# Theoretical Machine Learning

Lectured by Zhihua Zhang LaTeXed by Chengxin Gong

June 3, 2024

## Contents

# 1 Introduction

**Outline** 1.1 (Main tasks in machine learning) Generation, prediction, decision. Generation: $X_1, \cdots, X_n \sim F$, infer and analyse $F$, unsupervised learning, e.g. GAN, GPT, $\cdots$. Prediction: data pairs $(X^{(1)}, Y^{(1)}), \cdots, (X^{(n)}, Y^{(n)})$, input variables $X^{(i)} \in \mathbb{R}^d$, $f : \mathcal{X} \to \mathcal{Y}, x \in \mathcal{X}, y \in \mathcal{Y}$, ascribe, supervised learning. Decision: Reinforcement learning, Agent $\leftarrow$ action, state, reward $\rightarrow$ environment.

**Outline** 1.2 (Methods for solving tasks) Parameterized/Non-parameterized, frequency(MLE)/Bayesian.

**Outline** 1.3 (Modeling error) Supervised: Fix $X = (X_1, \cdots, X_d)^T \in \mathbb{R}^d$, for regression $Y \in \mathbb{R}$, for classificataion $Y \in \{0, 1\}$(also $\{-1, 1\}, \{1, \cdots, M\}, \{0, 1\}^M$). Random design for $X$(known as generative models): $Y^{(i)} = g(X^{(i)}, Z^{(i)})$. Fixed design for $X$(known as discriminative models): $Y^{(i)} = g(x^{(i)}, Z^{(i)})$. Unsupervised: $X = g(Z)$(e.g. factor model: $X = AZ + \varepsilon, Z \in \mathcal{N}(0, 1), \varepsilon \sim \mathcal{N}(0, \Sigma)$).

# 2 Statistical Decision Theory

**Definition** 2.1 (Basic concepts) Consider a state space $\Omega$, 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 $\delta$: $\mathcal{R}(\theta, \delta) = \mathbb{E}_\theta \mathcal{L}(\theta, A) = \mathbb{E}_\theta \mathbb{E}_a \mathcal{L}(\theta, A|X) = \iint \mathcal{L}(\theta, a) \mathrm{d}\delta_x(a) \mathrm{d}P_\theta(x)$.

**Example** 2.1 (Parameter estimation) $\theta \in \Omega, \mathscr{A} = \Omega, \mathcal{L}(\theta, a) = \|\theta - a\|_p^p (p \geq 1) \overset{\text{or}}{=} \int \log \frac{P_\theta(x)}{P_a(x)} P_\theta(x) \mathrm{d}m(x)$(KL divergence). $\mathcal{R} = \mathrm{Var}(a) + \mathrm{bias}^2(a)$. 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": $H_1 : \theta \in \Omega_1$. $\delta_x$ 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 $\delta$ is called inadmissible if a competing rule $\delta^*$ 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, $\delta$ 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 $\theta$, if $\delta$ is Bayesian for $\Lambda$ and has finite integrated risk $r(\Lambda, \delta) < \infty$, and if the support of $\Lambda$ is the whole state space $\Omega$, then $\delta$ is admissible.

**Property** 2.1 $p(\theta|x) = \dfrac{p_\theta(x)\lambda(\theta)}{\int p_\theta(x)\lambda(\theta)\mathrm{d}\theta} := \dfrac{p_\theta(x)\lambda(\theta)}{m(x)}$. Define the posterior risk of $\delta$: $r(\delta|X = x) = \int \mathcal{L}(\theta, \delta(x)) \mathrm{d}\mathbb{P}(\theta|x)$. 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 $\delta$ 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 $p = 2$, $r(f|X = x) = \int \mathcal{L}(y, f(x)) \mathrm{d}\mathbb{P}(y|x) = \int |y - f(x)|^2 \mathrm{d}\mathbb{P}(y|x)$. Regression function is $g(x) := \int y \mathrm{d}\mathbb{P}(y|x) \Rightarrow R_f = \mathbb{E}|Y - f(X)|^2 = \mathbb{E}|Y - g(X) + g(X) - f(X)|^2 = \mathbb{E}|Y - g(X)|^2 + \mathbb{E}|g(X) - f(X)|^2 \geq \mathbb{E}|Y - g(X)|^2$.

**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** $\mathbb{E}[\mathcal{L}(Y, f(X))|X = x] = \begin{cases} \mathbb{E}[\mathcal{L}(Y, 0)|X = x] = \mathcal{L}(0, 0)\mathbb{P}(Y = 0|X = x) + \mathcal{L}(1, 0)\mathbb{P}(Y = 1|X = x) \\ \mathbb{E}[\mathcal{L}(Y, 1)|X = x] = \mathcal{L}(0, 1)\mathbb{P}(Y = 0|X = x) + \mathcal{L}(1, 1)\mathbb{P}(Y = 1|X = x) \end{cases}$, compare the sizes of the two. □

**Property** 2.2 (Continuation) $\mathbb{P}(Y=1|X=x)=\mathbb{E}(Y|X=x):=g(x)$, $f(x)=1_{\{g(x)\geq\frac{1}{2}\}}$. Then $0\leq\mathbb{P}(\hat{f}(X)\neq Y)-\mathbb{P}(f(X)\neq Y)\leq 2\int_{\mathcal{X}}|\hat{g}(x)-g(x)|\mu(\mathrm{d}x)\leq 2(\int_{\mathcal{X}}|\hat{g}(x)-g(x)|^2\mu(dx))^{\frac{1}{2}}$. In Example 2.4, $f(x)=1_{\{\frac{p(x|y=1)}{p(x|y=0)}\geq\frac{p_0(\mathcal{L}(0,1)-\mathcal{L}(0,0))}{p_1(\mathcal{L}(1,0)-\mathcal{L}(1,1))}\}}$, which takes the same form as the likelihood ratio test (LRT): Likelihood $L(X):=\frac{p(X|Y=1)}{p(X|Y=0)}$ and $f(x)=1_{\{L(x)\geq\eta\}}$.

**Definition** 2.4 (Confusion table) Ture Positive Rate: $\mathrm{TPR}=\mathbb{P}(\hat{Y}=1|Y=1)$; False Negative Rate: $\mathrm{FNR}=1-\mathrm{TPR}$, type II error; False Positive Rate: $\mathrm{FPR}=\mathbb{P}(\hat{Y}=1|Y=0)$, type I error; True Negative Rate: $\mathrm{TNR}=1-\mathrm{FPR}$. Precision: $\mathbb{P}(Y=1|\hat{Y}=1)=\frac{p_1\mathrm{TPR}}{p_0\mathrm{FPR}+p_1\mathrm{TPR}}$. $F_1$-score: $F_1$ is the harmonic mean of precision and recall, which can be written as $F_1=\frac{2\mathrm{TPR}}{1+\mathrm{TPR}+\frac{p_0}{p_1}\mathrm{FPR}}$.

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

**Theorem** 2.2 (N-P lemma) Optimization: maximize TPR subject to $\mathrm{FPR}\leq\alpha,\alpha\in[0,1]$. 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.

**Proof** Let $\eta$ be the threshold for an LRT such that the predictor $Q_\eta(x)=1\{\alpha(x)\geq\eta\}$ has $\mathrm{FPR}=\alpha$. Such an LRT exists because likelihood functions are continuous. Let $\beta$ denote the TPR of $Q_\eta$. Prove that $Q_\eta$ 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 $\mathcal{R}_Q=p_0\mathrm{FPR}(Q)\mathcal{L}(0,1)+p_1(1-\mathrm{TPR}(Q))\mathcal{L}(1,0)=p_1\eta\mathrm{FPR}(Q)+p_1(1-\mathrm{TPR}(Q))$. Now let $Q$ be any other rule with $\mathrm{FPR}(Q)\leq\alpha$, $\mathcal{R}_{Q_\eta}=p_1\eta\alpha+p_1(1-\beta)\leq p_1\eta\mathrm{FPR}(Q)+p_1(1-\mathrm{TPR}(Q))\leq p_1\eta\alpha+p_1(1-\mathrm{TPR}(Q))\Rightarrow\mathrm{TPR}(Q)\leq\beta$. □

**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** We only prove (2). Fix $\alpha\in(0,1)$ and consider a randomized rate $\mathrm{TPR}=\mathrm{FPR}=\alpha$, $Q(x)\equiv\alpha$; (3): Consider two rules $(\mathrm{FPR}(\eta_1),\mathrm{TPR}(\eta_1))$ and $(\mathrm{FPR}(\eta_2),\mathrm{TPR}(\eta_2))$. Flip a biased coin and use the first rule with probability $t$ and the second rule with probability $1-t$. Then this yields a randomized rule with $(\mathrm{FPR},\mathrm{TPR})=(t\mathrm{FPR}(\eta_1)+(1-t)\mathrm{FPR}(\eta_2),t\mathrm{TPR}(\eta_1)+(1-t)\mathrm{FPR}(\eta_2))$. Fixing $\mathrm{FPR}\leq t\mathrm{FPR}(\eta_1)+(1-t)\mathrm{FPR}(\eta_2)$, $\mathrm{TPR}\geq t\mathrm{TPR}(\eta_1)+(1-t)\mathrm{TPR}(\eta_2)$. □

# 3 Markov Decision Process

**Definition** 3.1 (Basic concepts) 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** 3.2 (Decision rules) Prescribe a procedure for action selection in each state at a specified decision epoch. Four cases: (1) Markovian and Deterministic (MD): $\delta_t:\mathcal{S}\to\mathcal{A}$; (2) M and Randomized (MR): $\delta_t:\mathcal{S}\to\Delta(\mathcal{A})(q_{\delta_t(s)}(a))$; (3) History-dependent and D (HD): $h_t=(s_1,a_1,\cdots,s_{t-1},a_{t-1},s_t)=(h_{t-1},a_{t-1},s_t),\mathcal{H}_1=\mathcal{S},\mathcal{H}_2=\mathcal{S}\times\mathcal{A}\times\mathcal{S},\cdots,\delta_t:\mathcal{H}_t\to\mathcal{A}$; (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** 3.3 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 $U_N^\pi(s):=\mathbb{E}^\pi\{\sum_{t=1}^{N-1}r_t(X_t,Y_t)+r_N(X_N)|X_1=s\}$. Assume $|r_t(s,a)|\leq M<\infty$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$. Optimal policy: $U_N^{\pi^*}(s)\geq U_N^\pi(s),s\in\mathcal{S}$. $\varepsilon$-optimal policy: $U_N^{\pi_\varepsilon^*}(s)+\varepsilon>U_N^\pi(s),s\in\mathcal{S}$. The value of the MDP: $U_N^*(s)=\sup_{\pi\in\mathcal{D}^{\mathrm{HR}}}U_N^\pi(s),s\in\mathcal{S}$.

**Property** 3.1 (Finite-Horizon Policy Evaluation) $V_t^\pi(h_t) = \mathbb{E}^\pi\{\sum_{k=t}^{N-1} r_k(X_k, Y_k) + r_N(X_N)|h_t\}, V_N^\pi(h_N) = r_N(s), \pi \in \mathcal{D}^{\text{HD}}$. By the formula of total expectation,

$$V_t^\pi(h_t) = r_t(s_t, \delta_t(h_t)) + \mathbb{E}_{h_t}^\pi V_{t+1}^\pi(h_t, \delta_t(h_t), X_{t+1}) = r_t(s_t, \delta_t(h_t)) + \sum_{j\in\mathcal{S}} V_{t+1}^\pi(h_t, \delta_t(h_t), j)p(j|s_t, \delta_t(h_t)).$$

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 $\pi$, this yields a valid and accurate calculation method for $U_N^\pi(s)$.

**Theorem** 3.1 (The Bellman Equations) Let $V_t^*(h_t) = \sup_{\pi\in\mathcal{D}^{\text{HR}}} V_t^\pi(h_t)$. The optimality equations:

$$V_t(h_t) = \sup_{a\in\mathcal{A}}\{r_t(s_t, a) + \sum_{j\in\mathcal{S}} V_{t+1}(h_t, a, j)p_t(j|s_t, a)\} \text{ for } t = 1, 2, \cdots, N-1 \text{ and } h_t = (h_{t-1}, a_{t-1}, s_t) \in \mathcal{H}_t.$$

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** We divide the proof into two parts.

Step 1: Prove $V_n(h_n) \geq V_n^*(h_n)$ for all $h_n \in \mathcal{H}_n$. By induction: For $t = N$, $V_N(h_N) = r_N(s_N) = V_N^*(h_N)$ for all $h_t, \pi$. 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 $\mathcal{D}^{\text{HR}}$. On the one hand, for $t = n$, it is trivial that

$$\begin{aligned} V_n(h_n) &= \sup_{a\in\mathcal{A}}\{r_n(s_t, a_t) + \sum_{j\in\mathcal{S}} p(j|s_n, a)V_{n+1}(h_n, a, j)\} \geq \sup_{a\in\mathcal{A}}\{r_n(s_n, a) + \sum_{j\in\mathcal{S}} p_n(j|s_n, a)V_{n+1}^*(h_n, a, j)\} \\ &\geq \sup_{a\in\mathcal{A}}\{r_n(s_n, a) + \sum_{j\in\mathcal{S}} p_n(j|s_n, a)V_{n+1}^{\pi'}(h_n, a, j)\} \geq V_n^{\pi'}(h_n). \end{aligned}$$

Step 2: Prove that for any $\varepsilon > 0$, there exists a $\pi \in \mathcal{D}^{\text{HD}}$ such that

$$V_n^{\pi'}(h_n) + (N-n)\varepsilon \geq V_n(h_n) \Rightarrow V_n^*(h_n) + (N-n)\varepsilon \geq V_n^{\pi'}(h_n) + (N-n)\varepsilon \geq V_n(h_n) \geq V_n^*(h_n).$$

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: For $t = N$, $V_N^{\pi'}(h_N) = V_N(h_N)$. Assume $V_t^{\pi'}(h_t) + (N-t)\varepsilon \geq V_t(h_t)$ for $t = n+1, \cdots, N$. For $t = n$,

$$V_n^{\pi'}(h_n) = r_n(s_n, \pi_n'(h_n)) + \sum_{j\in\mathcal{S}} p_n(j|s_n, \delta_n^{\pi'}(h_n))V_{n+1}^{\pi'}(h_n, \delta_n^{\pi'}(h_n), j) \geq V_n(h_n) - (N-n)\varepsilon. \qquad \square$$

**Remark** 3.1 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. $U_N^*(s) = \sup_{\pi\in\mathcal{D}^{\text{HR}}} U_N^\pi(s) = \sup_{\pi\in\mathcal{D}^{\text{HD}}} U_N^\pi(s) \stackrel{?}{=} \sup_{\pi\in\mathcal{D}^{\text{MD}}} U_N^\pi(s)$. We will answer "?" in the following theorem.

**Theorem** 3.2 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 $\varepsilon$-optimal policy which is D and M; (c) Maximum can be achieved, it is optimal, which is MD.

**Proof** We only prove (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** 3.4 (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 $s_t \in \mathcal{S}$ according to

$$V_t^*(s_t) = \max_{a\in\mathcal{A}}\{r_t(s_t, a) + \sum_{j\in\mathcal{S}} p_t(j|s_t, a)V_{t+1}^*(s_t)\},$$

and 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** 3.2 (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** 3.5 (Infinite-Horizon MDPs) Assumptions: Stationary reward and transition probabilities, i.e. $r_t(s, a) \equiv r(s, a), p_t(j|s, a) \equiv p(j|s, a)$; Bounded rewards, i.e. $|r(s, a)| \le M < \infty$ for all $a \in \mathcal{A}$ and $s \in \mathcal{S}$; Discounting coefficient $\lambda, 0 \le \lambda < 1$; Discrete state space $\mathcal{S}$. The expected total reward of policy $\pi = (\delta_1, \delta_2, \cdots) \in \mathcal{D}^{\mathrm{HR}}$:

$$U^\pi(s) = \lim_{N\to+\infty} \mathbb{E}_s^\pi\{\sum_{t=1}^{N} \lambda^{t-1} r(X_t, Y_t)\} = \mathbb{E}_s^\pi\{\sum_{t=1}^{+\infty} \lambda^{t-1} r(X_t, Y_t)\}.$$

We say that a policy $\pi^*$ is optimal when $U^{\pi^*}(s) \ge U^\pi(s)$ for each $s \in \mathcal{S}$ and all $\pi \in \mathcal{D}^{\mathrm{HR}}$. Define the value of the MDP $U^*(s) = \sup_{\pi\in\mathcal{D}^{\mathrm{HR}}} U^\pi(s)$. Let $U_\nu^\pi(s)$ denote the expected reward obtained by using $\pi$ when the horizon $\nu$ is random. Then $U_\nu^\pi(s) = \mathbb{E}_s^\pi\{\mathbb{E}_{\nu\sim P} \sum_{t=1}^{\nu} r(X_t, Y_t)\}$.

**Theorem** 3.3 Suppose $\nu$ has a GD($\lambda$), i.e. $\mathbb{P}(\nu = n) = \lambda^{n-1}(1-\lambda)$. 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** 3.4 Suppose $\pi \in \mathcal{D}^{\mathrm{HR}}$, then for each $s \in \mathcal{S}$, there exists a $\pi' \in \mathcal{D}^{\mathrm{MR}}$ for which $U^{\pi'}(s) = U^\pi(s)$.

