高維統計學

干燥症

2024 年 11 月 11 日

目录

記號表

$\lceil x \rceil$	不小于 x 的最小整数
$\lfloor x \rfloor$	不大于 x 的最小整数
$\boldsymbol{x}_j^k$, $\boldsymbol{X}_j^k$	$(x_j, x_{j+1} \ldots, x_k)$, $(X_j, X_{j+1} \ldots, X_k)$
$\boldsymbol{X}^{\setminus k}$, $\boldsymbol{X}^{\setminus k}$	除去第 k 个分量后的向量 $(x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n)$, $(X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_n)$
$\mathbb{P}(f) = \mathbb{E}_{\mathbb{P}}[f]$	$\int f \, \mathrm{d}\mathbb{P} = \mathbb{E}_{X\sim\mathbb{P}}[f(X)]$
$\mathbb{E}_k$	保持其他分量不变，对第 k 个分量求逐坐标期望 $\mathbb{E}_k[f(\boldsymbol{X})] = \mathbb{E}[f(\boldsymbol{X})\|\boldsymbol{X}^{\setminus k}]$
vol	体积
$\Pi(\mathbb{Q}, \mathbb{P})$	分布对 $(\mathbb{Q}, \mathbb{P})$ 的所有可能的耦合

1 测度集中

> 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**: 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

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

对于随机变量 $Y \in L^p$, 取 $X = |Y - \mathbb{E}[Y]|$, $f(u) = u^p$ $(u \geq 0)$, 可以得到基于高阶矩的集中不等式

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

特别地, $p = 2$ 时就是经典的 Chebyshev 不等式.

要进入高维统计, 我们自然地考虑多个随机变量的函数 $f(X_1, \ldots, X_n)$ 的集中不等式. 例如当 f 是求和函数时, 大数定律告诉

这是渐进结果

非渐进视角

1-范数-通常与维数 n 有关

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.

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

$$\mathbb{P}[X - \mathbb{E}X \geq t] \leq \frac{\mathbb{E}[e^{\lambda(X - \mathbb{E}X)}]}{e^{\lambda t}}, \ \forall \lambda \in I \cap [0, \infty).$$

通过选取最优的参数 λ, 我们可以得到 **Chernoff 界**[1]

$$\mathbb{P}[X \geq \mathbb{E}X + t] \leq \exp\left[\inf_{\lambda \in I \cap [0,\infty)} \left\{\log \mathbb{E}[e^{\lambda(X-\mathbb{E}X)}] - \lambda t\right\}\right].$$

于是随机变量的集中不等式可以从其矩生成函数的界得到.

1.1 示例 (*Gauss 随機變量的上偏差不等式*). *Gauss 随机变量 $X \sim N(\mu, \sigma^2)$ 的矩生成函数*

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

在 $\lambda \in \mathbb{R}$ 总是存在. 通过简单的求导可以看到最优参数 $\lambda^* = \frac{t}{\sigma^2}$, 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 次高斯随機變量與 Hoeffding 界

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生成函数能被中心 Gauss 随机变量的矩生成函数所控制, 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 (*次高斯随機變量*). *称期望为 μ 的随机变量 X 为为 σ-次高斯随机的, 如果存在 $\sigma > 0$, 使得*

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

容易看到, 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 注 (*輔助函數的構造*). *定义1.2实际上是说, 次高斯随机变量的累积生成函数满足 $K(\lambda) := \log \mathbb{E}[e^{\lambda X}] = O(\lambda^2)$, 等价地, $K''(\lambda) = O(1)$; 或者说 $G(\lambda) := \lambda^{-1}K(\lambda) = O(\lambda)$, 等价地 $G'(\lambda) = O(1)$. 引理1.7和定理1.20的证明分别利用了这两种等价表述来构造辅助函数.*

类似的方法还可以用来得到次高斯随机变量最大值的上界.

1.4 定理 (*次高斯隨機變量最大值的上界*). *设 $\{X_i\}_{i=1}^n$ 为均值为 0, 参数为 σ 的次高斯随机变量 (对独立性未做要求).*

(1) $\mathbb{E}\left[\max_i X_i\right] \leq \sigma\sqrt{2\log n}, \forall n \geq 1;$

[1] 在最优的 k 下, 基于高阶矩的 Markov 不等式不会比 Chernoff 界中基于矩生成函数获得的界更差. 但在实际应用中, 由于矩生成函数计算的便利性, Chernoff 界仍有着广泛的应用.

(2) $\mathbb{E}\left[\max_i |X_i|\right] \leq \sigma\sqrt{2\log(2n)} \leq 2\sigma\sqrt{\log n}$, $\forall n \geq 2$.

證明. (1) 由 Jensen 不等式, 对任意 $\lambda > 0$,

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

于是 $\mathbb{E}\left[\max_i X_i\right] \leq \inf_{\lambda>0}\left\{\frac{\log n}{\lambda} + \lambda\frac{\sigma^2}{2}\right\} = \sqrt{2\sigma^2\log n}$, $\forall n \geq 1$.

(2) 由于上述结果的建立对独立性未做要求, 我们考虑 $X_{n+i} := -X_i$, $i = 1, \ldots, n$, 有

$$\max_{1\leq i\leq n} |X_i| = \max_{1\leq i\leq 2n} X_i,$$

于是将结果中 n 替换为 $2n$ 可见结果成立. □

1.5 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

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

利用独立性和 Chernoff 方法, 容易看到偏差不等式关于次高斯参数具有可加性:

1.6 命题 *(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*

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

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.

1.7 引理 (有界隨機變量). *若随机变量 $X \in [a,b]$ a.s., 那么它是 $\frac{b-a}{2}$-次高斯的.*

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 $\left|X_\lambda - \frac{a+b}{2}\right| \le \frac{b-a}{2}$ a.s.,

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

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

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

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

1.8 推論. *经典的 Hoeffding 不等式.*

1.1.2 次指數隨機變量與 Bernstein 界

很多时候随机变量的中心矩生成函数只会在 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.

1.9 定義 (次指數隨機變量). *称期望为 μ 的随机变量 X 为 (ν,α)-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

1.10 定理 (次指數隨機變量的上偏差不等式). *设随机变量 X 是 (ν,α)-次指数的, 那么*

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

1.11 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) 存在非负参数 (ν, α) 使得

$$\mathbb{E}[e^{\lambda X}] \le e^{\frac{\lambda^2\sigma^2}{2}},\ \forall |\lambda| < \frac{1}{\alpha}.$$

(II) 存在常数 $c_0 > 0$, 使得对于任意 $|\lambda| \le c_0$, 都有 $\mathbb{E}[e^{\lambda X}] < \infty$.

(III) 存在常数 $c_1, c_2 > 0$, 使得

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

(IV) 量 $\gamma := \sup_{k\ge 2}\left[\frac{\mathbb{E}[X^k]}{k!}\right]^{1/k}$ 是有限的.

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利用鞅差分解, 再逐个做估计.

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.

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

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

那么和式 $\sum_k D_k$ 是 $\left(\sqrt{\sum_k \nu_k^2}, \alpha_*\right)$-次指数随机变量, 其中 $\alpha_* := \max_k \alpha_k$. 再沿用

Chernoff 方法, 可以得到集中不等式

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

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

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

□

1.13 定理 (*Azuma–Hoeffding 不等式*)**.** *若鞅差序列 $\{(D_k, \mathcal{F}_k)\}_{k=1}^{\infty}$ 满足 $D_k \in [a_k, b_k]$ a.s., 则 $\sum_k D_k$ 是 $\frac{1}{2}\sqrt{\sum_k (b_k - a_k)^2}$-次高斯的, 且有集中不等式*

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

證明. 由引理1.7, 条件随机变量 $e^{\lambda D_k}|\mathcal{F}_{k-1}$ 是 $\frac{b_k - a_k}{2}$-次高斯的. 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.

1.14 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, 那么 $f(\boldsymbol{X})$ 是 $\frac{1}{2}\sqrt{\sum_k L_k^2}$-次高斯的, 从而有集中不等式*

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

證明. 定义随机变量

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

于是 $D_k \in [A_k, B_k]$ a.s. 且区间长度

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

于是由定理1.13可以得到集中不等式 □

1.15 示例. *若函数 f 关于 Hamming 距离 $d_H(\boldsymbol{X}, \boldsymbol{Y}) = \sum_k \mathbb{I}_{\{x_k \neq y_k\}}$ 是 L-Lipschitz 的, 那么 f 满足参数为 $(L, \dots, L)$ 的有界差不等式. 于是对于各分量独立的随机向量 $\boldsymbol{X}$, $f(\boldsymbol{X})$ 是 $\frac{\sqrt{n}L}{2}$-次高斯的, 从而有 $\mathbb{P}[|f(\boldsymbol{X}) - \mathbb{E}[f(\boldsymbol{X})]| \ge t] \le 2e^{-\frac{2t^2}{nL^2}}$, $\forall t \ge 0$.*

1.16 示例 (點集的 Rademacher 復雜度). *Rademacher 随机变量 ϵ 是指等概率地取 $\{-1, 1\}$ 的随机变量, Rademacher 随机向量 $\boldsymbol{\epsilon} \subset \mathbb{R}^n$ 的各分量是相互独立的 Rademacher 随机变量. 给定 $\mathbb{R}^n$ 的子集 A, 随机变量*

$$Z(A) := \sup_{\boldsymbol{a} \in A} \left[\sum_{i=1}^n a_i \epsilon_i \right] = \sup_{\boldsymbol{a} \in A} \langle \boldsymbol{a}, \boldsymbol{\epsilon} \rangle$$

在某种意义下度量了集合 A 的大小, 它的期望 $\mathbb{E}_\epsilon[Z(A)] =: \mathcal{R}(A)$ 称为点集 A 的 Rademacher 复杂度.

