# 目录

# 1 Outline

总体来说我感觉高维可算是经典的渐进统计的一个发展。在渐进统计中我们研究的是 $n \to \infty$ 时的性质，最主要的就是各类的依概率收敛和 a.s. 收敛，或者最多到渐进正态性的问题。但是对于更复杂的一些情况，我们还是要回到何谓“$\infty$”这个东西本身：我们总是需要先认识有限的情况，再将之推广到无穷，所以在高维统计中，我们需要关心“从无穷撤回一些”，也就是 $n$ 大但不完全大的情况，关心一些参考量（最典型如 loss，或者说 excess risk）随 $n$ 增大的时候的变化趋势，从而了解一些统计模型的样本性质。渐进统计能够成立所依赖的 trick 是：对于性质足够好的“平均行为”（比如大数定律），其中总会有东西被随机性给平均掉/互相抵消掉，对于 $n \to \infty$ 的情况，我们简单地发现细粒度的细节全部消失了，只留下我们经常关心的粗粒度性质；而在高维统计中，我们进一步关心“在多大程度上（as a function of $n$）”这些细节被冲洗掉，以控制尾概率（tail probability）的形式出现，例如：

$$\mathbb{P}\left(|\cdot - \mathbb{E}[\cdot]| \geq t\right) \leq \text{decaying function of } t$$

一个典型的例子就是高维回归，比如 $p$ 与 $n$ 同步增长的情况。在传统渐进统计框架中 ($p = \text{const}$) 是无法处理这种问题的，而应该先撤回到有限的 $n$，通过包含 $n$ 的理论，将 $p$ 引入之后才能解决高维回归的统计性质问题。

所以对高维问题的理解大致有几步，就像把大象塞进冰箱一样：首先比如我们要研究一个复杂的对象，比如一串随机变量的最大值，一组数据的 empirical loss，一个高维矩阵的 norm（在某种意义上的，比如秩/operator norm/vector norm/trace）；之后大致确定 $\mathbb{E}[\cdot]$ 的 bound（as a function of e.g. $n, p$）；然后控制 $\cdot - \mathbb{E}[\cdot]$ 的 concentration。虽然只有几步，但是过程中需要处理很多细节：研究的对象（函数）需要有怎么样的性质才足够好，能够产生有效的 concentration（以及如何 relax 对函数性质的限制）？期望如何估计？（因为可能很难计算，没有解析表达式）

**参考书目：**

MJW *High-Dimensional Statistics: A Non-Asymptotic Viewpoint* by Martin J. Wainwright

RV *High-Dimensional Probability: An Introduction with Applications in Data Science* by Roman Vershynin

RH *High-Dimensional Statistics Lecture Notes* by Philippe Rigollet and Jan-Christian Hütter

RvH *Probability in High Dimension - APC 550 Lecture Notes* by Ramon van Handel

YW *Information Theoretic Methods for High-Dimensional Statistics* by yihong Wu

# 2 Concentration Inequalities

Concentration 指的是随机变量在“某个值”附近不会偏离太远，最典型的例子就是 Markov 不等式：与期望的偏差被方差控制。综合来说有很多 concentration 的方法，核心基本上就是“控制尾部”(tail behaviour)，以及一些主旨相似的方法，比如控制矩（或矩母函数）。除了直接控制随机变量，我们还希望研究控制随机变量的函数，以及相关随机过程的行为。这里给出的基本都是在最 general 的情况下的 bound，对于一些特殊的问题则常常可以通过更仔细地处理得到更好的 bound。

## 2.1 Sub-Gaussian

对于随机变量的尾部控制一个典型的例子就是研究所谓的“亚高斯型”(sub-Gaussian) 型随机变量，它有如下几个等价表述：$X$ is sub-Gaussian if and only if there exists some $K_i > 0$ such that

1. 控制尾部概率：

$$\mathbb{P}\left(|X - \mathbb{E}[X]| \geq t\right) \leq 2e^{-t^2/K_1^2}, \quad \forall t \geq 0$$

2. 控制所有阶矩：

$$\left(\mathbb{E}\left[|X - \mathbb{E}[X]|^p\right]\right)^{1/p} \leq K_2\sqrt{p}, \quad \forall p \geq 1$$

亚高斯随机变量的参数也由此处的最小的 $K_2$ 给出。

$$\|X\|_{\psi_2} := \sup_{p\geq 1} \frac{1}{\sqrt{p}} \left(\mathbb{E}\left[|X - \mathbb{E}[X]|^p\right]\right)^{1/p}$$

3. 控制矩母函数：

$$\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leq e^{\lambda^2 K_3^2/2}, \quad \forall \lambda \in \mathbb{R}$$

这样我们就有了一类“性质比较好的”随机变量，也就是尾部行为不会特别离谱，我们能够在均值附近解决问题。这样的随机变量在高维统计中有很多应用，最重要的是 Hoeffding bound：对于亚高斯零均值的 $X = (X_1, \ldots, X_n)$，我们有 $\forall a \in \mathbb{R}^n$:

$$\mathbb{P}\left(|\langle a, X\rangle| \geq t\right) \leq 2\exp\left(-\frac{t^2}{\|a\|_2^2 \max_i \|X_i\|_{\psi_2}}^2\right)$$

取 $a = \frac{1}{n}\mathbf{1}$，就得到关于均值的 Hoeffding bound。

类似的“好的”随机变量还有 sub-exponential 型，它的尾部行为稍差一些：

$$\mathbb{P}\left(|X - \mathbb{E}[X]|\right) \leq 2e^{-t/K_1}, \quad \forall t \geq 0$$
$$\left(\mathbb{E}\left[|X - \mathbb{E}[X]|^p\right]\right)^{1/p} \leq K_2 p, \quad \forall p \geq 1$$

其亚指数参数为

$$\|X\|_{\psi_1} := \sup_{p \geq 1} \frac{1}{p} \left(\mathbb{E}\left[|X - \mathbb{E}[X]|^p\right]\right)^{1/p}$$

由此可以直接得到亚指数型与亚高斯型的关系。对于亚高斯的 $X$：

$$\|X\|_{\psi_2} \leq \left\|X^2\right\|_{\psi_1} \leq 2\|X\|_{\psi_2}$$

仔细地展开矩母函数很容易证。

## 2.2 Hoeffding's Inequality

从随机变量出发我们可以 bound 相应的随机过程以及随机变量的函数。重要例子是 Hoefdding 不等式。对于独立的 $\{X_i \sim \text{Sub-Gaussian}\}_{i=1}^n$ 我们有

$$\mathbb{P}\left(\left|\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{\sum_{i=1}^n \|X_i\|_{\psi_2}^2}\right)$$

## 2.3 McDiarmid's Inequality

另一个常用的不等式是 McDiarmid 不等式，适用于逐坐标 bounded 的函数（bounded-difference function），是鞅差 bound 的推论。

$$|f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, \tilde{x}_i, x_{i+1}, \ldots, x_n)| \leq c_i$$

，有

$$\mathbb{P}\left(f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)] \geq t\right) \leq \exp\left[-\frac{2t^2}{\sum c_i^2}\right]$$

## 2.4 Bounds for Lipschitz Functions

一个更为实用/常用的情形是 bound Lipschitz 函数的行为。因为很多时候我们想要 bound 住的是类似 norm 的东西，这些东西很多时候都是 Lipschitz 函数。对于 $L$-Lipschitz 函数 $f : \mathcal{X}^n \mapsto \mathbb{R}$ 和 $X \sim \text{Gaussian}(0, I)$

$$\mathbb{P}\left(|f(X) - \mathbb{E}[f(X)]| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2nL^2}\right)$$

也即：高斯 r.v. 经过 $L$-Lipschitz 函数后的行为类似于亚高斯 subGau($L^2$)。

## 2.5 Maximal Inequality

对于亚高斯变量一个非常好的用处是：既然其尾部行为比较好，那么一组亚高斯变量的 max 的尾部行为也会比较好，表现在我们可以 bound$\mathbb{E}\left[\max_{i\le n} X_i\right]$。

对于 $X_1,\dots,X_n \sim \text{subGau}(\sigma^2)$ 有

$$\begin{aligned}\mathbb{E}\left[\max_{i\le n} X_i\right] =& \mathbb{E}\left[\frac{1}{\lambda}\log\max e^{\lambda X_i}\right]\\ \le& \frac{1}{\lambda}\log\mathbb{E}\left[\max e^{\lambda X_i}\right]\\ \le& \frac{1}{\lambda}\log\mathbb{E}\left[\sum_{i=1}^n e^{\lambda X_i}\right]\\ \le& \frac{1}{\lambda}\log n + \frac{1}{\lambda}\frac{\lambda^2\sigma^2}{2}\\ =& \frac{1}{\lambda}\log n + \frac{\lambda\sigma^2}{2},\quad \forall\lambda>0\end{aligned}$$

优化 $\lambda$ 得到

$$\mathbb{E}\left[\max_{i\le n} X_i\right] \le \sigma\sqrt{2\log n}$$

由于上面我们并没有施加任何独立性条件，所以我们可以创建一个 $X_1,\dots,X_n,-X_1,\dots,-X_n$ 来得到对 $\max|X_i|$ 的 bound

$$\mathbb{E}\left[\max_{i\le n} |X_i|\right] \le \sigma\sqrt{2\log 2n}$$

尾部概率则是简单地通过 union bound 得到

$$\begin{aligned}\mathbb{P}\left(\max_{i\le n} X_i > t\right) =& \mathbb{P}\left(\bigcup X_i > t\right)\\ \le& \sum_{i=1}^n \mathbb{P}\left(X_i>t\right) = n\mathbb{P}\left(X_1>t\right)\\ \le& n\exp\left(-\frac{t^2}{2\sigma^2}\right)\end{aligned}$$

**Idea**: 对有限个 subGau 的 maximal 的 bound 是很实用的，因为对于某个我们想要研究的无限集，我们可以通过寻找其有限大代表元素集合（比如 covering number）来将整个无限集上的 maximal bound 转化为有限集上的 maximal bound。见4.3节的相关讨论。

## 2.6 Others

- Bernstein's Inequality: 稍微更小心地处理 sub-exp 的尾部即可。对于零均值 $X_i \sim \text{subexp}(\lambda)$ 有

  $$\mathbb{P}\left(\bar{X}>t\right)\vee\mathbb{P}\left(\bar{X}<-t\right) \le \exp[-\frac{n}{2}(\frac{t^2}{\lambda^2}\wedge\frac{t}{\lambda})]$$

  反映了这样一件事：亚指数在近处的行为类似于高斯（$t^2$），在远处的行为类似于指数（$t$）。

- 鞅差序列 $\{(D_k,\mathcal{F}_k)\}_{k=1}^\infty$ 的 bound。如果 $D_k|\mathcal{F}_k$ 是亚指数的 sujbexp$(\nu_k,\alpha_k)$，那么和会有 bound

  $$\sum_{k=1}^n D_k \sim \text{subexp}(\sqrt{\sum_{k=1}^n \nu_k^2}, \max\alpha_k)$$

# 3 Random Process and Complexity Control

一个随机过程

$$X(t), \quad t \in T$$

是一个由指标集 $T$ index 的随机变量，其中 $T$ 可能是连续/离散，可以是多维度的，甚至有一些其它几何结构。在高维统计的场景中 $T$ 经常是模型假设类 $\mathcal{H}$（的参数化 $\Theta$），而 $X_i(t)$ 可以是 $X_i(t) = \hat{f}(x_i) - \mathbb{E}\left[\hat{f}(x_i)\right]$ 或是 $X_i(t) = \ell(h(x_i;\theta), y_i) - \mathbb{E}\left[\ell(h(x_i;\theta), y_i)\right]$ 等等。我们通常关心的对象也是形如

$$\mathbb{E}\left[\sup_{t\in T}\sum_{i=1}^{N} X_i(t)\right], \qquad \mathbb{E}\left[\sup_{t\in T}\left|\sum_{i=1}^{N} X_i(t)\right|\right]$$

的东西（一族零均值随机过程的上界）. 笼统来说这和指标集 $T$ 的“大小”和“形状”有关，我们有诸多手段来控制（upper or lower bound）这些东西。

## 3.1 Complexity Control Via Metric Entropy

### 3.1.1 Rademacher Complexity and Symmetrization

Complexity 是从随机过程的角度对一个集合的“大小”和“形状良好”的描述，关于其应用见4.3节。对于一列**零均值**随机过程 $[X_1(t), \ldots, X_N(t)] := \vec{X}(t)$，我们定义其 Rademacher 复杂度为[1]

$$\mathcal{R}(T) = \mathbb{E}\left[\sup_{t\in T}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i X_i(t)\right|\right], \quad \varepsilon_i \sim \text{Rademacher}(\{\pm 1\})$$

可以如此理解：$\vec{X}(t) \mapsto \left\langle \varepsilon, \vec{X}(t)\right\rangle$ 相当于是做了一个对称化的操作，我的理解是这让我们可以利用上零均值的条件，得到下面的 bound：

$$\frac{1}{2}\mathbb{E}\left[\sup\left|\left\langle \varepsilon, \vec{X}(t)\right\rangle\right|\right] \le \mathbb{E}\left[\sup\left|\sum_i X_i(t)\right|\right] \le 2\mathbb{E}\left[\sup\left|\left\langle \varepsilon, \vec{X}(t)\right\rangle\right|\right]$$

证明思路和4.3中的类似，通过反复将对称话操作拿出拿进，制备 $\vec{X}_t$ 的 copies 来实现转换。[2]

[1] 也有一些相关概念，包括 Gaussian complexity:

$$\mathcal{G}(T) := \mathbb{E}\left[\sup_{t\in T}\left|\sum_{i=1}^{N}\varepsilon_i X_i(t)\right|\right], \quad \varepsilon_i \sim \mathcal{N}(0,1)$$

或者 Gaussian width:

$$w(T) := \mathbb{E}\left[\sup_{t\in T}\sum_{i=1}^{N}\varepsilon_i X_i(t)\right], \quad \varepsilon_i \sim \mathcal{N}(0,1)$$