**Proof** Note that

$$U^\pi(s) = \mathbb{E}_s^\pi\{\sum_{t=1}^{+\infty} \lambda^{t-1} r(X_t, Y_t)\} = \sum_{t=1}^{+\infty}\sum_{j\in\mathcal{S}}\sum_{a\in\mathcal{A}} \lambda^{t-1} r(j, a) p^\pi(X_t = j, Y_t = a|X_1 = s).$$

Fixing $s \in \mathcal{S}$, 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 $\delta_t'$ 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

$$\begin{aligned} p^{\pi'}(X_n = j, Y_n = a|X_1 = s) &= p^{\pi'}(Y_n = a|X_n = j, X_1 = s)p^{\pi'}(X_n = j|X_1 = s) \\ &= p^\pi(Y_n = a|X_n = j, X_1 = s)p^{\pi'}(X_n = j|X_1 = s). \end{aligned}$$

Then by induction assumption,

$$\begin{aligned} p^\pi(X_n = j|X_1 = s) &= \sum_{k\in\mathcal{S}}\sum_{a\in\mathcal{A}} p^\pi(X_{n-1} = k, Y_{n-1} = a|X_1 = s)p(j|k, a) \\ &= \sum_{k\in\mathcal{S}}\sum_{a\in\mathcal{A}} p^{\pi'}(X_{n-1} = k, Y_{n-1} = a|X_1 = s)p(j|k, a) = p^{\pi'}(X_n = j|X_1 = s) \end{aligned}$$

□

**Proposition** 3.1 (Vector expression for MDP) Let $\delta$ be 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 $\delta$, 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 $p_\pi^t$ satisfies

$$\begin{aligned} p_\pi^t(j|s) &= [p_{\delta_1} p_{\delta_2}\cdots p_{\delta_t}](j|s) = p^\pi(X_{t+1}=j|X_1=s) \\ \mathbb{E}_s^\pi g(X_t) &= \sum_{j\in\mathcal{S}} p_\pi^{t-1}(j|s) g(j) = (p_\pi^t g)_s \\ U^\pi &= \sum_{t=1}^{+\infty} \lambda^{t-1} p_\pi^{t-1} r_{\delta_t} = r_{\delta_1} + \lambda p_{\delta_1}(r_{\delta_1} + \lambda p_{\delta_2} r_{\delta_2} + \cdots) = r_{\delta_1} + \lambda p_{\delta_1} U^{\pi_1}. \end{aligned}$$

When $\pi$ is stationary, $U = r_\delta + \lambda p_\delta U$.

**Theorem** 3.5 Define $\mathscr{L}U = \sup_{d\in\mathcal{D}^{\text{MD}}}\{r_d + \lambda p_d U\}$. Suppose there exists a $U \in \mathcal{U}$ for which (a) $U \geq \mathscr{L}U$, then $U \geq U^*$; (b) $U \leq \mathscr{L}U$, then $U \leq U^*$; (c) $U = \mathscr{L}U$, then $U = U^*$.

**Proof** (a) By the given conditions,

$$\begin{aligned} U &\geq \sup_{\delta\in\mathcal{D}^{\text{MR}}}\{r_d + \lambda p_d U\} \geq r_{\delta_1} + \lambda p_{\delta_1} U \geq r_{\delta_1} + \lambda p_{\delta_1}(r_{\delta_2} + \lambda p_{\delta_2} U) \\ &\geq r_{\delta_1} + \lambda p_{\delta_1} r_{\delta_2} + \cdots + \lambda^{n-1} p_{\delta_1} p_{\delta_2}\cdots p_{\delta_{n-1}} r_{\delta_n} + \lambda^n p_\pi^n U \\ \Rightarrow U - U^\pi &\geq \lambda^n p_\pi^n U - \sum_{k=n}^{+\infty} \lambda^k p_\pi^k r_{\delta_{k+1}} \geq 0. \end{aligned}$$

(b) $U \leq \mathscr{L}U \Rightarrow U \leq r_d + \lambda p_d U + \varepsilon 1 \Rightarrow (I - \lambda p_d)U \leq r_d + \varepsilon 1 \Rightarrow U \leq (I-\lambda p_d)^{-1}(r_d + \varepsilon 1) = U^\pi + \varepsilon(1-\lambda)^{-1} 1_{|\mathcal{S}|}$.
(c) Omitted. □

**Theorem** 3.6 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 $\mathscr{L}v(s) \geq \mathscr{L}u(s)$ 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

$$\begin{aligned} 0 \leq \mathscr{L}v(s) - \mathscr{L}u(s) &\leq r(s,a_s^*) + \sum_{j\in\mathcal{S}} \lambda p(j|s,a_j^*) v(j) - r(s,a_j^*) - \sum_{j\in\mathcal{S}} \lambda p(j|s,a_s^*) u(j) \\ &= \lambda \sum_{j\in\mathcal{S}} p(j|s,a_s^*)(v(j) - u(j)) \leq \lambda \sum_{j\in\mathcal{S}} p(j|s,a_s^*) \|u - v\| = \lambda \|u - v\|. \end{aligned}$$

□

# 4 Statistical Learning Theory

**Definition** 4.1 (Basic concepts) $(X,Y) \sim P \in \mathcal{P}$, definite $(X_1,Y_1),\cdots,(X_n,Y_n)$ i.i.d., $\mathcal{D}_n = \{(X_1,Y_1),\cdots,(X_n,Y_n)\}$, risk $\mathcal{R}_n(f) = \mathbb{E}_{(X,Y)\in\mathcal{D}_n} l(X,Y)$. An algorithm $A$ is a mapping from $\mathcal{D}_n$ to a function $\mathcal{X} \to \mathcal{Y}$. Excess risk: $\mathcal{R}_P(A(\mathcal{D}_n)) - \mathcal{R}_P^*$. 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 $\varepsilon > 0$, $\mathbb{P}(\mathcal{R}_P(A(\mathcal{D}_n)) - \mathcal{R}_P^* \leq \varepsilon) \geq 1 - \delta$.

**Definition** 4.2 (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 (a) weakly consistent for a certain distribution of $(X,Y)$ if $\lim_{n\to+\infty} \mathbb{E}\{\int [g_n(x) - g(x)]\mu(\mathrm{d}x)\} = 0$; (b) strongly consistent for a certain distribution if $\lim_{n\to+\infty} \int [g_n(x) - g(x)]^2 \mu(\mathrm{d}x) = 0$ with probability 1; (c) weakly universally consistent if for all distributions of $(X,Y)$ with $\mathbb{E}[Y^2] < \infty$, $\cdots$; (d) strongly universally consistent $\cdots$.

**Definition** 4.3 (Penalized model) $g_n = \arg\min_f\{\frac{1}{n}\sum_{i=1}^n |f(X_i) - Y_i|^2 + J_n(f)\}$. Penalized term for $f$:

$$J_n(f) = \lambda_n \int |f''(t)|^2 \mathrm{d}t \text{ or } J_{n,k}(f) = \lambda_n \int \sum_{t_1,\cdots,t_k\in\{1,\cdots,d\}} \left|\frac{\partial f^k}{\partial x_{t_1}\cdots\partial x_{t_d}}\right|^2 \mathrm{d}t, \cdots$$

**Proposition** 4.1 (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 $\mathbb{P}\{\min_i \|X - X_i\|_\infty < t\} \le n\mathbb{P}(\|X - X_1\|_\infty \le t) \le n(2t)^d$, $d_\infty(d,n) \ge \frac{d}{2(d+1)} n^{-\frac{1}{d}}$.

**Theorem** 4.1 (No-Free lunch theorem) 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 $\pm 1$ valued, and $\limsup_{n\to+\infty} \frac{\mathbb{E}\|g_n - g\|^2}{a_n} \ge 1$.

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

$$\int_{A_j} |g_n(x) - g^{(c)}(x)|^2 \mu(\mathrm{d}x) = \int_{A_j} |g_n(x) - \bar{g}_n(x)|^2 \mu(\mathrm{d}x) + \int_A |\bar{g}_n(x) - g^{(c)}(x)|^2 \mu(\mathrm{d}x)$$
$$\ge \int_A |\bar{g}_n(x) - g^{(c)}(x)|^2 \mu(\mathrm{d}x).$$

Set $\hat{c}_{nj} = \begin{cases} 1 & \text{if } \int_{A_j} g_n(z)\mu(\mathrm{d}z) \ge 0 \\ -1 & \text{otherwise} \end{cases}$. For $x \in A_j$, if $\hat{c}_{nj} = 1$ and $c_j = -1$, then $\bar{g}_n(x) \ge 0$ and $g^{(c)}(x) = -1$, implying $|\bar{g}_n(x) - g^{(c)}(x)| \ge 1$; if $\hat{c}_{nj} = -1$ and $c_j = 1$, then $\bar{g}_n(x) < 0$ and $g^{(c)}(x) = 1$, also implying $|\bar{g}_n(x) - g^{(c)}(x)|^2 \ge 1$. Therefore,

$$\int_A |\bar{g}_n(x) - g^{(c)}(x)|^2 \mu(\mathrm{d}x) \ge 1_{\{\hat{c}_{nj} \ne c_j\}} \int_{A_j} 1\mu(\mathrm{d}x) \ge 1_{\{\hat{c}_{nj} \ne c_j\}} p_j \ge 1_{\{\hat{c}_{nj} \ne c_j\}} 1_{\{\mu_n(A_j) = 0\}} p_j$$
$$\Rightarrow \mathbb{E}\left\{\int |g_n(x) - g^{(c)}(x)|^2 \mu(\mathrm{d}x)\right\} \ge \sum_{j=1}^{+\infty} \mathbb{P}(\hat{c}_{nj} \ne c_j, \mu_n(A_j) = 0) p_j := R_n(c).$$

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

$$\mathbb{E}R_n(C) = \sum_{j=1}^{+\infty} \mathbb{E}\mathbb{P}(\hat{C}_{nj} \ne C_j, \mu_n(A_j) = 0) p_j \overset{\text{total expectation}}{=} \sum_{j=1}^{+\infty} \mathbb{E}\{1_{\{\mu_n(A_j)=0\}} \mathbb{P}(\hat{C}_{nj} \ne C_j | X_1, \cdots, X_n)\} p_j$$
$$= \frac{1}{2}\sum_{j=1}^{+\infty} \mathbb{P}(\mu_n(A_j) = 0) p_j = \frac{1}{2}\sum_{j=1}^{+\infty} (1 - p_j)^n p_j.$$

On the other hand,

$$R_n(c) \le \sum_{j=1}^{+\infty} \mathbb{P}(\mu_n(A_j) = 0) p_j = \sum_{j=1}^{+\infty} (1 - p_j)^n p_j \Rightarrow \frac{R_n(c)}{\mathbb{E}R_n(C)} \le 2.$$

By Fatou's lemma,

$$\mathbb{E}\left\{\limsup_{n\to+\infty} \frac{R_n(C)}{\mathbb{E}R_n(C)}\right\} \ge \limsup_{n\to+\infty} \left\{\frac{R_n(C)}{\mathbb{E}R_n(C)}\right\} = 1,$$

which implies that there exists $c \in C$ such that

$$\limsup_{n\to+\infty} \frac{R_n(C)}{\mathbb{E}R_n(C)} \ge 1 \Rightarrow \limsup_{n\to+\infty} \frac{\mathbb{E}\{\int |g_n(x) - g(x)|^2 \mu(\mathrm{d}x)\}}{\frac{1}{2}\sum_{j=1}^{+\infty} (1 - p_j)^n p_j} \ge 1.$$

Let $\{a_n\}$ be a sequence of positive numbers converging to 0 with $\frac{1}{2} \ge a_1 \ge a_2 \ge \cdots$, then there exists a probability $\{p_j\}$ such that $\sum_{j=1}^{+\infty} (1 - p_j)^n p_j \ge a_n, \forall n$. □

**Definition** 4.4 (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** 4.5 (Smoothness) Let $q=k+\beta$ for some $k\in\mathbb{N}$ and $0<\beta\le 1$ and let $\rho>0$. A function $f:\mathbb{R}^d\to\mathbb{R}$ is called $(q,\rho)$-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 $\left|\frac{\partial^k f}{\partial x_1^{\alpha_1}\cdots\partial x_d^{\alpha_d}}(x)-\frac{\partial^k f}{\partial x_1^{\alpha_1}\cdots\partial x_d^{\alpha_d}}(z)\right|\le\rho\|x-z\|^\beta$. Let $\mathscr{F}^{(q,\rho)}$ be the set of all $(q,\rho)$-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** 4.1 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,

$$\mathbb{P}(C=1|Z=z)=\frac{\mathbb{P}(C=1)\mathbb{P}(Z|C=1)}{\mathbb{P}(C=1)\mathbb{P}(Z|C=1)+\mathbb{P}(C=-1)\mathbb{P}(Z|C=-1)}=\frac{1}{1+\exp\left(\frac{\|Z-u\|^2}{2}-\frac{\|Z+u\|^2}{2}\right)}=\frac{1}{1+\exp(-2Z^Tu)}.$$

Therefore, the optimal Bayes decision is $g^*(Z)=\operatorname{sgn}(Z^Tu)$, and the risk is

$$\begin{aligned}\mathcal{R}^*&=\mathbb{P}(g^*(Z)\neq C)=\mathbb{P}(Z^Tu<0,C=1)+\mathbb{P}(Z^Tu>0,C=-1)\\&=\mathbb{P}(\|u\|^2+u^TN<0,C=1)+\mathbb{P}(-\|u\|^2+u^TN>0,C=-1)\\&=\frac{1}{2}\mathbb{P}(u^TN\le-\|u\|^2)+\frac{1}{2}\mathbb{P}(u^TN>\|u\|^2)=\Phi(-\|u\|).\end{aligned}$$

□

**Theorem** 4.2 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,

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

**Proof** Step 1: Construct an auxiliary function $g^{(c)}(x)$. Set $M_n=\lceil(\rho^2n)^{\frac{1}{2q+d}}\rceil$. Partition $[0,1]^d$ into $M_n^d$ 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 $\{\pm1\}$. Define $g^{(c_n)}(x)=\sum_{j=1}^{M_n^d}c_{n,j}f_{n,j}(x)$ 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 $\alpha=(\alpha_1,\cdots,\alpha_d),\alpha_i\in\mathbb{N},\sum_{j=1}^d\alpha_j=k$ and $D^\alpha=\frac{\partial^k}{\partial x_1^{\alpha_1}\cdots\partial x_d^{\alpha_d}}$. If $x,z\in A_{n,j}$,

$$|D^\alpha g^{c_n}(x)-D^\alpha g^{(c_n)}(z)|=|c_{n,k}||D^\alpha f_{n,j}(x)-D^\alpha f_{n,j}(z)|\le\rho\|x-z\|^\beta.$$

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 $A_{n,j}$. Then

$$\begin{aligned}|D^\alpha g^{(c_n)}(x)-D^\alpha g^{(c_n)}(z)|&\le|c_{n,i}D^\alpha f_{n,i}(x)|+|c_{n,j}D^\alpha f_{n,j}(z)|\\&=|c_{n,i}||D^\alpha f_{n,i}(x)-D^\alpha f_{n,i}(\bar{x})|+|c_{n,j}||D^\alpha f_{n,j}(z)-D^\alpha f_{n,j}(\bar{z})|\\&\le\rho2^{\beta-1}(\|x-\bar{x}\|^\beta+\|z-\bar{z}\|^\beta)=\rho2^\beta\left(\frac{\|x-\bar{x}\|^\beta}{2}+\frac{\|z-\bar{z}\|^\beta}{2}\right)\\&\le\rho2^\beta\left(\frac{\|x-\bar{x}\|}{2}+\frac{\|z-\bar{z}\|}{2}\right)^\beta\le\rho\|x-z\|^\beta.\end{aligned}$$

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 $\bar{c}_{n,j} = \mathrm{sgn}(\tilde{c}_{n,j})$, then

$$|\tilde{c}_{n,j} - c_{n,j}| \geq \frac{|\bar{c}_{n,j} - c_{n,j}|}{2} \Rightarrow \|g_n - g^{(c)}\|^2 \geq \int f^2(x) \mathrm{d}x \frac{1}{4} \frac{1}{M_n^{2q+d}} \sum_{j=1}^{M_n^d} (\bar{c}_{n,j} - c_{n,j})^2 = \frac{\rho^2}{M_n^{2q}} \int \bar{f}^2(x) \mathrm{d}x \frac{1}{M_n^d} \sum_{j=1}^{M_n^d} 1_{\{\bar{c}_{n,j} \neq c_{n,j}\}}.$$

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 4.1,

$$\mathbb{E}\{\mathbb{P}(\bar{C}_{n,1} \neq C_{n,1} | X_1, \cdots, X_n)\} = \mathbb{E}\Phi\left(-\sqrt{\sum_{r=1}^{t} f_{n,1}^2(X_{i_r})}\right) = \mathbb{E}\Phi\left(-\sqrt{\sum_{i=1}^{n} f_{n,1}^2(X_i)}\right)$$
$$\geq \Phi\left(-\sqrt{\mathbb{E}\sum_{i=1}^{n} f_{n,1}^2(X_i)}\right) \geq \Phi\left(-\sqrt{\int f^2(x) \mathrm{d}x}\right) > 0. \qquad \square$$

# 5 Uniform Laws of Large Numbers

**Definition** 5.1 (Background) 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** 5.1 (Hoeffding's inequality) For $g : \mathbb{R}^d \to [0, B]$, the following inequalities hold:

$$\mathbb{P}\left( \left| \frac{1}{n} \sum_{i=1}^{n} g(Z_i) - \mathbb{E}\{g(Z)\} \right| > \varepsilon \right) \leq 2e^{-\frac{2n\varepsilon^2}{B^2}} \Rightarrow \mathbb{P}\left( \sup_{g \in \mathscr{G}_n} \left| \frac{1}{n} \sum_{i=1}^{n} g(Z_i) - \mathbb{E}\{g(Z)\} \right| > \varepsilon \right) \leq 2|\mathscr{G}_n| e^{-\frac{2n\varepsilon^2}{B^2}}.$$

For finite class $\mathscr{G}$ satisfying $\sum_{n=1}^{+\infty} |\mathscr{G}_n| e^{-\frac{2n\varepsilon^2}{B^2}} < \infty$ for all $\varepsilon > 0$, by Borel-Cantelli lemma,

$$\mathbb{P}\left( \sup_{g \in \mathscr{G}_n} \left| \frac{1}{n} \sum_{i=1}^{n} g(Z_i) - \mathbb{E}\{g(Z)\} \right| > \varepsilon \text{ i.o.} \right) = 0$$

**Definition** 5.2 (Covering number) Let $\varepsilon > 0$ 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 $\|g - g_j\|_\infty < \varepsilon$ is called an $\varepsilon$-cover of $\mathscr{G}$ w.r.t. $\|\cdot\|_\infty$. Let $\mathscr{N}(\varepsilon, \mathscr{G}, \|\cdot\|_\infty)$ or $\mathscr{N}_\infty(\varepsilon, \mathscr{G})$ be the smallest $\varepsilon$-cover of $\mathscr{G}$ w.r.t. $\|\cdot\|_\infty$.

