Quantitative convergence analysis of kernel based large-margin unified machines

Fan, Jun; Xiang, Dao Hong

Published in:
Communications on Pure and Applied Analysis

DOI:
10.3934/cpaa.2020180

Published: 01/08/2020

Document Version:
Peer reviewed version

Quantitative Convergence Analysis of Kernel Based Large-margin Unified Machines

Jun Fan[1], **Dao-Hong Xiang**[2 †]

[1] Department of Mathematics, Hong Kong Baptist University, Kowloon, Hong Kong, China

[2] Department of Mathematics, Zhejiang Normal University, Jinhua, Zhejiang 321004, China

Abstract

High-dimensional binary classification has been intensively studied in the community of machine learning in the last few decades. Support vector machine (SVM), one of the most popular classifier, depends on only a portion of training samples called support vectors which leads to suboptimal performance in the setting of high dimension and low sample size (HDLSS). Large-margin unified machines (LUMs) are a family of margin-based classifiers proposed to solve the so-called "data piling" problem which is inherent in SVM under HDLSS settings. In this paper we study the binary classification algorithms associated with LUM loss functions in the framework of reproducing kernel Hilbert spaces. Quantitative convergence analysis has been carried out for these algorithms by means of a novel application of projection operators to overcome the technical difficulty. The rates are explicitly derived under priori conditions on approximation and capacity of the reproducing kernel Hilbert space.

Keywords: LUMs, convergence rates, kernel methods, regularization, projection operator

AMS Subject Classification Numbers. 68Q32, 62J02, 41A46

1 Introduction

In this paper we consider large-margin unified machines (LUMs) for binary classification problems and investigate the consistency of kernel based LUMs within the framework of learning theory.

Let the input space X be a compact domain of $\mathbb{R}^d$ and the output space $Y = \{-1, 1\}$ representing the two classes. We assume that a sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \in Z^m$ is generated by a probability

†Corresponding author: Dao-Hong Xiang, Email: daohongxiang@zjnu.cn.

The work by J. Fan is partially supported by the Hong Kong RGC ECS grant 22303518, HKBU FRG grant FRG2/17-18/091 and the NSF grant of China (No. 11801478). The work by D. H. Xiang is supported by the National Natural Science Foundation of China under Grant 11871438 and 11771120.

measure P on $Z := X \times Y$ in an i.i.d.fashion. The learning target in binary classification is to find a classifier $\mathcal{C} : X \to Y$ based on the sample data such that for a new observation (x, y) we have $\mathcal{C}(x) = y$ with high probability. We define $\mathcal{R}(\mathcal{C}) = \text{Prob}\{\mathcal{C}(x) \neq y\} = \int_X P(y \neq \mathcal{C}(x)|x)dP_X$ as the misclassification error which is used to measure the prediction power of a classifier $\mathcal{C}$. Here P_X is the marginal distribution of P on X and $P(y|x)$ is the conditional distribution at $x \in X$. The classifier minimizing the misclassification error is called the Bayes rule f_c defined as $f_c(x) = 1$ if $P(y = 1|x) \geq P(y = -1|x)$, and $f_c(x) = -1$ otherwise.

The classifiers considered here are induced by real-valued functions $f : X \to \mathbb{R}$ as $\mathcal{C}_f$ defined by $\text{sgn}(f)(x) = 1$ if $f(x) \geq 0$ and $\text{sgn}(f)(x) = -1$ otherwise. The real-valued functions are generated from different classification algorithms. By definition of the classification rule, it is clear that correct classification occurs if and only if $yf(x) > 0$. The quantity $yf(x)$ is referred as the functional margin and it plays an essential role in large-margin classification algorithms.

Among various margin-based methods, support vector machine (SVM) [4, 6] is the most well-known one. SVM falls into the group of the so-called hard classification since it tends to directly estimate the classification boundary. In addition, Boosting [10, 11] is also a typical hard classification method. Differing from the hard classification, soft classification rule aims at estimating the class conditional probabilities explicitly and then predicting the class based on the largest estimated probability. Fisher linear discriminant analysis and logistic regression are two typical soft classification methods [12]. These two kinds of methods are based on different philosophiles and each has its own merits. In a recent work [14], Liu and his coauthors proposed a unified framework of large-margin classifiers, namely large-margin unified machines (LUMs), which offers a unique transition from hard to soft classifiers. In addition, it was pointed out in [15] that the SVM may suffer from "data piling" phenomena for high dimension low sample size (HDLSS) data, that is, a large portion of data points are support vectors and they will pile up on top of each other while projected onto the normal vector of the separating hyperplane. Data piling is usually not desirable for a classifier since it may reduce generalizability and lead to suboptimal performance in the HDLSS setting. See [15] for real data example and more details. To solve the data piling problem, a large-margin classifier called distance-weighted discrimination (DWD) was proposed therein. Note that DWD is a special case of LUMs. The corresponding LUM loss functions are defined as follows:

Definition 1. *For given $p \geq 0$ and $q > 0$, define LUM loss functions as*

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

Here p and q play different roles. The parameter p controls the connecting point between the two pieces of the loss function as well as the shape of the right piece. The parameter q determines

the decaying speed for the right piece. The LUM loss defined in (1.1) represents a large family of loss functions. Many often used loss functions fall into LUM form. For example, if $p \to \infty$ and $q > 0$, the LUM loss reduces to $V^{(h)}(t) = (1-t)_+$, the hinge loss widely used in support vector machine (SVM). If $p = 1$ and $q = 1$, the LUM loss reduces to the DWD loss function proposed in [24] as follows:

$$V^{(D)}(t) = \begin{cases} 1-t, & \text{if } t < \frac{1}{2}, \\ \frac{1}{4t}, & \text{if } t \geq \frac{1}{2}, \end{cases}$$

