Theoretical Machine Learning

Lectured by Zhihua Zhang LaTeXed by Chengxin Gong

April 23, 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

Outline 1.1 (**机器学习的主要任务**) 生成、预测、决策. 生成: $X_1, \cdots, X_n \sim F$, 推断分析 F, 无监督学习, GAN, GPT, $\cdots$. 预测: 数据对 $(X^{(1)}, Y^{(1)}), \cdots, (X^{(n)}, Y^{(n)})$, $X^{(i)} \in \mathbb{R}^d$ 输入变量, $f: \mathcal{X} \to \mathcal{Y}, x \in \mathcal{X}, y \in \mathcal{Y}$, 归因, 有监督学习. 决策: 强化学习, Agent$\leftarrow$action, state, reward$\rightarrow$ 环境.

Outline 1.2 (**求解问题的途径**) 参数/非参数, 频率 (MLE)/贝叶斯.

Outline 1.3 (**误差模型**) 有监督: $X = (X_1, \cdots, X_d)^T \in \mathbb{R}^d$, 回归: $Y \in \mathbb{R}$; 分类: $Y \in \{0,1\}(\{-1,1\}, \{1, \cdots, M\}, \{0,1\}^M)$; X 随机, Random design(生成模型), $Y = g(X) + \varepsilon \overset{\text{or}}{=} g(X, Z), Y^{(i)} = g(X^{(i)}, Z^{(i)})$; 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

Definition 2.1 (Statistical decision theory) Consider a state space Ω, data space $\mathcal{D}$, model $\mathcal{P} = \{p(\theta, x)\}$, action space $\mathscr{A}$. Loss function: $\mathcal{L}: \Omega \times \mathscr{A} \to [-\infty, +\infty]$, measurable, nonnegative. A measurable function $\delta: \mathcal{D} \to \mathscr{A}$ is called a nonrandomized decision rule. Risk function is defined as $\mathcal{R}(\theta, \delta) = \int \mathcal{L}(\theta, \delta(x)) \mathrm{d}P_\theta(x) = \mathbb{E}_\theta \mathcal{L}(\theta, \delta(X))$. Randomized decision: for each $X = x$, $\delta(x)$ is a probability distribution: $[A|X = x] \sim \delta_x$. Risk function for δ: $\mathcal{R}(\theta, \delta) = \mathbb{E}_\theta \mathcal{L}(\theta, A) = \mathbb{E}_\theta \mathbb{E}_a \mathcal{L}(\theta, A|X) = \iint \mathcal{L}(\theta, a) \mathrm{d}\delta_x(a) \mathrm{d}P_\theta(x)$.

Example 2.1 (Parameter estimation) $\theta \in \Omega, \mathscr{A} = \Omega, \mathcal{L}(\theta, a) = \|\theta - a\|_2^2 \overset{\text{or}}{=} \|\theta - a\|_p^p (p \geq 1) \overset{\text{or}}{=} \int \log \frac{P_\theta(x)}{P_a(x)} P_\theta(x) \mathrm{d}m(x)$(KL). $\mathcal{R} = \mathrm{Var}(a) + \mathrm{bias}^2(a)$. Bregmass loss: $\phi: \mathbb{R}^d \to \mathbb{R}$ describe any strictly convex differentiable function. Then $\mathcal{L}_\phi(\theta, a) = \phi(a) - \phi(\theta) - (\phi - a)^T \nabla \phi(a)$.

Example 2.2 (Testing) $\mathscr{A} = \{0, 1\}$ 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 $\mathcal{R}(\theta, \delta) = \mathbb{P}_\theta(A = 1) 1_{\theta \in \Omega_0} + \mathbb{P}_\theta(A = 0) 1_{\theta \in \Omega_1}$.

Definition 2.2 (Admissibility) 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.

Definition 2.3 (Bayes rule) The maximum risk $\bar{\mathcal{R}}(\delta) = \sup_{\theta \in \Omega} \mathcal{R}(\theta, \delta)$ and the Bayes risk $r(\Lambda, \delta) = \int \mathcal{R}(\theta, \delta) \mathrm{d}\Lambda(\theta) = \int \mathcal{L}(\theta, \delta) \mathrm{d}\mathbb{P}(x, \theta)$ ($\Lambda(\theta)$ 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 $\delta^*: \sup_{\theta \in \Omega} \mathcal{R}(\theta, \delta^*) = \inf_\delta \sup_{\theta \in \Omega} \mathcal{R}(\theta, \delta)$.

Theorem 2.1 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.

Property 2.1 $p(\theta|x) = \frac{p_\theta(x)\lambda(\theta)}{\int p_\theta(x)\lambda(\theta)\mathrm{d}\theta} := \frac{p_\theta(x)\lambda(\theta)}{m(x)}$. Define the posterior risk of δ: $r(\delta|X = x) = \int \mathcal{L}(\theta, \delta(x)) \mathrm{d}\mathbb{P}(\theta|x)$. The Bayes risk $r(\Lambda, \delta)$ satisfies that $r(\Lambda, \delta) = \int r(\delta|x) \mathrm{d}M(x)$. Let $\hat{\delta}(x)$ be the value of δ that minimizes $r(\delta|x)$. Then $\hat{\delta}$ is the Bayes rule.

Example 2.3 (Application to supervised learning: regression) $(X, Y) \in \mathcal{X} \times \mathcal{Y}, f: \mathcal{X} \to \mathcal{Y}, \mathscr{A} = \Omega = \mathcal{Y}, \mathcal{D} = \mathcal{X}, \delta = f, \mathcal{L}(Y, f(X)) = \|Y - f(X)\|_p^p, p \geq 1$, risk $R_f = \iint \mathcal{L}(y, f(x)) \mathrm{d}\mathbb{P}(x, y) = \mathbb{E}[\mathcal{L}(Y, f(X))] = \mathbb{E}[\mathbb{E}\mathcal{L}(Y, f(X))|X]$. When $p = 2$, $r(f|X = x) = \int \mathcal{L}(y, f(x)) \mathrm{d}\mathbb{P}(y|x) = \int |y - f(x)|^2 \mathrm{d}\mathbb{P}(y|x)$. 回归函数 $g(x) := \int y \mathrm{d}\mathbb{P}(y|x) \Rightarrow R_f = \mathbb{E}|Y - f(X)|^2 = \mathbb{E}|Y - g(X) + g(X) - f(X)|^2 = \mathbb{E}|Y - g(X)|^2 + \mathbb{E}|g(X) - f(X)|^2 \geq \mathbb{E}|Y - g(X)|^2$.

Example 2.4 (Application to supervised learning: pattern classification) $Y \in \{0, 1\}, p_0 = P(Y = 0), p_1 = \mathbb{P}(Y = 1) = 1 - p_0, \mathbb{E}[\mathcal{L}(Y, f(X))] = \mathbb{P}(Y \neq f(X))$. The Bayesian 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 $\mathbb{E}[\mathcal{L}(Y, f(X))|X = x] = \begin{cases} \mathbb{E}[\mathcal{L}(Y, 0)|X = x] = \mathcal{L}(0,0)\mathbb{P}(Y = 0|X = x) + \mathcal{L}(1,0)\mathbb{P}(Y = 1|X = x) \\ \mathbb{E}[\mathcal{L}(Y, 1)|X = x] = \mathcal{L}(0,1)\mathbb{P}(Y = 0|X = x) + \mathcal{L}(1,1)\mathbb{P}(Y = 1|X = x) \end{cases}$, 比较两者大小. □

Property 2.2 (**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 $0 \leq \mathbb{P}(\hat{f}(X) \neq Y) - \mathbb{P}(f(X) \neq Y) \leq 2\int_{\mathcal{X}} |\hat{g}(x) - g(x)| \mu(\mathrm{d}x) \leq 2(\int_{\mathcal{X}} |\hat{g}(x) - g(x)|^2 \mu(dx))^{\frac{1}{2}}$. 回到 Example 2.4. $f(x) = 1_{\{\frac{p(x|y=1)}{p(x|y=0)} \geq \frac{p_0(\mathcal{L}(0,1) - \mathcal{L}(0,0))}{p_1(\mathcal{L}(1,0) - \mathcal{L}(1,1))}\}}$, 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 $f(x) = 1\{L(x) \geq \eta\}$.

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

Definition 2.4 (Confusion table) Ture Positive Rate: TPR = $\mathbb{P}(\hat{Y}=1|Y=1)$; False Negative Rate: FNR = 1 - TPR, type II error; False Positive Rate: FPR = $\mathbb{P}(\hat{Y}=1|Y=0)$, type I error; True Negative Rate: TNR = 1 - 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 $F_1=\frac{2\text{TPR}}{1+\text{TPR}+\frac{p_0}{p_1}\text{FPR}}$.

Theorem 2.2 (N-P lemma) Optimization: maximize TPR subject to FPR $\le \alpha, \alpha\in[0,1]$. 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 $\mathbb{E}[Q(x)|Y=0]\le\alpha$. Suppose the likelihood functions $p(x|y)$ are continuous. Then the optimal predictor is a deterministic LRT.

Proof Let η be the threshold for an LRT such that the predictor $Q_\eta(x)=1\{\alpha(x)\ge\eta\}$ has 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 $\text{FPR}(Q)\le\alpha$, $\mathcal{R}_{Q_\eta}=p_1\eta\alpha+p_1(1-\beta)\le p_1\eta\text{FPR}(Q)+p_1(1-\text{TPR}(Q))\le p_1\eta\alpha+p_1(1-\text{TPR}(Q))\Rightarrow\text{TPR}(Q)\le\beta$. □

Definition 2.5 (ROC (Receiver operating character) curve) y-axis is TPR and x-axis is FPR.

Proposition 2.1 (1) The points $(0,0)$ and $(1,1)$ are on the ROC curve; (2) The ROC must lie above the main diagnal; (3) The ROC curve is concave.

Proof (2): Fix $\alpha\in(0,1)$ and consider a randomized rate TPR = FPR = α, $Q(x)\equiv\alpha$; (3): Consider two rules (FPR(η_1), TPR(η_1)) and (FPR(η_2), TPR(η_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 (FPR, TPR) = $(t\text{FPR}(\eta_1)+(1-t)\text{FPR}(\eta_2), t\text{TPR}(\eta_1)+(1-t)\text{FPR}(\eta_2))$. Fixing FPR $\le t\text{FPR}(\eta_1)+(1-t)\text{FPR}(\eta_2)$, TPR $\ge t\text{TPR}(\eta_1)+(1-t)\text{TPR}(\eta_2)$. □

Definition 2.6 (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: $\{1,2,\cdots,N\}$; 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 $p_t(\cdot|s_t,a)$.

Definition 2.7 (Decision rules) Prescribe a procedure for action selection in each state at a specified decision epoch. Four cases: (1) Markovian and Deterministic: $\delta_t:\mathcal{S}\to\mathcal{A}$; (2) M and Randomized: $\delta_t:\mathcal{S}\to\Delta(\mathcal{A})(q_{\delta_t(s)}(a))$; (3) History-dependent and D: $h_t=(s_1,a_1,\cdots,s_{t-1},a_{t-1},s_t)=(h_{t-1},a_{t-1},s_t),\mathcal{H}_1=\mathcal{S},\mathcal{H}_2=\mathcal{S}\times\mathcal{A}\times\mathcal{S},\cdots,\delta_t:\mathcal{H}_t\to\mathcal{A}$; (4) HR: $\delta_t:\mathcal{H}_t\times\Delta(\mathcal{A})$. A policy $\pi=(\delta_1,\delta_2,\cdots,\delta_{N-1})$ is stationary if $\delta_1=\delta_2=\cdots=\delta$ for $t\in T$.

