 满足界

$$\log\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}]\leqslant\frac{\sigma^2\lambda^2}{1-b\lambda},\quad \forall\lambda\in[0,1/b). \tag{3.9}$$

证明 由重尺度化和中心化技巧，不失一般性，不妨设 $\mathbb{E}[X]=0$，$b=1$，类似命题 3.1 即证。□

3.1.3 可分凸函数和熵方法

如果对 $\forall k\in\{1,\cdots,n\}$，单变量函数 $y_k\mapsto f(x_1,\cdots,x_{k-1},y_k,x_{k+1},\cdots,x_n)$ 对任意给定的向量 $(x_1,\cdots,x_{k-1},x_{k+1},\cdots,x_n)\in\mathbb{R}^{n-1}$ 是凸的，则称函数 $f:\mathbb{R}^n\to\mathbb{R}$ 是**可分凸函数**。

定理 3.3

令 $\{X_i\}_{i=1}^n$ 是区间 $[a,b]$ 上的独立随机变量，$f:\mathbb{R}^n\to\mathbb{R}$ 是可分凸且关于欧几里得范数是 L-Lipschitz 的，那么对 $\forall\delta>0$，有

$$\mathbb{P}\left[f(X)\geqslant\mathbb{E}[f(X)]+\delta\right]\leqslant\exp\left(-\frac{\delta^2}{4L^2(b-a)^2}\right). \tag{3.10}$$

定理 3.3 可以用来获得一系列重要问题的最优界 (order-optimal bound)。

例 3.2 (Rademacher 复杂度的更优界)

给定一个有界子集 $\mathcal{A}\subset\mathbb{R}^n$，考虑随机变量 $Z=\sup_{a\in\mathcal{A}}\sum_{k=1}^n a_k\varepsilon_k$，其中 $\varepsilon_k\in\{-1,+1\}$ 为独立同分布的 Rademacher 随机变量，求 $Z-\mathbb{E}[Z]$ 的尾部概率界。

解 易验证 $Z=f(\boldsymbol{\varepsilon})=f(\varepsilon_1,\cdots,\varepsilon_n)$ 是完全凸的，因此也是可分凸的。对 $\forall \boldsymbol{a}\in\mathcal{A}$，有

$$\langle \boldsymbol{a},\boldsymbol{\varepsilon}\rangle - f(\boldsymbol{\varepsilon}') = \langle \boldsymbol{a},\boldsymbol{\varepsilon}\rangle - \sup_{\boldsymbol{a}'\in\mathcal{A}}\langle \boldsymbol{a}',\boldsymbol{\varepsilon}'\rangle \leqslant \langle \boldsymbol{a},\boldsymbol{\varepsilon}-\boldsymbol{\varepsilon}'\rangle \leqslant \|\boldsymbol{a}\|_2\|\boldsymbol{\varepsilon}-\boldsymbol{\varepsilon}'\|_2.$$

从而可以推出 Z 是 Lipschitz 的，参数为 $\mathcal{W}(A):=\sup\limits_{a\in\mathcal{A}}\|\boldsymbol{a}\|_2$。由定理 3.3，有

$$\mathbb{P}[Z\geqslant \mathbb{E}[Z]+t]\leqslant \exp\left(-\frac{t^2}{16W^2(\mathcal{A})}\right). \tag{3.11}$$

注意这里的参数 $\mathcal{W}^2(\mathcal{A})$ 可能比例 2.7 中得到的 $\sum\limits_{k=1}^{n}\sup\limits_{a\in\mathcal{A}} a_k^2$ 小很多。 □

3.1.4 张量化和可分凸函数

为了证明定理 3.3，先引入两个引理，这两个引理本身也都很重要。

引理 3.4 (单变量函数基于熵的界)

设 $X,Y\sim\mathbb{P}$ 是一对独立同分布的随机变量，则对任意函数 $g:\mathbb{R}\to\mathbb{R}$，有

$$\mathbb{H}\left(e^{\lambda g(X)}\right)\leqslant \lambda^2\mathbb{E}\left[(g(X)-g(Y))^2 e^{\lambda g(X)}\mathbb{I}[g(X)\geqslant g(Y)]\right],\quad \forall\lambda>0. \tag{3.12}$$

另外，如果 X 以 $[a,b]$ 为支撑集，g 是凸且 Lipschitz 的，则

$$\mathbb{H}\left(e^{\lambda g(X)}\right)\leqslant \lambda^2(b-a)^2\mathbb{E}\left[(g'(X))^2e^{\lambda g(X)}\right],\quad \forall\lambda>0, \tag{3.13}$$

其中 g' 为 g 的导数。

注 由 **Rademacher** 定理，任意一个凸的 Lipschitz 函数几乎处处有导数。进一步，如果 g 是参数为 L 的 Lipschitz 函数，则有 $\|g'\|_\infty\leqslant L$，于是可以进一步推出

$$\mathbb{H}\left(e^{\lambda g(X)}\right)\leqslant \lambda^2L^2(b-a)^2\mathbb{E}\left[e^{\lambda g(X)}\right],\quad \forall\lambda>0.$$

应用命题 3.1，可以推出定理 3.3 的单变量版本。

为了将一元的结果推广至多元的情形，熵的**张量化性质**会起到关键作用。考虑一个函数 $f:\mathbb{R}^n\to\mathbb{R}$，下标 $k\in\{1,\cdots,n\}$ 和向量 $\boldsymbol{x}_{\backslash k}=(x_i,i\neq k)\in\mathbb{R}^{n-1}$，定义坐标 k 的**条件熵**为

$$\mathbb{H}(e^{\lambda f_k(X_k)}\mid x_{\backslash k}):=\mathbb{H}(e^{\lambda f(x_1,\dots,x_{k-1},X_k,x_{k+1},\dots,x_n)}), \tag{3.14}$$

其中 $f_k:\mathbb{R}\to\mathbb{R}$ 是坐标函数 $x_k\mapsto f(x_1,\cdots,x_k,\cdots,x_n)$。

定理 3.5 (熵的张量化)