记 $f(\boldsymbol{\epsilon}) := \sup_{\boldsymbol{a} \in A} \langle \boldsymbol{a}, \boldsymbol{\epsilon} \rangle$, 我们对第 k 个坐标分量进行扰动: 考虑 $\boldsymbol{\epsilon}' \in \{-1, 1\}^n$ 满足 $\epsilon_i' = \epsilon_i$, $i \neq k$, 于是对任意 $\boldsymbol{a} \in A$,

$$\langle \boldsymbol{a}, \boldsymbol{\epsilon} \rangle - f(\epsilon') = \langle \boldsymbol{a}, \boldsymbol{\epsilon} \rangle - \sup_{\boldsymbol{a} \in A} \langle \boldsymbol{a}, \boldsymbol{\epsilon}' \rangle \le \langle \boldsymbol{a}, \epsilon - \epsilon' \rangle = a_k(\epsilon_k - \epsilon_k') \le 2|a_k|,$$

两边同时对 $\boldsymbol{a}$ 在集合 A 上取上确界有 $f(\boldsymbol{\epsilon}) - f(\boldsymbol{\epsilon}') \le 2 \sup_{\boldsymbol{a} \in A} |a_k|$, 交换 $\boldsymbol{\epsilon}$ 和 $\boldsymbol{\epsilon}'$ 的位置可以得到相同的结论.

对于 $k = 1, \dots, n$, 考虑鞅序列 $Y_0 := \mathbb{E}[f(\boldsymbol{\epsilon})]$, $Y_k := \mathbb{E}[f(\boldsymbol{\epsilon})|\boldsymbol{\epsilon}_1^k]$, 鞅差 $D_k := Y_k - Y_{k-1}$, 随机变量

$$A_k := \min_{s \in \{-1,1\}} \mathbb{E}[f(\boldsymbol{\epsilon})|\boldsymbol{\epsilon}_1^{k-1}, \epsilon_k = s] - \mathbb{E}[f(\boldsymbol{\epsilon})|\boldsymbol{\epsilon}_1^{k-1}],$$

$$B_k := \max_{s \in \{-1,1\}} \mathbb{E}[f(\boldsymbol{\epsilon})|\boldsymbol{\epsilon}_1^{k-1}, \epsilon_k = s] - \mathbb{E}[f(\boldsymbol{\epsilon})|\boldsymbol{\epsilon}_1^{k-1}].$$

于是 $D_k \in [A_k, B_k]$ a.s. 且区间长度

$$B_k - A_k = \max_{s,t \in \{-1,1\}} \mathbb{E}[f(\boldsymbol{\epsilon}_1^k, s, \boldsymbol{\epsilon}_{k+1}^n) - f(\boldsymbol{\epsilon}_1^k, t, \boldsymbol{\epsilon}_{k+1}^n)|\boldsymbol{\epsilon}_1^{k-1}] \le 2 \sup_{\boldsymbol{a} \in A} |a_k|.$$

于是由定理1.13, $f(\boldsymbol{\epsilon})$ 是 $\sqrt{\sum_i \sup_{\boldsymbol{a} \in A} a_i^2}$-次高斯的.

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, 则称概率分布空间上的泛函

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

为 X 的 Φ-熵. 由 Jensen 不等式易见 $\mathcal{H} \ge 0$, 它可以用来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累计生成函数, 可以用来计算尾部概率的 Chernoff 界.

本节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:

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

其中 $M_X(\lambda)=\mathbb{E}[e^{\lambda X}]$ 为矩生成函数.

1.17 示例 *(Gauss 随机变量的熵)*. 随机变量 $X\sim\mathcal{N}(\mu,\sigma^2)$ 的矩生成函数为 $M_X(\lambda)=e^{\lambda^2\sigma^2/2+\lambda\mu}$, 于是

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

1.18 定理 *(熵的变分表示)*.

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

证明. 考虑测度 $\mathbb{P}^g[A]=\mathbb{E}^g[\mathbb{I}_A]:=\mathbb{E}[e^{g(X)};A]$, 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, 容易验证函数 $\tilde{g}(x)=\lambda f(x)-\log\mathbb{E}[e^{\lambda f(X)}]$ 恰好使得等式成立. □

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.

1.19 引理 *(單變量函數熵的界). 设独立随机变量 $X, Y \sim \mathbb{P}$, 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

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

謐明. 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.9, g' 几乎处处存在. 于是在 $\{g(X) \ge g(Y)\}$, $\lambda > 0$, X 支撑集为 $[a,b]$ 的条件下, 我们有估计

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

代入(9)中可见引理成立. □

1.3.1 Herbst 方法

和次高斯随机变量的想法类似, 如果 $e^{\lambda X}$ 的熵能像 Gauss 随机变量的熵那样被控制, 相应的矩生成函数、尾部概率也会被控制.

1.20 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 $\mathcal{H}(e^{\lambda X}) \le \frac{\lambda^2\sigma^2}{2} \cdot \mathbb{E}[e^{\lambda X}]$, 那么*

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

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; 当 $I = \mathbb{R}$ 时, X 是 σ-次高斯的, 从而满足集中不等式 (4).

謐明. 对于 $\lambda \in I \backslash \{0\}$, 令 $G(\lambda) := \frac{\log M_X(\lambda)}{\lambda}$. 结合 (6), 条件 $\mathcal{H}(e^{\lambda X}) \le \frac{\lambda^2\sigma^2}{2} \cdot M_X(\lambda)$ 意味着

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

当 $\lambda > 0$ 时, 对任意的 $0 < \lambda_0 < \lambda$, 在区间 $[\lambda_0, \lambda]$ 上积分有 $G(\lambda) - G(\lambda_0) \le \frac{\sigma^2(\lambda-\lambda_0)}{2}$. 由洛必达法则, $\lim_{\lambda\downarrow 0} G(\lambda) = \lim_{\lambda\downarrow 0} \frac{M_X'(\lambda)}{M_X(\lambda)} = \mathbb{E}X$. 于是令 $\lambda_0 \to 0^+$ 有

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

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

1.21 注. *反之, 若 X 为参数是 $\sigma/2$-的次高斯的, 那么有 $\mathcal{H}(e^{\lambda X}) \le \frac{\lambda^2\sigma^2}{2} \cdot \mathbb{E}[e^{\lambda X}]$.*

證明. 考虑随机变量 $Z := e^{\lambda X}/\mathbb{E}[e^{\lambda X}]$, 那么

$$\mathbb{E}[Z \log Z] = \frac{1}{\mathbb{E}[e^{\lambda X}]}\mathbb{E}\left[e^{\lambda X}\left(\lambda X - \log \mathbb{E}[e^{\lambda X}]\right)\right] = \frac{\mathcal{H}(e^{\lambda X})}{\mathbb{E}[e^{\lambda X}]}.$$

考虑指数加权期望, 我们还有

$$\mathbb{E}[Z\log Z] = \frac{\mathbb{E}[\log Z \cdot e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} = \mathbb{E}_\lambda[\log Z] \le \log \mathbb{E}_\lambda[Z] = \log \mathbb{E}[e^{2\lambda X}] - 2\log \mathbb{E}[e^{\lambda X}]$$

由假设, 其中 $\log \mathbb{E}[e^{2\lambda X}] \le \lambda \mathbb{E}[X] + \frac{\lambda^2\sigma^2}{2}$, 而 $\log \mathbb{E}[e^{\lambda X}] \ge \lambda \mathbb{E}[X]$, 于是

$$\mathcal{H}(e^{\lambda X}) = \mathbb{E}[Z\log Z] \cdot \mathbb{E}[e^{\lambda X}] \le \frac{\lambda^2\sigma^2}{2} \cdot \mathbb{E}[e^{\lambda X}].$$

□

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.