Definition 2.8 Let $\pi=(\delta_1,\cdots,\delta_{N-1})$ in HR and $R_t:=r_t(X_t,Y_t)$ denote the random reward, $R_N:=r_N(X_N)$, $R:=(R_1,\cdots,R_N)$. The expected total reward $U_N^\pi(s):=\mathbb{E}^\pi\{\sum\limits_{t=1}^{N-1}r_t(X_t,Y_t)+r_N(X_N)|X_1=s\}$. Assume $|r_t(s,a)|\le M<\infty$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$. Optimal policy: $U_N^{\pi^*}(s)\ge U_N^\pi(s),s\in\mathcal{S}$. ε-optimal policy: $U_N^{\pi_\varepsilon^*}(s)+\varepsilon>U_N^\pi(s),s\in\mathcal{S}$. The value of the MDP: $U_N^*(s)=\sup_{\pi\in\mathcal{D}^{\text{HR}}}U_N^\pi(s),s\in\mathcal{S}$.

Definition 2.9 (Finite-Horizon Policy Evaluation) $V_t^\pi(h_t)=\mathbb{E}^\pi\{\sum\limits_{k=t}^{N-1}r_k(X_k,Y_k)+r_N(X_N)|h_t\},V_N^\pi(h_N)=r_N(s),\pi\in\mathcal{D}^{\text{HD}}$. 由重期望公式, $V_t^\pi(h_t)=r_t(s_t,\delta_t(h_t))+\mathbb{E}_{h_t}^\pi V_{t+1}^\pi(h_t,\delta_t(h_t),X_{t+1})=r_t(s_t,\delta_t(h_t))+\sum\limits_{j\in\mathcal{S}}V_{t+1}^\pi(h_t,\delta_t(h_t),j)p(j|s_t,\delta_t(h_t))$. Consider randomness (i.e. $\pi\in\mathcal{D}^{\text{HR}}$): $V_t^\pi(h_t)=\sum\limits_{a\in\mathcal{A}}q_{\delta_t(h_t)}(a)\{r_t(s_t,a)+\sum\limits_{j\in\mathcal{S}}V_{t+1}^\pi(h_t,a,j)p(j|s_t,a)\}$. Computational complexity: let $K=|\mathcal{S}|,L=|\mathcal{A}|$, at decision epoch t, $K^{t+1}L^t$ histories, $K^2\sum\limits_{i=0}^{N-1}(KL)^i$ 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 $U_N^\pi(s)$.

Theorem 2.3 (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 $t = 1, 2, \cdots, N-1$ 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 $h_t \in \mathcal{H}_t, t = 1, \cdots, N$ 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_t, π. Now assume that $V_t(h_t) \geq V_t^*(h_t)$ for all $h_t \in \mathcal{H}_t$ for $t = n+1, \cdots, N$. Let $\pi' = (\delta_1', \cdots, \delta_{N-1}')$ be an arbitrary policy in $\mathcal{D}^{\text{HR}}$. For $t = n$, the Bellman equations $V_n(h_n) = \sup_{a\in\mathcal{A}}\{r_n(s_t, a_t) + \sum_{j\in\mathcal{S}} p(j|s_n, a)V_{n+1}(h_n, a, j)\} \geq \sup_{a\in\mathcal{A}}\{r_n(s_n, a) + \sum_{j\in\mathcal{S}} p_n(j|s_n, a)V_{n+1}^*(h_n, a, j)\} \geq \sup_{a\in\mathcal{A}}\{r_n(s_n, a) + \sum_{j\in\mathcal{S}} p_n(j|s_n, a)V_{n+1}^{\pi'}(h_n, a, j)\} \geq V_n^{\pi'}(h_n)$. 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 $\pi' = (\delta_1', \cdots, \delta_{N-1}')$ by choosing $\delta_n'(h_n)$ to satisfy $r_n(s_n, \delta_n'(h_n)) + \sum_{j\in\mathcal{S}} p_n(j|s_n, \delta_n'(h_n))V_{n+1}(h_n, \delta_n'(h_n)) + \varepsilon \geq V_n(h_n)$. 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 $t = n+1, \cdots, N$. For $t = n$, $V_n^{\pi'}(h_n) = r_n(s_n, \pi_n'(h_n)) + \sum_{j\in\mathcal{S}} p_n(j|s_n, \delta_n^{\pi'}(h_n))V_{n+1}^{\pi'}(h_n, \delta_n^{\pi'}(h_n), j) \geq V_n(h_n) - (N-n)\varepsilon$. □

Remark 2.1 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. $U_N^*(s) = \sup_{\pi\in\mathcal{D}^{\text{HR}}} U_N^\pi(s) = \sup_{\pi\in\mathcal{D}^{\text{HD}}} U_N^\pi(s) \overset{?}{=} \sup_{\pi\in\mathcal{D}^{\text{MD}}} U_N^\pi(s)$.

Theorem 2.4 Let $V_t^*, t = 1, \cdots, N$ be solutions of Bellman Equations. Then (a) For each $t = 1, \cdots, N, V_t^*(h_t)$ 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 $t = n+1, \cdots, N$. Then $V_n^*(h_n) = \sup_{a\in\mathcal{A}}\{r_t(s_t, a) + \sum_{j\in\mathcal{S}} p_t(j|s_t, a)V_{t+1}^*(j)\} = V_n^*(s_t)$. □

Definition 2.10 (Backward Indcution (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.

Remark 2.2 (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 eact t, only $(N-1)LK^2$ multiplications, ease computation and storage cost (because there are $(L^K)^{N-1}$ DM policies).

Definition 2.11 (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 $\pi = (\delta_1, \delta_2, \cdots) \in \mathcal{D}^{\text{HR}} : U^\pi(s) = \lim_{N\to+\infty} \mathbb{E}_s^\pi\{\sum_{t=1}^N \lambda^{t-1}r(X_t, Y_t)\} = \mathbb{E}_s^\pi\{\sum_{t=1}^{+\infty} \lambda^{t-1}r(X_t, Y_t)\}$. 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 $\lambda : \mathbb{P}(\nu = n) = (1-\lambda)\lambda^{n-1}, n = 1, 2, \cdots$.

Theorem 2.5 Suppose ν has a GD(λ). Then $U^\pi(s) = U_\nu^\pi(s)$ for all $s \in \mathcal{S}$.

Proof $\mathbb{E}_\nu^\pi(s) = \mathbb{E}_s^\pi\{\sum_{n=1}^{+\infty}\sum_{t=1}^{n} r(X_t, Y_t)(1-\lambda)\lambda^{n-1}\} = \mathbb{E}_s^\pi\{\sum_{t=1}^{+\infty}\sum_{n=t}^{+\infty} r(X_t, Y_t)(1-\lambda)\lambda^{n-1}\} = \mathbb{E}_s^\pi\{\sum_{t=1}^{+\infty} \lambda^{t-1}r(X_t, Y_t)\}$ □

Theorem 2.6 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 $U^\pi(s) = \mathbb{E}_s^\pi\{\sum_{t=1}^{+\infty} \lambda^{t-1}r(X_t, Y_t)\} = \sum_{t=1}^{+\infty}\sum_{j\in\mathcal{S}}\sum_{a\in\mathcal{A}} \lambda^{t-1}r(j, a)p^\pi(X_t = j, Y_t = a|X_1 = s)$. 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 $t=0,1,\cdots,n-1$. 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, $p^\pi(X_n=j|X_1=s)=\sum\limits_{k\in\mathcal{S}}\sum\limits_{a\in\mathcal{A}}p^\pi(X_{n-1}=k,Y_{n-1}=a|X_1=s)p(j|k,a)=\sum\limits_{k\in\mathcal{S}}\sum\limits_{a\in\mathcal{A}}p^{\pi'}(X_{n-1}=k,Y_{n-1}=a|X_1=s)p(j|k,a)=p^{\pi'}(X_n=j|X_1=s)$. $\square$

Definition 2.12 (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 $r_\delta=(r_\delta(1),\cdots,r_\delta(|\mathcal{S}|))^T\in\mathbb{R}^{|\mathcal{S}|},p_\delta=(p_\delta)_{(s,j)}=p(j|s,\delta(s))$. For MR δ, define $r_\delta(s)=\sum\limits_{a\in\mathcal{A}}q_{\delta(s)}(a)r(s,a),p_\delta(j|s)=\sum\limits_{a\in\mathcal{A}}q_{\delta(s)}(a)p(j|s,a)$. The (s,j)-th component of the t-step transition probability matrix p_π^t satisfies $p_\pi^t(j|s)=[p_{\delta_1}p_{\delta_2}\cdots p_{\delta_t}](j|s$ $p^\pi(X_{t+1}=j|X_1=s),\mathbb{E}_s^\pi g(X_t)=\sum\limits_{j\in\mathcal{S}}p_\pi^{t-1}(j|s)g(j)=(p_\pi^t g)_s$, and $U^\pi=\sum\limits_{t=1}^{+\infty}\lambda^{t-1}p_\pi^{t-1}r_{\delta_t}=r_{\delta_1}+\lambda p_{\delta_1}(r_{\delta_1}+\lambda p_{\delta_2}r_{\delta_2}+\cdots)=r_{\delta_1}+\lambda p_{\delta_1}U^{\pi_1}$. When π is stationary, $U=r_\delta+\lambda p_\delta U$.

Theorem 2.7 Define $\mathscr{L}U=\sup_{d\in\mathcal{D}^{\mathrm{MD}}}\{r_d+\lambda p_d U\}$. Suppose there exists a $U\in\mathcal{U}$ for which (a) $U\geq\mathscr{L}U$, then $U\geq U^*$; (b) $U\leq\mathscr{L}U$, then $U\leq U^*$; (c) $U=\mathscr{L}U$, then $U=U^*$.

Proof (a) $U\geq\sup_{\delta\in\mathcal{D}^{\mathrm{MR}}}\{r_d+\lambda p_d U\}\geq r_{\delta_1}+\lambda p_{\delta_1}U\geq r_{\delta_1}+\lambda p_{\delta_1}(r_{\delta_2}+\lambda p_{\delta_2}U)\geq r_{\delta_1}+\lambda p_{\delta_1}r_{\delta_2}+\cdots+\lambda^{n-1}p_{\delta_1}p_{\delta_2}\cdots p_{\delta_{n-1}}r_{\delta_n}+\lambda^n p_\pi^n U\Rightarrow U-U^\pi\geq\lambda^n p_\pi^n U-\sum\limits_{k=n}^{+\infty}\lambda^k p_\pi^k r_{\delta_{k+1}}\geq 0$; (b) $U\leq\mathscr{L}U\Rightarrow U\leq r_d+\lambda p_d U+\varepsilon 1\Rightarrow(I-\lambda p_d)U\leq r_d+\varepsilon 1\Rightarrow U\leq(I-\lambda p_d)^{-1}(r_d+\varepsilon 1)=U^\pi+\varepsilon(1-\lambda)^{-1}1_{|\mathcal{S}|}$. $\square$

Theorem 2.8 If $0\leq\lambda<1$, $\mathscr{L}$ is a contraction mapping on $\mathcal{U}$.

Proof Let u and v in $\mathcal{U}$. For each $s\in\mathcal{S}$, assume that $\mathscr{L}v(s)\geq\mathscr{L}u(s)$ and let $a_s^*=\arg\max_{a\in\mathcal{A}}\{r(s,a)+\sum\limits_{j\in\mathcal{S}}\lambda p(j|s,a)v(j)\}$. Then $0\leq\mathscr{L}v(s)-\mathscr{L}u(s)\leq r(s,a_s^*)+\sum\limits_{j\in\mathcal{S}}\lambda p(j|s,a_j^*)v(j)-r(s,a_j^*)-\sum\limits_{j\in\mathcal{S}}\lambda p(j|s,a_s^*)u(j)=\lambda\sum\limits_{j\in\mathcal{S}}p(j|s,a_s^*)(v(j)-u(j))\leq\lambda\sum\limits_{j\in\mathcal{S}}p(j|s,a_s^*)||u-v\|=\lambda\|u-v\|$. $\square$

3 统计学习理论

Definition 3.1 $(X,Y)\sim P\in\mathcal{P}$, definite $(X_1,Y_1),\cdots,(X_n,Y_n)$ i.i.d., $\mathcal{D}_n=\{(X_1,Y_1),\cdots,(X_n,Y_n)\}$, $\mathcal{R}_n(f)=\mathbb{E}_{(X,Y)\in\mathcal{D}_n}l(X,Y)$. An algorithm A is a mapping from $\mathcal{D}_n$ to function from $\mathcal{X}\to\mathcal{Y}$. 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 $\mathbb{E}[\mathcal{R}_P(A(\mathcal{D}_n))]-\mathcal{R}_P^*\to 0$. PAC (probability approximately correct): for a given $\delta\in(0,1)$ and $\epsilon>0$, $\mathbb{P}(\mathcal{R}_P(A(\mathcal{D}_n))-\mathcal{R}_P^*\leq\epsilon)\geq 1-\delta$.

