Journal of Applied Analysis and Computation
Volume 12, Number 5, October 2022, 2110–2132
Website:http://www.jaac-online.com
DOI:10.11948/20220222

LARGE MARGIN UNIFIED MACHINES WITH NON-I.I.D. PROCESS*

Amina Benabid[1], Dan Su[2] and Dao-Hong Xiang[1,†]

Abstract In this paper, we investigate the convergence theory of large margin unified machines (LUMs) in a non-i.i.d. sampling. We decompose the total error into sample error, regularization error and drift error. The appearance of drift error is caused by the non-identical sampling. Independent blocks sequences are constructed to transform the analysis of the dependent sample sequences into the analysis of independent blocks sequences under some mixing conditions. We also require the assumption of polynomial convergence of the marginal distributions to deal with the non-identical sampling. A novel projection operator is introduced to overcome the technical difficulty caused by the unbounded target function. The learning rates are explicitly derived under some mild conditions on approximation and capacity of the reproducing kernel Hilbert space.

Keywords Large margin unified machines, β-mixing sequence, projection operator, reproducing kernel Hilbert spaces.

MSC(2010) 68Q32, 41A46.

1. Introduction

Many machine learning models suppose that the data are independent and identically distributed (i.i.d.). However, this ideal hypothesis is not always satisfied in the real cases. For example, financial predictions, signal processing, system observation and diagnosis, and speech or text recognition.

There is an extensive literature on investigating learning problems with the case of non-independent process or non-identical distributed process, or both cases. For example, Yu in [32] extended the classical empirical process theory for Vapnik-Cervonenkis classes which deals mainly with sequence of independent random variables to dependent cases, which inspired subsequent research works such as [15,18,20,23,25,31]. Smale and Zhou in [19] studied learning performance of online algorithm with independent but non-identically distributed data. Guo and Shi in [13] investigated learning algorithms for binary classification problems in a non-i.i.d process. Guo and Ye [10] studied the l^q-regularized regression learning algorithm with the non-identical and dependent samples.

[†] The corresponding author. Email: daohongxiang@zjnu.cn(D. Xiang)
[1] Department of Mathematics, Zhejiang Normal University, No.688 Yingbin Road, 321004 Jinhua, Zhejiang, China
[2] Department of Basic Sciences, Yangzhou Polytechnic Institute Laboratory, No.199 Huayang Road, 225127 Yangzhou, Jiangsu , China

[*] The authors is partially supported by the National Natural Science Foundation of China(Nos. 11871438 and U20A2068).

In this paper, we aim at studying the binary classification algorithms associated with large margin unified machines (LUMs) loss function with the non-independent and non-identically distributed sampling. Denote by X the input space which is a compact subset of $\mathbb{R}^d$ and the output space $Y = \{-1, 1\}$ represents two classes. Let P be an unknown probability measure defined on $Z := X \times Y$. The goal of binary classification is to learn a classifier $\mathcal{C} : X \to Y$ by minimizing the following misclassification error:

$$\mathcal{R}(\mathcal{C}) = \text{Prob}\{\mathcal{C}(x) \neq y\} = \int_X P(y \neq \mathcal{C}(x)|x) dP_X.$$

Here $P(y|x)$ is the conditional distribution at a given $x \in X$ and P_X is the marginal distribution of P on X. It is usual to consider the classifiers $\mathcal{C}_f = \text{sgn}(f)$ induced by real-valued functions $f : X \to \mathbb{R}$, where $\text{sgn}(f)(x) = 1$ if $f(x) \geq 0$ and $\text{sgn}(f)(x) = -1$ otherwise.

The Bayes rule f_c defined below is the ideal classifier which minimizes the misclassification error:

$$f_c(x) = \begin{cases} 1, & \text{if } \dfrac{1}{2} \leq \eta(x) \leq 1, \\ -1, & \text{if } 0 \leq \eta(x) < \dfrac{1}{2}, \end{cases}$$

where $\eta(x) = P(y = 1|x)$. Obviously, we can not use f_c directly in real applications since the population distribution P is unknown. There exist a huge amount of classification algorithms based on a finite samples $\mathbf{z} = \{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m \in Z^m$ to estimate f_c, including the classical distribution-based likelihood approaches such as logistic regression and Fisher linear discrimination analysis (LDA) [14], and the margin-based methods such as support vector machine (SVM) [3, 5]. As stated in [17], SVM suffers from data piling problems in the high-dimension low-sample size settings. Therefore, some new loss functions are invented to overcome the difficulty such as distance weighted discrimination (DWD) loss function [17] and LUMs loss function [16]. In this paper, LUMs loss function is adopted to measure the local error committed by the classifier $\mathcal{C}_f = \text{sgn}(f)$.

Definition 1.1. Let $0 \leq p \leq \infty$ and $0 < q \leq \infty$, the LUMs loss function $V : \mathbb{R} \to \mathbb{R}_+$ is defined by

$$V(t) = \begin{cases} \dfrac{1}{1+p}\Big(\dfrac{q}{(1+p)t - p + q}\Big)^q, & \text{if } t \geq \dfrac{p}{1+p}, \\ 1 - t, & \text{if } t < \dfrac{p}{1+p}. \end{cases} \tag{1.1}$$

LUMs loss function includes many popular loss functions such as DWD loss function when $p = 1$ and $q = 1$, the hinge loss function for SVM when $p = \infty$ and $q > 0$, and the Hybrid of SVM and AdaBoost function when $p = 0$ and $q = \infty$.

Denote by $\mathcal{H}_K$ the reproducing kernel Hilbert spaces (RKHS) generated by the Mercer kernel $K : X \times X \to \mathbb{R}$, which is a continuous, symmetric and positive semi-definite function. $\mathcal{H}_K$ is defined to be the completion of the linear span of functions $\{K_x = K(x, \cdot) : x \in X\}$ with the inner product $\langle \cdot, \cdot \rangle_K$ given by $\langle K_x, K_y \rangle_K = K(x, y)$. The reproducing property of $\mathcal{H}_K$ is given by $f(x) = \langle K_x, f \rangle_K$ which implies

$$\|f\|_{C(X)} \leqslant \kappa \|f\|_{\mathcal{H}_K}, \quad \forall f \in \mathcal{H}_K, \tag{1.2}$$

where $\kappa = \sup_{x\in X} \sqrt{K(x,x)}$ and $C(X)$ denotes the Banach space of continuous functions on X with norm $\|f\|_{C(X)} = \sup_{x\in X} |f(x)|$.

The estimator $f_{\mathbf{z},\lambda}$ is the minimizer of the following regularized LUMs scheme associated with LUMs loss function over $\mathcal{H}_K$:

$$f_{\mathbf{z},\lambda} := \arg\min_{f\in\mathcal{H}_K} \left\{ \frac{1}{m}\sum_{i=1}^{m} V(y_i f(x_i)) + \lambda \|f\|_{\mathcal{H}_K}^2 \right\}, \tag{1.3}$$

where $\lambda = \lambda(m) > 0$ is a regularization parameter balancing data fidelity and model complexity.

The purpose of this paper is to investigate the performance of the classifier $\mathcal{C}_{f_{\mathbf{z},\lambda}}$ induced by $f_{\mathbf{z},\lambda}$ with non i.i.d sampling, which can be measured by the excess misclassification error $\mathcal{R}(\mathrm{sgn}(f_{\mathbf{z},\lambda})) - \mathcal{R}(f_c)$. Assume that the sample is generated from a sequence of probability measures $\{P^{(t)}\}_{t=1}^{\infty}$ where the conditional distribution according to $\{P^{(t)}\}_{t=1}^{\infty}$ at each t is $P(y|x)$ while the marginal distributions $\{P_X^{(t)}\}_{t=1}^{\infty}$ change with t. Since the conditional distribution $P(y|x)$ keeps the same for each t, then the Bayes rule f_c is still well defined.

2. Notations and main results

Suppose that the sequence $\{P_X^{(t)}\}_{t=1}^{\infty}$ converges exponentially fast in the dual $(C^s(X))^*$ of the Hölder space $C^s(X)$ $(0 < s \leq 1)$, which is defined as the space of all continuous functions on X with the following norm:

$$\|f\|_{C^s(X)} = \|f\|_{C(X)} + |f|_{C^s(X)} < \infty, \text{ where } |f|_{C^s(X)} = \sup_{x\neq y} \frac{|f(x)-f(y)|}{|x-y|^s}.$$

Definition 2.1. Let $0 < s \leq 1$. The sequence $\{P_X^{(t)}\}_{t=1,2,\ldots}$ is said to converge exponentially fast to a probability measure P_X in $(C^s(X))^*$ if there exist $C > 0$ and $0 < \omega < 1$ such that

$$\|P_X^{(t)} - P_X\|_{(C^s(X))^*} \leq C\omega^t, \quad \forall t \in \mathbb{N}. \tag{2.1}$$

From the properties of the dual space, (2.1) is equivalent to

$$\left| \int_X f(x)dP_X^{(t)} - \int_X f(x)dP_X \right| \leq C\omega^t \|f\|_{C^s(X)}, \quad \forall f \in C^s(X), \quad t \in \mathbb{N}. \tag{2.2}$$

For the dependent sampling, we assume that the sequence of random variables satisfies the β-mixing condition.

Definition 2.2. Suppose $\mathbf{Z} = \{z_t\}_{t=1}^{\infty}$ is a sequence of random variables. For any i, $j \in \mathbb{N} \cup \{+\infty\}$, $\sigma_i^j = \sigma(z_i, \ldots, z_j)$ represents the σ-algebra generated by random variable $\{z_t : i \leq t \leq j\}$. Then for any $k \in \mathbb{N}$, the **β-mixing coefficient** of the random variable sequence $\mathbf{Z}$ is defined as

$$\beta(k) := \sup_{j\geq 1} \mathbb{E} \sup_{A\in\sigma_{j+k}^{\infty}} \left| \mathrm{Prob}(A \mid \sigma_1^j) - \mathrm{Prob}(A) \right|.$$

The random sequence $\mathbf{Z}$ is called β-mixing, if $\lim_{k\to\infty} \beta(k) = 0$. $\mathbf{Z}$ is called an **algebraically β-mixing sequence**, if for some $\beta_0 > 0$, $\vartheta > 0$, β-mixing coefficient satisfies

$$\beta(k) \leq \beta_0 k^{-\vartheta}, \text{ for } k \geq 1. \tag{2.3}$$

Z is called an **exponentially β-mixing sequence**, if for some $\beta_0 > 0$, $\beta_1 > 0$, $\vartheta > 0$, β-mixing coefficient satisfies

$$\beta(k) \le \beta_0 \exp(-\beta_1 k^{\vartheta}), \text{ for } k \ge 1. \tag{2.4}$$

Apart from the β-mixing condition used in this paper, there exist some other mixing conditions such as α-mixing and ϕ-mixing. In fact, the β-mixing condition is between the α-mixing and ϕ-mixing, which is not too weak like α-mixing condition nor too strong like ϕ-mixing condition (see [32] and the references within).

Denote the generalization error $\mathcal{E}^V(f)$ associated with the LUMs loss functions by

$$\mathcal{E}^V(f) = \int_{\mathcal{Z}} V(yf(x))\, dP(x,y) = \int_X \int_Y V(yf(x)) dP(y|x)\, dP_X(x). \tag{2.5}$$

It was shown in [16] that the minimizer f_P^V of $\mathcal{E}^V(f)$ over all measurable functions for $0 < q < \infty$ and $0 \le p < \infty$ is defined by

$$f_P^V(x) = \begin{cases} -\dfrac{1}{1+p}(R(x)^{-1}q - q + p), & \text{if } 0 \le \eta(x) < \dfrac{1}{2}, \\ \dfrac{1}{1+p}(R(x)q - q + p), & \text{if } \dfrac{1}{2} \le \eta(x) \le 1, \end{cases} \tag{2.6}$$

where $R(x) = \left(\frac{\eta(x)}{1-\eta(x)}\right)^{\frac{1}{q+1}}$. For $p \to \infty$, the LUMs loss reduces to the hinge loss for the SVM with the minimizer f_c.