令 $f:\mathbb{R}^n\to\mathbb{R}$，$\{X_k\}_{k=1}^n$ 为独立随机变量，则

$$\mathbb{H}\left(e^{\lambda f(X_1,\dots,X_n)}\right)\leqslant \mathbb{E}\left[\sum_{k=1}^{n}\mathbb{H}\left(e^{\lambda f_k(X_k)}\mid X^{\backslash k}\right)\right],\quad \forall\lambda>0. \tag{3.15}$$

注 这个结果说明，多变量熵可以被适当定义的单变量条件熵之和控制住上界。

证明 [定理 3.3 的证明] 由引理 3.4，对 $\forall\lambda>0$，有

$$\begin{aligned}\mathbb{H}\left(e^{\lambda f_k(X_k)} \mid x_{\setminus k}\right) &\leqslant \lambda^2(b-a)^2\mathbb{E}_{X_k}\left[(f_k'(X_k))^2 e^{\lambda f_k(X_k)} \mid x_{\setminus k}\right] \\ &= \lambda^2(b-a)^2\mathbb{E}_{X_k}\left[\left(\frac{\partial f(x_1,\ldots,X_k,\ldots,x_n)}{\partial x_k}\right)^2 e^{\lambda f(x_1,\ldots,X_k,\ldots,x_n)}\right],\end{aligned}$$

结合引理 3.5，

$$\mathbb{H}\left(e^{\lambda f(X)}\right) \leqslant \lambda^2(b-a)^2\mathbb{E}\left[\sum_{k=1}^n\left(\frac{\partial f(X)}{\partial x_k}\right)^2 e^{\lambda f(X)}\right] \leqslant \lambda^2(b-a)^2L^2\mathbb{E}\left[e^{\lambda f(X)}\right].$$

再由命题 3.1 即证。 □

3.2 集中度的几何观点

度量测度空间指的是一个度量空间 $(\mathcal{X},\rho)$ 在其 Borel 集上被赋予了一个概率测度 $\mathbb{P}$。经典的度量空间如 $\mathcal{X}=\mathbb{R}^n$ 上赋予了欧几里得度量 $\rho(\boldsymbol{x},\boldsymbol{y})=\|\boldsymbol{x}-\boldsymbol{y}\|_2$ 和离散方体 $\mathcal{X}=\{0,1\}^n$ 上赋予了 Hamming 度量 $\rho(\boldsymbol{x},\boldsymbol{y})=\sum_{i=1}^n\mathbb{I}[x_i\neq y_i]$。

3.2.1 集中度函数

与度量测度空间相关的是集中度函数，它是借助集合的 ϵ 扩张用几何方式定义的，详细刻画了 ϵ 扩张趋向于 1 的速度有多快。

给定一个集合 $A\subseteq\mathcal{X}$ 和一个点 $x\in\mathcal{X}$，定义点 x 到集合 A 的距离为

$$\rho(x,A) := \inf_{y\in A}\rho(x,y). \tag{3.16}$$

给定一个参数 $\epsilon>0$，定义 A 的 ϵ **扩张**为

$$A^\epsilon := \{x\in\mathcal{X} : \rho(x,A)<\epsilon\}. \tag{3.17}$$

定义 3.1 (集中度函数)

度量测度空间 $(\mathbb{P},\mathcal{X},\rho)$ 的集中度函数 $\alpha:[0,+\infty)\to\mathbb{R}_+$ 定义为

$$\alpha_{\mathbb{P},(\mathcal{X},\rho)}(\epsilon) := \sup_{A\subseteq\mathcal{X}}\left\{1-\mathbb{P}[A^\epsilon] : \mathbb{P}[A]\geqslant\frac{1}{2}\right\}, \tag{3.18}$$

其中上确界在所有可测子集 A 上取。有时也把上述记号简记为 $\alpha_{\mathbb{P}}$。

我们主要感兴趣的是，当 ϵ 增加时，集中度函数趋向于 0 的速度有多快。

例 3.3 (球面的集中度函数)

考虑 n 维欧几里得球面上的均匀分布定义的度量测度空间

$$\mathbb{S}^{n-1} := \{\boldsymbol{x}\in\mathbb{R}^n : \|\boldsymbol{x}\|_2=1\}, \tag{3.19}$$

其上的度量为测地线度量 $\rho(\boldsymbol{x},\boldsymbol{y})=\arccos\langle\boldsymbol{x},\boldsymbol{y}\rangle$。

3.2.2 与 Lipschitz 函数的联系

命题 3.6 (集中度函数的控制与 Lipschitz 函数的控制等价)

给定随机变量 $X \sim \mathbb{P}$ 和集中度函数 $\alpha_{\mathbb{P}}$，任意一个 $(\mathcal{X}, \rho)$ 上的 1-Lipschitz 函数 f 满足

$$\mathbb{P}[|f(X) - m_f| \geqslant \epsilon] \leqslant 2\alpha_{\mathbb{P}}(\epsilon), \tag{3.20}$$

其中 m_f 是 f 的任意中位数。

反之，假设有一个函数 $\beta : \mathbb{R}_+ \to \mathbb{R}_+$，满足对任意一个 $(\mathcal{X}, \rho)$ 上的 1-Lipschitz 函数 f 有

$$\mathbb{P}[f(X) \geqslant \mathbb{E}[f(X)] + \epsilon] \leqslant \beta(\epsilon), \quad \forall \epsilon \geqslant 0, \tag{3.21}$$

那么集中度函数满足界 $\alpha_{\mathbb{P}}(\epsilon) \leqslant \beta(\epsilon/2)$。

3.2.3 从几何到集中度

称函数 $\psi : \mathbb{R}^n \to \mathbb{R}$ 是**强凸函数**，如果存在常数 $\gamma > 0$，对 $\forall \lambda \in [0, 1]$ 和 $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$，有

$$\lambda\psi(x) + (1 - \lambda)\psi(y) - \psi(\lambda x + (1 - \lambda)y) \geqslant \frac{\gamma}{2}\lambda(1 - \lambda)\|x - y\|_2^2. \tag{3.22}$$

称一个分布 $\mathbb{P}$ 关于密度 p 是**强对数凹分布**，如果这个密度可以写成 $p(\boldsymbol{x}) = \exp(-\psi(\boldsymbol{x}))$ 的形式，其中函数 $\psi : \mathbb{R}^n \to \mathbb{R}$ 是强凸的。例如，n 维标准正态分布是参数 $\gamma = 1$ 的强对数凹分布。

