高維統計學

凌凌伍

2024 年 10 月 2 日

目録

1 測度集中

> A random variable that depends (in a 'smooth'way) on the influence of many independent variables (but not too much on any of them) is essentially constant.
>
> —Michel Talagrand (1996)

驯服随机!

在大尺度上, 无界随机变量的概率分布函数通常有着纤细、绵长的尾部, 这意味着集中现象:

例如正态分布的 3σ 原则告诉我们, "几乎所有"的值都在平均值正负三个标准差的范围内 (若 $X \sim \mathcal{N}(\mu, \sigma^2)$, 则 $\mathbb{P}(|X-\mu| \leq 3\sigma) \approx 0.9973$).

经典的结果是 (广义)Markov 不等式和 Chebyshev 不等式. 设函数 $f\colon \mathbb{R}\to\mathbb{R}^*$ 单调增, 对任意 $t>0$, 注意到

$$\mathbb{E}f(X)\geq \mathbb{E}\left[f(X);X\geq t\right]\geq \mathbb{E}\left[f(t)\mathbb{I}_{\{X\geq t\}}\right]=f(t)\cdot\mathbb{P}[X\geq t].$$

于是我们有随机变量的 *Markov* 不等式:

$$\mathbb{P}[X\geq t]\leq\frac{\mathbb{E}f(X)}{f(t)}. \tag{1}$$

特别地, 取 $f(u)=u^2$, $X=|Y-\mathbb{E}Y|$, 有 *Chebyshev* 不等式:

$$\mathbb{P}[|Y-\mathbb{E}Y|\geq t]\leq\frac{\operatorname{Var}Y}{t^2}.$$

更一般的, 考虑多个独立随机变量的函数 $f(X_1,\dots,X_n)$ 的集中不等式

1.1 Chernoff 方法

可以看到, Markov 不等式 (1) 中尾部概率由 f 的增长速度所控制——这意味着选取增长速度最快的函数, 可以得到更有效的尾部概率不等式. 自然地, 我们考虑指数函数.

若随机变量 X 在 0 的某个邻域 I 内有中心矩母函数, 即在 $\lambda\in I$ 上有 $\varphi_X(\lambda):=\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}]<\infty$, 那么 $\lambda\in I^*:=I\cap[0,\infty)$ 时, 取 $f(u)=e^{\lambda u}$ 可以得到下述的 *Chernoff* 不等式:

$$\mathbb{P}[X-\mathbb{E}X\geq t]\leq\frac{\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}]}{e^{\lambda t}},\ \forall\lambda\in I^*.$$

通过选取最优的 λ, 我们可以得到 *Chernoff* 界

$$\mathbb{P}[X\geq\mathbb{E}X+t]\leq\exp\left[\inf_{\lambda\in I^*}\left\{\log\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}]-\lambda t\right\}\right].$$

1.1.1 示例 *(Gauss* 隨機變量的上偏差不等式*).* 考虑最经典的 *Gauss* 随机变量 $X\sim N(\mu,\sigma^2)$, 其矩母函数

$$\mathbb{E}[e^{\lambda X}]=e^{\frac{\sigma^2\lambda^2}{2}+\mu\lambda} \tag{2}$$

对于 $\lambda\in\mathbb{R}$ 总是存在. 通过简单的求导可以看到最优的 $\lambda^*=\frac{t}{\sigma^2}$, 于是有

$$\mathbb{P}[X\geq\mu+t]\leq e^{-\frac{t^2}{2\sigma^2}},\ \forall t\geq 0. \tag{3}$$

1.2 次高斯随機變量

我们将 Gauss 随机变量作为"模版"来研究其他随机变量: 如果某个随机变量的中心矩母函数能被中心 Gauss 随机变量的矩母函数所控制, 利用 Chernoff 方法, 它的尾概率也会被中心 Gauss 随机变量的尾概率控制.

1.2.1 定義 (次高斯隨機變量). 称期望为 μ 的随机变量 X 为次高斯的, 如果存在 $\sigma > 0$, 使得

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\sigma^2\lambda^2/2},\ \forall \lambda \in \mathbb{R}.$$

称 σ 为 X 的次高斯参数.

可以看到, 参数为 σ 的次高斯随机变量 X 总是满足上偏差不等式 (3). 此外, 由于 $-X$ 也是次高斯随机变量, 可以得到下偏差不等式 $\mathbb{P}[X \leq \mu - t] \leq e^{-\frac{t^2}{2\sigma^2}}$ 对任意 $t \geq 0$ 成立, 于是次高斯随机变量满足集中不等式

$$\mathbb{P}[|X-\mu| \geq t] \leq 2e^{-\frac{t^2}{2\sigma^2}},\ \forall t > 0. \tag{4}$$

直观上, 一个有界随机变量没有无限的尾部, 因此具有集中性质. 事实上, 若 $X \in [a,b]$ a.e., 那么它是参数为 $b-a$ 的次高斯随机变量.

