高維統計學

干燥症

2024 年 10 月 12 日

目录

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 $f(X_1, \ldots, X_n)$ 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 示例 (*Gauss* 随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

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

在 $\lambda \in \mathbb{R}$ 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

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

1.1.1 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 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

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

1.3 定理 (次高斯隨機變量的等價定義). 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

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

直观上, 一个有界随机变量没有无限的尾部, 因此它应当是次高斯随机变量. 事实上, 我们有如下强结论.

1.4 引理 (有界隨機變量的次高斯參數)**.** 若随机变量 $X \in [a, b]$ *a.s.*, 那么它是参数为 $\frac{b-a}{2}$ 的次高斯随机变量.

證明. 定义随机变量 X_λ, 其关于 X 的分布的 Radon–Nikodym 导数为 $\frac{e^{\lambda X}}{\mathbb{E}[e^{\lambda X}]}$. 于是 $X_\lambda \in [a, b]$ a.s., 从而 $\left|X_\lambda - \frac{a+b}{2}\right| \leq \frac{b-a}{2}$ a.s.,

$$\operatorname{Var} X_\lambda = \operatorname{Var}\left(X_\lambda - \frac{a+b}{2}\right) \leq \mathbb{E}\left[X_\lambda - \frac{a+b}{2}\right]^2 \leq \left(\frac{b-a}{2}\right)^2.$$

记记 $\mathbb{E}[X] = \mu$, $\psi(\lambda) := \log \mathbb{E}[e^{\lambda X}]$, 它满足 $\psi(0) = 0$, $\psi'(0) = \frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]}\big|_{\lambda=0} = \mu$, 且对任意 λ,

$$\psi''(\lambda) = \frac{\mathbb{E}[X^2 e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} - \left(\frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]}\right)^2 = \mathbb{E}X_\lambda^2 - (\mathbb{E}X_\lambda)^2 = \operatorname{Var} X_\lambda.$$

于是 $\psi(\lambda)$ 的在原点的 Taylor 展开为 $\psi(\lambda) = \lambda\mu + \frac{\lambda^2}{2}\psi''(\xi) \leq \lambda\mu + \frac{\lambda^2}{2}\left(\frac{b-a}{2}\right)^2$. 从而 $\mathbb{E}[e^{\lambda(X-\mu)}] \leq \exp\left(\frac{\lambda^2}{2}\left(\frac{b-a}{2}\right)^2\right)$, 即 X 是参数为 $\frac{b-a}{2}$ 的次高斯随机变量. □

1.5 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, 我们有上偏差不等式

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

證明. □

1.1.2 次指數隨機變量

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.

1.6 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

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

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

1.7 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

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

1.2 鞅方法

之前我们讨论了独立随机变量和 $f(X_1, \ldots, X_n) = \sum_i X_i$ 的一些尾概率界. 对于更一般的函数 f, 建立尾概率界的经典方法方法是鞅的分解.

设随机向量 $\boldsymbol{X} = (X_1, \ldots, X_n)$ 各分量独立, 函数 $f\colon \mathbb{R}^n \to \mathbb{R}$ 满足 $\mathbb{E}[f(\boldsymbol{X})] < \infty$. 考虑随机变量序列 $Y_0 = \mathbb{E}[f(\boldsymbol{X})]$, $Y_k = \mathbb{E}[f(\boldsymbol{X})|\boldsymbol{X}_1^k]$, $k = 1, \ldots, n$. 由条件期望的性质易见 $Y_n = f(\boldsymbol{X})$ a.s. 且 $\{Y_k\}_{k=1}^n$ 是适应于 $\{\boldsymbol{X}_1^k\}_{k=1}^n$ 的鞅:

- 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.

1.8 引理. 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

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

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

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

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. □

1.9 定理 (Azuma–Hoeffding). 设鞅差序列 $\{(D_k, \mathcal{F}_k)\}_{k=1}^{\infty}$ 中 $D_k \in [a_k, b_k]$ *a.s.*, 我们有集中不等式

$$\mathbb{E}\left[\left|\sum_{k=1}^{n} D_k\right| \geq t\right] \leq 2\exp\left(-\frac{2t^2}{\sum_k (b_k - a_k)^2}\right), \quad \forall t \geq 0.$$

证明. 由引理1.4, 条件随机变量 $e^{\lambda D_k}|\mathcal{F}_{k-1}$ 是参数为 $\frac{b_k - a_k}{2}$ 的次高斯随机变量. 再根据上一证明中控制鞅差和的方法, 不难得到 Hoeffding 型集中不等式. □

称函数 $f\colon \mathbb{R}^n \to \mathbb{R}$ 满足参数为 $(L_1, \ldots, L_n)$ 的有界差不等式, 如果对指标 $k = 1, \ldots, n$, 总有

$$|f(\boldsymbol{x}_1^{k-1}, x_k, \boldsymbol{x}_{k-2}^{n}) - f(\boldsymbol{x}_1^{k-1}, x_k', \boldsymbol{x}_{k-2}^{n})| \leq L_k.$$

可以证明, 满足有界差不等式的函数一定有界, 而有界函数显然满足有界差不等式. 参数的好处只是为了做出更精确的估计罢了.