Definition 3.2 (Consistency) $g(x)=\mathbb{E}[Y|X=x],g_n(x,\mathcal{D}_n)=g_n(x),\mathbb{E}\{|g_n(X)-Y|^2|\mathcal{D}_n\}=\int_{\mathbb{R}^d}|g_n(x)-g(x)|^2\mu(\mathrm{d}x)+\mathbb{E}|g(X)-Y|^2$. A sequence of regression function estimates $\{g_n\}$ is called weakly consistent for a certain distribution of (X,Y) if $\lim\limits_{n\to+\infty}\mathbb{E}\{\int[g_n(x)-g(x)]\mu(\mathrm{d}x)\}=0$; strongly consistent for a certain distribution if $\lim\limits_{n\to+\infty}\int[g_n(x)-g(x)]^2\mu(\mathrm{d}x)=0$ with probability 1; weakly universally consistent if for all distributions of (X,Y) with $\mathbb{E}[Y^2]<\infty$, $\cdots$; strongly universally consistent $\cdots$.

Definition 3.3 (Penalized model) $g_n=\arg\min_f\{\frac{1}{n}\sum\limits_{i=1}^n|f(X_i)-Y_i|^2+J_n(f)\}$. Penalized term for f: $J_n(f)=\lambda_n\int|f''(t)|^2\mathrm{d}t,J$ $\lambda_n\int\sum\limits_{t_1,\cdots,t_k\in\{1,\cdots,d\}}|\frac{\partial f^k}{\partial x_{t_1}\cdots\partial x_{t_d}}|^2\mathrm{d}t$.

Property 3.1 (Curse of dimensionality) Let $X,X_1,\cdots,X_n$ i.i.d. $\mathbb{R}^d$ uniformly distributed in $[0,1]^d$. $d_\infty(d,n)=\mathbb{E}\{\min_{i=1,\cdots,n}\|X-X_i\|_\infty\}=\int_0^\infty\mathbb{P}\{\min_{i=1,\cdots,n}\|X-X_i\|_\infty>t\}\mathrm{d}t=\int_0^\infty(1-\mathbb{P}\{\min_{i=1,\cdots,n}\|X-X_i\|_\infty<t\})\mathrm{d}t$. 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 $\geq\frac{d}{2(d+1)}n^{-\frac{1}{d}}$.

Theorem 3.1 (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 $\limsup_{n\to+\infty}\frac{\mathbb{E}\|g_n-g\|^2}{a_n}\geq 1$.

Proof Let $\{p_i\}$ be a probability distribution and let $\mathscr{A}=\{\mathscr{A}_j\}$ be a partition of $[0,1]$ such that $\mathscr{A}_j$ is an interval of length p_j. Consider regression function indexed by a parameter c, $c=(c_1,c_2,\cdots)$ where $c_j\in\{\pm1\}$. Define $g^{(c)}:[0,1]\to\{-1,1\}$ by $g^{(c)}(x)=c_j$ if $x\in\mathscr{A}_j$ and $Y=g^{(c)}(x)$. For $x\in\mathscr{A}_j$, define $\bar{g}_n(x)=\frac{1}{p_j}\int_{\mathscr{A}_j}g_n(z)\mu(\mathrm{d}z)$ to be the projection of g_n on $\mathscr{A}$. Then $\int_{\mathscr{A}_j}|g_n(x)-g^{(c)}(x)|^2\mu(\mathrm{d}x)=\int_{\mathscr{A}_j}|g_n(x)-\bar{g}_n(x)|^2\mu(\mathrm{d}x)+\int_{\mathscr{A}}|\bar{g}_n(x)-g^{(c)}(x)|^2\mu(\mathrm{d}x)\geq$

$\int_{\mathscr{A}} |\bar{g}_n(x)-g^{(c)}(x)|^2\mu(\mathrm{d}x)$. Set $\hat{c}_{nj}=1$ if $\int_{\mathscr{A}_j} g_n(z)\mu(\mathrm{d}z)\geq 0$; $=-1$, otherwise. For $x\in\mathscr{A}_j$, 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 $g^{(c)}(x)=1\Rightarrow|\bar{g}_n(x)-g^{(c)}(x)|^2\geq 1$. Therefore $\int_{\mathscr{A}}|\bar{g}_n(x)-g^{(c)}(x)|^2\mu(\mathrm{d}x)\geq 1_{\{\hat{c}_{nj}\neq c_j\}}\int_{\mathscr{A}_j}1\mu(\mathrm{d}x)\geq 1_{\{\hat{c}_{nj}\neq c_j\}}p_j\geq 1_{\{\hat{c}_{nj}\neq c_j\}}1_{\{\mu_n(\mathscr{A}_j)=0\}}p_j\Rightarrow\mathbb{E}\{\int|g_n(x)-g^{(c)}(x)|^2\mu(\mathrm{d}x)\}\geq\sum_{j=1}^{+\infty}\mathbb{P}(\hat{c}_{nj}\neq c_j,\mu_n(\mathscr{A}_j)=0)p_j:=R_n(c)$. Now we randomize c. Let $C_1,C_2,\cdots$ be a sequence of i.i.d. random variables independent of $X_1,X_2,\cdots$ which satisfy $\mathbb{P}(c_1=1)=\mathbb{P}(c_1=-1)=\frac{1}{2}$. Thus $\mathbb{E}R_n(C)=\sum_{j=1}^{+\infty}\mathbb{E}\mathbb{P}(\hat{C}_{nj}\neq C_j,\mu_n(\mathscr{A}_j)=0)p_j\overset{\text{重期望}}{=}\sum_{j=1}^{+\infty}\mathbb{E}\{1_{\{\mu_n(\mathscr{A}_j)=0\}}\mathbb{P}(\hat{C}_{nj}\neq C_j|X_1,\cdots,X_n)\}p_j=\frac{1}{2}\sum_{j=1}^{+\infty}\mathbb{P}(\mu_n(\mathscr{A}_j)=0)p_j=\frac{1}{2}\sum_{j=1}^{+\infty}(1-p_j)^n p_j$. On the other hand, $R_n(c)\leq\sum_{j=1}^{+\infty}\mathbb{P}(\mu_n(\mathscr{A}_j)=0)p_j=\sum_{j=1}^{+\infty}(1-p_j)^n p_j\Rightarrow\frac{R_n(c)}{\mathbb{E}R_n(C)}\leq 2$. By Fatou's lemma, $\mathbb{E}\{\limsup_{n\to+\infty}\frac{R_n(C)}{\mathbb{E}R_n(C)}\}\geq\limsup_{n\to+\infty}\{\frac{R_n(C)}{\mathbb{E}R_n(C)}\}=1$, which implies that there exists $c\in C$ such that $\limsup_{n\to+\infty}\frac{R_n(C)}{\mathbb{E}R_n(C)}\geq 1\Rightarrow\limsup_{n\to+\infty}\frac{\mathbb{E}\{\int|g_n(x)-g(x)|^2\mu(\mathrm{d}x)\}}{\frac{1}{2}\sum_{j=1}^{+\infty}(1-p_j)^n p_j}\geq 1$. Let $\{a_n\}$ be a sequence of positive numbers converging to 0 with $\frac{1}{2}\geq a_1\geq a_2\geq\cdots$, then there exists a probability $\{p_j\}$ such that $\sum_{j=1}^{+\infty}(1-p_j)^n p_j\geq a_n,\forall n$.

□

Definition 3.4 (Minimax lower bounds) (a) The sequence of positive numbers a_n is called the lower minimax rate of convergence for the $\mathcal{P}$ if $\liminf_{n\to+\infty}\inf_{g_n}\sup_{P\in\mathcal{P}}\frac{\mathbb{E}\{\|g_n-g\|^2\}}{a_n}=c_1>0$. (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 $\limsup_{n\to+\infty}\sup_{P\in\mathcal{P}}\frac{\mathbb{E}\|g_n-g\|^2}{a_n}=c_n<\infty$.

Definition 3.5 (Smoothness) Let $q=k+\beta$ for some $k\in\mathbb{N}$ and $0<\beta\leq 1$ and let $\rho>0$. A function $f:\mathbb{R}^d\to\mathbb{R}$ is called (q,ρ)-smooth if for every $\alpha=(\alpha_1,\cdots,\alpha_d),\alpha_i\in\mathbb{N},\sum_{i=1}^{d}\alpha_i=k$, the partial derivative $\frac{\partial^k f}{\partial x_1^{\alpha_1}\cdots\partial x_d^{\alpha_d}}$ exists and satisfies $\left|\frac{\partial^k f}{\partial x_1^{\alpha_1}\cdots\partial x_d^{\alpha_d}}(x)-\frac{\partial^k f}{\partial x_1^{\alpha_1}\cdots\partial x_d^{\alpha_d}}(z)\right|\leq\rho\|x-z\|^\beta$. Let $\mathscr{F}^{(q,\rho)}$ 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) $g\in\mathscr{F}^{q,\rho}$.

Lemma 3.1 Let u be an l-dimensional real vector, let C be a zero means random variables takeing 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 $\mathcal{R}^*=\min_{g:\mathbb{R}^l\to\mathbb{R}}\mathbb{P}(g(Z)\neq C)=\Phi(-\|u\|)$.

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, $\mathbb{P}(C=1|Z=z)=\frac{\mathbb{P}(C=1)\mathbb{P}(Z|C=1)}{\mathbb{P}(C=1)\mathbb{P}(Z|C=1)+\mathbb{P}(C=-1)\mathbb{P}(Z|C=-1)}=\frac{1}{1+\exp(\frac{\|Z-u\|^2}{2}-\frac{\|Z+u\|^2}{2})}=\frac{1}{1+\exp(-2Z^Tu)}$. Therefore, the optimal Bayes decision is $g^*(Z)=\mathrm{sgn}(Z^Tu)$, the risk $\mathcal{R}^*=\mathbb{P}(g^*(Z)\neq C)=\mathbb{P}(Z^Tu<0,C=1)+\mathbb{P}(Z^Tu>0,C=-1)=\mathbb{P}(\|u\|^2+u^TN<0,C=1)+\mathbb{P}(-\|u\|^2+u^TN>0,C=-1)=\frac{1}{2}\mathbb{P}(u^TN\leq-\|u\|^2)+\frac{1}{2}\mathbb{P}(u^TN>\|u\|^2)=\Phi(-\|u\|)$. □

Theorem 3.2 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,

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

Proof Step 1: Construct an auxiliary function $g^{(c)}$. 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 $\bar{f}:\mathbb{R}^d\to\mathbb{R}$ such that the support of $\bar{f}$ is a subset of $[-\frac{1}{2},\frac{1}{2}]^d$, $\int\bar{f}^2(x)\mathrm{d}x>0$ and $\bar{f}\in\mathscr{F}^{(q,2^{\beta-1})}$. Define $f:\mathbb{R}^d\to\mathbb{R}$ by $f=\rho\bar{f}$. Let $c_n=(c_{n,1},\cdots,c_{n,M_n^d})\in\mathcal{C}_n$ take values in $\{\pm1\}$. $g^{(c_n)}(x)=\sum_{j=1}^{M_n^d}c_{n,j}f_{n,j}(x)$ where $f_{n_j}(x)=M_n^{-q}f(M_n(x-a_{n,j}))$.