定理 3.7

设 $\mathbb{P}$ 是参数 $\gamma > 0$ 的强对数凹分布，那么对任意欧几里得范数意义下的 L-Lipschitz 函数 $f : \mathbb{R}^n \to \mathbb{R}$ 有

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geqslant t] \leqslant 2\exp\left(-\frac{\gamma t^2}{4L^2}\right). \tag{3.23}$$

3.3 Wasserstein 距离和传输成本不等式

3.3.1 Wasserstein 距离

给定一个度量空间 $(\mathcal{X}, \rho)$，称一个函数 $f : \mathcal{X} \to \mathbb{R}$ 关于度量 ρ 是 **L-Lipschitz**的，如果

$$|f(x) - f(x')| \leqslant L\rho(x, x'), \quad \forall x, x' \in \mathcal{X}. \tag{3.24}$$

记 $\|f\|_{\text{Lip}}$ 为满足上述不等式的最小的 L。

给定 $\mathcal{X}$ 上的两个概率分布 $\mathbb{Q}, \mathbb{P}$，称由 ρ **诱导出的 Wasserstein 度量**为

$$W_\rho(\mathbb{Q}, \mathbb{P}) = \sup_{\|f\|_{\text{Lip}} \leqslant 1} \left[\int f \mathrm{d}\mathbb{Q} - \int f \mathrm{d}\mathbb{P}\right]. \tag{3.25}$$

对于每一个 ρ 的选取，这定义了一个概率测度空间上的距离。

例 3.4 (Hamming 度量和全变差距离)

考虑 Hamming 度量 $\rho(x,x')=\mathbb{I}[x\neq x']$，我们断言，与此相对应的 Wasserstein 距离等价于全变差距离

$$\|\mathbb{Q}-\mathbb{P}\|_{\mathrm{TV}}:=\sup_{A\subseteq\mathcal{X}}|\mathbb{Q}(A)-\mathbb{P}(A)|, \tag{3.26}$$

这里的上确界在所有可测子集 A 上取。

由**对偶理论**中一个经典的结果，任意 Wasserstein 距离都有一个基于耦合距离类型的等价定义。称乘积空间 $\mathcal{X}\times\mathcal{X}$ 上的一个分布 $\mathbb{M}$ 为关于 $(\mathbb{Q},\mathbb{P})$ 的一个**耦合**，如果其第一和第二坐标的边缘分布分别与 $\mathbb{Q}$ 和 $\mathbb{P}$ 相吻合。

事实上，关于 Wasserstein 距离和耦合，我们有等价性

$$W_\rho(\mathbb{Q},\mathbb{P})=\sup_{\|f\|_{\mathrm{Lip}}\leqslant 1}\int f\,(\mathrm{d}\mathbb{Q}-\mathrm{d}\mathbb{P})=\inf_{\mathbb{M}}\int_{\mathcal{X}\times\mathcal{X}}\rho(x,x')\mathrm{d}\mathbb{M}(x,x')=\inf_{\mathbb{M}}\mathbb{E}_{\mathbb{M}}[\rho(X,X')], \tag{3.27}$$

其中下确界在所有 $(\mathbb{Q},\mathbb{P})$ 的耦合 $\mathbb{M}$ 上取。这个基于耦合的 Wasserstein 距离表达式在接下来的证明中非常重要。

“传输成本”的术语来自于对基于耦合的表达式 (3.27) 的另一种理解。考虑 $\mathcal{X}$ 上 $\mathbb{P},\mathbb{Q}$ 关于 Lebesgue 测度有密度 p,q，且在乘积空间上的耦合 $\mathbb{M}$ 关于乘积空间上的 Lebesgue 测度有密度 m。密度 p 可以看成是 $\mathcal{X}$ 上的一些初始的质量分布，而密度 q 则可以看成是一些想要得到的分布，我们的目标是对质量进行传输，从而将初始分布 p 变换为想要的分布 q。$\rho(x,x')\mathrm{d}x\mathrm{d}x'$ 可以理解为将一个质量的小增量 $\mathrm{d}x$ 传输到新的增量 $\mathrm{d}x'$ 的损失，联合密度 $m(x,x')$ 称为**传输计划**，是一个将 p 转换成 q 的质量变换体系，则计划 m 对应的**传输成本**为

$$\int_{\mathcal{X}\times\mathcal{X}}\rho(x,x')m(x,x')\mathrm{d}x\mathrm{d}x',$$

取下确界就是 Wasserstein 距离。

3.3.2 传输成本和集中不等式

给定两个分布 $\mathbb{Q},\mathbb{P}$，两者之间的 **Kullback-Leibler(KL) 散度**为

$$D(\mathbb{Q}\,\|\,\mathbb{P}):=\begin{cases}\mathbb{E}_{\mathbb{Q}}\left[\log\dfrac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right], & \mathbb{Q}\text{ 关于 }\mathbb{P}\text{ 绝对连续}\\ +\infty, & \text{否则}\end{cases} \tag{3.28}$$

如果对某个测度 ν，这些分布有密度 q,p，那么 KL 散度可以写成

$$D(\mathbb{Q}\,\|\,\mathbb{P})=\int_X q(x)\log\frac{q(x)}{p(x)}\nu(\mathrm{d}x). \tag{3.29}$$

传输成本不等式指的是 Wasserstein 距离能被 KL 散度平方根的一个倍数所控制。

定义 3.2 (传输成本不等式)

对于一个给定的度量 ρ，概率测度 $\mathbb{P}$ 称作满足参数为 $\gamma>0$ 的 ρ 传输成本不等式，如果

对任意概率测度 $\mathbb{Q}$ 有

$$W_\rho(\mathbb{Q},\mathbb{P}) \leqslant \sqrt{2\gamma D(\mathbb{Q}\,\|\,\mathbb{P})}. \tag{3.30}$$

由于 KL 散度在信息论中非常重要，这个结果也被称为**信息不等式**。

由 Wasserstein 距离的定义，传输成本不等式 (3.30) 可以通过 KL 散度 $D(\mathbb{Q}\,\|\,\mathbb{P})$ 来控制偏差 $\int f\mathrm{d}\mathbb{Q} - \int f\mathrm{d}\mathbb{P}$ 的上界。因此，可以选取一个特定的分布 $\mathbb{Q}$ 来推导 $\mathbb{P}$ 下 f 的集中度界。