1.2.2 命題 (Hoeffding 界). 设 $\{X_i\}_{i=1}^n$ 为均值为 μ_i, 参数为 σ_i 的独立次高斯随机变量, 我们有

$$\mathbb{P}\left[\sum_{i=1}^n (X_i - \mu_i) \geq t\right] \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right),\ \forall t \geq 0. \tag{5}$$

1.2.3 定理 (次高斯隨機變量定義的等價性). 对任意均值为 0 的随机变量 X, 下述几条命题等价:

(I) 存在常数 $\sigma \geq 0$ 使得

$$\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\lambda^2\sigma^2}{2}},\ \forall \lambda \in \mathbb{R};$$

(II) 存在常数 $c \geq 0$ 和 $Z \sim \mathcal{N}(0,\tau^2)$ 使得

$$\mathbb{P}[|X| \geq s] \leq c\mathbb{P}[|Z| \geq s],\ \forall s \geq 0;$$

(III) 存在常数 $\theta \geq 0$ 使得

$$\mathbb{E}[X^{2k}] \leq \frac{(2k)!}{2^k k!}\theta^{2k}, \forall k = 1,2,\ldots;$$

(IV) 存在常数 $\sigma \geq 0$ 使得

$$\mathbb{E}\left[\exp\left(\frac{\lambda X^2}{2\sigma^2}\right)\right] \leq \frac{1}{\sqrt{1-\lambda}},\ \forall \lambda \in [0,1).$$

1.3 次指数随機變量

很多时候随机变量的中心矩母函数只会在 0 的某个邻域内存在, 相应地, 我们将次高斯随机变量的条件放宽如下.

1.3.1 定義 (次指數隨機變量). 称期望为 μ 的随机变量 X 为次指数的, 如果存在非负参数对 (ν,α) 使得

$$\mathbb{E}[e^{\lambda(X-\mu)}] \le e^{\frac{\nu^2\lambda^2}{2}},\ \forall|\lambda| < \frac{1}{\alpha}.$$

如果记 $+\infty = \frac{1}{0}$, 那么参数为 σ 的次高斯随机变量是 $(\sigma,0)$-次指数的——参数 α 衡量了次指数随机变量与次高斯随机变量相差"多大". 和次高斯随机变量类似, 我们利用 Chernoff 方法可以得到它的尾部不等式

1.3.2 定理 (次指數隨機變量的上偏差不等式). 设 X 是参数为 (ν,α) 的次指数随机变量, 那么

$$\mathbb{P}[X \ge \mathbb{E}X + t] \le \begin{cases} e^{-\frac{t^2}{2\nu^2}}, & 0 \le t \le \frac{\nu^2}{\alpha}, \\ e^{-\frac{t}{2\alpha}}, & t > \frac{\nu^2}{\alpha}. \end{cases}$$

1.4 鞅方法

之前我们讨论了独立随机变量和 $f(X_1,\dots,X_n) = \sum_i X_i$ 的一些尾概率界. 对于更一般的函数 f, 经典的求尾概率界方法是鞅的分解, 我们先回顾一些概念.

给定概率空间 $(\Omega,\mathcal{F}_0,\mathbb{P})$, 子 σ-域 $\mathcal{F} \subset \mathcal{F}_0$, 随机变量 $X \in \mathcal{F}_0$ 可积. 称 Y 为 X 关于 $\mathcal{F}$ 的条件期望, 如果

$$(1)\ Y \in \mathcal{F}; \quad (2)\ \text{对任意 } A \in \mathcal{F},\ \mathbb{E}(Y;A) = \mathbb{E}(X;A).$$

可以证明这样的的 Y 存在唯一 (a.s.), 且 $E|Y| < \infty$, 记做 $\mathbb{E}(X|\mathcal{F})$. 条件期望具有许多性质, 这里我们主要使用以下几个:

(i) 特别地, 如果 $X \in \mathcal{F}$, 则 $\mathbb{E}(X|\mathcal{F}) = X$ a.s.;

(ii) **(全期望公式)** $\mathbb{E}(\mathbb{E}(X|\mathcal{F})) = \mathbb{E}X$; (取 $A = \Omega \in \mathcal{F}$ 即可)

(iii) **(Jensen 不等式)** 若 φ 为凸函数且 $\mathbb{E}X, \mathbb{E}\varphi(X) < \infty$, 则 $\mathbb{E}(\varphi(X)|\mathcal{F}) \ge \varphi(\mathbb{E}(X|\mathcal{F}))$;

(iv) **(塔性质)** 若 $\mathcal{F}_1 \subset \mathcal{F}_2$, 则 $\mathbb{E}(\mathbb{E}(X|\mathcal{F}_1)|\mathcal{F}_2) = \mathbb{E}(X|\mathcal{F}_1) = \mathbb{E}(\mathbb{E}(X|\mathcal{F}_2)|\mathcal{F}_1)$.

随机变量序列 $\{X_k\}$ 是适应于 $\{\mathcal{F}_k\}$ 的鞅, 如果满足

$$(1)\ \mathbb{E}|X_k|<\infty;\quad (2)\ X_k\in\mathcal{F}_k;\quad (3)\ \mathbb{E}(X_{k+1}|\mathcal{F}_k)=X_k.$$

如果我们记 $D_k:=X_k-X_{k-1}$, 容易验证 $\{D_k\}$ 期望为 0, 并且也是适应于 $\{\mathcal{F}_k\}$ 的鞅, 我们称其为鞅差.

设 $\{X_k\}_{k=1}^n$ 为一列独立随机变量, 记 $\boldsymbol{X}=(X_1,\dots,X_n)$, $\boldsymbol{X}_1^k=(X_1,\dots,X_k)$. 对于函数 $f\colon\mathbb{R}^n\to\mathbb{R}$, 考虑随机变量序列 $Y_0=\mathbb{E}f(\boldsymbol{X})$, $Y_k=\mathbb{E}[f(\boldsymbol{X})|\boldsymbol{X}_1^k]$, $k=1,\dots,n$. 由条件期望的性质易见 $Y_n=f(\boldsymbol{X})$. 这样的 $\{Y_k\}_{k=1}^n$ 构成了关于 $\{\boldsymbol{X}_1^k\}_{k=1}^n$ 鞅:

- 由 Jensen 不等式和重期望公式,
$$\mathbb{E}|Y_k|=\mathbb{E}\left[\left|\mathbb{E}[f(\boldsymbol{X})|\boldsymbol{X}_1^k]\right|\right]\le\mathbb{E}\left[\mathbb{E}[|f(\boldsymbol{X})||\boldsymbol{X}_1^k]\right]=\mathbb{E}|f(\boldsymbol{X})|<\infty$$
- $\mathbb{E}[Y_{k+1}|\boldsymbol{X}_1^k]=\mathbb{E}\left[\mathbb{E}[f(\boldsymbol{X})|\boldsymbol{X}_1^{k+1}]\middle|\boldsymbol{X}_1^k\right]=\mathbb{E}[f(\boldsymbol{X})|X_1^k]=Y_k.$

从而 $f(\boldsymbol{X})$ 和 $\mathbb{E}f(\boldsymbol{X})$ 的偏差可以表示为鞅差分解

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

我们先来证明一个一般的鞅差序列的 Bernstein 型不等式界.