The sample free version of (1.3) is

$$f_\lambda := \arg \min_{f \in \mathcal{H}_K} \{\mathcal{E}^V(f) + \lambda \|f\|^2_{\mathcal{H}_K}\}. \tag{2.7}$$

Define

$$\mathcal{D}(\lambda) := \mathcal{E}^V(f_\lambda) - \mathcal{E}^V(f_P^V) + \lambda \|f_\lambda\|^2_{\mathcal{H}_K} \tag{2.8}$$

as **regularization error** which is independent of sample and measures the approximation ability of $\mathcal{H}_K$.

Assumption 2.1. *For some constants $C_r > 0$, suppose the regularization error (2.8) satisfies*

$$\mathcal{D}(\lambda) \le C_r \lambda^r, \quad 0 < r \le 1. \tag{2.9}$$

Definition 2.3. Mercer kernel K satisfies a kernel condition of order s, if for some $\kappa_s > 0$, such that

$$\left|K(x,x) - 2K(x,x') + K(x',x')\right| \le \kappa_s^2 \left|x - x'\right|^{2s}, \ \forall x, x' \in X. \tag{2.10}$$

(2.10) holds true if $K \in C^{2s}(X \times X)$.

Definition 2.4. For a subset S of the metric space (E, d) and $u > 0$, the covering number $\mathcal{N}(S, u)$ is the minimal integer $l \in \mathbb{N}$ such that there exist l disks with radius u covering S.

The covering number is mainly used to describe the complexity of function space. In the literature of statistical learning theory, the covering numbers of unit balls of classical function spaces have been well investigated (see e.g. [1,7,24,33,34] In this paper, we use the covering number of the ball $B_R = \{f : \|f\|_{\mathcal{H}_K} \le R\}$ of the RKHS $\mathcal{H}_K$. Estimating uniform convergence in terms of covering numbers has been well developed in [4,6,9,11,12,27].

In our analysis we make the following assumption on the covering number.

Assumption 2.2. *Denote $B_1 = \{f \in \mathcal{H}_K : \|f\|_{\mathcal{H}_K} \le 1\}$. We assume the covering number satisfies the following capacity condition with some power $\iota > 0$ and $C_\iota > 0$,*

$$\log \mathcal{N}(B_1, u) \le C_\iota \left(\frac{1}{u}\right)^{\iota}, \forall\, u > 0. \tag{2.11}$$

By the fact that the LUMs loss V has no zero on $\mathbb{R}$ when $0 \le p < \infty$, we get an unbounded target function f_P^V. This leads to some difficulties in our analysis. In order to overcome these difficulties, we introduce a projection operator π_M.

Definition 2.5. For any $M > 0$, for any measurable function $f : X \to \mathbb{R}$, the projection operator π_M is defined as

$$\pi_M(f)(x) = \begin{cases} M, & \text{if } f(x) > M, \\ f(x), & \text{if } -M \le f(x) \le M, \\ -M, & \text{if } f(x) < -M. \end{cases} \tag{2.12}$$

The same projection operator was proposed to analyze the binary classification with logistic loss function in [28]. Since the LUMs loss without zero leads to an unbounded target function f_P^V, we assume that the projection operator has the form with varying levels, i.e., $M = M(m)$. The projection operator π_M involved in this work differs from the original one introduced for classifying loss with zero in [4, 21, 29, 30]. Since $\operatorname{sgn}(\pi_M(f_{\mathbf{z},\lambda})) = \operatorname{sgn}(f_{\mathbf{z},\lambda})$, then $\mathcal{R}(\operatorname{sgn}(f_{\mathbf{z},\lambda})) - \mathcal{R}(f_c) = \mathcal{R}(\operatorname{sgn}(\pi_M(f_{\mathbf{z},\lambda}))) - \mathcal{R}(f_c)$.

Before we demonstrate our main results by stating learning rates for the special case of $K \in C^\infty(X \times X)$, we introduce the following noise condition on the distribution P.

Definition 2.6. Let $0 \le \tau \le \infty$, we say that the probability measure P satisfies a Tsybakov noise condition with exponent τ if there exists a constant C_τ such that

$$P_X\big(\{x \in X : |2\eta(x) - 1| \le C_\tau t\}\big) \le t^\tau.\, \forall\, t > 0. \tag{2.13}$$

Note that (2.13) always holds for $\tau = 0$ with $C_\tau = 1$. If $|2\eta(x) - 1|$ is almost everywhere bounded, then $\tau = \infty$.

Theorem 2.1. *Let V be the LUMs loss with $0 \le p < \infty$ and $1/2 < q < \infty$. Let Assumption 1 and Assumption 2 be satisfied with $r = 1$ and $\iota > 0$. Assume that $K \in C^\infty(X \times X)$, the marginal distribution sequence $\{P_X^{(t)}\}_{t=1,2,\cdots}$ satisfies (2.2), the sample sequence $\{z_i\}_{i=1}^m$ satisfies the exponentially β-mixing condition (2.4) with $\vartheta > 0, \beta_0 > 0, \beta_1 > 0$. For any $\zeta > 0$ and $0 < \delta < 1$, let*

$$0 < \eta < \frac{(2q-1)\zeta}{2(1+2q)},$$

$$m \ge \max\left\{ \left(\frac{1}{\beta_1} \log\left(\frac{m^\zeta \left(2\log\frac{2}{\eta} + 1\right)\beta_0}{\delta}\right)\right)^{\frac{1}{(1-\zeta)\vartheta}}, 8^{\frac{1}{\zeta}} \right\}.$$

By taking $\lambda = m^{-\alpha}$ *with* $\alpha = \frac{(2q-1)\zeta}{2(1+2q)}$ *and* $M = m^{\beta}$ *with* $\beta = \frac{\zeta}{1+2q}$, *with confidence* $1-\delta$,

(i) for $0 < p < \infty$, *we have*

$$\begin{aligned}&\mathcal{R}(\operatorname{sgn}(\pi_M(f_{\mathbf{z},\lambda}))) - \mathcal{R}(f_c)\\ \leq & C'\left(\log\frac{2}{\eta}\right)^2 \log\left(\frac{8\left(2\log\frac{2}{\eta}+1\right)}{\delta}\right) m^{-\left(\frac{(2q-1)\zeta}{2(1+2q)}-\eta\right)};\end{aligned} \tag{2.14}$$

(ii) for $p = 0$, *if additionally the probability measure satisfies (2.13) with* $0 \leq \tau \leq \infty$, *we have*

$$\begin{aligned}&\mathcal{R}(\operatorname{sgn}(\pi_M(f_{\mathbf{z},\lambda}))) - \mathcal{R}(f_c)\\ \leq & C''\left(\log\frac{2}{\eta}\right)^2 \log\left(\frac{8\left(2\log\frac{2}{\eta}+1\right)}{\delta}\right) m^{-\left(\frac{(2q-1)\zeta}{2(1+2q)}-\eta\right)\frac{\tau+1}{\tau+2}},\end{aligned} \tag{2.15}$$

where the constants C' *and* C'' *are independent of* δ, m, η.

In fact (2.9) holds with $r = 1$ when $f_P^V \in \mathcal{H}_K$. The learning rate in (2.14) can be $O(m^{\epsilon - \frac{1}{2}})$ for arbitrarily small $\epsilon > 0$ by choosing $\zeta = 1 - 2\epsilon$ when q is large enough and η is small enough. So the learning rate can be very close to $\frac{1}{2}$. If we take $\tau = \infty$ leading to $\frac{\tau+1}{\tau+2} = 1$, we can achieve the same learning rates for $p = 0$.

In [8], the LUMs with i.i.d. was investigated. Suppose that Assumption 1 and Assumption 2 are satisfied with $r = 1$ and $\iota > 0$ and $K \in C^{\infty}(X \times X)$, the learning rates for $0 < p < \infty$ are $m^{-\left(\frac{q}{2(1+q)}-\eta\right)}$ which is sharper than the one in (2.14). It is reasonable because non-i.i.d. setting is much weaker than i.i.d. setting.

Theorem 2.2. *Let* V *be the LUMs loss with* $0 \leq p < \infty$ *and* $1/2 < q < \infty$. *Let Assumption 1 and Assumption 2 be satisfied with* $0 < r \leq 1$ *and* $\iota > 0$. *Assume that the marginal distribution sequence* $\{P_X^{(t)}\}_{t=1,2,\ldots}$ *satisfies (2.2), the sample sequence* $\{z_i\}_{i=1}^m$ *satisfies the algebraically* β*-mixing condition (2.3) with* $\vartheta > 0, \beta_0 > 0$, *and the kernel* K *satisfies (2.10) with* $s > 0$. *For* $0 < \delta < 1$ *and* $0 < \zeta < \frac{\vartheta}{1+\vartheta}$, *take* $\lambda = m^{-\alpha}$ *with* $0 < \alpha < \frac{4(2q-1)\zeta}{3(2+\iota)(1+2q)}$ *and* $M = m^{\beta}$ *with* $\beta = \frac{\zeta}{1+2q}$. *Let*

$$0 < \eta < \frac{2(2q-1)\zeta - (1+2q)(2+\iota)\alpha}{(2+\iota)(1+2q)} \tag{2.16}$$

and $m \geq \max\left\{\left(\frac{\left(2\log\frac{2}{\eta}+1\right)\beta_0}{\delta}\right)^{\frac{1}{\vartheta-\zeta(\vartheta+1)}}, 8^{\frac{1}{\zeta}}\right\}$.

(i) For $0 < p < \infty$, *with confidence* $1-\delta$, *we have*

$$\mathcal{R}(\operatorname{sgn}(\pi_M(f_{\mathbf{z},\lambda}))) - \mathcal{R}(f_c) \leq C'\left(\log\frac{2}{\eta}\right)^2 \log\left(\frac{8\left(2\log\frac{2}{\eta}+1\right)}{\delta}\right) m^{-\xi}. \tag{2.17}$$

(ii) For $p = 0$, if additionally the probability measure satisfies (2.13) with $0 \leq \tau \leq \infty$, with confidence $1 - \delta$, we have

$$\mathcal{R}(\operatorname{sgn}(\pi_M(f_{\mathbf{z},\lambda}))) - \mathcal{R}(f_c) \leq C'' \left(\log \frac{2}{\eta}\right)^2 \log \left(\frac{8\left(2\log\frac{2}{\eta}+1\right)}{\delta}\right) \left(m^{-\xi}\right)^{\frac{\tau+1}{\tau+2}}, \tag{2.18}$$

where

$$\xi = \min\left\{\alpha r, \frac{q\zeta}{1+2q}, \frac{\alpha(r-1)+\zeta}{2}, \frac{(2q-1)\zeta}{(2+\iota)(1+2q)} - \frac{\alpha(1-r)}{2}, \frac{(2q-1)(6+\iota)\zeta}{4(2+\iota)(1+2q)} - \frac{\alpha}{2}, \frac{(2q-1)\zeta}{(2+\iota)(1+2q)} - \frac{\alpha(3-r)-\zeta}{4}, \frac{2(2q-1)\zeta}{(2+\iota)(1+2q)} - \alpha - \eta\right\}. \tag{2.19}$$

Here the constant C' and C'' are independent of δ, m, η.

Theorem 2.3. *Under the assumptions of Theorem 2.2, if the sample sequence $\{z_i\}_{i=1}^m$ satisfies the exponentially β-mixing condition (2.4) with $\vartheta > 0$, $\beta_0 > 0$, $\beta_1 > 0$, for some $0 < \delta < 1$, $0 < \zeta < 1$ and*

$$m \geq \max\left\{\left(\frac{1}{\beta_1}\log\left(\frac{m^\zeta\left(2\log\frac{2}{\eta}+1\right)\beta_0}{\delta}\right)\right)^{\frac{1}{(1-\zeta)\vartheta}}, 8^{\frac{1}{\zeta}}\right\},$$

with confidence $1 - \delta$, the bounds (2.17) and (2.18) are still valid.