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

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

那么 X 满足上界

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

进一步地, 由 Chernoff 方法, X 满足上偏差不等式

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

證明. 类似地, 命题中的条件意味着

$$G'(\lambda) \le \sigma^2 - b\mathbb{E}X + b \cdot \frac{M_X'(\lambda)}{M_X(\lambda)}.$$

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

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

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

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

□

1.3.2 熵的張量化

鞅方法虽然可以得到随机变量函数的集中不等式, 但是并没有利用充分各分量 X_k 的分布信息, 例如定理1.13中有界随机变量 X_k 的次高斯参数可能远小于 $\frac{b_k - a_k}{2}$. 本节我们利用熵的张量化,

---

由于独立随机变量之和的方差等于方差的和, 利用 Chebyshev 不等式, 我们可以得到

$$\mathbb{P}\left[\left|\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right| \ge t\right] \le \frac{\mathrm{Var}\left(\sum_i X_i\right)}{t^2} = \frac{\sum_i \mathrm{Var}(X_i)}{t^2}.$$

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[2], 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 $\boldsymbol{X} = (X_1, \dots, X_n)$ 各分量独立. 条件随机变量 $e^{\lambda f(\boldsymbol{X})}|\boldsymbol{X}^{\setminus k}$ 可以看作其它分量 $\boldsymbol{X}^{\setminus k}$ 已经给定, 关于 X_k 的函数. 于是我们可以在保持 $\boldsymbol{X}^{\setminus k}$ 不变的同时, 计算关于第 k 个分量的**逐坐标条件熵**

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

1.23 引理 (熵的張量化). *在上述的假设和记号下,*

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

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

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

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

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

[2] 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.

注意到条件随机变量 $e^{g^k(\boldsymbol{X}_k^n)}|\boldsymbol{X}^{\setminus k}$ 的期望满足

$$\mathbb{E}[e^{g^k(\boldsymbol{X}_k^n)}|\boldsymbol{X}^{\setminus k}] = \mathbb{E}\left[\frac{\mathbb{E}[e^{g(\boldsymbol{X})}|\boldsymbol{X}_k^n]}{\mathbb{E}[e^{g(\boldsymbol{X})}|\boldsymbol{X}_{k+1}^n]}\middle|\boldsymbol{X}^{\setminus k}\right] = 1,$$

于是由定理1.18, $\mathbb{E}[g^k(\boldsymbol{X}_k^n)e^{\lambda f(\boldsymbol{X})}|\boldsymbol{X}^{\setminus k}] \le \mathcal{H}_k(e^{\lambda f(\boldsymbol{X})})$. 再次利用定理1.18, 对满足条件的 g 取上确界可知引理成立.

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

$$\mathbb{E}\left[\frac{\mathbb{E}[e^{g(\boldsymbol{X})}|\boldsymbol{X}_k^n]}{\mathbb{E}[e^{g(\boldsymbol{X})}|\boldsymbol{X}_{k+1}^n]}\middle|\boldsymbol{X}^{\setminus k}\right] \le 1$$

□

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 $\boldsymbol{X}^{\setminus k} := (x_1, \ldots, x_{k-1}, x_{k+1}, \ldots x_n) \in \mathbb{R}^{n-1}$, 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.

1.24 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 $t \ge 0$, 成立*

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

證明. 对于 $\lambda > 0$, 由引理1.23、1.19,

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

其中 $\sum_k \left(\frac{\partial f(\boldsymbol{X})}{\partial x_k}\right)^2 = \|\nabla f\|_2^2 \le L^2$ 由 Lipchitz 条件得到. 再由 Herbst 方法 (定理1.20) 可见命题成立. □

1.4 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. 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. 它的经典证明基于数学家 Minkowski 的天才洞见.

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——$\mathbb{R}^n$ 中的凸体 C 和 D 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 $\mathrm{vol}(A) = \mathrm{vol}(\mathbb{B}^n)$,

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

1.4.1 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.

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.

1.25 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

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

證明. 令 $A := \{x \in \mathcal{X} \colon f(x) \le m_f\}$, 于是 $\mathbb{P}[A] \ge 1/2$. 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 $\mathbb{P}[A^{\epsilon/L}] \le \mathbb{P}[f(X) < m_f + \epsilon]$. 取余集可以得到

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

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.

1.26 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*

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

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

證明. 对任意 $\mathcal{X}$ 中满足 $\mathbb{P}[A] \ge 1/2$ 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 $f_A \le \epsilon$, 所以 $\mathbb{E}f_A(X) \le \epsilon(1 - \mathbb{P}[A]) \le \epsilon/2$. 于是我们有

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

再对满足条件 $\mathbb{P}[A] \ge 1/2$ 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 $|f_A(x) - f_A(y)| = |\rho(x, A) - \rho(y, A)| \le \rho(x, y)$;
- 若 $x, y \in \bar{A}^\epsilon$, 则 $|f_A(x) - f_A(y)| = |\epsilon - \epsilon| = 0 \le \rho(x, y)$;
- 若 $x \in A^\epsilon, y \in \bar{A}^\epsilon$, 此时 $\rho(x, A) \ge \rho(y, A) - \rho(x, y) \ge \epsilon - \rho(x, y)$, 则 $|f_A(x) - f_A(y)| = \epsilon - \rho(x, A) \le \epsilon - (\epsilon - \rho(x, y)) = \rho(x, y)$.

□

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

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

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.

1.27 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 $A=\{x\in\mathcal{X} : q(x)\geq p(x)\}$, 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

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

1.5.2 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

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

1.28 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

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

證明. 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)

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

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

1. 示性函数记号的问题 2. Talagrand 不等式 3. 假设类的引入 (0-下水平集...

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

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

经典的 **Glivenko-Cantelli 定理**告诉我们, 经验分布函数 $\hat{F}_n$ 是 F 在一致范数下的强相合估计, 即 $\hat{F}_n$ 和 F 之间的 **Kolmogorov 距离**几乎处处收敛到 0:

$$\|\hat{F}_n - F\|_\infty := \sup_{t\in\mathbb{R}} |\hat{F}_n(t) - F(t)| \overset{a.s.}{\to} 0$$

2.1 經驗過程

设 $\{X_i\}_{i=1}^n$ 是来自分布 $\mathbb{P}$ 的 n 个独立同分布样本, **经验分布**为 $\mathbb{P}_n(A) := \frac{1}{n}\sum_{i=1}^n \mathbb{I}_A(X_i)$, 其中集合 $A \subseteq \mathcal{X}$. 考虑区域 $\mathcal{X}$ 上的 $\mathbb{P}$-可积实值函数类 $\mathscr{F}$, 函数 $f \in \mathscr{F}$ 关于初始测度 $\mathbb{P}$ 和经验测度 $\mathbb{P}_n$ 的积分分别为

$$\mathbb{P}(f) = \int f \,\mathrm{d}\mathbb{P} = \mathbb{E}_{X\sim\mathbb{P}}[f(X)], \quad \mathbb{P}_n(f) = \frac{1}{n}\sum_{i=1}^n f(X_i).$$

于是经验过程 $X_f = \mathbb{P}(f) - \mathbb{P}_n(f)$, $f \in \mathscr{F}$ 衡量了经验期望和总体期望的偏差. 经验过程理论的研究对象是 $\mathbb{P}$ 在函数类 $\mathscr{F}$ 上被 $\mathbb{P}_n$ 一致逼近的性质, 更为具体的, 研究随机量

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

的测度集中现象和随机过程 $\{(\mathbb{P} - \mathbb{P}_n)(f) \colon f \in \mathscr{F}\}$ 的概率极限理论.

第一个问题涉及 $|P_n - P|_F$ 在 F 一致有界的情况下围绕其均值的集中性。问题的关键在于：变量 $|P_n - P|_F$ 在其均值附近有多集中？是否可以得到关于差值 $|P_n - P|_F - \mathbb{E}|P_n - P|F$ 的指数不等式，且达到与经典不等式对 $\sum i = 1^n \xi_i$（其中 ξ_i 是中心化且有界的）同样的精度？还是需要为我们同时考虑无穷多个独立随机变量之和而付出额外的代价？

对此问题的令人惊讶的答案是：在适当定义参数（规模和方差）的情况下，经典的指数不等式在经验过程上依然成立。这是经验过程理论中最重要、最强大的结果之一，称为 Talagrand 不等式。在本节稍后部分，我们将回顾实随机变量的经典指数不等式，为后续的经验过程部分提供背景。

可测函数类的上确界未必可测, 恐有不测之忧

- 如果 $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} \xrightarrow{\mathbb{P}} 0$, 则称函数类 $\mathscr{F}$ 为分布 $\mathbb{P}$ 上的一个 **Glivenko-Cantelli 类**, 或者 $\mathscr{F}$ 满足 **Glivenko-Cantelli 律**;
- 如果 $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} \xrightarrow{a.s.} 0$, 则称函数类 $\mathscr{F}$ 满足**强 Glivenko-Cantelli 律**.