定理 3.8 (从传输成本到集中度)

考虑一个度量测度空间 $(\mathbb{P},\mathcal{X},\rho)$，假设 $\mathbb{P}$ 满足 ρ 传输成本不等式 (3.30)，那么它的集中度满足界

$$\alpha_{\mathbb{P},(X,\rho)}(t) \leqslant 2\exp\left(-\frac{t^2}{2\gamma}\right). \tag{3.31}$$

进一步，对任意 $X\sim\mathbb{P}$ 和任意 L-Lipschitz 函数 $f:\mathcal{X}\to\mathbb{R}$，有集中不等式

$$\mathbb{P}\left[|f(X)-\mathbb{E}[f(X)]|\geqslant t\right] \leqslant 2\exp\left(-\frac{t^2}{2\gamma L^2}\right). \tag{3.32}$$

证明 略。 □

3.3.3 传输成本的张量化

命题 3.9

假设对每个 $k=1,\cdots,n$，单变量分布 $\mathbb{P}_k$ 满足参数为 γ_k 的 ρ_k 传输成本不等式，那么乘积分布 $\mathbb{P}=\bigotimes_{k=1}^n\mathbb{P}_k$ 满足传输成本不等式

$$W_\rho(\mathbb{Q},\mathbb{P}) \leqslant \sqrt{2\left(\sum_{k=1}^n\gamma_k\right)D(\mathbb{Q}\,\|\,\mathbb{P})},\quad \text{对所有分布 } \mathbb{Q}, \tag{3.33}$$

其中 Wasserstein 度量的定义基于距离 $\rho(x,y):=\sum_{k=1}^n\rho_k(x_k,y_k)$。

证明 略。 □

例 3.5 (有界差不等式)

设 f 满足有界差性质 (2.24)，由三角不等式，f 是 1-Lipschitz 函数，对应的是重尺度化 Hamming 度量 $\rho(\boldsymbol{x},\boldsymbol{y})=\sum_{k=1}^n\rho_k(x_k,y_k)$，其中 $\rho_k(x_k,y_k):=L_k\mathbb{I}[x_i\neq y_i]$。

由 Pinsker-Csiszár-Kullback 不等式，每个单变量分布 $\mathbb{P}_k$ 满足参数为 $\gamma_k=\frac{L_k^2}{4}$ 的 ρ_k 传输

成本不等式，由命题 3.9，$\mathbb{P}=\bigotimes_{k=1}^{n}\mathbb{P}_k$ 满足参数为 $\gamma=\frac{1}{4}\sum_{k=1}^{n}L_k^2$ 的 ρ 传输成本不等式。因为 f 在度量 ρ 下是 1-Lipschitz 的，由定理 3.8 知

$$\mathbb{P}[|f(X)-\mathbb{E}[f(X)]|\geqslant t]\leqslant 2\exp\left(-\frac{2t^2}{\sum_{k=1}^{n}L_k^2}\right),\quad \forall t\geqslant 0. \tag{3.34}$$

3.3.4 马尔可夫链的传输成本不等式

设 $(X_1,\cdots,X_n)$ 是一个由马尔可夫链产生的随机向量，其中每个 X_i 在可数空间 $\mathcal{X}$ 上取值。其在 $\mathcal{X}^n$ 上的分布 $\mathbb{P}$ 由一个初始分布 $X_1\sim\mathbb{P}_1$ 和转移核函数

$$\mathbb{K}_{i+1}(x_{i+1}\mid x_i)=\mathbb{P}_{i+1}(X_{i+1}=x_{i+1}\mid X_i=x_i) \tag{3.35}$$

定义。考虑离散空间上 β **收缩的马尔可夫链**，即存在 $\beta\in[0,1)$ 满足

$$\max_{i=1,\dots,n-1}\sup_{x_i,x_i'}\|\mathbb{K}_{i+1}(\cdot\mid x_i)-\mathbb{K}_{i+1}(\cdot\mid x_i')\|_{\mathrm{TV}}\leqslant\beta. \tag{3.36}$$

定理 3.10

令 $\mathbb{P}$ 是离散空间 $\mathcal{X}^n$ 上的一个 β 收缩的马尔可夫链的分布函数，那么对于 $\mathcal{X}^n$ 上的任意其他分布 $\mathbb{Q}$，有

$$W_\rho(\mathbb{Q},\mathbb{P})\leqslant\frac{1}{1-\beta}\sqrt{\frac{n}{2}D(\mathbb{Q}\,\|\,\mathbb{P})}, \tag{3.37}$$

其中 Wasserstein 距离是基于 Hamming 范数 $\rho(\boldsymbol{x},\boldsymbol{y})=\sum_{i=1}^{n}\mathbb{I}[x_i\neq y_i]$ 定义的。

3.3.5 非对称耦合成本

和定理 2.10 的结果类似，基于 ℓ_2 范数的 Lipschitz 条件通常可以导出独立于维数的结果。

定理 3.11

考虑一个独立随机变量的向量 $(X_1,\cdots,X_n)$，每个都在 $[0,1]$ 上取值。设 $f:\mathbb{R}^n\to\mathbb{R}$ 是凸的，并在欧几里得范数下是 L-Lipschitz 的，则有

$$\mathbb{P}\left[|f(\boldsymbol{X})-\mathbb{E}[f(\boldsymbol{X})]|\geqslant t\right]\leqslant 2\exp\left(-\frac{t^2}{2L^2}\right),\quad \forall t\geqslant 0. \tag{3.38}$$

证明 略。 □

例 3.6 (回顾 Rademacher 复杂度)

由上述定理，可以得到 Rademacher 复杂度更精细的界

$$\mathbb{P}\left[|Z-\mathbb{E}[Z]|\geqslant t\right]\leqslant 2\exp\left(-\frac{t^2}{2\mathcal{W}^2(\mathcal{A})}\right). \tag{3.39}$$

3.4 经验过程的尾部概率界

本节我们使用熵方法来推导经验过程上确界的多种尾部概率界。

设 $\mathcal{F}$ 为一类函数 $f:\mathcal{X}\to\mathbb{R}$，设 $(X_1,\cdots,X_n)$ 来自一个乘积分布 $\mathbb{P}=\bigotimes_{i=1}^{n}\mathbb{P}_i$，其中每个 $\mathbb{P}_i$ 以某个 $\mathcal{X}_i\subset\mathcal{X}$ 为支撑集。考虑随机变量

$$Z=\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^{n}f(X_i)\right\}, \tag{3.40}$$