**Theorem** 5.1 For $n \in \mathbb{N}$, let $\mathscr{G}_n$ be a set of functions $g : \mathbb{R}^d \to [0, B]$ and let $\varepsilon > 0$. Then

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

**Proof** Let $\mathscr{G}_{n,\frac{\varepsilon}{3}}$ be an $\frac{\varepsilon}{3}$-cover of $\mathscr{G}_n$ w.r.t. $\|\cdot\|_\infty$ of minimal cardinality. Fix $g \in \mathscr{G}_n$, there exists $\bar{g} \in \mathscr{G}_{n,\frac{\varepsilon}{3}}$ such that $\|g - \bar{g}\|_\infty < \frac{\varepsilon}{3}$. Then

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^n g(Z_i) - \mathbb{E}g(Z)\right| &\le \left|\frac{1}{n}\sum_{i=1}^n (g(Z_i) - \bar{g}(Z_i))\right| + \left|\frac{1}{n}\sum_{i=1}^n \bar{g}(Z_i) - \mathbb{E}\{\bar{g}(Z)\}\right| + |\mathbb{E}\bar{g}(Z) - \mathbb{E}g(Z)| \\ &\le \frac{2\varepsilon}{3} + \left|\frac{1}{n}\sum_{i=1}^n \bar{g}(Z_i) - \mathbb{E}\{\bar{g}(Z)\}\right|, \\ \Rightarrow \mathbb{P}\left(\sup_{g\in\mathscr{G}_n}\left|\frac{1}{n}\sum_{i=1}^n g(Z_i) - \mathbb{E}\{g(Z)\}\right| > \varepsilon\right) &\le \mathbb{P}\left(\sup_{g\in\mathscr{G}_{n,\frac{\varepsilon}{3}}}\left|\frac{1}{n}\sum_{i=1}^n g(Z_i) - \mathbb{E}\{g(Z)\}\right| > \frac{\varepsilon}{3}\right)\end{aligned}$$

Then use Hoeffding's inequality. □

**Definition** 5.3 Let $\varepsilon > 0$, $\mathscr{G}$ be a set of functions $\mathbb{R}^d \to \mathbb{R}$, $1 \le p < \infty$, and $\nu$ 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 $\|g - g_j\|_{L_p(\nu)} < \varepsilon$ is called a $\varepsilon$-cover of $\mathscr{G}$. Similarly define $\mathscr{N}(\varepsilon, \mathscr{G}, \|\cdot\|_{L_p(\nu)})$. (b) Let $Z^{1:n} = (Z_1, \cdots, Z_n) \subset \mathbb{R}^d$ and $\nu_n$ 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 $\mathscr{N}_p(\varepsilon, \mathscr{G}, Z^{1:n})$.

**Definition** 5.4 (Packing number) (a) Every finite collection of functions $g_1, \cdots, g_N \in \mathscr{G}$ with $\|g_j - g_k\|_{L_p(\nu)} \ge \varepsilon$ for all $1 \le j < k \le N$ is called $\varepsilon$-packing of $\mathscr{G}$ with $\|\cdot\|_{L_p(\nu)}$. The largest $\varepsilon$-packing is denoted as $\mathscr{M}(\varepsilon, \mathscr{G}, \|\cdot\|_{L_p(\nu)})$. Similarly define $\mathscr{M}(\varepsilon, \mathscr{G}, Z^{1:n})$.

**Property** 5.1 (Covering number v.s. packing number)

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

**Theorem** 5.2 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 $R > 0, \varepsilon > 0$, and $z_1, \cdots, z_n \in \mathbb{R}^d$,

$$\mathscr{N}_2\left(\varepsilon, \left\{f \in \mathscr{F} : \frac{1}{n}\sum_{i=1}^n |f(z_i)|^2 \le R^2\right\}, Z^{1:n}\right) \le \left(\frac{4R + \varepsilon}{\varepsilon}\right)^D.$$

**Definition** 5.5 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** 5.6 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 $n$th 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** 5.7 (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** 5.1 $S(\mathscr{A}, n) \le \sum_{i=0}^{V_{\mathscr{A}}} \binom{n}{i}$.

**Theorem** 5.3 Let $\mathscr{G}$ be a set of functions $g : \mathbb{R}^d \to [0, B]$. For any $n \in \mathbb{N}$ and $\varepsilon > 0$,

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

**Proof** Step 1: Symmetrization. Let $Z'^{1:n}$ be i.i.d. samples from the same distribution and independent of $Z^{1:n}$ and $g^* \in \mathscr{G}$ be a function such that $\left|\frac{1}{n}\sum_{i=1}^n g(Z_i) - \mathbb{E}g(Z)\right| > \varepsilon$ if there exists such one. Otherwise, let $g^*$ be an arbitrary function in $\mathscr{G}$. $g^*(z)$ depends on $Z^{1:n}$ and $\mathbb{P}\left\{\left|\mathbb{E}[g^*(Z)|Z^{1:n}] - \frac{1}{n}\sum_{i=1}^n g^*(Z'_i)\right| > \frac{\varepsilon}{2}\middle| Z^{1:n}\right\} \le \frac{\mathrm{Var}(g^*(Z)|Z^{1:n})}{n(\frac{\varepsilon}{2})^2} \le \frac{B^2/4}{n\varepsilon^2/4} = \frac{B^2}{n\varepsilon^2} \le \frac{1}{2}$ holds for $n \ge \frac{2B^2}{\varepsilon^2}$. Thus we have

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

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

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 $Z'^{1:n}$.

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

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

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

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

**Theorem** 5.4 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 $p \ge 1$, $\nu$ be a probability measure on $\mathbb{R}^d$ and $0 < \varepsilon < \frac{B}{4}$. Then

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

**Proof** Step 1: Set $p = 1$. Relate $\mathscr{M}(\varepsilon, \mathscr{G}, \|\cdot\|_{L_p(\nu)})$ to a shatter coefficient of $\mathscr{G}^+$. Set $m = \mathscr{M}(\varepsilon, \mathscr{G}, \|\cdot\|_{L_p(\nu)})$ and let $\bar{\mathscr{G}} = \{g_1, \cdots, g_m\}$ be a $\varepsilon$-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 $\nu$. 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

$$S(\mathscr{G}^+, K) = \max_{\{z_1,\cdots,z_K\}\in\mathbb{R}^d\times\mathbb{R}} s(\mathscr{G}^+, \{z_1, \cdots, z_K\}) \ge \mathbb{E}s(\mathscr{G}_+, \{R_1, \cdots, R_K\}) \ge \mathbb{E}s(\{\mathscr{G}_f : f \in \mathscr{G}\}, \{R_1, \cdots, R_K\})$$

$$\geq \mathbb{E}s(\{\mathscr{G}_f : f \in \mathscr{G}, \mathscr{G}_f \cap R^{1:K} \neq \mathscr{G}_g \cap R^{1:K} \text{ for all } g \in \bar{\mathscr{G}}, g \neq f\}, R^{1:K})$$

$$= \mathbb{E}\left\{\sum_{f\in\bar{\mathscr{G}}} 1_{\{\mathscr{G}_f \cap R^{1:K} \neq \mathscr{G}_g \cap R^{1:K} \text{ for all } g\in\mathscr{G}, g\neq f\}}\right\} = \sum_{f\in\bar{\mathscr{G}}} \mathbb{P}(\mathscr{G}_f \cap R^{1:K} \neq \mathscr{G}_g \cap R^{1:K} \text{ for all } g \in \mathscr{G}, g \neq f)$$

$$= \sum_{f\in\bar{\mathscr{G}}} \left(1 - \mathbb{P}(\exists g \in \bar{\mathscr{G}}, g \neq f, \mathscr{G}_f \cap R^{1:K} = \mathscr{G}_g \cap R^{1:K})\right) \geq \sum_{f\in\bar{\mathscr{G}}} \left(1 - m \max_{g\in\bar{\mathscr{G}}, g\neq f} \mathbb{P}(\mathscr{G}_f \cap R^{1:K} = \mathscr{G}_g \cap R^{1:K})\right).$$

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

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

and

$$\mathbb{P}(\mathscr{G}_f \cap \{R_1\} = \mathscr{G}_g \cap \{R_1\}) = 1 - \mathbb{P}(\mathscr{G}_f \cap \{R_1\} \neq \mathscr{G}_g \cap \{R_1\}) = 1 - \mathbb{E}[\mathbb{P}(\mathscr{G}_f \cap \{R_1\} \neq \mathscr{G}_g \cap \{R_1\}|Q_1)]$$

$$= 1 - \mathbb{E}[\mathbb{P}(f(Q_1) < T \leq g(Q_1) \text{ or } g(Q_1) < T \leq f(Q_1)|Q_1)] = 1 - \mathbb{E}\left[\frac{|f(Q_1) - g(Q_1)|}{B}\right]$$

$$= 1 - \frac{1}{B}\int |f(x) - g(x)|\nu(\mathrm{d}x) \leq 1 - \frac{\varepsilon}{B} \Rightarrow \mathbb{P}(\mathscr{G}_f \cap \{R_1\} = \mathscr{G}_g \cap \{R_1\})^K \leq \left(1 - \frac{\varepsilon}{B}\right)^K \leq \exp\left(-\frac{\varepsilon K}{B}\right)$$

$$\Rightarrow S(\mathscr{G}^+, K) \geq m\left(1 - m\exp\left(-\frac{\varepsilon K}{B}\right)\right).$$

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

$$1 - m\exp\left(-\frac{\varepsilon K}{B}\right) \geq 1 - m\exp\left(-\frac{\varepsilon}{B}\left(\frac{B}{\varepsilon}\log(2m) - 1\right)\right) = 1 - \frac{1}{2}\exp\left(\frac{\varepsilon}{B}\right) \geq 1 - \frac{1}{2}\exp\left(\frac{1}{4}\right) \geq \frac{1}{3} \Rightarrow m \leq 3S(\mathscr{G}_+, K).$$

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

**Lemma** 5.2 Let $\mathscr{A} \in \mathbb{R}^d$ and $V_{\mathscr{A}} < \infty$. Then $\forall n \in \mathbb{N}, S(\mathscr{A}, n) \leq (n+1)^{V_{\mathscr{A}}}$ and $\forall n \geq V_{\mathscr{A}}, S(\mathscr{A}, n) \leq (\frac{en}{V_{\mathscr{A}}})^{V_{\mathscr{A}}}$.

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

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

Step 4: Let $1 < p < \infty$. Then for any $g_j, g_k \in \mathscr{G}$,

$$\|g_j - g_k\|^p_{L_p(\nu)} \leq B^{p-1}\|g_j - g_k\|_{L_1(\nu)} \Rightarrow \mathscr{M}(\varepsilon, \mathscr{G}, \|\cdot\|_{L_p(\nu)}) \leq \mathscr{M}\left(\frac{\varepsilon^p}{B^{p-1}}, \mathscr{G}, \|\cdot\|_{L_p(\nu)}\right). \qquad \square$$

**Theorem** 5.5 (ULLN) 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 $\mathscr{G}_L := \{g \cdot 1_{\{G\leq L\}} : g \in \mathscr{G}\}$. For $g \in \mathscr{G}$,

$$\left|\frac{1}{n}\sum_{i=1}^n g(Z_i) - \mathbb{E}g(Z)\right| \leq \left|\frac{1}{n}\sum_{i=1}^n g(Z_i) - \frac{1}{n}\sum_{i=1}^n g(Z_i)1_{\{G(Z_i)\leq L\}}\right|$$

$$+ \left|\frac{1}{n}\sum_{i=1}^n g(Z_i)1_{\{G(Z_i)\leq L\}} - \mathbb{E}\{g(Z)1_{G(Z)\leq L}\}\right| + \left|\mathbb{E}\{g(Z)1_{G(Z)\leq L}\} - \mathbb{E}\{g(Z)\}\right|$$

$$= \left|\frac{1}{n}\sum_{i=1}^n g(Z_i)1_{\{G(Z_i)>L\}}\right| + \mathbb{E}|g(Z)|1_{\{G(Z)>L\}} + \left|\frac{1}{n}\sum_{i=1}^n g(Z_i)1_{\{G(Z_i)\leq L\}} - \mathbb{E}\{g(Z)1_{G(Z)\leq L}\}\right|$$

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

# 6 Least Square Estimates: Consistency and Convergence Rate

**Definition** 6.1 (Notation) $\mathbb{E}\{(m(X)-Y)^2\} = \inf_f \mathbb{E}\{(f(X)-Y)^2\} \Rightarrow m(X) = \mathbb{E}[Y|X]$. Define

$$m_n = \arg\min_{f\in\mathscr{F}_n} \frac{1}{n}\sum_{i=1}^n |f(X_i)-Y_i|^2, m^* = \arg\min_{f\in\mathscr{F}_n} \mathbb{E}\{(f(X)-Y)^2\}.$$

**Theorem** 6.1 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:

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

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

**Proposition** 6.1 (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 the set of functions $\cup_{k=1}^{+\infty}\{\sum_{j=1}^k a_j\psi_j(x) : a_1,\cdots,a_k\in\mathbb{R}\}$ is dense in $L_2(\mu)$ for any probability measure $\mu$ on $\mathbb{R}^d$. Define the regression function estimate $m_n$ as a function minimizing the empirical $L_2$ risk $\frac{1}{n}\sum_{i=1}^n (f(X_i)-Y_i)^2$ over the function form $f(x) = \sum_{j=1}^{k_n} a_j\psi_j(x)$ with $\sum_{j=1}^{k_n}|a_j|\le\beta_n$. If $\mathbb{E}(Y^2)<\infty$ and $k_n$ and $\beta_n$ 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.

**Proposition** 6.2 Consider $\mathscr{F}_n = \left\{\sum_{j=1}^{k_n} a_j\psi_j(x) : \sum_{j=1}^{k_n}|a_j|\le\beta_n\right\}$ and $\widetilde{\mathscr{F}}_n = \left\{\sum_{j=1}^{k_n} a_j\psi_j(x) : a_j\in\mathbb{R}\right\}$. 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 $T_L u = \begin{cases} u, & \text{if } |u|\le L \\ L\mathrm{sgn}(u), & \text{otherwise}\end{cases}$. (a) If $\mathbb{E}(Y^2)<\infty$ and $k_n$ and $\beta_n$ satisfy $k_n\to\infty, \beta_n\to\infty, \frac{k_n\beta_n^4\log\beta_n}{n}\to 0$, then $\mathbb{E}\left\{\int (m_n(x)-m(x))^2\mu(\mathrm{d}x)\right\}\to 0$. (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** 6.3 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** 6.2 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 $\beta_n\to+\infty, \mathbb{E}\{\sim\}\to 0, \mathbb{E}\{\sim\}\to 0$, then $\mathbb{E}\{\sim\}\to 0$.

**Definition** 6.2 (Piecewise polynomial partition estimate) $\mathcal{P}_n = \{A_{n,1}, A_{n,2},\cdots\}$ be a partition of $\mathbb{R}^d$,

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

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

**Theorem** 6.3 Let $\mathscr{F}$ be a class of function $f : \mathbb{R}^d \to \mathbb{R}$ bounded in abolute value by $B$. Let $\varepsilon > 0$. Then

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

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

**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 $\|f\|_{n'}^2 = \frac{1}{n}\sum_{i=n+1}^{2n} |f(X_i)|^2$. Let $f^*$ be a function $f \in \mathscr{F}$ such that $\|f\|_2 - 2\|f\|_n > \varepsilon$ if there exists any such function, and let $f^*$ be an arbitrary function in $\mathscr{F}$ if such a function does not exist. Then

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

Therefore,

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

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

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

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

Step 3: Conditioning and introduction of a covery. Let $\mathscr{G} = \{g_j : j = 1, \cdots, \mathscr{N}_2(\frac{\sqrt{2}}{24}\varepsilon, \mathscr{F}, X^{1:2n})\}$ be a $\frac{\sqrt{2}}{24}\varepsilon$-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 $f \in \mathscr{F}$, $\|f - g\|_{2n} \le \frac{\sqrt{2}}{24}\varepsilon$. Then

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

In this way,

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

Step 4: Application of Hoeffding's inequality.

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

Then

$$\begin{aligned}\mathbb{P}\{(\frac{1}{n}\sum_{i=n+1}^{2n}|g(Z_i)|^2)^{\frac{1}{2}}-(\frac{1}{n}\sum_{i=1}^{n}|g(Z_i)|^2)^{\frac{1}{2}}>\frac{\varepsilon}{12}|X^{1:2n}\}&\le2\exp\left(-\frac{2n^2\frac{\varepsilon^2}{144}(\frac{1}{n}\sum_{i=1}^{2n}|g(X_i)|^2)}{\sum_{i=1}^{n}4(|g(X_i)|^2-|g(X_{i+n})|^2)^2}\right)\\ &\le2\exp\left(-\frac{2n^2\frac{\varepsilon^2}{144}(\frac{1}{n}\sum_{i=1}^{2n}|g(X_i)|^2)}{\sum_{i=1}^{n}4B^2(|g(X_i)|^2+|g(X_{i+n})|^2)}\right)\\ &=\exp\left(-\frac{n\varepsilon^2}{288B^2}\right).\end{aligned}$$

□

**Theorem** 6.4 Assume $\sigma^2=\sup_{x\in\mathbb{R}^d}\mathrm{Var}(Y|X=x)<\infty$. 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

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

Write that $\widetilde{m}_n=\sum_{j=1}^{k_n}a_jf_{j,n}$ 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 $\frac{1}{n}B^TBa=\frac{1}{n}B^TY$, $B=(f_{j,n}(X_i))_{1\le i\le n,1\le j\le k_n}$ and $Y=(Y_1,\cdots,Y_n)^T$. Then

$$\mathbb{E}^*\{\widetilde{m}_n\}=\sum_{j=1}^{k_n}\mathbb{E}^*\{a_j\}f_{j,n}\text{ and }\frac{1}{n}B^TB\mathbb{E}^*a=\frac{1}{n}B^T\mathbb{E}^*Y=\frac{1}{n}B^T(m(X_1),\cdots,m(X_n))^T$$

$$\Rightarrow \|\mathbb{E}^*(\widetilde{m}_n) - m\|_n^2 = \min_{f\in\mathscr{F}_n} \|f-m\|_n^2.$$

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 $\{X_1,\cdots,X_n\}$, $\mathrm{span}\{f_1,\cdots,f_k\}\subset\mathscr{F}_n$, $\widetilde{m}_n(x) = f(x)^T\frac{1}{n}B^TY$ where $B=(f_j(X_i))_{1\le j\le n,1\le j\le k}$, $B^TB=I$. Therefore,