1.10 推論 (有界差不等式). 设函数 $f\colon \mathbb{R}^n \to \mathbb{R}$ 满足参数为 $(L_1, \ldots, L_n)$ 的有界差不等式, 随机向量 $\boldsymbol{X} = (X_1, \ldots, X_n)$ 各分量独立, 我们有集中不等式

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2\exp\left(-\frac{2t^2}{\sum_k L_k^2}\right), \quad \forall t \geq 0.$$

证明. 定义随机变量

$$A_k := \inf_x \mathbb{E}[f(\boldsymbol{X})|\boldsymbol{X}_1^{k-1}, X_k = x] - \mathbb{E}[f(\boldsymbol{X})|\boldsymbol{X}_1^{k-1}],$$
$$B_k := \sup_x \mathbb{E}[f(\boldsymbol{X})|\boldsymbol{X}_1^{k-1}, X_k = x] - \mathbb{E}[f(\boldsymbol{X})|\boldsymbol{X}_1^{k-1}].$$

于是 $D_k \in [A_k, B_k]$ a.s. 且区间长度

$$B_k - A_k = \sup_{x,y} \mathbb{E}[f(\boldsymbol{X}_1^k, x, \boldsymbol{X}_{k+1}^n) - f(\boldsymbol{X}_1^k, y, \boldsymbol{X}_{k+1}^n)|X_1^{k-1}] \leq L_k.$$

于是由定理1.9可以得到集中不等式 □

1.3 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

$$\mathcal{H}_\Phi(X) := \mathbb{E}[\Phi(X)] - \Phi(\mathbb{E}[X])$$

称为 X 的 Φ 熵. 由 Jensen 不等式易见 $\mathcal{H} \geq 0$. 这样的泛函可以衡量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, 我们可以利用方差熵来控制尾部概率;
- 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, 我们总是考虑非负随机变量 $e^{\lambda X}$ 由凸函数 $\Phi\colon [0, \infty) \to \mathbb{R}$:

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

诱导的熵:

$$\mathcal{H}(e^{\lambda X}) = \lambda \mathbb{E}[X e^{\lambda X}] - \mathbb{E}[e^{\lambda X}] \log \mathbb{E}[e^{\lambda X}] = \lambda \varphi_X'(\lambda) - \varphi_X(\lambda) \log \varphi_X(\lambda), \quad (6)$$

其中 $\varphi_X(\lambda) = \mathbb{E}[e^{\lambda X}]$ 为矩母函数.

1.11 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

$$\mathcal{H}(e^{\lambda X}) = (\lambda^2 \sigma^2 + \lambda \mu) \varphi_X(\lambda) - \left(\frac{\lambda^2 \sigma^2}{2} + \lambda \mu \right) \varphi_X(\lambda) = \frac{\lambda^2 \sigma^2}{2} \cdot \varphi_X(\lambda). \quad (7)$$

1.12 定理 (熵的變分表示).

$$\mathcal{H}(e^{\lambda f(X)}) = \sup_{g\colon \mathbb{E}[e^{g(X)}]\le 1} \left\{\mathbb{E}\left[g(X)e^{\lambda f(X)}\right]\right\} \tag{8}$$

證明. 考虑测度 $\mathbb{E}^g[A] = \mathbb{E}^g[\mathbb{I}_A] := \mathbb{E}[e^{g(X)}; A]$, 此时期望的 Jensen 不等式依然成立. 一方面, 我们有

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

另一方面, 容易验证 $g(x) = \lambda f(x) - \log\mathbb{E}[e^{\lambda f(X)}]$ 使得等式成立. □

使用独立复制的方法可以对单变量函数熵的界做如下估计.