[2] 不带绝对值的 width 版本也是可证的：

$$\frac{1}{2}\mathbb{E}\left[\sup\left\langle \varepsilon, \vec{X}(t)\right\rangle\right] \le \mathbb{E}\left[\sup\sum_i X_i(t)\right] \le 2\mathbb{E}\left[\sup\left\langle \varepsilon, \vec{X}(t)\right\rangle\right]$$

### 3.1.2 Covering Number

本 subsection 余下部分本来接在 ULLN control 后，所以使用的是 $(X_\theta, \Theta)$ 或 $X_h, \mathcal{H}$ 这套记号，基本上作 $\theta \mapsto t$, $\Theta \mapsto T$ 即可接续上文。

复杂度笼统来说就是一个集合的“大小”。在高维问题中我们很多时候关心的是一个无限集合，这个时候没办法用 card 来处理问题，退而求其次得我们研究 covering number 和 packing number。也就是将集合 $\mathcal{H}$ 近似到某个精确度 $\varepsilon$ 来研究。

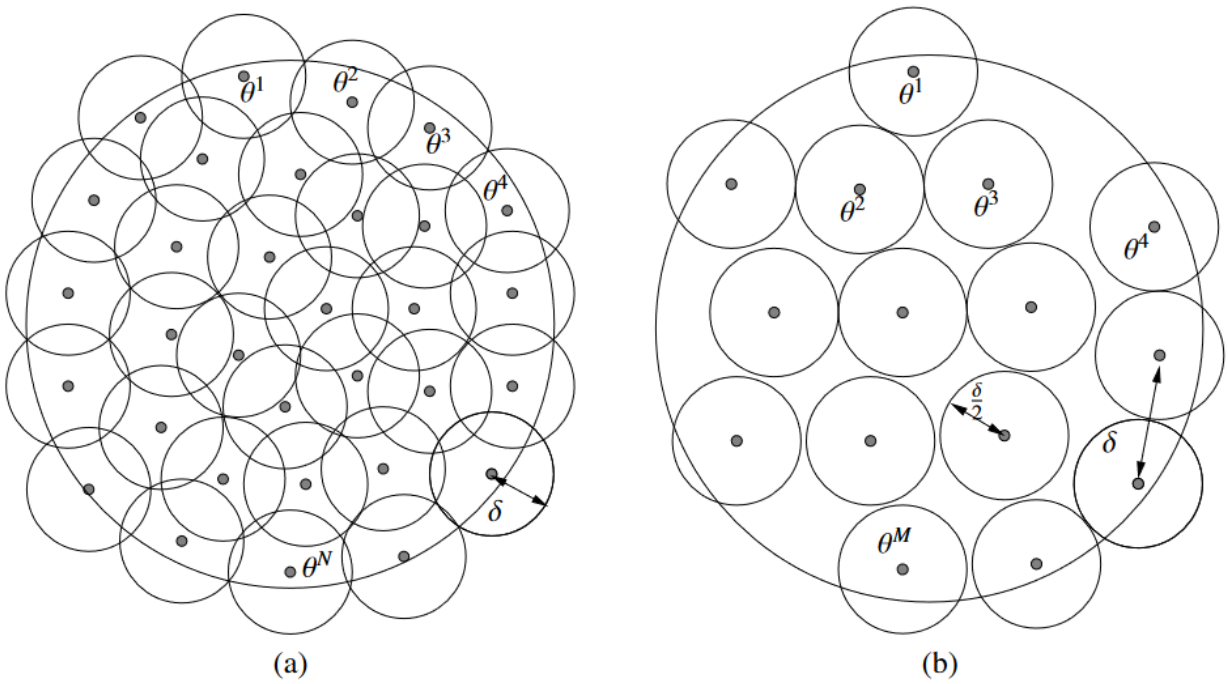



**Figure 5.1** Illustration of packing and covering sets. (a) A $\delta$-covering of $\mathbb{T}$ is a collection of elements $\{\theta^1, \ldots, \theta^N\} \subset \mathbb{T}$ such that for each $\theta \in \mathbb{T}$, there is some element $j \in \{1, \ldots, N\}$ such that $\rho(\theta, \theta^j) \leq \delta$. Geometrically, the union of the balls with centers $\theta^j$ and radius $\delta$ cover the set $\mathbb{T}$. (b) A $\delta$-packing of a set $\mathbb{T}$ is a collection of elements $\{\theta^1, \ldots, \theta^M\} \subset \mathbb{T}$ such that $\rho(\theta^j, \theta^k) > \delta$ for all $j \neq k$. Geometrically, it is a collection of balls of radius $\delta/2$ with centers contained in $\mathbb{T}$ such that no pair of balls have a non-empty intersection.

- Covering Number: 对于一个集合 $\mathcal{H}$，我们定义其 $\varepsilon$-covering number 为

$$N(\varepsilon, \mathcal{H}, \|\cdot\|) := \min\{m \in \mathbb{N} : \exists h_1, \ldots, h_m \in \mathcal{H}, \forall h \in \mathcal{H}, \exists i \in [m], \|h - h_i\| \leq \varepsilon\}$$

- Packing Number: 对于一个集合 $\mathcal{H}$，我们定义其 $\varepsilon$-packing number 为

$$M(\varepsilon, \mathcal{H}, \|\cdot\|) := \max\{m \in \mathbb{N} : \exists h_1, \ldots, h_m \in \mathcal{H}, \forall i \neq j, \|h_i - h_j\| \geq \varepsilon\}$$

他们之间有如下很好的关系：

$$M(2\varepsilon, \mathcal{H}, \|\cdot\|) \overset{(i)}{\leq} N(\varepsilon, \mathcal{H}, \|\cdot\|) \overset{(ii)}{\leq} M(\varepsilon, \mathcal{H}, \|\cdot\|)$$

证明. (i) 考虑满足 $M(2\varepsilon,\mathcal{H},\|\cdot\|)$ 的 $\mathcal{M}(2\varepsilon,\mathcal{H},\|\cdot\|)\subset\mathcal{H}$ 及其中任意两个不同元素 $h_i,h_j$。取某个在 $\mathbb{B}_{\|\cdot\|}(h_i,\varepsilon)$ 中的点 $h$ 我们即有

$$\|h-h_j\|\geq\|h_i-h_j\|-\|h-h_i\|>2\varepsilon-\varepsilon=\varepsilon,\quad\forall h\in\mathbb{B}_{\|\cdot\|}(h_i,\varepsilon)$$

也就是说 $\mathcal{M}(2\varepsilon,\mathcal{H},\|\cdot\|)\setminus\{h_i\}$ 不满足 $\varepsilon$-covering，又因为 $\varepsilon$-covering 是所有如此集合的 card 下界，所以 $N(\varepsilon,\mathcal{H},\|\cdot\|)\geq M(2\varepsilon,\mathcal{H},\|\cdot\|)$。

(ii) 考虑满足 $M(\varepsilon,\mathcal{H},\|\cdot\|)$ 的 $\mathcal{M}(\varepsilon,\mathcal{H},\|\cdot\|)\subset\mathcal{H}$。由于 $\mathcal{M}(\varepsilon,\mathcal{H},\|\cdot\|)$ 是满足 $\varepsilon$-packing 的最大集合，所以任意 $h\in\mathcal{H}$ 必满足

$$\exists h_i\in\mathcal{M}(\varepsilon,\mathcal{H},\|\cdot\|),\ \|h-h_i\|\leq\varepsilon$$

（也就是 $\varepsilon$-covering）。因为如果不如此的话，$M(\varepsilon,\mathcal{H},\|\cdot\|)\cup\{h\}$ 就是一个更大的 $\varepsilon$-packing，这与 $\mathcal{M}(\varepsilon,\mathcal{H},\|\cdot\|)$ 的最大性矛盾。所以 $N(\varepsilon,\mathcal{H},\|\cdot\|)\leq M(\varepsilon,\mathcal{H},\|\cdot\|)$。

□

所以我们只需能 bound 住其中之一即可控制集合的尺寸，一般是 bound covering。

一个应用是通过这个方法控制某个集合的 Gaussian 复杂度，见 Wainwright 的书 P. 136。核心方法是：对于零均值亚高斯过程 $\{X_\theta:\theta\in\Theta\}$（w.r.t metric $\rho_X$），有如下的 discretization bound

$$\mathbb{E}\left[\sup_{\theta,\tilde{\theta}\in\Theta}(X_\theta-X_{\tilde{\theta}})\right]\leq\underbrace{2\mathbb{E}\left[\sup_{\gamma,\gamma'\in\Theta;\rho_X(\gamma,\gamma')\leq\varepsilon}(X_\gamma-X_{\gamma'})\right]}_{\text{approximation error}}+\underbrace{2\sqrt{D^2\log N(\varepsilon,\Theta,\rho_X)}}_{\text{estimation error}},\quad\forall\varepsilon$$

其中 $D=\sup_{\theta,\tilde{\theta}\in\Theta}\rho_X(\theta,\tilde{\theta})$。亚高斯过程指的是

$$\mathbb{E}\left[e^{\lambda(X_\theta-X_{\tilde{\theta}})}\right]\leq e^{\lambda^2\rho_X(\theta,\tilde{\theta})^2/2}$$

上面的 discretization bound 的右边是一个 $\varepsilon$ 的函数，我们再对 $\varepsilon$ 优化就能 trade-off 两个 error 得到一个更好的 bound。

以 $\mathcal{T}\subset\mathbb{R}^n$ with $\ell_2$ norm 为例：

$$\mathcal{G}(\mathcal{H})\leq\min_{\varepsilon\in[0,D]}\left\{\varepsilon\sqrt{d}+2\sqrt{D^2\log N(\varepsilon,\mathcal{H},\|\cdot\|_2)}\right\}$$

具体例子见 Wainwright 的书 P. 137。

证明. (Sketch) 本质上大致是这么件事情：在 $\theta,\tilde{\theta}$ 之间插入 $\gamma,\gamma'\in\mathcal{N}(\varepsilon,\Theta,\rho_X)$

$$\begin{aligned}\sup_{\theta,\tilde{\theta}\in\Theta}(X_\theta-X_{\tilde{\theta}})\leq&\sup_{\theta;\gamma\in\mathcal{N}(\varepsilon,\Theta,\rho_X)}(X_\theta-X_\gamma)+\sup_{\gamma,\gamma'\in\mathcal{N}(\varepsilon,\Theta,\rho_X)}(X_\gamma-X_{\gamma'})+\sup_{\tilde{\theta};\gamma'\in\mathcal{N}(\varepsilon,\Theta,\rho_X)}(X_{\gamma'}-X_{\tilde{\theta}})\\\leq&2\sup_{\gamma,\gamma'\in\Theta;\rho_X(\gamma,\gamma')\leq\varepsilon}(X_\gamma-X_{\gamma'})+2\max_{\theta_i\in\mathcal{N}(\varepsilon,\Theta,\rho_X)}|X_{\theta_i}-X_{\theta_1}|\\\leq&2\sup_{\gamma,\gamma'\in\Theta;\rho_X(\gamma,\gamma')\leq\varepsilon}(X_\gamma-X_{\gamma'})+2\sqrt{D^2\log N(\varepsilon,\Theta,\rho_X)}\end{aligned}$$

□

### 3.1.3 Chaining

上面的红色部分实际上是一个过于粗糙的近似，我们可以用 chaining 方法来改进，得到 Dudley's entropy integral bound.

$$\mathbb{E}\left[\sup_{\theta,\tilde{\theta}\in\Theta}(X_\theta - X_{\tilde{\theta}})\right] \le \underbrace{2\mathbb{E}\left[\sup_{\gamma,\gamma'\in\Theta;\rho_X(\gamma,\gamma')\le\varepsilon}(X_\gamma - X_{\gamma'})\right]}_{\text{approximation error}} + \underbrace{64\sqrt{2}\int_{\varepsilon/4}^{D/2}\sqrt{\log N(u,\Theta,\rho_X)}\,\mathrm{d}u}_{\text{estimation error}}, \quad \forall\varepsilon$$

证明. (Sketch)Chaining 方法的主旨是，将 $\Theta$ 在 $\varepsilon_m = D\cdot 2^{-m}$ 精度下依次划分，然后对于每一个递进层 $\mathcal{N}(D2^{-m},\Theta,\rho_X) \rightsquigarrow \mathcal{N}(D2^{-(m+1)},\Theta,\rho_X)$，bound 粗细粒度改进时 approximation error:

$$\begin{aligned}
\sup_{\theta,\tilde{\theta}\in\Theta}(X_\theta - X_{\tilde{\theta}}) &\le \max_{\gamma,\gamma'\in\mathcal{N}(D2^{-1},\Theta,\rho_X)}|X_\gamma - X_{\gamma'}| + \sum_{m=2}^{\sim\log_2 D}\max_{\beta\in\mathcal{N}(D2^{-m},\Theta,\rho_X)}|X_\beta - X_{\text{sth in previous level}}| \\
&\le \sum_{m=1}^{\sim\log_2 D} 2D\cdot 2^{-m+1}\sqrt{\log N(D2^{-m},\Theta,\rho_X)} \\
&\lessapprox \text{const}\cdot\int_{\varepsilon/4}^{D/2}\sqrt{\log N(u,\Theta,\rho_X)}\,\mathrm{d}u \qquad \text{(precise)} \\
&\lesssim C\cdot\int_0^{D/2}\sqrt{\log N(u,\Theta,\rho_X)}\,\mathrm{d}u \qquad \text{(rough)}
\end{aligned}$$

其中的红色部分是通过更精确的 trivial 计算可算出来的常数。 □

## 3.2 Sudakov-Fernique's Inequality for Gaussian Processes

对于特别的 $X(t)\sim\mathcal{N}(0,\Sigma)$ 有特殊的比较定理，Sudakov-Fernique's 不等式。若 $(X(t))_{t\in T}$ 和 $(Y(t))_{t\in T}$ 是零均值**高斯**过程，有

$$\|X_t - X_s\|_2 \le \|Y_t - Y_s\|_2,\ \forall t,s\in T \Rightarrow \mathbb{E}\left[\sup_t X_t\right] \le \mathbb{E}\left[\sup_t Y_t\right]$$