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

□

**Theorem** 6.5 Assume $\sigma^2=\sup_{x\in\mathbb{R}^d}\mathrm{Var}(Y|X=x)<\infty$ 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,

$$\begin{aligned}\mathbb{P}\left(2\max\{\|m_n-m\|_2-2\|m_n-m\|_n,0\}>u\right)&\le\mathbb{P}\left(\exists f\in T_L\mathscr{F}_n:\|f-m\|_2-2\|f-m\|_n>\sqrt{\frac{u}{2}}\right)\\ &\le 3\mathbb{E}\mathscr{N}_2\left(\frac{\sqrt{u}}{24},\mathscr{F}_n,X^{1:2n}\right)\exp\left(-\frac{nu}{576(2L)^2}\right)\\ &\le 9(12en)^{2(k_n+1)}\exp\left(-\frac{nu}{2304L^2}\right)\\ \Rightarrow\mathbb{E}(2\max\{\|m_n-m\|_2-2\|m_n-m\|_n,0\})&\le u+\int_u^\infty\mathbb{P}(2\max\{\|m_n-m\|_2-2\|m_n-m\|_n,0\}>t)\mathrm{d}t\\ \left(\text{take } u\ge\frac{576L^2}{n}\right)&\le CL^2\frac{\log(n)+1}{n}k_n.\end{aligned}$$

Combine these two bounds together. □

**Property** 6.1 (Nonlinear LSE) $|Y|\le L\le\beta_n$ a.s., $m_n(\cdot)=T_{\beta_n}\widetilde{m}_n(\cdot), \widetilde{m}_n(\cdot)=\arg\min_{f\in\mathscr{F}_n}\frac{1}{n}\sum_{i=1}^n|f(X_i)-Y_i|^2$. We do the following decomposition:

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

On the one hand,

$$\mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^n[|m_n(X_i)-Y_i|^2-|m(X_i)-Y_i|^2]\right\}\le\mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^n|\widetilde{m}_n(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\right\}$$

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

On the other hand,

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

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

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

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

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

**Theorem** 6.6 Let $n\in\mathbb{N}$ and $1\le L<\infty$. Assume $|Y|\le L$ a.s. Let estimate $m_n$ be defined by minimization of the empirical $l_2$ risk over a set of functions $\mathscr{F}_n$ and truncation at $L$. Then one has

$$
\mathbb{E}\int|m_n(x)-m(x)|^2\mu(\mathrm{d}x)\le\frac{c_1}{n}+\frac{(c_2+c_3\log n)V_{\mathscr{F}_n^+}}{n}+2\inf_{f\in\mathscr{F}_n}\int|f(x)-m(x)|^2\mu(\mathrm{d}x)
$$

**Proof** We first introduce a theorem/lemma.

**Theorem** 6.7 **Lemma** 6.1 Assume $|Y|\le B$ a.s. and $B\ge 1$. Let $\mathscr{F}$ be a set of functions $f:\mathbb{R}^d\to\mathbb{R}$ and $|f(x)|\le B$. Then for any $n\ge 1$, $\alpha,\beta>0$ and $0<\varepsilon\le\frac{1}{2}$,

$$
\begin{aligned}
&\mathbb{P}\Bigg\{\exists f\in\mathscr{F}:\mathbb{E}|f(X)-Y|^2-\mathbb{E}|m(X)-Y|^2-\frac{1}{n}\sum_{i=1}^n\left[|f(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\right]\\
&\qquad\qquad\ge\varepsilon(\alpha+\beta+\mathbb{E}|f(X)-Y|^2-\mathbb{E}|m(X)-Y|^2)\Bigg\}\\
&\le 14\sup_{X^{1:n}}\mathscr{N}_1\left(\frac{\beta\varepsilon}{20B},\mathscr{F},X^{1:n}\right)\exp\left(-\frac{\varepsilon^2(1-\varepsilon)\alpha n}{214(1+\varepsilon)B^2}\right).
\end{aligned}
$$

Now let's return to the original **Theorem** 6.6.

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

Since

$$
\mathbb{E}(T_{2,n})\le 2\inf_{f\in\mathscr{F}_n}\int|f(x)-m(x)|^2\mu(\mathrm{d}x),\mathbb{E}(T_{1,n})=\int_0^\infty\mathbb{P}(T_{1,n}>t)\mathrm{d}t\le\varepsilon+\int_\varepsilon^\infty\mathbb{P}(T_{1,n}>t)\mathrm{d}t
$$

and

$$\begin{aligned}\mathbb{P}(T_{1,n}>t)&=\mathbb{P}\Bigg\{\mathbb{E}[|m_n(X)-Y|^2|\mathcal{D}_n]-\mathbb{E}[|m(X)-Y|^2]-\frac{1}{n}\sum_{i=1}^n[|m_n(X_i)-Y_i|^2-|m(X_i)-Y_i|^2]\\&\qquad\geq\frac{1}{2}\left(\frac{t}{2}+\frac{t}{2}+\mathbb{E}[|m_n(X)-Y|^2|\mathcal{D}_n]-\mathbb{E}[|m(X)-Y|^2]\right)\Bigg\}\\&\leq\mathbb{P}\Bigg\{\exists f\in T_L\mathscr{F}_n:\mathbb{E}|f(X)-Y|^2-\mathbb{E}|m(X)-Y|^2-\frac{1}{n}\sum_{i=1}^n[|f(X_i)-Y_i|^2-|m(X_i)-Y_i|^2]\\&\qquad\geq\frac{1}{2}\left(\frac{t}{2}+\frac{t}{2}+\mathbb{E}|f(X)-Y|^2-\mathbb{E}|m(X)-Y|^2\right)\Bigg\}\\&\text{(by lemma 6.1)}\leq 14\sup_{X^{1:n}}\mathscr{N}_1\left(\frac{1}{80Ln},T_L\mathscr{F}_n,X^{1:n}\right)\exp\left(-\frac{nt}{24\cdot 214L^4}\right)\\&\leq 3(480eL^2n)^{2V_{(T_L\mathscr{F}_n)^+}}\exp\left(-\frac{nt}{24\cdot 214L^4}\right).\end{aligned}$$

Plug this bound into the integral in the previous expectation bound,

$$\mathbb{E}(T_{1,n})\leq\varepsilon+\frac{24\cdot 214L^4}{n}42(480eL^2n)^{2V_{\mathscr{F}_n^+}}\exp\left(-\frac{n\varepsilon}{24\cdot 214L^4}\right).$$

□

**Lemma** 6.2 Let $V_1,\cdots,V_n$ i.i.d. r.v.'s, $0\leq V_i\leq B, 0<\alpha<1$ and $\nu>0$. Then

$$\mathbb{P}\left\{\frac{\left|\frac{1}{n}\sum_{i=1}^n V_i-\mathbb{E}V_1\right|}{\nu+\frac{1}{n}\sum_{i=1}^n V_i+\mathbb{E}V_1}>\alpha\right\}\leq\mathbb{P}\left\{\frac{\left|\frac{1}{n}\sum_{i=1}^n V_i-\mathbb{E}V_1\right|}{\nu+\mathbb{E}V_1}>\alpha\right\}<\frac{B}{4\alpha^2\nu n}.$$

**Proof** The first inequality is trivial. For the second, note that

$$\mathbb{P}\left\{\left|\sum_{i=1}^n(V_i-\mathbb{E}V_1)\right|>\alpha n(\nu+\mathbb{E}V_1)\right\}\leq\frac{\mathbb{E}|\sum_{i=1}^n(V_i-\mathbb{E}V_1)|^2}{[\alpha n(\nu+\mathbb{E}V_1)]^2}=\frac{\mathrm{Var}(V_1)}{n\alpha^2(\nu+\mathbb{E}V_1)^2}\leq\frac{\mathbb{E}V_1(B-\mathbb{E}V_1)}{n\alpha^2(\nu+\mathbb{E}V_1)^2}$$

where the last inequality holds since

$$\mathrm{Var}(V_1)=\mathbb{E}\{(V_1-\mathbb{E}V_1)(V_1-\mathbb{E}V_1)=\mathbb{E}V_1(V_1-\mathbb{E}V_1)\leq\mathbb{E}V_1(B-\mathbb{E}V_1)\}.$$

In addition,

$$\frac{\mathbb{E}V_1(B-\mathbb{E}V_1)}{n\alpha^2(\nu+\mathbb{E}V_1)^2}\leq\max_{x\in[0,\beta]}\frac{x(B-x)}{n\alpha^2(\nu+x)^2}=\frac{B^2}{4\alpha^2\nu n(B+\nu)}<\frac{B}{4\alpha^2\nu n}.$$

□

**Theorem** 6.8 Let $B\geq 1$ and $G$ be a set of functions $g:\mathbb{R}^d\to[0,B]$. Let $Z_1,\cdots,Z_n$ be i.i.d. $\mathbb{R}^d$-valued r.v.'s. Assume $\alpha>0, 0<\varepsilon<1$ and $n\geq 1$. Then

$$\mathbb{P}\left\{\sup_{g\in\mathscr{G}}\frac{\frac{1}{n}\sum_{i=1}^n g(Z_i)-\mathbb{E}g(Z)}{\alpha+\frac{1}{n}\sum_{i=1}^n g(Z_i)+\mathbb{E}g(Z)}>\varepsilon\right\}\leq 4\mathbb{E}\mathscr{N}_1\left(\frac{2\varepsilon}{5},G,Z^{1:n}\right)\exp\left(-\frac{3\varepsilon^2\alpha n}{40B}\right).$$

**Proof** Step 1: Replace the expectation with empirical mean. Ghost sample $Z'_{1:n}=(Z'_1,\cdots,Z'_n)$ i.i.d. Let $g^*$ be a function $g\in\mathscr{G}$ such that

$$\frac{1}{n}\sum_{i=1}^n g(Z_i)-\mathbb{E}g(Z)>\varepsilon\left(\alpha+\frac{1}{n}\sum_{i=1}^n g(Z_i)+\mathbb{E}g(Z)\right)$$

if there exists any such function. Otherwise, let $g^*$ be an arbitrary function in $G$. $g^*$ depenods on $Z^{1:n}$. Since

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n g(Z_i)-\mathbb{E}g(Z)>\varepsilon\left(\alpha+\frac{1}{n}\sum_{i=1}^n g(Z_i)+\mathbb{E}g(Z)\right)\text{ and }\frac{1}{n}\sum_{i=1}^n g(Z'_i)-\mathbb{E}g(Z)\leq\frac{\varepsilon}{4}\left(\alpha+\frac{1}{n}\sum_{i=1}^n g(Z'_i)+\mathbb{E}g(Z)\right)\\&\Rightarrow\frac{1}{n}\sum_{i=1}^n g(Z_i)-\frac{1}{n}\sum_{i=1}^n g(Z'_i)>\frac{3}{4}\varepsilon\alpha+\frac{\varepsilon}{n}\sum_{i=1}^n g(Z_i)-\frac{\varepsilon}{n}\sum_{i=1}^n g(Z'_i)+\frac{3\varepsilon}{4}\mathbb{E}g(Z)\\&\Leftrightarrow\left(1-\frac{5}{8}\varepsilon\right)\left(\frac{1}{n}\sum_{i=1}^n g(Z_i)-\frac{1}{n}\sum_{i=1}^n g(Z'_i)\right)>\frac{3}{8}\varepsilon\left(2\alpha+\frac{1}{n}\sum_{i=1}^n g(Z_i)+\frac{1}{n}\sum_{i=1}^n g(Z'_i)\right)+\frac{3\varepsilon}{4}\mathbb{E}g(Z)\end{aligned}$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^n g(Z_i) - \frac{1}{n}\sum_{i=1}^n g(Z_i') > \frac{3\varepsilon}{8}\left(2\alpha + \frac{1}{n}\sum_{i=1}^n g(Z_i) + \frac{1}{n}\sum_{i=1}^n g(Z_i')\right),$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left\{\exists g \in \mathscr{G} : \frac{1}{n}\sum_{i=1}^n g(Z_i) - \frac{1}{n}\sum_{i=1}^n g(Z_i') > \frac{3}{8}\varepsilon\left(2\alpha + \frac{1}{n}\sum_{i=1}^n g(Z_i) + \frac{1}{n}\sum_{i=1}^n g(Z_i')\right)\right\} \\
\geq\ &\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n g^*(Z_i) - \frac{1}{n}\sum_{i=1}^n g^*(Z_i') > \frac{3\varepsilon}{8}\left(2\alpha + \frac{1}{n}\sum_{i=1}^n g^*(Z_i) + \frac{1}{n}\sum_{i=1}^n g^*(Z_i')\right)\right\} \\
\geq\ &\mathbb{P}\Bigg\{\frac{1}{n}\sum_{i=1}^n g^*(Z_i) - \mathbb{E}[g^*(Z)|Z^{1:n}] > \varepsilon\left(\alpha + \frac{1}{n}\sum_{i=1}^n g^*(Z_i) + \mathbb{E}[g^*(Z)|Z^{1:n}]\right) \text{ and} \\
&\quad \frac{1}{n}\sum_{i=1}^n g^*(Z_i') - \mathbb{E}[g^*(Z)|Z^{1:n}] \leq \frac{\varepsilon}{4}\left(\alpha + \frac{1}{n}\sum_{i=1}^n g^*(Z_i') + \mathbb{E}[g^*(Z)|Z^{1:n}]\right)\Bigg\} \\
=\ &\mathbb{E}1_{\left\{\frac{1}{n}\sum_{i=1}^n g^*(Z_i) - \mathbb{E}[g^*(Z)|Z^{1:n}] > \varepsilon(\cdots)\right\}} \underbrace{\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n g^*(Z_i') - \mathbb{E}[g^*(Z)|Z^{1:n}] \leq \frac{\varepsilon}{4}(\cdots)\right\}}_{>1-\frac{B}{4(\frac{\varepsilon}{4})^2\alpha n} = 1-\frac{4B}{\varepsilon^2\alpha n} \geq \frac{1}{2} \text{ for } n > \frac{8B}{\varepsilon^2\alpha} \text{ by lemma 6.2}} \\
\geq\ &\frac{1}{2}\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n g^*(Z_i) - \mathbb{E}[g^*(Z)|Z^{1:n}] > \varepsilon\left(\alpha + \frac{1}{n}\sum_{i=1}^n g^*(Z_i) + \mathbb{E}[g^*(Z)|Z^{1:n}]\right)\right\} \\
=\ &\frac{1}{2}\mathbb{P}\left\{\exists g \in \mathscr{G} : \frac{1}{n}\sum_{i=1}^n g(Z_i) - \mathbb{E}[g(Z)] > \varepsilon\left(\alpha + \frac{1}{n}\sum_{i=1}^n g(Z_i) + \mathbb{E}[g(Z)]\right)\right\}
\end{aligned}$$

Step 2: Symmetrization. Let $U_1, \cdots, U_n \overset{\text{i.i.d.}}{\sim} U\{-1,1\}$ independent of $Z_{1:n}', Z^{1:n}$. Therefore,

$$\begin{aligned}
&\mathbb{P}\left\{\exists g \in \mathscr{G} : \frac{1}{n}\sum_{i=1}^n g(Z_i) - \frac{1}{n}\sum_{i=1}^n g(Z_i') > \frac{3}{8}\varepsilon\left(2\alpha + \frac{1}{n}\sum_{i=1}^n g(Z_i) + \frac{1}{n}\sum_{i=1}^n g(Z_i')\right)\right\} \\
=\ &\mathbb{P}\left\{\exists g \in \mathscr{G} : \frac{1}{n}\sum_{i=1}^n U_i[g(Z_i) - \frac{1}{n}\sum_{i=1}^n g(Z_i')] > \frac{3}{8}\varepsilon\left(2\alpha + \frac{1}{n}\sum_{i=1}^n g(Z_i) + \frac{1}{n}\sum_{i=1}^n g(Z_i')\right)\right\} \\
\leq\ &\mathbb{P}\left\{\exists g \in \mathscr{G} : \frac{1}{n}\sum_{i=1}^n U_i g(Z_i) > \frac{3\varepsilon}{8}\left(\alpha + \frac{1}{n}\sum_{i=1}^n g(Z_i)\right)\right\} + \mathbb{P}\left\{\exists g \in \mathscr{G} : \frac{1}{n}\sum_{i=1}^n U_i g(Z_i') < -\frac{3\varepsilon}{8}\left(\alpha + \frac{1}{n}\sum_{i=1}^n g(Z_i')\right)\right\} \\
=\ &2\mathbb{P}\left\{\exists g \in \mathscr{G} : \frac{1}{n}\sum_{i=1}^n U_i g(Z_i) > \frac{3\varepsilon}{8}\left(\alpha + \frac{1}{n}\sum_{i=1}^n g(Z_i)\right)\right\}.
\end{aligned}$$

Step 3: Conditioning and introduction of a covering. Let $\delta > 0$ and $\mathscr{G}_\delta$ be an $L_1$ $\delta$-cover of $\mathscr{G}$ on $z^{1:n}$. For $g \in \mathscr{G}$, there exists a $\bar{g} \in \mathscr{G}_\delta$ such that $\frac{1}{n}\sum_{i=1}^n |g(z_i) - \bar{g}(z_i)| < \delta$. Therefore,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n U_i g(z_i) &= \frac{1}{n}\sum_{i=1}^n U_i g(z_i) - \frac{1}{n}\sum_{i=1}^n U_i \bar{g}(z_i) + \frac{1}{n}\sum_{i=1}^n U_i \bar{g}(z_i) \\
&\leq \frac{1}{n}\sum_{i=1}^n U_i \bar{g}(z_i) + \frac{1}{n}\sum_{i=1}^n |g(z_i) - \bar{g}(z_i)| \leq \frac{1}{n}\sum_{i=1}^n U_i \bar{g}(z_i) + \delta.
\end{aligned}$$

On the other hand, $\frac{1}{n}\sum_{i=1}^n g(z_i) \geq \frac{1}{n}\sum_{i=1}^n \bar{g}(z_i) - \frac{1}{n}\sum_{i=1}^n |g(z_i) - \bar{g}(z_i)| \geq \frac{1}{n}\sum_{i=1}^n \bar{g}(z_i) - \delta$. Therefore,

$$\begin{aligned}
&\mathbb{P}\left\{\exists g \in \mathscr{G} : \frac{1}{n}\sum_{i=1}^n U_i g(z_i) > \frac{3\varepsilon}{8}\left(\alpha + \frac{1}{n}\sum_{i=1}^n g(z_i)\right)\right\} \\
\leq\ &\mathbb{P}\left\{\exists g \in \mathscr{G}_\delta : \frac{1}{n}\sum_{i=1}^n U_i g(z_i) + \delta > \frac{3\varepsilon}{8}\left(\alpha + \frac{1}{n}\sum_{i=1}^n g(z_i) - \delta\right)\right\} \\
\leq\ &|\mathscr{G}_\delta| \max_{g \in \mathscr{G}_\delta} \mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n U_i g(z_i) > \frac{3\varepsilon\alpha}{8} - \frac{3\varepsilon\delta}{8} - \delta + \frac{3\varepsilon}{8}\frac{1}{n}\sum_{i=1}^n g(z_i)\right\}
\end{aligned}$$

$$\left(\text{Take } \delta = \frac{\varepsilon\alpha}{5}\right) \le |\mathscr{G}_\delta| \max_{g\in\mathscr{G}_\delta} \mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n U_i g(z_i) > \frac{\varepsilon\alpha}{10} + \frac{3\varepsilon}{8}\frac{1}{n}\sum_{i=1}^n g(z_i)\right\}$$