3. Error analysis

This section is devoted to the error analysis. Recall the comparison theorem associated with LUMs loss funtion investigated in [2] as follows, which plays a key role in our analysis.

Lemma 3.1. *(i) Let V be the LUMs loss with $0 < p < \infty$ and $0 < q \leq \infty$. For any probability measure P, any measurable function $f : X \to \mathbb{R}$, and some constant $C_p > 0$, it holds*

$$\mathcal{R}(\operatorname{sgn}(f)) - \mathcal{R}(f_c) \leq C_p\left(\mathcal{E}^V(f) - \mathcal{E}^V(f_P^V)\right). \tag{3.1}$$

(ii) Let V be the LUMs loss with $p = 0$. Under the assumption (2.13) with $0 \leq \tau \leq \infty$, the following comparison theorem holds true with some constant $C_{q,\tau} > 0$,

$$\mathcal{R}(\operatorname{sgn}(f)) - \mathcal{R}(f_c) \leq C_{q,\tau}\left(\mathcal{E}^V(f) - \mathcal{E}^V(f_P^V)\right)^{\frac{\tau+1}{\tau+2}}. \tag{3.2}$$

The above lemma implies that estimating the excess misclassification error $\mathcal{R}(\operatorname{sgn}(f))-\mathcal{R}(f_c)$ for the classifier $\operatorname{sgn}(f)$ can be done by bounding the excess generalization error $\mathcal{E}^V(f) - \mathcal{E}^V(f_P^V)$.

3.1. Error decomposition

Denote the error caused by the non-identically distributed sampling as

$$\mathcal{E}_m^V(f) = \frac{1}{m}\sum_{i=1}^{m}\int_Z V(yf(x))dP^{(i)}. \tag{3.3}$$

Let

$$\mathcal{E}_{\mathbf{z}}^V(f) = \frac{1}{m}\sum_{i=1}^{m} V(y_i f(x_i))$$

be the empirical error of f.

The following error decomposition is helpful to bound the excess generalization error $\mathcal{E}^V(\pi_M(f_{\mathbf{z},\lambda})) - \mathcal{E}^V(f_P^V)$ caused by the estimator $f_{\mathbf{z},\lambda}$.

Lemma 3.2. *Let $f_{\mathbf{z},\lambda} \in \mathcal{H}_K$ be defined by (1.3), $f_\lambda \in \mathcal{H}_K$ defined by (2.7) and $M > 0$. Then*

$$\mathcal{E}^V(\pi_M(f_{\mathbf{z},\lambda})) - \mathcal{E}^V(f_P^V) + \lambda\|f_{\mathbf{z},\lambda}\|_{\mathcal{H}_K}^2 \le \mathcal{P}(m,\lambda) + \mathcal{S}(\mathbf{z},\lambda) + \mathcal{D}(\lambda) + V(M), \tag{3.4}$$

where the **drift error** *denoted by $\mathcal{P}(m,\lambda)$ can be expressed as*

$$\mathcal{P}(m,\lambda) = \left\{\mathcal{E}^V(\pi_M(f_{\mathbf{z},\lambda})) - \mathcal{E}_m^V(\pi_M(f_{\mathbf{z},\lambda}))\right\} + \left\{\mathcal{E}_m^V(f_\lambda) - \mathcal{E}^V(f_\lambda)\right\},$$

the **sample error** *denoted by $\mathcal{S}(\mathbf{z},\lambda)$ can be expressed as*

$$\begin{aligned}\mathcal{S}(\mathbf{z},\lambda) = &\left\{\left[\mathcal{E}_{\mathbf{z}}^V(f_\lambda) - \mathcal{E}_{\mathbf{z}}^V(\pi_M(f_P^V))\right] - \left[\mathcal{E}_m^V(f_\lambda) - \mathcal{E}_m^V(\pi_M(f_P^V))\right]\right\}\\ &+ \left\{\left[\mathcal{E}_m^V(\pi_M(f_{\mathbf{z},\lambda})) - \mathcal{E}_m^V(\pi_M(f_P^V))\right] - \left[\mathcal{E}_{\mathbf{z}}^V(\pi_M(f_{\mathbf{z},\lambda})) - \mathcal{E}_{\mathbf{z}}^V(\pi_M(f_P^V))\right]\right\}\\ := &\,\mathcal{S}_1 + \mathcal{S}_2,\end{aligned}$$

the **regularization error** *$\mathcal{D}(\lambda)$ is defined by (2.8).*

Proof. Since the sample is non-identically distributed and the marginal distribution $\left\{P_X^{(t)}\right\}_{t=1,2,\ldots}$ for each sample point is different, then the excess generalization error can be written as

$$\begin{aligned}&\mathcal{E}^V(\pi_M(f_{\mathbf{z},\lambda})) - \mathcal{E}^V(f_P^V) + \lambda\left\|f_{\mathbf{z},\lambda}\right\|_{\mathcal{H}_K}^2\\ \le&\left\{\mathcal{E}^V(\pi_M(f_{\mathbf{z},\lambda})) - \mathcal{E}_m^V(\pi_M(f_{\mathbf{z},\lambda}))\right\} + \left\{\mathcal{E}_m^V(\pi_M(f_{\mathbf{z},\lambda})) - \mathcal{E}_{\mathbf{z}}^V(\pi_M(f_{\mathbf{z},\lambda}))\right\}\\ &+ \left\{\mathcal{E}_{\mathbf{z}}^V(\pi_M(f_{\mathbf{z},\lambda})) + \lambda\|f_{\mathbf{z},\lambda}\|_{\mathcal{H}_K}^2 - \left(\mathcal{E}_{\mathbf{z}}^V(f_\lambda) + \lambda\|f_\lambda\|_{\mathcal{H}_K}^2\right)\right\}\\ &+ \left\{\mathcal{E}_{\mathbf{z}}^V(f_\lambda) - \mathcal{E}_m^V(f_\lambda)\right\} + \left\{\mathcal{E}_m^V(f_\lambda) - \mathcal{E}^V(f_\lambda)\right\}\\ &+ \left\{\mathcal{E}^V(f_\lambda) - \mathcal{E}^V(f_P^V) + \lambda\|f_\lambda\|_{\mathcal{H}_K}^2\right\}.\end{aligned}$$

Since V is a decreasing function on $\mathbb{R}$, the projection operator induces that for $t \le M$, $V(\pi_M(t)) \le V(t)$, while for $t > M$, $V(\pi_M(t)) > V(t)$. Hence for any $t \in \mathbb{R}$, $V(\pi_M(t)) - V(t) \le V(M)$. This fact together with the definition of $f_{\mathbf{z},\lambda}$ yields that

$$\begin{aligned}&\left\{\mathcal{E}_{\mathbf{z}}^V(\pi_M(f_{\mathbf{z},\lambda})) + \lambda\|f_{\mathbf{z},\lambda}\|_{\mathcal{H}_K}^2 - \left(\mathcal{E}_{\mathbf{z}}^V(f_\lambda) + \lambda\|f_\lambda\|_{\mathcal{H}_K}^2\right)\right\}\\ \le&\left\{\mathcal{E}_{\mathbf{z}}^V(f_{\mathbf{z},\lambda}) + \lambda\|f_{\mathbf{z},\lambda}\|_{\mathcal{H}_K}^2 + V(M) - \left(\mathcal{E}_{\mathbf{z}}^V(f_\lambda) + \lambda\|f_\lambda\|_{\mathcal{H}_K}^2\right)\right\} \le V(M).\end{aligned}$$

Finally, the lemma is proved by adding and subtracting $\mathcal{E}_{\mathbf{z}}^V(\pi_M(f_P^V))$ and $\mathcal{E}_m^V(\pi_M(f_P^V))$ to the first and third term of the inequality. □

The appearance of $V(M)$ comes from the fact that the LUMs loss function is strictly decreasing and positive. The drift error $\mathcal{P}(m,\lambda)$ is caused by the non-identically distributed sampling. The above error decomposition is different from the standard one in [4, 21, 29, 30] for i.i.d. sample and convex loss functions with zero.

3.2. Bounds on drift error

In order to estimate the drift error we present the following lemma.

Lemma 3.3. *It follows from (2.8) with* $0 < r \leq 1$ *that*

$$\|f_\lambda\|_{\mathcal{H}_K} \leq \sqrt{\frac{\mathcal{D}(\lambda)}{\lambda}} \leq \sqrt{C_r}\lambda^{\frac{r-1}{2}}. \tag{3.5}$$

This lemma comes immediately from the fact

$$\lambda\|f_\lambda\|_{\mathcal{H}_K}^2 \leq \mathcal{E}^V(f_\lambda) - \mathcal{E}^V(f_P^V) + \lambda\|f_\lambda\|_{\mathcal{H}_K}^2 = \mathcal{D}(\lambda).$$

Lemma 3.4. *Let Assumptions 1 be satisfied. Assume that the kernel* K *satisfy (2.10) with* $s > 0$*. Then we have*

$$\begin{aligned}&\|V(f)\|_{C(X)} \leq 1 + \|f\|_{C(X)},\\&\|V(\pi_M(f_{\mathbf{z},\lambda}))\|_{C^s(X)} \leq \kappa_s\|f_{\mathbf{z},\lambda}\|_{\mathcal{H}_K} + 1 + M,\\&\|V(f_\lambda)\|_{C^s(X)} \leq 1 + (\kappa+\kappa_s)\sqrt{C_r}\lambda^{\frac{r-1}{2}}.\end{aligned} \tag{3.6}$$

Proof. The first inequality (3.6) can be obtained directly from the definition of LUM loss function.

Since $V(t)$ is a Lipschitz continuous function and $|V'(t)| \leq 1$, we find that

$$|V(f(x)) - V(f(x'))| \leq |f(x) - f(x')|.$$

Then

$$|V(f)|_{C^s(X)} = \sup_{x\neq x'} \frac{|V(f(x)) - V(f(x'))|}{|x-x'|^s} \leq \sup_{x\neq x'} \frac{|f(x) - f(x')|}{|x-x'|^s} = |f|_{C^s(X)}. \tag{3.7}$$

For any $x \in X,\ f \in \mathcal{H}_K$, recalling the reproducing property of $\mathcal{H}_K$, we know that

$$|f(x) - f(x')| = |\langle f, K_x - K_{x'}\rangle_K| \leq \|f\|_{\mathcal{H}_K}\sqrt{|K(x,x) - 2K(x,x') + K(x',x')|}.$$

Under the kernel condition (2.10), we obtain

$$|f|_{C^s(X)} = \sup_{x\neq x'\in X} \frac{|\ f(x) - f(x')\ |}{|\ x - x'\ |^s} \leq \kappa_s\|f\|_{\mathcal{H}_K}. \tag{3.8}$$

Firstly, we bound $\|V(\pi_M(f_{\mathbf{z},\lambda}))\|_{C(X)}$. Since $|\pi(f)(x) - \pi(f)(x')| \leq |f(x) - f(x')|$, then

$$|\pi(f)|_{C^s(X)} \leq |f|_{C^s(X)} \leq \kappa_s\|f\|_{\mathcal{H}_K}.$$

Based on (3.7) and (3.8), we can derive

$$|V(\pi_M(f_{\mathbf{z},\lambda}))|_{C^s(X)} \leq |\pi_M(f_{\mathbf{z},\lambda})|_{C^s(X)} \leq |f_{\mathbf{z},\lambda}|_{C^s(X)} \leq \kappa_s \|f_{\mathbf{z},\lambda}\|_{\mathcal{H}_K}.$$

In addition,

$$\|V(\pi_M(f_{\mathbf{z},\lambda}))\|_{C(X)} \leq V(-M) = 1 + M.$$

Consequently, we have

$$\|V(\pi_M(f_{\mathbf{z},\lambda}))\|_{C^s(X)} \leq \kappa_s \|f_{\mathbf{z},\lambda}\|_{\mathcal{H}_K} + 1 + M.$$

In the rest of the proof, we focus on estimating $\|V(f_\lambda)\|_{C^s(X)}$. Combining (1.2), (2.9) and (3.5), we have

$$\|f_\lambda\|_{C(X)} \leqslant \kappa \|f_\lambda\|_{\mathcal{H}_K} \leq \kappa\sqrt{C_r}\lambda^{\frac{r-1}{2}}.$$

The above inequality in connection with (3.6) yields

$$\|V(f_\lambda)\|_{C(X)} \leq 1 + \|f_\lambda\|_{C(X)} \leq 1 + \kappa\sqrt{C_r}\lambda^{\frac{r-1}{2}}.$$

Applying (3.7) and (3.8) again, we get that

$$|V(f_\lambda)|_{C^s(X)} \leq |f_\lambda|_{C^s(X)} \leq \kappa_s \|f_\lambda\|_{\mathcal{H}_K} \leq \kappa_s\sqrt{C_r}\lambda^{\frac{r-1}{2}}.$$

Therefore, we obtain that

$$\|V(f_\lambda)\|_{C^s(X)} \leq 1 + \kappa\sqrt{C_r}\lambda^{\frac{r-1}{2}} + \kappa_s\sqrt{C_r}\lambda^{\frac{r-1}{2}}.$$

□

Lemma 3.4 will be used in the following proposition to estimate the drift error. Recalling the definition of the regression function $f_P(x) = 2\eta(x) - 1, \eta(x) = P(y = 1|x)$.