1.4.1 定理. *设鞅差序列 $\{(D_k,\mathcal{F}_k)\}_{k=1}^\infty$ 满足次指数条件*

$$\mathbb{E}\left[e^{\lambda D_k}|\mathcal{F}_{k-1}\right]\le e^{\frac{\lambda^2\nu_k^2}{2}}\ \ a.e.\ ,\ \forall|\lambda|<\frac{1}{\alpha_k},$$

那么

(1) 和式 $\sum_k D_k$ 是参数为 $\left(\sqrt{\sum_k\nu_k^2},\alpha_*\right)$ 的次指数随机变量, 其中 $\alpha_*:=\max_k\alpha_k$;

(2) 和式 $\sum_k D_k$ 满足集中不等式

$$\mathbb{P}\left[\left|\sum_{k=1}^n D_k\right|\ge t\right]\le\begin{cases}2\exp\left(-\frac{t^2}{2\sum_k\nu_k^2}\right), & 0\le t\le\alpha_*^{-1}\sum_k\nu_k^2,\\ 2\exp\left(-\frac{t}{2\alpha_*}\right), & t>\alpha_*^{-1}\sum_k\nu_k^2.\end{cases}$$

證明. 我们首先使用控制鞅差和的标准方法, 对于 $|\lambda|<\alpha_*^{-1}$,

$$\begin{aligned}\mathbb{E}\left[e^{\lambda\sum_k D_k}\right]&=\mathbb{E}\left[\mathbb{E}\left[e^{\lambda\sum_k D_k}\middle|\mathcal{F}_{n-1}\right]\right]=\mathbb{E}\left[e^{\lambda\sum_{k=1}^{n-1}D_k}\mathbb{E}[e^{\lambda D_n}|\mathcal{F}_{n-1}]\right]\\&\le\mathbb{E}\left[e^{\lambda\sum_{k=1}^{n-1}D_k}\right]e^{\frac{\lambda^2\nu_n^2}{2}}\le\cdots\le e^{\lambda^2\sum_k\nu^2/2}\end{aligned}$$

于是 $\sum_k D_k$ 是次指数随机变量, 再沿用 Chernoff 方法, 可以得到集中不等式. □

1.5 熵方法

设随机变量 $X \sim \mathbb{P}$, 给定凸函数 $\phi\colon \mathbb{R} \to \mathbb{R}$, 若 X 和 $\phi(X)$ 的期望存在, 则概率分布空间上的泛函

$$\mathcal{H}_\phi(X) := \mathbb{E}\phi(X) - \phi(\mathbb{E}X) \geq 0,$$

称为 X 的 ϕ 熵. 这样的泛函衡量了随机变量随机性的大小:

- 若 $X = C$ a.e., 那么 $\mathcal{H}_\phi(X) = 0$;
- 特别地, 取 $\phi(u) = u^2$ 时, 此时熵 $\mathcal{H}_\phi(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2$ 就是方差, 根据 Chebyshev 不等式, 我们可以利用熵来控制尾部概率;
- 此外, 取 $\phi(u) = -\log u$, 随机变量 $Z = e^{\lambda X}$ 时,

  $$\mathcal{H}_\phi(Z) = -\lambda\mathbb{E}X + \log\mathbb{E}[e^{\lambda X}] = \log\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}],$$

  这样的熵对应的是中心化的矩母函数, 可以利用这样的熵来计算尾部概率的 Chernoff 界.

之后, 我们总是考虑一个具体的凸函数 $\phi\colon [0,\infty) \to \mathbb{R}$:

$$\phi(u) = \begin{cases} u\log u, & u > 0; \\ 0, & u = 0. \end{cases}$$

对于非负随机变量 $Z \geq 0$, 它的 ϕ 熵为

$$\mathcal{H}(Z) = \mathbb{E}[Z\log Z] - \mathbb{E}[Z]\log\mathbb{E}[Z], \tag{6}$$

这里我们略去下标 ϕ. 对于随机变量 $Z = e^{\lambda X}$, 不难看出此时熵可以由矩母函数 $\varphi_X(\lambda) = \mathbb{E}[e^{\lambda X}]$ 来表示:

$$\mathcal{H}(e^{\lambda X}) = \lambda\varphi_X'(\lambda) - \varphi_X(\lambda)\log\varphi_X(\lambda). \tag{7}$$

1.5.1 示例 (Gauss 随機變量的熵). 对于 $X \sim \mathcal{N}(\mu,\sigma^2)$, 其矩母函数为 $\varphi_X(\lambda) = e^{\lambda^2\sigma^2/2+\lambda\mu}$, 于是

$$\mathcal{H}(e^{\lambda X}) = (\lambda^2\sigma^2 + \lambda\mu)\varphi_X(\lambda) - \left(\frac{\lambda^2\sigma^2}{2} + \lambda\mu\right)\varphi_X(\lambda) = \frac{\lambda^2\sigma^2}{2}\cdot\varphi_X(\lambda). \tag{8}$$

1.5.1 将次高斯、次指数随機變量作爲比較對象

和次高斯随机变量的想法类似, 如果 $e^{\lambda X}$ 的熵能被 Gauss 随机变量的熵所控制, 相应的矩母函数、尾部概率也会被控制.