Step 4: Application of Hoeffding's inequality. □

**Proof** (**Theorem** 6.7 **Lemma** 6.1) Let $Z = (X, Y), Z_i = (X_i, Y_i), g_f(x, y) = |f(x) - y|^2 - |m(x) - y|^2$ where $m(x) = \mathbb{E}[Y|X = x]$. Since $|f(x)| \le B, |Y| \le B$ and $|m(x)| \le B$, $-4B^2 \le g_f(x, y) \le 4B^2$. Therefore, LHS can be written as

$$\mathbb{P}\left\{\exists f \in \mathscr{F} : \mathbb{E}g_f(Z) - \frac{1}{n}\sum_{i=1}^n g_f(Z_i) \ge \varepsilon(\alpha + \beta + \mathbb{E}g_f(Z))\right\}.$$

Step 1: Symmetrization by a ghost sample $Z'_{1:n} \overset{\text{i.i.d.}}{\sim} Z_{1:n}$. Consider a function $f_n \in \mathscr{F}$ depending on $Z_{1:n}$ such that

$$\mathbb{E}[g_{f_n}(Z)|Z_{1:n}] - \frac{1}{n}\sum_{i=1}^n g_{f_n}(Z_i) \ge \varepsilon(\alpha + \beta + \mathbb{E}[g_{f_n}(Z)|Z_{1:n}])$$

if such a function exists in $\mathscr{F}$. Since $\mathrm{Var}(g_{f_n}(Z)|Z_{1:n}) \le 16B^2\mathbb{E}[g_{g_n}(Z)|Z_{1:n}]$,

$$\begin{aligned}
&\mathbb{P}\left\{\mathbb{E}[g_{f_n}(Z)|Z_{1:n}] - \frac{1}{n}\sum_{i=1}^n g_{f_n}(Z'_i) > \frac{\varepsilon}{2}(\alpha+\beta) + \frac{\varepsilon}{2}\mathbb{E}[g_{f_n}(Z)|Z_{1:n}]\middle| Z_{1:n}\right\} \\
\le& \frac{\mathrm{Var}(g_{f_n}(Z)|Z_{1:n})}{n\left(\frac{\varepsilon}{2}(\alpha+\beta) + \frac{\varepsilon}{2}\mathbb{E}[g_{f_n}(Z)|Z_{1:n}]\right)^2} \le \frac{16B^2\mathbb{E}[g_{f_n}(Z)|Z_{1:n}]}{n\left(\frac{\varepsilon}{2}(\alpha+\beta) + \frac{\varepsilon}{2}\mathbb{E}[g_{f_n}(Z)|Z_{1:n}]\right)^2} \\
\le& \frac{16B^2}{\varepsilon^2(\alpha+\beta)n} \le \frac{1}{8}\left(\text{for } n > \frac{128B^2}{\varepsilon^2(\alpha+\beta)}\right).
\end{aligned}$$

That's to say,

$$\mathbb{P}\left\{\mathbb{E}[g_{f_n}(Z)|Z_{1:n}] - \frac{1}{n}\sum_{i=1}^n g_{f_n}(Z'_i) \le \frac{\varepsilon}{2}(\alpha+\beta) + \frac{\varepsilon}{2}\mathbb{E}[g_{f_n}(Z)|Z_{1:n}]\middle| Z_{1:n}\right\} \ge \frac{7}{8}.$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left\{\exists f \in \mathscr{F} : \frac{1}{n}\sum_{i=1}^n g_f(Z'_i) - \frac{1}{n}\sum_{i=1}^n g_f(Z_i) \ge \frac{\varepsilon}{2}(\alpha+\beta) + \frac{\varepsilon}{2}\mathbb{E}g_f(Z)\right\} \\
\ge& \mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n g_{f_n}(Z'_i) - \frac{1}{n}\sum_{i=1}^n g_{f_n}(Z_i) \ge \frac{\varepsilon}{2}(\alpha+\beta) + \frac{\varepsilon}{2}\mathbb{E}[g_{f_n}(Z)|Z_{1:n}]\right\} \\
\ge& \mathbb{P}\Bigg\{\mathbb{E}[g_{f_n}(Z)|Z_{1:n}] - \frac{1}{n}\sum_{i=1}^n g_{f_n}(Z_i) \ge \varepsilon(\alpha+\beta) + \varepsilon\mathbb{E}[g_{f_n}(Z)|Z_{1:n}], \\
&\qquad \mathbb{E}[g_{f_n}(Z)|Z_{1:n}] - \frac{1}{n}\sum_{i=1}^n g_{f_n}(Z'_i) \le \frac{\varepsilon}{2}(\alpha+\beta) + \frac{\varepsilon}{2}\mathbb{E}[g_{f_n}(Z)|Z_{1:n}]\Bigg\} \\
\ge& \frac{7}{8}\mathbb{P}\left\{\mathbb{E}[g_{f_n}(Z)|Z_{1:n}] - \frac{1}{n}\sum_{i=1}^n g_{f_n}(Z_i) \ge \varepsilon(\alpha+\beta) + \varepsilon\mathbb{E}[g_{f_n}(Z)|Z_{1:n}]\right\} \\
=& \frac{7}{8}\mathbb{P}\left\{\exists f \in \mathscr{F} : \mathbb{E}g_f(Z) - \frac{1}{n}\sum_{i=1}^n g_f(Z_i) \ge \varepsilon(\alpha+\beta+\mathbb{E}g_f(Z))\right\}.
\end{aligned}$$

In other words,

$$\text{LHS} \le \frac{8}{7}\mathbb{P}\left\{\exists f \in \mathscr{F} : \frac{1}{n}\sum_{i=1}^n g_f(Z'_i) - \frac{1}{n}\sum_{i=1}^n g_f(Z_i) \ge \frac{\varepsilon}{2}(\alpha+\beta) + \frac{\varepsilon}{2}\mathbb{E}g_f(Z)\right\}.$$

Step 2: Replacement of the expectation by an empirical mean of the ghost sample. That's to say,

$$\begin{aligned}
&\mathbb{P}\left\{\exists f \in \mathscr{F} : \frac{1}{n}\sum_{i=1}^n g_f(Z'_i) - \frac{1}{n}\sum_{i=1}^n g_f(Z_i) \ge \frac{\varepsilon}{2}(\alpha+\beta) + \frac{\varepsilon}{2}\mathbb{E}g_f(Z)\right\} \\
\le& \mathbb{P}\Bigg\{\exists f \in \mathscr{F} : \frac{1}{n}\sum_{i=1}^n g_f(Z'_i) - \frac{1}{n}\sum_{i=1}^n g_f(Z_i) \ge \frac{\varepsilon}{2}(\alpha+\beta) + \frac{\varepsilon}{2}\mathbb{E}g_f(Z),
\end{aligned}$$

$$\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i) - \mathbb{E}g_f^2(Z) \le \varepsilon\left(\alpha+\beta+\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i)+\mathbb{E}g_f^2(Z)\right),$$
$$\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i') - \mathbb{E}g_f^2(Z) \le \varepsilon\left(\alpha+\beta+\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i)+\mathbb{E}g_f^2(Z)\right)\Bigg\}$$
$$+2\mathbb{P}\left\{\exists f\in\mathscr{F}: \frac{\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i)-\mathbb{E}g_f^2(Z)}{\alpha+\beta+\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i)+\mathbb{E}g_f^2(Z)} > \varepsilon\right\}.$$

The second inequality follows from the fact that

$$\mathbb{P}(A) = \mathbb{P}((A\cap B\cap C)\cup\bar{B}\cup\bar{C}) \le \mathbb{P}(A\cap B\cap C)+\mathbb{P}(\bar{B})+\mathbb{P}(\bar{C}).$$

By theorem 6.8,

$$\mathbb{P}\left\{\exists f\in\mathscr{F}: \frac{\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i)-\mathbb{E}g_f^2(Z)}{\alpha+\beta+\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i)+\mathbb{E}g_f^2(Z)} > \varepsilon\right\} \le 4\mathbb{E}\mathscr{N}_1\left(\frac{\alpha+\beta}{5}\varepsilon, \{g_f: f\in\mathscr{F}, Z_{1:n}\}\right)\exp\left(-\frac{\varepsilon^2(\alpha+\beta)n}{40\times 16B^4}\right).$$

Now we consider the first probability on the RHS. The second inequality inside the probability implies

$$(1+\varepsilon)\mathbb{E}g_f^2(Z) \ge (1-\varepsilon)\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i) - \varepsilon(\alpha+\beta)$$
$$\Leftrightarrow \frac{1}{32B^2}\mathbb{E}g_f^2(Z) \ge \frac{1-\varepsilon}{32B^2(1+\varepsilon)}\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i) - \frac{\varepsilon(\alpha+\beta)}{32B^2(1+\varepsilon)}.$$

Since

$$\mathbb{E}g_f(Z) \ge \frac{1}{16B^2}\mathbb{E}g_f^2(Z) = \frac{2}{32B^2}\mathbb{E}g_f^2(Z),$$

the first probability on the RHS can be bounded by

$$\mathbb{P}\Bigg\{\exists f\in\mathscr{F}: \frac{1}{n}\sum_{i=1}^n g_f(Z_i') - \frac{1}{n}\sum_{i=1}^n g_f(Z_i) \ge \frac{\varepsilon}{2}(\alpha+\beta) + \frac{\varepsilon}{2}\Bigg(\frac{1-\varepsilon}{32B^2(1+\varepsilon)}\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i) - \frac{\varepsilon(\alpha+\beta)}{32B^2(1+\varepsilon)}$$
$$+\frac{1-\varepsilon}{32B^2(1+\varepsilon)}\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i') - \frac{\varepsilon(\alpha+\beta)}{32B^2(1+\varepsilon)}\Bigg)\Bigg\}.$$

Step 3: Additional randomization by random signs. The bound is equal to

$$\mathbb{P}\Bigg\{\exists f\in\mathscr{F}: \frac{1}{n}\sum_{i=1}^n U_i[g_f(Z_i') - g_f(Z_i)] \ge \frac{\varepsilon}{2}(\alpha+\beta) + \frac{\varepsilon}{2}\Bigg(\frac{1-\varepsilon}{32B^2(1+\varepsilon)}\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i) - \frac{\varepsilon(\alpha+\beta)}{32B^2(1+\varepsilon)}$$
$$+\frac{1-\varepsilon}{32B^2(1+\varepsilon)}\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i') - \frac{\varepsilon(\alpha+\beta)}{32B^2(1+\varepsilon)}\Bigg)\Bigg\},$$

which is bounded by

$$2\mathbb{P}\left\{\exists f\in\mathscr{F}: \left|\frac{1}{n}\sum_{i=1}^n U_i g_f(Z_i)\right| \ge \frac{\varepsilon}{4}(\alpha+\beta) - \frac{\varepsilon^2(\alpha+\beta)}{64B^2(1+\varepsilon)} + \frac{\varepsilon(1-\varepsilon)}{64B^2(1+\varepsilon)}\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i)\right\}.$$

Step 4: Conditioning and using covering:

$$\mathbb{P}\left\{\exists f\in\mathscr{F}: \left|\frac{1}{n}\sum_{i=1}^n U_i g_f(z_i)\right| \ge \frac{\varepsilon(\alpha+\beta)}{4} - \frac{\varepsilon^2(\alpha+\beta)}{64B^2(1+\varepsilon)} + \frac{\varepsilon(1-\varepsilon)}{64B^2(1+\varepsilon)}\frac{1}{n}\sum_{i=1}^n g_f^2(z_i)\right\}$$
$$\le \mathbb{P}\left\{\exists g\in\mathscr{G}_\delta: \left|\frac{1}{n}\sum_{i=1}^n U_i g(z_i)\right| + \delta \ge \frac{\varepsilon(\alpha+\beta)}{4} - \frac{\varepsilon^2(\alpha+\beta)}{64B^2(1+\varepsilon)} + \frac{\varepsilon(1-\varepsilon)}{64B^2(1+\varepsilon)}\left(\frac{1}{n}\sum_{i=1}^n g_f^2(z_i) - 8B^2\delta\right)\right\}$$
$$\le |\mathscr{G}_\delta|\max_{g\in\mathscr{G}_\delta}\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n U_i g(z_i)\right| \ge \frac{\varepsilon(\alpha+\beta)}{4} - \frac{\varepsilon^2(\alpha+\beta)}{64B^2(1+\varepsilon)} - \delta - \delta\frac{\varepsilon(1-\varepsilon)}{8(1+\varepsilon)} + \frac{\varepsilon(1-\varepsilon)}{64B^2(1+\varepsilon)}\frac{1}{n}\sum_{i=1}^n g_f^2(z_i)\right\}$$
$$\left(\text{Take } \delta = \frac{\varepsilon\beta}{5}\right) \le |\mathscr{G}_\delta|\max_{g\in\mathscr{G}_\delta}\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n U_i g(z_i)\right| \ge \frac{\varepsilon\alpha}{4} - \frac{\varepsilon^2\alpha}{64B^2(1+\varepsilon)} + \frac{\varepsilon(1-\varepsilon)}{64B^2(1+\varepsilon)}\frac{1}{n}\sum_{i=1}^n g^2(z_i)\right\}.$$

Step 5: Application of Bernstein's inequality:

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}U_i g(z_i)\right| \geq \frac{\varepsilon\alpha}{4} - \frac{\varepsilon^2\alpha}{64B^2(1+\varepsilon)} + \frac{\varepsilon(1-\varepsilon)}{64B^2(1+\varepsilon)}\frac{1}{n}\sum_{i=1}^{n}g^2(z_i)\right\} \leq 2\exp\left(-\frac{\varepsilon^2(1-\varepsilon)\alpha n}{140B^2(1+\varepsilon)}\right).$$

Step 6: Bounding the covering number:

$$\mathscr{N}_1\left(\frac{\varepsilon\beta}{5}, \{g_f, f\in\mathscr{F}\}, Z^{1:n}\right) \leq \mathscr{N}_1\left(\frac{\varepsilon\beta}{20B}, \mathscr{F}, X^{1:n}\right).$$

□

# 7 Advanced Techniques from Empirical Process Theory

**Theorem** 7.1 Let $L\in\mathbb{R}_+$ and $\varepsilon_1,\cdots,\varepsilon_n$ be independent random variables with expectation zero and values in $[-L,L]$. Let $z_1,\cdots,z_n\in\mathbb{R}^d, R>0$ and $\mathscr{F}$ be a class of functions $f:\mathbb{R}^d\to\mathbb{R}$ with the property $\|f\|_n^2 = \frac{1}{n}\sum_{i=1}^{n}|f(z_i)|^2 \leq R^2 (\forall f\in\mathscr{F})$. Then $\sqrt{n}\delta \geq 48\sqrt{2}L\int_{\frac{\delta}{8L}}^{\frac{R}{2}}(\log\mathscr{N}_2(u,\mathscr{F},z_{1:n}))^{\frac{1}{2}}\,\mathrm{d}u$ and $\sqrt{n}\delta\geq 36RL$ imply

$$\mathbb{P}\left\{\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}f(z_i)\varepsilon_i\right| > \delta\right\} \leq 5\exp\left(-\frac{n\delta^2}{2304L^2R^2}\right).$$

**Proof** For $R\leq\frac{\delta}{2L}$,

$$\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}f(z_i)\varepsilon_i\right| \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}f(z_i)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2} \leq \sup_{f\in\mathscr{F}}\|f\|_n\sqrt{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2} \leq RL \leq \delta,$$

so WLOG we assume $R>\frac{\delta}{2L}$.

