

Theoretical Machine Learning

Lectured by [Zhihua Zhang](#)

L^AT_EXed by [Chengxin Gong](#)

April 1, 2024

Contents

1	简介	2
2	统计决策理论	2
3	统计学习理论	5

1 简介

- 机器学习的主要任务: 生成、预测、决策. 生成: $X_1, \dots, X_n \sim F$, 推断分析 F , 无监督学习, GAN, GPT, \dots . 预测: 数据对 $(X^{(1)}, Y^{(1)}), \dots, (X^{(n)}, Y^{(n)})$, $X^{(i)} \in \mathbb{R}^d$ 输入变量, $f: \mathcal{X} \rightarrow \mathcal{Y}, x \in \mathcal{X}, y \in \mathcal{Y}$, 归因, 有监督学习. 决策: 强化学习, Agent \leftarrow action, state, reward \rightarrow 环境.
- 求解问题的途径: 参数/非参数, 频率 (MLE)/贝叶斯.
- 误差模型: 有监督: $X = (X_1, \dots, X_d)^T \in \mathbb{R}^d$, 回归: $Y \in \mathbb{R}$; 分类: $Y \in \{0, 1\}(\{-1, 1\}, \{1, \dots, M\}, \{0, 1\}^M)$; X 随机, Random design(生成模型), $Y = g(X) + \varepsilon \stackrel{\text{or}}{=} g(X, Z), Y^{(i)} = g(X^{(i)}, Z^{(i)})$; X 固定 $X = x$, Fixed design(判别模型), $Y^{(i)} = g(x^{(i)}, Z^{(i)})$. 无监督: $X = g(Z)$ (因子模型: $X = AZ + \varepsilon, Z \in \mathcal{N}(0, 1), \varepsilon \sim \mathcal{N}(0, \Sigma)$).

2 统计决策理论

- Consider a state space Ω , data space \mathcal{D} , model $\mathcal{P} = \{p(\theta, x)\}$, action space \mathcal{A} . Loss function: $\mathcal{L}: \Omega \times \mathcal{A} \rightarrow [-\infty, +\infty]$, measurable, nonnegative. A measurable function $\delta: \mathcal{D} \rightarrow \mathcal{A}$ is called a nonrandomized decision rule. Risk function is defined as $\mathcal{R}(\theta, \delta) = \int \mathcal{L}(\theta, \delta(x)) dP_\theta(x) = \mathbb{E}_\theta \mathcal{L}(\theta, \delta(X))$. Randomized decision: for each $X = x$, $\delta(x)$ is a probability distribution: $[A|X = x] \sim \delta_x$. Risk function for δ : $\mathcal{R}(\theta, \delta) = \mathbb{E}_\theta \mathcal{L}(\theta, A) = \mathbb{E}_\theta \mathbb{E}_a \mathcal{L}(\theta, A|X) = \iint \mathcal{L}(\theta, a) d\delta_x(a) dP_\theta(x)$.
- Example [参数估计]: $\theta \in \Omega, \mathcal{A} = \Omega, \mathcal{L}(\theta, a) = \|\theta - a\|_2^2 \stackrel{\text{or}}{=} \|\theta - a\|_p^p (p \geq 1) \stackrel{\text{or}}{=} \int \log \frac{P_\theta(x)}{P_a(x)} P_\theta(x) dm(x) (\text{KL})$. $\mathcal{R} = \text{Var}(a) + \text{bias}^2(a)$. Bregmass loss: $\phi: \mathbb{R}^d \rightarrow \mathbb{R}$ describe any strictly convex differentiable function. Then $\mathcal{L}_\phi(\theta, a) = \phi(a) - \phi(\theta) - (\phi - a)^T \nabla \phi(a)$.
- Example [Testing]: $\mathcal{A} = \{0, 1\}$ with action “0” associated with accepting $H_0: \theta \in \Omega_0$ and “1”: $H_1: \theta \in \Omega_1$. δ_x is a Bernolli distribution. $\mathcal{L}(\theta, a) = I\{a = 1, \theta \in \Omega_0\} + I\{a = 0, \theta \in \Omega_1\}$. Risk $\mathcal{R}(\theta, \delta) = \mathbb{P}_\theta(A = 1)1_{\theta \in \Omega_0} + \mathbb{P}_\theta(A = 0)1_{\theta \in \Omega_1}$.
- A decision rule δ is called inadmissible if a competing rule δ^* such that $\mathcal{R}(\theta, \delta^*) \leq \mathcal{R}(\theta, \delta)$ for all $\theta \in \Omega$ and $\mathcal{R}(\theta, \delta^*) < \mathcal{R}(\theta, \delta)$ for at least one $\theta \in \Omega$. Otherwise, δ is admissible.
- The maximum risk $\bar{\mathcal{R}}(\delta) = \sup_{\theta \in \Omega} \mathcal{R}(\theta, \delta)$ and the Bayes risk $r(\Lambda, \delta) = \int \mathcal{R}(\theta, \delta) d\Lambda(\theta) = \int \mathcal{L}(\theta, \delta) d\mathbb{P}(x, \theta)$ ($\Lambda(\theta)$ is a prior). A decision rule that minimizes the Bayes risk is called a Bayes rule, that is, $\hat{\delta}: r(\Lambda, \hat{\delta}) = \inf_\delta r(\Lambda, \delta)$. Minimax rule $\delta^*: \sup_{\theta \in \Omega} \mathcal{R}(\theta, \delta^*) = \inf_\delta \sup_{\theta \in \Omega} \mathcal{R}(\theta, \delta)$.
- If risk functions for all decision rules are continuous in θ , if δ is Bayesian for Λ and has finite integrated risk $r(\Lambda, \delta) < \infty$, and if the support of Λ is the whole state space Ω , then δ is admissible.
- $p(\theta|x) = \frac{p_\theta(x)\lambda(\theta)}{\int p_\theta(x)\lambda(\theta)d\theta} := \frac{p_\theta(x)\lambda(\theta)}{m(x)}$. Define the posterior risk of δ : $r(\delta|X = x) = \int \mathcal{L}(\theta, \delta(x)) d\mathbb{P}(\theta|x)$. The Bayes risk $r(\Lambda, \delta)$ satisfies that $r(\Lambda, \delta) = \int r(\delta|x) dM(x)$. Let $\hat{\delta}(x)$ be the value of δ that minimizes $r(\delta|x)$. Then $\hat{\delta}$ is the Bayes rule.
- Application to supervised learning. Case 1: Regression. $(X, Y) \in \mathcal{X} \times \mathcal{Y}, f: \mathcal{X} \rightarrow \mathcal{Y}, \mathcal{A} = \Omega = \mathcal{Y}, \mathcal{D} = \mathcal{X}, \delta = f, \mathcal{L}(Y, f(X)) = \|Y - f(X)\|_p^p, p \geq 1$, risk $R_f = \iint \mathcal{L}(y, f(x)) d\mathbb{P}(x, y) = \mathbb{E}[\mathcal{L}(Y, f(X))] = \mathbb{E}[\mathbb{E}\mathcal{L}(Y, f(X))|X]$. When $p = 2$, $r(f|X = x) = \int \mathcal{L}(y, f(x)) d\mathbb{P}(y|x) = \int |y - f(x)|^2 d\mathbb{P}(y|x)$. 回归函数 $g(x) := \int y d\mathbb{P}(y|x) \Rightarrow R_f = \mathbb{E}|Y - f(X)|^2 = \mathbb{E}|Y - g(X) + g(X) - f(X)|^2 = \mathbb{E}|Y - g(X)|^2 + \mathbb{E}|g(X) - f(X)|^2 \geq \mathbb{E}|Y - g(X)|^2$.
- Case 2: Pattern classification. $Y \in \{0, 1\}, p_0 = P(Y = 0), p_1 = P(Y = 1) = 1 - p_0, \mathbb{E}[\mathcal{L}(Y, f(X))] = P(Y \neq f(X))$. The Bayesian rule (predictor) is given by $f(x) = 1\{\mathbb{P}(Y = 1|X = x) \geq \frac{\mathcal{L}(1,0) - \mathcal{L}(0,0)}{\mathcal{L}(0,1) - \mathcal{L}(1,1)} \mathbb{P}(Y = 0|X = x)\}$. (Proof: $\mathbb{E}[\mathcal{L}(Y, f(X))|X = x] = \begin{cases} \mathbb{E}[\mathcal{L}(Y, 0)|X = x] = \mathcal{L}(0,0)\mathbb{P}(Y = 0|X = x) + \mathcal{L}(1,0)\mathbb{P}(Y = 1|X = x) \\ \mathbb{E}[\mathcal{L}(Y, 1)|X = x] = \mathcal{L}(0,1)\mathbb{P}(Y = 0|X = x) + \mathcal{L}(1,1)\mathbb{P}(Y = 1|X = x) \end{cases}$, 比较大小)
- 连续化: $\mathbb{P}(Y = 1|X = x) = \mathbb{E}(Y|X = x) := g(x)$ (回归), $f(x) = 1\{g(x) \geq \frac{1}{2}\}$. Then $0 \leq \mathbb{P}(\hat{f}(X) \neq Y) - \mathbb{P}(f(X) \neq Y) \leq 2 \int_{\mathcal{X}} |\hat{g}(x) - g(x)| \mu(dx) \leq 2(\int_{\mathcal{X}} |\hat{g}(x) - g(x)|^2 \mu(dx))^{\frac{1}{2}}$.