1.5.2 定理 *(Herbst 方法).* 若对任意的 $\lambda \in I$ (这里 I 取 $[0,\infty)$ 或者 $\mathbb{R}$) 总有熵 $\mathcal{H}(e^{\lambda X}) \leq \frac{\lambda^2\sigma^2}{2} \cdot \varphi_X(\lambda)$, 那么

$$\log \mathbb{E}\left[e^{\lambda(X-\mathbb{E}X)}\right] \leq \frac{\lambda^2\sigma^2}{2},\ \forall \lambda \in I.$$

證明. 对于 $\lambda \in I\backslash\{0\}$, 令 $G(\lambda) := \frac{\log \varphi_X(\lambda)}{\lambda}$. 结合 (7), 条件 $\mathcal{H}(e^{\lambda X}) \leq \frac{\lambda^2\sigma^2}{2} \cdot \varphi_X(\lambda)$ 意味着

$$G'(\lambda) \leq \frac{\sigma^2}{2},\ \forall \lambda \in I\backslash\{0\}.$$

当 $\lambda > 0$ 时, 对任意的 $0 < \lambda_0 < \lambda$, 在区间 $[\lambda_0, \lambda]$ 上积分有 $G(\lambda) - G(\lambda_0) \leq \frac{\sigma^2(\lambda-\lambda_0)}{2}$. 再令 $\lambda_0 \to 0^+$ 有

$$\log \mathbb{E}\left[e^{\lambda(X-\mathbb{E}X)}\right] = \lambda(G(\lambda) - \mathbb{E}X) \leq \frac{\lambda^2\sigma^2}{2},\ \forall \lambda \geq 0.$$

其中

$$G(0) = \lim_{\lambda\to 0} G(\lambda) = \lim_{\lambda\to 0} \frac{\varphi'_X(\lambda)}{\varphi_X(\lambda)} = \mathbb{E}X.$$

类似的, 我们可以证明目标不等式在 $\lambda \leq 0$ 时成立. □

1.5.3 推論. 若熵 $\mathcal{H}(e^{\lambda X}) \leq \frac{\lambda^2\sigma^2}{2} \cdot \varphi_X(\lambda)$ 在 $I = [0,\infty)$ 上成立时, X 满足次高斯随机变量的上偏差不等式. 进一步地, 若 $I = \mathbb{R}$, X 满足集中不等式 (4).

自然地, 我们可以将上述方法由次高斯随机变量推广至次指数随机变量.

1.5.4 命題 *(Bernstein 熵的界).* 若存在正常数 b, σ 使得

$$\mathcal{H}(e^{\lambda X}) \leq \lambda^2\left[b\varphi'_X(\lambda) + \varphi_X(\lambda)(\sigma^2 - b\mathbb{E}X)\right],\ \forall \lambda \in [0, 1/b),$$

那么 X 满足上界

$$\log \mathbb{E}\left[e^{\lambda(X-\mathbb{E}X)}\right] \leq \sigma^2\lambda^2(1-b\lambda)^{-1},\ \forall \lambda \in [0, 1/b).$$

證明. 类似地, 命题中的条件意味着

$$G'(\lambda) \leq \sigma^2 - b\mathbb{E}X + b \cdot \frac{\varphi'_X(\lambda)}{\varphi_X(\lambda)}.$$

在任意的区间 $[\lambda_0, \lambda] \subseteq (0, 1/b)$ 上积分有

$$G(\lambda) - G(\lambda_0) \leq (\sigma^2 - b\mathbb{E}X)(\lambda - \lambda_0) + b(\log \varphi_X(\lambda) - \log \varphi_X(\lambda_0)).$$

再令 $\lambda_0 \to 0^+$ 有 $G(\lambda) - \mathbb{E}X \leq \lambda\sigma^2 + b\lambda(G(\lambda) - \mathbb{E}X)$, 于是

$$\log \mathbb{E}\left[e^{\lambda(X-\mathbb{E}X)}\right] = \lambda(G(\lambda) - \mathbb{E}X) \leq \frac{\lambda^2\sigma^2}{1 - b\lambda}, \ \forall \lambda \in [0, 1/b).$$

□

1.5.5 推論. 利用 *Chernoff* 方法, 命题 *1.5.4* 中的熵条件意味着 X 满足上偏差不等式

$$\mathbb{P}[X \geq \mathbb{E}X + t] \leq \exp\left(-\frac{t^2}{4\sigma^2 + 2bt}\right), \ \forall t \geq 0.$$

1.5.2 熵的張量化

多变量函数的熵可以被适当的单变量的熵的和控制上界.