这一节的主要目标是推导关于尾事件 $\{Z\geqslant\mathbb{E}[Z]+\delta\}$ 的上界。

3.4.1 一个泛函 Hoeffding 不等式

定理 3.12 (泛函 Hoeffding 定理) 对于每个 $f\in\mathcal{F}$ 和 $i=1,\cdots,n$，假设存在实数 $a_{i,f}\leqslant b_{i,f}$，对 $\forall x\in\mathcal{X}_i$，满足 $f(x)\in[a_{i,f},b_{i,f}]$，那么对 $\forall\delta\geqslant 0$，有

$$\mathbb{P}\left[Z\geqslant\mathbb{E}[Z]+\delta\right]\leqslant\exp\left(-\frac{n\delta^2}{4L^2}\right), \tag{3.41}$$

其中 $L^2:=\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^{n}(b_{i,f}-a_{i,f})^2\right\}$。

证明 略。 □

3.4.2 一个泛函 Bernstein 不等式

定理 3.13 (经验过程的 Talagrand 集中度) 考虑一个被 b 一致控制的可数函数类 $\mathcal{F}$，那么对 $\forall\delta>0$，随机变量 Z 满足上尾部概率界

$$\mathbb{P}[Z\geqslant\mathbb{E}[Z]+\delta]\leqslant 2\exp\left(\frac{-n\delta^2}{8e\mathbb{E}[\Sigma^2]+4b\delta}\right), \tag{3.42}$$

其中 $\Sigma^2=\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^{n}f^2(X_i)\right\}$。

证明 略。 □

第 4 章 一致大数定律

第 5 章 2024 高维期末范围

第 2 章

例 5.1 (习题 2.4：有界随机变量紧的次高斯参数)

考虑一个均值为 $\mu = \mathbb{E}[X]$ 的随机变量 X，满足 $X \in [a, b]$, a.s. 。

(a) 定义函数 $\psi(\lambda) = \log \mathbb{E}[e^{\lambda X}]$，证明：$\psi(0) = 0$ 且 $\psi'(0) = \mu$。

(b) 证明 $\psi''(\lambda) = \mathbb{E}_\lambda[X^2] - (\mathbb{E}_\lambda[X])^2$，其中 $\mathbb{E}_\lambda[f(X)] := \dfrac{\mathbb{E}[f(X)e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]}$，并由此得到一个 $\sup\limits_{\lambda \in \mathbb{R}} |\psi''(\lambda)|$ 的上界。

(c) 用 (a) 和 (b) 证明 X 是次高斯的，且参数至多为 $\sigma = \dfrac{b-a}{2}$。

证明 易证。由 X 的有界性可得 $|\psi''(\lambda)| = \mathrm{Var}_\lambda(X) \leqslant \left(\dfrac{b-a}{2}\right)^2$。 □

例 5.2 (习题 2.5：次高斯界和均值方差)

设随机变量 X 满足

$$\mathbb{E}[e^{\lambda X}] \leqslant \exp\left(\frac{\lambda^2\sigma^2}{2} + \lambda\mu\right), \quad \forall \lambda \in \mathbb{R}.$$

(a) 证明 $\mathbb{E}[X] = \mu$。

(b) 证明 $\mathrm{Var}(X) \leqslant \sigma^2$。

(c) 假设 σ 的取值是使条件中的不等式成立的最小的 σ，那么 $\mathrm{Var}(X) = \sigma^2$ 是否成立？

证明 (a)(b) 泰勒展开分别令 $\lambda \to 0^+$ 和 $\lambda \to 0^-$ 即可。

(c) 否，反例考虑以较小概率取较大值的随机变量，如 $\mathrm{Bernoulli}(1, p)$ 中取 $p = \dfrac{1}{4}$。 □

例 5.3 (习题 2.11：正态随机变量最大值的上下界)

设 $\{X_i\}_{i=1}^n$ 为 i.i.d. 的 $\mathcal{N}(0, \sigma^2)$ 随机变量序列，考虑随机变量 $Z_n := \max\limits_{i=1,\cdots,n} |X_i|$。

(a) 证明

$$\mathbb{E}[Z_n] \leqslant \sqrt{2\sigma^2 \log n} + \frac{4\sigma}{\sqrt{2\log n}}, \quad \forall n \geqslant 2.$$

(提示：标准正态随机变量的尾部概率不等式 $\mathbb{P}[U \geqslant \delta] \leqslant \sqrt{\frac{2}{\pi}}\frac{1}{\delta}e^{-\delta^2/2}$)

(b) 证明

$$\mathbb{E}[\mathbb{Z}_n] \geqslant (1 - 1/e)\sqrt{2\sigma^2 \log n}, \quad \forall n \geqslant 5.$$

(c) 证明 $\dfrac{\mathbb{E}[Z_n]}{\sqrt{2\sigma^2 \log n}} \to 1$ 当 $n \to \infty$。

证明 (a) 记 $c=\sqrt{2\log n}$，则

$$\begin{aligned}
\mathbb{E}[Z_n] &= \int_0^\infty \mathbb{P}[Z_n\geqslant t]dt \leqslant c+\int_c^\infty \mathbb{P}[Z_n\geqslant t]dt \\
&\leqslant c+2n\int_c^\infty \mathbb{P}[X_1\geqslant t]dt \leqslant c+4n\int_c^\infty \frac{1}{\sqrt{2\pi}}t^{-1}e^{-t^2/2}dt \\
&\leqslant c+\frac{4n}{c}\int_c^\infty \frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt \\
&= c+\frac{4}{c}\left[n(1-\Phi(c))\right]
\end{aligned}$$

由 $n(1-\Phi(\sqrt{2\log n}))\leqslant 1$ 即证。

(b)(c) ??? □

例 5.4 (习题 2.12：次高斯随机变量最大值的上界)