**Remark:** 注意到这里的条件相当于在说 $\mathbb{E}[X_tX_s] \ge \mathbb{E}[Y_tY_s]$，这是一个“协方差的单调性”条件：内部“更不一致的”（$Y$）更容易有大的 supremum。

证明. 使用如下的引理：对于 $W\sim\mathcal{N}(0,\Xi)$ 和可导的 $h$，有

$$\mathbb{E}[Wh(W)] = \Sigma\mathbb{E}[\nabla h(W)]$$

用密度函数做分部积分即可得到。

然后对 $X(t),Y(t)$ 做插值：

$$Z_u(t) = \sqrt{u}X(t) + \sqrt{1-u}Y(t)$$

然后对合适的 $f$ 尝试证明 $\mathbb{E}[f(Z_u(t))]$ 关于 $u$ 的单调性即可。具体来说取 $f(x) = \frac{1}{\beta}\log\sum_i e^{\beta x_i} \underset{\beta\to\infty}{\to} \max x_i$ 即

可。

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}u}\mathbb{E}\left[f(Z(u))\right] =& \mathbb{E}\left[\sum_{i=1}^{n}\frac{\mathrm{d}f}{\mathrm{d}\xi_i}\left(\frac{X(t)}{2\sqrt{u}}-\frac{Y(t)}{\sqrt{1-u}}\right)\right]\\ \overset{\text{Lemma}}{=}& \frac{1}{2\sqrt{u}}\sum_{i=1}^{n}\Sigma_X\cdot\mathbb{E}\left[\sqrt{u}\nabla\frac{\mathrm{d}f}{\mathrm{d}\xi_i}\right]-\frac{1}{2\sqrt{1-u}}\sum_{i=1}^{n}\Sigma_Y\cdot\mathbb{E}\left[\sqrt{1-u}\nabla\frac{\mathrm{d}f}{\mathrm{d}\xi_i}\right]\\ =&\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\Sigma_{X,ij}-\Sigma_{Y,ij})\mathbb{E}\left[\frac{\mathrm{d}^2f}{\mathrm{d}\xi_i\xi_j}\right]\end{aligned}$$

trivially 可以发现 $\dfrac{\mathrm{d}^2f}{\mathrm{d}\xi_i\xi_j}\le 0$，这样就能用单调性得到结论。

□

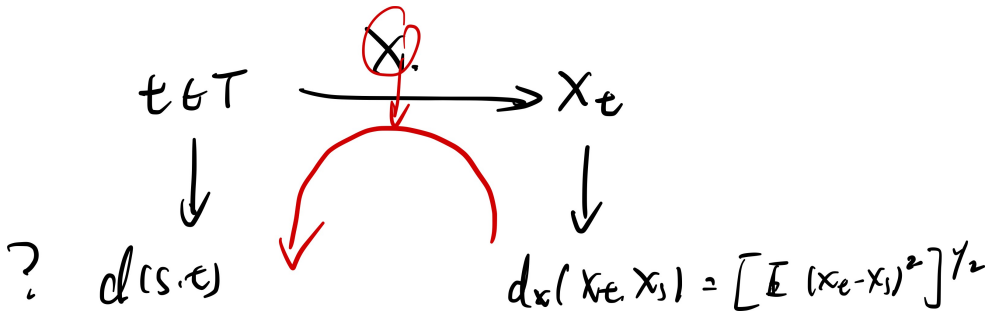
**Application**: $L$-Lipshitz 函数的 supermum bound。

**Example**: 对随机矩阵 $\{X_{ij}\overset{i.i.d.}{\sim}\mathcal{N}(0,1)\}_{m\times n}$ 推 bound $\mathbb{E}\left[|||X|||\right]\le\sqrt{m}+\sqrt{n}$。

**Application**: 将 $\vec{X}(t)$ 设为我们想研究的过程，$Y$ 是另一个相关的，容易研究的过程（or vice versa），通过 Sudakov-Fernique's 不等式我们可以得到 $\mathbb{E}\left[\sup X(t)\right]$ 的 upper(或 lower) bound。具体来说，

$$(s,t)\mapsto\mathbb{E}\left[(X(t)-X(s))^2\right]^{1/2}$$

是 $\{X\}$ 空间中的正则度量（欧式度量），我们可以通过下面这个反向路径诱导出 $T$ 空间的（伪）度量：




也就是 $X(\cdot)\mapsto d_T(\cdot,\cdot)$，这允许我们通过在 $X$ 欧式空间中研究 $T$ 的几何和尺寸（比如这里就可以应用 Sudakov Ineq 了）。

比如对于这样诱导出的 $T$ 空间的度量，我们可以进行 $T$ 的 packing$\mathcal{M}(\delta,T,d_T(\cdot,\cdot))=\mathcal{M}(\delta,X(T),|||_2)$。进而通过 $Y_t=$ independent $N(0,\delta^2/2)$ 即可得到一个 $X_t$ 的 lower bound：

$$\forall s,t\in\mathcal{M}(\delta,T,d_T(\cdot,\cdot)):\ d^2(s,t)=\mathbb{E}\left[(X_s-X_t)^2\right]\ge\varepsilon^2=\mathbb{E}\left[N(0,\delta^2)^2\right]=\mathbb{E}\left[(Y_t-Y_s)^2\right]$$

故而得到 (use MJW Exercise 2.11 or Vershynin Exercise 2.5.11)

$$\mathbb{E}\left[\sup X_t\right]\ge\mathbb{E}\left[\sup Y_i\right]\ge c\delta\sqrt{\log M(\delta,T,d_T(\cdot,\cdot))}$$

进一步由于有 $M(\delta,T,\cdot)\ge N(\delta,T,\cdot)$ 我们也可以将 $X_t$ 构造成一个容易的过程来 bound 覆盖数。

**Example:** 比如对于直径 $D\le 1$ 的 $N$ 顶点多边形 $P$

$$\delta\sqrt{\log N(\delta,T,\cdot)}\le c\mathbb{E}\left[\sup_{t\in P}\langle\varepsilon,t\rangle\right]=c\mathbb{E}\left[\sup_{t\in\text{vertices}}\langle t,\varepsilon\rangle\right]\lesssim\delta\sqrt{\log N},\qquad\varepsilon\sim N(0,1)$$

以此得到

$$N(\delta, T, \cdot) \lesssim N^{c/\varepsilon^2}$$

（更精确的结果见 Stat461-2023Fall exam Q2，得到 $C = 1$）

### 3.3 本章小结

我们有如下的 bound：对于零均值高斯过程 $[X_1(t), \ldots, X_N(t)] := \vec{X}(t)\,, t \in T$，和 $\delta$ covering number $N(\delta, T, \cdot)$，有

$$c \sup_{\delta>0} \delta \sqrt{\log N(\delta, T, \cdot)} \overset{\text{Sudakov}}{\leq} \mathbb{E}\left[\sup_{t\in T} X_t\right] \overset{\text{Dudley}}{\leq} C \int_0^{\text{Diam}(T)/2} \sqrt{\log N(u, T, \cdot)}\, \mathrm{d}u$$

其中更具体来说，Dudley side 对于 subGau 也有效，上界 $\propto \|X_t\|_{\psi_2}$.

## 4 Supervised Learning and Gereralization Bounds

### 4.1 Model and Empirical Risk

机器学习中的典型框架是有监督学习 (supervised learning)，数据为 $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$，$X$ 与 $Y$ 之间由函数 $h : \mathcal{X} \mapsto \mathcal{Y}$ 刻画，我们即希望学习这样一个 predictor（也就是我们的模型）。模型的好坏由一个损失函数 (loss function)$\ell : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$ 来刻画，以反映估计值和真实值之间的差距。

- Expected Loss (of a model $h$)：

$$L(h) = \mathop{\mathbb{E}}_{X,Y} \left[\ell\left(h(X), Y\right)\right]$$

- Excess Risk:

$$E(h) = L(h) - \inf_{h'\in\mathcal{H}} L(h')$$

其中 $\mathcal{H}$ 是我们的假设空间/假设类（hypothesis class）。我们的任务就是在 $\mathcal{H}$ 中找到一个 $h$，使得 $E(h)$ 尽可能小，即取得最小的 excess risk（对应泛化误差）。进一步，表达式后半部分的 minimizer 即为 ground truth 的 $h^*$

$$h^* = \arg\inf_{h'\in\mathcal{H}} L(h')$$

为了强调这一点我们 overload$E$ 为 $E(h, h^*)$。

- Emperical Loss: 实际上我们并不能求出 Expected Loss，只能用数据来近似之。

$$\hat{L}_n = \hat{L}(\mathcal{D}_n) = \frac{1}{n}\sum_{i=1}^n \ell\left(h(x_i), y_i\right)$$

已经提到，我们的目标是找到一个 $h$，使得 $E(h, h^*)$（或等价地，使 $L(h)$）尽可能小，我们的估计量则自然是通过最小化 $\hat{L}_n$ 来实现的，此估计量为 $\hat{h}_n$。

$$\hat{h}_n = \arg\inf_{h'\in\mathcal{H}} \hat{L}_n(h')$$

- 参数化：对于假设类 $\mathcal{H}$，很多情况下我们可以将其参数化到有限维空间上（比如线性模型 $\{x \mapsto \beta \cdot x,\, \beta \in \mathbb{R}\}$，CDF 经验估计 $\{x \mapsto \mathbf{1}_{(-\infty,t]}(x),\, t \in \mathbb{R}\}$），此时我们直接 overload 上述的 $L$ 和 $E$，用函数 $h \in \mathcal{H}$ 的参数化 $\theta \in \Theta$ 来表示 $h$，即 $L(\theta)$ 和 $E(\theta, \theta^*)$。（非参数化情形暂不讨论）

$$
\begin{aligned}
L(\theta) =& \mathop{\mathbb{E}}_{X,Y}\left[\ell\left(h(X;\theta), Y\right)\right] \\
E(\theta, \theta^*) =& L(\theta) - \inf_{\theta' \in \Theta} L(\theta') = L(\theta) - L(\theta^*) \\
\hat{L}_n(\theta) =& \frac{1}{n}\sum_{i=1}^{n} \ell\left(h(x_i;\theta), y_i\right)
\end{aligned}
$$

一个好的模型应该能使得 emperical risk $E(\hat{\theta}, \theta^*)$ 尽可能小。在高维统计中我们关心：**随着样本量 $n$ 的增大，$E(\hat{\theta}, \theta^*)$ 以何种趋势减小**，即我们关心的是 $E(\hat{\theta}, \theta^*)$ 的性质，它可以作如下分解：

$$
E(\hat{\theta}, \theta^*) = \underbrace{L(\hat{\theta}) - \hat{L}(\hat{\theta})}_{\text{I}} + \underbrace{\hat{L}(\hat{\theta}) - \hat{L}(\theta^*)}_{\text{II}} + \underbrace{\hat{L}(\theta^*) - L(\theta^*)}_{\text{III}}
$$

其中：

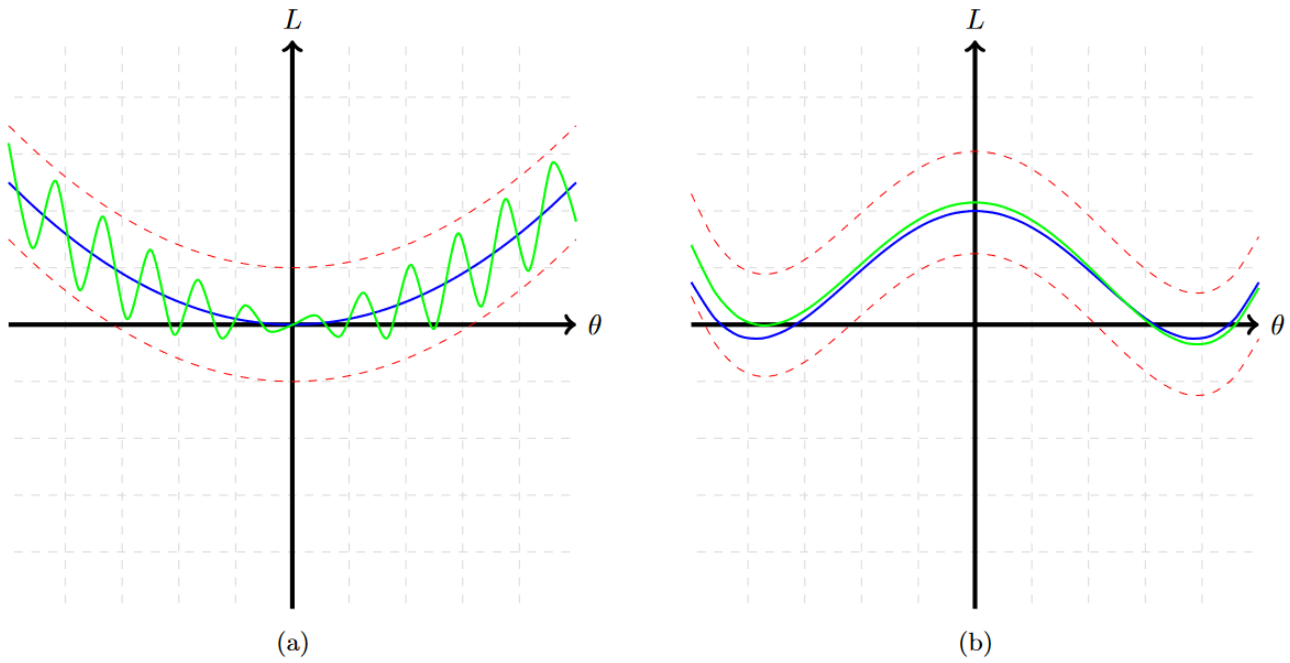
- II 非正
- III 由于 $\theta^*$ 是不带随机性的参数，并注意到 $\mathbb{E}_{\mathcal{D}\sim\theta^*}\left[\hat{L}(\theta^*)\right] = L(\theta^*)$，这一项可以用经典的对均值的 bound 来控制，比如 Hoeffding bound。

  I 比较麻烦，因为 $\hat{L}$ 是与样本有关的（有随机性）

  $$
  \text{I} = \mathbb{E}\left[\ell(h(X;\hat{\theta}), Y)\right] - \frac{1}{n}\sum_{i=1}^{n} \ell(h(x_i;\hat{\theta}), y_i)
  $$

  这需要 **ULLN**（uniform law of large numbers）来控制

  $$
  \text{I} \le \sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n} \ell(h(x_i;\theta), y_i) - \mathbb{E}\left[\ell(h(X;\theta), Y)\right]\right| := \left\|\mathbb{P}_n - \mathbb{P}\right\|_{\mathcal{H}(\Theta)}
  $$




(a) (b)

Figure 4.1: These curves demonstrate how we apply uniform convergence to bound the population risk. The blue curves are the unobserved population risk we aim to bound. The green curves denote the empirical risk we observe. Though this curve is often depicted as the fluctuating curve used in Figure 4.1a, it is more often a smooth curve whose shape mimics that of the population risk (Figure 4.1b). Uniform convergence allows us to construct additive error bounds for the excess risk, which are depicted using the red, dashed lines.