Proposition 3.1. *Let (2.2) and the assumptions in Lemma 3.4 be satisfied. Then we have*

$$\begin{aligned} \left|\mathcal{E}_m^V(f_\lambda) - \mathcal{E}^V(f_\lambda)\right| &\leq \frac{C_1\omega}{1-\omega}\left(\lambda^{\frac{r-1}{2}} + 1\right)m^{-1}, \\ \left|\mathcal{E}^V(\pi_M(f_{\mathbf{z},\lambda})) - \mathcal{E}_m^V(\pi_M(f_{\mathbf{z},\lambda}))\right| &\leq \frac{C_2\omega}{1-\omega}\left(\|f_{\mathbf{z},\lambda}\|_{\mathcal{H}_K} + M\right)m^{-1}, \end{aligned} \tag{3.9}$$

where the constants C_1, C_2 are given by

$$\begin{aligned} C_1 &= C\max\left\{\left(4\kappa + 2\kappa_s + \kappa|f_P|_{C^s(X)}\right)\sqrt{C_r}, 4 + |f_P|_{C^s(X)}\right\}, \\ C_2 &= C\max\left\{2\kappa_s, 8 + 2|f_P|_{C^s(X)}\right\}. \end{aligned}$$

Proof. Let

$$\Gamma_i = \int_Z V(yf_\lambda(x))d(P^{(i)} - P) = \int_X\int_Y V(yf_\lambda(x))dP(y|x)d(P_X^{(i)} - P_X).$$

Notice that $\eta(x) = (1 + f_P(x))/2$ and $1 - \eta(x) = (1 - f_P(x))/2$. It follows by (2.2) that

$$\begin{aligned} |\Gamma_i| &= \left|\int_X \left\{\frac{1+f_P(x)}{2}V(f_\lambda) + \frac{1-f_P(x)}{2}V(-f_\lambda)\right\}d(P_X^{(i)} - P_X)\right| \\ &\leq \frac{1}{2}C\omega^i\left\{\|(1+f_P)V(f_\lambda)\|_{C^s(X)} + \|(1-f_P)V(-f_\lambda)\|_{C^s(X)}\right\}. \end{aligned}$$

For any $f, g \in C^s(X)$, it holds

$$\|fg\|_{C^s(X)} \leq \|f\|_{C(X)}\|g\|_{C^s(X)} + \|f\|_{C^s(X)}\|g\|_{C(X)}.$$

Since $\|1 + f_P\|_{C(X)} \leq 2$ and $\|1 + f_P\|_{C^s(X)} \leq 2 + |f_P|_{C^s(X)}$, we have

$$\|(1 + f_P)V(f_\lambda)\|_{C^s(X)} \leq 2\|V(f_\lambda)\|_{C^s(X)} + (2 + |f_P|_{C^s(X)})\|V(f_\lambda)\|_{C(X)}$$

and

$$\|(1 - f_P)V(-f_\lambda)\|_{C^s(X)} \leq 2\|V(-f_\lambda)\|_{C^s(X)} + (2 + |f_P|_{C^s(X)})\|V(-f_\lambda)\|_{C(X)}.$$

By Lemma 3.4, we obtain

$$\begin{aligned}|\Gamma_i| &\leq \frac{1}{2}C\omega^i \left\{4\left\|V(f_\lambda)\right\|_{C^s(X)} + \left(4 + 2|f_P|_{C^s(X)}\right)\left\|V(f_\lambda)\right\|_{C(X)}\right\}\\ &\leq C\omega^i\left\{2\left(1 + (\kappa + \kappa_s)\sqrt{C_r}\lambda^{\frac{r-1}{2}}\right) + \left(2 + |f_P|_{C^s(X)}\right)\left(1 + \|f_\lambda\|_{C(X)}\right)\right\}\\ &\leq C\omega^i\left\{2\left(1 + (\kappa + \kappa_s)\sqrt{C_r}\lambda^{\frac{r-1}{2}}\right) + \left(2 + |f_P|_{C^s(X)}\right)\left(1 + \kappa\sqrt{C_r}\lambda^{\frac{r-1}{2}}\right)\right\}\\ &\leq C\omega^i\left\{\left(4\kappa + 2\kappa_s + \kappa|f_P|_{C^s(X)}\right)\sqrt{C_r}\lambda^{\frac{r-1}{2}} + \left(4 + |f_P|_{C^s(X)}\right)\right\}.\end{aligned}$$

Therefore, it follows that

$$\begin{aligned}&\left|\mathcal{E}_m^V(f_\lambda) - \mathcal{E}^V(f_\lambda)\right| = \left|\frac{1}{m}\sum_{i=1}^m \Gamma_i\right|\\ \leq&\frac{C\omega}{1-\omega}\frac{1}{m}\left\{\left(4\kappa + 2\kappa_s + \kappa|f_P|_{C^s(X)}\right)\sqrt{C_r}\lambda^{\frac{r-1}{2}} + \left(4 + |f_P|_{C^s(X)}\right)\right\}\\ \leq&\frac{C_1\omega}{1-\omega}\left(\lambda^{\frac{r-1}{2}} + 1\right)m^{-1},\end{aligned}$$

where $C_1 = C\max\left\{\left(4\kappa + 2\kappa_s + \kappa|f_P|_{C^s(X)}\right)\sqrt{C_r}, 4 + |f_P|_{C^s(X)}\right\}.$

Similarly, we have

$$\begin{aligned}&\left|\mathcal{E}^V(\pi_M(f_{\mathbf{z},\lambda})) - \mathcal{E}_m^V(\pi_M(f_{\mathbf{z},\lambda}))\right|\\ \leq&\frac{C\omega}{1-\omega}\frac{1}{m}\left\{2\kappa_s\left\|f_{\mathbf{z},\lambda}\right\|_{\mathcal{H}_K} + 2(1+M) + \left(2 + |f_P|_{C^s(X)}\right)(1+M)\right\}\\ \leq&\frac{C\omega}{1-\omega}\frac{1}{m}\left\{2\kappa_s\left\|f_{\mathbf{z},\lambda}\right\|_{\mathcal{H}_K} + \left(8 + 2|f_P|_{C^s(X)}\right)M\right\}\\ \leq&\frac{C_2\omega}{1-\omega}\left(\|f_{\mathbf{z},\lambda}\|_{\mathcal{H}_K} + M\right)m^{-1},\end{aligned}$$

where $C_2 = C\max\left\{2\kappa_s, 8 + 2|f_P|_{C^s(X)}\right\}.$ □

Applying Proposition 3.1, it is easy to esimate the drift error $\mathcal{P}(m, \lambda)$ as follows

$$\begin{aligned}\mathcal{P}(m,\lambda) &= \left\{\mathcal{E}^V(\pi_M(f_{\mathbf{z},\lambda})) - \mathcal{E}_m^V(\pi_M(f_{\mathbf{z},\lambda}))\right\} + \left\{\mathcal{E}_m^V(f_\lambda) - \mathcal{E}^V(f_\lambda)\right\}\\ &\leq \frac{C_1\omega}{1-\omega}\left(\lambda^{\frac{r-1}{2}} + 1\right)m^{-1} + \frac{C_2\omega}{1-\omega}\left(\|f_{\mathbf{z},\lambda}\|_{\mathcal{H}_K} + M\right)m^{-1}.\end{aligned}$$

3.3. Bounds on sample error

We are now in a position to estimate the sample error $\mathcal{S}_1$ and $\mathcal{S}_2$. [32] introduced the blocking technique to deal with the original weakly dependent sequence. The basic idea is to construct an independent block (IB) sequence, which transforms the analysis of the dependent sample sequences into the analysis of independent block sequences.

Let $(z_1, z_2, \ldots, z_m)$ be an absolutely regular sequence. Given any integer pair (a_m, μ_m) with $\mu_m = [m/2a_m]$, the sequence is divided into $2\mu_m$ blocks of length a_m and a remainder block of length $m - 2a_m\mu_m$. For $1 \leq k \leq 2\mu_m$, let $Q_k^{a_m}$ be the marginal distribution of block $(z_{(k-1)a_m+1}, z_{(k-1)a_m+2}, \ldots, z_{ka_m})$ and $(z'_1, \ldots, z'_{2\mu_m a_m})$ a random sequence with the product distribution $\prod_{k=1}^{2\mu_m} Q_k^{a_m}$. Using the independent block technique, we define the following sequence

$$\begin{aligned}
Z_1 &= (z_1, \ldots, z_{a_m}, z_{2a_m+1}, \ldots, z_{3a_m}, \ldots, z_{2(\mu_m-1)a_m+1}, \ldots, z_{(2\mu_m-1)a_m}),\\
Z_2 &= (z_{a_m+1}, \ldots, z_{2a_m}, z_{3a_m+1}, \ldots, z_{4a_m}, \ldots, z_{(2\mu_m-1)a_m+1}, \ldots, z_{2\mu_m a_m}).
\end{aligned}$$

Correspondingly, we can define

$$\begin{aligned}
Z'_1 &= (z'_1, \ldots, z'_{a_m}, z'_{2a_m+1}, \ldots, z'_{3a_m}, \ldots, z'_{2(\mu_m-1)a_m+1}, \ldots, z'_{(2\mu_m-1)a_m}),\\
Z'_2 &= (z'_{a_m+1}, \ldots, z'_{2a_m}, z'_{3a_m+1}, \ldots, z'_{4a_m}, \ldots, z'_{(2\mu_m-1)a_m+1}, \ldots, z'_{2\mu_m a_m}).
\end{aligned}$$

The following lemma from [32] plays a key role to connect the original mixing sequences to the independent block sequences.

Lemma 3.5. *Assume that $Z^{\mu_m a_m}$ be a β-mixing sequence. For any bounded measurable function h on $Z^{\mu_m a_m}$, we have*

$$|\mathbb{E}h(Z_i) - \mathbb{E}h(Z'_i)| \leq \|h\|_\infty (\mu_m - 1)\beta(a_m), \ \forall\, i = 1, 2. \tag{3.10}$$

According to Lemma 3.5, we can transfer the problem of analyzing the weakly dependent sequences to analyzing the independent block sequences.

The following lemma is a corollary of Lemma 3.5, which plays an important role in our sample error estimation. Although the proof of Lemma 3.6 is similar to the one in [13], we still provide the proof here to make the paper self-contained.