设 $\{X_i\}_{i=1}^n$ 为均值为 0，参数为 σ 的次高斯随机变量序列（不要求独立性）。

(a) 证明

$$\mathbb{E}\left[\max_{i=1,\ldots,n} X_i\right]\leqslant\sqrt{2\sigma^2\log n},\quad \forall n\geqslant 1.$$

(提示：指数函数为凸函数)

(b) 证明随机变量 $Z:=\max\limits_{i=1,\cdots,n}|X_i|$ 满足

$$\mathbb{E}[Z]\leqslant\sqrt{2\sigma^2\log(2n)}\leqslant 2\sqrt{\sigma^2\log n},\quad \forall n\geqslant 2.$$

证明

(a) 由 Jensen 不等式，对 $\forall\lambda>0$

$$\exp\left(\lambda\mathbb{E}\left[\max_{1\leqslant i\leqslant n}X_i\right]\right)\leqslant\mathbb{E}\left[\max_{1\leqslant i\leqslant n}e^{\lambda X_i}\right]\leqslant\sum_{i=1}^n\mathbb{E}\left[e^{\lambda X_i}\right]\leqslant ne^{\frac{\lambda^2\sigma^2}{2}},$$

从而

$$\mathbb{E}\left[\max_{1\leqslant i\leqslant n}X_i\right]\leqslant\inf_{\lambda>0}\left\{\frac{\log n}{\lambda}+\frac{\lambda\sigma^2}{2}\right\}=\sqrt{2\sigma^2\log n}.$$

(b) 记 $Y_i=X_i, Y_{n+1}=-X_i, i=1,\cdots,n$，则 $Z=\mathbb{E}\left[\max\limits_{1\leqslant i\leqslant 2n}Y_i\right]$，由 (a) 易证。

□

例 5.5 (习题 2.14：中位数和均值的集中不等式)

给定一元随机变量 X，假设存在正常数 c_1,c_2 满足

$$\mathbb{P}[|X-\mathbb{E}[X]|\geqslant t]\leqslant c_1e^{-c_2t^2},\quad\forall t\geqslant 0.$$

(a) 证明：$\mathrm{Var}(X)\leqslant\frac{c_1}{c_2}$。

(b) 设中位数 m_X 是任意满足 $\mathbb{P}[X\geqslant m_X]\geqslant 1/2$ 和 $\mathbb{P}[X\leqslant m_X]\geqslant 1/2$ 的数，给出一个中位数不唯一的例子。

(c) 证明只要均值的集中不等式成立，那么对任意中位数 m_X，也有集中不等式成立

$$\mathbb{P}[|X-m_X|\geqslant t]\leqslant c_3e^{-c_4t^2},\quad\forall t\geqslant 0,$$

其中 $c_3=4c_1, c_4=\dfrac{c_2}{8}$。

(d) 反过来，证明只要关于中位数的集中不等式成立，那么均值的集中不等式也成立，对应的参数为 $c_1=2c_3, c_2=\dfrac{c_4}{4}$。

证明

(a)
$$\mathrm{Var}(X)=\int_0^\infty \mathbb{P}((X-\mu)^2>t)\mathrm{d}t\leqslant c_1\int_0^\infty e^{-c_2t}\mathrm{d}t=\frac{c_1}{c_2}.$$

(b) Rademacher 随机变量显然满足要求。

(c) 记 $\Delta=|\mu_X-m_X|$，引入待定参数 $\alpha>0$，对 $t>\alpha\Delta$，有
$$\begin{aligned}\mathbb{P}(|X-m_X|\geqslant t)&=\mathbb{P}\left(|X-m_X|\geqslant\frac{t}{\alpha}+\left(1-\frac{1}{\alpha}\right)t\right)\\&\leqslant\mathbb{P}\left(|X-m_X|\geqslant\Delta+\left(1-\frac{1}{\alpha}\right)t\right)\\&\leqslant\mathbb{P}\left(|X-\mu_X|\geqslant\left(1-\frac{1}{\alpha}\right)t\right)\leqslant c_1e^{-c_2(1-\alpha^{-1})^2t^2}.\end{aligned}$$
对 $0\leqslant t\leqslant\alpha\Delta$，再引入待定参数 $\beta>1$，我们希望下式成立
$$\beta c_1e^{-c_2(1-\alpha^{-1})^2t^2}\geqslant\beta c_1e^{-c_2(\alpha-1)^2\Delta^2}\geqslant1\geqslant\mathbb{P}(|X-m_X|\geqslant t).$$
因此，需要估计 Δ 的上界。由
$$\begin{aligned}\mathbb{P}(X\geqslant\mathbb{E}[X]+t)\leqslant c_1e^{-c_2t^2},\\\mathbb{P}(X\leqslant\mathbb{E}[X]-t)\leqslant c_1e^{-c_2t^2},\end{aligned}$$
如果 μ_X+t 或 μ_X-t 为中位数，则必有 $c_1e^{-c_2t^2}\geqslant\dfrac{1}{2}$，从而知 $\Delta\leqslant\sqrt{\dfrac{1}{c_2}\log(2c_1)}$。为使下式成立
$$\beta c_1e^{-c_2(\alpha-1)^2\Delta^2}\geqslant\beta c_1e^{-(\alpha-1)^2\log 2c_1}=\beta c_1\left(\frac{1}{2c_1}\right)^{(\alpha-1)^2}\geqslant1.$$
取 $\alpha=2$ 和 $\beta=2$ 即可。进而可得结论对 $c_3=2c_1, c_4=\dfrac{c_2}{4}$ 成立。

(d) ???