## 4.2 ULLN Control

(To be modified)

## 4.3 Via Rademacher Complexity

控制 $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{H}(\Theta)}$ 的一种方法是用 Rademacher 复杂度（Rademacher complexity）

定义：对于一个集合 $\mathcal{I} \equiv \{\imath\} \subset \mathbb{R}^n$，我们定义其 Rademacher 复杂度 $\mathcal{R}(\mathcal{I})$ 为

$$\mathcal{R}(\mathcal{I}) := \mathbb{E}\left[\sup_{\imath\in\mathcal{I}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\imath_i\right|\right], \quad \varepsilon_i \sim \text{Rademacher}(\{\pm 1\})$$

直观来说，如果集合 $\mathcal{I}$ “很大”，那么我们就容易地总能找到 $\imath \in \mathcal{I}$ 很大，使得 $\langle\varepsilon, \imath\rangle$ 很大。更多关于 Complexity 的讨论见3.1.1节。

我们会看到，risk 就可以通过 Rademacher 复杂度来控制。这里的 loss 来自前面 I 部分，即我们希望控制

$$\mathrm{I} \le \sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\ell(h(x_i;\theta), y_i) - \mathbb{E}\left[\ell(h(X;\theta), Y)\right]\right| := \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{H}(\Theta)}$$

对于任意给定的数据 $\mathcal{D}$ 和假设空间 $\mathcal{H}$ 都会生成一个集合，为了克服掉数据的随机性我们再套一个期望，进一步定义这个场景下的 Rademacher 复杂度为

$$\mathcal{R}(\mathcal{H}) := \mathbb{E}_{\mathcal{D}\sim\mathbb{P}}\left[\mathcal{R}(\{\ell(h(x_i), y_i)\}_{h\in\mathcal{H}})\right] = \mathbb{E}_{\mathcal{D}\sim\mathbb{P}}\left[\mathbb{E}_{\varepsilon\sim\text{Unif}\{\pm1\}}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\ell(h(x_i), y_i)\right|\right]\right]$$

这个情形下可以用 Rademacher 复杂度来控制 risk 的收敛性。定理如下：

$$\frac{1}{2}\mathcal{R}(\mathcal{H})-\frac{\sup_{h\in\mathcal{H}}\mathbb{E}\left[|\ell(h(x),y)|\right]}{2\sqrt{n}}\leq\mathbb{E}\left[\|\mathbb{P}_n-\mathbb{P}\|_{\mathcal{H}}\right]\leq 2\mathcal{R}(\mathcal{H})$$

定理的证明用了一个神乎其技的对称话 trick

上界 ：

证明. 我们制备一个 $\mathcal{D}=\{(x_i,y_i)\}_{i=1}^n$ 的 i.i.d. copy$\tilde{\mathcal{D}}=\{(\tilde{x}_i,\tilde{y}_i)\}_{i=1}^n$. 于是可以在一些地方把 $\mathcal{D}$ 和 $\tilde{\mathcal{D}}$ 混合/替换，然后再用对称属性将混合的 $X$-$Y$ 换成 $\varepsilon X$ 从而由 risk gap 过渡到 complexity，实在是非常巧妙：

$$\begin{aligned}
\mathbb{E}\left[\|\mathbb{P}_n-\mathbb{P}\|_{\mathcal{H}}\right]=&\frac{1}{n}\mathbb{E}_{\mathcal{D}}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^n\ell\big(h(x_i),y_i\big)-\mathbb{E}_{\tilde{\mathcal{D}}}\left[\ell\big(h(\tilde{x}_i,\tilde{y}_i)\big)\right]\right|\right]\\
=&\frac{1}{n}\mathbb{E}_{\mathcal{D}}\left[\sup_{h\in\mathcal{H}}\left|\mathbb{E}_{\tilde{\mathcal{D}}}\sum_{i=1}^n\ell\big(h(x_i),y_i\big)-\ell\big(h(\tilde{x}_i,\tilde{y}_i)\big)\right|\right]\\
\overset{\text{Jensen}}{\leq}&\frac{1}{n}\mathbb{E}_{\mathcal{D},\tilde{\mathcal{D}}}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^n\ell\big(h(x_i),y_i\big)-\ell\big(h(\tilde{x}_i,\tilde{y}_i)\big)\right|\right]\\
\overset{\text{symmetrize}}{=}&\frac{1}{n}\mathbb{E}_{\mathcal{D},\tilde{\mathcal{D}},\varepsilon}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^n\varepsilon_i\left[\ell\big(h(x_i),y_i\big)-\ell\big(h(\tilde{x}_i,\tilde{y}_i)\big)\right]\right|\right]\\
\leq&2\mathbb{E}_{\mathcal{D},\varepsilon}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^n\varepsilon_i\ell\big(h(x_i),y_i\big)\right|\right]\\
=&2\mathcal{R}(\mathcal{H})
\end{aligned}$$

□

下界 ：

证明. 再用一次对称化得到另一边的 bound

$$
\begin{aligned}
\mathbb{E}\left[\|\mathbb{P}_n-\mathbb{P}\|_{\mathcal{H}}\right] =& \frac{1}{n}\mathbb{E}_{\mathcal{D}}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^{n}\ell\big(h(x_i),y_i\big)-\mathbb{E}_{\mathcal{D}}\left[\ell\big(h(x_i,y_i)\big)\right]\right|\right] \\
=& \frac{1}{2n}\mathbb{E}_{\mathcal{D}}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^{n}\ell\big(h(x_i),y_i\big)-\mathbb{E}_{\tilde{D}}\left[\ell\big(h(\tilde{x_i},\tilde{y_i})\big)\right]\right|\right] \\
&+\frac{1}{2n}\mathbb{E}_{\tilde{\mathcal{D}}}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^{n}\ell\big(h(\tilde{x_i}),\tilde{y_i}\big)-\mathbb{E}_{\tilde{D}}\left[\ell\big(h(\tilde{x_i},\tilde{y_i})\big)\right]\right|\right] \\
\overset{\text{triangular}}{\geq}& \frac{1}{2n}\mathbb{E}_{\mathcal{D},\tilde{\mathcal{D}}}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^{n}\ell\big(h(x_i),y_i\big)-\ell\big(h(\tilde{x_i}),\tilde{y_i}\big)\right|\right] \\
\overset{\text{symmetrize}}{=}& \frac{1}{2n}\mathbb{E}_{\mathcal{D},\tilde{\mathcal{D}},\varepsilon}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^{n}\varepsilon_i\big[\ell\big(h(x_i),y_i\big)-\ell\big(h(\tilde{x_i}),\tilde{y_i}\big)\big]\right|\right] \\
\overset{\text{convex}}{\geq}& \frac{1}{2n}\mathbb{E}_{\mathcal{D},\varepsilon}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^{n}\varepsilon_i\big(\ell\big(h(x_i),y_i\big)-\mathbb{E}_{\mathcal{D}}\left[\ell(h(x_i),y_i)\right]\big)\right|\right] \\
\geq& \frac{1}{2n}\mathbb{E}_{\mathcal{D},\varepsilon}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^{n}\varepsilon_i\ell\big(h(x_i),y_i\big)\right|-\frac{1}{2n}\sup_{h\in\mathcal{H}}\mathbb{E}\left[|\ell(h(x),y)|\right]\left|\sum_{i=1}^{n}\varepsilon_i\right|\right] \\
\geq& \frac{1}{2}\mathcal{R}(\mathcal{H})-\frac{\sup\limits_{h\in\mathcal{H}}\mathbb{E}\left[|\ell(h(x),y)|\right]}{2\sqrt{n}}
\end{aligned}
$$

□

综合以上我们就得到了关于 Rademacher 复杂度的 risk bound

$$
\frac{1}{2}\mathcal{R}(\mathcal{H})-\frac{\sup\limits_{h\in\mathcal{H}}\mathbb{E}\left[|\ell(h(x),y)|\right]}{2\sqrt{n}}\leq\mathbb{E}\left[\|\mathbb{P}_n-\mathbb{P}\|_{\mathcal{H}}\right]\leq 2\mathcal{R}(\mathcal{H})
$$

所以如果我们能上下 bound 住 Rademacher 复杂度，比如比较好的情况是 $\mathcal{R}(\mathcal{H})=o(1)$，那么自然 risk 就会被控制到 0，进一步 $\|\mathbb{P}_n-\mathbb{P}\|_{\mathcal{H}}$ 如果是 $\frac{b}{n}$-bounded differnce（容易满足的，比如只要 $h\in\mathcal{H}$ 有界）

$$
\mathbb{E}\left[\|\mathbb{P}_n-\mathbb{P}\|_{\mathcal{H}}\right]-\|\mathbb{P}_n-\mathbb{P}\|_{\mathcal{H}}\leq\delta,\quad w.p.\geq 1-\exp\left(-\frac{n\delta^2}{b^2}\right)
$$

那么我们就能控制 $\|\mathbb{P}_n-\mathbb{P}\|_{\mathcal{H}}$

# 5 Sparse Regression

## 5.1 OLS Regression

回归中最为经典的情形。过往情况下我们都要求 $n>d$，最好能是 $\gg$，以此来保证 $X^TX$ 是可逆的，进而有唯一解。渐进统计中我们已经有了很多关于 OLS 的结果，比如在高斯误差下

$$
\mathbb{E}\left[\text{MSE}\right]=\mathbb{E}\left[\frac{1}{n}\left\|\hat{Y}-Y\right\|_2^2\right]=\sigma^2\frac{d}{n}
$$

（或者在 $X'X$ 不满秩的情况下 $r<d$ 将 $d$ 替换为 $r$）。

我们可以用高维的思路 recover 这个问题，并顺便将 error 拓展到更一般的 subGau 情况。具体来说：对于预测问题

$$y = x'\theta^* + \varepsilon, \quad \varepsilon \sim \text{SubGau}(\sigma^2)$$

的解

$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \|Y - X\theta\|^2$$

我们希望研究 $\mathbb{E}[\text{MSE}] = \mathbb{E}\left[\frac{1}{n}\left\|\hat{Y} - Y\right\|_2^2\right] \lesssim ?$.(这里 $Y \in \mathbb{R}^n$ 指数据)

思路如下: 首先将 $\theta^*$ 和 $\hat{\theta}$ 联系起来（即将真实情况 $\varepsilon$ 和估计情况 $\hat{\varepsilon}$ 联系起来）：$\hat{\theta}$ 作为 MSE 的 minimizer，我们有

$$\|\hat{\varepsilon}\| = \left\|Y - X\hat{\theta}\right\|_2^2 \leq \|Y - X\theta^*\|_2^2 = \|\varepsilon\|_2^2$$

展开 $Y = X\theta^* + \varepsilon$ 整理，并记 $\hat{\Delta} = \hat{\theta} - \theta^*$

$$\begin{aligned}
\left\|X\hat{\Delta}\right\|_2^2 \leq & 2\varepsilon' X\hat{\Delta} \\
= & 2\left\|X\hat{\Delta}\right\|_2 \frac{\varepsilon' X\hat{\Delta}}{\left\|X\hat{\Delta}\right\|_2} \\
\Rightarrow \left\|X\hat{\Delta}\right\|_2 \leq & 2\frac{\varepsilon' X\hat{\Delta}}{\left\|X\hat{\Delta}\right\|_2} \\
\leq & 2 \sup_{u \in \mathbb{S}_{n-1}^{r-1}} \varepsilon' u
\end{aligned}$$

其中由于 $X$ 是秩 $r$ 的，所以 $u$ 只能在 $r$ 维度子空间和 $\mathbb{S}^{n-1}$ 的交里。这样我们就有

$$\mathbb{E}\left[\left\|X\hat{\Delta}\right\|_2^2\right] \leq 4\mathbb{E}\left[\sup_{u \in \mathbb{S}_{n-1}^{r-1}} (\varepsilon' u)^2\right] = 4\mathbb{E}\left[\text{subGau}(r\sigma^2)\right]$$

而 subGau 自动包含了对所有 $p$ 阶矩的 bound，所以我们有

$$\begin{aligned}
\mathbb{E}[\text{MSE}] = & \mathbb{E}\left[\frac{1}{n}\left\|X\hat{\Delta}\right\|_2^2\right] \lesssim \sigma^2 \frac{r}{n} \\
\mathbb{P}(\text{MSE} \geq t) \leq & \exp\left(-\frac{nt}{4r\sigma^2}\right)
\end{aligned}$$

顺便我们能得到 $\hat{\Delta}$ 的尺度估计

$$\left\|\hat{\Delta}\right\|_2^2 \leq \frac{\left\|X\hat{\Delta}\right\|_2^2}{\lambda_{\min}(X'X)}$$