Lemma 3.6. *Let $\mathscr{G}$ be a class of measurable functions on Z such that for each $g \in \mathscr{G}$, $\|g - \int_Z g dP^{(i)}\|_\infty \leq G$, then*

$$\text{Prob}\left(\sup_{g\in\mathscr{G}} \left|\frac{1}{m}\sum_{i=1}^m \left(g(z_i) - \int_Z g(z)dP^{(i)}\right)\right| > \epsilon + \frac{G}{\mu_m}\right) \leq \prod_1 + \prod_2 + 2\mu_m\beta(a_m), \tag{3.11}$$

where

$$\prod_1 = \text{Prob}\left(\sup_{g\in\mathscr{G}}\left|\frac{1}{\mu_m}\sum_{j=1}^{\mu_m}\frac{2\mu_m}{m}\sum_{i=2(j-1)a_m+1}^{(2j-1)a_m}\left(g(z'_i) - \int_Z g(z)dP^{(i)}\right)\right| \geq \epsilon\right),$$

$$\prod_2 = \text{Prob}\left(\sup_{g\in\mathscr{G}}\left|\frac{1}{\mu_m}\sum_{j=1}^{\mu_m}\frac{2\mu_m}{m}\sum_{i=(2j-1)a_m+1}^{2ja_m}\left(g(z'_i) - \int_Z g(z)dP^{(i)}\right)\right| \geq \epsilon\right).$$

Proof. Since

$$I := \sup_{g \in \mathscr{G}} \left| \frac{1}{m} \sum_{i=1}^{m} \left(g(z_i) - \int_Z g(z) dP^{(i)} \right) \right| \leq I_1 + I_2 + I_3,$$

where

$$\begin{aligned} I_1 &= \sup_{g \in \mathscr{G}} \left| \frac{1}{\mu_m} \sum_{j=1}^{\mu_m} \left(\frac{\mu_m}{m} \sum_{i=2(j-1)a_m+1}^{(2j-1)a_m} \left(g(z_i) - \int_Z g(z) dP^{(i)} \right) \right) \right|, \\ I_2 &= \sup_{g \in \mathscr{G}} \left| \frac{1}{\mu_m} \sum_{j=1}^{\mu_m} \left(\frac{\mu_m}{m} \sum_{i=(2j-1)a_m+1}^{2ja_m} \left(g(z_i) - \int_Z g(z) dP^{(i)} \right) \right) \right|, \\ I_3 &= \sup_{g \in \mathscr{G}} \left| \frac{1}{m} \sum_{i=2a_m\mu_m+1}^{m} \left(g(z_i) - \int_Z g(z) dP^{(i)} \right) \right|, \end{aligned}$$

and

$$\|I_3\|_\infty \leq \frac{1}{m} \sum_{i=2a_m\mu_m+1}^{m} \left\| g - \int_Z g(z) dP^{(i)} \right\|_\infty < \frac{G}{\mu_m}.$$

Applying Lemma 3.5 with $h = \chi_{\{2I_i > \epsilon\}}, i = 1, 2$, it yields that for all $\epsilon > 0$,

$$\begin{aligned} \text{Prob}\left(I_i > \frac{\epsilon}{2} \right) &= \mathbb{E}\chi_{\{I_i(z_i) > \frac{\epsilon}{2}\}} \\ &\leq \mathbb{E}\chi_{\{2I_i(z_i') > \epsilon\}} + \left\| \chi_{\{2I_i > \epsilon\}} \right\|_\infty (\mu_m - 1)\beta(a_m) \\ &\leq \mathbb{E}\chi_{\{2I_i(z_i') > \epsilon\}} + \mu_m \beta(a_m). \end{aligned}$$

Since $I_1 + I_2 + I_3 \geq I > \epsilon + \frac{G}{\mu_m}$, then $I_1 + I_2 + I_3 - \frac{G}{\mu_m} \geq I - \frac{G}{\mu_m} > \epsilon$, which leads to $I_1 + I_2 > \epsilon$. Therefore, we have

$$\begin{aligned} \text{Prob}\left(I > \epsilon + \frac{G}{\mu_m} \right) &\leq \text{Prob}\left(I_1 + I_2 > \epsilon \right) \leq \text{Prob}\left(2I_1 > \epsilon \right) + \text{Prob}\left(2I_2 > \epsilon \right) \\ &\leq \mathbb{E}\chi_{\{2I_1(z_i') > \epsilon\}} + \mathbb{E}\chi_{\{2I_2(z_i') > \epsilon\}} + 2(\mu_m - 1)\beta(a_m) \\ &\leq \prod_1 + \prod_2 + 2\mu_m \beta(a_m). \end{aligned}$$

□

Obviously, to bound $\sup_{g \in \mathscr{G}} \left| \frac{1}{m} \sum_{i=1}^{m} \left(g(z_i) - \int_Z g(z) dP^{(i)} \right) \right|$, we just need to estimate $\prod_1$ and $\prod_2$ respectively. Notice that $(z_1', \ldots, z_{2b_m a_m}')$ is an independent block sequence, so we can use the standard techniques for the independent case.

In order to bound the sample error, we need the following one-side Hoeffding inequality.

Lemma 3.7. *Let ξ be a random variable on a probability space Z with $\mathbb{E}(\xi) = \mu$, and satisfying $|\xi - \mu| \leq B$ for almost all $\mathbf{z} \in Z$. Then for all $\epsilon > 0$,*

$$\operatorname*{Prob}_{\mathbf{z} \in Z^m} \left\{ \frac{1}{m} \sum_{i=1}^{m} \xi(z_i) - \mu \ \geq \epsilon \right\} \leq \exp\left(-\frac{m\epsilon^2}{2B^2} \right).$$

If $\mathscr{G}$ is a singular function set, we use the above one-side Hoeffding inequality to get the following result.

Lemma 3.8. *Let g be a measurable function on Z satisfying $\|g(z) - \int_Z g(z)dP^{(i)}\|_\infty \leq G$. For any $\delta > 0$, the following inequality is established with confidence $1-\delta$*

$$\frac{1}{m}\sum_{i=1}^{m}\left(g(z_i) - \int_Z g(z)dP^{(i)}\right) \leq \frac{2\sqrt{2}G}{\sqrt{\mu_m}}\sqrt{\log\left(\frac{2}{\delta - 2\mu_m\beta(a_m)}\right)} + \frac{G}{\mu_m}. \tag{3.12}$$

Proof. For $k \in \mathbb{N}$, define $\xi_k = \frac{2\mu_m}{m}\sum_{i=(k-1)a_m+1}^{ka_m}\left(g(z_i') - \int_Z g(z)dP^{(i)}\right)$. Then $\mathbb{E}(\xi_k) = 0$ and

$$\begin{aligned}|\xi_k| &\leq \frac{2\mu_m}{m}\sum_{i=(k-1)a_m+1}^{ka_m}\left|g(z_i') - \int_Z g(z)dP^{(i)}\right| \\ &\leq \frac{2\mu_m}{m}\sum_{i=(k-1)a_m+1}^{ka_m}\left\|g(z) - \int_Z gdP^{(i)}\right\|_\infty \leq \frac{2\mu_m}{m}Ga_m \leq G.\end{aligned}$$

Applying the one-side Hoeffding inequality, for any $\epsilon > 0$, we have

$$\prod_1 = \text{Prob}\left(\frac{1}{\mu_m}\sum_{j=1}^{\mu_m}\xi_{2j-1} > \frac{\epsilon}{2}\right) \leq \exp\left(-\frac{\mu_m\epsilon^2}{8G^2}\right), \tag{3.13}$$

$$\prod_2 = \text{Prob}\left(\frac{1}{\mu_m}\sum_{j=1}^{\mu_m}\xi_{2j} > \frac{\epsilon}{2}\right) \leq \exp\left(-\frac{\mu_m\epsilon^2}{8G^2}\right).$$

According to (3.11), we get

$$\text{Prob}\left(\frac{1}{m}\sum_{i=1}^{m}\left(g(z_i) - \int_Z g(z)dP^{(i)}\right) > \epsilon + \frac{G}{\mu_m}\right) \leq 2\exp\left(-\frac{\mu_m\epsilon^2}{8G^2}\right) + 2\mu_m\beta(a_m).$$

For $\epsilon > 0$, solving the following equation

$$\exp\left(-\frac{\mu_m\epsilon^2}{8G^2}\right) = \frac{\delta}{2} - \mu_m\beta(a_m),$$

it follows

$$\epsilon = \frac{2\sqrt{2}G}{\sqrt{\mu_m}}\sqrt{\log\left(\frac{2}{\delta - 2\mu_m\beta(a_m)}\right)}.$$

Therefore, the desired result is proved. □

Now we apply Lemma 3.8 to estimate the sample error

$$\mathcal{S}_1 = \left\{\mathcal{E}_{\mathbf{z}}^V(f_\lambda) - \mathcal{E}_{\mathbf{z}}^V\left(\pi_M(f_P^V)\right)\right\} - \left\{\mathcal{E}_m^V(f_\lambda) - \mathcal{E}_m^V\left(\pi_M(f_P^V)\right)\right\}.$$

Proposition 3.2. *Let $M > 0$ and Assumption 1 be satisfied with $0 < r \leq 1$. If the sample sequence $\{z_i\}_{i=1}^m$ is a β-mixing sequence, then for any $\lambda > 0$ and any $0 < \delta < 1$, with confidence $1 - \delta/2$, we have*

$$\begin{aligned}&\left\{\mathcal{E}_{\mathbf{z}}^V(f_\lambda) - \mathcal{E}_{\mathbf{z}}^V\left(\pi_M(f_P^V)\right)\right\} - \left\{\mathcal{E}_m^V(f_\lambda) - \mathcal{E}_m^V\left(\pi_M(f_P^V)\right)\right\} \\ \leq &(2\sqrt{2}+1)\kappa\sqrt{C_r}\frac{\lambda^{\frac{r-1}{2}}}{\sqrt{\mu_m}}t + 3(2\sqrt{2}+1)\frac{M}{\sqrt{\mu_m}}t,\end{aligned}$$

where $t = \log\left(\frac{4}{\delta - 4\mu_m \beta(a_m)}\right)$.

Proof. Let $g(z) = V\left(yf_\lambda(x)\right) - V\left(y\pi_M(f_P^V)(x)\right)$ for $z = (x, y) \in Z$. Then the quantity $\mathcal{S}_1$ can be expressed as $\frac{1}{m}\sum_{i=1}^m \left(g(z_i) - \int_Z g(z)dP^{(i)}\right)$. Since $-V(-M) \leq g(z) \leq V\left(-\|f_\lambda\|_\infty\right)$, then it yields

$$\left\|g - \int_Z g(z)dP^{(i)}\right\|_\infty \leq V\left(-\|f_\lambda\|_\infty\right) + V(-M).$$

The definition of V and (3.5) imply that $V(-M) = 1 + M$ and $V(-\|f_\lambda\|_{C(X)}) \leq 1 + \kappa\sqrt{C_r}\lambda^{\frac{r-1}{2}}$.

Replacing δ and G in Lemma 3.8 by $\delta/2$ and $2 + M + \kappa\sqrt{C_r}\lambda^{\frac{r-1}{2}}$ respectively, the conclusion of this proposition is proved. □

The sample error $\mathcal{S}_2 = \mathcal{E}_m^V(\pi_M(f_{\mathbf{z},\lambda})) - \mathcal{E}_m^V(\pi_M(f_P^V)) - \left(\mathcal{E}_{\mathbf{z}}^V(\pi_M(f_{\mathbf{z},\lambda})) - \mathcal{E}_{\mathbf{z}}^V(\pi_M(f_P^V))\right)$ involves the function $f_{\mathbf{z},\lambda}$, which is not a singe function because it changes with the sample $\mathbf{z}$. Hence the analysis to bound $\mathcal{S}_2$ is more difficult. Here we overcome the difficulty by a covering number argument over the ball $B_R = \{f \in \mathcal{H}_K : \|f\|_{\mathcal{H}_K} \leq R\}$ where $f_{\mathbf{z},\lambda}$ belongs.