- 回到 Case 2. $f(x) = 1\{\frac{p(x|y=1)}{p(x|y=0)} \geq \frac{p_0(\mathcal{L}(0,1)-\mathcal{L}(0,0))}{p_1(\mathcal{L}(1,0)-\mathcal{L}(1,1))}\}$, 这与似然比检验 (LRT) 相同: Likelihood $L(X) := \frac{p(X|Y=1)}{p(X|Y=0)}$, 形式为 $f(x) = 1\{L(x) \geq \eta\}$.

- Confusion table:

	Y = 0	Y = 1
$\hat{Y} = 0$	true negative	false negative
$\hat{Y} = 1$	false positive	true positive

True Positive Rate: $\text{TPR} = \mathbb{P}(\hat{Y} = 1|Y = 1)$; False Negative Rate: $\text{FNR} = 1 - \text{TPR}$, type II error; False Positive Rate: $\text{FPR} = \mathbb{P}(\hat{Y} = 1|Y = 0)$, type I error; True Negative Rate: $\text{TNR} = 1 - \text{FPR}$. Precision: $\mathbb{P}(Y = 1|\hat{Y} = 1) = \frac{p_1 \text{TPR}}{p_0 \text{FPR} + p_1 \text{TPR}}$. F_1 -score: F_1 is the harmonic mean of precision and recall, which can be written as $F_1 = \frac{2\text{TPR}}{1 + \text{TPR} + \frac{p_0}{p_1} \text{FPR}}$.

- Optimization: maximize TPR subject to $\text{FPR} \leq \alpha, \alpha \in [0, 1]$. Randomized rule: Q return 1 with probability $Q(x)$ and 0 with probability $1 - Q(x)$. Maximize $\mathbb{E}[Q(x)|Y = 1]$ subject to $\mathbb{E}[Q(x)|Y = 0] \leq \alpha$. Suppose the likelihood functions $p(x|y)$ are continuous. Then the optimal predictor is a deterministic LRT (N-P lemma). (Proof: Let η be the threshold for an LRT such that the predictor $Q_\eta(x) = 1\{\alpha(x) \geq \eta\}$ has $\text{FPR} = \alpha$. Such an LRT exists because likelihood are continuous. Let β denote the TPR of Q_η . Prove that Q_η is optimal for risk minimization problem corresponding to the loss functions $\mathcal{L}(0, 1) = \eta \frac{p_1}{p_0}, \mathcal{L}(1, 0) = 1, \mathcal{L}(1, 1) = \mathcal{L}(0, 0) = 0$ since $\frac{p_0(\mathcal{L}(0,1)-\mathcal{L}(0,0))}{p_1(\mathcal{L}(1,0)-\mathcal{L}(1,1))} = \frac{p_0 \mathcal{L}(0,1)}{p_1 \mathcal{L}(1,0)} = \eta$. Under these loss functions, the risk of Bayes predictor for Q is $\mathcal{R}_Q = p_0 \text{FPR}(Q) \mathcal{L}(0, 1) + p_1(1 - \text{TPR}(Q)) \mathcal{L}(1, 0) = p_1 \eta \text{FPR}(Q) + p_1(1 - \text{TPR}(Q))$. Now let Q be any other rule with $\text{FPR}(Q) \leq \alpha, \mathcal{R}_{Q_\eta} = p_1 \eta \alpha + p_1(1 - \beta) \leq p_1 \eta \text{FPR}(Q) + p_1(1 - \text{TPR}(Q)) \leq p_1 \eta \alpha + p_1(1 - \text{TPR}(Q)) \Rightarrow \text{TPR}(Q) \leq \beta$)
- ROC (Receiver operating character) curve: y -axis is TPR and x -axis is FPR. Proposition: (1) The points (0,0) and (1,1) are on the ROC curve; (2) The ROC must lie above the main diagonal; (3) The ROC curve is concave. (Proof: (2): Fix $\alpha \in (0, 1)$ and consider a randomized rate $\text{TPR} = \text{FPR} = \alpha, Q(x) \equiv \alpha$; (3): Consider two rules $(\text{FPR}(\eta_1), \text{TPR}(\eta_1))$ and $(\text{FPR}(\eta_2), \text{TPR}(\eta_2))$. If we flip a biased coin and use the first rule with probability t and use the second rule with probability $1 - t$. Then this yields a randomized rule with $(\text{FPR}, \text{TPR}) = (t\text{FPR}(\eta_1) + (1 - t)\text{FPR}(\eta_2), t\text{TPR}(\eta_1) + (1 - t)\text{TPR}(\eta_2))$. Fixing $\text{FPR} \leq t\text{FPR}(\eta_1) + (1 - t)\text{FPR}(\eta_2)$, $\text{TPR} \geq t\text{TPR}(\eta_1) + (1 - t)\text{TPR}(\eta_2)$.)
- Markov Decision Processes (MDPs): Five elements: decision epoches, states, actions, transition probabilities and rewards. (1) Decision epoches: Let T denote the set of decision epoches, discrete: $\{1, 2, \dots, N\}$; continuous: $[0, N]$; $N < / = \infty$: finite or infinite. (2) State and action sets: decision epoch $t \in T$, the system occupies a state $S_t \in \mathcal{S}$, the decision maker $a \in \mathcal{A}$. (3) Reward and transition probabilities: t , in state s , choose action a , (i) the decision maker receives a reward $r_t(s, a)$, (ii) the system state at the next decision epoch is determined by the probability distribution $p_t(\cdot|s_t, a)$.
- Decision rules: Prescribe a procedure for action selection in each state at a specified decision epoch. Four cases: (1) Markovian and Deterministic: $\delta_t : \mathcal{S} \rightarrow \mathcal{A}$; (2) M and Randomized: $\delta_t : \mathcal{S} \rightarrow \Delta(\mathcal{A})(q_{\delta_t(s)}(a))$; (3) History-dependent and D: $h_t = (s_1, a_1, \dots, s_{t-1}, a_{t-1}, s_t) = (h_{t-1}, a_{t-1}, s_t), \mathcal{H}_1 = \mathcal{S}, \mathcal{H}_2 = \mathcal{S} \times \mathcal{A} \times \mathcal{S}, \dots, \delta_t : \mathcal{H}_t \rightarrow \mathcal{A}$; (4) HR: $\delta_t : \mathcal{H}_t \times \Delta(\mathcal{A})$. A policy $\pi = (\delta_1, \delta_2, \dots, \delta_{N-1})$ is stationary if $\delta_1 = \delta_2 = \dots = \delta$ for $t \in T$.
- Let $\pi = (\delta_1, \dots, \delta_{N-1})$ in HR and $R_t := r_t(X_t, Y_t)$ denote the random reward, $R_N := r_N(X_N), R := (R_1, \dots, R_N)$. The expected total reward $U_N^\pi(s) := \mathbb{E}^\pi\{\sum_{t=1}^{N-1} r_t(X_t, Y_t) + r_N(X_N)|X_1 = s\}$. Assume $|r_t(s, a)| \leq M < \infty$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. Optimal policy: $U_N^*(s) \geq U_N^\pi(s), s \in \mathcal{S}$. ε -optimal policy: $U_N^{\pi^*}(s) + \varepsilon > U_N^\pi(s), s \in \mathcal{S}$. The value of the MDP: $U_N^*(s) = \sup_{\pi \in \mathcal{D}^{\text{HR}}} U_N^\pi(s), s \in \mathcal{S}$.
- Finite-Horizon Policy Evaluation: $V_t^\pi(h_t) = \mathbb{E}^\pi\{\sum_{k=t}^{N-1} r_k(X_k, Y_k) + r_N(X_N)|h_t\}, V_N^\pi(h_N) = r_N(s), \pi \in \mathcal{D}^{\text{HD}}$. 由重期望公式, $V_t^\pi(h_t) = r_t(s_t, \delta_t(h_t)) + \mathbb{E}_{h_t} V_{t+1}^\pi(h_t, \delta_t(h_t), X_{t+1}) = r_t(s_t, \delta_t(h_t)) + \sum_{j \in \mathcal{S}} V_{t+1}^\pi(h_t, \delta_t(h_t), j) p(j|s_t, \delta_t(h_t))$.