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

Step 3: Prove that $\liminf_{n\to+\infty}\inf_{g_n}\sup_{Y=g^{(c)}(X)+N,c\in\mathcal{C}_n}\frac{M_n^{2q}}{\rho^2}\mathbb{E}\|g_n-g^{(c)}\|^2>0$. $\{f_{n,j}\}$ 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 $\bar{g}_n = \sum_{j=1}^{M_n} \tilde{c}_{n,j} f_{n,j}(x)$. Then $\|g_n - g^{(c)}\|^2 = \|g_n - \bar{g}_n\|^2 + \|g_n - g^{(c)}\|^2 \geq \|\bar{g}_n - g^{(c)}\|^2 = \sum_{j=1}^{M_n^d} \int_{A_{n,j}} (\tilde{c}_{n,j} f_{n,j}(x) - c_{n,j} f_{n,j}(x))^2 \mathrm{d}x = \sum_{j=1}^{M_n^d} \int_{A_{n,j}} (\tilde{c}_{n,j} - c_{n,k})^2 f_{n,j}^2(x) \mathrm{d}x = \int f^2(x)\mathrm{d}x \sum_{j=1}^{M_n^d} (\tilde{c}_{n,j} - c_{n,j})^2 \frac{1}{M_n^{2q+d}}$. Define $\bar{c}_{n,j} = \mathrm{sgn}(\tilde{c}_{n,j}) \Rightarrow |\tilde{c}_{n,j} - c_{n,j}| \geq \frac{|\bar{c}_{n,j} - c_{n,j}|}{2} \Rightarrow \|g_n - g^{(c)}\|^2 \geq \int f^2(x)\mathrm{d}x \frac{1}{4} \frac{1}{M_n^{2q+d}} \sum_{j=1}^{M_n^d} (\bar{c}_{n,j} - c_{n,j})^2 = \frac{\rho^2}{M_n^{2q}} \int \bar{f}^2(x)\mathrm{d}x \frac{1}{M_n^d} \sum_{j=1}^{M_n^d} 1_{\{\bar{c}_{n,j} \neq c_{n,j}\}}$.

Step 4: Prove that $\liminf_{n\to+\infty} \inf_{\bar{c}_n} \sup_{c_n} \frac{1}{M_n^d} \sum_{j=1}^{M_n^d} \mathbb{P}(\bar{c}_{n,j} \neq c_{n,j}) > 0$. Now we randomize c_n. Let $c_{n,1}, \cdots, c_{n,M_n^d}$ be i.i.d. random variables independent of $(X_1, N_1), \cdots, (X_n, N_n)$, $\mathbb{P}(C_{n,1} = 1) = \mathbb{P}(C_{n,1} = -1) = \frac{1}{2}$. $\bar{c}_{n,j}$ 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, $\inf_{\bar{c}_n} \sup_{c_n} \frac{1}{M_n^d} \sum_{j=1}^{M_n^d} \mathbb{P}(\bar{c}_{n,j} \neq c_{n,,j}) \geq \inf_{\bar{c}_n} \frac{1}{M_n^d} \sum_{j=1}^{M_n^d} \mathbb{P}(\bar{c}_{n,j} \neq C_{n,j}) \geq \frac{1}{M_n^d} \sum_{j=1}^{M_n^d} \mathbb{P}(\bar{C}_{n,j} \neq C_{n,j}) = \mathbb{P}(\bar{C}_{n,1} \neq C_{n,1}) = \mathbb{E}\{\mathbb{P}(\bar{C}_{n,1} \neq C_{n,1}|X_1, \cdots, X_n)\}$. Let $X_{i_1}, \cdots, X_{i_t}$ be those $X_i \in A_{n,1}$, $(Y_{i,1}, \cdots, Y_{i_t}) = C_{n,1}(f_{n,1}(X_{i_1}), \cdots, f_{n,1}(X_{i_t})) + (N_{i_1}, \cdots, N_{i_t})$. By lemma 3.1, $\mathbb{E}\{\mathbb{P}(\bar{C}_{n,1} \neq C_{n,1}|X_1, \cdots, X_n)\} = \mathbb{E}\Phi\left(-\sqrt{\sum_{r=1}^{t} f_{n,1}^2(X_{i_r})}\right) = \mathbb{E}\Phi\left(-\sqrt{\sum_{i=1}^{n} f_{n,1}^2(X_i)}\right) \geq \Phi\left(-\sqrt{\mathbb{E}\sum_{i=1}^{n} f_{n,1}^2(X_i)}\right) \geq \Phi\left(-\sqrt{\int f^2(x)\mathrm{d}x}\right) > 0$. □

Definition 3.6 (Uniform laws of large numbers) Set $Z = (X, Y), Z_i = (X_i, Y_i), g_f(x, y) = |f(x) - y|^2$ for $f \in \mathscr{F}_n, G_n = \{g_f : f \in \mathscr{F}_n\}$, consider the limit $\lim_{n\to+\infty} \sup_{g\in\mathscr{G}_n} \left|\frac{1}{n}\sum_{i=1}^{n} g(Z_i) - \mathbb{E}g(Z)\right|$.

Lemma 3.2 (Hoeffding's inequality) $g : \mathbb{R}^d \to [0, B]$, $\begin{cases} \mathbb{P}\left(|\frac{1}{n}\sum_{i=1}^{n} g(Z_i) - \mathbb{E}\{g(Z)\}| > \epsilon\right) \leq 2e^{-\frac{2n\epsilon^2}{B^2}} \\ \mathbb{P}\left(\sup_{g\in\mathscr{G}_n} |\frac{1}{n}\sum_{i=1}^{n} g(Z_i) - \mathbb{E}\{g(Z)\}| > \epsilon\right) \leq 2|\mathscr{G}_n|e^{-\frac{2n\epsilon^2}{B^2}} \end{cases}$. For finite class $\mathscr{G}$ satisfying $\sum_{n=1}^{+\infty} |\mathscr{G}_n|e^{-\frac{2n\epsilon^2}{B^2}} < \infty$ for all $\epsilon > 0$, by Borel-Cantelli lemma, the event $\sup_{g\in\mathscr{G}_n} |\frac{1}{n}\sum_{i=1}^{n} g(Z_i) - \mathbb{E}\{g(Z)\}| > \epsilon$ occurs f.o.

Definition 3.7 (Covering number) Let $\epsilon > 0$ and $\mathscr{G}$ be a set of functions $\mathbb{R}^d \to \mathbb{R}$. Every finite collection of functions $g_1, \cdots, g_N : \mathbb{R}^d \to \mathbb{R}$ with the property that for every $g \in \mathscr{G}$ there is a $j = j(g) \in [N]$ such that $\|g - g_j\|_\infty < \epsilon$ is called an ϵ-cover of $\mathscr{G}$ w.r.t. $\|\cdot\|_\infty$. Let $\mathscr{N}(\epsilon, \mathscr{G}, \|\cdot\|_\infty)$ or $\mathscr{N}_\infty(\epsilon, \mathscr{G})$ be the smallest ϵ-cover of $\mathscr{G}$ w.r.t. $\|\cdot\|_\infty$.

Theorem 3.3 For $n \in \mathbb{N}$, let $\mathscr{G}_n$ be a set of functions $g : \mathbb{R}^d \to [0, B]$ and let $\epsilon > 0$. Then

$$\mathbb{P}\left(\sup_{g\in\mathscr{G}_n} |\frac{1}{n}\sum_{i=1}^{n} g(Z_i) - \mathbb{E}\{g(Z)\}| > \epsilon\right) \leq 2\mathscr{N}_\infty\left(\frac{\epsilon}{3}, \mathscr{G}_n\right) \exp\left(-\frac{2n\epsilon^2}{9B^2}\right).$$

Proof Let $\mathscr{G}_{n,\frac{\epsilon}{3}}$ be an $\frac{\epsilon}{3}$-cover of $\mathscr{G}_n$ w.r.t. $\|\cdot\|_\infty$ of minimal cardinality. Fix $g \in \mathscr{G}_n$, there exists $\bar{g} \in \mathscr{G}_{n,\frac{\epsilon}{3}}$ such that $\|g - \bar{g}\|_\infty < \frac{\epsilon}{3}$. Since $|\frac{1}{n}\sum_{i=1}^{n} g(Z_i) - \mathbb{E}g(Z)| \leq |\frac{1}{n}\sum_{i=1}^{n}(g(Z_i) - \bar{g}(Z_i))| + |\frac{1}{n}\sum_{i=1}^{n} \bar{g}(Z_i) - \mathbb{E}\{\bar{g}(Z)\}| + |\mathbb{E}\bar{g}(Z) - \mathbb{E}g(Z)| \leq \frac{2\epsilon}{3} + |\frac{1}{n}\sum_{i=1}^{n} \bar{g}(Z_i) - \mathbb{E}\{\bar{g}(Z)\}|$. Thus $\mathbb{P}\left(\sup_{g\in\mathscr{G}_n} |\frac{1}{n}\sum_{i=1}^{n} g(Z_i) - \mathbb{E}\{g(Z)\}| > \epsilon\right) \leq \mathbb{P}\left(\sup_{g\in\mathscr{G}_{n,\frac{\epsilon}{3}}} |\frac{1}{n}\sum_{i=1}^{n} g(Z_i) - \mathbb{E}\{g(Z)\}| > \frac{\epsilon}{3}\right)$. Then use Hoeffding's inequality. □

Definition 3.8 Let $\epsilon > 0$ and $\mathscr{G}$ be a set of functions $\mathbb{R}^d \to \mathbb{R}$, $1 \leq p < \infty$, and ν be a probability measure on $\mathbb{R}^d$. (a) Every finite collection of functions $g_1, \cdots, g_N : \mathbb{R}^d \to \mathbb{R}$ with the property that for every $g \in \mathscr{G}$ there is a $j = j(g) \in [N]$ such that $\|g - g_j\|_{L_p(\nu)} < \epsilon$ is called a ϵ-cover of $\mathscr{G}$. Similarly define $\mathscr{N}(\epsilon, \mathscr{G}, \|\cdot\|_{L_p(\nu)})$. (b) Let $Z^{1:n} = (Z_1, \cdots, Z_n) \subset \mathbb{R}^d$ and ν_n be the corresponding empirical measure, then $\|f\|_{L_p(\nu_n)} := \{\frac{1}{n}\sum_{i=1}^{n} |f(Z_i)|^p\}^{\frac{1}{p}}$ and similarly define $\mathscr{N}_p(\epsilon, \mathscr{G}, Z^{1:n})$.

Definition 3.9 (Packing number) (a) Every finite collection of functions $g_1, \cdots, g_N \in \mathscr{G}$ with $\|g_j - g_k\|_{L_p(\nu)} \geq \epsilon$ for all $1 \leq j < k \leq N$ is called ϵ-packing of $\mathscr{G}$ with $\|\cdot\|_{L_p(\nu)}$. The largest ϵ-packing is denoted as $\mathscr{M}(\epsilon, \mathscr{G}, \|\cdot\|_{L_p(\nu)})$. Similarly define $\mathscr{M}(\epsilon, \mathscr{G}, Z^{1:n})$.