1.13 引理 (單變量函數熵的界). 设独立随机变量 $X, Y \sim \mathbb{E}$, 对任意函数 $g\colon \mathbb{R}\to\mathbb{R}$, 我们有

$$\mathcal{H}(e^{\lambda g(X)}) \le \lambda^2\mathbb{E}[(g(X) - g(Y))^2 e^{\lambda g(X)}; g(X)\ge g(Y)], \quad \forall\lambda > 0. \tag{9}$$

特别地, 当 X 支撑集为 $[a,b]$, g 为凸的 *Lipschitz* 函数时, 我们有

$$\mathcal{H}(e^{\lambda g(X)}) \le \lambda^2(b-a)^2\mathbb{E}[(g'(X))^2 e^{\lambda g(X)}], \quad \forall\lambda > 0.$$

證明. 由 Jensen 不等式, 不难得出

$$\mathcal{H}(e^{\lambda g(X)}) = \mathbb{E}[\lambda g(X)e^{\lambda g(X)}] - \mathbb{E}[e^{\lambda g(X)}]\log\mathbb{E}[e^{\lambda g(Y)}] \le \mathbb{E}[\lambda g(X)e^{\lambda g(X)}] - \mathbb{E}[\lambda g(X)e^{\lambda g(Y)}].$$

再利用 X 和 Y 的对称性,

$$\begin{aligned}\mathcal{H}(e^{\lambda g(X)}) &\le \lambda\mathbb{E}[g(X)(e^{\lambda g(X)} - e^{\lambda g(Y)})] = -\lambda\mathbb{E}[g(Y)(e^{\lambda g(X)} - e^{\lambda g(Y)})]\\ &= \frac{\lambda}{2}\mathbb{E}\left[(g(X) - g(Y))(e^{\lambda g(X)} - e^{\lambda g(Y)})\right]\\ &= \lambda\mathbb{E}\left[(g(X) - g(Y))(e^{\lambda g(X)} - e^{\lambda g(Y)}); g(X)\ge g(Y)\right].\end{aligned}$$

而在 $\{g(X)\ge g(Y)\}$, $\lambda > 0$ 时, 有 $e^{\lambda g(X)} - e^{\lambda g(Y)} \le \lambda(g(X) - g(Y))e^{\lambda g(X)}$, 从而(9)成立. 进一步地, 当 g 为凸的 Lipschitz 函数时, 由 Rademacher 定

理A.8, g' 几乎处处存在. 于是在 $\{g(X) \geq g(Y)\}$, $\lambda > 0$, X 支撑集为 $[a, b]$ 的条件下,

$$(g(X) - g(Y))^2 \leq (g'(X))^2 (X - Y)^2 \leq (b - a)^2 (g'(X))^2 \quad \text{a.s..}$$

□

1.3.1 Herbst 方法

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.

1.14 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*

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

进一步地, 由 Chernoff 方法, $I = [0, \infty)$ 时, X 满足次高斯随机变量的上偏差不等式; $I = \mathbb{R}$ 时, X 满足集中不等式 (4).

证明. 对于 $\lambda \in I \backslash \{0\}$, 令 $G(\lambda) := \frac{\log \varphi_X(\lambda)}{\lambda}$. 结合 (6), 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. □

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.

1.15 命题 (*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

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

进一步地, 由 *Chernoff* 方法, X 满足上偏差不等式

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

证明. 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

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

□

1.3.2 熵的张量化

由于独立随机变量的联合分布是边缘分布的张量积, 我们可以计算每个随机变量对函数的贡献得到随机变量的函数熵的界[1], 这种被称为熵的张量化, 或者说, 熵具有次可加性.

具体地, 设 $f\colon \mathbb{R}^n \to \mathbb{R}$, 随机向量 $\boldsymbol{X} = (X_1, \ldots, X_n)$ 各分量独立, 我们在保持 $\boldsymbol{X}^{\setminus k}$ 不变的同时, 计算关于 X_k 的逐坐标条件熵

$$\mathcal{H}_k(e^{\lambda f(\boldsymbol{X})}) := \mathcal{H}(e^{\lambda f(\boldsymbol{X})} | \boldsymbol{X}^{\setminus k}).$$

不难看出这是 $\boldsymbol{X}^{\setminus k}$ 的函数, 它的期望就是

[1] 最为经典的例子是独立随机变量和的方差熵是随机变量的方差和.

1.16 引理 (熵的张量化). 在上述的假设和记号下,

$$\mathcal{H}(e^{\lambda f(\boldsymbol{X})}) \leq \mathbb{E}\left[\sum_{k=1}^{n} \mathcal{H}_k(e^{\lambda f(\boldsymbol{X})})\right], \ \forall \lambda > 0.$$

证明. 考虑满足 $\mathbb{E}[e^{g(\boldsymbol{X})}] \leq 1$ 的函数 g, 定义函数序列 $\{g^1, \ldots, g^n\}$:

$$g^k(\boldsymbol{X}_k^n) := \log \mathbb{E}[e^{g(\boldsymbol{X})} | \boldsymbol{X}_k^n] - \log \mathbb{E}[e^{g(\boldsymbol{X})} | \boldsymbol{X}_{k+1}^n].$$