经典的 Glivenko-Cantelli 定理实际上是在示性函数类 $\mathscr{F} = \{\mathbb{I}_{(-\infty,t]} \colon t \in \mathbb{R}\}$ 上的强一致定律. 对更为广义的函数类的研究则始于 Vapnik-Čhervonenkis(1971) 和 Dudley(1978) 的工作, 这在统计中有着十分重要的作用.

量 $\mathbb{E}\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}\right]$ 实际上是经验测度 $\mathbb{P}_n$ 和 $\mathbb{P}$ 之间的 Wasserstein 距离 $W(\mathbb{P}, \mathbb{P}_n)$.

2.1 示例 *(M 估計?[GN15] P110)*. *若分布 $\mathbb{P}_{\theta^*}$ 中的 $\theta^* \in \Theta$ 未知, 其中空间 Θ 可能是 $\mathbb{R}^d$, 对应参数估计问题; 或者是某个函数类 $\mathscr{G}$, 对应非参数问题.*

θ^* 的估计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 $R(\theta, \theta^*) := \mathbb{E}_{\mathbb{P}_{\theta^*}}[\mathcal{L}_\theta(X)]$ 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 Θ 的某个子集 Θ_0 上最小化经验风险得到估计

$$\hat{\theta} = \arg\min_{\theta\in\Theta_0} \hat{R}_n(\theta, \theta^*) = \arg\min_{\theta\in\Theta_0} \frac{1}{n}\sum_{i=1}^n \mathcal{L}_\theta(X_i).$$

我们希望经验风险和总体风险足够接近，即控制过度风险 $\mathbb{E}(\hat{\theta},\theta^*) := R(\hat{\theta},\theta^*) - \inf_{\theta\in\Theta_0} R(\theta,\theta^*)$.

为了方便起见，我们假设存在某个 $\theta_0 \in \Theta_0$ 满足 $R(\theta_0,\theta^*) = \inf_{\theta\in\Theta_0} R(\theta,\theta^*)$. 于是过度风险可以做如下估计

$$\mathbb{E}(\hat{\theta},\theta^*) = \underbrace{\left[R(\hat{\theta},\theta^*) - \hat{R}_n(\hat{\theta},\theta^*)\right]}_{T_1} + \underbrace{\left[\hat{R}_n(\hat{\theta},\theta^*) - \hat{R}_n(\theta_0,\theta^*)\right]}_{T_2} + \underbrace{\left[\hat{R}_n(\theta_0,\theta^*) - R(\theta_0,\theta^*)\right]}_{T_3}.$$

其中 $|T_1| = \left|\frac{1}{n}\sum_i \mathcal{L}_\theta(X_i) - \mathbb{E}_{\mathbb{P}_{\theta^*}}[\mathcal{L}_\theta(X)]\right| \le \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{L}_{\Theta_0}}$，这需要需要考虑损失函数类 $\mathcal{L}_{\Theta_0} := \{\mathcal{L}_\theta(\cdot) \colon \theta\in\Theta_0\}$ 的一致大数定律；而 $\hat{\theta}$ 最小化了经验风险，可以看到 $T_2 \le 0$; T_3 对应的则是控制随机变量 $\frac{1}{n}\sum_i \mathcal{L}_{\theta_0}(X_i)$ 及其期望 $\mathbb{E}_{\mathbb{P}_{\theta_0}}[\mathcal{L}_\theta(X)]$ 之间的偏差，这里 θ_0 是一个未知但非随机的值，因此可以用测度集中的方法来得到. 当然 $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{L}_{\Theta_0}}$ 也是 T_3 的上界，于是过度风险至多是 $2\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{L}_{\Theta_0}}$.

2.2 經驗過程的尾部概率界

设 $(X_1,\dots,X_n)$ 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 $\boldsymbol{X} \in \mathcal{X}^n$ 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:

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

我们把 Hoeffding

2.2 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 $\delta \ge 0$，成立

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

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 $Z = \sup_{f\in\mathscr{F}} \left\{\sum_i f(X_i)\right\}$，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1.23、1.19,

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

2.3 定理（經驗過程的 *Talagrand* 集中度）. 若可数函数类 $\mathscr{F}$ 是 b 一致有界的，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

Peter L.Bartlett 与 Shahar Mendelson （此人是 Empirical Process 的专家）提出了用 Rademacher / Gaussian Complexity 来研究 Risk Bounds 的方法

一致大数定律的的一个重要度量是函数类 $\mathscr{F}$ 的 **Rademacher 复杂度**. 设 $\mathscr{F}$ 为区域 $\mathcal{X}$ 上的函数类, 对于点集 $\boldsymbol{x}_1^n \in \mathcal{X}^n$, 记 $\mathscr{F}(\boldsymbol{x}_1^n) := \{(f(x_1), \ldots, f(x_n))\colon f \in \mathscr{F}\}$. $\mathscr{F}$ 关于点集 $\boldsymbol{x}_1^n$ 的 Rademacher 复杂度为

$$\mathcal{R}_n(\mathscr{F}) := \mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\boldsymbol{a}\in\mathscr{F}(\boldsymbol{x}_1^n)} \langle f(\boldsymbol{x}_1^n), \boldsymbol{\epsilon}\rangle\right] = \mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)\right|\right].$$

进一步地, 函数类 $\mathscr{F}$ 关于经验分布 $\mathbb{P}_n$ 的 Rademacher 复杂度为

$$\mathcal{R}_{\mathbb{P}_n}(\mathscr{F}) := \mathbb{E}_{\boldsymbol{X},\boldsymbol{\epsilon}}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{k=1}^n \epsilon_i f(X_i)\right|\right].$$

这可以看作随机向量 $(f(X_1), \cdots, f(X_n))_{f\in\mathscr{F}}$ 和噪声向量 $\boldsymbol{\epsilon}$ 之间最大相关关系的平均值.

对于一致有界函数类 $\mathscr{F}$, 我们将看到“$\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} \approx \mathcal{R}_{\mathbb{P}_n}(\mathscr{F})$”, 而 Rademacher 复杂度的界是更为容易得到的.

2.4 引理. 若函数类 $\mathscr{F}$ 是 b-一致有界的，那么随机变量 $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}$ 是 $\frac{b}{\sqrt{n}}$-次高斯的.

證明. 引入函数的中心化记号 $\bar{f}(x) := f(x) - \mathbb{E}[f(X)]$, 则 $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}$ 可以简记为 $\sup_{f\in\mathscr{F}} |\frac{1}{n}\sum_i \bar{f}(X_i)|$. 考虑函数 $G(x_1, \ldots, x_n) = \sup_{f\in\mathscr{F}} |\frac{1}{n}\sum_i \bar{f}(x_i)|$, 它与各坐标的顺序置换无关. 于是只需对第一个坐标分量进行扰动, 就可以说明它满足参数为 $(\frac{2b}{n}, \ldots, \frac{2b}{n})$ 的有界差不等式: 设向量 $\boldsymbol{X} = (x_1, \ldots, x_n)$, $\boldsymbol{Y} = (y_1, \ldots, y_n)$ 满足 $x_i = y_i$, $i > 1$, 说明 $|G(\boldsymbol{X}) - G(\boldsymbol{Y})| < \frac{2b}{n}$ 即可. 对任意 $f \in \mathscr{F}$, 由于 $\|f\|_\infty \leq b$,

$$\left|\frac{1}{n}\sum_{i=1}^n \bar{f}(x_i)\right| - \sup_{h\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \bar{h}(y_i)\right| \leq \left|\frac{1}{n}\sum_{i=1}^n \bar{f}(x_i)\right| - \left|\frac{1}{n}\sum_{i=1}^n \bar{f}(y_i)\right| \leq \frac{1}{n}|\bar{f}(x_1) - \bar{f}(y_1)| \leq \frac{2b}{n}.$$

结合推论1.14, 可以看到 $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}$ 是 $\frac{b}{\sqrt{n}}$-次高斯的. □

2.5 定理. 设函数类 $\mathscr{F}$ 是 b-一致有界的, 对任意 $n \in \mathbb{Z}_+$, $\delta \geq 0$, 我们有

$$\mathbb{P}\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} \leq 2\mathcal{R}_{\mathbb{P}_n}(\mathscr{F}) + \delta\right] \geq 1 - \exp\left(-\frac{n\delta^2}{2b^2}\right).$$

于是如果函数类 $\mathscr{F}$ 满足 $\mathcal{R}_{\mathbb{P}_n}(\mathscr{F}) = o(1)$, 就可以得到 $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}$ 以指数速度几乎确定收敛到 0, 即 $\mathscr{F}$ 为 $\mathbb{P}$ 上的 Glivenko-Cantelli 类.