Property 3.2 (Covering number v.s. packing number)

$$\mathscr{M}(2\epsilon, \mathscr{G}, \|\cdot\|_{L_p(\nu)}) \le \mathscr{N}(\epsilon, \mathscr{G}, \|\cdot\|_{L_p(\nu)}) \le \mathscr{M}(\epsilon, \mathscr{G}, \|\cdot\|_{L_p(\nu)}),$$
$$\mathscr{M}(2\epsilon, \mathscr{G}, Z^{1:n}) \le \mathscr{N}(\epsilon, \mathscr{G}, Z^{1:n}) \le \mathscr{M}(\epsilon, \mathscr{G}, Z^{1:n}).$$

Theorem 3.4 Let $\mathscr{F}$ be a set of functions $\mathbb{R}^d \to \mathbb{R}$. Assume that $\mathscr{F}$ is a linear vector space of dimension D. Then for arbitrary $R > 0, \epsilon > 0$, and $z_1, \cdots, z_n \in \mathbb{R}^d$, $\mathscr{N}_2(\epsilon, \{f \in \mathscr{F} : \frac{1}{n}\sum\limits_{i=1}^{n} |f(z_i)|^2 \le R^2\}, Z^{1:n}) \le \left(\frac{4R+\epsilon}{\epsilon}\right)^D$.

Definition 3.10 Let $\mathscr{A}$ be a class of subsets of $\mathbb{R}^d$ and $n \in \mathbb{N}$. For $z_1, \cdots, z_n \in \mathbb{R}^d$, define $s(\mathscr{A}, \{z_1, \cdots, z_n\}) = |\{A \cap \{z_1, \cdots, z_n\} : A \in \mathscr{A}\}|$.

Definition 3.11 Let $\mathscr{G}$ be a subset of $\mathbb{R}^d$ of size n. We say $\mathscr{A}$ shatters $\mathscr{G}$ if $s(\mathscr{A}, \mathscr{G}) = 2^n$. The nth shatter coefficient of $\mathscr{A}$ is $S(\mathscr{A}, n) = \max_{\{z_1, \cdots, z_n\} \subset \mathbb{R}^d} s(\mathscr{A}, \{z_1, \cdots, z_n\})$, the maximum number of different subsets of n points that can be picked out by set from $\mathscr{A}$.

Definition 3.12 (VC dimension) Let $\mathscr{A}$ be a class of subsets of $\mathbb{R}^d$ with $\mathscr{A} \ne \emptyset$. The VC dimension $V_{\mathscr{A}}$ of $\mathscr{A}$ is defined by $V_{\mathscr{A}} = \sup\{n \in \mathbb{N}, S(\mathscr{A}, n) = 2^n\}$.

Proposition 3.1 $S(\mathscr{A}, n) \le \sum\limits_{i=0}^{V_{\mathscr{A}}} \binom{n}{i}$.

Theorem 3.5 Let $\mathscr{G}$ be a set of functions $g : \mathbb{R}^d \to [0, B]$. For any $n \in \mathbb{N}$ and $\epsilon > 0$,

$$\mathbb{P}\left\{\sup_{g\in\mathscr{G}} \left|\frac{1}{n}\sum_{i=1}^{n} g(Z) - \mathbb{E}[g(Z)]\right| > \epsilon\right\} \le 8\mathbb{E}\mathscr{N}_1(\frac{\epsilon}{8}, \mathscr{G}, Z^{1:n}) e^{-\frac{n\epsilon^2}{128B^2}}.$$

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 $g \in \mathscr{G}$ $\left|\frac{1}{n}\sum\limits_{i=1}^{n} g(Z_i) - \mathbb{E}g(Z)\right| > \epsilon$ if there exists such a function. Otherwise, let g^* be an arbitrary function in $\mathscr{G}$. g^* depends on $Z^{1:n}$. $\mathbb{P}\left\{\left|\mathbb{E}[g^*(Z)|Z^{1:n}] - \dfrac{1}{n}\sum\limits_{i=1}^{n} g^*(Z_i')\right| > \dfrac{\epsilon}{2}|Z^{1:n}\right\} \le \dfrac{\mathrm{Var}(g^*(Z)|Z^{1:n})}{n(\frac{\epsilon}{2})^2} \le \dfrac{B^2/4}{n\epsilon^2/4} = \dfrac{B^2}{n\epsilon^2} \le \dfrac{1}{2}$ for $n \ge \dfrac{2B^2}{\epsilon^2}$. Thus we have

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

Therefore, $2\mathbb{P}\left\{\sup\limits_{g\in\mathscr{G}} \left|\dfrac{1}{n}\sum\limits_{i=1}^{n} g(Z_i) - \dfrac{1}{n}\sum\limits_{i=1}^{n} g(Z_i')\right| > \dfrac{\epsilon}{2}\right\} \ge \mathbb{P}\left\{\sup\limits_{g\in\mathscr{G}} \left|\dfrac{1}{n}\sum\limits_{i=1}^{n} g(Z_i) - \mathbb{E}[g(Z)]\right| > \epsilon\right\}$.

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

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

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

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

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

Theorem 3.6 Let $\mathscr{G}$ be a class of functions $g : \mathbb{R}^d \to [0, B]$ with $V_{\mathscr{G}^+} \ge 2$ where $\mathscr{G}^+ := \{(z,t) \in \mathbb{R}^d \times \mathbb{R} : t \le g(z), g \in \mathscr{G}\}$. Let $p \ge 1$, ν be a probability measure on $\mathbb{R}^d$ and $0 < \epsilon < \frac{B}{4}$. Then

$$\mathscr{M}(\epsilon, \mathscr{G}, \|\cdot\|_{L_p(\nu)}) \le 3\left(\frac{2eB^p}{\epsilon^p}\log\frac{3eB^p}{\epsilon^p}\right)^{V_{\mathscr{G}^+}}.$$

Proof Step 1: Set $p = 1$. Relate $\mathscr{M}(\epsilon, \mathscr{G}, \|\cdot\|_{L_p(\nu)})$ to a shatter coefficient of $\mathscr{G}^+$. Set $m = \mathscr{M}(\epsilon, \mathscr{G}, \|\cdot\|_{L_p(\nu)})$ and let $\bar{\mathscr{G}} = \{g_1, \cdots, g_m\}$ be a ϵ-packing of $\mathscr{G}$ w.r.t. $\|\cdot\|_{L_p(\nu)}$. Let $Q_1, \cdots, Q_K \in \mathbb{R}^d$ be K independent r.v.'s with common ν. Generate K independent r.v.'s $T_1, \cdots, T_K$ uniformly distributed on $[0, B]$. Denote $R_i = (Q_i, T_i), i = 1, \cdots, K, \mathscr{G}_f = \{(x,t) : t \le f(x)\}$ for $f : \mathbb{R}^d \to [0, B]$. Then

$$\begin{aligned}
S(\mathscr{G}^+, K) &= \max_{\{z_1,\cdots,z_K\}\in\mathbb{R}^d\times\mathbb{R}} s(\mathscr{G}^+, \{z_1, \cdots, z_K\}) \ge \mathbb{E}s(\mathscr{G}_+, \{R_1, \cdots, R_K\}) \ge \mathbb{E}s(\{\mathscr{G}_f : f \in \mathscr{G}\}, \{R_1, \cdots, R_K\}) \\
&\ge \mathbb{E}s(\{\mathscr{G}_f : f \in \mathscr{G}, \mathscr{G}_f \cap R^{1:K} \ne \mathscr{G}_g \cap R^{1:K} \text{ for all } g \in \bar{\mathscr{G}}, g \ne f\}, R^{1:K}) \\
&= \mathbb{E}\left\{\sum_{f\in\bar{\mathscr{G}}} 1_{\{\mathscr{G}_f\cap R^{1:K}\ne\mathscr{G}_g\cap R^{1:K} \text{ for all } g\in\mathscr{G}, g\ne f\}}\right\} = \sum_{f\in\bar{\mathscr{G}}} \mathbb{P}(\mathscr{G}_f \cap R^{1:K} \ne \mathscr{G}_g \cap R^{1:K} \text{ for all } g \in \mathscr{G}, g \ne f) \\
&= \sum_{f\in\bar{\mathscr{G}}} (1 - \mathbb{P}(\exists g \in \bar{\mathscr{G}}, g \ne f, \mathscr{G}_f \cap R^{1:K} = \mathscr{G}_g \cap R^{1:K})) \ge \sum_{f\in\bar{\mathscr{G}}}\left(1 - m\max_{g\in\bar{\mathscr{G}}, g\ne f} \mathbb{P}(\mathscr{G}_f \cap R^{1:K} = \mathscr{G}_g \cap R^{1:K})\right).
\end{aligned}$$

For $f, g \in \bar{\mathscr{G}}, f \ne g$,

$$\mathbb{P}(\mathscr{G}_f \cap R^{1:K} = \mathscr{G}_g \cap R^{1:K}) = \mathbb{P}(\mathscr{G}_f \cap \{R_1\} = \mathscr{G}_g \cap \{R_1\})^K,$$

and

$$\begin{aligned}
\mathbb{P}(\mathscr{G}_f \cap \{R_1\} = \mathscr{G}_g \cap \{R_1\}) &= 1 - \mathbb{P}(\mathscr{G}_f \cap \{R_1\} \ne \mathscr{G}_g \cap \{R_1\}) = 1 - \mathbb{E}[\mathbb{P}(\mathscr{G}_f \cap \{R_1\} \ne \mathscr{G}_g \cap \{R_1\}|Q_1)] \\
&= 1 - \mathbb{E}[\mathbb{P}(f(Q_1) < T \le g(Q_1) \text{ or } g(Q_1) < T \le f(Q_1)|Q_1)] = 1 - \mathbb{E}[\frac{|f(Q_1) - g(Q_1)|}{B}] \\
&= 1 - \frac{1}{B}\int |f(x) - g(x)|\nu(\mathrm{d}x) \le 1 - \frac{\epsilon}{B} \Rightarrow \mathbb{P}(\mathscr{G}_f \cap \{R_1\} = \mathscr{G}_g \cap \{R_1\})^K \le (1 - \frac{\epsilon}{B})^K \le \exp(-\frac{\epsilon K}{B})
\end{aligned}$$

$$\Rightarrow S(\mathscr{G}^+, K) \ge m(1 - m\exp(-\frac{\epsilon K}{B})).$$

Set $K = \lfloor \frac{B}{\epsilon}\log(2m) \rfloor$. Then

$$1 - m\exp(-\frac{\epsilon K}{B}) \ge 1 - m\exp(-\frac{\epsilon}{B}(\frac{B}{\epsilon}\log(2m) - 1)) = 1 - \frac{1}{2}\exp(\frac{\epsilon}{B}) \ge 1 - \frac{1}{2}\exp(\frac{1}{4}) \ge \frac{1}{3} \Rightarrow m \le 3S(\mathscr{G}_+, K).$$

Step 2: Relate $S(\mathscr{G}_+, K)$ to $V_{\mathscr{G}_+}$. Set $K = \lfloor \frac{B}{\epsilon}\log(2\mathscr{M}(\epsilon, \mathscr{G}, \|\cdot\|_{L_p(\nu)})) \rfloor \le V_{\mathscr{G}_+} \Rightarrow \mathscr{M}(\epsilon, \mathscr{G}, \|) \cdot \|_{L_p(\nu)} \le \frac{\epsilon}{2}\exp(V_{\mathscr{G}_+}) \le 3\left(\frac{2eB}{\epsilon}\log\frac{3eB}{\epsilon}\right)^{V_{\mathscr{G}_+}}$. In the case $K > V_{\mathscr{G}_+}$, use the lemma:

Lemma 3.3 Let $\mathscr{A} \in \mathbb{R}^d$ and $V_{\mathscr{A}} < \infty$. Then $\forall n \in \mathbb{N}, S(\mathscr{A}, n) \le (n+1)^{V_{\mathscr{A}}}$ and $\forall n \ge V_{\mathscr{A}}, S(\mathscr{A}, n) \le (\frac{en}{V_{\mathscr{A}}})^{V_{\mathscr{A}}}$.

Then $\mathscr{M}(\epsilon, \mathscr{G}, \|\cdot\|_{L_p(\nu)}) \le 3\left(\frac{eK}{V_{\mathscr{G}_+}}\right)^{V_{\mathscr{G}_+}} \le 3\left(\frac{eB}{\epsilon V_{\mathscr{G}_+}}\log(2\mathscr{M}(\epsilon, \mathscr{G}, \|\cdot\|_{L_p(\nu)}))\right)^{V_{\mathscr{G}_+}}$.