1.5.6 引理 (熵的張量化). 设 $f\colon \mathbb{R}^n \to \mathbb{R}$, $\{X_i\}_{i=1}^n$ 为独立随机变量, 那么

$$\mathcal{H}\left(e^{\lambda f(X_1, \cdots, X_n)}\right) \leq \mathbb{E}\left[\sum_{k=1}^n \mathcal{H}\left(e^{\lambda f_k(X_k)} \middle| X_k\right)\right], \ \forall \lambda > 0.$$

證明. □

下面我们说明, 熵方法

称函数 $f\colon \mathbb{R}^n \to \mathbb{R}$ 为可分凸函数, 如果对任意指标 $k \in \{1, \ldots, n\}$, 给定向量 $(x_1, \ldots, x_{k-1}, x_{k+1}, \ldots x_n) \in \mathbb{R}^{n-1}$, 单变量函数

$$y_k \mapsto f(x_1, \ldots x_{k-1}, y_k, x_{k+1}, \ldots x_n)$$

总是凸函数. 凸函数总是可分凸的,

1.5.7 定理. 令 $\{X_i\}_{i=1}^n$ 为区间 $[a, b]$ 上的独立随机变量, $f\colon \mathbb{R}^n \to \mathbb{R}$ 为可分凸函数且关于 2 范数为 *L-Lipschitz* 的, 那么对任意 $t \geq 0$, 成立

$$\mathbb{P}[f(X) \geq \mathbb{E}f(X) + t] \leq \exp\left(-\frac{t^2}{4L^2(b-a)^2}\right).$$

1.6 集中的幾何觀點

1.6.1 經典等周不等式

经典的等周不等式断言, 欧氏空间 $(\mathbb{R}^n, \rho)$ 中相同体积的子集中, 球的表面积最小. 一种等价表述是, 使得给定体积的子集的 (一致) ϵ-扩张

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

的体积 (作为 ϵ 的函数) 最小的集合 A 一定是球体[1]. 这种表述避免了表面积的概念, 并且可以推广至任意度量空间 $(\mathcal{X}, \rho)$ 上.

Minkowski 将空间的向量加法这一代数结构同凸体的体积这一几何结构联系在一起, 提出了 *Minkowski 和*的概念: 对于 $\mathbb{R}^n$ 中的凸体 C 和 D, 它们的 Minkowski 和定义为

$$\lambda C + (1-\lambda)D := \{\lambda c + (1-\lambda)d \colon c \in C,\ d \in D\},$$

并证明了混合体积的 *Brunn-Minkowski 不等式*

$$[\mathrm{vol}(\lambda C + (1-\lambda)D)]^{\frac{1}{n}} \geq \lambda[\mathrm{vol}(C)]^{\frac{1}{n}} + (1-\lambda)[\mathrm{vol}(D)]^{\frac{1}{n}},\ \forall \lambda \in [0,1].$$

从这一定理出发, 很容易证明经典等周不等式: 对任意 $A \subseteq \mathbb{R}^n$ 满足 $\mathrm{vol}(A) = \mathrm{vol}(\mathbb{B}^n)$,

$$\begin{aligned}[\mathrm{vol}(A^\epsilon)]^{\frac{1}{n}} &= [\mathrm{vol}(A + \epsilon\mathbb{B}^n)]^{\frac{1}{n}} = (1+\epsilon)\left[\mathrm{vol}\left(\frac{1}{1+\epsilon}A + \frac{\epsilon}{1+\epsilon}\mathbb{B}^n\right)\right]^{\frac{1}{n}} \\ &\geq [\mathrm{vol}(A)]^{\frac{1}{n}} + \epsilon[\mathrm{vol}(\mathbb{B}^n)]^{\frac{1}{n}} = (1+\epsilon)[\mathrm{vol}(\mathbb{B}^n)]^{\frac{1}{n}} = [\mathrm{vol}((\mathbb{B}^n)^\epsilon)]^{\frac{1}{n}}.\end{aligned}$$

1.6.2 Lipschitz 函數的集中性

赋予 $(\mathcal{X}, \rho)$ 赋予一个概率测度 $\mathbb{P}$, 我们称三元组 $(\mathcal{X}, \rho, \mathbb{P})$ 为度量测度空间. 考虑随机变量 $X \sim \mathbb{P}$, 此时等周不等式表述为, 确定满足 $\mathbb{P}[X \in A] \geq 1/2$、使得测度 $\mathbb{P}[X \in A^\epsilon]$ 最小的集合 $A \subseteq \mathcal{X}$.

我们引入 $(\mathcal{X}, \rho, \mathbb{P})$ 上的集中度函数 $\alpha \colon \mathbb{R}^* \to [0, 1/2]$

$$\alpha_{\mathbb{P}}(\epsilon) := \sup_{A \subseteq \mathcal{X} \colon \mathbb{P}[A] \geq 1/2} \{1 - \mathbb{P}[A^\epsilon]\}.$$

于是等周不等式相当于确定 $\alpha_{\mathbb{P}}$ 的上界.

[1] 直观上, 膨胀得最慢的是球体.

下面的定理说明, 集中度函数可以控制 Lipschitz 函数的尾部. 回忆 $f(X)$ 的中位数是指满足 $\mathbb{P}[f(X) \geq m_f] \geq 1/2, \mathbb{P}[f(X) \leq m_f] \geq 1/2$ 的某个常数 m_f.

1.6.1 定理 (Lévy 不等式). *设 $f\colon \mathcal{X} \to \mathbb{R}$ 关于 ρ 是 L-Lipschitz 连续的函数, $X \sim \mathbb{P}$, 有 $\mathbb{P}[|f(X) - m_f| \geq \epsilon] \leq 2\alpha(\epsilon/L)$. 特别地, 当 f 是 1-Lipschitz 连续函数时, 我们有*

$$\mathbb{P}[|f(X) - m_f| \geq \epsilon] \leq 2\alpha(\epsilon/L).$$

证明. 令 $A := \{x \in \mathcal{X}\colon f(x) \leq m_f\}$, 于是 $\mathbb{P}[A] \geq 1/2$. 由扩张的定义, 对任意 $x \in A^{\epsilon/L}$, 存在 $y \in A$ 使得 $\rho(x, y) < \epsilon/L$. 于是 $f(x) < f(y) + |f(x) - f(y)| < m_f + \varepsilon$, 进一步地, 我们有 $\mathbb{P}[A^{\epsilon/L}] \leq \mathbb{P}[f(X) < m_f + \epsilon]$. 取余集可以得到

$$\mathbb{P}[f(X) \geq m_f + \epsilon] \leq 1 - \mathbb{P}[A^{\epsilon/L}] \leq \alpha_{\mathbb{P}}(\epsilon/L).$$

对 $-f$ 运用相同的方法可以得到下偏差不等式, 结合起来可得集中不等式. □

反过来, Lipschitz 函数的集中不等式也蕴含着等周不等式. 换而言之, 两种对尾部的控制是等价的.