□

例 5.6 (习题 2.15：集中度和核密度估计)

设 $\{X_i\}_{i=1}^n$ 是密度函数为 f 的独立同分布随机变量序列，密度函数 f 的一个标准估计是核密度估计
$$\widehat{f}_n(x):=\frac{1}{nh}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right),$$
其中 $K:\mathbb{R}\to[0,\infty)$ 是满足 $\displaystyle\int_{-\infty}^\infty K(t)\mathrm{d}t=1$ 的核函数，$h>0$ 为窗宽参数。考虑用 L_1

范数 $\|\widehat{f}_n - f\|_1 := \int_{-\infty}^{\infty} |\widehat{f}_n(t) - f(t)|\mathrm{d}t$ 来评估统计量 $\widehat{f}_n$，证明：

$$\mathbb{P}\left(\|\widehat{f}_n - f\|_1 \geqslant \mathbb{E}\left[\|\widehat{f}_n - f\|_1\right] + \delta\right) \leqslant e^{-\frac{n\delta^2}{8}}.$$

证明 记 $g(x_1, \dots, x_n) = \|\hat{f}_n - f\|_1$，则 g 满足参数为 $L_k = \dfrac{2}{n}$ 的有界差性质，由推论 2.9 可得更优的上界为 $e^{-\frac{1}{2}nt^2}$。 □

第 3 章

例 5.7 (习题 3.2：链式法则和 Kullback-Leibler 散度)

给定两个 n 元分布 $\mathbb{Q}$ 和 $\mathbb{P}$，证明 Kullback-Leibler 散度可以被分解为

$$D(\mathbb{Q} \,\|\, \mathbb{P}) = D(\mathbb{Q}_1 \,\|\, \mathbb{P}_1) + \sum_{j=2}^{n} \mathbb{E}_{\mathbb{Q}_1^{j-1}}\left[D(\mathbb{Q}_j(\cdot \mid X_1^{j-1}) \,\|\, \mathbb{P}_j(\cdot \mid X_1^{j-1}))\right],$$

其中 $\mathbb{Q}_j(\cdot \mid X_1^{j-1})$ 为在 $\mathbb{Q}$ 下给定 $(X_1, \cdots, X_{j-1})$ 时 X_j 的条件分布，$\mathbb{P}_j(\cdot \mid X_1^{j-1})$ 类似。

证明 由定义得，

$$\begin{aligned}
D(\mathbb{Q} \,\|\, \mathbb{P}) &= \int \log \frac{q(x_1^n)}{p(x_1^n)} q(x_1^n)\nu(dx_1^n) = \int \log \frac{q_n(x_n|x_1^{n-1}) \cdots q_2(x_2|x_1)q_1(x_1)}{p_n(x_n|x_1^{n-1}) \cdots p_2(x_2|x_1)p_1(x_1)} q(x_1^n)\nu(dx_1^n) \\
&= \int \log \frac{q_1(x_1)}{p_1(x_1)} q(x_1)\nu(dx_1) + \sum_{j=2}^{n} \int \log \frac{q_j(x_j|x_1^{j-1})}{p_j(x_j|x_1^{j-1})} q(x_1^n)\nu(dx_1^n) \\
&= D(\mathbb{Q}_1 \,\|\, \mathbb{P}_1) + \sum_{j=2}^{n} \int \left[\int \log \frac{q_j(x_j|x_1^{j-1})}{p_j(x_j|x_1^{j-1})} q_j(x_j|x_1^{j-1})\nu(dx_j)\right] q(x_1^{j-1})\nu(dx_1^{j-1}) \\
&= D(\mathbb{Q}_1 \,\|\, \mathbb{P}_1) + \sum_{j=2}^{n} \mathbb{E}_{\mathbb{Q}_1^{j-1}}\left[D(\mathbb{Q}_j(\cdot|X_1^{j-1}) \,\|\, \mathbb{P}_j(\cdot|X_1^{j-1}))\right].
\end{aligned}$$

□

例 5.8 (习题 3.7：有界变量的熵)

设 $X \in [a, b]$, a.s. 是均值为 0 的随机变量，证明 $\mathbb{H}\left(e^{\lambda X}\right) \leqslant \dfrac{\lambda^2\sigma^2}{2}\varphi_X(\lambda)$，其中 $\sigma = \dfrac{b-a}{2}$。(提示：可以利用习题 3.3 的结果，熵有变分表示 $\mathbb{H}\left(e^{\lambda X}\right) = \inf_{t \in \mathbb{R}} \mathbb{E}\left[\psi(\lambda(X-t))e^{\lambda X}\right]$，其中 $\psi(u) := e^{-u} - 1 + u$。)

证明 由 $\psi(u) \leqslant \dfrac{u^2}{2}, \forall u > 0$，在熵的变分表示中令 $t = a$ 得，

$$\mathbb{H}(e^{\lambda X}) \leqslant \psi(\lambda(b-a))\varphi_X(\lambda) \leqslant \frac{1}{2}\lambda^2(b-a)^2\varphi_X(\lambda),$$

该结果和结论相差一个系数。 □

例 5.9 (习题 3.9：熵的另一种变分表示)

证明熵的下列变分表达形式

$$\mathbb{H}\left(e^{\lambda f(X)}\right)=\sup_{g}\left\{\mathbb{E}\left[g(X)e^{\lambda f(X)}\right]:\mathbb{E}\left[e^{g(X)}\right]\leqslant 1\right\},$$

其中上确界在所有可测函数上取，并给出一个函数 g 使得上确界被取到。

证明 设函数 g 满足 $\mathbb{E}\left[e^{g(X)}\right]\leqslant 1$，记 $\mathbb{E}_g$ 为在密度函数 $\dfrac{e^{g(x)}f_X(x)}{\mathbb{E}\left[e^{g(X)}\right]}$ 下求期望，则有

$$\begin{aligned}
&\mathbb{H}(e^{\lambda f(X)})-\mathbb{E}[g(X)e^{\lambda f(X)}]\\
&=\mathbb{E}[(\lambda f(X)-g(X))e^{\lambda f(X)}]-\mathbb{E}[e^{\lambda f(X)}]\log\mathbb{E}[e^{\lambda f(X)}]\\
&=\mathbb{E}[e^{g(X)}]\left\{\mathbb{E}_g[(\lambda f(X)-g(X))e^{\lambda f(X)-g(X)}]-\mathbb{E}_g[e^{\lambda f(X)-g(X)}]\log\left(\mathbb{E}[e^{g(X)}]\mathbb{E}_g[e^{\lambda f(X)-g(X)}]\right)\right\}\\
&\geqslant\mathbb{E}[e^{g(X)}]\left\{\mathbb{E}_g[(\lambda f(X)-g(X))e^{\lambda f(X)-g(X)}]-\mathbb{E}_g[e^{\lambda f(X)-g(X)}]\log\mathbb{E}_g[e^{\lambda f(X)-g(X)}]\right\}\\
&=\mathbb{E}[e^{g(X)}]\cdot\mathbb{H}_g(e^{\lambda f(X)-g(X)})\geqslant 0
\end{aligned}$$