For $s\in\mathbb{N}_+$, let $\{f_1^s,\cdots,f_{N_s}^s\}$ be a $\|\cdot\|_n$-cover of $\mathscr{F}$ of radius $\frac{R}{2^s}$ of size $N_s=\mathscr{N}_2\left(\frac{R}{2^s},\mathscr{F},z_{1:n}\right)$. Set $S=\min\{s\geq 1:\frac{R}{2^s}\leq\frac{\delta}{2L}\}$. Since

$$\begin{aligned}
\left|\frac{1}{n}\sum_{i=1}^{n}f(z_i)\varepsilon_i\right| &= \left|\frac{1}{n}\sum_{i=1}^{n}(f(z_i)-f^S(z_i))\varepsilon_i + \sum_{s=1}^{S}\frac{1}{n}\sum_{i=1}^{n}(f^s(z_i)-f^{s-1}(z_i))\varepsilon_i\right| \\
&\leq \left|\frac{1}{n}\sum_{i=1}^{n}(f(z_i)-f^S(z_i))\varepsilon_i\right| + \sum_{s=1}^{S}\left|\frac{1}{n}\sum_{i=1}^{n}(f^s(z_i)-f^s(z_{i-1}))\varepsilon_i\right| \\
&\leq \|f-f^S\|_n\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2} + \sum_{s=1}^{S}\left|\frac{1}{n}\sum_{i=1}^{n}(f^s(z_i)-f^s(z_{i-1}))\varepsilon_i\right| \\
&\leq \frac{\delta}{2} + \sum_{s=1}^{S}\left|\frac{1}{n}\sum_{i=1}^{n}(f^s(z_i)-f^s(z_{i-1}))\varepsilon_i\right|,
\end{aligned}$$

for $\eta_1,\cdots,\eta_S\geq 0$ satisfy $\eta_1+\cdots+\eta_S\leq 1$, we have

$$\begin{aligned}
\mathbb{P}\left\{\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}f(z_i)\varepsilon_i\right|>\delta\right\} &\leq \mathbb{P}\left\{\exists f\in\mathscr{F}:\frac{\delta}{2}+\sum_{s=1}^{S}\left|\frac{1}{n}\sum_{i=1}^{n}(f^s(z_i)-f^{s-1}(z_i))\varepsilon\right|>\frac{\delta}{2}+\frac{\delta}{2}\sum_{s=1}^{S}\eta_s\right\} \\
&\leq \sum_{s=1}^{S}\mathbb{P}\left\{\exists f\in\mathscr{F}:\left|\frac{1}{n}\sum_{i=1}^{n}(f^s(z_i)-f^{s-1}(z_i))\varepsilon_i\right|>\frac{\delta}{2}\eta_s\right\} \\
&\leq \sum_{s=1}^{S}N_sN_{s-1}\max_{f\in\mathscr{F}}\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}(f^s(z_i)-f^{s-1}(z_i))\varepsilon_i\right|>\frac{\delta}{2}\eta_s\right\}.
\end{aligned}$$

For $s\in\{1,\cdots,S\}$ and $f\in\mathscr{F}$, $(f^s(z_1)-f^{s-1}(z_1))\varepsilon_1,\cdots,(f^s(z_n)-f^{s-1}(z_n)\varepsilon_n)$ are independent, have zero means and take values in $[-L|f^s(z_i)-f^{s-1}(z_i)|, L|f^s(z_i)-f^{s-1}(z_i)|]$. Therefore,

$$\frac{1}{n}\sum_{i=1}^{n}(2L|f^s(z_i)-f^{s-1}(z_i)|)^2 = 4L^2\|f^s-f^{s-1}\|_n^2 \leq 4L^2\left(\|f^s-f\|_n+\|f^{s-1}-f\|_n\right)^2 \leq 4L^2\left(\frac{R}{2^s}+\frac{R}{2^{s-1}}\right)^2 = \frac{36R^2L^2}{2^{2s}}.$$

By Hoeffding's inequality,

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}(f^s(z_i)-f^{s-1}(z_i))\varepsilon_i\right|>\frac{\delta}{2}\eta_s\right\}\le 2\exp\left(-\frac{2n\left(\frac{\eta_s\delta}{2}\right)^2}{36\frac{R^2L^2}{2^{2s}}}\right)$$
$$\Rightarrow\mathbb{P}\left\{\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}f(z_i)\varepsilon_i\right|>\delta\right\}\le\sum_{s=1}^{S}2N_s^2\exp\left(-\frac{n\delta^2\eta_s^2 2^{2s}}{72R^2L^2}\right)=\sum_{s=1}^{S}2\exp\left(2\log N_s-\frac{n\delta^2\eta_s^2 2^{2s}}{72R^2L^2}\right).$$

Choose $\eta_s$ such that

$$2\log N_s\le\frac{1}{2}\frac{n\delta^2\eta_s^2 2^{2s}}{72R^2L^2}\Leftrightarrow\eta_s\ge\bar{\eta}_s:=\frac{12\sqrt{2}RL}{2^s\delta\sqrt{n}}(\log N_s)^{\frac{1}{2}}.$$

More precisely, set

$$\eta_s:=\max\left\{\bar{\eta}_s,\frac{2^{-s}\sqrt{s}}{4}\right\}.$$

Because of

$$\sum_{s=1}^{S}\frac{2^{-s}\sqrt{s}}{4}\le\frac{1}{8}\sum_{s=1}^{+\infty}s\left(\frac{1}{2}\right)^{s-1}=\frac{1}{2}$$

and

$$\sum_{s=1}^{S}\bar{\eta}_s=\sum_{s=1}^{S}\frac{24\sqrt{2}L}{\delta\sqrt{n}}\frac{R}{2^{s+1}}\left\{\log\mathscr{N}_2\left(\frac{R}{2^s},\mathscr{F},z_{1:n}\right)\right\}^{\frac{1}{2}}\le\sum_{s=1}^{S}\frac{24\sqrt{2}L}{\delta\sqrt{n}}\int_{\frac{R}{2^{s+1}}}^{\frac{R}{2^s}}\{\log\mathscr{N}_2(u,\mathscr{F},z_{1:n})\}^{\frac{1}{2}}\,\mathrm{d}u\le\frac{1}{2},$$

we have

$$\sum_{s=1}^{S}\eta_s\le\sum_{s=1}^{S}\frac{2^{-s}\sqrt{s}}{4}+\sum_{s=1}^{S}\bar{\eta}_s\le 1.$$

Therefore,

$$\mathbb{P}\left\{\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}f(z_i)\varepsilon_i\right|>\delta\right\}\le\sum_{s=1}^{S}2\exp\left(-\frac{n\delta^2\eta_s^2 2^{2s}}{144R^2L^2}\right)\le\sum_{s=1}^{S}2\exp\left(-\frac{n\delta^2}{16\cdot 144R^2L^2}\cdot s\right)\le 5\exp\left(-\frac{n\delta^2}{2304L^2R^2}\right).\ \square$$

**Theorem** 7.2 (Extension of **Theorem** 6.8) Let $Z,Z_1,\cdots,Z_n$ be i.i.d. with values in $\mathbb{R}^d$. Let $K\ge 1$ and $\mathscr{F}$ be a class of functions $f:\mathbb{R}^d\to[0,K]$. Let $0<\varepsilon<1$ and $\alpha>0$. Assume that $\sqrt{n}\varepsilon\sqrt{\alpha}\ge 576\sqrt{K}$ and that for all $z_1,\cdots,z_n\in\mathbb{R}^d$ and all $\delta\ge\frac{\alpha K}{2}$, $\frac{\sqrt{n}\varepsilon\delta}{192\sqrt{2}K}\ge\int_{\frac{\varepsilon\delta}{32K}}^{\sqrt{\delta}}\left(\log\mathscr{N}_2\left(u,\left\{f\in\mathscr{F}:\frac{1}{n}\sum_{i=1}^{n}f(z_i)\le\frac{4\delta}{K}\right\},z_{1:n}\right)\right)^{\frac{1}{2}}\mathrm{d}u$. Then

$$\mathbb{P}\left\{\sup_{f\in\mathscr{F}}\frac{|\mathbb{E}[f(Z)]-\frac{1}{n}\sum_{i=1}^{n}f(Z_i)|}{\alpha+\mathbb{E}[f(Z)]+\frac{1}{n}\sum_{i=1}^{n}f(Z_i)}>\varepsilon\right\}\le 15\exp\left(-\frac{n\alpha\varepsilon^2}{128\cdot 2304K}\right).$$

**Proof** Step 1: $Z'_{1:n}=(Z'_1,\cdots,Z'_n)$,

$$\mathbb{P}\left\{\sup_{f\in\mathscr{F}}\frac{|\mathbb{E}[f(Z)]-\frac{1}{n}\sum_{i=1}^{n}f(Z_i)|}{\alpha+\mathbb{E}[f(Z)]+\frac{1}{n}\sum_{i=1}^{n}f(Z_i)}>\varepsilon\right\}$$
$$\le\frac{100}{99}\mathbb{P}\left\{\exists f\in\mathscr{F}:\left|\frac{1}{n}\sum_{i=1}^{n}f(Z_i)-\frac{1}{n}\sum_{i=1}^{n}f(Z'_i)\right|>\frac{\varepsilon}{8}\left(2\alpha+\frac{1}{n}\sum_{i=1}^{n}(Z_i)+\frac{1}{n}\sum_{i=1}^{n}f(Z'_i)\right)\right\}.$$

Step 2: $U_1,\cdots,U_n\overset{\text{i.i.d.}}{\sim}U\{-1,1\}$,

$$\le\frac{100}{99}\cdot 2\mathbb{P}\left\{\exists f\in\mathscr{F}:\left|\frac{1}{n}\sum_{i=1}^{n}U_if(Z_i)\right|>\frac{\varepsilon}{8}\left(\alpha+\frac{1}{n}\sum_{i=1}^{n}f(Z_i)\right)\right\}.$$

Step 3: Peeling,

$$\le\sum_{k=1}^{+\infty}\mathbb{P}\left\{\exists f\in\mathscr{F}:1_{\{k\ne 1\}}2^{k-1}\alpha\le\frac{1}{n}\sum_{i=1}^{n}f(Z_i)<2^k\alpha,\left|\frac{1}{n}\sum_{i=1}^{n}U_if(Z_i)\right|>\frac{\varepsilon}{8}\left(\alpha+\frac{1}{n}\sum_{i=1}^{n}f(Z_i)\right)\right\}$$
$$\le\sum_{k=1}^{+\infty}\mathbb{P}\left\{\exists f\in\mathscr{F}:\frac{1}{n}\sum_{i=1}^{n}f(Z_i)\le 2^k\alpha,\left|\frac{1}{n}\sum_{i=1}^{n}U_if(Z_i)\right|>\frac{\varepsilon}{8}\alpha 2^{k-1}\right\}.$$

Step 4: Application of Theorem 7.1. $R^2=\alpha 2^k K, L=1, \delta=\frac{\varepsilon}{8}\alpha 2^{k-1}$. $\square$

**Theorem** 7.3 Let $Z, Z_1, \cdots, Z_n$ be i.i.d. with values in $\mathbb{R}^d$. Let $K_1, K_2 \geq 1$ and $\mathscr{F}$ be a class of functions $\mathbb{R}^d \to \mathbb{R}$ with properties $|f(z)| \leq K_1 (\forall z \in \mathbb{R}^d)$ and $\mathbb{E}f(Z)^2 \leq K_2 \mathbb{E}f(Z)$. Let $0 < \varepsilon < 1$ and $\alpha > 0$. Assume that $\sqrt{n}\varepsilon\sqrt{1-\varepsilon}\sqrt{\alpha} \geq 288 \max\{2K_1, \sqrt{2K_2}\}$ and that for all $z_1, \cdots, z_n \in \mathbb{R}^d$ and $\delta \geq \frac{\alpha}{8}$,

$$\frac{\sqrt{n}\varepsilon(1-\varepsilon)\delta}{96\sqrt{2}\max\{K_1, 2K_2\}} \geq \int_{\frac{\varepsilon(1-\varepsilon)\delta}{16\max\{K_1,2K_2\}}}^{\sqrt{\delta}} \left(\log \mathscr{N}_2\left(u, \left\{f \in \mathscr{F} : \frac{1}{n}\sum_{i=1}^n f(z_i)^2 \leq 16\delta\right\}, z_1^n\right)\right)^{\frac{1}{2}} \mathrm{d}u.$$

Then

$$\mathbb{P}\left\{\sup_{f\in\mathscr{F}} \frac{|\mathbb{E}f(Z) - \frac{1}{n}\sum_{i=1}^n f(Z_i)|}{\alpha + \mathbb{E}f(Z)} > \varepsilon\right\} \leq 60 \exp\left(-\frac{n\alpha\varepsilon^2(1-\varepsilon)}{128 \cdot 2304 \max\{K_1^2, K_2\}}\right).$$

**Proof** Step 1: Ghost samples $Z'_{1:n} = (Z'_1, \cdots, Z'_n)$,

$$\mathbb{P}\left\{\sup_{f\in\mathscr{F}} \frac{|\mathbb{E}f(Z) - \frac{1}{n}\sum_{i=1}^n f(Z_i)|}{\alpha + \mathbb{E}f(Z)} > \varepsilon\right\} \leq \frac{10}{9}\mathbb{P}\left\{\exists f \in \mathscr{F} : \left|\frac{1}{n}\sum_{i=1}^n f(Z_i) - \frac{1}{n}\sum_{i=1}^n f(Z'_i)\right| > \frac{\varepsilon}{2}\alpha + \frac{\varepsilon}{2}\mathbb{E}f(Z)\right\}.$$

Step 2:

$$\leq \mathbb{P}\left\{\exists f \in \mathscr{F} : \left|\frac{1}{n}\sum_{i=1}^n f(Z_i) - \frac{1}{n}\sum_{i=1}^n f(Z'_i)\right| > \frac{\varepsilon\alpha}{4} + \frac{(1-\varepsilon)\varepsilon}{4(1+\varepsilon)K_2}\left(\frac{1}{n}\sum_{i=1}^n f(Z_i)^2 + \frac{1}{n}\sum_{i=1}^n f(Z'_i)^2\right)\right\}$$
$$+ 2\mathbb{P}\left\{\sup_{f\in\mathscr{F}} \frac{\left|\frac{1}{n}\sum_{i=1}^n f(Z_i)^2 - \mathbb{E}f(Z)^2\right|}{\alpha + \frac{1}{n}\sum_{i=1}^n f(Z_i)^2 + \mathbb{E}f(Z)^2} > \varepsilon\right\}.$$

Step 3:

$$\mathbb{P}\left\{\sup_{f\in\mathscr{F}} \frac{\left|\frac{1}{n}\sum_{i=1}^n f(Z_i)^2 - \mathbb{E}f(Z)^2\right|}{\alpha + \frac{1}{n}\sum_{i=1}^n f(Z_i)^2 + \mathbb{E}f(Z)^2} > \varepsilon\right\} \leq 15 \exp\left(-\frac{n\varepsilon^2\alpha}{128 \cdot 2304 K_1^2}\right).$$

Step 4: $U_1, \cdots, U_n \overset{\text{i.i.d.}}{\sim} U\{-1, 1\}$,

$$\mathbb{P}\left\{\exists f \in \mathscr{F} : \left|\frac{1}{n}\sum_{i=1}^n f(Z_i) - \frac{1}{n}\sum_{i=1}^n f(Z'_i)\right| > \frac{\varepsilon\alpha}{4} + \frac{(1-\varepsilon)\varepsilon}{4(1+\varepsilon)K_2}\left(\frac{1}{n}\sum_{i=1}^n f(Z_i)^2 + \frac{1}{n}\sum_{i=1}^n f(Z'_i)^2\right)\right\}$$
$$\leq 2\mathbb{P}\left\{\exists f \in \mathscr{F} : \left|\frac{1}{n}\sum_{i=1}^n U_i f(Z_i)\right| > \frac{\varepsilon\alpha}{8} + \frac{(1-\varepsilon)\varepsilon}{4(1+\varepsilon)K_2}\frac{1}{n}\sum_{i=1}^n f(Z_i)^2\right\}.$$

Step 5: Peeling,

$$\leq \sum_{k=1}^{+\infty} \mathbb{P}\left\{\exists f \in \mathscr{F} : \frac{1}{n}\sum_{i=1}^n f(Z_i)^2 \leq 2^k \frac{K_2(1+\varepsilon)\alpha}{2(1-\varepsilon)}, \left|\frac{1}{n}\sum_{i=1}^n U_i f(Z_i)\right| > \frac{\varepsilon\alpha}{8}2^{k-1}\right\}.$$

Step 6: Application of Theorem 7.1. □

**Definition** 7.1 (Piecewise polynomial partitioning estimates) $X \in [0,1]$ a.s. and $|Y| \leq L$ a.s.. The piecewise polynomial partitioning estimate is defined by minimizing the empirical $L_2$ risk over the set $\mathscr{F}_{K,M}$ of all piecewise polynomials of degree $M$ (or less) with respect to an equidistant partition of $[0,1]$ into $K$ intervals. More precisely, set

$$m_{n,(K,M)}(\cdot) = \operatorname*{arg\,min}_{f\in\mathscr{F}_{K,M}(L+1)} \frac{1}{n}\sum_{i=1}^n |f(X_i) - Y_i|^2$$

where $\mathscr{F}_{K,M}(L+1) = \{f \in \mathscr{F}_{K,M} : \sup_{x\in[0,1]} |f(x)| \leq L+1\}$.

**Theorem** 7.4 Let $M \in \mathbb{N}, K \in \mathbb{N}_+, x \in [0,1]$ and $L > 0$. Then

$$\mathbb{E}\int |m_{n,(K,M)}(x) - m(x)|^2 \mu(\mathrm{d}x) \leq c_1 \cdot \frac{(M+1)K}{n} + 2 \inf_{f\in\mathscr{F}_{K,M}(L+1)} \int |f(x) - m(x)|^2 \mu(\mathrm{d}x).$$

**Proof** Conduct the following decomposition:

$$\int |m_{n,(K,M)}(x) - m(x)|^2 \mu(\mathrm{d}x) = \mathbb{E}\{|m_{n,(K,M)}(X) - Y|^2 | \mathcal{D}_n\} - \mathbb{E}\{|m(X) - Y|^2\} := T_{1,n} + T_{2,n}$$

where

$$T_{1,n} = 2\frac{1}{n}\sum_{i=1}^{n}\{|m_{n,(K,M)}(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2\},$$
$$T_{2,n} = \mathbb{E}\{|m_{n,(K,M)}(X) - Y|^2|\mathcal{D}_n\} - \mathbb{E}\{|m(X) - Y|^2\} - T_{1,n}.$$

By definition of the estimate,

$$T_{1,n} = 2\min_{f\in\mathscr{F}_{K,M}(L+1)}\frac{1}{n}\sum_{i=1}^{n}\{|f(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2\} \Rightarrow \mathbb{E}T_{1,n} \le 2\inf_{f\in\mathscr{F}_{K,M}(L+1)}\int |f(x) - m(x)|^2\mu(\mathrm{d}x).$$

On the other hand,

$$\mathbb{E}\{T_{2,n}\} \le \int_0^{+\infty}\mathbb{P}\{T_{2,n} > t\}\mathrm{d}t$$

and

$$\mathbb{P}\{T_{2,n} > t\} \le \mathbb{P}\Bigg\{\exists f \in \mathscr{F}_{K,M}(L+1) : 2\mathbb{E}\{|f(X) - Y|^2 - |m(X) - Y|^2\} - \frac{2}{n}\sum_{i=1}^{n}\{|f(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2\}$$
$$> t + \mathbb{E}\{|f(X) - Y|^2 - |m(X) - Y|^2\}\Bigg\}.$$

Then use the following lemma. □

**Lemma** 7.1 Let $(X,Y),(X_1,Y_1),\cdots,(X_n,Y_n)$ i.i.d., $|Y| \le L$ a.s., $L \ge 1$. Let $K \in \mathbb{N}_+$ and $M \in \mathbb{N}$. Then for $\alpha \ge c_3\frac{(M+1)K}{n}$ ($c_3$ depends on $L$),

$$\mathbb{P}\Bigg\{\exists f \in \mathscr{F}_{K,M}(L+1) : \mathbb{E}\{|f(X) - Y|^2 - |m(X) - Y|^2\} - \frac{1}{n}\sum_{i=1}^{n}\{|f(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2\}$$
$$> \frac{1}{2}(\alpha + \mathbb{E}\{|f(X) - Y|^2 - |m(X) - Y|^2\})\Bigg\} \le 60\exp\left(-\frac{n\alpha}{128\cdot 2304\cdot 800\cdot L^4}\right).$$

**Proof** Set $Z = (X,Y), Z_i = (X_i,Y_i)$, define $g_f : \mathbb{R}\times\mathcal{R} \to \mathbb{R}, g_f(x,y) = (|f(x) - y|^2 - |m(x) - y|^2)1_{\{y\in[-L,L]\}}$. Set $\mathscr{G} = \{g_f : f \in \mathscr{F}_{K,M}(L+1)\}$, then the LHS can be written as

$$\mathbb{P}\left\{\exists g \in \mathscr{G} : \frac{\mathbb{E}g(Z) - \frac{1}{n}\sum_{i=1}^{n}g(Z_i)}{\alpha + \mathbb{E}g(Z)} > \frac{1}{2}\right\}.$$

Furthermore,

$$\mathbb{E}\{g_f(Z)^2\} = \mathbb{E}\{(|f(X) - Y|^2 - |m(X) - Y|^2)^2\} \le \mathbb{E}\{|f(X) + m(X) - 2Y| \times |f(X) - Y|\}$$
$$\le (L + 1 + 3L)^2\mathbb{E}\{|f(X) - Y|^2\} \le 25L^2\mathbb{E}\{g_f(Z)\}.$$

By the defintion of piecewise polynomials, $\mathscr{G}$ is a subset of a linear vector space of dimension

$$D = K\cdot((2M+1) + (M+1)) + 1.$$

Therefore,

$$\log\mathscr{N}_2\left(u, \left\{g \in \mathscr{G} : \frac{1}{n}\sum_{i=1}^{n}g(z_i)^2 \le 16\delta\right\}, Z_{1:n}\right) \le D\log\frac{16\sqrt{\delta} + u}{u}.$$

Then use Theorem 7.3. □

**Corollary** 7.1 $C, L > 0$ and $p = k + \beta$ with $k \in \mathbb{N}$ and $\beta \in (0,1]$. Set $M = k$ and $K_n = \lceil C^{\frac{2}{2q+1}}n^{\frac{1}{2q+1}}\rceil$. Then there exists a constant $c_2$ which only depends on $p$ and $L$ such that for all $n \ge \max\{C^{\frac{1}{q}}, C^{-2}\}$,

$$\mathbb{E}\int |m_{n,(K_n,M)}(x) - m(x)|^2\mu(\mathrm{d}x) \le c_2C^{\frac{2}{2q+1}}n^{-\frac{2q}{2q+1}}$$

for every distribution of $(X,Y)$ with $X \in [0,1]$ a.s., $|Y| \le L$ a.s. and $m(x)$ $(q,C)$-smooth.