Step 3: Setting $a = \frac{eB}{\epsilon}$ and $b = V_{\mathscr{G}_+}, \mathscr{M}(\epsilon, \mathscr{G}, \|\cdot\|_{L_p(\nu)}) := x \le 3(\frac{a}{b}\log(2x))^b \Rightarrow x \le 3(2a\log(3a))^b$.

Step 4: Let $1 < p < \infty$. Then for any $g_j, g_k \in \mathscr{G}, \|g_j - g_k\|_{L_p(\nu)}^p \le B^{p-1}\|g_j - g_k\|_{L_1(\nu)} \Rightarrow \mathscr{M}(\epsilon, \mathscr{G}, \|\cdot\|_{L_p(\nu)}) \le \mathscr{M}(\frac{\epsilon^p}{B^{p-1}}, \mathscr{G}, \|\cdot\|_{L_p(\nu)})$. □

Theorem 3.7 (A uniform law of large numbers) Let $\mathscr{G}$ be a class of functions $g:\mathbb{R}^d\to\mathbb{R}$ and $G:\mathbb{R}^d\to\mathbb{R}, G(x)=\sup_{g\in\mathscr{G}}|g(x)|$ be an envelope of $\mathscr{G}$. Assume $\mathbb{E}G(Z)<\infty$ and $V_{\mathscr{G}^+}<\infty$. Then

$$\sup_{g\in\mathscr{G}}\left|\frac{1}{n}\sum_{i=1}^n g(Z_i)-\mathbb{E}g(Z)\right|\to 0 \text{ a.s. as } n\to+\infty$$

Proof For $L>0$, set $\mathscr{G}_L:=\{g\cdot 1_{\{G\le L\}}:g\in\mathscr{G}\}$. For $g\in\mathscr{G}$,

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^n g(Z_i)-\mathbb{E}g(Z)\right| &\le \left|\frac{1}{n}\sum_{i=1}^n g(Z_i)-\frac{1}{n}\sum_{i=1}^n g(Z_i)1_{\{G(Z_i)\le L\}}\right|\\ &\quad+\left|\frac{1}{n}\sum_{i=1}^n g(Z_i)1_{\{G(Z_i)\le L\}}-\mathbb{E}\{g(Z)1_{G(Z)\le L}\}\right|+\left|\mathbb{E}\{g(Z)1_{G(Z)\le L}\}-\mathbb{E}\{g(Z)\}\right|\\ &=\left|\frac{1}{n}\sum_{i=1}^n g(Z_i)1_{\{G(Z_i)>L\}}\right|+\mathbb{E}|g(Z)|1_{\{G(Z)>L\}}+\left|\frac{1}{n}\sum_{i=1}^n g(Z_i)1_{\{G(Z_i)\le L\}}-\mathbb{E}\{g(Z)1_{G(Z)\le L}\}\right|\end{aligned}$$

Since $\mathbb{P}(\sup_{g\in\mathscr{G}_L}\left|\frac{1}{n}\sum_{i=1}^n g(Z_i)-\mathbb{E}g(Z)\right|>\epsilon)\le 8\mathbb{E}\{\mathscr{M}_1(\frac{\epsilon}{8},\mathscr{G}_L,Z^{1:n})\exp\left(-\frac{n\epsilon^2}{128(2L)^2}\right)e\}$, use B-C lemma. □

Definition 3.13 (Least square estimates) $\mathbb{E}\{(m(X)-Y)^2\}=\inf_f\mathbb{E}\{(f(X)-Y)^2\}\Rightarrow m(X)=\mathbb{E}[Y|X]$. Define $m_n=\arg\min_{f\in\mathscr{F}_n}\frac{1}{n}\sum\limits_{i=1}^n|f(X_i)-Y_i|^2, m^*=\arg\min_{f\in\mathscr{F}_n}\mathbb{E}\{(f(X)-Y)^2\}$.

Theorem 3.8 Let $\mathscr{F}_n$ be a class of functions $f:\mathbb{R}^d\to\mathbb{R}$ depending on th data $\mathcal{D}_n=\{(X_1,Y_1),\cdots,(X_n,Y_n)\}$. Then

$$\int|m_n(x)-m(x)|^2\nu(\mathrm{d}x)\le 2\sup_{f\in\mathscr{F}_n}\left|\frac{1}{n}\sum_{i=1}^n|f(X_i)-Y_i|^2-\mathbb{E}\{(f(X)-Y)^2\}\right|+\inf_{f\in\mathscr{F}_n}\int|f(x)-m(x)|^2\nu(\mathrm{d}x).$$

Proof We do the following decomposition:

$$\begin{aligned}\int|m_n(x)-m(x)|^2\nu(\mathrm{d}x)&=\mathbb{E}[|m_n(X)-Y|^2|\mathcal{D}_n]-\mathbb{E}[|m(X)-Y|^2]\\ &=\{\mathbb{E}[|m_n(X)-Y|^2|\mathcal{D}_n]-\inf_{f\in\mathscr{F}_n}\mathbb{E}|f(X)-Y|^2\}+\{\inf_{f\in\mathscr{F}_n}\mathbb{E}|f(X)-Y|^2-\mathbb{E}|m(X)-Y|^2\}\\ &:=I_1+I_2.\end{aligned}$$

$I_1\le 2\sup\limits_{f\in\mathscr{F}_n}\left|\frac{1}{n}\sum\limits_{i=1}^n|f(X_i)-Y_i|^2-\mathbb{E}|f(X)-Y|^2\right|$. $I_2=\inf\limits_{f\in\mathscr{F}_n}\int(f(x)-m(x))^2\nu(\mathrm{d}x)$. □

Proposition 3.2 (Method of Sieves) Let $\psi_1,\psi_2,\cdots,\mathbb{R}^d\to\mathbb{R}$ be bounded functions such that $|\psi_j(x)|\le 1$. Assume that the set of functions $\cup_{k=1}^{+\infty}\{\sum\limits_{j=1}^k a_j\psi_j(x):a_1,\cdots,a_k\in\mathbb{R}\}$ is dense in $L_2(\mu)$ for any probability measure μ on $\mathbb{R}^d$. Define the regression function estimate m_n as a function minimizing the empirical L_2 risk $\frac{1}{n}\sum\limits_{i=1}^n(f(X_i)-Y_i)^2$ over function $f(x)=\sum\limits_{j=1}^{k_n}a_j\psi_j(x)$ with $\sum\limits_{j=1}^{k_n}|a_j|\le\beta_n$. If $\mathbb{E}(Y^2)<\infty$ and k_n and β_n satisfy $k_n\to\infty,\beta_n\to\infty,\frac{k_n\beta_n^4\log\beta_n}{n}\to 0$ and $\frac{\beta_n^4}{n^{1-\delta}}\to 0$ for some $\delta>0$, then $\int(m_n(x)-m(x))^2\mu(\mathrm{d}x)\to 0$ with probability 1.

Theorem 3.9 Consider $\mathscr{F}_n=\{\sum\limits_{j=1}^{k_n}a_j\psi_j(x):\sum\limits_{j=1}^{k_n}|a_j|\le\beta_n\}$ and $\widetilde{\mathscr{F}}_n=\{\sum\limits_{j=1}^{k_n}a_j\psi_j(x):a_j\in\mathbb{R}\}$. Step 1: derive $\widetilde{m}_n$ by using $\widetilde{\mathscr{F}}_n$. Step 2: Trancation of $\widetilde{m}_n$, $m_n(x)=T_{\beta_n}\widetilde{m}_n(x)$ where $T_Lu=\begin{cases}u, & \text{if } |u|\le L\\ L\mathrm{sgn}(u), & \text{otherwise}\end{cases}$. (a) If $\mathbb{E}(Y^2)<\infty$ and k_n and β_n satisfy $k_n\to\infty,\beta_n\to\infty,\frac{k_n\beta_n^4\log\beta_n}{n}\to 0$, then $\mathbb{E}\int(m_n(x)-m(x))^2\mu(\mathrm{d}x)\to 0$. (b) If adding the extra condition $\frac{\beta_n^4}{n^{1-\delta}}\to 0$ for some $\delta>0$, then $\int(m_n(x)-m(x))^2\mu(\mathrm{d}x)\to 0$ a.s.

Proposition 3.3 Let $\widetilde{F}_n=\widetilde{F}_n(\mathcal{D}_n)$ be a class of functions $f:\mathbb{R}^d\to\mathbb{R}$. If $|Y|\le\beta_n$ a.s., then

$$\int(m_n(x)-m(x))^2\mu(\mathrm{d}x)\le 2\sup_{f\in T_{\beta_n}\widetilde{F}_n}\left|\frac{1}{n}\sum_{i=1}^n|f(X_i)-Y_i|^2-\mathbb{E}|f(X)-Y|^2\right|+\inf_{f\in\widetilde{F}_n,\|f\|_\infty\le\beta_n}\int|f(x)-m(x)|^2\mu(\mathrm{d}x)$$

Theorem 3.10 Let $\widetilde{\mathscr{F}}_n = \widetilde{\mathscr{F}}_n(\mathcal{D}_n)$ be a class of functions $f: \mathbb{R}^d \to \mathbb{R}$ and $Y_L = T_L Y, Y_{i,L} = T_L Y_i$. (a) If

$$\lim_{n\to+\infty} \beta_n = \infty, \lim_{n\to+\infty} \inf_{f\in\widetilde{F}_n, \|f\|_\infty \le \beta_n} \int |f(x) - m(x)|^2 \mu(\mathrm{d}x) = 0 \text{ a.s.}$$

$$\lim_{n\to+\infty} \sup_{f\in T_{\beta_n}\widetilde{F}_n} \left| \frac{1}{n}\sum_{i=1}^n |f(X_i) - Y_{i,L}|^2 - \mathbb{E}(f(X) - Y_L)^2 \right| = 0 \text{ a.s. for all } L > 0$$

, then $\lim_{n\to+\infty} \int |m_n(x) - m(x)|^2 \mu(\mathrm{d}x) = 0$ a.s. (b) If $\beta_n \to +\infty, \mathbb{E}\{\cdot\} \to 0, \mathbb{E}\{\cdot\} \to 0$, then $\mathbb{E}\{\cdot\} \to 0$.

Definition 3.14 (Piecewise polynomial partition estimate) $\mathcal{P}_n = \{A_{n,1}, A_{n,2}, \cdots\}$ be a partition of $\mathbb{R}^d$,

$$\hat{m}_n(x) := \frac{\sum\limits_{i=1}^n Y_i I_{\{X_i \in A_n(x)\}}}{\sum\limits_{i=1}^n I_{\{X_i \in A_n(x)\}}}$$

where $A_n(x)$ denotes the cell $A_{n,j} \in \mathcal{P}_n$ which contains x.

Theorem 3.11 Let $\mathscr{F}$ be a class of function $f: \mathbb{R}^d \to \mathbb{R}$ bounded in abolute value by B. Let $\epsilon > 0$. Then

$$\mathbb{P}\{\exists f \in \mathscr{F} \text{ s.t.} \|f\|_2 - 2\|f\|_n > \epsilon\} \le \mathbb{E}\mathscr{N}_2\left(\frac{\sqrt{2}}{24}\epsilon, \mathscr{F}, X^{1:2n}\right) \exp\left(-\frac{n\epsilon^2}{288B^2}\right)$$

where $\|f\|_n^2 = \frac{1}{n}\sum\limits_{i=1}^n |f(X_i)|^2$.