于是 $\sum_{k=1}^n g^k(\boldsymbol{X}_k^n) = g(\boldsymbol{X}) - \log \mathbb{E}[e^{g(\boldsymbol{X})}] \geq g(\boldsymbol{X})$, 利用这一分解可得

$$\mathbb{E}[g(\boldsymbol{X}) e^{\lambda f(\boldsymbol{X})}] \leq \sum_{k=1}^{n} \mathbb{E}[g^k(\boldsymbol{X}_k^n) e^{\lambda f(\boldsymbol{X})}] = \sum_{k=1}^{n} \mathbb{E}\left[\mathbb{E}[g^k(\boldsymbol{X}_k^n) e^{\lambda f(\boldsymbol{X})} | \boldsymbol{X}^{\backslash k}]\right].$$

条件随机变量 $g^k(\boldsymbol{X}_k^n) | \boldsymbol{X}^{\backslash k}$ 的期望满足

$$\mathbb{E}\left[g^k(\boldsymbol{X}_k^n) | \boldsymbol{X}^{\backslash k}\right] \geq \log \mathbb{E}\left[\exp(g^k(\boldsymbol{X}_k^n)) | \boldsymbol{X}_{k+1}^n\right] = 0.$$

$\mathbb{E}\left[\exp(g^k(\boldsymbol{X}_k^n)) | \boldsymbol{X}_{k+1}^n\right] = 1.$

$\mathbb{E}\left[g^k(\boldsymbol{X}_k^n) | \boldsymbol{X}^{\backslash k}\right] = 1$

$\boldsymbol{X} = (X_1, \ldots, X_n)$ 为独立随机变量构成的随机向量, 记 $\boldsymbol{X}_k^n = (X_k, \ldots, X_n)$, $\boldsymbol{X}^{\backslash k} = (X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_n)$. 考虑满足 $\mathbb{E}[e^{g(\boldsymbol{X})}] \leq 1$ 的函数 g, 定义函数序列 $\{g^1, \ldots, g^n\}$:

$$g^k(\boldsymbol{X}_k^n) := \log \mathbb{E}[e^{g(\boldsymbol{X})} | \boldsymbol{X}_k^n] - \log \mathbb{E}[e^{g(\boldsymbol{X})} | \boldsymbol{X}_{k+1}^n].$$

说明 $\mathbb{E}\left[g^k(\boldsymbol{X}_k^n) | \boldsymbol{X}^{\backslash k}\right] = 1$

由定理1.12, 对左侧满足条件的 g 取上确界可知引理成立. □

特别地, 熵的张量化在处理可分凸函数时, 可以得到很好的结论. 称函数 $f\colon \mathbb{R}^n \to \mathbb{R}$ 是可分凸的, 如果对任意指标 $k \in \{1, \ldots, n\}$, 给定向量 $\boldsymbol{x}^{\backslash k} := (x_1, \ldots, x_{k-1}, x_{k+1}, \ldots x_n) \in \mathbb{R}^{n-1}$, 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.

1.17 命题. 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

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

证明. 对于 $\lambda > 0$, 由引理1.16、1.13,

$$\begin{aligned}\mathcal{H}(e^{\lambda f(\boldsymbol{X})}) &\leq \mathbb{E}\left[\sum_{k=1}^n \mathcal{H}_k(e^{\lambda f(\boldsymbol{X})})\right] \leq \lambda^2(b-a)^2\mathbb{E}\left[\sum_{k=1}^n \mathbb{E}_k\left[\left(\frac{\partial f(\boldsymbol{X})}{\partial X_k}\right)^2 e^{\lambda f(\boldsymbol{X})}\right]\right] \\ &= \lambda^2(b-a)^2\mathbb{E}\left[\sum_{k=1}^n \left(\frac{\partial f(\boldsymbol{X})}{\partial X_k}\right)^2 e^{\lambda f(\boldsymbol{X})}\right] \leq \lambda^2(b-a)^2L^2\mathbb{E}[e^{\lambda f(\boldsymbol{X})}].\end{aligned}$$

再由 Herbst 方法 (定理1.14) 可见命题成立. □

1.4 幾何觀點

1.4.1 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[2]. 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

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

[2]直观上, 膨胀得最慢的是球体.

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 $\mathrm{vol}(A) = \mathrm{vol}(\mathbb{B}^n)$,

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

1.4.2 Lipschitz 函数的集中性

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.

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.

1.18 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.

1.19 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 $|f_A(x) - f_A(y)| = \epsilon - \rho(x,A) \leq \epsilon - (\epsilon - \rho(x,y)) = \rho(x,y)$.