**Proof** There exists a piecewise polynomial $g \in \mathscr{F}_{K_n,M}$ such that

$$\begin{aligned}
&\sup_{x\in[0,1]} |g(x) - m(x)| \le \frac{1}{2^q k!} \cdot \frac{C}{K_n^q} \\
\Rightarrow &\sup_{x\in[0,1]} |g(x)| \le \sup_{x\in[0,1]} |m(x)| + \frac{1}{2^q k!}\frac{C}{K_n^q} \le L + (Cn^{-q})^{\frac{1}{2q+1}} \le L+1 \\
\Rightarrow &g \in \mathscr{F}_{K_n,M}(L+1) \\
\Rightarrow &\inf_{f\in\mathscr{F}_{K,M}(L+1)} \int |f(x) - m(x)|^2 \mu(\mathrm{d}x) \le \sup_{x\in[0,1]} |g(x) - m(x)|^2 \le \frac{1}{(2^p k!)^2} C^2 \frac{1}{K_n^{2q}}.
\end{aligned}$$

□

# 8 Rademacher Complexity

**Definition** 8.1 (Rademacher complexity) $z_1, \cdots, z_n \in \mathcal{Z}$, $\mathcal{H}$ of functions from $\mathcal{Z} \to \mathbb{R}$. In our context, $z = (x, y)$, $\mathcal{H} = \{(x, y) \mapsto l(y, f(x)) : f \in \mathscr{F}\}, \mathcal{D} = \{z_1, \cdots, z_n\}$. Rademacher complexity of $\mathcal{H}$:

$$\mathscr{R}_n(\mathcal{H}) = \mathbb{E}_{\varepsilon,\mathcal{D}}\left(\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i h(z_i)\right)$$

where $\varepsilon_1, \cdots \varepsilon_n \overset{\text{i.i.d.}}{\sim} U\{-1, 1\}$ are called Rademacher random variables.

**Proposition** 8.1 (Symmetrization)

$$\mathbb{E}\left\{\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^n h(z_i) - \mathbb{E}[h(Z)]\right)\right\} \le 2\mathscr{R}_n(\mathcal{H}), \mathbb{E}\left\{\sup_{h\in\mathcal{H}}\left(\mathbb{E}[h(Z)] - \frac{1}{n}\sum_{i=1}^n h(z_i)\right)\right\} \le 2\mathscr{R}_n(\mathcal{H}).$$

**Proof** Let $\mathcal{D}' = (z_1', \cdots, z_n')$ independent of $\mathcal{D}$. Then

$$\begin{aligned}
\mathbb{E}\left\{\sup_{h\in\mathcal{H}}\left(\mathbb{E}[h(Z)] - \frac{1}{n}\sum_{i=1}^n h(z_i)\right)\right\} &= \mathbb{E}\left\{\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}[h(z_i')\mathcal{D}] - \frac{1}{n}\sum_{i=1}^n h(z_i)\right)\right\} \\
&= \mathbb{E}\left\{\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}[(h(z_i') - h(z_i))|\mathcal{D}]\right)\right\} \\
&\le \mathbb{E}\left\{\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^n [h(z_i') - h(z_i)]\right)\middle|\mathcal{D}\right]\right\} \\
&= \mathbb{E}\left\{\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^n (h(z_i') - h(z_i))\right)\right\} \\
&= \mathbb{E}\left\{\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^n [\varepsilon_i(h(z_i') - h(z_i))]\right)\right\} \le 2\mathscr{R}_n(\mathcal{H}).
\end{aligned}$$

□

**Proposition** 8.2 (Contraction principle) Given any functions $b, a_i : \Theta \to \mathbb{R}$ and $\phi_i : \mathbb{R} \to \mathbb{R}$ any 1-Lipschitz function for $i = 1, \cdots, n$, we have, for $\varepsilon \in \mathbb{R}^n$,

$$\mathbb{E}_\varepsilon\left[\sup_{\theta\in\Theta}\left(b(\theta) + \sum_{i=1}^n \varepsilon_i \phi_i(a_i(\theta))\right)\right] \le \mathbb{E}_\varepsilon\left[\sup_{\theta\in\Theta}\left(b(\theta) + \sum_{i=1}^n \varepsilon_i a_i(\theta)\right)\right].$$

**Proof** Induction on $n$. $n = 0$ is trivial. $n \ge 0$ to $n+1$:

$$\begin{aligned}
\mathbb{E}_{\varepsilon_1,\cdots,\varepsilon_{n+1}}\left[\sup_{\theta\in\Theta}\left(b(\theta) + \sum_{i=1}^{n+1} \varepsilon_i \phi_i(a_i(\theta))\right)\right] &= \frac{1}{2}\mathbb{E}_{\varepsilon_1,\cdots,\varepsilon_n}\left[\sup_{\theta\in\Theta}\left(b(\theta) + \sum_{i=1}^{n} \varepsilon_i \phi_i(a_i(\theta)) + \phi_{n+1}(a_{n+1}(\theta))\right)\right] \\
&\quad + \frac{1}{2}\mathbb{E}_{\varepsilon_1,\cdots,\varepsilon_n}\left[\sup_{\theta\in\Theta}\left(b(\theta) + \sum_{i=1}^{n} \varepsilon_i \phi_i(a_i(\theta)) - \phi_{n+1}(a_{n+1}(\theta))\right)\right] \\
= \mathbb{E}_{\varepsilon_1,\cdots,\varepsilon_n}\left[\sup_{(\theta,\theta')\in\Theta^2}\left(\frac{b(\theta) + b(\theta')}{2} + \sum_{i=1}^n \varepsilon_i \frac{\phi_i(a_i(\theta)) + \phi_i(a_i(\theta'))}{2} + \frac{\phi_{n+1}(a_{n+1}(\theta)) - \phi_{n+1}(a_{n+1}(\theta'))}{2}\right)\right] \\
\le \mathbb{E}_{\varepsilon_1,\cdots,\varepsilon_n}\left[\sup_{(\theta,\theta')\in\Theta^2}\left(\frac{b(\theta) + b(\theta')}{2} + \sum_{i=1}^n \varepsilon_i \frac{\phi_i(a_i(\theta)) + \phi_i(a_i(\theta'))}{2} + \frac{|a_{n+1}(\theta) - a_{n+1}(\theta')|}{2}\right)\right]
\end{aligned}$$

$$= \mathbb{E}_{\varepsilon_1,\cdots,\varepsilon_{n+1}}\left[\sup_{\theta\in\Theta}\left(b(\theta)+\varepsilon_{n+1}a_{n+1}(\theta)+\sum_{i=1}^{n}\varepsilon_i a_i(\theta)\right)\right]. \qquad \square$$

**Example** 8.1 Let $u_i \mapsto l(y_i,u_i)$ be $G$-Lipschitz continuous, $b=0$, $\Theta=\{(f(x_1),\cdots,f(x_n)) : f\in\mathscr{F}\}\subset\mathbb{R}^n$, $a_i(\theta)=\theta_i, \phi_i(u_i)=l(y_i,u_i)$. Then

$$\mathbb{E}_\varepsilon\left[\sup_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i l(y,f(x_i))\middle|\mathcal{D}\right]\le G\mathbb{E}_\varepsilon\left[\sup_{f\in\mathscr{F}}\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(x_i)\right)\middle|\mathcal{D}\right]\Rightarrow \mathscr{R}_n(\mathcal{H})\le G\mathscr{R}_n(\mathcal{F}).$$

**Proposition** 8.3 (Abolute contraction principle) $\phi_i:\mathbb{R}\to\mathbb{R}$ 1-Lipschitz continuous, $\phi_i(0)=0$, then

$$\mathbb{E}_\varepsilon\left[\sup_{\theta\in\Theta}\left|\sum_{i=1}^{n}\varepsilon_i\phi_i(a_i(\theta))\right|\right]\le 2\mathbb{E}_\varepsilon\left[\sup_{\theta\in\Theta}\left|\sum_{i=1}^{n}\varepsilon_i a_i(\theta)\right|\right].$$

**Definition** 8.2 A function $\phi:\mathbb{R}\to\mathbb{R}$ is contraction vanishing at 0 if it satisfies $|\phi(s)-\phi(t)|\le|s-t|$ for all $s,t\in\mathbb{R}$ and $\phi(0)=0$.

**Theorem** 8.1 (Contraction principle for Rademacher processes) Let $F$ be a nonnegative, convex and nondecreasing function defined on $[0,\infty)$. Let $\phi_i:\mathbb{R}\to\mathbb{R}$ contraction vanishing at 0, and let $T$ be a bounded set of $\mathbb{R}^n, n<\infty$. Then

$$\mathbb{E}F\left(\frac{1}{2}\|\sum_{i=1}^{n}\varepsilon_i\phi_i(t_i)\|_T\right)\le\mathbb{E}F\left(\|\sum_{i=1}^{n}\varepsilon_i t_i\|_T\right)$$

where $t=(t_1,\cdots,t_n)\in T$ and $\|X\|_T:=\sup_{t\in T}|X(t)|$.

**Example** 8.2 Let $X_1,\cdots,X_n\in D\subset\mathbb{R}^d$ and $\mathscr{F}$ be a countable class of measurable functions with $F(x)=\sup_{f\in\mathscr{F}}|f(x)|$ is finite for all $x\in D$. Set $U=\max_{i=1,\cdots,n}|F(X_i)|$ and $\sigma^2=\sup_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}f^2(X_i)<\infty$. For $X_1,\cdots,X_n$ fixed, let $t_i=Uf(X_i), i=1,\cdots,n$, $T=\{(Uf(X_i))_{i=1,\cdots,n}, f\in\mathscr{F}\}$ and $\phi_i(s)\equiv\phi(s)=\frac{s^2}{2U^2}\wedge\frac{U^2}{2}$. Then by Theorem 8.1,

$$\frac{1}{4}\mathbb{E}_\varepsilon\|\sum_{i=1}^{n}\varepsilon_i f^2(X_i)\|_{\mathscr{F}}\le U\mathbb{E}_\varepsilon\|\sum_{i=1}^{n}\varepsilon_i f(X_i)\|_{\mathscr{F}}$$
$$\Rightarrow \mathbb{E}_X\mathbb{E}_\varepsilon\|\sum_{i=1}^{n}\varepsilon_i f^2(X_i)\|_{\mathscr{F}}\le 4\mathbb{E}_X U\mathbb{E}_\varepsilon\|\sum_{i=1}^{n}\varepsilon_i f(X_i)\|_{\mathscr{F}}.$$

# 9 Optimization for Machine Learning

**Definition** 9.1 $(x_i,y_i)$ i.i.d., $f:\mathcal{X}\to\mathbb{R}$, $\mathcal{R}(f)=\mathbb{E}[l(y,f(x))]$, objective function $F(\theta)=\frac{1}{n}\sum_{i=1}^{n}l(y_i,f_\theta(x_i))+\lambda_n\Omega(\theta)$ where $\Omega(\theta)$ is regularization, $\theta^*\in\arg\min_\theta\mathcal{R}(f_\theta)$, $\eta^*\in\arg\min_\theta\hat{\mathcal{R}}(f_\theta)$.

**Definition** 9.2 (Gradient descent (GD)) Pick $\theta_0\in\mathbb{R}^d$ and for $t\ge1$, $\theta_t=\theta_{t-1}-\gamma_t F'(\theta_{t-1})$.

**Example** 9.1 (Ordinary least-squares) $F(\theta)=\frac{1}{2n}\|\Phi\theta-y\|_2^2, F'(\theta)=\frac{1}{n}\Phi^T(\Phi\theta-y), H=\frac{1}{n}\Phi^T\Phi\in\mathbb{R}^{d\times d}$ Hessian matrix. $F'(\eta^*)=0\Rightarrow H\eta^*=\frac{1}{n}\Phi^T y, F(\theta)-F(\eta^*)=F'(\eta^*)^T(\theta-\eta^*)+\frac{1}{2}(\theta-\eta^*)^T H(\theta-\eta^*)$. Condition number $\kappa=\frac{L}{\mu}\ge1$ where $\mu/L$ is the smallest/largest eigenvalue of $H$, respectively. Then

$$\theta_t-\eta^*=\theta_{t-1}-\gamma\left(\frac{1}{n}\Phi^T(\Phi\theta_{t-1}-y)\right)-\eta^*=\theta_{t-1}-\gamma(H\theta_{t-1}-H\eta^*)-\eta^*$$
$$=(I-\gamma H)(\theta_{t-1}-\eta^*)=(I-\gamma H)^t(\theta_0-\eta^*)$$
$$\Rightarrow\|\theta_t-\eta^*\|^2=(\theta_0-\eta^*)(I-\gamma H)^{2t}(\theta_0-\eta^*), F(\theta_t)-F(\eta^*)=\frac{1}{2}(\theta_0-\eta^*)^T(I-\gamma H)^{2t}H(\theta_0-\eta^*).$$

To derive the fastest convergence, we need to consider $\min_\gamma\left(\max_{\lambda\in[\mu,L]}|1-\gamma\lambda|\right)^{2t}\Rightarrow\gamma^*=\frac{2}{\mu+L}, \text{rate}^*=1-\frac{2}{\kappa+1}$. To ensure convergence, we need to bound $\max\{|1-\gamma L|,|1-\gamma\mu|\}<1\Leftrightarrow\gamma<\frac{2}{L}$. If we take $\gamma=\frac{1}{L}$ (i.e. independent of $\mu$), $\|\theta_t-\eta^*\|^2\le\left(1-\frac{1}{\kappa}\right)^{2t}\|\theta_0-\eta^*\|^2, F(\theta_t)-F(\eta^*)\le\left(1-\frac{1}{\kappa}\right)^{2t}[F(\theta_0)-F(\eta^*)]\le\exp\left(-\frac{2t}{\kappa}\right)[F(\theta_0)-F(\eta^*)]$.

**Definition** 9.3 (Convex functions) Convex: $\forall\alpha\in(0,1), F(\alpha\theta+(1-\alpha)\eta)\le\alpha F(\theta)+(1-\alpha)F(\eta)$. Strictly convex: replace "$\le$" with "$<$". $\mu$-strongly(uniformly) convex: $\alpha F(\theta)+(1-\alpha)F(\eta)-F(\alpha\theta+(1-\alpha)\eta)\ge\mu\alpha(1-\alpha)\|\theta-\eta\|^2\Leftrightarrow F(\theta)-\frac{\mu}{2}\|\theta\|^2$ convex $\overset{\text{if } F \text{ differentiable}}{\Leftrightarrow} F(\theta)-F(\eta)\ge F'(\eta)(\theta-\eta)+\frac{\mu}{2}\|\theta-\eta\|^2\Leftrightarrow\eta^T F''(\theta)\eta\ge\mu\|\eta\|^2$.

**Property** 9.1 (Lojasiewicz inequality) If $F$ is differentiable and $\mu$-strongly convex with unique minimizer $\eta^*$, then we have $\|F'(\theta)\|_2^2 \geq 2\mu(F(\theta) - F(\eta^*)), \forall \theta \in \mathbb{R}^d$.

**Definition** 9.4 (Smoothness) A differentiable $F$ is said $L$-smooth iff

$$|F(\eta) - F(\theta) - F'(\theta)^T(\eta - \theta)| \leq \frac{L}{2}\|\theta - \eta\|^2, \forall \theta, \eta \in \mathbb{R}^d \Leftrightarrow \|F'(\theta) - F'(\eta)\|_2 \leq L\|\theta - \eta\|.$$

**Remark** 9.1 If $F$ is $L$-smooth and $\mu$-strongly convex, then $\mu \leq L$ and we can define the condition number as $\kappa = \frac{L}{\mu}$.