Proposition 3.3. *Let* $M > 0$ *and Assumption 2 be satisfied with* $\iota > 0$. *If the sample sequence* $\{z_i\}_{i=1}^m$ *is a* β*-mixing sequence, then for any* $\lambda > 0$ *and any* $0 < \delta < 1$, *with confidence* $1 - \delta/2$, *we have*

$$\begin{aligned}&\left\{\mathcal{E}_m^V(\pi_M(f_{\mathbf{z},\lambda})) - \mathcal{E}_m^V(\pi_M(f_P^V))\right\} - \left\{\mathcal{E}_{\mathbf{z}}^V(\pi_M(f_{\mathbf{z},\lambda})) - \mathcal{E}_{\mathbf{z}}^V(\pi_M(f_P^V))\right\}\\ \leq &\frac{36M}{\sqrt{\mu_m}}t + 32\sqrt{C_\iota}\left(\frac{M}{\sqrt{\mu_m}}\right)^{\frac{2}{2+\iota}} R^{\frac{\iota}{2+\iota}},\end{aligned}$$

where $t = \log\left(\frac{4}{\delta - 4\mu_m \beta(a_m)}\right)$.

Proof. Let $\mathscr{F}_R = \left\{V(y\pi_M(f)(x)) - V(y\pi_M(f_P^V)(x)) : f \in B_R\right\}$. The quantity $\mathcal{S}_2$ can be expressed as $\mathcal{S}_2 = \frac{1}{m}\sum_{i=1}^m \left(\int_Z g(z)dP^{(i)} - g(z_i)\right)$ for $g \in \mathscr{F}_R$. Let

$$\prod_1 = \text{Prob}\left(\sup_{g \in \mathscr{F}_R} \frac{1}{\mu_m}\sum_{j=1}^{\mu_m}\left(\frac{\mu_m}{m}\sum_{i=2(j-1)a_m+1}^{(2j-1)a_m}\left(\int_Z g(z)dP^{(i)} - g(z_i')\right)\right) \geq \frac{\epsilon}{2}\right).$$

To shorten the notation in this proof, we define

$$\Delta(g) = \frac{1}{\mu_m}\sum_{j=1}^{\mu_m}\frac{\mu_m}{m}\sum_{i=2(j-1)a_m+1}^{(2j-1)a_m}\left(g(z_i') - \int_Z g(z)dP^{(i)}\right).$$

Obviously, $-V(-M) \leq g(z) \leq V(-M)$. Then $\left\|g(z) - \int_Z g(z)dP^{(i)}\right\|_\infty \leq 2V(-M)$. Let $\ell = \mathcal{N}(\mathscr{F}_R, \frac{\epsilon}{4})$ and consider a set of functions $\{g_k\}_{k=1}^\ell \subset \mathscr{F}_R$ such that the disks $B^{(k)}, k = 1, 2, \cdots, \ell$ centered at g_k and with radius $\frac{\epsilon}{4}$ cover $\mathscr{F}_R$. For all $g \in B^{(k)}$,

$$\begin{aligned}|\Delta(g) - \Delta(g_k)| \leq &\frac{1}{\mu_m}\sum_{j=1}^{\mu_m}\frac{\mu_m}{m}\sum_{i=2(j-1)a_m+1}^{(2j-1)a_m}\left|g(z_i') - g_k(z_i') - \int_Z \left(g(z) - g_k(z)\right)dP^{(i)}\right|\\ \leq &\frac{1}{\mu_m}\sum_{j=1}^{\mu_m}\frac{\mu_m}{m}\sum_{i=2(j-1)a_m+1}^{(2j-1)a_m} 2\|g - g_k\|_{C(X)} \leq \|g - g_k\|_{C(X)} \leq \frac{\epsilon}{4}.\end{aligned}$$

It yields that

$$\sup_{g\in B^{(k)}} \Delta(g) \le -\frac{\epsilon}{2} \quad \Rightarrow \; \Delta(g_k) \le -\frac{\epsilon}{4}.$$

Therefore, for any $\epsilon > 0$, the above fact together with (3.13) implies that

$$\begin{aligned}\prod_1 =& \mathrm{Prob}\left\{\sup_{g\in\mathscr{F}_R} -\Delta(g) \ge \frac{\epsilon}{2}\right\} = \mathrm{Prob}\left\{\sup_{g\in\mathscr{F}_R} \Delta(g) \le -\frac{\epsilon}{2}\right\} \\ \le& \sum_{k=1}^{\ell} \mathrm{Prob}\left\{\sup_{g\in B^{(k)}} \Delta(g) \le -\frac{\epsilon}{2}\right\} \le \ell \ \mathrm{Prob}\left\{\Delta(g_k) \le -\frac{\epsilon}{4}\right\} \\ \le& \mathcal{N}(\mathscr{F}_R, \frac{\epsilon}{4}) \exp\left\{-\frac{\mu_m \epsilon^2}{128V^2(-M)}\right\}.\end{aligned}$$

In the same way, we can conclude that

$$\begin{aligned}\prod_2 =& \mathrm{Prob}\left\{\sup_{g\in\mathscr{F}_R} \frac{1}{\mu_m} \sum_{j=1}^{\mu_m} \frac{\mu_m}{m} \sum_{i=(2j-1)a_m+1}^{2ja_m} \left(\int_Z g(\mathrm{z})dP^{(i)} - g(\mathrm{z}'_i)\right) \ge \frac{\epsilon}{2}\right\} \\ \le& \mathcal{N}(\mathscr{F}_R, \frac{\epsilon}{4}) \exp\left\{-\frac{\mu_m \epsilon^2}{128\, V^2(-M)}\right\}.\end{aligned}$$

In addition, for any g_1 , $g_2 \in \mathscr{F}_R$, we observe that

$$\left|g_1 - g_2\right| \le \|f_1 - f_2\|_{C(X)}\,, \quad f_1, f_2 \in B_R.$$

Hence

$$\mathcal{N}(\mathscr{F}_R, \frac{\epsilon}{4}) \le \mathcal{N}(B_R, \frac{\epsilon}{4}) \le \mathcal{N}(B_1, \frac{\epsilon}{4R}).$$

(2.11) yields that

$$\log \mathcal{N}(B_1, \frac{\epsilon}{4R}) \le C_\iota \left(\frac{4R}{\epsilon}\right)^\iota.$$

Applying Lemma 3.6, we have

$$\begin{aligned}& \mathrm{Prob}\left\{\sup_{g\in\mathscr{F}_R} \frac{1}{m} \sum_{i=1}^{m} \left(\int_Z g(z)dP^{(i)} - g(z_i)\right) > \epsilon + \frac{2V(-M)}{\mu_m}\right\} \\ \le& 2\mathcal{N}(B_1, \frac{\epsilon}{4R}) \exp\left(-\frac{\mu_m \epsilon^2}{128\, V^2(-M)}\right) + 2\mu_m \beta(a_m).\end{aligned}$$

Let $\epsilon^*(\mu_m, R, M, \delta/2)$ be the positive root of the following equation:

$$C_\iota \left(\frac{4R}{\epsilon}\right)^\iota - \frac{\mu_m \epsilon^2}{128\ V^2(-M)} = \log\left(\frac{\delta - 4\mu_m\beta(a_m)}{4}\right),$$

which can be rewritten as

$$\epsilon^{2+\iota} - \frac{128\, V^2(-M)}{\mu_m} \log\left(\frac{4}{\delta - 4\mu_m\beta(a_m)}\right) \epsilon^\iota - \frac{2^{2\iota+7} C_\iota V^2(-M)}{\mu_m} R^\iota = 0.$$

Applying Lemme 7.2 in [6], $\varepsilon^*(\mu_m, R, M, \delta/2)$ can be bounded as

$$\begin{aligned}
&\epsilon^*(\mu_m, R, M, \delta/2)\\
\leq &\max\left\{\frac{16\,V(-M)}{\sqrt{\mu_m}}\sqrt{\log\left(\frac{4}{\delta-4\mu_m\beta(a_m)}\right)},\ 16\sqrt{C_\iota}\left(\frac{V^2(-M)}{\mu_m}\right)^{\frac{1}{2+\iota}}R^{\frac{\iota}{2+\iota}}\right\}\\
\leq &\frac{16\,V(-M)}{\sqrt{\mu_m}}\log\left(\frac{4}{\delta-4\mu_m\beta(a_m)}\right)+16\sqrt{C_\iota}\left(\frac{V^2(-M)}{\mu_m}\right)^{\frac{1}{2+\iota}}R^{\frac{\iota}{2+\iota}}.
\end{aligned}$$

Substituting $V(-M) = 1 + M$ into the above formula, the conclusion is proved. □

3.4. Fast learning rates by iteration

In this section we conduct the iterative algorithm to improve the bound of $||f_{\mathbf{z},\lambda}||_{\mathcal{H}_K}$. Let

$$\mathcal{W}_R = \{\mathbf{z} \in Z^m : \|f_{\mathbf{z},\lambda}\|_{\mathcal{H}_K} \leq R\},\ \forall\ R > 0.$$

The following decay property of the LUMs loss function is required in our error analysis, which is proved in [8].

Lemma 3.9. *Let V be the LUM loss functions with $0 \leq p < \infty$ and $0 < q < \infty$. For $t \geq \frac{p}{1+p}$, it holds*

$$V(t) \leq C_{p,q}t^{-q} \tag{3.14}$$

where $C_{p,q} = (1/(1+p))^{q+1}(\max\{p,q\})^q$.

To shorten the notation, define $\Delta_{\mathbf{z}} := \mathcal{E}^V(\pi_M(f_{\mathbf{z},\lambda})) - \mathcal{E}^V(f_P^V) + \lambda\|f_{\mathbf{z},\lambda}\|^2_{\mathcal{H}_K}$. Applying Lemma 3.3, 3.9, Proposition 3.1, 3.2 and 3.3, the total error can be derived.

Proposition 3.4. *Let V be the LUMs loss with $0 \leq p < \infty$ and $0 < q < \infty$. Let Assumption 1 and Assumption 2 be satisfied with $0 < r \leq 1$ and $\iota > 0$. Assume that the marginal distribution sequence $\{P_X^{(t)}\}_{t=1,2,\ldots}$ satisfies (2.2), the sample sequence $\{z_i\}_{i=1}^m$ is a β-mixing sequence, and the kernel K satisfies (2.10) with $s > 0$. Let $R \geq 1$, $M \geq 1$, $0 < \lambda \leq 1$. For $0 < \delta < 1$, there exists a subset V_R of Z^m with measure at most δ such that*

$$\begin{aligned}
&\mathcal{E}^V(\pi_M(f_{\mathbf{z},\lambda})) - \mathcal{E}^V(f_P^V) + \lambda\|f_{\mathbf{z},\lambda}\|^2_{\mathcal{H}_K}\\
\leq &C_r\lambda^r + C_{p,q}M^{-q} + \frac{2C_1\omega}{1-\omega}\frac{\lambda^{\frac{r-1}{2}}}{m} + \frac{C_2\omega}{1-\omega}\frac{M}{m} + 4\kappa\sqrt{C_r}\frac{\lambda^{\frac{r-1}{2}}}{\sqrt{\mu_m}}t\\
&+ 48\frac{M}{\sqrt{\mu_m}}t + \frac{C_2\omega}{1-\omega}\frac{R}{m} + 32\sqrt{C_\iota}\left(\frac{M}{\sqrt{\mu_m}}\right)^{\frac{2}{2+\iota}}R^{\frac{\iota}{2+\iota}}, \qquad \forall\,\mathbf{z} \in \mathcal{W}_R \setminus V_R,
\end{aligned} \tag{3.15}$$

where $t = \log\left(\frac{4}{\delta-4\mu_m\beta(a_m)}\right)$.