Consider randomness (i.e. $\pi \in \mathcal{D}^{\text{HR}}$): $V_t^\pi(h_t) = \sum_{a \in \mathcal{A}} q_{\delta_t(h_t)}(a) \{r_t(s_t, a) + \sum_{j \in \mathcal{S}} V_{t+1}^\pi(h_t, a, j) p(j|s_t, a)\}$. Computational complexity: let $K = |\mathcal{S}|, L = |\mathcal{A}|$, at decision epoch t , $K^{t+1}L^t$ histories, $K^2 \sum_{i=0}^{N-1} (KL)^i$ multiplications. If $\pi \in \mathcal{D}^{\text{MD}}$, $V_t^\pi(s_t) = r_t(s_t, \delta_t(s_t)) + \sum_{j \in \mathcal{S}} V_{t+1}^\pi(j) p(j|s_t, \delta_t(s_t))$, only $(N-1)K^2$ multiplications. On the other hand, given π , this yields a valid and accurate calculation method for $U_N^\pi(s)$.

- The Bellman Equations: Let $V_t^*(h_t) = \sup_{\pi \in \mathcal{D}^{\text{HR}}} V_t^\pi(h_t)$. The optimality equations: $V_t(h_t) = \sup_{a \in \mathcal{A}} \{r_t(s_t, a) + \sum_{j \in \mathcal{S}} V_{t+1}(h_t, a, j) p_t(j|s_t, a)\}$ for $t = 1, 2, \dots, N-1$ and $h_t = (h_{t-1}, a_{t-1}, s_t) \in \mathcal{H}_t$. For $t = N$, $V_N(h_N) = r_N(s_N)$. Suppose V_t is a solution and V_N satisfies $V_N(h_N) = r_N(s_N)$. Then $V_t(h_t) = V_t^*(h_t)$ for all $h_t \in \mathcal{H}_t, t = 1, \dots, N$ and $V_1(s_1) = V_1^*(s_1) = U_N^*(s_1)$ for all $s_1 \in \mathcal{S}$. (**Proof:** Two parts. First prove $V_n(h_n) \geq V_n^*(h_n)$ for all $h_n \in \mathcal{H}_n$. By induction: $N : V_N(h_N) = r_N(s_N) = V_N^*(h_N)$ for all h_N, π . Now assume that $V_t(h_t) \geq V_t^*(h_t)$ for all $h_t \in \mathcal{H}_t$ for $t = n+1, \dots, N$. Let $\pi' = (\delta'_1, \dots, \delta'_{N-1})$ be an arbitrary policy in \mathcal{D}^{HR} . For $t = n$, the Bellman equations $V_n(h_n) = \sup_{a \in \mathcal{A}} \{r_n(s_n, a) + \sum_{j \in \mathcal{S}} p(j|s_n, a) V_{n+1}(h_n, a, j)\} \geq \sup_{a \in \mathcal{A}} \{r_n(s_n, a) + \sum_{j \in \mathcal{S}} p_n(j|s_n, a) V_{n+1}^*(h_n, a, j)\} \geq \sup_{a \in \mathcal{A}} \{r_n(s_n, a) + \sum_{j \in \mathcal{S}} p_n(j|s_n, a) V_{n+1}^{\pi'}(h_n, a, j)\} \geq V_n^{\pi'}(h_n)$. Second prove for any $\varepsilon > 0$, there exists a $\pi \in \mathcal{D}^{\text{HD}}$ for which $V_n^{\pi'}(h_n) + (N-n)\varepsilon \geq V_n(h_n) \Rightarrow V_n^*(h_n) + (N-n)\varepsilon \geq V_n^{\pi'}(h_n) + (N-n)\varepsilon \geq V_n(h_n) \geq V_n^*(h_n)$. Construct a policy $\pi' = (\delta'_1, \dots, \delta'_{N-1})$ by choosing $\delta'_n(h_n)$ to satisfy $r_n(s_n, \delta'_n(h_n)) + \sum_{j \in \mathcal{S}} p_n(j|s_n, \delta'_n(h_n)) V_{n+1}(h_n, \delta'_n(h_n), j) + \varepsilon \geq V_n(h_n)$. By induction: $N : V_N^{\pi'}(h_N) = V_N(h_N)$. Assume that $V_t^{\pi'}(h_t) + (N-t)\varepsilon \geq V_t(h_t)$ for $t = n+1, \dots, N$. For $t = n$, $V_n^{\pi'}(h_n) = r_n(s_n, \pi'_n(h_n)) + \sum_{j \in \mathcal{S}} p_n(j|s_n, \delta'_n(h_n)) V_{n+1}^{\pi'}(h_n, \delta'_n(h_n), j) \geq V_n(h_n) - (N-n)\varepsilon$. The equations yield that $\delta'_t(h_t) \in \arg \max_{a \in \mathcal{A}} \{r_t(s_t, a) + \sum_{j \in \mathcal{S}} p_t(s_t, a) V_{t+1}^*(h_t, a, j)\}$, which means it is HD, i.e. $U_N^*(s) = \sup_{\pi \in \mathcal{D}^{\text{HR}}} U_N^\pi(s) = \sup_{\pi \in \mathcal{D}^{\text{HD}}} U_N^\pi(s) \stackrel{?}{=} \sup_{\pi \in \mathcal{D}^{\text{MD}}} U_N^\pi(s)$.
- Let $V_t^*, t = 1, \dots, N$ be solutions of Bellman Equations. Then (a) For each $t = 1, \dots, N, V_t^*(h_t)$ depends on h_t only through s_t ; (b) For any $\varepsilon > 0$, there exists an ε -optimal policy which is D and M; (c) Max can be achieved, it is optimal, which is MD. (**Proof:** (a): By induction, $V_N^*(h_N) = V_N^*(h_{N-1}, a_{N-1}, s) = r_N(s)$ for all $h_{N-1} \in \mathcal{H}_{N-1}$. Assume (a) is valid for $t = n+1, \dots, N$. Then $V_n^*(h_n) = \sup_{a \in \mathcal{A}} \{r_t(s_t, a) + \sum_{j \in \mathcal{S}} p_t(j|s_t, a) V_{t+1}^*(j)\} = V_n^*(s_t)$.
- Backward Induction (Dynamic Programming) Algorithm: 1. Set $t = N$ and $V_N^*(s_N) = r_N(s_N)$ for all $s_N \in \mathcal{S}$; 2. Substitute $t-1$ for t and compute $V_t^*(s_t)$ for each $s_t \in \mathcal{S}$: $V_t^*(s_t) = \max_{a \in \mathcal{A}} \{r_t(s_t, a) + \sum_{j \in \mathcal{S}} p_t(j|s_t, a) V_{t+1}^*(s_t)\}$, set $\mathcal{A}_{s_t} = \arg \max_{a \in \mathcal{A}} \{r_t(s_t, a) + \sum_{j \in \mathcal{S}} p_t(j|s_t, a) V_{t+1}^*(s_t)\}$; 3. If $t = 1$, stop. Otherwise return to Step 2.
- Other remarks: (1) At time t , specialized \mathcal{S}_t and \mathcal{A}_s , special structure for r_t and p_t ; (2) $K = |\mathcal{S}|$ and $L = |\mathcal{A}|$, at each t , only $(N-1)LK^2$ multiplications, ease computation and storage cost (because there are $(L^K)^{N-1}$ DM policies).
- Infinite-Horizon MDPs: Assumptions: Stationary reward and transition probabilities $r_t(s, a) \equiv r(s, a), p_t(j|s, a) \equiv p(j|s, a)$; Bounded rewards $|r(s, a)| \leq M < \infty$ for all $a \in \mathcal{A}$ and $s \in \mathcal{S}$; Discounting $\lambda, 0 \leq \lambda < 1$; Discrete state space \mathcal{S} . The expected total reward of policy $\pi = (\delta_1, \delta_2, \dots) \in \mathcal{D}^{\text{HR}}$: $U^\pi(s) = \lim_{N \rightarrow +\infty} \mathbb{E}_s^\pi \{ \sum_{t=1}^N \lambda^{t-1} r(X_t, Y_t) \} = \mathbb{E}_s^\pi \{ \sum_{t=1}^{+\infty} \lambda^{t-1} r(X_t, Y_t) \}$. We say that a policy π^* is optimal when $U^{\pi^*}(s) \geq U^\pi(s)$ for each $s \in \mathcal{S}$ and all $\pi \in \mathcal{D}^{\text{HR}}$. Define the value of the MDP $U^*(s) = \sup_{\pi \in \mathcal{D}^{\text{HR}}} U^\pi(s)$. Let $U_\nu^\pi(s)$ denote the expected reward obtained by using π when the horizon ν is random. Then $U_\nu^\pi(s) = \mathbb{E}_s^\pi \{ \mathbb{E}_{\nu \sim P} \sum_{t=1}^\nu r(X_t, Y_t) \}$. Let's recall geometric distribution with parameter $\lambda : \mathbb{P}(\nu = n) = (1-\lambda)\lambda^{n-1}, n = 1, 2, \dots$.
- Suppose ν has a GD(λ). Then $U^\pi(s) = U_\nu^\pi(s)$ for all $s \in \mathcal{S}$. (**Proof:** $\mathbb{E}_\nu^\pi(s) = \mathbb{E}_s^\pi \{ \sum_{n=1}^{+\infty} \sum_{t=1}^n r(X_t, Y_t) (1-\lambda)\lambda^{n-1} \} = \mathbb{E}_s^\pi \{ \sum_{t=1}^{+\infty} \sum_{n=t}^{+\infty} r(X_t, Y_t) (1-\lambda)\lambda^{n-1} \} = \mathbb{E}_s^\pi \{ \sum_{t=1}^{+\infty} \lambda^{t-1} r(X_t, Y_t) \}$)