在证明之前, 我们给出一个函数类上确界期望的不等式, 它和 Fatou 不等式或者 Jensen 不等式类似: 对于可积函数类 $\mathscr{G}$, 有 $\mathbb{E}[g(X)] \leq \mathbb{E}\left[\sup_{g\in\mathscr{G}} |g(X)|\right]$, 于是再对左侧取上确界可以得到

$$\sup_{g\in\mathscr{G}} \mathbb{E}[g(X)] \leq \mathbb{E}\left[\sup_{g\in\mathscr{G}} |g(X)|\right]. \tag{14}$$

进一步地, 对于凸的非减函数 Φ, 结合 Jensen 不等式, 我们有

$$\sup_{g\in\mathscr{G}} \Phi(\mathbb{E}[|g(X)|]) \leq \Phi\left(\mathbb{E}\left[\sup_{g\in\mathscr{G}} |g(X)|\right]\right) \leq \mathbb{E}\left[\Phi\left(\sup_{g\in\mathscr{G}} |g(X)|\right)\right] \tag{15}$$

證明. 我们首先证明 $2\mathcal{R}_{\mathbb{P}_n}(\mathscr{F})$ 是 $\mathbb{E}\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}\right]$ 的上界, 这可以使用**对称化技巧**来得到: 设 $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ 与 $\boldsymbol{X}$ 独立同分布, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$ 为独立的 Rademacher 向量, 于是 $\epsilon_i(f(X_i) - f(Y_i))$ 和 $f(X_i) - f(Y_i)$ 有相同的分布:

$$\begin{aligned}
&\mathbb{P}[f(X_i) - f(Y_i) \leq t] = \frac{1}{2}\mathbb{P}[f(X_i) - f(Y_i) \leq t] + \frac{1}{2}\mathbb{P}[f(Y_i) - f(X_i) \leq t] \\
=&\mathbb{P}[\epsilon_i = 1] \cdot \mathbb{P}[\epsilon_i(f(X_i) - f(Y_i)) \leq t | \epsilon_i = 1] + \mathbb{P}[\epsilon_i = -1] \cdot \mathbb{P}[\epsilon_i(f(X_i) - f(Y_i)) \leq t | \epsilon_i = -1] \\
=&\mathbb{P}[\epsilon_i(f(X_i) - f(Y_i)) \leq t].
\end{aligned}$$

结合不等式(14)、三角不等式

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}\right] = \mathbb{E}_{\boldsymbol{X}}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}_{Y_1}[f(Y_1)]\right|\right] \\
=&\mathbb{E}_{\boldsymbol{X}}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n (f(X_i) - \mathbb{E}_{Y_i}[f(Y_i)])\right|\right] = \mathbb{E}_{\boldsymbol{X}}\left[\sup_{f\in\mathscr{F}}\left|\mathbb{E}_{\boldsymbol{Y}}\left[\frac{1}{n}\sum_{i=1}^n (f(X_i) - f(Y_i))\right]\right|\right] \\
\leq&\mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n (f(X_i) - f(Y_i))\right|\right] = \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\epsilon}}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\left(f(X_i) - f(Y_i)\right)\right|\right] \\
\leq&\mathbb{E}_{\boldsymbol{X},\boldsymbol{\epsilon}}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right|\right] + \mathbb{E}_{\boldsymbol{Y},\boldsymbol{\epsilon}}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(Y_i)\right|\right] = 2\mathcal{R}_{\mathbb{P}_n}(\mathscr{F}).
\end{aligned}$$

于是根据 $\frac{b}{\sqrt{n}}$-次高斯随机变量 $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}$ 的上偏差不等式, 对任意 $\delta \geq 0$, 我们有

$$\mathbb{P}\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} \leq 2\mathcal{R}_{\mathbb{P}_n}(\mathscr{F}) + \delta\right] \geq \mathbb{P}\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} \leq \mathbb{E}\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}\right] + \delta\right] \geq 1 - \exp\left(-\frac{nt^2}{2b^2}\right).$$

$\square$

对称化技巧实际上考虑了随机变量 $\|\mathbb{S}_n\|_{\mathscr{F}} := \sup_{f\in\mathscr{F}} \left|\frac{1}{n}\sum_i \epsilon_i f(X_i)\right|$——它的期望就是 Rademacher 复杂度, 我们有更强的结论.

2.6 命题. 对于任意非减的凸函数 $\Phi\colon \mathbb{R}\to\mathbb{R}$, 我们有

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{\epsilon}}\left[\Phi\left(\frac{1}{2}\|\mathbb{S}\|_{\bar{\mathscr{F}}}\right)\right] \le \mathbb{E}_{\boldsymbol{X}}\left[\Phi\left(\|\mathbb{P}_n-\mathbb{P}\|_{\mathscr{F}}\right)\right] \le \mathbb{E}_{\boldsymbol{X},\boldsymbol{\epsilon}}\left[\Phi\left(2\|\mathbb{S}\|_{\mathscr{F}}\right)\right]$$

2.7 注. 特别地, 取 $\Phi(t)=t$ 可以得到

$$\frac{1}{2}\mathbb{E}_{\boldsymbol{X},\boldsymbol{\epsilon}}\left[\|\mathbb{S}\|_{\bar{\mathscr{F}}}\right] \le \mathbb{E}_{\boldsymbol{X}}\left[\|\mathbb{P}_n-\mathbb{P}\|_{\mathscr{F}}\right] \le 2\mathbb{E}_{\boldsymbol{X},\boldsymbol{\epsilon}}\left[\|\mathbb{S}\|_{\mathscr{F}}\right] \tag{16}$$

證明. 右侧不等式可以看作是上一定理的证明的简单推广：结合不等式(15)、三角不等式, 利用 Φ 的凸性, 我们有

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}}\left[\Phi\left(\|\mathbb{P}_n-\mathbb{P}\|_{\mathscr{F}}\right)\right] = \mathbb{E}_{\boldsymbol{X}}\left[\Phi\left(\sup_{f\in\mathscr{F}}\left|\mathbb{E}_{\boldsymbol{Y}}\left[\frac{1}{n}\sum_{i=1}^n (f(X_i)-f(Y_i))\right]\right|\right)\right]\\
\le&\mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}}\left[\Phi\left(\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n (f(X_i)-f(Y_i))\right|\right)\right] = \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\epsilon}}\left[\Phi\left(\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\left(f(X_i)-f(Y_i)\right)\right|\right)\right]\\
\le&\mathbb{E}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\epsilon}}\left[\Phi\left(\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right| + \sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(Y_i)\right|\right)\right]\\
\le&\frac{1}{2}\mathbb{E}_{\boldsymbol{X},\boldsymbol{\epsilon}}\left[\Phi\left(2\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right|\right)\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{Y},\boldsymbol{\epsilon}}\left[\Phi\left(2\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(Y_i)\right|\right)\right]\\
=&\mathbb{E}_{\boldsymbol{X},\boldsymbol{\epsilon}}\left[\Phi\left(2\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right|\right)\right] = \mathbb{E}_{\boldsymbol{X},\boldsymbol{\epsilon}}\left[\Phi\left(2\|\mathbb{S}\|_{\mathscr{F}}\right)\right].
\end{aligned}$$

下面我们证明左侧不等式, 由不等式(15)、三角不等式和 Φ 的非减性、Φ 的凸性, 我们有

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X},\boldsymbol{\epsilon}}\left[\Phi\left(\frac{1}{2}\|\mathbb{S}\|_{\bar{\mathscr{F}}}\right)\right] = \mathbb{E}_{\boldsymbol{X},\boldsymbol{\epsilon}}\left[\Phi\left(\frac{1}{2}\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\left(f(X_i)-\mathbb{E}_{Y_i}[f(Y_i)]\right)\right|\right)\right]\\
\le&\mathbb{E}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\epsilon}}\left[\Phi\left(\frac{1}{2}\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\left(f(X_i)-f(Y_i)\right)\right|\right)\right] = \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}}\left[\Phi\left(\frac{1}{2}\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n (f(X_i)-f(Y_i))\right|\right)\right]\\
\le&\mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}}\left[\Phi\left(\frac{1}{2}\left\{\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n (f(X_i)-\mathbb{E}_{\mathbb{P}}[f]) + \sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n (f(Y_i)-\mathbb{E}_{\mathbb{P}}[f])\right|\right\}\right|\right)\right]\\
\le&\frac{1}{2}\mathbb{E}_{\boldsymbol{X}}\left[\Phi\left(\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n (f(X_i)-\mathbb{E}_{\mathbb{P}}[f])\right|\right)\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{Y}}\left[\Phi\left(\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n (f(Y_i)-\mathbb{E}_{\mathbb{P}}[f])\right|\right)\right]\\
=&\mathbb{E}_{\boldsymbol{X}}\left[\Phi\left(\|\mathbb{P}_n-\mathbb{P}\|_{\mathscr{F}}\right)\right].
\end{aligned}$$

$\square$

2.8 命题. 设函数类 $\mathscr{F}$ 是 b-一致有界的, 对任意 $n \in \mathbb{Z}_+$, $\delta \geq 0$, 我们有

$$\mathbb{P}\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} \geq \frac{1}{2}\mathcal{R}_{\mathbb{P}_n}(\mathscr{F}) - \frac{\sup_{f\in\mathscr{F}}|\mathbb{E}_{\mathbb{P}}[f]|}{2\sqrt{n}} - \delta\right] \geq 1 - \exp\left(-\frac{n\delta^2}{2b^2}\right).$$

于是当函数类 $\mathscr{F}$ 的 *Rademacher* 复杂度 $\mathcal{R}_{\mathbb{P}_n}(\mathscr{F})$ 存在远离 0 的下界时, $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}$ 不可能以概率收敛于零, 即 $\mathscr{F}$ 不可能为 $\mathbb{P}$ 上的 *Glivenko-Cantelli* 类.