1.6.2 定理. *若存在函数 $\beta\colon \mathbb{R}^* \to [0, 1]$ 使得对任意的 $(\mathcal{X}, \rho)$ 上的 1-Lipschitz 函数都有*

$$\mathbb{P}[f(X) \geq \mathbb{E}f(X) + \epsilon] \leq \beta(\epsilon), \ \forall \epsilon \geq 0,$$

那么 $\alpha_{\mathbb{P}}(\epsilon) \leq \beta(\epsilon/2)$.

证明. 对任意 $\mathcal{X}$ 中满足 $\mathbb{P}[A] \geq 1/2$ 的可测集 A, 构造 $f_A(x) := \rho(x, A) \wedge \epsilon$. 注意到在 A 上有 $f_A = 0$, 在 A 外有 $f_A \leq \epsilon$, 所以 $\mathbb{E}f_A(X) \leq \epsilon(1 - \mathbb{P}[A]) \leq \epsilon/2$. 于是我们有

$$1 - \mathbb{P}[A^\epsilon] = \mathbb{P}[X \in \bar{A^\epsilon}] = \mathbb{P}[f_A(X) \geq \epsilon] \leq \mathbb{P}\left[f_A(X) \geq \mathbb{E}f_A(X) + \frac{\epsilon}{2}\right] \leq \beta\left(\frac{\epsilon}{2}\right),$$

再对满足条件 $\mathbb{P}[A] \geq 1/2$ 的 A 取上确界即可. 其中最后一个不等式是由于 f_A 是一个 1-Lipschitz 函数:

- 若 $x, y \in A^\epsilon$, 则 $|f_A(x) - f_A(y)| = |\rho(x, A) - \rho(y, A)| \leq \rho(x, y)$;
- 若 $x, y \in \bar{A^\epsilon}$, 则 $|f_A(x) - f_A(y)| = |\epsilon - \epsilon| = 0 \leq \rho(x, y)$;
- 若 $x \in A^\epsilon$, $y \in \bar{A^\epsilon}$, 此时 $\rho(x, A) \geq \rho(y, A) - \rho(x, y) \geq \epsilon - \rho(x, y)$, 则 $|f_A(x) - f_A(y)| = \epsilon - \rho(x, A) \leq \epsilon - (\epsilon - \rho(x, y)) = \rho(x, y)$.

□

1.7 Wasserstein 距離和信息不等式

给定 $(\mathcal{X},\rho)$ 上的两个概率分布 $\mathbb{Q}$ 和 $\mathbb{P}$, 它们之间的 *Wasserstein* 距离为

$$W_\rho(\mathbb{Q},\mathbb{P}) := \sup_{\|f\|_{Lip}\le 1} [\mathbb{E}_\mathbb{Q} f - \mathbb{E}_\mathbb{P} f] = \sup_{\|f\|_{Lip}\le 1} \int f(\mathrm{d}\mathbb{Q} - \mathrm{d}\mathbb{P}).$$

可以证明这样的 W_ρ 构成了一个度量: 存在满足 $\|f\|_{Lip}\le 1$ 的 f 使得 $W_\rho(\mathbb{Q},\mathbb{P}) = \int f(\mathrm{d}\mathbb{Q} - \mathrm{d}\mathbb{P})$, 于是 $W_\rho(\mathbb{P},\mathbb{Q}) \ge \int(-f)(\mathrm{d}\mathbb{P} - \mathrm{d}\mathbb{Q}) = W_\rho(\mathbb{Q},\mathbb{P})$. 类似地, 还有 $W_\rho(\mathbb{Q},\mathbb{P}) \ge W_\rho(\mathbb{P},\mathbb{Q})$, 从而二者相等. 我们将其称为 ρ 诱导的 Wasserstein 度量.

Wasserstein 距离

1.7.1 示例 (*Hamming* 度量和全變差距離). 关于 *Hamming* 度量的 *Wasserstein* 距离 $W_{Ham}(\mathbb{Q},\mathbb{P})$ 等价于全变差距离 $\|\mathbb{Q}-\mathbb{P}\|_{TV} := \sup_{A\subseteq\mathcal{X}} |\mathbb{Q}(A) - \mathbb{P}(A)|$.

此时 f 是关于 *Hamming* 度量的 1-*Lipschitz* 连续函数等价于 f 值域为某个区间 $[c, c+1]$, 不失一般性地, 我们假定 $c = 0$. 记 $\mathbb{Q}$ 和 $\mathbb{P}$ 关于 *Lebesgue* 测度 ν 的密度分别为 q, p, 集合 $A = \{x\in\mathcal{X}\colon q(x)\ge p(x)\}$, 于是有

$$W_{Ham}(\mathbb{Q},\mathbb{P}) = \sup_{f\colon \mathcal{X}\to[0,1]} \int_\mathcal{X} f(q-p)\,\mathrm{d}\nu \le \int_A (\mathrm{d}\mathbb{Q} - \mathrm{d}\mathbb{P}) \le \|\mathbb{Q}-\mathbb{P}\|_{TV}.$$

另一方面, 对任意可测集 $B\subseteq\mathcal{X}$, 注意到 $\mathbb{I}_B$ 是 1-*Lipschitz* 连续的, 于是

$$\mathbb{Q}(B) - \mathbb{P}(B) = \int \mathbb{I}_B(\mathrm{d}\mathbb{Q} - \mathrm{d}\mathbb{P}) \le W_{Ham}(\mathbb{Q},\mathbb{P}).$$

于是有 $\|\mathbb{Q}-\mathbb{P}\|_{TV} \le W_{Ham}(\mathbb{Q},\mathbb{P})$, 从而二者等价.