□

1.5 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

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

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

1.20 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 $A=\{x\in\mathcal{X}\colon q(x)\geq p(x)\}$, 于是有

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

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

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

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

1.5.1 *Kantorovich–Rubinstein* 對偶

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

1.5.2 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

$$D(\mathbb{Q}\|\mathbb{P}) := \mathbb{E}_{\mathbb{Q}}\left[\log\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right] = \int_{\mathcal{X}} q(x)\log\frac{q(x)}{p(x)}\nu(\mathrm{d}x), \tag{10}$$

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

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

$$W_\rho(\mathbb{Q},\mathbb{P}) \leq \sqrt{2\gamma D(\mathbb{Q}\|\mathbb{P})}. \tag{11}$$

1.21 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(11), 那么它的集中度满足

$$\alpha_{\mathbb{P}}(\epsilon) \leq 2\exp\left(-\frac{\epsilon^2}{2\gamma}\right). \tag{12}$$

证明. 考虑任意满足 $\mathbb{P}[A] \geq \frac{1}{2}$ 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(12)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

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

对所有可能的耦合取下确界, 可得 $W_\rho(A,B) \geq \epsilon$. 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)

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

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 $\epsilon^2 \leq -2\gamma\log(\mathbb{P}[A]\mathbb{P}[B])$, 等价地

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

□

1.5.3 非對稱耦合成本

定义

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

2 一致大數定律

设 $\{X_i\}_{i=1}^n$ 是来自分布 F 的 n 个独立同分布样本, F 经典的无偏估计是经验分布函数

$$\hat{F}_n(t) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{(-\infty],t]}(X_i).$$

强大数定律告诉我们 $\hat{F}_n(t) \xrightarrow{a.s.} F(t)$, 这是一种逐点收敛.

2.1 定理 *(Glivenko-Cantelli* 定理*).* 对任意分布 F, 经验分布 $\hat{F}_n$ 是 F 在一致范数下的强相合估计, 即 $\|\hat{F}_n - F\|_\infty := \sup_{t\in\mathbb{R}} |\hat{F}_n(t) - F(t)| \xrightarrow{a.s.} 0$.

在统计背景下, 将 $\hat{F}_n$ 代入 F 的泛函 $\gamma(F)$ 可以得到估计 $\gamma(\hat{F}_n)$, 例如

2.2 示例*.* 给定可积函数 g, 定义期望泛函 $\gamma_g(F) := \int g\,\mathrm{d}F$, 代入估计 $\hat{F}_n$

$$\gamma_g(\hat{F}_n) = \frac{1}{n}\sum_{n=1}^{n} g(X_i)$$

可以作为 $\mathbb{E}g(X)$ 的估计.

若泛函 γ 连续, 估计 $\gamma(\hat{F}_n)$ 的相合性可以很好地被研究: 称泛函 γ 在 F 关于极大范数 $\|\cdot\|_\infty$ 是连续的, 如果对于任意 $\epsilon > 0$, 存在 $\delta > 0$, 使得对任意 $\|G - F\|_\infty < \delta$ 的函数 G, 总有 $|\gamma(G) - \gamma(F)| < \epsilon$.

2.1 函數類的一致大數定律

设 $\mathscr{F}$ 为在区域 $\mathcal{X}$ 上可积的实值函数类, $\{X_i\}_{i=1}^n$ 是来自分布 $\mathbb{P}$ 的 n 个独立同分布样本.

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} := \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} f(X_i) - Ef(X) \right|$$

在函数类 $\mathscr{F}$ 上衡量了样本平均 $\frac{1}{n}\sum_i f(X_i)$ 和总体平均 $\mathbb{E}f(X)$ 间的偏差. 我们称 $\mathcal{F}$ 为 $\mathbb{P}$ 上的一个 *Glivenko-Cantelli* 类, 如果 $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} \xrightarrow{\mathbb{P}} 0$. 经典的 Glivenko-Cantelli 定理可以看作是在示性函数类 $\mathscr{F} = \{\mathbb{I}_{(-\infty,t]} \colon t \in \mathbb{R}\}$ 上的强一致定律.

未知分布 $\mathbb{P}_{\theta^*}$, 其中 $\theta^* \in \Omega$ 未知, $\{\mathbb{P}_\theta \colon \theta \in \Omega\}$ 为概率分布族. 这里的 Ω 可能是 $\mathbb{R}^d$, 对应参数估计问题; 或者是函数类 $\mathscr{G}$, 对应非参数问题.