證明. 根据三角不等式容易得到估计

$$\|\mathbb{S}_n\|_{\bar{\mathscr{F}}} = \sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i) - \frac{1}{n}\sum_{i=1}^n \epsilon_i \mathbb{E}_{\mathbb{P}}[f]\right| \geq \sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right| - \sup_{f\in\mathscr{F}}|\mathbb{E}_{\mathbb{P}}[f]|\cdot\frac{|\sum_{i=1}^n \epsilon_i|}{n}.$$

再由 Cauchy–Schwarz 不等式可得 $\mathbb{E}\left[\left|\sum_{i=1}^n \epsilon_i\right|\right] \leq \sqrt{\mathbb{E}\left[\left(\sum_{i=1}^n \epsilon_i\right)^2\right]} = \sqrt{\mathbb{E}\left[\sum_{i=1}^n \epsilon_i^2\right]} = \sqrt{n}$. 于是由不等式(16), $\frac{b}{\sqrt{n}}$-次高斯随机变量 $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}$ 的下偏差不等式, 可见

$$\begin{aligned}
&1 - \exp\left(-\frac{n\delta^2}{2b^2}\right) \leq \mathbb{P}\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} \geq \mathbb{E}\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}\right] - \delta\right] \\
\leq &\mathbb{P}\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} \geq \frac{1}{2}\mathbb{E}_{\boldsymbol{X},\boldsymbol{\epsilon}}\left[\|\mathbb{S}\|_{\bar{\mathscr{F}}}\right] - \delta\right] \leq \mathbb{P}\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} \geq \frac{1}{2}\mathcal{R}_{\mathbb{P}_n}(\mathscr{F}) - \frac{\sup_{f\in\mathscr{F}}|\mathbb{E}_{\mathbb{P}}[f]|}{2\sqrt{n}} - \delta\right]
\end{aligned}$$

□

于是 Rademacher 复杂度是 1 的高阶无穷小为满足 Glivenko-Cantelli 性质提供了一个充分必要条件. 下面两节我们寻求控制 Rademacher 复杂度的方法.

2.4 多項式識別函數類

2.9 定義 (多項式識別). 称区域 $\mathcal{X}$ 上的函数类 $\mathscr{F}$ 有阶为 $\nu \geq 1$ 的多项式识别, 如果对任意正整数 n 和点集 $\boldsymbol{x}_1^n = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$, 集合

$$\mathscr{F}(\boldsymbol{x}_1^n) := \{(f(x_1), \ldots, f(x_n)) \colon f \in \mathscr{F}\} \subseteq \mathbb{R}^n$$

的基数 $\operatorname{card}(\mathscr{F}(\boldsymbol{x}_1^n)) \leq (n+1)^\nu$.

很多时候函数类的基数 $\operatorname{card}(\mathscr{F}) = \infty$, 但是

2.10 引理. 若区域 $\mathcal{X}$ 上的函数类 $\mathscr{F}$ 有阶为 $\nu \geq 1$ 的多项式识别, 那么对任意正整数 n, 函数类 $\mathscr{F}$ 关于点集 $\boldsymbol{x}_1^n = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$ 的 *Rademacher* 复杂度

$$\mathcal{R}_n(\mathscr{F}) = \mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)\right|\right] \leq 2D(\boldsymbol{x}_1^n)\sqrt{\frac{\nu\log(n+1)}{n}},$$

其中 $D(\boldsymbol{x}_1^n)=\sup_{f\in\mathscr{F}}\sqrt{\frac{1}{n}\sum_i f(x_i)^2}$ 为向量集合 $\mathscr{F}(\boldsymbol{x}_1^n)/\sqrt{n}$ 的 ℓ^2 半径.

證明. 考虑集合 $A=\left\{\frac{1}{n}(f(x_1),\ldots,f(x_n))\colon f\in\mathscr{F}\cup(-\mathscr{F})\right\}\subseteq\mathbb{R}^n$, 由于 $\mathscr{F}$ 有阶为 ν 的多项式识别, 我们有 $\operatorname{card}(A)\le 2(n+1)^\nu$. 回忆示例1.16, 不等式左侧等价于点集 A 的 Rademacher 复杂度, 即

$$\mathcal{R}(A)=\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\boldsymbol{a}\in A}\langle\boldsymbol{a},\boldsymbol{\epsilon}\rangle\right]=\mathbb{E}_{\boldsymbol{\epsilon}}\left[\max_{\boldsymbol{a}\in A}\langle\boldsymbol{a},\boldsymbol{\epsilon}\rangle\right],$$

其中 $\langle\boldsymbol{a},\boldsymbol{\epsilon}\rangle$ 总是 $D(\boldsymbol{x}_1^n)/\sqrt{n}$-次高斯的, 于是由定理1.4

$$\mathbb{E}_{\boldsymbol{\epsilon}}\left[\max_{\boldsymbol{a}\in A}\langle\boldsymbol{a},\boldsymbol{\epsilon}\rangle\right]\le D(\boldsymbol{x}_1^n)/\sqrt{n}\cdot\sqrt{2\log(\operatorname{card}(A))}\le 2D(\boldsymbol{x}_1^n)\sqrt{\frac{\nu\log(n+1)}{n}}.$$

□

2.11 注. *若函数类 $\mathscr{F}$ 是 b-一致有界的, 那么它关于经验分布 $\mathbb{P}_n$ 的 Rademacher 复杂度*

$$\begin{aligned}\mathcal{R}_{\mathbb{P}_n}(\mathscr{F})&=\mathbb{E}_{\boldsymbol{X}}\left[\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n\epsilon_i f(X_i)\right|\right]\right]=\mathbb{E}_{\boldsymbol{X}}\left[\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n\epsilon_i f(X_i)\right|\,\Big|\boldsymbol{X}\right]\right]\\ &\le\sup_{\boldsymbol{x}}\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n\epsilon_i f(X_i)\right|\,\Big|\boldsymbol{X}=\boldsymbol{x}\right]\le 2\sup_{\boldsymbol{x}}\sup_{f\in\mathscr{F}}\sqrt{\frac{1}{n}\sum_{i=1}^n f(x_i)^2}\cdot\sqrt{\frac{\nu\log(n+1)}{n}}\\ &\le 2b\sqrt{\frac{\nu\log(n+1)}{n}}.\end{aligned}$$

结合定理2.5, 我们可以看到具有多项式识别的有界函数类总是 Glivenko-Cantelli 的. 例如, 经典的 Glivenko-Cantelli 定理考虑的函数类 $\mathscr{F}=\{\mathbb{I}_{(-\infty,t]}\colon t\in\mathbb{R}\}$ 是 1-一致有界的, 我们可以得到 Glivenko-Cantelli 定理的定量版本.

2.12 推論 (Glivenko-Cantelli 定理-定量版本)**.** *对任意 $\delta\ge 0$,*

$$\mathbb{P}\left[\|\hat{F}_n-F\|_\infty\ge 8\sqrt{\frac{\log(n+1)}{n}}+\delta\right]\le\exp\left(-\frac{n\delta^2}{2}\right).$$

于是 $\|\hat{F}_n-F\|_\infty$ 以指数速度几乎确定收敛于 0.

證明. 对于任意样本 $\boldsymbol{x}_1^n$, 考虑次序样本 $x_{(1)}\le x_{(2)}\le\cdots\le x_{(n)}$, 我们有

$$\mathscr{F}(\boldsymbol{x}_1^n)\subseteq\{(0,0,\ldots,0),(1,0,\ldots,0),(1,1,\ldots,0),\ldots,(1,1,\ldots,1)\}.$$

于是 $\operatorname{card}(\mathscr{F}(\boldsymbol{x}_1^n))\le n+1$, 即多项式识别阶数 $\nu=1$, 于是 $\mathcal{R}_{\mathbb{P}_n}(\mathscr{F})\le 2\sqrt{\frac{\log(n+1)}{n}}$. 结合定理2.5可见不等式成立. □

2.5 Vapnik-Čhervonenkis 維數

对于布尔值函数类, 即值域为 $\{0,1\}$ 的函数构成的类, 我们常用 **Vapnik-Chervonenkis 维数** (简称 VC 维数) 来衡量它的复杂度.

hypothesis class/ concept class

例如集合类 $\mathcal{S}$ 的示性函数类 $\mathbb{I}_{\mathcal{S}} := \{\mathbb{I}_S : S \in \mathcal{S}\}$, 为了记号的方便, 我们将集合类 $\mathcal{S}$ 等价于函数类 $\mathbb{I}_{\mathcal{S}}$.