- Suppose $\pi \in \mathcal{D}^{\text{HR}}$, then for each $s \in \mathcal{S}$, there exists a $\pi' \in \mathcal{D}^{\text{MR}}$ for which $U^{\pi'}(s) = U^\pi(s)$. (**Proof:** Note that $U^\pi(s) = \mathbb{E}_s^\pi \left\{ \sum_{t=1}^{+\infty} \lambda^{t-1} r(X_t, Y_t) \right\} = \sum_{t=1}^{+\infty} \sum_{j \in \mathcal{S}} \sum_{a \in \mathcal{A}} \lambda^{t-1} r(j, a) p^\pi(X_t = j, Y_t = a | X_1 = s)$. Fix $s \in \mathcal{S}$, so we only need to check $p^\pi(X_t = j, Y_t = a | X_1 = s) = p^{\pi'}(X_t = j, Y_t = a | X_1 = s)$. For each $j \in \mathcal{S}$ and $a \in \mathcal{A}$, define the randomized Markov decision rule δ'_t by $q_{\delta'_t(j)}(a) = p^\pi(Y_t = a | X_t = j, X_1 = s)$. Then $p^{\pi'}(Y_t = a | X_t = j) = p^\pi(Y_t = a | X_t = j, X_1 = s)$. Assume the conclusion holds for $t = 0, 1, \dots, n-1$. Then $p^{\pi'}(X_n = j, Y_n = a | X_1 = s) = p^{\pi'}(Y_n = a | X_n = j, X_1 = s) p^{\pi'}(X_n = j | X_1 = s) = p^\pi(Y_n = a | X_n = j, X_1 = s) p^{\pi'}(X_n = j | X_1 = s)$. Then by induction assumption, $p^\pi(X_n = j | X_1 = s) = \sum_{k \in \mathcal{S}} \sum_{a \in \mathcal{A}} p^\pi(X_{n-1} = k, Y_{n-1} = a | X_1 = s) p(j|k, a) = \sum_{k \in \mathcal{S}} \sum_{a \in \mathcal{A}} p^{\pi'}(X_{n-1} = k, Y_{n-1} = a | X_1 = s) p(j|k, a) = p^{\pi'}(X_n = j | X_1 = s)$.)
- Vector express for MDP: δ MD, define $r_\delta(s)$ and $p_\delta(j|s)$ by $r_\delta(s) := r(s, \delta(s))$, $p_\delta(j|s) = p(j|s, \delta(s))$. Denote $r_\delta = (r_\delta(1), \dots, r_\delta(|\mathcal{S}|))^T \in \mathbb{R}^{|\mathcal{S}|}$, $p_\delta = (p_\delta)_{(s,j)} = p(j|s, \delta(s))$. For MR δ , define $r_\delta(s) = \sum_{a \in \mathcal{A}} q_{\delta(s)}(a) r(s, a)$, $p_\delta(j|s) = \sum_{a \in \mathcal{A}} q_{\delta(s)}(a) p(j|s, a)$. The (s, j) -th component of the t -step transition probability matrix p_π^t satisfies $p_\pi^t(j|s) = [p_{\delta_1} p_{\delta_2} \dots p_{\delta_t}](j|s) = p^\pi(X_{t+1} = j | X_1 = s)$, $\mathbb{E}_s^\pi g(X_t) = \sum_{j \in \mathcal{S}} p_\pi^{t-1}(j|s) g(j) = (p_\pi^t g)_s$, and $U^\pi = \sum_{t=1}^{+\infty} \lambda^{t-1} p_\pi^{t-1} r_{\delta_t} = r_{\delta_1} + \lambda p_{\delta_1} (r_{\delta_1} + \lambda p_{\delta_2} r_{\delta_2} + \dots) = r_{\delta_1} + \lambda p_{\delta_1} U^{\pi_1}$. When π is stationary, $U = r_\delta + \lambda p_\delta U$.
- Define $\mathcal{L}U = \sup_{d \in \mathcal{D}^{\text{MD}}} \{r_d + \lambda p_d U\}$. Suppose there exists a $U \in \mathcal{U}$ for which (a) $U \geq \mathcal{L}U$, then $U \geq U^*$; (b) $U \leq \mathcal{L}U$, then $U \leq U^*$; (c) $U = \mathcal{L}U$, then $U = U^*$. (**Proof:** (a) $U \geq \sup_{d \in \mathcal{D}^{\text{MR}}} \{r_d + \lambda p_d U\} \geq r_{\delta_1} + \lambda p_{\delta_1} U \geq r_{\delta_1} + \lambda p_{\delta_1} (r_{\delta_2} + \lambda p_{\delta_2} U) \geq r_{\delta_1} + \lambda p_{\delta_1} r_{\delta_2} + \dots + \lambda^{n-1} p_{\delta_1} p_{\delta_2} \dots p_{\delta_{n-1}} r_{\delta_n} + \lambda^n p_\pi^n U \Rightarrow U - U^\pi \geq \lambda^n p_\pi^n U - \sum_{k=n}^{+\infty} \lambda^k p_\pi^k r_{\delta_{k+1}} \geq 0$; (b) $U \leq \mathcal{L}U \Rightarrow U \leq r_d + \lambda p_d U + \varepsilon 1 \Rightarrow (I - \lambda p_d)U \leq r_d + \varepsilon 1 \Rightarrow U \leq (I - \lambda p_d)^{-1} (r_d + \varepsilon 1) = U^\pi + \varepsilon (1 - \lambda)^{-1} 1_{|\mathcal{S}|}$.)
- If $0 \leq \lambda < 1$, \mathcal{L} is a contraction mapping on \mathcal{U} . (**Proof:** Let u and v in \mathcal{U} . For each $s \in \mathcal{S}$, assume that $\mathcal{L}v(s) \geq \mathcal{L}u(s)$ and let $a_s^* = \arg \max_{a \in \mathcal{A}} \{r(s, a) + \sum_{j \in \mathcal{S}} \lambda p(j|s, a) v(j)\}$. Then $0 \leq \mathcal{L}v(s) - \mathcal{L}u(s) \leq r(s, a_s^*) + \sum_{j \in \mathcal{S}} \lambda p(j|s, a_s^*) v(j) - r(s, a_s^*) - \sum_{j \in \mathcal{S}} \lambda p(j|s, a_s^*) u(j) = \lambda \sum_{j \in \mathcal{S}} p(j|s, a_s^*) (v(j) - u(j)) \leq \lambda \sum_{j \in \mathcal{S}} p(j|s, a_s^*) \|u - v\| = \lambda \|u - v\|$.)