**Proposition** 9.1 (Convergence of GD for smooth strongly-convex functions) Let $\gamma_t = \frac{1}{L}$, then

$$F(\theta_t) - F(\eta^*) \leq \left(1 - \frac{1}{\kappa}\right)^t (F(\theta_0) - F(\eta^*)) \leq \exp\left(-\frac{t}{\kappa}\right)(F(\theta_0) - F(\eta^*)).$$

**Proposition** 9.2 (Convergence of GD for smooth convex functions) Let $\gamma_t = \frac{1}{L}$, then

$$F(\theta_t) - F(\eta^*) \leq \frac{L}{2t}\|\theta_0 - \eta^*\|_2^2.$$

**Proof** Define $V_t(\theta_t) = t[F(\theta_t) - F(\eta^*)] + \frac{L}{2}\|\theta_t - \eta^*\|^2$ as the Lyapunov function. Then

$$V_t(\theta_t) - V_{t-1}(\theta_{t-1}) = t[F(\theta_t) - F(\theta_{t-1})] + F(\theta_{t-1}) - F(\eta^*) + \frac{L}{2}\|\theta_t - \eta^*\|^2 - \frac{L}{2}\|\theta_{t-1} - \eta^*\|^2.$$

Since (1) $F(\theta_t) - F(\theta_{t-1}) \leq -\frac{1}{2L}\|F'(\theta_{t-1})\|_2^2$; (2) $F(\theta_{t-1}) - F(\eta^*) \leq F'(\theta_{t-1})^T(\theta_{t-1} - \eta^*)$; (3) $\frac{L}{2}\|\theta_t - \eta^*\|^2 - \frac{L}{2}\|\theta_{t-1} - \eta^*\|^2 = -L\gamma(\theta_{t-1} - \eta^*)^T F'(\theta_{t-1}) + \frac{L\gamma^2}{2}\|F'(\theta_{t-1})\|^2$, we have

$$\begin{aligned} V_t(\theta_t) - V_{t-1}(\theta_{t-1}) &\leq t\left[-\frac{1}{2L}\|F'(\theta_{t-1})\|_2^2\right] + F'(\theta_{t-1})^T(\theta_{t-1} - \eta^*) - L\gamma(\theta_{t-1} - \eta^*)^T + \frac{L\gamma^2}{2}\|F'(\theta_{t-1})\|^2 \\ &= -\frac{t-1}{2L}\|F'(\theta_{t-1})\|_2^2 \leq 0 \\ \Rightarrow t[F(\theta_t) - F(\eta^*)] &\leq V_t(\theta_t) \leq V_0(\theta_0) = \frac{L}{2}\|\theta_0 - \eta^*\|_2^2. \end{aligned}$$

□

**Definition** 9.5 (Nesterov acceleration)

$$\theta_{t+\frac{1}{2}} = \theta_t - \frac{1}{L}F'(\theta_t), \theta_{t+1} = \theta_{t+\frac{1}{2}} + \frac{1 - \sqrt{\frac{\mu}{L}}}{1 + \sqrt{\frac{\mu}{L}}}(\theta_{t+\frac{1}{2}} - \theta_t) = \left(1 + \frac{1 - \sqrt{\frac{\mu}{L}}}{1 + \sqrt{\frac{\mu}{L}}}\right)\theta_{t+\frac{1}{2}} - \left(\frac{1 - \sqrt{\frac{\mu}{L}}}{1 + \sqrt{\frac{\mu}{L}}}\right)\theta_t.$$

**Proposition** 9.3 (Convergence of Nesterov acceleration for smooth convex functions)

$$F(\theta_t) - F(\eta^*) \leq \frac{2L\|\theta_0 - \eta^*\|^2}{(t+1)^2}.$$

**Definition** 9.6 (Proximal GD) $F = G + H$ while $G$ is smooth and $H$ is non-smooth. Define

$$\theta_t = \arg\max_\theta G(\theta_{t+1}) + (\theta - \theta_{t-1})^T G'(\theta_{t-1}) + \frac{L}{2}\|\theta - \theta_{t-1}\|_2^2 + H(\theta).$$

**Definition** 9.7 (Subgradients) $\partial F(\theta) = \{z \in \mathbb{R}^d : \forall \eta \in \mathbb{R}^d, F(\eta) \geq F(\theta) + z^T(\eta - \theta)\}$. For a convex function defined on $\mathbb{R}^d$, the subdifferential is a non-empty. If $F$ is differentiable, $\partial F(\theta) = \{F'(\theta)\}$.

**Proposition** 9.4 $F$ is convex, $B$-Lipschitz continuous and admits a minimizer $\eta^*$ that satisfies $\|\eta^* - \theta_0\|_2 \leq D$. By setting $\gamma_t = \frac{D}{B\sqrt{t}}$, the iterates $(\theta_n)_{n \geq 0}$ of GD on $F$ satisfy

$$\min_{0 \leq s \leq t-1} F(\theta_s) - F(\eta^*) \leq DB\frac{2 + \log(t)}{2\sqrt{t}}.$$

**Definition** 9.8 (Stochastic gradient descent (SGD)) Assume objective function $F(\theta) = \frac{1}{n}\sum_{i=1}^n l(y_i, f_\theta(x_i)) + \Omega(\theta)$. Let

$$\theta_t = \theta_{t-1} - \gamma_t g_t(\theta_{t-1})$$

where $\mathbb{E}[g_t(\theta_{t-1})|\theta_{t-1}] = F'(\theta_{t-1})$ (and $\|g_t(\theta_{t-1})\|_2^2 \leq B^2$ a.s.) for all $t \geq 1$.

**Proposition** 9.5 (Convergence of SGD) $F$ is convex, $B$-Lipschitz and admits a minimizer $\eta^*$ that satisfies $\|\eta^* - \theta_0\| \leq D$. Set $\gamma_t = \frac{D}{B\sqrt{t}}$. Then

$$\mathbb{E}[F(\bar{\theta}_t) - F(\eta^*)] \leq DB\frac{2+\log(t)}{2\sqrt{t}}$$

where $\bar{\theta}_t = \frac{\sum_{s=1}^t \gamma_s \theta_{s-1}}{\sum_{s=1}^t \gamma_s}$.

**Proposition** 9.6 (Convergence of SGD for strongly-convex problems) $G(\theta) = F(\theta) + \frac{\mu}{2}\|\theta\|^2, \theta_t = \theta_{t-1} - \gamma_t[g_t(\theta_{t-1}) + \mu\theta_{t-1}], \gamma_t = \frac{1}{\mu t}$, then

$$\mathbb{E}[G(\bar{\theta}_t) - G(\eta^*)] \leq \frac{2B^2(1+\log t)}{\mu t}$$

where $\bar{\theta}_t = \frac{1}{t}\sum_{s=1}^t \theta_{s-1}$.

**Definition** 9.9 (Variance reduction: SAGA) Consider a finite sum $F(\theta) = \frac{1}{n}\sum_{i=1}^n f_i(\theta)$ where each $f_i$ is $R^2$-smooth and $F$ is $\mu$-strongly convex. Let

$$\theta_t = \theta_{t-1} - \gamma\left[f'_{i(t)}(\theta_{t-1}) + \frac{1}{n}\sum_{i=1}^n z_i^{(t-1)} - z_{i(t)}^{(t-1)}\right]$$

where $i(t)$ is selected uniformly at random in $\{1, \cdots, n\}$ and $z_{i(t)}^{(t)} = f'_{i(t)}(\theta_{t-1})$.

**Proposition** 9.7 $z_i^{(0)} := f'_i(\theta_0)$ for all $i \in \{1, \cdots, n\}, \gamma = \frac{1}{4R^2}$, then

$$\mathbb{E}[\|\theta_t - \eta^*\|_2^2] \leq \left(1 - \min\left\{\frac{1}{3n}, \frac{3\mu}{16R^2}\right\}\right)^t \left(1 + \frac{n}{4}\right)\|\theta_0 - \eta^*\|_2^2.$$

# 10 From Online Learning to Bandits

**Definition** 10.1 $F'_t(\theta_{t-1}) = \frac{\partial l(y_t, f_\theta(x_t))}{\partial \theta}|_{\theta=\theta_{t-1}}$. Performance measure is $\mathbb{E}[F(\theta_t)] - F^*$ where $F(\theta) = \mathbb{E}[l(y, f_\theta(x))]$ and $F^* = \inf_{\theta\in\mathcal{C}} F(\theta)$. Regret: $\frac{1}{t}\sum_{s=1}^t F(\theta_{s-1}) - \inf_{\theta\in\mathcal{C}} F(\theta)$. Adversarial: $\frac{1}{t}\sum_{s=1}^t F_s(\theta_{s-1}) - \inf_{\theta\in\mathcal{C}} \frac{1}{t}\sum_{s=1}^t F_s(\theta)$.

**Proposition** 10.1 (First-order online convex optimization) $F_s : \mathbb{R}^d \to \mathbb{R}$, compact set $\mathcal{C}$, $\theta_0 \in \mathcal{C}$. Let $\frac{1}{t}\sum_{s=1}^t F_s(\theta_{s-1}) - \inf_{\theta\in\mathcal{C}} \frac{1}{t}\sum_{s=1}^t F_s(\theta)$ be small as possible. Unbiased vergion $g_s$, $\mathscr{F}_s$ denotes the information up to (and including) time $s$. Assume (1) $\mathbb{E}[g_s|\mathscr{F}_{s-1}] = F'_s(\theta_{s-1})$; (2) $\|g_s\|_2^2 \leq B^2$ a.s.. Projected SGD: $\theta_s = \prod_{\mathcal{C}}(\theta_{s-1} - \gamma_s g_s)$. Then for any $\theta \in \mathcal{C}$,

$$\begin{aligned}
\|\theta_s - \theta\|_2^2 &\leq \|\theta_{s-1} - \theta\|_2^2 - 2\gamma_s g_s^T(\theta_{s-1} - \theta) + \gamma_s^2 B^2 \text{ by contractivity of projections,}\\
\Rightarrow \mathbb{E}[\|\theta_s - \theta\|_2^2|\mathscr{F}_{s-1}] &\leq \|\theta_{s-1} - \theta\|_2^2 - 2\gamma_s F'_s(\theta_{s-1})^T(\theta_{s-1} - \theta) + \gamma_s^2 B^2 \text{ by the unbiasedness}\\
&\leq \|\theta_{s-1} - \theta\|_2^2 - 2\gamma_s[F_s(\theta_{s-1}) - F_s(\theta)] + \gamma_s^2 B^2 \text{ by the convexity}\\
\Rightarrow \mathbb{E}[F_s(\theta_{s-1}) - F_s(\theta)] &\leq \frac{1}{2\gamma_s}\left(\mathbb{E}[\|\theta_{s-1} - \theta\|_2^2] - \mathbb{E}[\|\theta_s - \theta\|_2^2]\right) + \frac{\gamma_s}{2}B^2 \text{ by taking full expectations}\\
\Rightarrow \frac{1}{t}\sum_{s=1}^t \mathbb{E}[F_s(\theta_{s-1})] - \frac{1}{t}\sum_{s=1}^t F_s(\theta) &\leq \frac{1}{t}\sum_{s=1}^t \frac{1}{2\gamma_s}\left(\mathbb{E}[\|\theta_{s-1} - \theta\|_2^2] - \mathbb{E}[\|\theta_s - \theta\|_2^2]\right) + \frac{1}{t}\sum_{s=1}^t \frac{\gamma_s}{2}B^2\\
\Rightarrow \frac{1}{t}\sum_{s=1}^t \mathbb{E}[F_s(\theta_{s-1})] - \frac{1}{t}\sum_{s=1}^t F_s(\theta) &\leq \frac{1}{t}\sum_{s=1}^t \frac{1}{2\gamma_s}(\delta_{s-1} - \delta_s) + \frac{1}{t}\sum_{s=1}^t \frac{\gamma_s}{2}B^2 \text{ by letting } \delta_s = \mathbb{E}[\|\theta_s - \theta\|_2^2]\\
&= \frac{1}{t}\sum_{s=1}^{t-1}\delta_s\left(\frac{1}{2\gamma_{s+1}} - \frac{1}{2\gamma_s}\right) + \frac{\delta_0}{2t\gamma_1} - \frac{\delta_t}{2t\gamma_t} + \frac{1}{t}\sum_{s=1}^t \frac{\gamma_s}{2}B^2 \text{ by using Abel's formula}\\
&\leq \frac{1}{t}\sum_{s=1}^{t-1}\mathrm{diam}(\mathcal{C})^2\left(\frac{1}{2\gamma_{s+1}} - \frac{1}{2\gamma_s}\right) + \frac{\mathrm{diam}(\mathcal{C})^2}{2t\gamma_1} + \frac{1}{t}\sum_{s=1}^t \frac{\gamma_s}{2}B^2\\
&= \frac{\mathrm{diam}(\mathcal{C})^2}{2t\gamma_t} + \frac{1}{t}\sum_{s=1}^t \frac{\gamma_s}{2}B^2.
\end{aligned}$$

By choosing $\gamma_s = \frac{\mathrm{diam}(\mathcal{C})}{B\sqrt{s}}$, we have

$$\frac{1}{t}\sum_{s=1}^t \mathbb{E}[F_s(\theta_{s-1})] - \frac{1}{t}\sum_{s=1}^t F_s(\theta) \leq \frac{3B\mathrm{diam}(\mathcal{C})}{2\sqrt{t}}.$$

In the strongly-convex case, we can replace $[F_s(\theta_{s-1})-F_s(\theta)]$ in the third row with $[F_s(\theta_{s-1})-F_s(\theta)+\frac{\mu}{2}\|\theta_{s-1}-\theta\|_2^2]$ and $\frac{1}{2\gamma_s}(\mathbb{E}[\|\theta_{s-1}-\theta\|_2^2]-\mathbb{E}[\|\theta_s-\theta\|_2^2])$ in the fourth row with $\left(\frac{1}{2\gamma_s}-\frac{\mu}{2}\right)\mathbb{E}[\|\theta_{s-1}-\theta\|_2^2]-\frac{1}{2\gamma_s}\mathbb{E}[\|\theta_s-\theta\|_2^2]$. By summing between $s=1$ to $s=t$ and choosing $\gamma_s=\frac{1}{\mu s}$, we obtain

$$\frac{1}{t}\sum_{s=1}^{t}\mathbb{E}[F_s(\theta_{s-1})]-\frac{1}{t}\sum_{s=1}^{t}F_s(\theta)\le\frac{1}{t}\sum_{s=1}^{t}\frac{1}{2\mu s}B^2\le\frac{1}{2\mu t}(1+\log t)B^2.$$

**Definition** 10.2 (Mirror map) A differentiable and $\mu$-strongly convex function $\Phi:\mathcal{C}_\Phi\to\mathbb{R}$ w.r.t. a norm $\|\cdot\|$, that is,

$$\Phi(\eta)\ge\Phi(\theta)+\Phi'(\theta)^T(\eta-\theta)+\frac{\mu}{2}\|\eta-\theta\|^2,\forall\eta,\theta\in\mathcal{C}.$$

**Proposition** 10.2 (Online mirror descent) Consider the same setup of Proposition 10.1. We have Lipschitz-continuous functions $F_s$ for $s\ge1$, $\mathbb{E}[g_s|\mathscr{F}_{s-1}]=F_s'(\theta_{s-1})$ and

$$\theta_t:=\arg\min_{\theta\in\mathcal{C}}\left[g_t^T(\theta-\theta_{t-1})+\frac{1}{\gamma}D_\Phi(\theta,\theta_{t-1})\right]$$

where $D_\Phi(\theta,\eta)=\Phi(\theta)-\Phi(\eta)-\Phi'(\eta)^T(\theta-\eta)$ is the Bregman divergence. Assume $\mathbb{E}[\|g_s\|^2|\mathscr{F}_{s-1}]\le B$ for all $s\ge1$. Then for every $\theta\in\mathcal{C}$, we have

$$\frac{1}{t}\sum_{s=1}^{t}\mathbb{E}[F_s(\theta_{s-1})-F_s(\theta)]\le\frac{1}{\gamma t}D_\Phi(\theta,\theta_0)+\frac{B^2\gamma}{2\mu}.$$

**Definition** 10.3 (Zero-th order convex optimization) We consider the task of unconstrained minimization of a convex function $F$, given only access to function values. Finite difference:

$$\hat{F}'(\theta)=\sum_{i=1}^{d}\frac{1}{\delta}[F(\theta++\delta e_i)-F(\theta)]e_i.$$

Assume no noise and the function is $L$-smooth. Then

$$\|\hat{F}'(\theta)-F'(\theta)\|_2^2=\frac{1}{\delta^2}\sum_{i=1}^{d}[F(\theta+\delta e_i)-F(\theta)-F'(\theta)\delta e_i]^2\le\frac{d}{\delta^2}\left(\frac{L\delta^2}{2}\right)^2=\frac{dL^2\delta^2}{4}.$$

Now consider the general noise case. Then

$$\theta_t=\theta_{t-1}-\gamma\left[\frac{1}{\delta}\left(F(\theta_{t-1}+\delta z_t)+\zeta_t-F(\theta_{t-1})-\zeta_t'\right)z_t\right].$$

Write $\varepsilon_t=\zeta_t-\zeta_t'\sim(0,2\sigma^2)$,

$$\theta_t=\theta_{t-1}-\gamma\left[\frac{1}{\delta}\left(F(\theta_{t-1}+\delta z_t)-F(\theta_{t-1})+\varepsilon_t\right)z_t\right]$$

where $\mathbb{E}z_t=0,\mathbb{E}(z_tz_t^T)=I$. Two choices of $z$: (1) a signed canonical basis vectors: $\pm\sqrt{d}e_i$ with $i$ selected uniformly at random in $\{1,\cdots,d\}$; (2) a standard Gaussian vector.

The key in analyzing the iteration is to study

$$\begin{aligned}g&=\frac{1}{\delta}\left(F(\theta+\delta z)-F(\theta)\right)z=\frac{1}{\delta}\left(\delta z^TF'(\theta)+O(\delta^2)\right)z\\&=z^TF'(\theta)z+O(\delta)=zz^TF'(\theta)+O(\delta)\\\Rightarrow\mathbb{E}[g]&=F'(\theta)+O(\delta).\end{aligned}$$

**Proposition** 10.3 (Smooth stochastic GD) $F$ is $L$-smooth, $F_\delta(\theta):=\mathbb{E}_{z\sim\mathcal{N}(0,I)}F(\theta+\delta z)$. Then

$$0\le F_\delta(\theta)-F(\theta)=\mathbb{E}_{z\sim\mathcal{N}(0,I)}[F(\theta+\delta z)-F(\theta)-\delta F'(\theta)z]\le\frac{L\delta^2}{2}\mathbb{E}_{z\sim\mathcal{N}(0,I)}\|z\|_2^2=\frac{L\delta^2d}{2}.$$

When $F$ is not $L$-smooth but $B$-Lipschitz continuous, the function $F_\delta$ is still $B$-Lipschitz continuous. Moreover, $F_\delta$ is $(\frac{\sqrt{d}}{\delta}B)$-smooth with gradient equal to

$$F_\delta'(\theta)=\frac{1}{\delta}\mathbb{E}_{z\sim\mathcal{N}(0,I)}(F(\theta+\delta z)-F(\theta))z.$$

Also, $|F_\delta(\theta)-F(\theta)|\le B\delta\sqrt{d}$.