集合 Λ 被函数类**打散**是指无论我们对每个点如何赋予布尔值标签, 都有函数类中的一个函数将它实现. 例如对于点集 $\boldsymbol{x}_1^n = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$, 共有 2^n 种赋予布尔值标签的方式, 换而言之, 集合 $\mathscr{F}(\boldsymbol{x}_1^n)$ 至多有 2^n 个元素, 于是 $\boldsymbol{x}_1^n$ 被打散等价于 $\operatorname{card}(\mathscr{F}(\boldsymbol{x}_1^n)) = 2^n$. 函数类的 VC 维定义为可以被其成员打散的点的最多数目.

2.13 定義 *(VC 維數)***.** *我们称集合 $\Lambda \subseteq \mathcal{X}$ 被 $\mathscr{F}$ 打散, 如果对任意映射 $g\colon \Lambda \to \{0,1\}$, 都存在某个 $f \in \mathscr{F}$ 使得 $f|_\Lambda = g$. 函数类 $\mathscr{F}$ 的 VC 维数为能被 $\mathscr{F}$ 打散的集合的最大基数: 如果*

$$\operatorname{vc}(\mathscr{F}) = \max\{\operatorname{card}(\Lambda) \colon \operatorname{card}(\mathscr{F}(\Lambda)) = 2^{\operatorname{card}(\Lambda)}\} < \infty$$

*则称函数类 $\mathscr{F}$ 为 **VC 类**, 否则, 记 $\operatorname{vc}(\mathscr{F}) = \infty$.*

2.14 示例 *($\mathbb{R}$ 上的區間)***.** *推论2.12的证明本质上考虑了 $\mathbb{R}$ 上左侧半区间类 $\mathcal{S}_{left} := \{(-\infty, t] \colon t \in \mathbb{R}\}$ 的示性函数类, 它的多项式识别的阶为 1. 对于 $x_1 < x_2$, $\mathcal{S}_{left}(x_1, x_2) = \{(0,0), (0,1), (1,1)\}$, 于是 $\operatorname{vc}(\mathcal{S}_{left}) = 1$.*

进一步地, 双侧区间类 $\mathcal{S}_{two} : \{(b,a] \colon a, b \in \mathbb{R}, b < a\}$ 的示性函数类可以打散任意的两点集, 但是对于三个不同的点 $x_1 < x_2 < x_3$, 它不能选出集合 $\{x_1, x_3\}$: 即 $(1,0,1) \notin \mathcal{S}_{two}(\boldsymbol{x}_1^3)$, 于是 $\operatorname{vc}(\mathcal{S}_{two}) = 2$. 此外, $\operatorname{card}(\mathcal{S}_{two}(\boldsymbol{x}_1^n)) \le (n+1)^2$, 于是 $\mathcal{S}_{two}$ 多项式识别的阶为 2.

根据定义, 若 $\operatorname{vc}(\mathscr{F}) < n$, 我们只能得到指数增长的结果 $\operatorname{card}(\mathscr{F}(\boldsymbol{x}_1^n)) \le 2^n - 1$. 但是在上述示例中, 我们看到 VC 维数和多项式识别的阶数似乎存在着一定的联系. 事实上, 利用组合的方法, 我们可以得到如下结论.

2.15 引理 *(Sauer-Shelah)***.** *设 $\mathcal{S}$ 是 VC 类, 对任意点集 $\boldsymbol{x}_1^n$, 其中 $n > \operatorname{vc}(\mathcal{S})$, 我们有*

$$\operatorname{card}(\mathcal{S}(\boldsymbol{x}_1^n)) \le \sum_{i=1}^{\operatorname{vc}(\mathcal{S})} \binom{n}{i} \le (n+1)^{\operatorname{vc}(\mathcal{S})}.$$

證明. □

VC 类在有限次集合运算下保持不变, 这被称为 VC 稳定性.

2.16 命题 (VC 稳定性). 若 $\mathcal{S}$ 和 $\mathcal{T}$ 是 VC 类, 那么下述集合类也是 VC 类:

(1) $\mathcal{S}^c := \{S^c : S \in \mathcal{S}\}$;

(2) $\mathcal{S} \sqcap \mathcal{T} := \{S \cap T : S \in \mathcal{S}, T \in \mathcal{T}\}$;

(3) $\mathcal{S} \sqcup \mathcal{T} := \{S \cup T : S \in \mathcal{S}, T \in \mathcal{T}\}$.

證明. (1) 若点集 $\boldsymbol{x}_1^n$ 可以被 $\mathcal{S}$ 打散, 对任意 $S \in \mathcal{S}$, 由于 S^c 会给 $\boldsymbol{x}_1^n$ 和 S 完全相反的布尔值标签, 于是 $\mathrm{card}(\mathcal{S}^c(\boldsymbol{x}_1^n)) = 2^n = \mathrm{card}(\mathcal{S}(\boldsymbol{x}_1^n))$, 即 $\mathrm{vc}(\mathcal{S}^c) = \mathrm{vc}(\mathcal{S})$.

(2) 注意到对任意的 $S \in \mathcal{S}, T \in \mathcal{T}$, 我们有 $\mathbb{I}_{S\cap T} = \mathbb{I}_S \cdot \mathbb{I}_T$, 结合引理2.15,

$$\mathrm{card}\left(\mathcal{S} \sqcap \mathcal{T}(\boldsymbol{x}_1^n)\right) \le \mathrm{card}(\mathcal{S}(\boldsymbol{x}_1^n)) \cdot \mathrm{card}(\mathcal{T}(\boldsymbol{x}_1^n)) \le (n+1)^{\mathrm{vc}(\mathcal{S})+\mathrm{vc}(\mathcal{T})}.$$

(3) 由 $S \cup T = (S^c \cap T^c)^c$ 可见成立.

□

实值函数类 $\mathscr{F}$ 可以通过取 0-下水平集来定义**相伴集合类** $\mathcal{S}(\mathscr{F}) := \{S_f : f \in \mathscr{F}\}$, 其中 $S_f := \{x \in \mathcal{X} : f(x) \le 0\}$ 称为函数 f 的 0-下水平集. 很多重要的集合类, 例如半平面、椭球体, 都可以用这种方式来表达.

2.17 命题. 设函数类 $\mathscr{G}$ 为 $\mathbb{R}^d$ 上实值函数的线性空间, 其中 $\dim(\mathscr{G}) < \infty$, 那么 $\mathcal{S}(\mathscr{G})$ 的 VC 维数至多为 $\dim(\mathscr{G})$.

證明. 采用反证法, 假设存在点集 $\boldsymbol{x}_1^n = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d$ 可以被 $\mathcal{S}(\mathscr{G})$ 打散, 其中 $n = \dim(\mathscr{G}) + 1$. 由于 $\mathscr{G}$ 构成了线性空间, 集合 $\mathscr{G}(\boldsymbol{x}_1^n) = \{(g(x_1), \ldots, g(x_n)) : g \in \mathscr{G}\}$ 构成了 $\mathbb{R}^n$ 的子空间, 并且它的维数至多为 $\dim(\mathscr{G}) = n - 1 < n$. 因此, 存在某个非零向量 $\boldsymbol{\gamma} \in \mathbb{R}^n$ 满足 $\langle \boldsymbol{\gamma}, g(\boldsymbol{x}_1^n) \rangle = \sum_i \gamma_i g(x_i) = 0, \forall g \in \mathscr{G}$. 不失一般性地, 假定 $\boldsymbol{\gamma}$ 至少存在一个正的分量, 于是

$$\sum_{i:\ \gamma_i \le 0} (-\gamma_i) g(x_i) = \sum_{i:\ \gamma_i > 0} \gamma_i g(x_i), \quad \forall g \in \mathscr{G}.$$

由于 $\mathcal{S}(\mathscr{G})$ 可以将 $\boldsymbol{x}_1^n$ 打散, 于是存在 $g \in \mathscr{G}$ 使得 $S_g \cap \boldsymbol{x}_1^n = \{x_i : \gamma_i \le 0\}$, 此时等式右侧严格正而左侧非正, 推出矛盾. □

2.18 示例 ($\mathbb{R}^d$ 中的半平面). $\mathbb{R}^n$ 中的半平面 $H_{\boldsymbol{a},b} := \{\boldsymbol{x} \in \mathbb{R}^d : \langle \boldsymbol{a}, \boldsymbol{x} \rangle \le b\}$ 可以看作线性函数 $l_{\boldsymbol{a},b}(\boldsymbol{x}) = \langle \boldsymbol{a}, \boldsymbol{x} \rangle - b$ 的 0-下水平集. 全体线性函数构成类 $\mathscr{L}^d := \{l_{\boldsymbol{a},b} : (\boldsymbol{a}, b) \in \mathbb{R}^d \times \mathbb{R}\}$, 由线性代数的知识, 不难看出它是 $d+1$ 维的线性空间, 于是 $\mathcal{S}(\mathscr{L}^d)$ 的 VC 维数不会大于 $d+1$.