Proof. From Proposition 3.1, for any $\mathbf{z} \in Z^m$, we get

$$\mathcal{P}(m,\lambda) \leq \frac{2C_1\omega}{1-\omega}\frac{\lambda^{\frac{r-1}{2}}}{m} + \frac{C_2\omega}{1-\omega}\frac{R}{m} + \frac{C_2\omega}{1-\omega}\frac{M}{m}.$$

Proposition 3.2 ensures the existence of V_1 of Z^m with measure at most $\delta/2$ such that

$$\mathcal{S}_1 \leq (2\sqrt{2}+1)\kappa\sqrt{C_r}\,\frac{\lambda^{\frac{r-1}{2}}}{\sqrt{\mu_m}}\,t + 3(2\sqrt{2}+1)\frac{M}{\sqrt{\mu_m}}\,t, \quad \forall\, \mathbf{z} \in \mathcal{W}_R \setminus V_1.$$

Proposition 3.3 tells us that there exists a subset V_2 of Z^m with measure at most $\delta/2$ such that

$$\mathcal{S}_2 \leq \frac{36M}{\sqrt{\mu_m}}\,t + 32\sqrt{C_\iota}\left(\frac{M}{\sqrt{\mu_m}}\right)^{\frac{2}{2+\iota}} R^{\frac{\iota}{2+\iota}}, \quad \forall\, \mathbf{z} \in \mathcal{W}_R \setminus V_2.$$

Let $V_R = V_1 \cup V_2$. The above estimations in connection with Lemma 3.2, 3.9 and Assumption 1 yield that

$$\begin{aligned}\Delta_{\mathbf{z}} \leq & C_r\lambda^r + C_{p,q}M^{-q} + \frac{2C_1\omega}{1-\omega}\,\frac{\lambda^{\frac{r-1}{2}}}{m} + \frac{C_2\omega}{1-\omega}\,\frac{M}{m} + 4\kappa\sqrt{C_r}\,\frac{\lambda^{\frac{r-1}{2}}}{\sqrt{\mu_m}}\,t \\ & + 48\frac{M}{\sqrt{\mu_m}}\,t + \frac{C_2\omega}{1-\omega}\,\frac{R}{m} + 32\sqrt{C_\iota}\left(\frac{M}{\sqrt{\mu_m}}\right)^{\frac{2}{2+\iota}} R^{\frac{\iota}{2+\iota}}, \quad \forall\, \mathbf{z} \in \mathcal{W}_R \setminus V_R.\end{aligned}$$

The desired result is proved. □

Recalling the definition (1.3) of $f_{\mathbf{z},\lambda}$, it is easy to get that by taking $f = 0$

$$\|f_{\mathbf{z},\lambda}\|_{\mathcal{H}_K} \leq \lambda^{-\frac{1}{2}}, \quad \forall\, \mathbf{z} \in Z^m. \tag{3.16}$$

We observe that the bound in (3.5) is much better than the one in (3.16). This motivates us to get a similar tight bound for $f_{\mathbf{z},\lambda}$. We will apply Proposition 3.4 iteratively to achieve this target which in turn improves learning rates. This iteration technique has been used in [22, 26].

Lemma 3.10. *Suppose that all assumptions in Proposition 3.4 are satisfied. Take $\lambda = m^{-\alpha}$ with $0 < \alpha \leq 1$ and $M = m^\beta$ with $0 < \beta \leq \infty$. Let $0 < \zeta < 1$, $0 < \eta < 1$ and $m \geq 8^{\frac{1}{\zeta}}$. For any $0 < \delta < 1$, with confidence $1 - 2\log\frac{2}{\eta}\,\delta$, there holds*

$$\|f_{\mathbf{z},\lambda}\|_{\mathcal{H}_K} \leq R^{(J)} \leq C_5\left(\log\frac{2}{\eta}\right)^2\sqrt{\log\left(\frac{4}{\delta - 4\mu_m\beta(a_m)}\right)}\, m^{\theta_\eta}, \tag{3.17}$$

where

$$\theta_\eta = \max\left\{\frac{\alpha(1-r)}{2}, \frac{\alpha-\beta q}{2}, \frac{\alpha(3-r)-\zeta}{4}, \frac{\alpha+\beta-\zeta/2}{2}, \frac{\alpha(2+\iota)+2\beta-v\zeta}{2+\iota}+\eta\right\} \geq 0$$

and

$$C_5 = \left(\frac{2C_2\omega}{1-\omega} + 64(2C_\iota)^{\frac{1}{2}}\right)\left(1 + \sqrt{C_r} + \sqrt{C_{p,q}} + 2\sqrt{\frac{C_2\omega}{1-\omega}} + 13 + 4\sqrt{\kappa\sqrt{C_r}}\right).$$

Proof. Take a_m satisfying $m^{1-\zeta} \leq a_m \leq m^{1-\zeta} + 1, 0 < \zeta < 1$. Since $m \geq 8^{\frac{1}{\zeta}}$, it follows by $\mu_m = [\frac{m}{2a_m}]$ that

$$\frac{1}{\mu_m} \leq \frac{1}{\frac{m}{2a_m} - 1} \leq \frac{2(m^{1-\zeta}+1)}{m - 2(m^{1-\zeta}+1)} \leq 8m^{-\zeta}.$$

Thus $\frac{1}{\sqrt{\mu_m}} \le 2\sqrt{2}m^{-\zeta/2}$. Let $\lambda = m^{-\alpha}$, $M = m^{\beta}$ and $t > 1$, we obtain from (3.4) that for all $\mathbf{z} \in \mathcal{W}_R \setminus V_R$,

$$\begin{aligned}\Delta_{\mathbf{z}} \le & C_r m^{-\alpha r} + C_{p,q} m^{-\beta q} + \frac{2C_1\omega}{1-\omega} m^{-\frac{\alpha(r-1)+2}{2}} + 8\kappa\sqrt{2C_r} m^{-\frac{\alpha(r-1)+\zeta}{2}} + \frac{C_2\omega}{1-\omega} m^{-(1-\beta)} \\ & + 96\sqrt{2} m^{-\left(\frac{\zeta}{2}-\beta\right)} t + \frac{C_2\omega}{1-\omega} m^{-1} R + 64\sqrt{2C_\iota} m^{-\frac{\zeta-2\beta}{2+\iota}} R^{\frac{\iota}{2+\iota}} \\ \le & C_r m^{-\alpha r} + C_{p,q} m^{-\beta q} + \left(\frac{2C_1\omega}{1-\omega} + 8\kappa\sqrt{2C_r}\right) m^{-\frac{\alpha(r-1)+\zeta}{2}} \\ & + \left(\frac{C_2\omega}{1-\omega} + 96\sqrt{2}\right) m^{-\left(\frac{\zeta}{2}-\beta\right)} t + \left(\frac{C_2\omega}{1-\omega} + 64\sqrt{2C_\iota}\right) m^{-\frac{\zeta-2\beta}{2+\iota}} R. \end{aligned} \tag{3.18}$$

Therefore, we have

$$\|f_{\mathbf{z},\lambda}\|_{\mathcal{H}_K} \le \sqrt{\frac{\Delta_{\mathbf{z}}}{\lambda}} \le c_m R^{\frac{1}{2}} + d_m, \tag{3.19}$$

where

$$\begin{aligned} & c_m = C_3 m^{\frac{\alpha}{2} + \frac{2\beta-\zeta}{2(2+\iota)}}, \ d_m = C_4\sqrt{t} m^{\theta}, \\ & \theta = \max\left\{\frac{\alpha(1-r)}{2}, \ \frac{\alpha-\beta q}{2}, \ \frac{\alpha(3-r)-\zeta}{4}, \ \frac{\alpha+\beta-\zeta/2}{2}\right\} \ge 0, \\ & C_3 = (C_2\omega/(1-\omega))^{1/2} + 8(2C_\iota)^{1/4}, \\ & C_4 = \left(\sqrt{C_r} + \sqrt{C_{p,q}} + 2\sqrt{\frac{C_2\omega}{1-\omega}} + 13 + 4\sqrt{\kappa\sqrt{C_r}}\right). \end{aligned}$$

It follows that

$$\mathcal{W}_R \subseteq \mathcal{W}\left(c_m R^{\frac{1}{2}} + d_m\right) \cup V_R. \tag{3.20}$$

Let us apply (3.20) iteratively to a sequence $\{R^{(j)}\}_{j=0}^{J}$ defined by $R^{(0)} = \lambda^{-1/2} = m^{\alpha/2}$ and $R^{(j)} = c_m \left(R^{(j-1)}\right)^{\frac{1}{2}} + d_m$, where $j \ge 2$. Then $\mathcal{W}_{R^{(j-1)}} \subseteq \mathcal{W}_{R^{(j)}} \cup V_{R^{(j-1)}}$. (3.16) implies that $\mathcal{W}_{R^{(0)}} = Z^m$. Hence we have

$$Z^m = \mathcal{W}_{R^{(0)}} \subseteq \mathcal{W}_{R^{(1)}} \cup V_{R^{(0)}} \subseteq \cdots \subseteq \mathcal{W}_{R^{(J)}} \cup \left(\cup_{j=0}^{J-1} V_{R^{(j)}}\right).$$

Due to the measure of $V_{R^{(j)}}$ is at most δ, we get that the measure of $\cup_{j=0}^{J-1} V_{R^{(j)}}$ is at most $J\delta$. Hence, $\mathcal{W}_{R^{(J)}}$ has measure at least $1 - J\delta$.

Denote $\Delta = \frac{1}{2}$. It follows from the definition of the sequence $\{R^{(j)}\}_{j=1}^{J}$ that

$$R^{(J)} \le c_m^{1+\Delta+\Delta^2+\cdots+\Delta^{J-1}} \left(R^{(0)}\right)^{\Delta^J} + \sum_{j=1}^{J-1} c_m^{1+\Delta+\Delta^2+\cdots+\Delta^{j-1}} d_m^{\Delta^j} + d_m. \tag{3.21}$$

Now we need to bound the two terms on the right-hand side of (3.21).

The first term equals

$$C_3^{\frac{1-\Delta^J}{1-\Delta}} m^{\frac{\alpha(2+\iota)+2\beta-\zeta}{2(2+\iota)} \frac{1-\Delta^J}{1-\Delta}} m^{\frac{\alpha\Delta^J}{2}}$$

which can be bounded by

$$C_3^2 m^{\frac{\alpha(2+\iota)+2\beta-\zeta}{2(2+s)(1-\Delta)}} m^{\left(\frac{\alpha}{2} - \frac{\alpha(2+\iota)+2\beta-\zeta}{2(2+\iota)(1-\Delta)}\right)\Delta^J} \le C_3^2 m^{\frac{\alpha(2+\iota)+2\beta-\zeta}{2+\iota}} m^{\frac{\zeta}{2+\iota}} 2^{-J}.$$

Take J to be the smallest integer such that $J \geq \log(1/\eta)/\log 2$. The above inequality can be bounded by

$$C_3^2\, m^{\frac{\alpha(2+\iota)+2\beta-\zeta}{2+\iota}+\eta}.$$

The second term of (3.21) can be bounded by

$$\begin{aligned}&\sum_{j=1}^{J-1} C_3^2\, m^{\frac{\alpha(2+\iota)+2\beta-\zeta}{2(2+\iota)}\frac{1-\Delta^j}{1-\Delta}} \left(C_4\sqrt{t}\right)^{\Delta^j} m^{\theta\Delta^j} + C_4\sqrt{t}m^{\theta},\\ &\leq C_3^2\, C_4\sqrt{t}\, m^{\frac{\alpha(2+\iota)+2\beta-\zeta}{2+\iota}} \sum_{j=0}^{J-1} m^{\left(\theta-\frac{\alpha(2+\iota)+2\beta-\zeta}{2+\iota}\right)2^{-j}}.\end{aligned}$$

When $\theta \leq \frac{\alpha(2+\iota)+2\beta-\zeta}{2+\iota}$, the above expression can be bounded by

$$C_3^2\, C_4\sqrt{t}\, J\, m^{\frac{\alpha(2+\iota)+2\beta-\zeta}{2+\iota}}.$$

When $\theta \geq \frac{\alpha(2+\iota)+2\beta-\zeta}{2+\iota}$, the bound is

$$C_3^2\, C_4\sqrt{t}\, J\, m^{\theta}.$$

According to the above discussion, we finally obtain that

$$R^{(J)} \leq C_3^2 \left(1 + C_4\sqrt{t}\, J\right) m^{\theta_\eta},$$

where $\theta_\eta = \max\left\{\theta,\ \frac{\alpha(2+\iota)+2\beta-\zeta}{2+\iota} + \eta\right\}$. Hence with confidence $1 - J\delta$, there holds

$$\|f_{\mathbf{z},\lambda}\|_{\mathcal{H}_K} \leq C_3^2\ (1 + C_4)\sqrt{t}\, J m^{\theta_\eta}.$$

Finally, the desired result follows by taking J to satisfy $J \leq 2\log\frac{2}{\eta}$. □

4. Proofs of main results

In this section, we will prove our main results. We first provide the proof for the general case.