一些 Conherence 条件能够改善这个 bound。

## 5.2 Regression with Constraints

对于高维情况，比如可能会有 $d \sim n$ 甚至 $d > n$ 时，我们很难得到或得不到唯一解，但有时我们对问题会有一些先验知识，体现为一些“对解的结构的假设”，典型且实用的一种是稀疏性（sparsity）。这时我们可以将问题转化为一个带约束的优化问题

$$\hat{\theta} \in \arg\min_{\theta \in K} \|Y - X\theta\|^2, \quad K \text{ is some sparisty constraint set, e.g. } \begin{cases} \|\theta\|_0 \le k, & \text{L0 constraint} \\ \|\theta\|_1 \le R, & \text{L1 constraint} \end{cases}$$

我们可以用和 OLS 类似的方法来 bound MSE：由于 $K$ 是对称的，我们有

$$\left\|X\hat{\Delta}\right\|_2^2 \le 2\varepsilon' X(\hat{\theta} - \theta) \le 2 \sup_{\hat{\theta}, \theta \in K} \varepsilon' X(\hat{\theta} - \theta) = 2 \sup_{\theta \in 2K} \varepsilon' X\theta = 4 \sup_{v \in XK} \varepsilon' v$$

这样我们注意到

$$\begin{aligned} \mathbb{E}\left[\frac{1}{n}\left\|X\hat{\Delta}\right\|_2^2\right] \le & \frac{4}{n}\mathbb{E}\left[\sup_{v \in XK} \varepsilon' v\right] = \frac{4\sigma}{n}\mathcal{G}(XK) \\ \mathbb{P}\left(\frac{1}{n}\left\|X\hat{\Delta}\right\|_2^2 \ge t\right) \le & \mathbb{P}\left(\frac{4}{n}\sup_{v \in XK} \varepsilon' v \ge t\right) \end{aligned}$$

恰好是 Gaussian Complexity 的形式，所以我们可以相应集合的 Gaussian Complexity 来 bound 这个 MSE。由于 $K$ 是预先给定的行为很好的“球”，所以一个影响 bound 效果的重要因素就是 $X$ 的行为。这里我们先只限制 $X$ 的列的范数，即 $\|X_j\|_2 \le \sqrt{n}$

- L1 Constraint: 这个情况下 $XK_1$ 是一个 polytope，所以我们只需研究其顶点就足够决定 $XK_1$ 的 Gaussian Complexity。

  $$\mathcal{G}(XK_1) = \mathcal{G}(\text{vertices of } XK_1) \lesssim R\sqrt{n \log 2d}$$

  这样我们就有

  $$\mathbb{E}\left[\text{MSE}_{\ell_1}\right] \lesssim \sigma R \frac{\sqrt{\log d}}{\sqrt{n}}$$

  以及

  $$\begin{aligned} \mathbb{P}\left(\text{MSE}_{\ell_1} \ge t\right) \le & \mathbb{P}\left(\frac{4}{n}\sup_{v \in XK_1} \varepsilon' v \ge t\right) \\ = & \mathbb{P}\left(\frac{4}{n}\exists_{v \in \text{vertices of } XK_1} \varepsilon' v \ge t\right) \\ \le & 2d\mathbb{P}\left(\frac{4}{n}R\sqrt{n}\mathcal{N}(0, \sigma^2) \ge t\right) \\ \le & 2d\exp\left(-\frac{nt^2}{16R^2\sigma^2}\right) \end{aligned}$$

  then set $2d\exp\left(-\frac{nt^2}{16R^2\sigma^2}\right) = \delta$ we have

  $$w.p. \ge 1 - \delta, \quad \text{MSE}_{\ell_1} \lesssim \sigma R \frac{\sqrt{\log d/\delta}}{\sqrt{n}}$$

- L0 Constraint：见 Stat461-2023Fall exam Q3，有

$$\mathbb{E}\left[\mathrm{MSE}_{\ell_0}\right] \lesssim \sigma^2 k/n$$

关于 MSE 的 bound 推导见 Stat461-2023Fall exam Q3. 结论为

$$w.p. \geq 1-\delta, \quad \mathrm{MSE}_{\ell_0} \lesssim \frac{\sigma^2}{n}\left(k + k\log\frac{d}{k} + \log\frac{1}{\delta}\right)$$

这里我们能看到：L1 的 bound 更差一些，但是是一个凸优化问题，容易求解；而 L0 的 bound 更好，但是是一个 NP-hard 问题，难以求解。总而言之收敛速度的 bound 总结如下：

| Model | $1-\delta$ MSE Bound |
|---|---|
| OLS | $\lesssim \frac{\sigma^2}{n}\left(r + \log\frac{1}{\delta}\right)$ |
| L1 $\|\theta\|_1 \leq R$ | $\lesssim \sigma R\sqrt{\frac{\log d/\delta}{n}}$ |
| L0 $\|\theta\|_0 \leq k$ | $\lesssim \frac{\sigma^2}{n}\left(k + k\log\frac{d}{k} + \log\frac{1}{\delta}\right)$ |

# 6 Random Matrix

# 7 Minimax Risk

更多内容见9.3.

## 7.1 Problem Formulation

对于一批假设分布类 $\mathcal{P} = \{\mathbb{P}\}$，我们感兴趣于估计某个所谓的“参数”

$$\theta(\mathbb{P}) : \mathcal{P} \mapsto \Theta$$

比如对一族随机矩阵估计其谱范数，对一族分布估计其均值之类。我们的数据是从某个真值 $\mathbb{P}^*$ i.i.d. 抽取的，并有某种方法来估计这个参数

$$\hat{\theta} : \mathcal{X}^n \mapsto \Theta$$

并有某个（伪）metric 来衡量估计的好坏

$$\rho(\,\cdot\,,\,\cdot\,) : \Theta \times \Theta \mapsto [0, \infty)$$

那么如何来认为我们得到了一个好的 estimator 呢？Minimax risk 陈述的是这样一个标准：

$$\mathfrak{M}\big(\theta(\mathcal{P}); \rho\big) := \inf_{\hat{\theta}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}\left[\rho\big(\hat{\theta}, \theta(\mathbb{P})\big)\right]$$

即：在所有可能的估计器中（$\inf_{\hat{\theta}}$），最坏的情况下（$\sup_{\mathbb{P}\in\mathcal{P}}$）的 expected risk 最小的。这个量是一个很好的标准，因为它告诉我们在通过合理选取估计器后，我们至少能达到的 risk。

## 7.2 From Minimax Risk to Testing

这里我们再次涉及 ranging over 一个无穷集合的极值问题，思路和 upper bound side 是一样的：挑选其中的代表元素，不过这次使用的则是 packing number。在有了有限个 $M := M(2\delta, \mathbb{P}, \cdot)$ 个代表元素后，下一步方法是通过 Markov 不等式转化为一些 probability 的 bound，这时问题就很像是对一个有限集合的分类问题了（分类到由 packing 选出的 $M$ 个代表性参数上）。

具体如下：构造 $M(2\delta, \mathbb{P}, \cdot)$ packing $\{\theta_j\}_{j=1}^M$

$$\begin{aligned}\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}\left[\rho\big(\hat{\theta}, \theta(\mathbb{P})\big)\right] \geq & \sup_{\mathbb{P}\in\mathcal{P}} \delta\mathbb{P}\left(\rho\big(\hat{\theta}, \theta(\mathbb{P})\big) \geq \delta\right) \\ \geq & \frac{\delta}{M}\sum_{j=1}^{M} \mathbb{P}_{\theta_j}\left(\rho\big(\hat{\theta}, \theta_j\big) \geq \delta\right) \\ := & \delta\mathbb{Q}[\rho\big(\hat{\theta}, \theta_J\big) \geq \delta]\end{aligned}$$

其中的 $\mathbb{Q}$ 类似于是一个“均匀分布”在 $\{\theta_j\}_{j=1}^M$ 上的分布。对于任意的 $\hat{\theta}(\cdot)$ 我们可以构造一个 testing（或者说判别）

$$\hat{\psi}(\cdot) := \arg\min_{j\in[M]} \rho(\hat{\theta}, \theta_j)$$

这样我们有 $\rho(\text{not } \hat{\psi}(\cdot), \hat{\theta}) \geq \delta$（$\hat{\theta}$ 离未被判别中的点足够远），那么 $\{\rho(\hat{\theta}, \theta^{\psi}) < \delta\} \subset \{\psi(\cdot) = \hat{\psi}(\cdot)\}$，即

$$\begin{aligned}\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}\left[\rho\big(\hat{\theta}, \theta(\mathbb{P})\big)\right] \geq & \delta\mathbb{Q}[\rho\big(\hat{\theta}, \theta_J\big) \geq \delta] \\ \geq & \delta\mathbb{Q}[\hat{\psi}(\cdot) \neq \psi(\cdot)], \qquad \cdot \text{ is the data from } \mathbb{P}_J\end{aligned}$$

这里的 $\mathbb{Q}[\hat{\psi}(\cdot) \neq \psi(\cdot)]$ 看起来就像是一个 $M$ 类判别问题的错误率。Intuitively，较小的 $\delta$ 会导致有更多类 $M$ 需要处理，类与类之间也更接近，所以错误率会上升，所以我们需要进一步对 $\delta$ optimize 来得到最优的 lower bound, 这个过程就是通过 packing number 来完成的。自动是一个 variance-bias trade-off 的过程。

## 7.3 K-L Divergence Method

这里介绍基于信息论的方法。我们的问题是：在观测数据 $Z \sim \mathbb{Q} = \frac{1}{M}\sum_{j=1}^M \mathbb{P}_{\theta_j}$ 情况下，用 $\hat{\psi} = \arg\min_{j\in[M]} \rho(\hat{\theta}, \theta_j)$ 进行判别的错误率（的 lower bound）。如果用隐变量模型的角度来说就是：数据是 $(D, J)$，其中 $D|J \sim \mathbb{P}_{\theta_J}$，但我们只能观察到 $Z$ 的 marginal。而如果 $\mathbb{P}_Z$ 和 $\mathbb{J}$ 越互相独立，我们就会越难区分 $J$，这个“越互相独立”可以用互信息衡量：

$$I(Z, J) := \mathrm{KL}\left(\mathbb{Q}_{Z,J} \| \mathbb{Q}_Z\mathbb{Q}_J\right)$$

如果 $Z \perp\!\!\!\perp J$ 我们就会有 $I(Z, J) = 0$。

对于我们的情况，$J \sim \mathrm{Unif}[M]$，互信息可以写成

$$I(Z, J) = \frac{1}{M}\sum_{j=1}^{M} \mathrm{KL}\left(\mathbb{P}_{\theta_j} \| \mathbb{Q}_Z\right)$$

且可以实用如下的 Fano's inequality：

$$\mathbb{P}\left(\hat{\psi}(Z)=J\right)\leq\frac{I(Z,J)+\log 2}{\log M}$$

这直接给出了 lower bound:

$$\begin{aligned}\mathfrak{M}\big(\theta(\mathcal{P});\rho\big)=&\inf_{\hat{\theta}}\sup_{\mathbb{P}\in\mathcal{P}}\mathbb{E}_{\mathbb{P}}\left[\rho\big(\hat{\theta},\theta(\mathbb{P})\big)\right]\\=&\inf_{\hat{\theta}}\delta\mathbb{Q}[\rho\big(\hat{\theta},\theta_J\big)\geq\delta]\\\geq&\delta\big\{1-\frac{I(Z,J)+\log 2}{\log M}\big\}\\=&\delta\big\{1-\frac{\frac{1}{M}\sum_{j=1}^{M}\mathrm{KL}(\mathbb{P}_j\|\mathbb{Q})+\log 2}{\log M}\big\}\end{aligned}$$

进一步只需：1) upper bound $I(Z,J)$; 2) 选择合适的 $\delta$，即相应的 $2\delta$ packing.

未完成，余下部分转移到9.3进行了。

# 8 Non-Parametric LS

非参数回归聚焦的是 predictor 本身：

$$\mathcal{L}_f=\mathbb{E}_{X,Y}\left[(Y-f(X))^2\right]$$

其中的 $f$ 是某种非线性的函数，理论最优是 Bayes predictor$f^*(x)=\mathbb{E}\left[Y|X=x\right]$，所以我们使用的 loss 也变为 risk w.r.t. $f^*$：

$$\mathcal{L}_f-\mathcal{L}_{f^*}=\mathbb{E}_{X,Y}\left[(Y-f(X))^2\right]-\mathbb{E}_{X,Y}\left[(Y-f^*(X))^2\right]=\mathbb{E}_X\left[(f(X)-f^*(X))^2\right]:=\|f^*-f\|^2_{L^2(\mathbb{P})}$$

估计值则使用 $\mathbb{P}_n=\frac{1}{n}\sum_{i=1}^{n}\delta_i$:

$$\hat{L}_f-\hat{L}_{f^*}=\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i)-f^*(x_i)\right)^2:=\|f-f^*\|^2_{L^2(\mathbb{P}_n)}$$

（简化起见也写作 $\|f-f^*\|_2^2$ 和 $\|f-f^*\|_n^2$.）

$\hat{f}$ 的求解是在一个 RKHS 的子集中 minimize empirical risk，即

$$\hat{f}\in\arg\min_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-f(X_i)\right)^2+\lambda_n\|f\|^2_{\mathscr{F}}$$

与前面 ULLN 一节相似，estimate loss$\|f-f^*\|_n^2$ 的行为也与 $\mathscr{F}:=\{f-f^*\}$ 的 complexity 有关，这一节我们要研究一个进阶对象：localized form of complexity，以 Gaussian complexity 为例：

$$\mathcal{G}_n(\delta;\mathscr{F}^*):=\mathbb{E}_{\varepsilon\sim\mathcal{N}(0,I)}\left[\sup_{g\in\mathscr{F}^*,\|g\|_n\leq\delta}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i g(x_i)\right|\right]$$

也即 $\mathscr{F}^*(\vec{x})$ 在 $f^*$ 附近的[3]的复杂度。进一步只要我们这个“附近”（$\delta$）尺寸取得合适，我们就能只用 $\mathscr{F}^*$ 的局部信息来 bound 整体的 complexity。这个“合适”依赖于两点：

[3]in the sense that $\|g\|_n\leq\delta$

- 1. $\mathscr{F}^*$ 的行为不能太差，具体来说是 star shape around 0: 一个稍微弱于 convex 的条件，即 $\forall g \in \mathscr{F}^*, \lambda \in [0,1], \lambda g \in \mathscr{F}^*$. 这个条件的作用是下面 point 2 的 $\delta$ 存在，笼统来说我感觉这个性质是确保 global 上远端不会有太差的行为导致用 localize form 的 complexity bound 不佳。
- 2. 在我们实际估计的时候，$\hat{g} = \hat{f} - f^*$ 的尺涨落是由模型 $Y_i = f(X_i) + \sigma\varepsilon_i$ 中的误差 $\varepsilon_i \sim \mathcal{N}(0,1)$引起的，所以笼统来说只要 $\delta$ 大于某种意义上 $\varepsilon_i$ 引起的尺度涨落即可包括足够多的 $g$。这个“合适”具体如下：

  注意到 $\sum(y_i - \hat{f}(x_i))^2 \leq \sum(y_i - f^*(x_i))^2$，得到

  $$\frac{1}{2}\left\|\hat{f} - f^*\right\|_n^2 \leq \frac{\sigma}{n}\sum_{i=1}^n \varepsilon_i(\hat{f}(x_i) - f^*(x_i))$$

  由于我们研究 $\|g\|_n \leq \delta$ 区域，所以笼统来说对 $\delta$ 我们期望有：[4]

  $$\frac{\delta^2}{2} \geq \sigma\mathcal{G}_n(\delta, \mathscr{F}^*)$$

  为 $\delta$ 应该满足的条件。（这个 justify 不严格，只感受其直觉即可）

  进一步对于 $\mathcal{G}_n(\delta, \mathscr{F}^*)$ 可以应用 metric entropy 来 bound（Dudley size），就能估出一个可用的 $\delta_n$，见 MJW P426.