2.19 示例 ($\mathbb{R}^d$ 中的球). 考虑 $\mathbb{R}^d$ 中全体的欧氏球 $\mathcal{S}_{euc}^d := \{S_{\boldsymbol{a},b} : (\boldsymbol{a}, b) \in \mathbb{R}^d \times \mathbb{R}_+\}$, 其中 $S_{\boldsymbol{a},b} := \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x} - \boldsymbol{a}\|_2 \le b\}$ 为以 $\boldsymbol{a}$ 为圆心、b 为半径的欧氏球, 可以看作是 $s_{\boldsymbol{a},b} := \|\boldsymbol{x} - \boldsymbol{a}\|_2 - b$ 的 0-下水平集. 考虑特征映射 $\phi : \mathbb{R}^d \to \mathbb{R}^{d+2}$, $\boldsymbol{x} \mapsto \left(1, x_1, \ldots, x_n, \|\boldsymbol{x}\|_2^2\right)$, 函

数 $g_{\boldsymbol{c}}(\boldsymbol{x}) := \langle \boldsymbol{c}, \boldsymbol{x} \rangle$ 构成的函数类 $\mathscr{G} = \{g_{\boldsymbol{c}} \colon \boldsymbol{c} \in \mathbb{R}^{d+2}$ 是一个 $d+2$ 维的线性空间，并且 $s_{\boldsymbol{a},b}$ 就属于此类．于是 $\mathrm{vc}(\mathcal{S}^d_{euc}) \le d+2$．[3]

8.3.5 Empirical processes via VC dimension (HDP)

2.20 定理 *(VC 類的一致大數定律)*. 设 $\mathscr{F}$ 为 VC 类,

$$\mathbb{E}\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}\right] \le C\sqrt{\frac{\mathrm{vc}(\mathscr{F})}{n}}.$$

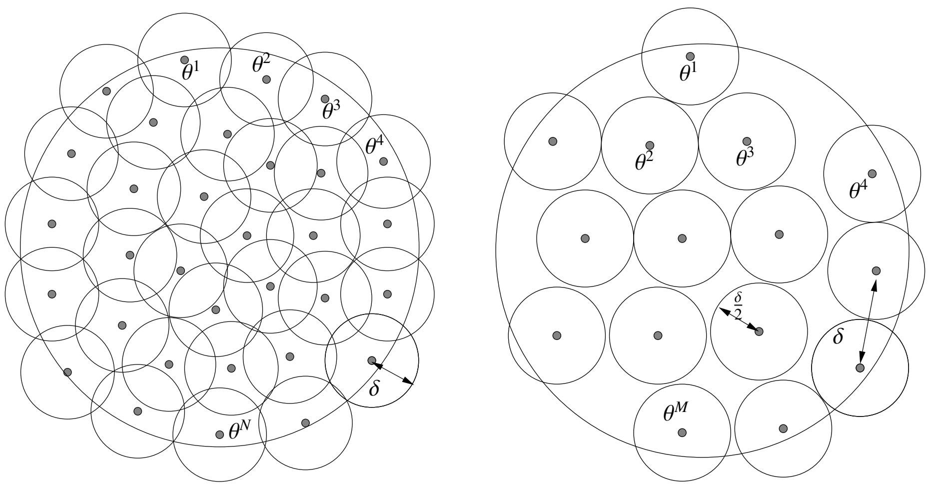
3 度量熵

对于次高斯

在度量空间理论中, ϵ-网, ϵ-覆盖,

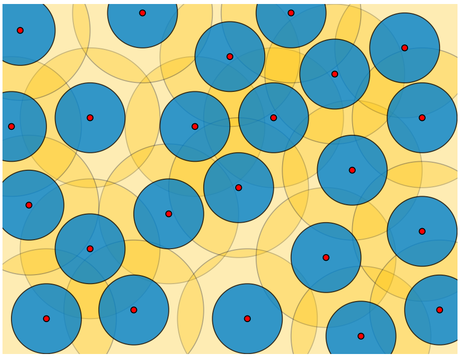
网、 包、 覆盖、均匀离散集、相对密集集和德尔尼集（以鲍里斯·德尔尼命名）是几个紧密相关的定义，用于描述点集的良好分布，这些集合的包络半径和覆盖半径衡量了它们的分布情况。这些集合在编码、近似算法等理论中有着应用.

度量空间 $(\mathcal{X}, \rho)$ 的子集 T



(a) (b)

[3]一个更加细致的分析可以得到它的 VC 维数实际上是 $d+1$.



A 預備知識

A.1 Landau 記號

A.2 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.2.1 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

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

證明. 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.2.2 矩生成函數、纍計生成函數

对于随机变量 X, 若函数 $M_X(\lambda) := \mathbb{E}[e^{\lambda X}]$ 在 0 的开邻域 I 内存在, 则称 M_X 为 X 的**矩生成函数** (moment generating function). 我们可以利用矩生成函数来导出 X 的各阶矩:

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

于是 $\mathbb{E}X^n = M_X^{(n)}(0)$. 随机变量 X 的**累计生成函数** (cumulant generating function) 是它的矩生成函数的自然对数

$$
K_X(\lambda) := \log \mathbb{E}[e^{\lambda X}],
$$

它的一个优点是它是加性函数——对于互相独立的随机变量 X 和 Y,

$$
K_{X+Y}(\lambda) = \log\left(\mathbb{E}[e^{\lambda X}] \cdot \mathbb{E}[e^{\lambda Y}]\right) = K_X(\lambda) + K_Y(\lambda).
$$

此外还可以按如下方式计算随机变量的各阶矩.

A.3 引理. 若非负随机变量 $X \in L^p$, $p > 0$, 则有

$$
\mathbb{E}X^p = \int_0^\infty p x^{p-1} \mathbb{P}(X > x) \,\mathrm{d}x. \tag{17}
$$

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.

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

□

A.2.3 Randon-Nikodym 導數、密度

Randon-Nikodym 导数是定义密度和条件期望的关键.

A.4 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.5 示例 *(分布的密度)*. 随机变量 $X : (\mathcal{X}, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}, \mu)$ 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:

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

从而由积分变换定理A.2,

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

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

A.6 示例 *(指數加權)*. 若随机变量 X 的矩生成函数 $\mathbb{E}[e^{\lambda X}]$ 在某个开区间 I 存在, 考虑指数加权期望 $\mathbb{E}_\lambda$

$$\mathbb{E}_\lambda[f(X)] := \frac{\mathbb{E}[f(X)e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} = \int f(X) \frac{e^{\lambda X}}{\mathbb{E}[e^{\lambda X}]} \,\mathrm{d}\mathbb{P}$$

它借用了 *Gibbs* 分布的思想, 用指数权重调整对随机变量的关注, 并将矩生成函数作为配分函数. 也可以看作定义了新的概率测度 $\mathbb{P}^\lambda$, 它关于 $\mathbb{P}$ 的 *Randon-Nikodym* 导数即为 $\frac{e^{\lambda X}}{\mathbb{E}[e^{\lambda X}]}$. 这被用在引理1.7和定理1.18的证明.

A.2.4 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 $X|\{\emptyset, \Omega\}$.

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

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 M 为凸函数且 $\mathbb{E}X, \mathbb{E}M(X) < \infty$, 则 $\mathbb{E}(M(X)|\mathcal{F}) \geq M(\mathbb{E}(X|\mathcal{F}))$;

(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.2.5 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.7 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.8 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.2.6 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.3 凸分析

A.3.1 Rademacher 定理

A.9 定理 *(Rademacher)*. 任意凸的 *Lipschitz* 函数几乎处处有导数

A.3.2 Fenchel 共轭

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 $\mathrm{dom} f := \{x\in\mathcal{X}\colon f(x)\in\mathbb{R}\neq\emptyset$, 其在 $u\in\mathcal{X}$ 的 **Fenchel 共轭**为

$$f^*(u)=\sup_{x\in\mathcal{X}}\left\{\langle x,u\rangle-f(x)\right\}. \tag{18}$$

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

$$f(x)+f^*(u)\geq\langle x,u\rangle. \tag{19}$$

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

王家卫在《一代宗师》里寄出一句台词:

人生要是无憾，那多无趣?

而我说：算法要是无憾，那应该是过拟合了。

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.

[GN15] Evarist Giné and Richard Nickl, *Mathematical foundations of infinite-dimensional statistical models*, Cambridge University Press, 2015.

[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.

[vH16] Ramon van Handel, *Probability in high dimension*, APC 550 Lecture Notes Princeton University, December 2016.

[Wai19] Martin J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*, Cambridge University Press, February 2019.