1.7.1 Kantorovich–Rubinstein 對偶

$$\inf_{\mathbb{M}\in\Pi(\mathbb{Q},\mathbb{P})} \mathbb{E}_\mathbb{M}\left[\rho(X,X')\right] = \inf_{\mathbb{M}\in\Pi(\mathbb{Q},\mathbb{P})} \int_{\mathcal{X}\times\mathcal{X}} \rho(x,x')\,\mathrm{d}\mathbb{M}(x,x').$$

1.7.2 信息不等式

给定 $(\mathcal{X},\rho)$ 上的分布 $\mathbb{Q}\ll\mathbb{P}$, 它们之间的 Kullback–Leibler 散度 (相对熵) 定义为

$$D(\mathbb{Q}\|\mathbb{P}) := \mathbb{E}_\mathbb{Q}\left[\log\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right] = \int_\mathcal{X} q(x)\log\frac{q(x)}{p(x)}\nu(\mathrm{d}x), \tag{9}$$

其中 q,p 为 $\mathbb{Q},\mathbb{P}$ 的密度, ν 为 $\mathcal{X}$ 上的 Lebesgue 测度. 它不是一个度量: 不满足对称性、三角不等式.

称 $(\mathcal{X},\rho)$ 上的概率测度 $\mathbb{P}$ 满足参数为 $\gamma>0$ 的ρ-传输成本不等式, 如果对任意的概率测度 $\mathbb{Q}$ 总有

$$W_\rho(\mathbb{Q},\mathbb{P})\le\sqrt{2\gamma D(\mathbb{Q}\|\mathbb{P})}. \tag{10}$$

1.7.2 定理. *若度量测度空间 $(\mathcal{X},\rho,\mathbb{P})$ 中的概率测度满足 ρ-传输成本不等式(10), 那么它的集中度满足*

$$\alpha_{\mathbb{P}}(\epsilon)\le 2\exp\left(-\frac{\epsilon^2}{2\gamma}\right). \tag{11}$$

证明. 考虑任意满足 $\mathbb{P}[A]\ge\frac{1}{2}$ 的集合 A 和 $\epsilon>0$, 只需证明 $B:=\bar{A}^\epsilon$ 的测度总是小于不等式(11)的右侧. 若 $\mathbb{P}[B]=0$, 则不等式显然成立, 下面我们总假设 $\mathbb{P}[B]>0$.

考虑 $\mathbb{P}_A,\mathbb{P}_B$ 为在 A 和 B 上的条件分布, $\mathbb{M}$ 为它们的任意耦合, 于是

$$\begin{aligned}\int_{\mathcal{X}\times\mathcal{X}}\rho(x,x')\,\mathrm{d}\mathbb{M}(x,x')&=\int_{A\times B}\rho(x,x')\,\mathrm{d}\mathbb{M}(x,x')\\&\ge\rho(A,B)\int_{A\times B}\mathrm{d}\mathbb{M}=\rho(A,B)\ge\epsilon.\end{aligned}$$

对所有可能的耦合取下确界, 可得 $W_\rho(A,B)\ge\epsilon$. 再由三角不等式和 ρ-传输成本不等式, 我们有 (这里根号下似乎少了个 2)

$$\begin{aligned}\epsilon&\le W_\rho(\mathbb{P},\mathbb{P}_A)+W_\rho(\mathbb{P},\mathbb{P}_B)\le\sqrt{\gamma D(\mathbb{P}_A\|\mathbb{P})}+\sqrt{\gamma D(\mathbb{P}_B\|\mathbb{P})}\\&\le\sqrt{2\gamma}\left[D(\mathbb{P}_A\|\mathbb{P})+D(\mathbb{P}_B\|\mathbb{P})\right]^{1/2}.\end{aligned}$$

另一方面, $\mathbb{P}_A$ 的密度为 $p_A(x)=\frac{p(x)\mathbb{I}_A(x)}{\mathbb{P}[A]}$, 于是 $D(\mathbb{P}_A\|\mathbb{P})=-\log\mathbb{P}[A]$, $D(\mathbb{P}_B\|\mathbb{P})=-\log\mathbb{P}[B]$, 从而有 $\epsilon^2\le-2\gamma\log(\mathbb{P}[A]\mathbb{P}[B])$, 等价地

$$\mathbb{P}[B]\le(\mathbb{P}[A])^{-1}\exp\left(-\frac{\epsilon^2}{2\gamma}\right)\le 2\exp\left(-\frac{\epsilon^2}{2\gamma}\right).$$

□

1.7.3 非對稱耦合成本

定义

$$\mathcal{C}(\mathbb{Q},\mathbb{P})=\sqrt{\int\left(1-\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right)_+^2\mathrm{d}\mathbb{P}}$$

A 預備知識

A.1 矩的另一種求法

A.1.1 引理. 若非负随机变量 $X \in L^p$, $p > 0$, 则有

$$\mathbb{E}X^p = \int_0^\infty px^{p-1}\mathbb{P}(X > x)\,\mathrm{d}x. \tag{12}$$

特别的, 对于 $X \geq 0$, 有

$$\mathbb{E}X = \int_0^\infty \mathbb{P}(X > x)\,\mathrm{d}x.$$

进一步地, 若 X 取值范围为 $\mathbb{N}$, 则有

$$\mathbb{E}X = \sum_{k=0}^\infty \mathbb{P}(X \geq k).$$

證明.