更严格来说，$\delta_n^2/2 \geq \sigma\mathcal{G}_n(\delta, \mathscr{F}^*)$ 确保如下的 bound（MJW Thm 13.5）：

$$\mathbb{P}\left(\left\|\hat{f} - f^*\right\|_n^2 \geq 16t\delta_n\right) \leq \exp\left[-\frac{nt\delta_n}{2\sigma^2}\right], \quad \forall t \geq \delta_n$$

或者通过取定 $t = \delta_n$ 得到如下形式：

$$\left\|\hat{f} - f^*\right\|_n^2 \lesssim \delta_n^2, \quad w.p.1 - \exp\left[-\frac{n\delta_n^2}{2\sigma^2}\right]$$

## 8.1 Oracle Inequality

更进一步，如果 $f^* \notin \mathscr{F}$ 那么我们其实只能在 $\mathscr{F}$ 中找到一个 $f^*$ 的“投影”，最后的估计误差也由两部分组成，具体来说由下面的 bound 保证（MJW P433 Thm 13.13）：对于 $\delta_n^2 \geq 2\sigma\mathcal{G}_n(\delta, \mathscr{F} - \mathscr{F})$, 取任意 $t \geq \delta_n$ 有

$$\left\|\hat{f} - f^*\right\|_n^2 \leq \inf_{\gamma\in[0,1]} \frac{1+\gamma}{1-\gamma}\|f - f^*\|_n^2 + \frac{c_0}{\gamma(1-\gamma)}t\delta_n, \quad \forall f \in \mathscr{F}, \quad w.p.1 - c_1\exp\left[-c_2 nt\delta_n/\sigma^2\right]$$

更详细来说，取定 $t = \delta_n$ 并对 $\gamma$ 优化得到

$$\left\|\hat{f} - f^*\right\|_n^2 \lesssim \underbrace{\inf_{f\in\mathscr{F}}\|f - f^*\|_n^2}_{\text{approximation error}} + \underbrace{\delta_n^2}_{\text{estimation error}}$$

容易发现如果 $f \in \mathscr{F}$ 的话，上面的 bound 自然退化成 $\left\|\hat{f} - f^*\right\|_n^2 \lesssim \delta_n^2$。

**Example**: $k$-sparse 的估计误差前面给出过，这里对于 oracle 情形则是多出一项 approximation error:

$$\left\|\hat{f}_{k \text{ sparse}} - f^*\right\| \lesssim \inf_{k \text{ sparse } f}\|f - f^*\|_n + \underbrace{\frac{\sigma^2}{n}\left(k\log\frac{ed}{k} + \log\frac{1}{\delta}\right)}_{\delta_n^2}$$

[4] 实际计算的时候我们一般用 $\delta/2\sigma \geq \mathcal{G}_n(\delta, \cdot)/\delta$，这时不等式右边类似于是“$\delta$ 范围的平均复杂度”（本质上是因为 star-shape 保证了 $\mathscr{F}$ 的尺度是 $O(\delta)$ 的）

# 9 Review of Information Theory and Related Bounds

本部分主要基于 YW。

## 9.1 $f$-Divergence

$f$-divergence（散度）是一类衡量两个分布之间差异的方法（值得注意的是一般而言散度不是对称的，所以不能称之为“距离”），一种常见的 divergence 是广为熟知的 KL 散度。Intuitively, 散度越高，两个分布的差异越大，也就越难以将之区分开，进而我们可以用散度构建一些 bound。

最 general 的散度定义是：

$$D_f(P\|Q) := \mathbb{E}_Q\left[f(\frac{\mathrm{d}P}{\mathrm{d}Q})\right]$$

其中 $f$ 需要满足一个正则化条件：$f(1)=0$ 且 $f$ 在 1 处强凸。真正需要计算的时候我们会借用一个 dominating measure $\mu$，写成大家熟知的基于密度的形式

$$D_f(P\|Q) = \mathbb{E}_Q\left[f(\frac{\mathrm{d}P/\mathrm{d}\mu}{\mathrm{d}Q/\mathrm{d}\mu})\right] = \int f(\frac{\mathrm{d}P/\mathrm{d}\mu}{\mathrm{d}Q/\mathrm{d}\mu})\frac{\mathrm{d}Q}{\mathrm{d}\mu}\,\mathrm{d}\mu = \int f(\frac{p}{q})q\,\mathrm{d}\mu$$

► “Not all $f$-divergences are born equal”

$f$ 的不同取法导致不同的散度，而不同的任务其实会 boil down to 不同的散度，同时不同的散度之间会有彼此 bound 的关系，所以：我们可以针对不同的任务选取特定的（好计算的/性质的好的）散度来构建 bound。下面是四个常见的散度：

- KL 散度：$f(t)=t\log t$

$$D(P\|Q) := D_{KL}(P\|Q) = \mathbb{E}_Q\left[\frac{P}{Q}\log\frac{P}{Q}\right] = \mathbb{E}_P\left[\log\frac{P}{Q}\right]$$

- 全变差（TV）散度：$f(t)=\frac{1}{2}|t-1|$，这个是对称的

$$d_{\mathrm{TV}}(P,Q) := \frac{1}{2}\int|p-q|$$

- $\chi^2$ 散度：$f(t)=(t-1)^2$，容易 tensorize

$$\chi^2(P\|Q) := \mathbb{E}_Q\left[\frac{(P-Q)^2}{Q^2}\right] = \int\frac{p^2}{q}\,\mathrm{d}\mu - 1$$

- Hellinger 散度：$f(t)=(\sqrt{t}-1)^2$，这个也是对称的，且容易 tensorize

$$H^2(P,Q) := \int(\sqrt{p}-\sqrt{q})^2\,\mathrm{d}\mu$$

它可以导出一个常用量：

$$1-\frac{1}{2}H^2(P,Q) = \int\sqrt{pq}\,\mathrm{d}\mu$$

► 为什么 $f$ 散度能度量信息量？

- 考虑如下的“条件散度”：

$$D_f\left(P_{Y|X}\|Q_{Y|X}\middle|P_X\right) := \mathbb{E}_{X\sim P_X}\left[D_f\left(P_{Y|X}\|Q_{Y|X}\right)\right]$$

具有如下性质：

$$D_f\left(P_{Y|X}\|Q_{Y|X}\right) \le D_f\left(P_{Y|X}\|Q_{Y|X}\middle|P_X\right)$$

即有了“额外的”信息 $P_X$ 后，$D_f$ 上升

- 考虑某种 channel model：将 $X$ 通过 $P_{Y|X}$ 生成 $Y$，具体来说，能够将 $P_X$ 随机变量生成为 $P_Y$, similarly 将 $Q_X$ 生成为 $Q_Y$，那么有

$$D_f(P_X\|Q_X) \ge D_f(P_Y\|Q_Y) \tag{9.1}$$

即通过了信息 channel 后，信息减少，$D_f$ 下降。

**Remark**: 这个性质的一个延申是：对于任意给定的 region $E$，我们总能诱导出关于 $P$，$Q$ 的 Bernoulli 分布 $P(E) = \mathrm{Id}_{P\in E}$，$Q(E) = \mathrm{Id}_{Q\in E}$，这相当于是一种 channel，所以自然的结果是

$$D_f(P\|Q) \ge D_f(P(E)\|Q(E))$$

而 Bernoulli 是一类比较好研究的随机变量，我们可以通过研究 Bernoulli 变量 + 遍历/optimize over 各种 $E$ 来研究 $D_f$ 的 bound。

### 9.1.1 Relation between $f$-Divergence and Other Properties

一些常用的，$f$-divergence 之间的关系：

- Sandwiched between TV and Heillinger:

$$0 \le \frac{1}{2}H^2(P,Q) \le d_{\mathrm{TV}}(P,Q) \le H(P,Q)\sqrt{1-\frac{H^2(P,Q)}{4}} \le 1$$

- Pinsker inequality:

$$d_{\mathrm{TV}}(P,Q) \le \sqrt{\frac{1}{2}D_{KL}(P\|Q)}$$

- Tensorization of $\chi^2$ and Hellinger:

$$\begin{aligned}\chi^2\left(\prod_{i=1}^n P_i, \prod_{i=1}^n Q_i\right) &= \prod_{i=1}^n\left(1+\chi^2(P_i,Q_i)\right)-1\\ H^2\left(\prod_{i=1}^n P_i, \prod_{i=1}^n Q_i\right) &= 2-2\prod_{i=1}^n\left(1-\frac{H^2(P_i,Q_i)}{2}\right)\end{aligned}$$

### 9.1.2 Mutual Information

实际上信息量本身并不是重要的，信息如何在 processing 过程中流动才是。引入 Mutual Information 就是用来描述这种情况。

具体来说，我们有两个随机变量 $X, Y$，他们的联合分布是 $P_{XY}$, 边缘分布分别是 $P_X$ 和 $P_Y$，那么他们的 Mutual Information 定义为：

$$I(X;Y) := D_{KL}(P_{XY} \| P_X P_Y)$$

i.e. 联合分布与边缘分布的 KL 散度。

它有几种等价定义（或者称之为性质/理解方法）：

- $Y|X$（或 vice versa）这个 channel 的信息：

$$I(X;Y) = D_{KL}(P_{Y|X} \| P_Y \big| P_X) = D_{KL}(P_{X|Y} \| P_X \big| P_Y)$$

- 信息流对称：

$$I(X;Y) = I(Y;X)$$

- 熵的差：同样是反映 channel 的内涵

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X), \qquad H(\xi) := \mathbb{E}\left[-\log f(\xi)\right]$$

- 优化视角：

$$I(X;Y) = \inf_{Q} D_{KL}(P_{Y|X} \| Q \big| P_X)$$

既然 Mutual Information 是一个 channel 的信息，那么对于一个 channel 的 pipeline 我们自然会有单调关系。具体来说我们 formalize 成一个 Markov chain $X \leftrightarrow Y \leftrightarrow Z$，那么有

$$I(X;Z) \le I(X;Y)$$

这个结果是很自然的，通过 Markov chain 这种“概率化”的 channel 传递时，信息会减少。

**Note:** 基于 KL 散度的互信息是最常见的版本，但当然也可以定义 general 的 $f$-divergence 的互信息：

$$I_f(X;Y) := D_f(P_{Y|X} \| P_Y \big| P_X)$$

## 9.2 Review of Minimax Theory

前面已经提到过 Minimax risk 的定义

$$\mathfrak{M}(\theta(\mathcal{P}); \rho) := \inf_{\hat{\theta}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}} \left[ \rho(\hat{\theta}, \theta(\mathbb{P})) \right]$$

这里我们直接采用一个更具体的版本：对于检验一个参数类 $\Theta$，以及使用损失函数 $\rho = \ell$，即

$$\mathfrak{M}(\Theta; \ell) := \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_{\theta} \left[ \ell(\hat{\theta}, \theta) \right]$$

Minimax risk 有几种 trivial 的 bound：

- 直接研究一个特定的估计量得到 Minimax risk 的 upper bound，例如对于高斯均值 $X_i \overset{i.i.d.}{\sim} N(\mu, I_p\sigma^2)$，直接固定研究 $\hat{\mu} = \bar{X}$，得到

$$\mathfrak{M} \leq \sup_{\mu\in\mathbb{R}} \mathbb{E}_\mu \left[\ell(\hat{\mu} = \bar{X}, \mu)\right] = \frac{p\sigma^2}{n}$$

- 缩小参数类得到 Minimax risk 的 lower bound，例如本节即将介绍的 Le Cam's method：直接将参数空间缩小到两点，然后就能够和 $d_{\mathrm{TV}}$ 挂钩。
- Worst case Bayes risk. 定义为

$$R_B^* := \sup_{\pi} \inf_{\hat{\theta}} \mathbb{E}_{\theta\sim\pi} \left[\ell(\hat{\theta}, \theta)\right]$$

即：对于每个 $\pi$ 找到最优的估计量 $\hat{\theta}$，由于对于每个 $\pi$ 得到的 Bayes risk$R_\pi$ 都是在先验均值意义上的 risk，所以有

$$\mathfrak{M} \geq R_\pi, \quad \forall \pi$$

遍历所有参数先验 $\pi = \pi_\theta$[5]最后取最大值就是 Worst case Bayes risk，自然构成了 Minimax risk 的一个 lower bound

$$\mathfrak{M} \geq \sup_{\pi} R_\pi = R_B^*$$

另一种理解方式就是 $\inf\sup \geq \sup\inf$.