3 统计学习理论

- $(X, Y) \sim P \in \mathcal{P}$, definite $(X_1, Y_1), \dots, (X_n, Y_n)$ i.i.d., $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$, $\mathcal{R}_n(f) = \mathbb{E}_{(X,Y) \in \mathcal{D}_n} l(X, Y)$. An algorithm A is a mapping from \mathcal{D}_n to function from $\mathcal{X} \rightarrow \mathcal{Y}$. Excess risk of A : $\mathcal{R}_P(A(\mathcal{D}_n)) - \mathcal{R}_P^*$. Expected error $\mathbb{E}[\mathcal{R}_P(A(\mathcal{D}_n))]$. An algorithm is called consistent in expectation for P iff $\mathbb{E}[\mathcal{R}_P(A(\mathcal{D}_n))] - \mathcal{R}_P^* \rightarrow 0$. PAC (probability approximately correct): for a given $\delta \in (0, 1)$ and $\epsilon > 0$, $\mathbb{P}(\mathcal{R}_P(A(\mathcal{D}_n)) - \mathcal{R}_P^* \leq \epsilon) \geq 1 - \delta$.
- 回归: $g(x) = \mathbb{E}[Y|X = x]$, $g_n(x, \mathcal{D}_n) = g_n(x)$, $\mathbb{E}\{[g_n(X) - Y]^2 | \mathcal{D}_n\} = \int_{\mathbb{R}^d} |g_n(x) - g(x)|^2 \mu(dx) + \mathbb{E}[g(X) - Y]^2$. A sequence of regression function estimates $\{g_n\}$ is called weakly consistent for a certain distribution of (X, Y) if $\lim_{n \rightarrow +\infty} \mathbb{E}\{\int [g_n(x) - g(x)] \mu(dx)\} = 0$; strongly consistent for a certain distribution if $\lim_{n \rightarrow +\infty} \int [g_n(x) - g(x)]^2 \mu(dx) = 0$ with probability 1; weakly universally consistent if for all distributions of (X, Y) with $\mathbb{E}[Y^2] < \infty$, \dots ; strongly universally consistent \dots .
- Penalized model: $g_n = \arg \min_f \left\{ \frac{1}{n} \sum_{i=1}^n |f(X_i) - Y_i|^2 + J_n(f) \right\}$. Penalized term for f : $J_n(f) = \lambda_n \int |f''(t)|^2 dt$, $J_{n,k}(f) = \lambda_n \int \sum_{t_1, \dots, t_k \in \{1, \dots, d\}} \left| \frac{\partial^k f}{\partial x_{t_1} \dots \partial x_{t_d}} \right|^2 dt$.
- Curse of dimensionality: let X, X_1, \dots, X_n i.i.d. \mathbb{R}^d uniformly distributed in $[0, 1]^d$. $d_\infty(d, n) = \mathbb{E}\{\min_{i=1, \dots, n} \|X - X_i\|_\infty\} = \int_0^\infty \mathbb{P}\{\min_{i=1, \dots, n} \|X - X_i\|_\infty > t\} dt = \int_0^\infty (1 - \mathbb{P}\{\min_{i=1, \dots, n} \|X - X_i\|_\infty < t\}) dt$. Since $\mathbb{P}\{\min_i \|X - X_i\|_\infty < t\} \leq n \mathbb{P}(\|X - X_1\|_\infty \leq t) \leq n(2t)^d$, 原式 $\geq \frac{d}{2(d+1)} n^{-\frac{1}{d}}$.
- No-Free lunch: Let $\{a_n\}$ be a sequence of positive numbers converging to 0. For every sequence of regression estimates, there exists a distribution of (X, Y) such that X is uniformly distributed on $[0, 1]$, $Y = g(X)$, g is ± 1 valued, and $\limsup_{n \rightarrow +\infty} \frac{\mathbb{E}\|g_n - g\|^2}{a_n} \geq 1$. (**Proof:** Let $\{p_i\}$ be a probability distribution and let $\mathcal{A} = \{\mathcal{A}_j\}$

be a partition of $[0, 1]$ such that \mathcal{A}_j is an interval of length p_j . Consider regression function indexed by a parameter c , $c = (c_1, c_2, \dots)$ where $c_j \in \{\pm 1\}$. Define $g^{(c)} : [0, 1] \rightarrow \{-1, 1\}$ by $g^{(c)}(x) = c_j$ if $x \in \mathcal{A}_j$ and $Y = g^{(c)}(x)$. For $x \in \mathcal{A}_j$, define $\bar{g}_n(x) = \frac{1}{p_j} \int_{\mathcal{A}_j} g_n(z) \mu(dz)$ to be the projection of g_n on \mathcal{A} . Then $\int_{\mathcal{A}_j} |g_n(x) - g^{(c)}(x)|^2 \mu(dx) = \int_{\mathcal{A}_j} |g_n(x) - \bar{g}_n(x)|^2 \mu(dx) + \int_{\mathcal{A}_j} |\bar{g}_n(x) - g^{(c)}(x)|^2 \mu(dx) \geq \int_{\mathcal{A}_j} |\bar{g}_n(x) - g^{(c)}(x)|^2 \mu(dx)$. Set $\hat{c}_{nj} = 1$ if $\int_{\mathcal{A}_j} g_n(z) \mu(dz) \geq 0$; $= -1$, otherwise. For $x \in \mathcal{A}_j$, if $\hat{c}_{nj} = 1$ and $c_j = -1$, then $\bar{g}_n(x) \geq 0$ and $g^{(c)}(x) = -1$, implying $|\bar{g}_n(x) - g^{(c)}(x)| \geq 1$; if $\hat{c}_{nj} = -1$ and $c_j = 1$, then $\bar{g}_n(x) < 0$ and $g^{(c)}(x) = 1 \Rightarrow |\bar{g}_n(x) - g^{(c)}(x)| \geq 1$. Therefore $\int_{\mathcal{A}} |\bar{g}_n(x) - g^{(c)}(x)|^2 \mu(dx) \geq 1_{\{\hat{c}_{nj} \neq c_j\}} \int_{\mathcal{A}_j} 1 \mu(dx) \geq 1_{\{\hat{c}_{nj} \neq c_j\}} p_j \geq 1_{\{\hat{c}_{nj} \neq c_j\}} 1_{\{\mu_n(\mathcal{A}_j) = 0\}} p_j \Rightarrow \mathbb{E}\{\int |g_n(x) - g^{(c)}(x)|^2 \mu(dx)\} \geq \sum_{j=1}^{+\infty} \mathbb{P}(\hat{c}_{nj} \neq c_j, \mu_n(\mathcal{A}_j) = 0) p_j := R_n(c)$. Now we randomize c . Let C_1, C_2, \dots be a sequence of i.i.d. random variables independent of X_1, X_2, \dots which satisfy $\mathbb{P}(c_1 = 1) = \mathbb{P}(c_1 = -1) = \frac{1}{2}$. Thus $\mathbb{E}R_n(C) = \sum_{j=1}^{+\infty} \mathbb{E}\mathbb{P}(\hat{C}_{nj} \neq C_j, \mu_n(\mathcal{A}_j) = 0) p_j \stackrel{\text{重期望}}{=} \sum_{j=1}^{+\infty} \mathbb{E}\{1_{\{\mu_n(\mathcal{A}_j)=0\}} \mathbb{P}(\hat{C}_{nj} \neq C_j | X_1, \dots, X_n)\} p_j = \frac{1}{2} \sum_{j=1}^{+\infty} \mathbb{P}(\mu_n(\mathcal{A}_j) = 0) p_j = \frac{1}{2} \sum_{j=1}^{+\infty} (1 - p_j)^n p_j$. On the other hand, $R_n(c) \leq \sum_{j=1}^{+\infty} \mathbb{P}(\mu_n(\mathcal{A}_j) = 0) p_j = \sum_{j=1}^{+\infty} (1 - p_j)^n p_j \Rightarrow \frac{R_n(c)}{\mathbb{E}R_n(C)} \leq 2$. By Fatou's lemma, $\mathbb{E}\{\limsup_{n \rightarrow +\infty} \frac{R_n(C)}{\mathbb{E}R_n(C)}\} \geq \limsup_{n \rightarrow +\infty} \{\frac{R_n(C)}{\mathbb{E}R_n(C)}\} = 1$, which implies that there exists $c \in C$ such that $\limsup_{n \rightarrow +\infty} \frac{R_n(C)}{\mathbb{E}R_n(C)} \geq 1 \Rightarrow \limsup_{n \rightarrow +\infty} \frac{\mathbb{E}\{\int |g_n(x) - g(x)|^2 \mu(dx)\}}{\frac{1}{2} \sum_{j=1}^{+\infty} (1 - p_j)^n p_j} \geq 1$. Let $\{a_n\}$ be a sequence of positive numbers converging to 0 with $\frac{1}{2} \geq a_1 \geq a_2 \geq \dots$, then there exists a probability $\{p_j\}$ such that $\sum_{j=1}^{+\infty} (1 - p_j)^n p_j \geq a_n, \forall n$.