估计 θ^* 的决策方法总是基于最小化损失函数 $\mathcal{L}_\theta(X)$: 最优的 θ 应当使得总体风险 $R(\theta, \theta^*) := \mathbb{E}_{\mathbb{P}_{\theta^*}} \mathcal{L}_\theta$ 达到最小. 然而在实践中, 我们通常无法获得总体数据, 只能根据有限个样本 $\{X_i\}_{i=1}^n$, 在 Ω 的某个子集 Ω_0 上最小化经验风险得到估计

$$\hat{\theta} = \arg\min_{\theta \in \Omega_0} \hat{R}_n(\theta, \theta^*) = \arg\min_{\theta \in \Omega_0} \frac{1}{n} \sum_{i=1}^n \mathcal{L}_\theta(X_i).$$

控制过度风险

$$\mathbb{E}(\hat{\theta}, \theta^*) := R(\hat{\theta}, \theta^*) - \inf_{\theta \in \Omega_0} R(\theta, \theta^*).$$

为了方便起见, 我们假设存在某个 $\theta_0 \in \Omega_0$ 满足 $R(\theta_0, \theta^*) = \inf_{\theta \in \Omega_0} R(\theta, \theta^*)$. 于是过度风险可以做如下估计

$$\begin{aligned}\mathbb{E}(\hat{\theta}, \theta^*) &= \left[R(\hat{\theta}, \theta^*) - \hat{R}_n(\hat{\theta}, \theta^*)\right] + \left[\hat{R}_n(\hat{\theta}, \theta^*) - \hat{R}_n(\theta_0, \theta^*)\right] + \left[\hat{R}_n(\theta_0, \theta^*) - R(\theta_0, \theta^*)\right] \\ &\leq \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{L}_{\Omega_0}} + \left[\hat{R}_n(\hat{\theta}, \theta^*) - \hat{R}_n(\theta_0, \theta^*)\right] + \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{L}_{\Omega_0}},\end{aligned}$$

其中函数类 $\mathcal{L}_{\Omega_0} := \{\mathcal{L}_\theta(\cdot) \colon \theta \in \Omega_0\}$. 而 $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{L}_{\Omega_0}}$ 具体的

2.2 經驗過程的尾部概率界

设 $(X_1, \ldots, X_n)$ 来自乘积分布 $\mathbb{P} = \otimes_{i=1}^n \mathbb{P}_i$, 其中 $\mathbb{P}_i$ 的支撑集 $\mathcal{X}_i \subseteq \mathcal{X}$. 对于定义域为 $\mathcal{X}$ 函数类 $\mathscr{F}$, 考虑随机变量

$$Z := \sup_{f \in \mathscr{F}} \left\{ \frac{1}{n} \sum_{i=1}^n f(X_i) \right\}.$$

注意这里的 sup 是对每一点 $x \in \mathcal{X}^n$ 取极大值. 若要考虑 $\sup_{f \in \mathscr{F}} \left|\frac{1}{n} \sum_i f(X_i)\right|$, 只需考虑在函数类 $\tilde{\mathscr{F}} := \mathscr{F} \cup (-\mathscr{F})$ 上考虑上确界即可:

$$\sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_i f(X_i) \right| = \sup_{f \in \mathscr{F}} \left\{ \max \left\{ \frac{1}{n} \sum_{i=1}^n f(X_i), -\frac{1}{n} \sum_{i=1}^n f(X_i) \right\} \right\} = \sup_{\tilde{\mathscr{F}}} \left\{ \frac{1}{n} \sum_{i=1}^n f(X_i) \right\}$$

我们把 Hoeffding

2.3 定理 (泛函 Hoeffding 不等式). 若对每个 $f \in \mathscr{F}$, 都有 $f(\mathcal{X}_i) \subseteq [a_{i,f}, b_{i,f}]$, $i = 1, \dots, n$, 那么对任意 $\delta \geq 0$, 成立

$$\mathbb{P}[Z \geq \mathbb{E}[Z] + \delta] \leq \exp\left(-\frac{n\delta^2}{4L^2}\right), \tag{13}$$

其中 $L^2 = \sup_{f\in\mathscr{F}} \left\{\frac{1}{n}\sum_i (b_{i,f} - a_{i,f})^2\right\}$.

證明. 为了简便, 我们使用非重尺度化的 $Z = \sup_{f\in\mathscr{F}} \{\sum_i f(X_i)\}$, 它是 $\boldsymbol{X} = (X_1, \dots, X_n)$ 的泛函.

定义 $Z_j \colon x_j \mapsto Z(X_1, \dots, X_{j-1}, x_j, X_{j+1}, \dots, X_n)$

对于 $\lambda > 0$, 由引理1.16、1.13,

对 $f \in \mathscr{F}$, 定义 $\mathcal{A}(f) := \{(x_1, \dots, x_n) \colon Z = \sum_i f(x_i)\}$ □

2.4 定理 (經驗過程的 Talagrand 集中度). 若可数函数类 $\mathscr{F}$ 被 b 一致控制, 那么对任意 $\delta > 0$, 成立

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

其中 $\Sigma^2 = \sup_{f\in\mathscr{F}} \frac{1}{n} f^2(X_i)$.