**Example:** $d$ 维 OLS 的 Minimax risk 的一个 bound 可以被精确地计算出来，为

$$\mathfrak{M}_{\mathrm{OLS}(n,d,\mathbf{X})}(\mathbb{R}^d; \|\cdot\|_2) = \sigma^2 \frac{\mathrm{rank}(\mathbf{X})}{n}$$

但是对于更多/更复杂的情况，我们很难精确计算出 Minimax bound，(而且 minimax bound) 的具体值与 $\ell$ 的尺度变换有关，所以经常我们会关心下面这种情况：

**Goal:** 对于 Minimax bound, i.e. 我们构建的统计量“最好情况下”能将“最坏情况的”risk bound 住，我们希望研究它的rate，也就是

$$\mathfrak{M}(\Theta; \ell) \asymp \text{ func of } (\dim(\Theta), N, \text{ structure params, etc.})$$

这要求我们分别研究 upper bound 和 lower bound，Hopefully 如果两个 bound 是 match 的，我们就能对 Minimax risk 的趋势有一个比较好的理解。

## 9.3 Minimax Lower Bound by Le Cam's Method

一般来说在 Minimax 的 bound 中我们经常关心的问题是 lower bound，因为 minimax lowed bound 可以自然制作一个：

$$R(\hat{\theta}) \geq \mathfrak{M}(\Theta; \ell) \geq \text{minimax lower bound}$$

[5]在实操上只要取一族参数先验一般就能得到一个不错的 bound，不一定要取便整个先验假设类。

从而得到某个估计量的 risk 下界。结合前面关于 bound$\mathbb{E}\left[\|\mathbb{P}_n - \mathbb{P}\|\right]$ 的讨论，其实我们就能够得到对某个给定估计量的好坏的两边的 bound 了。

得到 Minimax lower bound 的基本思路就是

$$\mathfrak{M}(\Theta;\ell) = \inf_{\hat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_\theta \left[\ell(\hat{\theta},\theta)\right] \geq \inf_{\hat{\theta}} \sup_{\theta\in\Theta'} \mathbb{E}_\theta \left[\ell(\hat{\theta},\theta)\right]$$

其中 $\Theta' \subseteq \Theta$ 一般是某种好计算的情况，我们就能够算出 lower bound 了。

### 9.3.1 Le Cam's Two-Point Method

Le Cam 是最为简单粗暴的，即用上述的第二条，直接将参数空间变成两点，用假设检验的视角来解决。

具体来说，我们研究将参数空间缩小到 $\tilde{\Theta} := \{\theta_0, \theta_1\}$，这样得到

$$\mathfrak{M}(\Theta) \geq \mathfrak{M}(\tilde{\Theta}) = \inf_{\hat{\theta}} \sup_{\theta\in\tilde{\Theta}} \mathbb{E}_\theta \left[\ell(\hat{\theta},\theta)\right]$$

并对损失函数施加一个 $\alpha$-triangle inequality，即

$$\ell(\theta_0,\theta_1) \leq \alpha\left(\ell(\theta_0,\theta) + \ell(\theta,\theta_1)\right)$$

在这种情况下，$\ell$ 能自然诱导出一个最优检验 depending on $\ell(\hat{\theta},\theta_0) - \ell(\hat{\theta},\theta_1)$，用这个检验得到

$$\begin{aligned}\mathfrak{M}(\tilde{\Theta}) =& \inf_{\theta\in\tilde{\Theta}} \sup_{\hat{\theta}} \mathbb{E}_\theta \left[\ell(\hat{\theta},\theta)\right] \\ =& \mathbb{E}_{\theta_0}\left[\ell(\hat{\theta},\theta_1)\right] \vee \mathbb{E}_{\theta_1}\left[\ell(\hat{\theta},\theta_0)\right]\end{aligned}$$

然后用 Markov 不等式链接期望和概率，然后再联系到 $d_{\mathrm{TV}}$ 即可得到 Le Cam's method 的 bound:

$$\begin{aligned}1 - d_{\mathrm{TV}}(\mathbb{P}_{\theta_0},\mathbb{P}_{\theta_1}) \leq& \mathbb{P}_{\theta_0}(\text{test gives } \theta_1) + \mathbb{P}_{\theta_1}(\text{test gives } \theta_0) \\ =& \mathbb{P}_{\theta_0}\left(\ell(\hat{\theta},\theta_0) \geq \ell(\hat{\theta},\theta_1)\right) + \mathbb{P}_{\theta_1}\left(\ell(\hat{\theta},\theta_1) \geq \ell(\hat{\theta},\theta_0)\right) \\ \leq& \mathbb{P}_{\theta_0}\left(\ell(\hat{\theta},\theta_0) \geq \frac{\ell(\theta_0,\theta_1)}{2\alpha}\right) + \mathbb{P}_{\theta_1}\left(\ell(\hat{\theta},\theta_1) \geq \frac{\ell(\theta_0,\theta_1)}{2\alpha}\right) \\ \leq& \frac{2\alpha}{\ell(\theta_0,\theta_1)}\left(\mathbb{E}_{\theta_0}\left[\ell(\hat{\theta},\theta_1)\right] + \mathbb{E}_{\theta_1}\left[\ell(\hat{\theta},\theta_0)\right]\right) \\ \leq& \frac{4\alpha}{\ell(\theta_0,\theta_1)}\left(\mathbb{E}_{\theta_0}\left[\ell(\hat{\theta},\theta_1)\right] \vee \mathbb{E}_{\theta_1}\left[\ell(\hat{\theta},\theta_0)\right]\right) \\ =& \frac{4\alpha}{\ell(\theta_0,\theta_1)}\mathfrak{M}(\tilde{\Theta})\end{aligned}$$

于是得到了Le Cam's two-point method的 bound：

$$\begin{aligned}&\mathfrak{M}(\Theta) \geq \mathfrak{M}(\tilde{\Theta}) \geq \frac{\ell(\theta_0,\theta_1)}{4\alpha}\left(1 - d_{\mathrm{TV}}(\theta_0,\theta_1)\right), \quad \forall\theta_0,\theta_1 \in \Theta \\ \Rightarrow\ &\mathfrak{M}(\Theta) \geq \sup_{\theta_0,\theta_1\in\Theta} \frac{\ell(\theta_0,\theta_1)}{4\alpha}\left(1 - d_{\mathrm{TV}}(\theta_0,\theta_1)\right)\end{aligned}$$

▶ 关于 $\alpha$ 的取法

对于常用的平方损失 $\ell(\theta,\theta') = \|\theta - \theta'\|_2^2$, $\alpha = 2$.

▶ Le Cam two-point on Bayesian setting

Le Cam 方法的核心是将参数空间缩小到两点。基于这个哲学，我们也可以利用 Bayesian risk 来做 bound。具体来说，**使用平方损失时**，我们采用先验类 $\{\theta_0, \theta_1\}$ 上的 Bernoulli，并得到取等概率时的 Bayesian risk bound

$$\mathfrak{M}(\Theta) \geq \mathfrak{M}(\tilde{\Theta}) \geq \frac{\|\theta_0 - \theta_1\|_2^2}{4} \left(1 - d_{\mathrm{TV}}(\mathbb{P}_{\theta_0}, \mathbb{P}_{\theta_1})\right)$$

（比前述的 $\alpha = 2$ 版本更紧）。

### 9.3.2 Upgrades Le Cam's Two-Point Method by Assouad's Lemma

直接缩小到两点疑似有点太极端了，而且对于高维情况下，其实是完全省略掉了高维空间带来的结构，所以没那么极端的方式（i.e. 也就是一种 two-point method 的升级版）是在每个维度上各有两个点，即 $2^p$ 个点的检验。当然额外我们要求 $\ell$ 是可以直接 tensorize 的，比如 $L_1$ 或平方损失。这里以平方损失 $\|\cdot\|_2^2$ 为例，取的点是边长为 $\varepsilon$ 的 cube,

$$\begin{aligned}
\mathbb{E}\left[\ell(\theta, \hat{\theta})\right] &\overset{\text{tensorize}}{=} \sum_{i=1}^{p} \mathbb{E}\left[\ell_i(\hat{\theta}_i, \theta_i)\right] \\
&\geq \frac{1}{4} \sum_{i=1}^{p} \mathbb{E}\left[(\theta_{i0} - \theta_{i1})^2\right] \\
&\geq \frac{\varepsilon^2}{4} \sum_{i=1}^{p} \inf_{\hat{\theta}_i} \mathbb{P}\left(\theta_i \neq \hat{\theta}_i\right) \\
&\geq \frac{\varepsilon^2}{8} \sum_{i=1}^{p} \left(1 - d_{\mathrm{TV}}(\mathbb{P}_{X|\theta_{i0}}, \mathbb{P}_{X|\theta_{i1}})\right)
\end{aligned}$$

对于 $d_{\mathrm{TV}}$ 有如下的 tensorize 关系[6]：

$$d_{\mathrm{TV}}(\mathbb{P}_{X|\theta_{i0}}, \mathbb{P}_{X|\theta_{i1}}) \leq \max_{\text{Hamming distance}(\theta,\theta')=1} d_{\mathrm{TV}}(\mathbb{P}_{X|\theta}, \mathbb{P}_{X|\theta'})$$

得到升级版的 bound：Assouad's lemma

$$\mathfrak{M}(\Theta) \geq \frac{\varepsilon^2 p}{8} \left(1 - \max_{\substack{\mathrm{HamDist}(\theta,\theta')=1 \\ \theta,\theta' \in \mathrm{cube}_p(\varepsilon)}} d_{\mathrm{TV}}(\mathbb{P}_{X|\theta}, \mathbb{P}_{X|\theta'})\right)$$

可以看到这个引理能将 MiniMax risk 从 $\gtrsim 1$ 升级到 $\gtrsim p$，能够具有一些高维性质。

### 9.3.3 Minimax by Fano's Method

Fano's method 可称之为是一种“更纯净”的方法，因为基本的原则就是：我们但凡进行估计（数据处理）就会损失信息，所以如果估计任务 $\theta(\mathcal{P})$ 非常复杂，那么估计过程中损失的信息量也会更大，进而导致误差，这是（我认为的）基于信息的 bound 的理解。Fano's method 的 idea 是：找到一个合适的数据 processing pipeline，让丢失信息量不大的情况下将原本的模型简化到可以简单计算的情况（简单 in the sense that 可以构建直接的概率和信息之间的计算关系），然后再通过信息的单调性来得到 Minimax risk 的 bound。

▶ Roadmap

[6] 见 YW P. 69

1. 对 $\theta$ 和 $\hat{\theta}$ 作离散化（一个 channel）：考虑假设类 $\Theta$ 的一个 $2\delta$-packing $\mathcal{M}=\mathcal{M}(2\delta,\Theta,\ell(\cdot))=\{\theta_i\}_{i=1}^{|\mathcal{M}|}$，这个 packing 诱导出一个一个变换 $f$，将任意 $\tilde{\theta}\in\Theta$ 映射到 $\mathcal{M}$ 中最近的点：

$$f(\hat{\theta})=\arg\min_{\theta_i\in\mathcal{M}}\ell(\hat{\theta},\theta_i)$$

利用这个 pipe 我们有下面的 chain：

$$f(\theta)\leftrightarrow\theta\leftrightarrow X\leftrightarrow\hat{\theta}\leftrightarrow f(\hat{\theta})$$

2. 用离散化的集合做假设检验：$\mathbf{1}\{f(\theta)\neq f(\hat{\theta})\}$，相当于是将 $(f(\theta),f(\hat{\theta}))$ 通过一个 channel 传递，所以可以直接用信息的单调性 (9.1) 处理

$$\begin{aligned}I(\theta;X)\geq&I(f(\theta);f(\hat{\theta}))\\=&D_{KL}(P_{f(\theta),f(\hat{\theta})}\|P_{f(\theta)}\otimes P_{f(\hat{\theta})})\\\geq&D_{KL}(\text{error rate}\|\text{Bernoulli}(1-\frac{1}{|\mathcal{M}|}))\end{aligned}$$

3. 通过如上两个 channel，我们已经能将互信息简化到两个 Bernoulli 的情形了，这允许我们直接搭建互信息和概率之间的关系。通过一些 trivial 的计算可以得到

$$I(\theta;X)\geq D_{KL}(\text{error rate}\|\text{Bernoulli}(1-\frac{1}{|\mathcal{M}|}))\geq-\log 2+\log|\mathcal{M}|\,(1-\text{error rate})$$

具体再研究里面的 error rate 是怎么倒的可以得到：

$$\begin{aligned}\mathbb{P}_{\theta\sim\text{Unif}(\mathcal{M})}\left(\ell(\hat{\theta},\theta)\geq\delta\right)\geq&\mathbb{P}_{\theta\sim\text{Unif}(\mathcal{M})}\left(\ell(f(\hat{\theta}),\theta)\geq 2\delta\right)\\=&\mathbb{P}_{\theta\sim\text{Unif}(\mathcal{M})}\left(\ell(f(\hat{\theta}),f(\theta))\geq 2\delta\right)\\=&\mathbb{P}_{f(\theta),f(\hat{\theta})}\left(f(\theta)\neq f(\hat{\theta})\right)\\=&\text{error rate}\\\geq&1-\frac{I(\theta;X)+\log 2}{\log|\mathcal{M}|}\end{aligned}$$

d

4. 用与 two-point method 类似的 Markov 不等式即可得到 $\mathbb{E}$ 的 Minimax bound[7]：

$$\begin{aligned}\mathfrak{M}=&\inf_{\hat{\theta}}\sup_{\theta\in\Theta}\mathbb{E}_{\theta}\left[\ell(\hat{\theta},\theta)\right]\\\geq&\inf_{\hat{\theta}}\mathbb{E}_{\theta\sim\text{Unif}(\mathcal{M})}\left[\ell(\hat{\theta},\theta)\right]\\\geq&\inf_{\hat{\theta}}\delta\mathbb{P}_{\theta\sim\text{Unif}(\mathcal{M})}\left(\ell(\hat{\theta},\theta)\geq\delta\right)\\\geq&\inf_{\hat{\theta}}\delta\big(1-\frac{I(\theta;X)+\log 2}{\log|\mathcal{M}|}\big)\\=&\delta\big(1-\frac{I(\theta;X)+\log 2}{\log|\mathcal{M}|}\big)\end{aligned}$$