Proof Step 1: Replace $L_2(\mu)$ norm by the empirical norm. Let $\widetilde{X}^{1:n} = (X_{n+1}, \cdots, X_{2n})$ be a ghost sample of i.i.d. r.v.'s as X and independent of $X^{1:n}$. Define $\|f\|_{n'}^2 = \frac{1}{n}\sum\limits_{i=n+1}^{2n} |f(X_i)|^2$. Let f^* be a function $f \in \mathscr{F}$ such that $\|f\|_2 - 2\|f\|_n > \epsilon$ if there exists any such function, and let f^* be an arbitrary function in $\mathscr{F}$ if such a function does not exist. Then

$$\begin{aligned}
\mathbb{P}\{2\|f^*\|_{n'} + \frac{\epsilon}{2} > \|f^*\|_2 | X^{1:n}\} &\ge \mathbb{P}\{4\|f^*\|_{n'}^2 + \frac{\epsilon^2}{4} > \|f^*\|_2^2 | X^{1:n}\} = 1 - \mathbb{P}\{4\|f^*\|_{n'}^2 + \frac{\epsilon^2}{4} \le \|f^*\|_2^2 | X^{1:n}\} \\
&= 1 - \mathbb{P}\{3\|f^*\|_2^2 + \frac{\epsilon^2}{4} \le 4(\|f^*\|_2^2 - \|f^*\|_{n'}^2) | X^{1:n}\} \ge 1 - \frac{16\mathrm{Var}\left(\frac{1}{n}\sum\limits_{i=n+1}^{2n} |f^*(X_i)|^2 | X^{1:n}\right)}{(3\|f^*\|_2^2 + \frac{\epsilon^2}{4})^2} \\
&\ge 1 - \frac{\frac{16}{n}B^2\|f^*\|_2^2}{(3\|f^*\|_2^2 + \frac{\epsilon^2}{4})^2} \ge 1 - \frac{\frac{16}{3}\frac{B^2}{n}}{3\|f^*\|_2^2 + \frac{\epsilon^2}{4}} \ge 1 - \frac{64}{3\epsilon^2}\frac{B^2}{n} \ge \frac{2}{3} \text{ for } n \ge \frac{64B^2}{\epsilon^2}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{P}\{\exists f \in \mathscr{F} : \|f\|_{n'} - \|f\|_n > \frac{\epsilon}{4}\} &\ge \mathbb{P}\{2\|f^*\|_{n'} - 2\|f^*\|_n > \frac{\epsilon}{2}\} \ge \mathbb{P}\{2\|f^*\|_{n'} + \frac{\epsilon}{2} - 2\|f^*\|_n > \epsilon, 2\|f^*\|_{n'} + \frac{\epsilon}{2} > \|f^*\|_2\} \\
&\ge \mathbb{P}\{\|f^*\|_2 - 2\|f^*\|_n > \epsilon, 2\|f^*\|_{n'} + \frac{\epsilon}{2} > \|f^*\|_2\} \\
&= \mathbb{E}\{1_{\{\|f^*\|_2 - 2\|f^*\|_n > \epsilon\}} \mathbb{P}\{2\|f^*\|_{n'} + \frac{\epsilon}{2} > \|f^*\|_2 | X^{1:n}\}\} \\
&\ge \frac{2}{3}\mathbb{P}\{\|f^*\|_2 - 2\|f^*\|_n > \epsilon\} = \frac{2}{3}\mathbb{P}\{\exists f \in \mathscr{F} : \|f\|_2 - 2\|f\|_n > \epsilon\}.
\end{aligned}$$

This proves $\mathbb{P}\{\exists f \in \mathscr{F} : \|f\|_2 - 2\|f\|_n > \epsilon\} \le \frac{3}{2}\mathbb{P}\{\exists f \in \mathscr{F} : \|f\|_{n'} - \|f\|_n > \frac{\epsilon}{4}\}$.

Step 2: Introduction of additional randomness. Let $U_1, \cdots, U_n$ be independent and uniformly distributed on $\{-1, 1\}$ and independent of $X_1, \cdots, X_{2n}$. Set $Z_i = \begin{cases} X_{i+n} & \text{if } U_i = 1 \\ X_i & \text{if } U_i = -1 \end{cases}$ and $Z_{i+n} = \begin{cases} X_i & \text{if } U_i = 1 \\ X_{i+n} & \text{if } U_i = -1 \end{cases}$. Then

$$\begin{aligned}
\mathbb{P}\left\{\exists f \in \mathscr{F} : \|f\|_{n'} - \|f\|_n > \frac{\epsilon}{4}\right\} &= \mathbb{P}\left\{\exists f \in \mathscr{F} : \left(\frac{1}{n}\sum_{i=n+1}^{2n} |f(X_i)|^2\right)^{\frac{1}{2}} - \left(\frac{1}{n}\sum_{i=1}^n |f(X_i)|^2\right)^{\frac{1}{2}} > \frac{\epsilon}{4}\right\} \\
&= \mathbb{P}\left\{\exists f \in \mathscr{F} : \left(\frac{1}{n}\sum_{i=n+1}^{2n} |f(Z_i)|^2\right)^{\frac{1}{2}} - \left(\frac{1}{n}\sum_{i=1}^n |f(Z_i)|^2\right)^{\frac{1}{2}} > \frac{\epsilon}{4}\right\}
\end{aligned}$$

Step 3: Conditioning and introduction of a covery. Let $\mathscr{G}=\{g_j : j=1,\cdots,\mathscr{N}_2(\frac{\sqrt{2}}{24}\epsilon,\mathscr{F},X^{1:2n})\}$ be a $\frac{\sqrt{2}}{24}\epsilon$-cover of $\mathscr{F}$ w.r.t. $\|\cdot\|_{2n}$ of minimal size. $\|f\|_{2n}^2=\frac{1}{2n}\sum_{i=1}^{2n}|f(X_i)|^2$. Fix $f\in\mathscr{F}$, $\|f-g\|_{2n}\le\frac{\sqrt{2}}{24}\epsilon$. Then

$$
\begin{aligned}
&\left\{\frac{1}{n}\sum_{i=n+1}^{2n}|f(Z_i)|^2\right\}^{\frac12}-\left\{\frac{1}{n}\sum_{i=1}^{n}|f(Z_i)|^2\right\}^{\frac12}\\
&=\left\{\frac{1}{n}\sum_{i=n+1}^{2n}|f(Z_i)|^2\right\}^{\frac12}-\left\{\frac{1}{n}\sum_{i=n+1}^{2n}|g(Z_i)|^2\right\}^{\frac12}+\left\{\frac{1}{n}\sum_{i=n+1}^{2n}|g(Z_i)|^2\right\}^{\frac12}-\left\{\frac{1}{n}\sum_{i=1}^{n}|g(Z_i)|^2\right\}^{\frac12}+\left\{\frac{1}{n}\sum_{i=1}^{n}|g(Z_i)|^2\right\}^{\frac12}-\left\{\frac{1}{n}\sum_{i=1}^{n}|f(Z_i)|^2\right\}^{\frac12}\\
&\le\left\{\frac{1}{n}\sum_{i=n+1}^{2n}|f(Z_i)-g(Z_i)|^2\right\}^{\frac12}+\left\{\frac{1}{n}\sum_{i=n+1}^{2n}|g(Z_i)|^2\right\}^{\frac12}-\left\{\frac{1}{n}\sum_{i=1}^{n}|g(Z_i)|^2\right\}^{\frac12}+\left\{\frac{1}{n}\sum_{i=1}^{n}|g(Z_i)-f(Z_i)|^2\right\}^{\frac12}\\
&\le 2\sqrt{2}\|f-g\|_{2n}+\left\{\frac{1}{n}\sum_{i=n+1}^{2n}|g(Z_i)|^2\right\}^{\frac12}-\left\{\frac{1}{n}\sum_{i=1}^{n}|g(Z_i)|^2\right\}^{\frac12}\\
&\le\frac{\epsilon}{6}+\left\{\frac{1}{n}\sum_{i=n+1}^{2n}|g(Z_i)|^2\right\}^{\frac12}-\left\{\frac{1}{n}\sum_{i=1}^{n}|g(Z_i)|^2\right\}^{\frac12}
\end{aligned}
$$

In this way,

$$
\begin{aligned}
&\mathbb{P}\left\{\exists f\in\mathscr{F}:\left(\frac{1}{n}\sum_{i=n+1}^{2n}|f(Z_i)|^2\right)^{\frac12}-\left(\frac{1}{n}\sum_{i=1}^{n}|f(Z_i)|^2\right)^{\frac12}>\frac{\epsilon}{4}|X^{1:2n}\right\}\\
&\le\mathbb{P}\left\{\exists g\in\mathscr{G}:\left(\frac{1}{n}\sum_{i=n+1}^{2n}|g(Z_i)|^2\right)^{\frac12}-\left(\frac{1}{n}\sum_{i=1}^{n}|g(Z_i)|^2\right)^{\frac12}>\frac{\epsilon}{12}|X^{1:2n}\right\}\\
&\le|\mathscr{G}|\max_{g\in\mathscr{G}}\mathbb{P}\left\{\left(\frac{1}{n}\sum_{i=n+1}^{2n}|g(Z_i)|^2\right)^{\frac12}-\left(\frac{1}{n}\sum_{i=1}^{n}|g(Z_i)|^2\right)^{\frac12}>\frac{\epsilon}{12}|X^{1:2n}\right\}
\end{aligned}
$$

Step 4: Application of Hoeffding's inequality.

$$
\begin{aligned}
\left(\frac{1}{n}\sum_{i=n+1}^{2n}|g(Z_i)|^2\right)^{\frac12}-\left(\frac{1}{n}\sum_{i=1}^{n}|g(Z_i)|^2\right)^{\frac12}&\le\left|\frac{\frac1n\sum_{i=n+1}^{2n}|g(Z_i)|^2-\frac1n\sum_{i=1}^{n}|g(Z_i)|^2}{\left(\frac1n\sum_{i=n+1}^{2n}|g(Z_i)|^2\right)^{\frac12}+\left(\frac1n\sum_{i=1}^{n}|g(Z_i)|^2\right)^{\frac12}}\right|\\
&\le\frac{\left|\frac1n\sum_{i=n+1}^{2n}|g(Z_i)|^2-\frac1n\sum_{i=1}^{n}|g(Z_i)|^2\right|}{\left(\frac1n\sum_{i=1}^{2n}|g(Z_i)|^2\right)^{\frac12}}=\frac{\left|\frac1n\sum_{i=1}^{n}U_i|g(X_i)|^2-\frac1n\sum_{i=1}^{n}U_i|g(X_{i+n})|^2\right|}{\left(\frac1n\sum_{i=1}^{2n}|g(Z_i)|^2\right)^{\frac12}}
\end{aligned}
$$

Then

$$
\begin{aligned}
\mathbb{P}\{(\frac{1}{n}\sum_{i=n+1}^{2n}|g(Z_i)|^2)^{\frac12}-(\frac{1}{n}\sum_{i=1}^{n}|g(Z_i)|^2)^{\frac12}>\frac{\epsilon}{12}|X^{1:2n}\}&\le 2\exp\left(-\frac{2n^2\frac{\epsilon^2}{144}(\frac1n\sum_{i=1}^{2n}|g(X_i)|^2)}{\sum_{i=1}^{n}4(|g(X_i)|^2-|g(X_{i+n})|^2)^2}\right)\\
&\le 2\exp\left(-\frac{2n^2\frac{\epsilon^2}{144}(\frac1n\sum_{i=1}^{2n}|g(X_i)|^2)}{\sum_{i=1}^{n}4B^2(|g(X_i)|^2+|g(X_{i+n})|^2)}\right)\\
&=\exp\left(-\frac{n\epsilon^2}{288B^2}\right)
\end{aligned}
$$

□

Theorem 3.12 Assume $\sigma^2=\sup_{x\in\mathbb{R}^d}\operatorname{Var}(Y|X=x)<\infty$. Let $k_n=k_n(x_1,\cdots,x_n)$ be the vector space dimension of $\mathscr{F}_n$. Then

$$
\mathbb{E}\{\|\widetilde{m}_n-m\|_n^2|X^{1:n}\}\le\frac{\sigma^2k_n}{n}+\min_{f\in\mathscr{F}_n}\|f-m\|_n^2.
$$