2.3 函數類的 Rademacher 復雜度

2.4 Vapnik-Chervonenkis 維數

1

A 預備知識

A.1 概率論

A.1 定理 (*Borel–Cantelli* 引理). 设 $(A_n)_{n\in\mathbb{N}}$ 为事件序列.

1. 若 $\sum_n \mathbb{P}(A_n) < \infty$, 则 $\mathbb{P}(A_n \ i.o.) = 0$.
2. 若 A_n 相互独立且 $\sum_n \mathbb{P}(A_n) = \infty$, 则 $\mathbb{P}(A_n \ i.o.) = 1$.

證明. 1. 由 $\mathbb{P}$ 上半连续、σ-可加性和 Cauchy 收敛准则:

$$\mathbb{P}\left(\lim_{n\to\infty}\bigcup_{m\geq n} A_m\right) = \lim_{n\to\infty}\mathbb{P}\left(\bigcup_{m\geq n} A_m\right) \leq \lim_{n\to\infty}\sum_{m\geq n}\mathbb{P}(A_m) = 0.$$

2. 由 De Morgan 律和 $\mathbb{P}$ 下半连续

$$\mathbb{P}\left(\left(\bigcap_{m=1}^{\infty}\bigcup_{n=m}^{\infty} A_n\right)^c\right) = \mathbb{P}\left(\bigcup_{m=1}^{\infty}\bigcap_{n=m}^{\infty} A_n^c\right) = \lim_{m\to\infty}\mathbb{P}\left(\bigcap_{n=m}^{\infty} A_n^c\right).$$

而对任意 $m \in \mathbb{N}$, 由不等式 $\log(1-x) \leq -x$ 在 $x \in [0,1]$ 成立, 总有

$$\begin{aligned}\mathbb{P}\left(\bigcap_{n=m}^{\infty} A_n^c\right) &= \lim_{N\to\infty}\mathbb{P}\left(\bigcap_{n=m}^{N} A_n^c\right) = \prod_{n=m}^{\infty}(1-\mathbb{P}(A_n)) \\ &= \exp\left(\sum_{n=m}^{\infty}\log(1-\mathbb{P}(A_n))\right) \leq \exp\left(-\sum_{n=m}^{\infty}\mathbb{P}(A_n)\right) = 0.\end{aligned}$$

□

A.1.1 積分變換、Stieltjes 積分

A.2 定理 (積分變換). 设 $f\colon (\mathcal{X},\mathcal{A},\mu) \to (E,\mathcal{E})$ 为可测映射, g 为

$(E, \mathcal{E})$ 上的可测函数, 则

$$\int_{f^{-1}(B)} g \circ f \,\mathrm{d}\mu = \int_B g \,\mathrm{d} f_* \mu.$$

证明. 只需证明对可测示性成立即可. 对于 $g = \mathbb{I}_F$, $F \in \mathcal{E}$, 有

$$\begin{aligned}\int_B \mathbb{I}_F \,\mathrm{d} f_*\mu = f_*\mu(B \cap F) = \mu\left(f^{-1}(B) \cap f^{-1}(F)\right) = \int_{f^{-1}(B)} \mathbb{I}_{f^{-1}(F)} \,\mathrm{d}\mu \\ = \int_{f^{-1}(B)} \mathbb{I}_F(f(x))\mu(\mathrm{d}x) = \int_{f^{-1}(B)} \mathbb{I}_F \circ f \,\mathrm{d}\mu.\end{aligned}$$

□

A.1.2 Randon-Nikodym 導數、密度

Randon-Nikodym 导数是定义密度和条件期望的关键.

A.3 定理 (*Radon-Nikodym* 定理). 设 μ, ν 为可测空间 $(\mathcal{X}, \mathcal{A})$ 上的两个概率测度, ν 关于 μ 绝对连续, 即对于满足 $\mu(A) = 0$ 的 $A \in \mathcal{A}$, 一定有 $\nu(A) = 0$, 记做 $\nu \ll \mu$. 存在 $\mathcal{X}$ 上的非负函数 f, 使得 $\nu(A) = \int_A f \,\mathrm{d}\mu$, 且 f 在 μ-*a.e.* 意义下唯一, 记做 $f = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}$.

A.4 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 $X_*\mathbb{P}(B) := \mathbb{P} \circ X^{-1}(B) = \mathbb{P}[X \in B]$, 它的关于 μ 的密度由 *Randon-Nikodym* 导数给出:

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

从而由积分变换定理*A.2*,

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

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

此外, 由于分布总是 (前推) 测度, 我们可以通过给出关于随机变量分布的密度函数来定义新的随机变量, 例如引理*1.4*的证明.