- **Minimax lower Bounds:** (a) The sequence of positive numbers a_n is called the lower minimax rate of convergence for the \mathcal{P} if $\liminf_{n \rightarrow +\infty} \inf_{g_n} \sup_{P \in \mathcal{P}} \frac{\mathbb{E}\{\|g_n - g\|^2\}}{a_n} = c_1 > 0$. (b) a_n is called optimal rate of convergence for the class \mathcal{P} if it is a lower minimax rate of convergence and there is an estimate g_n such that $\limsup_{n \rightarrow +\infty} \sup_{P \in \mathcal{P}} \frac{\mathbb{E}\|g_n - g\|^2}{a_n} = c_n < \infty$.
- **Smoothness:** Let $q = k + \beta$ for some $k \in \mathbb{N}$ and $0 < \beta \leq 1$ and let $\rho > 0$. A function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is called (q, ρ) -smooth if for every $\alpha = (\alpha_1, \dots, \alpha_d), \alpha_i \in \mathbb{N}, \sum_{i=1}^d \alpha_i = k$, the partial derivative $\frac{\partial^k f}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}$ exists and satisfies $\left| \frac{\partial^k f}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}(x) - \frac{\partial^k f}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}(z) \right| \leq \rho \|x - z\|^\beta$. Let $\mathcal{F}^{(q, \rho)}$ be the set of all (q, ρ) -smooth functions f . Let $\mathcal{P}^{(q, \rho)}$ be the class of distributions (X, Y) such that (i) X is uniformly distributed on $[0, 1]^d$; (ii) $Y = g(X) + N$, where $X \perp\!\!\!\perp N$, and N is standard normal; (iii) $g \in \mathcal{F}^{q, \rho}$.
- Let u be an l -dimensional real vector, let C be a zero means random variables taking values in $\{-1, 1\}$ and let N be an l -dimensional standard normal independent of C . Set $Z = Cu + N$. Then the error probability of the Bayesian decision for C based on Z is $\mathcal{R}^* = \min_{g: \mathbb{R}^l \rightarrow \mathbb{R}} \mathbb{P}(g(Z) \neq C) = \Phi(-\|u\|)$. (**Proof:** $\mathbb{P}(C = 1) = \mathbb{P}(C = -1) = \frac{1}{2}, \mathbb{P}(Z|C = 1) = \mathcal{N}(u, I), \mathbb{P}(Z|C = -1) = \mathcal{N}(-u, I)$. By the Bayes formula, $\mathbb{P}(C = 1|Z = z) = \frac{\mathbb{P}(C=1)\mathbb{P}(Z|C=1)}{\mathbb{P}(C=1)\mathbb{P}(Z|C=1) + \mathbb{P}(C=-1)\mathbb{P}(Z|C=-1)} = \frac{1}{1 + \exp(\frac{\|Z - u\|^2}{2} - \frac{\|Z + u\|^2}{2})} = \frac{1}{1 + \exp(-2Z^T u)}$. Therefore, the optimal Bayes decision is $g^*(Z) = \text{sgn}(Z^T u)$, the risk $\mathcal{R}^* = \mathbb{P}(g^*(Z) \neq C) = \mathbb{P}(Z^T u < 0, C = 1) + \mathbb{P}(Z^T u > 0, C = -1) = \mathbb{P}(\|u\|^2 + u^T N < 0, C = 1) + \mathbb{P}(-\|u\|^2 + u^T N > 0, C = -1) = \frac{1}{2} \mathbb{P}(u^T N \leq -\|u\|^2) + \frac{1}{2} \mathbb{P}(u^T N > \|u\|^2) = \Phi(-\|u\|)$.)
- For the class $\mathcal{P}^{(q, \rho)}$, the sequence $a_n = n^{-\frac{2q}{2q+d}}$ is a lower minimax rate of convergence. In particular,

$$\liminf_{n \rightarrow \infty} \inf_{g_n} \sup_{P_{(X, Y)} \in \mathcal{P}^{(q, \rho)}} \frac{\mathbb{E}\|g_n - g\|^2}{\rho^{\frac{2d}{2q+d}} n^{-\frac{2q}{2q+d}}} \geq c_1 > 0.$$

证明分为 4 步. Step 1: 构造一个辅助函数 $g^{(c)}$. Set $M_n = \lceil (\rho^2 n)^{\frac{1}{2q+d}} \rceil$. Partition $[0, 1]^d$ by M_n^d cubes $\{A_{n,j}\}$ of side length $\frac{1}{M_n}$ and with centers $\{a_{n,j}\}$. Choose a function $\bar{f} : \mathbb{R}^d \rightarrow \mathbb{R}$ such that the support of \bar{f} is a subset of $[-\frac{1}{2}, \frac{1}{2}]^d, \int \bar{f}(x) dx > 0$ and $\bar{f} \in \mathcal{F}^{(q, 2^{\beta-1})}$. Define $f : \mathbb{R}^d \rightarrow \mathbb{R}$ by $f = \rho \bar{f}$. Let $c_n = (c_{n,1}, \dots, c_{n,M_n^d}) \in \mathcal{C}_n$ take values in $\{\pm 1\}$. $g^{(c_n)}(x) = \sum_{j=1}^{M_n^d} c_{n,j} f_{n,j}(x)$ where $f_{n,j}(x) = M_n^{-q} f(M_n(x - a_{n,j}))$.

Step 2: 证明 $g^{(c_n)} \in \mathcal{F}^{(q, \rho)}$. Let $\alpha = (\alpha_1, \dots, \alpha_d), \alpha_i \in \mathbb{N}$ and $\sum_{j=1}^d \alpha_j = k$. Set $D^\alpha = \frac{\partial^k}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}$. If $x, z \in A_{n,j}$, $|D^\alpha g^{(c_n)}(x) - D^\alpha g^{(c_n)}(z)| = |c_{n,j}| |D^\alpha f_{n,j}(x) - D^\alpha f_{n,j}(z)| \leq \rho \|x - z\|^\beta$. If $x \in A_{n,i}, z \in A_{n,j}$, choose \bar{x}, \bar{z} on the

line between x and z such that \bar{x} is on the boundary of $A_{n,i}$ and \bar{z} is on the boundary of $A_{n,j}$. $|D^\alpha g^{(c_n)}(x) - D^\alpha g^{(c_n)}(z)| \leq |c_{n,i} D^\alpha f_{n,i}(x)| + |c_{n,j} D^\alpha f_{n,j}(z)| = |c_{n,i}| |D^\alpha f_{n,i}(x) - D^\alpha f_{n,i}(\bar{x})| + |c_{n,j}| |D^\alpha f_{n,j}(z) - D^\alpha f_{n,j}(\bar{z})| \leq \rho 2^{\beta-1} (\|x - \bar{x}\|^\beta + \|z - \bar{z}\|^\beta) = \rho 2^\beta \left(\frac{\|x - \bar{x}\|^\beta}{2} + \frac{\|z - \bar{z}\|^\beta}{2} \right) \leq \rho 2^\beta \left(\frac{\|x - \bar{x}\|}{2} + \frac{\|z - \bar{z}\|}{2} \right)^\beta \leq \rho \|x - z\|^\beta$.

Step 3: Prove that $\liminf_{n \rightarrow +\infty} \inf_{g_n} \sup_{Y=g^{(c)}(X)+N, c \in \mathcal{C}_n} \frac{M_n^{2q}}{\rho^2} \mathbb{E} \|g_n - g^{(c)}\|^2 > 0$. $\{f_{n,j}\}$ forms a set of orthogonal basis.

Let g_n be an arbitrary estimate, and the projection \bar{g}_n of g_n to $\{g^{(c)} : c \in \mathcal{C}_n\}$ is given by $\bar{g}_n = \sum_{j=1}^{M_n} \tilde{c}_{n,j} f_{n,j}(x)$. $\|g_n - g^{(c)}\|^2 = \|g_n - \bar{g}_n\|^2 + \|g_n - g^{(c)}\|^2 \geq \|\bar{g}_n - g^{(c)}\|^2 = \sum_{j=1}^{M_n} \int_{A_{n,j}} (\tilde{c}_{n,j} f_{n,j}(x) - c_{n,j} f_{n,j}(x))^2 dx = \sum_{j=1}^{M_n} \int_{A_{n,j}} (\tilde{c}_{n,j} - c_{n,j})^2 f_{n,j}^2(x) dx = \int f^2(x) dx \sum_{j=1}^{M_n} (\tilde{c}_{n,j} - c_{n,j})^2 \frac{1}{M_n^{2q+d}}$. Define $\bar{c}_{n,j} = \text{sgn}(\tilde{c}_{n,j}) \Rightarrow |\tilde{c}_{n,j} - c_{n,j}| \geq \frac{|\tilde{c}_{n,j} - c_{n,j}|}{2} \Rightarrow \|g_n - g^{(c)}\|^2 \geq \int f^2(x) dx \frac{1}{4} \frac{1}{M_n^{2q+d}} \sum_{j=1}^{M_n} (\bar{c}_{n,j} - c_{n,j})^2 = \frac{\rho^2}{M_n^{2q}} \int \bar{f}^2(x) dx \frac{1}{M_n^d} \sum_{j=1}^{M_n} 1_{\{\bar{c}_{n,j} \neq c_{n,j}\}}$.

Step 4: Prove that $\liminf_{n \rightarrow +\infty} \inf_{\bar{c}_n} \sup_{c_n} \frac{1}{M_n^d} \sum_{j=1}^{M_n} \mathbb{P}(\bar{c}_{n,j} \neq c_{n,j}) > 0$. Now we randomize c_n . Let $c_{n,1}, \dots, c_{n,M_n^d}$ be i.i.d. random variables independent of $(X_1, N_1), \dots, (X_n, N_n)$, $\mathbb{P}(C_{n,1} = 1) = \mathbb{P}(C_{n,1} = -1) = \frac{1}{2}$. $\bar{c}_{n,j}$ can be interpreted as a decision on $C_{n,j}$ using \mathcal{D}_n . Let $\bar{C}_{n,j} = 1$ if $\mathbb{P}(\bar{C}_{n,j} = 1 | \mathcal{D}_n) \geq \frac{1}{2}$. Therefore, $\inf_{\bar{c}_n} \sup_{c_n} \frac{1}{M_n^d} \sum_{j=1}^{M_n} \mathbb{P}(\bar{c}_{n,j} \neq c_{n,j}) \geq \inf_{\bar{c}_n} \frac{1}{M_n^d} \sum_{j=1}^{M_n} \mathbb{P}(\bar{c}_{n,j} \neq C_{n,j}) \geq \frac{1}{M_n^d} \sum_{j=1}^{M_n} \mathbb{P}(\bar{C}_{n,j} \neq C_{n,j}) = \mathbb{P}(\bar{C}_{n,1} \neq C_{n,1}) = \mathbb{E}\{\mathbb{P}(\bar{C}_{n,1} \neq C_{n,1} | X_1, \dots, X_n)\}$. Let X_{i_1}, \dots, X_{i_t} be those $X_i \in A_{n,1}$, $(Y_{i_1}, \dots, Y_{i_t}) = C_{n,1}(f_{n,1}(X_{i_1}), \dots, f_{n,1}(X_{i_t})) + (N_{i_1}, \dots, N_{i_t})$. By the latest "•", $\mathbb{E}\{\mathbb{P}(\bar{C}_{n,1} \neq C_{n,1} | X_1, \dots, X_n)\} = \mathbb{E}\Phi\left(-\sqrt{\sum_{r=1}^t f_{n,1}^2(X_{i_r})}\right) = \mathbb{E}\Phi\left(-\sqrt{\sum_{i=1}^n f_{n,1}^2(X_i)}\right) \geq \Phi\left(-\sqrt{\mathbb{E} \sum_{i=1}^n f_{n,1}^2(X_i)}\right) \geq \Phi(-\sqrt{\int f^2(x) dx}) > 0$.