由上述证明过程知，取 $g(x)=\lambda f(x)-\log\mathbb{E}[e^{\lambda f(X)}]$ 时等号成立，故结论得证。 □

例 5.10 (习题 3.13：全变差和 Wasserstein 距离)

考虑基于 Hamming 度量的 Wasserstein 距离，即 $W_\rho(\mathbb{P},\mathbb{Q})=\inf_{\mathbb{M}}\mathbb{M}[X\neq Y]$，其中下确界在所有耦合 $\mathbb{M}$ 上取。证明：

$$\inf_{\mathbb{M}}\mathbb{M}[X\neq Y]=\|\mathbb{P}-\mathbb{Q}\|_{\mathrm{TV}}=\sup_{A\subseteq\mathcal{X}}|\mathbb{P}(A)-\mathbb{Q}(A)|,$$

其中上确界在 $\mathcal{X}$ 的所有可测子集 A 上取。

证明 思路：分别证明 $\|\mathbb{P}-\mathbb{Q}\|_{\mathrm{TV}}=\sup_{f:\mathcal{X}\mapsto[0,1]}\int f(\mathrm{d}\mathbb{P}-\mathrm{d}\mathbb{Q})$ 和 $W_\rho(\mathbb{P},\mathbb{Q})=\sup_{f:\mathcal{X}\mapsto[0,1]}\int f(\mathrm{d}\mathbb{P}-\mathrm{d}\mathbb{Q})$，从而得到 $\|\mathbb{P}-\mathbb{Q}\|_{\mathrm{TV}}=W_\rho(\mathbb{P},\mathbb{Q})$。 □

第 4 章

附录 A 常用分布表

分布名称	概率密度	期望	方差	特征函数
退化分布	$\begin{pmatrix} c \\ 1 \end{pmatrix}$	c	0	e^{ict}
Bernoulli 分布	$\begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix}$	p	pq	$q+pe^{it}$
二项分布 $B(n,p)$	$\binom{n}{k}p^k q^{n-k}$, $k=0,1,\cdots,n$	np	npq	$(q+e^{it})^n$
几何分布 $G(p)$	$q^{k-1}p$, $k=1,2,\cdots$	$\frac{1}{p}$	$\frac{q}{p^2}$	$\frac{pe^{it}}{1-qe^{it}}$
Pascal 分布 (负二项分布)	$\binom{k-1}{r-1}p^r q^{k-r}$, $k=r,r+1,\cdots$	$\frac{r}{p}$	$\frac{rq}{p^2}$	$\left(\frac{pe^{it}}{1-qe^{it}}\right)^r$
超几何分布	$\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$, $k=0,1,\cdots,n$	$\frac{nM}{N}$	$\frac{nM}{N}\left(1-\frac{M}{N}\right)\frac{N-n}{N-1}$	
Poisson 分布 $P(\lambda)$	$\frac{\lambda^k}{k!}e^{-\lambda}$, $k=0,1,2,\cdots$	λ	λ	$e^{\lambda(e^{it}-1)}$
均匀分布 $U(a,b)$	$\frac{1}{b-a}$, $a<x<b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{itb}-e^{ita}}{it(b-a)}$
正态分布 $N(a,\sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}$, $-\infty<x<+\infty$	a	σ^2	$e^{iat-\frac{1}{2}\sigma^2t^2}$

分布名称	概率密度	期望	方差	特征函数
指数分布	$\lambda e^{-\lambda x}$, $x>0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\left(1-\frac{it}{\lambda}\right)^{-1}$
Γ分布 $\Gamma(\lambda, r)$	$\frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}$, $x>0$	$\frac{r}{\lambda}$	$\frac{r}{\lambda^2}$	$\left(1-\frac{it}{\lambda}\right)^{-r}$
χ^2分布 $\chi^2(n)$	$\frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}$, $x>0$	n	$2n$	$(1-2it)^{-\frac{n}{2}}$
Cauchy 分布 $C(\lambda,\mu)$	$\frac{1}{\pi}\frac{\lambda}{\lambda^2+(x-\mu)^2}$, $-\infty<x<+\infty$	不存在	不存在	$e^{i\mu t-\lambda\lvert t\rvert}$
Rayleigh 分布	$\frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}}$, $x>0$	$\sqrt{\frac{\pi}{2}}\sigma$	$\left(2-\frac{\pi}{2}\right)\sigma^2$	
对数正态分布	$\frac{1}{\sigma x\sqrt{2\pi}} e^{-\frac{(\ln x-a)^2}{2\sigma^2}}$, $x>0$	$e^{a+\frac{\sigma^2}{2}}$	$e^{2a+\sigma^2}(e^{\sigma^2}-1)$	
Weibull 分布	$\alpha\lambda x^{\alpha-1} e^{\lambda x^\alpha}$, $x>0$	$\Gamma(\frac{1}{\alpha}+1)\lambda^{-\frac{1}{\alpha}}$	$\lambda^{-\frac{2}{\alpha}}[\Gamma\left(\frac{2}{\alpha}+1\right)$ $-\left(\Gamma(\frac{1}{\alpha}+1)\right)^2]$	
Laplace 分布	$\frac{1}{2\lambda} e^{-\frac{\lvert x-\mu\rvert}{\lambda}}$, $-\infty<x<+\infty$	μ	$2\lambda^2$	$\frac{e^{i\mu t}}{1+\lambda^2 t^2}$
t 分布 $t(n)$	$\frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\,\Gamma(\frac{n}{2})}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$, $-\infty<x<+\infty$	0 $(n>1)$	$\frac{n}{n-2}$ $(n>2)$	
F 分布 $F(m,n)$	$\frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} m^{\frac{m}{2}} n^{\frac{n}{2}} x^{\frac{m}{2}-1}\cdot$ $(n+mx)^{-\frac{m+n}{2}}, x>0$	$\frac{n}{n-2}$ $(n>2)$	$\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$ $(n>4)$	
贝塔分布	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$, $0<x<1$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	

索　　引

参考文献

[1] Wainwright M J. High-dimensional Statistics: A Non-Asymptotic Viewpoint: vol. 48. Cambridge university press, 2019.

[2] 王成, 刘卫东. 高维统计学: 非渐进视角. 机械工业出版社, 2023.