A.1.3 矩的求法

对于随机变量 X, 若函数 $\varphi_X(\lambda)=\mathbb{E}[e^{\lambda X}]$ 存在, 则称 φ_X 为 X 的矩母函数. 我们可以利用矩母函数来导出 X 的各阶矩:

$$\frac{\mathrm{d}^n}{\mathrm{d}\lambda^n}\varphi_X(\lambda)=\frac{\mathrm{d}^n}{\mathrm{d}\lambda^n}\mathbb{E}\left[1+\lambda X+\frac{\lambda^2}{2!}X^2+\dots\right]=\mathbb{E}X^n+\frac{\lambda}{n+1}\mathbb{E}X^{n+1}+\dots,$$

于是 $\mathbb{E}X^n=\varphi_X^{(n)}(0)$.

并非所有随机变量都具有矩生成函数,

很多时候它可能只在某个 0 的开邻域内存在

中心矩母函数

A.5 引理. 若非负随机变量 $X\in L^p$, $p>0$, 则有

$$\mathbb{E}X^p=\int_0^\infty px^{p-1}\mathbb{P}(X>x)\,\mathrm{d}x. \tag{14}$$

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

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

证明.

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

□

A.1.4 條件期望、鞅、鞅差

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 $\mathbb{E}(X|\mathcal{F})$. 我们可以把 $X|\mathcal{F}$ 看作随机变量, 称为条件随机变量. 在这样的记号下, X 等价于 $X|\{\emptyset, \Omega\}$.

条件期望 $\mathbb{P}(A|\mathcal{F}) = \mathbb{E}[\mathbb{I}_A|\mathcal{F}]$

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 $\mathbb{E}(\varphi(X)|\mathcal{F}) \geq \varphi(\mathbb{E}(X|\mathcal{F}))$;

(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*.

A.1.5 方差的表示

方差的通常计算方式为 $\operatorname{Var} X = \mathbb{E}[X - \mathbb{E}X]^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2$, 这里我们介绍两种其他的表示方式.

A.6 *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.7 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.1.6 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 $\Pi(\mathbb{Q}, \mathbb{P})$.

A.2 凸分析

A.2.1 Rademacher 定理

A.8 定理 (Rademacher). 任意凸的 *Lipschitz* 函数几乎处处有导数

A.2.2 Fenchel 共軛

Fenchel 共轭是 Fourier 变换在凸分析中的对应. 对于实 Hilbert 空间 $\mathcal{X}$ 上的正则函数 $g\colon \mathcal{X} \to (-\infty, +\infty]$, 即 $\operatorname{dom} f := \{x \in \mathcal{X}\colon f(x) \in \mathbb{R} \neq \emptyset$,

其在 $u \in \mathcal{X}$ 的 *Fenchel* 共轭为

$$f^*(u) = \sup_{x \in \mathcal{X}} \{\langle x, u\rangle - f(x)\}. \tag{15}$$

通过定义可以看到 Fenchel 共轭满足 *Fenchel-Young* 不等式

$$f(x) + f^*(u) \geq \langle x, u\rangle. \tag{16}$$

此外, f^* 是凸的、下半连续的, 这是由于它是放射连续函数族 $(\langle x, \cdot\rangle - f(x))_{x \in \mathcal{X}}$ 的上确界. 对偶 $f = f^{**}$ 当且仅当 f 是凸的、下半连续函数

B　定理證明

記號表

I^*	区间 I 的非负部分: $I \cap [0, \infty)$
$\boldsymbol{x}_j^k$, $\boldsymbol{X}_j^k$	$(x_j, x_{j+1} \ldots, x_k)$, $(X_j, X_{j+1} \ldots, X_k)$
$\mathbb{E}_{\mathbb{P}} f$	$\mathbb{E}_{\mathbb{P}}[f(X)]$, 其中 $X \sim \mathbb{P}$
$\mathbb{E}_i$	逐坐标期望
vol	体积
$\Pi(\mathbb{Q}, \mathbb{P})$	分布对 $(\mathbb{Q}, \mathbb{P})$ 的所有可能的耦合

參考文獻

[BL12] H Bauschke and Yves Lucet, *What is a fenchel conjugate*, Notices of the AMS **59** (2012), no. 1, 44–46.

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

[Çin11] Erhan Çinlar, *Probability and stochastics*, Graduate Texts in Mathematics, vol. 261, Springer New York, 2011.

[Tro23] Joel A. Tropp, *Acm 217: Probability in high dimensions*, August 2023.

[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.

[Wil91] David Williams, *Probability with martingales*, Cambridge University Press, February 1991.