$$\begin{aligned}\mathbb{E}X^p &= \int_\Omega X^p\,\mathrm{d}\mathbb{P} = \int_\Omega \int_0^Y px^{p-1}\,\mathrm{d}x\,\mathrm{d}\mathbb{P} = \int_\Omega \int_0^\infty px^{p-1}\mathbb{I}_{\{X>x\}}\,\mathrm{d}x\,\mathrm{d}\mathbb{P} \\ &= \int_0^\infty px^{p-1}\int_\Omega \mathbb{I}_{\{X>x\}}\,\mathrm{d}\mathbb{P}\,\mathrm{d}x = \int_0^\infty px^{p-1}\mathbb{P}(X > x)\,\mathrm{d}x.\end{aligned}$$

□

A.1.2 定理 (*Rademacher*).

A.2 方差的表示

A.2.1 引理 (方差的變分表示). 设随机变量 $X \in L^2$, 那么

$$\operatorname{Var} X = \inf_{a\in\mathbb{R}} \mathbb{E}(X-a)^2.$$

證明. 记 $f(a) = \mathbb{E}(X-a)^2 = a^2 - 2\mathbb{E}X \cdot a + \mathbb{E}X^2$ 为二次函数, 不难看出 f 在 $\mathbb{E}X$ 有最小值 $-(\mathbb{E}X)^2 + \mathbb{E}X^2 = \operatorname{Var} X$. □

A.2.2 引理 (獨立復制). 设随机变量 $X \in L^2$, X' 为 X 的独立复制, 那么

$$\operatorname{Var} X = \frac{1}{2}\mathbb{E}(X-X')^2 = \mathbb{E}(X-X')_+^2 = \mathbb{E}(X-X')_-^2.$$

證明. 由独立性, $\mathbb{E}(X-X')^2 = \mathbb{E}X - 2\mathbb{E}X \cdot \mathbb{E}X + \mathbb{E}X^2 = 2\operatorname{Var} X$. 另一方面, $X - X'$ 和 $X' - X$ 有相同的分布, 于是 $\mathbb{E}(X-X')_+^2 = \mathbb{E}(X-X')_-^2$ 且两者之和即 $\mathbb{E}(X-X')^2$. □

A.3 Randon-Nikodym 導數

Randon-Nikodym 导数是定义密度和条件概率的关键.

设 μ, ν 为可测空间 $(\mathcal{X}, \mathcal{A})$ 上的两个概率测度. 称 ν 关于 μ 绝对连续, 如果 μ-零集一定是 ν-零集, 即对于满足 $\mu(A) = 0$ 的 $A \in \mathcal{A}$, 一定有 $\nu(A) = 0$, 记做 $\nu \ll \mu$. 称 μ 和 ν 相互奇异, 如果存在 $A \in \mathcal{A}$ 使得 $\mu(A) = 0$, $\nu(\mathcal{X} \backslash A) = 0$

A.3.1 定理 *(Lebesgue 分解定理).* 设 μ 和 ν 为 $(\mathcal{X}, \mathcal{A})$ 上 σ-有限测度, 那么 ν 可以唯一地分解为关于 μ 绝对连续的部分 ν_a 和与 μ 相互奇异的部分 ν_s.

A.3.2 示例 *(分布的密度).* 随机变量 $X\colon (\mathcal{X}, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}, \mu)$ 的分布是前推测度 $X_*\mathbb{P}(A) := \mathbb{P} \circ X^{-1}(A) = \mathbb{P}[X \in A]$, 它的密度由 *Randon-Nikodym* 导数给出*:*

$$f_X = \frac{\mathrm{d}X_*\mathbb{P}}{\mathrm{d}\mu}.$$

从而

$$\mathbb{P}[X \in A] = \int_{X^{-1}(A)} \mathrm{d}\mathbb{P} = \int_A \mathrm{d}X_*\mathbb{P} = \int_A f_X \,\mathrm{d}\mu = \mathbb{E}[f_X; A]$$

$$\mathbb{E}[f(X); A] = \int_A f \,\mathrm{d}X_*\mathbb{P} = \int_A f(x) f_X(x) \,\mathrm{d}\mu(x)$$

A.4 耦合

耦合是一种应用广泛的概率技术: 比较两个概率测度 $\mathbb{Q}$, $\mathbb{P}$, 我们可以考虑具有边缘分布 $\mathbb{Q}$, $\mathbb{P}$ 的乘积概率空间.

为了比较概率空间 $\mathcal{X}$ 上两个概率测度 $\mathbb{Q}$, $\mathbb{P}$, 我们可以

很多情况下, 构造乘积空间

的耦合, 是指 $\mathcal{X} \times \mathcal{X}$ 上的联合分布 $\mathbb{M}$, 其边缘分布满足

满足第一和第二坐标的边缘分布分别是 $\mathbb{Q}$ 和 $\mathbb{P}$.

显然乘积测度 $\mathbb{Q} \otimes \mathbb{P}$ 是 $(\mathbb{Q}, \mathbb{P})$ 的耦合,

耦合并不唯一, 记为 $\Pi(\mathbb{Q}, \mathbb{P})$.

記號表

X_1^k	$(X_1, X_2 \ldots, X_k)$
$\mathbb{E}_{\mathbb{P}} f$	$\mathbb{E}_{\mathbb{P}}[f(X)]$

參考文獻

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 02 2013.

[Dur19] Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, 5 edition, April 2019.

[Kle20] Achim Klenke. *Probability Theory: A Comprehensive Course*. Springer International Publishing, 3 edition, 2020.

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, September 2018.

[Wai19] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, February 2019.