[7]这里的损失函数需要是齐次的，比如 $\ell=\|\cdot\|_p$，否则需要按照 power 数修成 $\varepsilon$ 的 power。比如对于平方损失需要修正成

$$\mathfrak{M}(\Theta)\geq\delta^2\big(1-\frac{I(\theta;X)+\log 2}{\log|\mathcal{M}|}\big)$$

### 9.3.4 Summary of Minimax Lower Bound

Roughly speaking，Minimax lower bound 希望找到某种假设类的近似 $\Theta' \subset \Theta$（比如 $\Theta'$ 包含两个点的 Le Cam's method，或者 $2^p$ 个点的 Assouad's lemma，或者为 $\Theta$ 的 packing set 的 Fano's method），使得

$$\mathfrak{M}(\Theta) \geq \mathfrak{M}(\Theta')$$

这样的离散化操作让 Minimax 是一个可以计算的问题。Information theoretic 的视角是：离散化操作是一种 information processing 的 channel，所以我们只要能对离散化的情形 bound 住某种信息（比如互信息），然后再通过信息的单调性来得到 Minimax risk 的 bound。（离散化的相比原本集合的“更简单”也体现在可以容易地直接将信息与概率对应起来，然后用 Markov 不等式来得到 Minimax risk (which is some $\mathbb{E}$) 的 bound）。

## 9.4 Minimax Upper Bound à la Le Cam-Birgé's comparison theorem

另一侧是 Minimax upper bound，思路也很直接，就是找到某种“好的”估计量 $\tilde{\theta}$，使得

$$\mathfrak{M}(\Theta) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[ \ell(\hat{\theta}, \theta) \right] \leq \mathbb{E}_\theta \left[ \ell(\tilde{\theta}, \theta^*) \right]$$

这个估计量最好是一种“通用”的方法，使得对于各种模型我们都能得到一个不错的 upper bound（因为如果对于每个模型都要重新构造一个估计量，那就有点显得画蛇添足了，这个 $\tilde{\theta}$ 的通用性使得我们可以研究一些我们不知道怎么解决的问题）（“不知道怎么解决”也可以是 in the sense of 不知道怎么在 computationally efficient 的情况下解决）。

### 9.4.1 Le Cam-Birgé's comparison theorem

Le Cam-Birgé's comparison theorem 事实上也是基于离散化 + 假设检验问题的，具体来说是用那个最优估计量[8]在一个 covering set 上进行一个"tournament"，选择数据最倾向于选择的点作为估计量。

Le Cam-Birgé's comparison 的原版版本使用的是 Hellinger 距离 $H^2(P,Q) := \left( \int \sqrt{p(x) - q(x)} dx \right)^2$（见 YW P110），这里暂且遵循这个版本，即

$$\ell(\hat{\theta}, \theta) = H^2(P_{\hat{\theta}}, P_\theta), \quad \theta \in \Theta$$

**Roadmap:** 首先假设我们有某种最优 decision rule（pairwise comparison）$\psi_{ij}(\mathcal{D})$ 可以评价“数据更 favor$\{\theta_i, \theta_j\}$ 中的哪个点”，即下面的假设检验问题：

$$\theta \in \text{Ball}(\theta_i) \mapsto \mathcal{P} \Leftrightarrow \theta \in \text{Ball}(\theta_j) \mapsto \mathcal{Q}$$

对于这种简单的假设检验问题，我们是知道最小的错误概率形式的：可以用 $d_{\text{TV}}$ 表示出来[9]：

$$\begin{aligned} \min_{\psi} \left\{ \sup_{P \in \mathcal{P}} P(\psi \text{ favor } Q) + \sup_{Q \in \mathcal{Q}} Q(\psi \text{ favor } P) \right\} &= 1 - d_{\text{TV}}(\text{conv}(\mathcal{P}^{\otimes n}), \text{conv}(\mathcal{Q}^{\otimes n})) && (9.2) \\ &\leq \left(1 - \frac{1}{2} H^2(\text{conv}(\mathcal{P}), \text{conv}(\mathcal{Q}))\right)^n && (9.3) \\ &\leq \exp\left[ -\frac{n}{2} H^2(\text{conv}(\mathcal{P}), \text{conv}(\mathcal{Q})) \right] && (9.4) \end{aligned}$$

[8]我们不一定知道，但我们会有一些关于这个最优估计量的一些性质。

[9]这里只给出结果，见 YW P. 108

也就是对于每一对给定的点 $\theta_i, \theta_j$ 和某种意义上的以之为中心的球，我们可以得到对这对点的一个理论最优区分方式 $\psi_{ij}$，将这个区分方法 replicate 到整个假设类的 covering set 上就能得到整体错误率的一个 bound。

1. 首先是 pairwise comparison: 对于两个点 $\theta_i, \theta_j$ 和它们对应的 Hellinger distance ball $B_H(\theta_i, \varepsilon)$, $B_H(\theta_j, \varepsilon)$，我们考虑应该判定 $\theta$ 是在哪个 ball 里面。想象我们有某种（**基于数据的**）比较方法 $\psi_{ij}(\mathcal{D})$，并基于这个 $\psi_{ij}$ 决定我们判定

$$\theta \in \begin{cases} B_H(\theta_i, \varepsilon) & \text{if } \psi_{ij}(\mathcal{D}) = 0, \qquad \text{v.s.} \\ B_H(\theta_j, \varepsilon) & \text{if } \psi_{ij}(\mathcal{D}) = 1 \end{cases}$$

   利用上面(9.2)的 bound，我们知道最优检验 $\psi_{ij}$ 满足

$$\begin{aligned} \sup_{\theta \in B_{H^2}(\theta_i, \varepsilon)} P(\psi_{ij} = 1) \vee \sup_{\theta \in B_{H^2}(\theta_j, \varepsilon)} P(\psi_{ij} = 0) &\leq \exp\left[-\frac{n}{2} H^2(\mathrm{conv}(\mathcal{P}, \mathrm{conv}(\mathcal{Q})))\right] \\ &= \exp\left[-\frac{n}{2} H^2(B_H(\theta_i, \varepsilon), B_H(\theta_j, \varepsilon))\right] \end{aligned}$$

2. 制作 packing set 使得上述的 pairwise comparison 可以扩展覆盖整个 $\Theta$。我们考虑 $\Theta$ 的一个（maximal）$\delta$-packing $\mathcal{M} = \mathcal{M}(\delta, \Theta, H(\cdot, \cdot)) = \{\theta_i\}_{i=1}^{|\mathcal{M}|}$，它自然也是一个 $\delta$ covering。然后我们取上面的 $\varepsilon = \delta$，这样我们就得到 $\forall i \neq j \in \mathcal{M}, H(P_{\theta_i}, P_{\theta_j}) \geq 4\delta$ :

$$\sup_{P \in B(\theta_i, \delta), Q \in B(\theta_j, \delta)} H(P, Q) \geq H(P_{\theta_i}, P_{\theta_j}) - 2\delta \geq H(P, Q) \geq H(P_{\theta_i}, P_{\theta_j})/2 \geq 2\delta$$

$$\Rightarrow \sup_{\theta \in B_{H^2}(\theta_i, \varepsilon)} P(\psi_{ij} = 1) \vee \sup_{\theta \in B_{H^2}(\theta_j, \varepsilon)} P(\psi_{ij} = 0) \leq \exp\left[-\frac{n}{8} H^2(P_{\theta_i}, P_{\theta_j})\right] \leq \exp\left[2n\delta^2\right], \quad H(P_{\theta_i}, P_{\theta_j}) \geq 4\delta$$

   直观来说：只要 $\theta_i, \theta_j$ “不是特别近”，我们就能大概率区分开来。

3. 升级到 packing set 并构造估计量。

   由于对任意的 $i \neq j \in \mathcal{M}$ 都有一个对应的 $\psi_{ij}$，我们可以用这组 $\{\psi_{ij}\}$ 来判断 $\theta$ 在哪个 ball 里面，具体如下：

$$\text{def. } T_i := \begin{cases} \max_{\theta_j \in \mathcal{M}} H(P_{\theta_i}, P_{\theta_j}) & \text{if } \psi_{ij}(\mathcal{D}) = 1, \quad H(P_{\theta_i}, P_{\theta_j}) \geq 4\delta \\ 0 & \text{otherwise} \end{cases}$$

$$\hat{\theta} \in \arg\min_{\theta_i \in \mathcal{M}} T_i$$

   也就是：对于每个 $\theta_i$，$T_i$ 统计的是：与 $\theta_i$ 距离足够远（所以大概率不会和 $\theta_i$ 由于随机误差而混淆），但又能够“打败”$\theta i$ 的那些 $\theta_i \in \mathcal{M}$，然后取与 opponents 最近的 $\theta_i$ 作为估计量（或者没有足够远的 opponents 能打败 $\theta_i$ 时 $T_i = 0$ 的话自然 $\theta_i$ 是最优）。

4. 研究升级后的估计量的 prob bound，具体来说，如果 ground truth 是 $\theta^* \in B_H(\theta^*, \delta)$，那么

$$\mathbf{1}_{H(P_{\hat{\theta}}, P_{\theta^*}) \geq 2\delta} \leq \mathbf{1}_{\max(T_{i^*}, T_{\hat{i}}) \geq 2\delta} = \mathbf{1}_{T_{i^*} \geq 2\delta}$$

   于是:$\forall t \geq 1$

$$\mathbb{P}_{\theta^*}\left(H(P_{\hat{\theta}}, P_{\theta^*}) \geq 4t\delta\right) \leq P_{\theta^*}\left(T_{i^*} \geq 4t\delta\right) \leq |\mathcal{M}| \exp\left[-2nt^2\delta^2\right] \qquad (\star)$$

（这里得到了一个包含 $t$ 的 tail bound，通过调整前面不等式的 packing/covering scale 的位置很容易得到）。同时需要一个一些技术条件 1：$n\delta^2 \geq 1$ 之类的。

积分就得到关于期望的 Minimax bound:

$$\begin{aligned}\mathfrak{M}(\Theta) &= \inf_{\hat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_\theta \left[\ell(\hat{\theta}, \theta)\right] \\ &\leq \sup_{\theta\in\Theta} \mathbb{E}_\theta \left[\ell(\hat{\theta}_{\{\psi_{ij}\}}, \theta)\right] \\ &\leq \int_0^\infty \mathbb{P}\left(H^2(P_{\hat{\theta}}, P_{\theta^*}) \geq \xi\right) \mathrm{d}\xi\end{aligned}$$

之后增加一个 $2n\delta^2 \geq \log|\mathcal{M}|$ 的技术条件 2使得我们可以分别积 $[0,\delta)$ 和 $[\delta,\infty)$ 就能得到 Minimax upper bound：

$$\mathfrak{M}(\Theta) \lesssim \delta^2, \quad n\delta^2 \gtrsim \log|\mathcal{M}(\delta, \Theta, H(\cdot,\cdot))| \vee 1$$

**Note:** 关于这两个技术条件，可以发现 R.H.S. 是一个 $\delta$ 的减函数，所以我们只要取最小的 possible $\delta$ 就能得到优化的 Minimax upper bound.

**Note:** Le Cam-Birgé's comparison theorem 中关键的一步是：不需要具体/显式地找出某个足够好的估计量，而是给出某个好的估计量可以满足什么样的性质（具体来说，一个关于 Hellinger distance 的 tournament 和相关的 prob bound），然后得到一个 Minimax upper bound。

**Note:** 我们也可以只部分使用 Le Cam-Birgé's comparison theorem，比如

- 我们已经显式构造出了一个不错的估计量 $\tilde{\theta}$，那么直接 bound
$$\sup_{\theta\in\Theta} \mathbb{E}_\theta \left[\ell(\tilde{\theta}, \theta)\right]$$
- 我们能设计出某种不错的 comparison 方法 $\psi_{ij}(\|\cdot\|)$ 来得到一个类似于
$$\text{error rate}(B_{\|\cdot\|}(\theta_i), B_{\|\cdot\|}(\theta_j)) \leq \exp\left[-\text{const}(n,\delta,...)\cdot\|\theta_i-\theta_j\|^2\right], \quad \|\theta_i-\theta_j\| \geq C\delta$$
的 bound，那后面复刻 upgrade 到 packing set 上的过程，也会能得到一个 Minimax upper bound。

# 10 Miscellaneous

## 10.1 Universality

Universality 指的是一类：某个 object 在进入高维的时候其细节变得不显著（irrelevent）的现象，一个典型的例子是中心极限定理：对于任意的（with some mild conditions）独立同分布的随机变量 $X_1, X_2, \cdots X_n \overset{i.i.d.}{\sim} X$，和具有相同前两阶矩的 Gaussian 随机变量 $G_1, G_2, \cdots, G_n$，我们有

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i - \mathbb{E}X_i \overset{(i)}{\approx} \frac{1}{\sqrt{n}}\sum_{i=1}^n (G_i - \mathbb{E}G_i)$$

其中 $(i)$ in the sense of closeness of the distribution, e.g. $\xrightarrow{\mathrm{d}}$. 在这个过程中所有关于 $X$ 更高阶矩的信息都 vanish 了。

Universality 提供了如下几个 points:

- 很多高维现象最后会归结为与一些少数几个参数有关，而细节被冲淡了（当然一定也会有些细节不重要的，和施加的假设强弱有关）
- 一些难以 explicitly 计算的东西可以通过 Universality 倒成为可以计算的东西（比如倒成高斯 r.v. 模型或其他一些易于计算的分布模型）
- 高维情况下有时候还会有相变（phase transition）现象，这时候 Universality 会有一些极端现象（incontinuity in some senses）