Proof of Theorem 2.2. We take a_m to satisfy $m^{1-\zeta} \leq a_m \leq m^{1-\zeta} + 1$ with $0 < \zeta < 1$ and $\mu_m = [\frac{m}{2a_m}]$. Let $R = C_5 J\sqrt{t}m^{\theta_\eta}$, $t = \sqrt{\log\left(\frac{4}{\delta - 4\mu_m\beta(a_m)}\right)}$, $\lambda = m^{-\alpha}$ and $M = m^{\beta}$. Combining Lemma 3.10 and Proposition (3.4), it holds with confidence $1 - \left(2\log\frac{2}{\eta} + 1\right)\delta$ that

$$\begin{aligned}&\mathcal{E}^V(\pi_M(f_{\mathbf{z},\lambda})) - \mathcal{E}^V(f_P^V)\\ \leq &C_r m^{-\alpha r} + C_{p,q}m^{-\beta q} + \left(\frac{2C_1\omega}{1-\omega} + 8\kappa\sqrt{2C_r}\right) m^{-\frac{\alpha(r-1)+\zeta}{2}} + \left(\frac{C_2\omega}{1-\omega} + 96\sqrt{2}\right) m^{-\left(\frac{\zeta}{2}-\beta\right)} t\\ &+ \left(\frac{C_2\omega}{1-\omega} + 64\sqrt{2C_\iota}\right) C_5 \left(\log\frac{2}{\eta}\right)^2 \sqrt{t}\, m^{-\frac{\zeta-2\beta}{2+\iota}+\theta_\eta}\\ \leq &C_6 \left(\log\frac{2}{\eta}\right)^2 t\, m^{-\xi},\end{aligned}$$

where

$$\xi = \min\left\{\alpha r,\, \beta q,\, \frac{\alpha(r-1)+\zeta}{2},\, \frac{\zeta-2\beta}{2},\, \frac{\zeta-2\beta}{2+\iota} - \theta_\eta\right\}, \quad \theta_\eta < \frac{\zeta-2\beta}{2+\iota}$$

and

$$C_6 = C_r + C_{p,q} + \frac{2C_1\omega}{1-\omega} + 8\kappa\sqrt{2C_r} + \frac{C_2\omega}{1-\omega} + 96\sqrt{2} + \left(\frac{C_2\omega}{1-\omega} + 64\sqrt{2C_\iota}\right) C_5.$$

Setting $\beta = \frac{\zeta}{1+2q}$, then $\theta_\eta < \frac{(2q-1)\zeta}{(2+\iota)(1+2q)}$. If $0 < \alpha < \frac{4(2q-1)\zeta}{3(2+\iota)(1+2q)}$ with $q > 1/2$, we get that $\frac{\alpha(r-1)}{2} < \frac{(2q-1)\zeta}{(2+\iota)(1+2q)}$, $\frac{\alpha-\beta q}{2} < \frac{\alpha+\beta-\zeta/2}{2} < \frac{(2q-1)\zeta}{(2+\iota)(1+2q)}$ and $\frac{\alpha(3-r)-\zeta}{4} < \frac{(2q-1)\zeta}{(2+\iota)(1+2q)}$. Additionally, η satisfies (2.16), it follows that $\frac{\alpha(2+\iota)+2\beta-\zeta}{2+\iota} + \eta < \frac{(2q-1)\zeta}{(2+\iota)(1+2q)}$. Therefore, the above discussion ensures $\theta_\eta < \frac{\zeta-2\beta}{2+\iota}$.

Since $\beta(k) \leq \beta_0 k^{-\vartheta}$ with $\vartheta > 0$ and $\beta_0 > 0$, we choose m such that $\delta - 4\left(2\log\frac{2}{\eta}+1\right)\mu_m\beta(a_m) \leq \frac{\delta}{2}$ and replace δ by $\frac{\delta}{2\log\frac{2}{\eta}+1}$. Applying Lemma 3.1, the proof of Theorem 2.2 is complete. □

Proof of Theorem 2.3. The proof is similar to the one for Theorem 2.2. Here we just need to notice that the sample sequence $\{z_i\}_{i=1}^m$ is an exponentially β-mixing sequence for $k \geq 1$, i.e. $\beta(k) \leq \beta_0 \exp(-\beta_1 k^\vartheta)$ with $\beta_0 > 0$, $\beta_1 > 0$, $\vartheta > 0$. Hence under this assumption, we choose m such that $\delta - 4\left(2\log\frac{2}{\eta}+1\right)\mu_m\beta(a_m) \leq \frac{\delta}{2}$. □

Now we are in a position to show Theorem 2.1 with the constants $C^{'} = C_6 C_p$, $C^{''} = C_6 C_{q,\tau}$.

Proof of Theorem 2.1. It was proved in [34] that (2.11) holds for any $\iota > 0$ if $K \in C^\infty(X \times X)$. With $0 < \eta < \frac{(2q-1)\zeta}{2(1+2q)}, q > 1/2$, let us choose ι to be a positive number satisfying the following four inequalities:

$$\begin{aligned}
&\frac{(2q-1)\zeta}{2(1+2q)} < \frac{(2q-1)\zeta}{(2+\iota)(1+2q)} < \frac{4(2q-1)\zeta}{3(2+\iota)(1+2q)},\\
&\frac{1}{3} < \frac{2(2q-1)\zeta - \frac{1}{2}(2+\iota)(2q-1)\zeta}{(2+\iota)(1+2q)},\\
&\frac{(2q-1)\zeta}{2(1+2q)} - \eta < \frac{2(2q-1)\zeta}{(2+\iota)(1+2q)} - \frac{(2q-1)\zeta}{2(1+2q)} - \frac{1}{3},\\
&\frac{(2q-1)\zeta}{2(1+2q)} - \eta < \frac{2(2q-1)\zeta}{(2+\iota)(1+2q)}.
\end{aligned}$$

The first inequality above tells us that the restriction on α is satisfied by choosing $\alpha = \frac{(2q-1)\zeta}{2(1+2q)}$. The second inequality shows that condition (2.16) for the parameter η renamed now as η^* is also satisfied by taking $\eta^* = 1/3$. Thus we apply Lemma 3.1 and Theorem 2.3 by taking $r = 1$ and $\alpha = \frac{(2q-1)\zeta}{2(1+2q)}$, the proof is complete. □

References

[1] P. L. Bartlett, *The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network*, IEEE Trans. Inf. Theory, 1998, 44(2), 525–536.

[2] A. Benabid, J. Fan and D. Xiang, *Comparison theorems on large-margin learning*, Int. J. Wavelets Multiresolution Inf. Process., 2021, 19(05), 2150015 (18 pages).

[3] B. E. Boser, I. M. Guyon and V. N. Vapnik, *A training algorithm for optimal margin classifiers*, in Proceedings of the fifth annual workshop on Computational learning theory, 1992, 144–152.

[4] D. Chen, Q. Wu, Y. Ying and D. Zhou, *Support vector machine soft margin classifiers: error analysis*, J. Mach. Learn. Res., 2004, 5, 1143–1175.

[5] C. Cortes and V. Vapnik, *Support-vector networks*, Machine learning, 1995, 20, 273–297.

[6] F. Cucker and D. Zhou, *Learning Theory: An Approximation Theory Viewpoint*, Cambridge University Press, 2007.

[7] D. E. Edmunds and H. Triebel, *Function spaces, entropy numbers, differential operators*, 180, Cambridge University Press, 1996.

[8] J. Fan and D. Xiang, *Quantitative convergence analysis of kernel based large-margin unified machines*, Commun. Pure Appl. Anal., 2020, 19(8), 4069–4083.

[9] Y. Feng, J. Fan and J. Suykens, *A statistical learning approach to modal regression*, J. Mach. Learn. Res., 2020, 21(2), 1–35.

[10] Q. Guo and P. Ye, *Error analysis of least-squares l^q-regularized regression learning algorithm with the non-identical and dependent samples*, IEEE Access, 2018, 6, 43824–43829.

[11] X. Guo, T. Hu, Q. Wu, et al., *Distributed minimum error entropy algorithms.*, J. Mach. Learn. Res., 2020, 21, 1–31.

[12] X. Guo, L. Li and Q. Wu, *Modeling interactive components by coordinate kernel polynomial models*, Math. Fund. Computing, 2020, 3(4), 263–277.

[13] Z. Guo and L. Shi, *Classification with non-i.i.d. sampling*, Math. Comput. Modell, 2011, 54(5–6), 1347–1364.

[14] T. Hastie, R. Tibshirani and J. Friedman, *The elements of statistical learning: Data Mining, Inference, and Prediction*, Springer-Verlag, New York, 2001.

[15] A. Khaleghi and G. Lugosi, *Inferring the mixing properties of an ergodic process*, arXiv: 2106.07054, 2021.

[16] Y. Liu, H. Zhang and Y. Wu, *Hard or soft classification? large-margin unified machines*, J. Am. Stat. Assoc., 2011, 106(493), 166–177.

[17] J. S. Marron, M. J. Todd and J. Ahn, *Distance-weighted discrimination*, Journal of the American Statistical Association, 2007, 102(480), 1267–1271.

[18] L. Peng, Y. Zhu and W. Zhong, *Lasso regression in sparse linear model with φ-mixing errors*, Metrika, 2022, 1–26.

[19] S. Smale and D. Zhou, *Online learning with markov sampling*, Anal. Appl., 2009, 7(01), 87–113.

[20] I. Steinwart and A. Christmann, *Fast learning from non-i.i.d. observations*, Adv. Neural Inf. Process. Syst., 2009, 22, 1–9.

[21] I. Steinwart and A. Christmann, *Estimating conditional quantiles with the help of the pinball loss*, Bernoulli, 2011, 17(1), 211–225.

[22] I. Steinwart and C. Scovel, *Fast rates for support vector machines using gaussian kernels*, Ann. Stat., 2007, 35(2), 575–607.

[23] H. Sun and Q. Wu, *Regularized least square regression with dependent samples*, Adv. Comput. Math., 2010, 32(2), 175–189.

[24] R. C. Williamson, A. J. Smola and B. Scholkopf, *Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators*, IEEE Trans. Inform. Theory, 2001, 47(6), 2516–2532.

[25] K. Wong, Z. Li and A. Tewari, *Lasso guarantees for β-mixing heavy-tailed time series*, Ann. Stat., 2020, 48(2), 1124–1142.

[26] Q. Wu, Y. Ying and D. Zhou, *Learning rates of least-square regularized regression*, Found. Comput. Math., 2006, 6(2), 171–192.

[27] Q. Wu, Y. Ying and D. Zhou, *Multi-kernel regularized classifiers*, J. Complex., 2007, 23(1), 108–134.

[28] D. Xiang, *Logistic classification with varying gaussians*, Comput. Math. Appl., 2011, 61(2), 397–407.

[29] D. Xiang, *Conditional quantiles with varying gaussians*, Adv. Comput. Math., 2013, 38(4), 723–735.

[30] D. Xiang and D. Zhou, *Classification with gaussians and convex loss*, J. Mach. Learn. Res., 2009, 10, 1447–1468.

[31] Y. Xu and D. Chen, *Learning rates of regularized regression for exponentially strongly mixing sequence*, J. Stat. Plan. Inference, 2008, 138(7), 2180–2189.

[32] B. Yu, *Rates of convergence for empirical processes of stationary mixing sequences*, Ann. Probab., 1994, 22, 94–116.

[33] D. Zhou, *The covering number in learning theory*, J. Complex., 2002, 18(3), 739–767.

[34] D. Zhou, *Capacity of reproducing kernel spaces in learning theory*, IEEE Trans. Inform. Theory, 2003, 49(7), 1743–1752.