- Uniform laws of large numbers: Set $Z = (X, Y)$, $Z_i = (X_i, Y_i)$, $g_f(x, y) = |f(x) - y|^2$ for $f \in \mathcal{F}_n$, $G_n = \{g_f : f \in \mathcal{F}_n\}$, consider the limit $\lim_{n \rightarrow +\infty} \sup_{g \in \mathcal{G}_n} \left| \frac{1}{n} \sum_{i=1}^n g(Z_i) - \mathbb{E}g(Z) \right|$.
- Hoeffding's inequality: $g : \mathbb{R}^d \rightarrow [0, B]$, $\begin{cases} \mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n g(Z_i) - \mathbb{E}\{g(Z)\}\right| > \epsilon\right) \leq 2e^{-\frac{2n\epsilon^2}{B^2}} \\ \mathbb{P}\left(\sup_{g \in \mathcal{G}_n} \left|\frac{1}{n} \sum_{i=1}^n g(Z_i) - \mathbb{E}\{g(Z)\}\right| > \epsilon\right) \leq 2|\mathcal{G}_n|e^{-\frac{2n\epsilon^2}{B^2}} \end{cases}$. For finite class \mathcal{G} satisfying $\sum_{n=1}^{+\infty} |\mathcal{G}_n|e^{-\frac{2n\epsilon^2}{B^2}} < \infty$ for all $\epsilon > 0$, by Borel-Cantelli lemma, the event $\sup_{g \in \mathcal{G}_n} \left|\frac{1}{n} \sum_{i=1}^n g(Z_i) - \mathbb{E}\{g(Z)\}\right| > \epsilon$ occurs f.o.
- Let $\epsilon > 0$ and \mathcal{G} be a set of functions $\mathbb{R}^d \rightarrow \mathbb{R}$. Every finite collection of functions $g_1, \dots, g_N : \mathbb{R}^d \rightarrow \mathbb{R}$ with the property that for every $g \in \mathcal{G}$ there is a $j = j(g) \in [N]$ such that $\|g - g_j\|_\infty < \epsilon$ is called an ϵ -cover of \mathcal{G} w.r.t. $\|\cdot\|_\infty$. Let $\mathcal{N}(\epsilon, \mathcal{G}, \|\cdot\|_\infty)$ or $\mathcal{N}_\infty(\epsilon, \mathcal{G})$ be the smallest ϵ -cover of \mathcal{G} w.r.t. $\|\cdot\|_\infty$.
- For $n \in \mathbb{N}$, let \mathcal{G}_n be a set of functions $g : \mathbb{R}^d \rightarrow [0, B]$ and let $\epsilon > 0$, then $\mathbb{P}\left(\sup_{g \in \mathcal{G}_n} \left|\frac{1}{n} \sum_{i=1}^n g(Z_i) - \mathbb{E}\{g(Z)\}\right| > \epsilon\right) \leq 2\mathcal{N}_\infty\left(\frac{\epsilon}{3}, \mathcal{G}_n\right)e^{-\frac{2n\epsilon^2}{9B^2}}$. (Proof: Let $\mathcal{G}_{n, \frac{\epsilon}{3}}$ be an $\frac{\epsilon}{3}$ -cover of \mathcal{G}_n w.r.t. $\|\cdot\|_\infty$ of minimal cardinality. Fix $g \in \mathcal{G}_n$, there exists $\bar{g} \in \mathcal{G}_{n, \frac{\epsilon}{3}}$ such that $\|g - \bar{g}\|_\infty < \frac{\epsilon}{3}$. Since $\left|\frac{1}{n} \sum_{i=1}^n g(Z_i) - \mathbb{E}g(Z)\right| \leq \left|\frac{1}{n} \sum_{i=1}^n (g(Z_i) - \bar{g}(Z_i))\right| + \left|\frac{1}{n} \sum_{i=1}^n \bar{g}(Z_i) - \mathbb{E}\{\bar{g}(Z)\}\right| + |\mathbb{E}\bar{g}(Z) - \mathbb{E}g(Z)| \leq \frac{2\epsilon}{3} + \left|\frac{1}{n} \sum_{i=1}^n \bar{g}(Z_i) - \mathbb{E}\{\bar{g}(Z)\}\right|$. Thus $\mathbb{P}\left(\sup_{g \in \mathcal{G}_n} \left|\frac{1}{n} \sum_{i=1}^n g(Z_i) - \mathbb{E}\{g(Z)\}\right| > \epsilon\right) \leq \mathbb{P}\left(\sup_{g \in \mathcal{G}_{n, \frac{\epsilon}{3}}} \left|\frac{1}{n} \sum_{i=1}^n g(Z_i) - \mathbb{E}\{g(Z)\}\right| > \frac{\epsilon}{3}\right)$. Then use Hoeffding's inequality.)
- Let $\epsilon > 0$ and \mathcal{G} be a set of functions $\mathbb{R}^d \rightarrow \mathbb{R}$, $1 \leq p < \infty$, and ν be a probability measure on \mathbb{R}^d . (a) Every finite collection of functions $g_1, \dots, g_N : \mathbb{R}^d \rightarrow \mathbb{R}$ with the property that for every $g \in \mathcal{G}$ there is a $j = j(g) \in [N]$ such that $\|g - g_j\|_{L_p(\nu)} < \epsilon$ is called a ϵ -cover of \mathcal{G} . Similarly define $\mathcal{N}(\epsilon, \mathcal{G}, \|\cdot\|_{L_p(\nu)})$. (b) Let $Z^{1:n} = (Z_1, \dots, Z_n) \subset \mathbb{R}^d$ and ν_n be the corresponding empirical measure, then $\|f\|_{L_p(\nu_n)} := \left\{\frac{1}{n} \sum_{i=1}^n |f(Z_i)|^p\right\}^{\frac{1}{p}}$ and similarly define $\mathcal{N}_p(\epsilon, \mathcal{G}, Z^{1:n})$.

- Packing numbers: (a) Every finite collection of functions $g_1, \dots, g_N \in \mathcal{G}$ with $\|g_j - g_k\|_{L_p(\nu)} \geq \epsilon$ for all $1 \leq j < k \leq N$ is called ϵ -packing of \mathcal{G} with $\|\cdot\|_{L_p(\nu)}$. The largest ϵ -packing is denoted as $\mathcal{M}(\epsilon, \mathcal{G}, \|\cdot\|_{L_p(\nu)})$. Similarly define $\mathcal{M}(\epsilon, \mathcal{G}, Z^{1:n})$.
- $\mathcal{M}(2\epsilon, \mathcal{G}, \|\cdot\|_{L_p(\nu)}) \leq \mathcal{N}(\epsilon, \mathcal{G}, \|\cdot\|_{L_p(\nu)}) \leq \mathcal{M}(\epsilon, \mathcal{G}, \|\cdot\|_{L_p(\nu)}), \mathcal{M}(2\epsilon, \mathcal{G}, Z^{1:n}) \leq \mathcal{N}(\epsilon, \mathcal{G}, Z^{1:n}) \leq \mathcal{M}(\epsilon, \mathcal{G}, Z^{1:n})$.
- Let \mathcal{F} be a set of functions $\mathbb{R}^d \rightarrow \mathbb{R}$. Assume that \mathcal{F} is a linear vector space of dimension D . Then for arbitrary $R > 0, \epsilon > 0$, and $z_1, \dots, z_n \in \mathbb{R}^d$ such that $\mathcal{N}_2(\epsilon, \{f \in \mathcal{F} : \frac{1}{n} \sum_{i=1}^n |f(z_i)|^2 \leq R^2\}, Z^{1:n}) \leq (\frac{4R+\epsilon}{\epsilon})^D$.
- Let \mathcal{A} be a class of subsets of \mathbb{R}^d and $n \in \mathbb{N}$. (a) For $z_1, \dots, z_n \in \mathbb{R}^d$, define $s(\mathcal{A}, \{z_1, \dots, z_n\}) = |\{A \cap \{z_1, \dots, z_n\} : A \in \mathcal{A}\}|$.
- Let \mathcal{G} be a subset of \mathbb{R}^d of size n . We say \mathcal{A} shatters \mathcal{G} if $s(\mathcal{A}, \mathcal{G}) = 2^n$. The n th shatter coefficient of \mathcal{A} is $S(\mathcal{A}, n) = \max_{\{z_1, \dots, z_n\} \subset \mathbb{R}^d} s(\mathcal{A}, \{z_1, \dots, z_n\})$, the maximum number of different subsets of n points that can be picked out by set from \mathcal{A} .
- Let \mathcal{A} be a class of subsets of \mathbb{R}^d with $\mathcal{A} \neq \emptyset$. The VC dimension $V_{\mathcal{A}}$ of \mathcal{A} is defined by $V_{\mathcal{A}} = \sup\{n \in \mathbb{N}, S(\mathcal{A}, n) = 2^n\}$.
- $S(\mathcal{A}, n) \leq \sum_{i=0}^{V_{\mathcal{A}}} \binom{n}{i}$.