Proof Denote $\mathbb{E}^*\{\cdot\} = \mathbb{E}\{\cdot|X^{1:n}\}$. Then

$$\begin{aligned}
EE^*\{\|\widetilde{m}_n - m\|_n^2\} &= \mathbb{E}^*\{\frac{1}{n}\sum_{i=1}^n |\widetilde{m}_n(X_i) - m(X_i)|^2\} \\
&= \mathbb{E}^*\{\frac{1}{n}\sum_{i=1}^n |\widetilde{m}_n(X_i) - \mathbb{E}^*(\widetilde{m}_n(X_i)) + \mathbb{E}^*(\widetilde{m}_n(X_i)) - m(X_i)|^2\} \\
&= \mathbb{E}^*\{\frac{1}{n}\sum_{i=1}^n |\widetilde{m}_n(X_i) - \mathbb{E}^*(\widetilde{m}_n(X_i))|^2\} + \mathbb{E}^*\{|\mathbb{E}^*(\widetilde{m}_n(X_i)) - m(X_i)|^2\} \\
&= \mathbb{E}^*\{\|\widetilde{m}_n - \mathbb{E}^*(\widetilde{m}_n)\|_n^2\} + \|\mathbb{E}^*(\widetilde{m}_n) - m\|_n^2.
\end{aligned}$$

Write that $\widetilde{m}_n = \sum_{j=1}^{k_n} a_j f_{j,n}$ where $f_{1,n}, \cdots, f_{k_n,n}$ is a basis of $\mathscr{F}_n$, and $a = (a_j)_{j=1,\cdots,k_n}$ satisfies that $\frac{1}{n}B^T Ba = \frac{1}{n}B^T Y$, $B = (f_{j,n}(X_i))_{1\le i\le n, 1\le j\le k_n}$ and $Y = (Y_1, \cdots, Y_n)^T$. Then

$$\begin{aligned}
&\mathbb{E}^*\{\widetilde{m}_n\} = \sum_{j=1}^{k_n} \mathbb{E}^*\{a_j\} f_{j,n} \text{ and } \frac{1}{n}B^T B\mathbb{E}^* a = \frac{1}{n}B^T \mathbb{E}^* Y = \frac{1}{n}B^T (m(X_1), \cdots, m(X_n))^T \\
&\Rightarrow \|\mathbb{E}^*(\widetilde{m}_n) - m\|_n^2 = \min_{f\in\mathscr{F}_n} \|f - m\|_n^2.
\end{aligned}$$

Choose a complete orthogonormal system $f_1, \cdots, f_k$ in $\mathscr{F}_n$ w.r.t. the empirical scalar proudct $\langle\cdot,\cdot\rangle_n$ where $\langle f, g\rangle_n = \frac{1}{n}\sum_{i=1}^n f(X_i)g(X_i), k \le k_n$. We remind our readers that such a system depends on $X_1, \cdots, X_n$. Then, on $\{X_1, \cdots, X_n\}$, $\text{span}\{f_1, \cdots, f_k\} \subset \mathscr{F}_n$, $\widetilde{m}_n(x) = f(x)^T \frac{1}{n}B^T Y$ where $B = (f_j(X_i))_{1\le j\le n, 1\le j\le k}, B^T B = I$. Therefore,

$$\begin{aligned}
\mathbb{E}^*\{|\widetilde{m}_n(x) - \mathbb{E}^*(\widetilde{m}_n(x))|^2\} &= \mathbb{E}^*\{|f(x)^T \frac{1}{n}B^T Y - f(x)^T\frac{1}{n}B^T(m(X_1), \cdots, m(X_n))^T|^2\} \\
&= f(x)^T \frac{1}{n}B^T(\mathbb{E}^*\{(Y_i - m(X_i))(Y_j - m(X_j))^T\})\frac{1}{n}Bf(x) \\
\Rightarrow \mathbb{E}^*\{\|\widetilde{m}_n - \mathbb{E}^*(\widetilde{m}_n)\|_n^2\} &\le \frac{1}{n^2} f^T B^T \sigma^2 I B f = \frac{\sigma^2}{n}\sum_{j=1}^k \|f_j\|_n^2 = \frac{\sigma^2}{n}k \le \frac{\sigma^2}{n}k_n.
\end{aligned}$$

□

Theorem 3.13 Assume $\sigma^2 = \sup_{x\in\mathbb{R}^d} \text{Var}(Y|X = x) < \infty$ and $\|m\|_\infty = \sup_{x\in\mathbb{R}^d} |m(x)| \le L \in \mathbb{R}_+, m_n(\cdot) = T_L\widetilde{m}_n(\cdot)$. Then

$$\mathbb{E}\int |m_n(x) - m(x)|^2\mu(\mathrm{d}x) \le C\cdot\max\{\sigma^2, L^2\}\frac{\log(n) + 1}{n}k_n + 8\inf_{f\in\mathscr{F}_n}\int |f(x) - m(x)|^2\mu(\mathrm{d}x).$$

Proof First we note that

$$\begin{aligned}
\int |m_n(x) - m(x)|^2\mu(\mathrm{d}x) &= (\|m_n - m\|_2 - 2\|m_n - m\|_n + 2\|m_n - m\|_n)^2 \\
&\le (\max\{\|m_n - m\|_2 - 2\|m_n - m\|_m, 0\} + 2\|m_n - m\|_n)^2 \\
&\le 2(\max\{\|m_n - m\|_2 - 2\|m_n - m\|_n, 0\})^2 + 8\|m_n - m\|_n^2.
\end{aligned}$$

On the one hand,

$$\begin{aligned}
\mathbb{E}\{8\|m_n - m\|_n^2\} &\le 8\mathbb{E}\{\mathbb{E}\{\|\widetilde{m}_n - m\|_n^2|X_1, \cdots, X_n\}\} \\
&\le 8\sigma^2\frac{k_n}{n} + 8\mathbb{E}\{\min_{f\in\mathscr{F}_n}\|f - m\|_n^2\} \\
&\le 8\sigma^2\frac{k_n}{n} + 8\inf_{f\in\mathscr{F}_n}\mathbb{E}\|f - m\|_n^2.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\mathbb{P}\left(2\max\{\|m_n - m\|_2 - 2\|m_n - m\|_n, 0\} > u\right) &\le \mathbb{P}\left(\exists f \in T_L\mathscr{F}_n : \|f - m\|_2 - 2\|f - m\|_n > \sqrt{\frac{u}{2}}\right) \\
&\le 3\mathbb{E}\mathscr{N}_2\left(\frac{\sqrt{u}}{24}, \mathscr{F}_n, X^{1:2n}\right)\exp\left(-\frac{nu}{576(2L)^2}\right)
\end{aligned}$$

$$\leq 9(12en)^{2(k_n+1)}\exp\left(-\frac{nu}{2304L^2}\right)$$

$$\Rightarrow \mathbb{E}(2\max\{\|m_n-m\|_2-2\|m_n-m\|_n,0\}) \leq u+\int_u^\infty \mathbb{P}(2\max\{\|m_n-m\|_2-2\|m_n-m\|_n,0\}>t)\mathrm{d}t$$

$$\left(\text{take } u \geq \frac{576L^2}{n}\right) \leq CL^2\frac{\log(n)+1}{n}k_n.$$

Combine these two bounds together. □

Property 3.3 (Nonlinear LSE) $|Y| \leq L \leq \beta_n$ a.s., $m_n(\cdot) = T_{\beta_n}\widetilde{m}_n(\cdot), \widetilde{m}_n(\cdot) = \arg\min_{f\in\mathscr{F}_n}\frac{1}{n}\sum\limits_{i=1}^n|f(X_i)-Y_i|^2$. We do the following decomposition:

$$\int|m_n(x)-m(x)|^2\mu(\mathrm{d}x) = \left\{\mathbb{E}\{|m_n(X)-Y|^2|\mathcal{D}_n\}-\mathbb{E}|m(X)-Y|^2-\frac{2}{n}\sum_{i=1}^n\left[|m_n(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\right]\right\}$$
$$+\frac{2}{n}\sum_{i=1}^n[|m_n(X_i)-Y_i|^2-|m(X_i)-Y_i|^2].$$

On the one hand,

$$\begin{aligned}
\mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^n[|m_n(X_i)-Y_i|^2-|m(X_i)-Y_i|^2]\right\} &\leq \mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^n|\widetilde{m}_n(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\right\}\\
&\leq \mathbb{E}\left\{\inf_{f\in\mathscr{F}_n}\frac{1}{n}\sum_{i=1}^n\left[|f(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\right]\right\}\\
&\leq \inf_{f\in\mathscr{F}_n}\mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^n\left[|f(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\right]\right\}\\
&= \inf_{f\in\mathscr{F}_n}\{\mathbb{E}|f(X)-Y|^2-\mathbb{E}|m(X)-Y|^2\}\\
&= \inf_{f\in\mathscr{F}n}\int|f(x)-m(x)|^2\mu(\mathrm{d}x)
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\mathbb{P}\left\{\mathbb{E}[|m_n(X)-Y|^2|\mathcal{D}_n]-\mathbb{E}|m(X)-Y|^2-\frac{2}{n}\sum_{i=1}^n\left[|m_n(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\right]>\epsilon\right\}\\
&=\mathbb{P}\left\{\mathbb{E}[|m_n(X)-Y|^2|\mathcal{D}_n]-\mathbb{E}|m(X)-Y|^2-\frac{1}{n}\sum_{i=1}^n\left[|m_n(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\right]>\frac{\epsilon}{2}+\frac{1}{2}\left[\mathbb{E}[|m_n(X)-Y|^2|\mathcal{D}_n]-\mathbb{E}|m(X)-Y|^2\right]\right\}\\
&\leq\mathbb{P}\left\{\exists f\in T_{\beta_n}\mathscr{F}_n:\mathbb{E}|f(X)-Y|^2-\mathbb{E}|m(X)-Y|^2-\frac{1}{n}\sum_{i=1}^n\left[|f(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\right]>\frac{\epsilon}{2}+\frac{1}{2}\left[\mathbb{E}|f(X)-Y|^2-\mathbb{E}|m(X)-Y|^2\right]\right\}
\end{aligned}$$

Set $Z=(X,Y), Z_i=(X_i,Y_i), g(Z)=|f(X)-Y|^2-|m(X)-Y|^2$. We can rewrite the above equation as

$$\mathbb{P}\left\{\mathbb{E}g(Z)-\frac{1}{n}\sum_{i=1}^n g(Z_i)>\frac{\epsilon}{2}+\frac{1}{2}\mathbb{E}g(Z)\right\}.$$

Since $|g(Z)|=|(f(X)+m(X)-2Y)(f(X)-m(X))|\leq 4\beta_n|f(X)-m(X)|, \sigma^2:=\mathrm{Var}(g(Z))\leq\mathbb{E}g(Z)^2\leq 16\beta_n^2\mathbb{E}|f(X)-m(X)|^2=16\beta_n^2(\mathbb{E}|f(X)-Y|^2-\mathbb{E}|m(X)-Y|^2)$, the above equation is upper-bounded by

$$\mathbb{P}\left\{\mathbb{E}g(Z)-\frac{1}{n}\sum_{i=1}^n g(Z_i)>\frac{\epsilon}{2}+\frac{1}{2}\frac{\mathrm{Var}(g(Z))}{16\beta_n^2}\right\}\overset{\text{Berstein's inequality}}{\leq}\exp\left(-\frac{n[\frac{\epsilon}{2}+\frac{\sigma^2}{32\beta_n^2}]^2}{2\sigma^2+2\frac{8\beta_n^2}{3}[\frac{\epsilon}{2}+\frac{\sigma^2}{32\beta_n^2}]}\right)\leq\exp\left(-\frac{1}{128+\frac{32}{3}}\frac{n\epsilon}{\beta_n^2}\right).$$