- Let \mathcal{G} be a set of functions $g : \mathbb{R}^d \rightarrow [0, B]$. For any $n \in \mathbb{N}$ and $\epsilon > 0$, $\mathbb{P}\left\{\sup_{g \in \mathcal{G}} \left|\frac{1}{n} \sum_{i=1}^n g(Z_i) - \mathbb{E}[g(Z)]\right| > \epsilon\right\} \leq 8\mathbb{E}\mathcal{N}_1(\frac{\epsilon}{8}, \mathcal{G}, Z^{1:n})e^{-\frac{n\epsilon^2}{128B^2}}$. (**Proof:** Step 1: Symmetrization. Let $Z'^{1:n}$ be i.i.d. samples from the same distribution and independent of $Z^{1:n}$ and g^* be a function $g \in \mathcal{G}$ $\left|\frac{1}{n} \sum_{i=1}^n g(Z_i) - \mathbb{E}g(Z)\right| > \epsilon$ if there exists such a function. Otherwise, let g^* be an arbitrary function in \mathcal{G} . g^* depends on $Z^{1:n}$. $\mathbb{P}\left\{\left|\mathbb{E}[g^*(Z)|Z^{1:n}] - \frac{1}{n} \sum_{i=1}^n g^*(Z'_i)\right| > \frac{\epsilon}{2} |Z^{1:n}\right\} \leq \frac{\text{Var}(g^*(Z)|Z^{1:n})}{n(\frac{\epsilon}{2})^2} \leq \frac{B^2/4}{n\epsilon^2/4} = \frac{B^2}{n\epsilon^2} \leq \frac{1}{2}$ for $n \geq \frac{2B^2}{\epsilon^2}$. Thus we have

$$\begin{aligned} \mathbb{P}\left\{\sup_{g \in \mathcal{G}} \left|\frac{1}{n} \sum_{i=1}^n g(Z_i) - \frac{1}{n} \sum_{i=1}^n g(Z'_i)\right| > \frac{\epsilon}{2}\right\} &\geq \mathbb{P}\left\{\left|\frac{1}{n} \sum_{i=1}^n g^*(Z_i) - \frac{1}{n} \sum_{i=1}^n g^*(Z'_i)\right| > \frac{\epsilon}{2}\right\} \\ &\geq \mathbb{P}\left\{\left|\frac{1}{n} \sum_{i=1}^n g^*(Z_i) - \mathbb{E}[g^*(Z)|Z^{1:n}]\right| > \epsilon, \left|\frac{1}{n} \sum_{i=1}^n g^*(Z'_i) - \mathbb{E}[g^*(Z)|Z^{1:n}]\right| \leq \frac{\epsilon}{2}\right\} \\ &= \mathbb{E}\left\{1_{\left\{\left|\frac{1}{n} \sum_{i=1}^n g^*(Z_i) - \mathbb{E}[g^*(Z)|Z^{1:n}]\right| > \epsilon\right\}} \mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n g^*(Z'_i) - \mathbb{E}[g^*(Z)|Z^{1:n}]\right| \leq \frac{\epsilon}{2} |Z^{1:n}\right)\right\} \\ &\geq \frac{1}{2} \mathbb{P}\left\{\left|\frac{1}{n} \sum_{i=1}^n g^*(Z_i) - \mathbb{E}[g^*(Z)|Z^{1:n}]\right| > \epsilon\right\} \end{aligned}$$

Therefore, $2\mathbb{P}\left\{\sup_{g \in \mathcal{G}} \left|\frac{1}{n} \sum_{i=1}^n g(Z_i) - \frac{1}{n} \sum_{i=1}^n g(Z'_i)\right| > \frac{\epsilon}{2}\right\} \geq \mathbb{P}\left\{\sup_{g \in \mathcal{G}} \left|\frac{1}{n} \sum_{i=1}^n g(Z_i) - \mathbb{E}[g(Z)]\right| > \epsilon\right\}$.

Step 2: Introduction of additive randomness by random signs. Let U_1, \dots, U_n be independent and uniformly distributed over $\{-1, 1\}$ and independent $Z^{1:n}$ and $Z'^{1:n}$.

$$\begin{aligned} \mathbb{P}\left\{\sup_{g \in \mathcal{G}} \left|\frac{1}{n} \sum_{i=1}^n [g(Z_i) - g(Z'_i)]\right| > \frac{\epsilon}{2}\right\} &= \mathbb{P}\left\{\sup_{g \in \mathcal{G}} \left|\frac{1}{n} \sum_{i=1}^n U_i [g(Z_i) - g(Z'_i)]\right| > \frac{\epsilon}{2}\right\} \\ &\leq \mathbb{P}\left\{\sup_{g \in \mathcal{G}} \left|\frac{1}{n} \sum_{i=1}^n U_i g(Z_i)\right| > \frac{\epsilon}{4}\right\} + \mathbb{P}\left\{\sup_{g \in \mathcal{G}} \left|\frac{1}{n} \sum_{i=1}^n U_i g(Z'_i)\right| > \frac{\epsilon}{4}\right\} \\ &= 2\mathbb{P}\left\{\sup_{g \in \mathcal{G}} \left|\frac{1}{n} \sum_{i=1}^n U_i g(Z_i)\right| > \frac{\epsilon}{4}\right\} \end{aligned}$$

Step 3: Conditioning and introduction of a covering on $Z^{1:n}$. Let $\mathcal{G}_{\frac{\epsilon}{8}}$ be an L_1 $\frac{\epsilon}{8}$ -cover of \mathcal{G} in $Z^{1:n}$. Fix $g \in \mathcal{G}$, then there exists $\bar{g} \in \mathcal{G}_{\frac{\epsilon}{8}}$ s.t. $\frac{1}{n} \sum_{i=1}^n |g(Z_i) - \bar{g}(Z_i)| < \frac{\epsilon}{8}$. $\left| \frac{1}{n} \sum_{i=1}^n U_i g(Z_i) \right| = \left| \frac{1}{n} \sum_{i=1}^n U_i \bar{g}(Z_i) + \frac{1}{n} \sum_{i=1}^n U_i [g(Z_i) - \bar{g}(Z_i)] \right| \leq \left| \frac{1}{n} \sum_{i=1}^n U_i \bar{g}(Z_i) \right| + \frac{\epsilon}{8}$. Thus

$$\mathbb{P} \left\{ \exists g \in \mathcal{G} : \left| \frac{1}{n} \sum_{i=1}^n U_i g(Z_i) \right| > \frac{\epsilon}{4} \right\} \leq \mathbb{P} \left\{ \exists g \in \mathcal{G}_{\frac{\epsilon}{8}} : \left| \frac{1}{n} \sum_{i=1}^n U_i \bar{g}(Z_i) \right| > \frac{\epsilon}{8} \right\} \leq |\mathcal{G}_{\frac{\epsilon}{8}}| \max_{g \in \mathcal{G}_{\frac{\epsilon}{8}}} \mathbb{P} \left\{ \left| \frac{1}{n} \sum_{i=1}^n U_i g(Z_i) \right| > \frac{\epsilon}{8} \right\}$$

Step 4: Application of Hoeffding's inequality: $-B \leq U_i g(Z_i) \leq B \Rightarrow \mathbb{P}\{|\frac{1}{n} \sum_{i=1}^n U_i g(Z_i)| > \frac{\epsilon}{8}\} \leq 2 \exp\left(-\frac{2n(\frac{\epsilon}{8})^2}{(2B)^2}\right) = 2 \exp\left(-\frac{n\epsilon^2}{128B^2}\right)$.

- Let \mathcal{G} be an r -dimensional vector space of real functions on \mathbb{R}^d and $\mathcal{A} = \{\{z : g(z) \geq 0\} : g \in \mathcal{G}\}$. Then $V_{\mathcal{A}} \leq r$.