If $p = q > 0$, then

$$V_q^{(D)}(t) = \begin{cases} 1-t, & \text{if } t < \frac{q}{q+1}, \\ \frac{1}{t^q}\frac{q^q}{(q+1)^{q+1}} & \text{if } t \geq \frac{q}{q+1} \end{cases}$$

is the so called generalized DWD loss defined in [24]. If $q \to \infty$ and $p \geq 0$, the LUM loss reduces to

$$V^{(he)}(t) = \begin{cases} 1-t, & \text{if } t < \frac{p}{1+p}, \\ \frac{1}{1+p}\exp\{-[(1+p)t-p]\}, & \text{if } t \geq \frac{p}{1+p}, \end{cases}$$

which is a hybrid of SVM and AdaBoost [11, 14]. In particular, when $q \to \infty$ and $p = 0$,

$$V^{(he)}(t) = \begin{cases} 1-t, & \text{if } t < 0, \\ e^{-t}, & \text{if } t \geq 0 \end{cases}$$

is the combination of the hinge loss and exponential loss.

We say that $K : X \times X \to \mathbb{R}$ is a Mercer kernel if it is continuous, symmetric and positive semidefinite in the sense that the matrix $(K(x_i, x_j))_{i,j=1}^l$ is positive semidefinite for any $\{x_1, \cdots, x_l\} \subset X$. The reproducing kernel Hilbert space (RKHS) (see [1]) $\mathcal{H}_K$ associated with the kernel K is defined to be the completion of the linear span of the set of functions $\{K_x = K(x, \cdot) : x \in X\}$ with the inner product $\langle \cdot, \cdot \rangle_K$ given by $\langle K_x, K_y \rangle_K = K(x, y)$. Denote $\kappa = \sup_{x \in X} \sqrt{K(x, x)}$. RKHS has the reproducing property

$$\langle K_x, f \rangle = f(x), \ x \in X, f \in \mathcal{H}_K. \tag{1.2}$$

With the LUM loss V and an RKHS $\mathcal{H}_K$, the kernel based LUMs can be formulated as the following regularization scheme [22]:

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

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

In [14], the Fisher consistency associated with LUM loss functions is provided. Later on, [24] formulates a kernel DWD approach in a reproducing kernel Hilbert space and further establishes the Bayes risk consistency of the kernel based DWD. To the best of our knowledge, there is no any

quantitative convergence analysis for the kernel based LUMs, even not for the kernel based DWD, except for the classical support vector machine associated with the hinge loss, which has been already well studied in a large literature. See [37, 5, 30, 34, 36, 18, 20] and references therein. The purpose of this paper is to provide quantitative convergence analysis for the kernel based LUMs, i.e., to estimate the excess misclassification error $\mathcal{R}(\text{sgn}(f_{\mathbf{z}})) - \mathcal{R}(f_c)$ as $m \to \infty$.

2 Key Properties and Main Results

2.1 Properties of LUM Loss

In this section we give some discussions on LUM loss. From the definition of the LUM loss, it is easy to verify the following property.

Lemma 1. *Let V be the LUM loss functions defined in (1.1), then $V(\cdot)$ is a convex function. It is differentiable everywhere for $0 \le p < \infty$. Moreover, $V(\cdot)$ has the smallest zero at 1 when $p \to \infty$. For $0 \le p < \infty$, the loss functions $V(\cdot)$ have no zero.*

The following decay property of the LUM loss is required in our error analysis, which is proved in Appendix.

Lemma 2. *Let V be the LUM loss functions with $0 \le p < \infty$ and $0 < q < \infty$. For $t \ge \frac{p}{1+p}$, it holds*

$$V(t) \le C_{p,q} t^{-q} \tag{2.1}$$

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

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

$$\mathcal{E}(f) = \int_{\mathcal{Z}} V(yf(x))\, dP(x,y) = \int_X \int_Y V(yf(x)) dP(y|x)\, dP_X(x). \tag{2.2}$$

Let $\eta(x) = P(y=1|x)$. It was shown in [14] that the minimizer f_P of $\mathcal{E}(f)$ over all measurable functions for $0 < q < \infty$ and $0 \le p < \infty$ is defined by

$$f_P(x) = \begin{cases} -\frac{1}{1+p}(R(x)^{-1}q - q + p), & \text{if } 0 \le \eta(x) < \frac{1}{2}, \\ \frac{1}{1+p}(R(x)q - q + p), & \text{if } \frac{1}{2} \le \eta(x) \le 1, \end{cases} \tag{2.3}$$

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

$$f_P(x) = \begin{cases} -1, & \text{if } 0 \le \eta(x) < \frac{1}{2}, \\ 1 & \text{if } \frac{1}{2} \le \eta(x) \le 1. \end{cases} \tag{2.4}$$

For $q \to \infty$, the LUM loss reduces to the hybrid loss of the hinge loss and exponential loss with the minimizer below:

$$f_P(x) = \begin{cases} \frac{1}{1+p}\big[\ln \frac{\eta(x)}{1-\eta(x)} - p\big], & \text{if } 0 \le \eta(x) < \frac{1}{2}, \\ \frac{1}{1+p}\big[\ln \frac{\eta(x)}{1-\eta(x)} + p\big] & \text{if } \frac{1}{2} \le \eta(x) \le 1. \end{cases} \tag{2.5}$$

Lemma 3. *Let V be the LUM loss functions. If $p \to \infty$, f_P defined in (2.4) is bounded by 1. If $0 \le p < \infty$, f_P defined in (2.3) and (2.5) are finite if and only if $\eta(x) \in (0, 1)$.*

The excess misclassification error $\mathcal{R}(\text{sgn}(f)) - \mathcal{R}(f_c)$ for the classifier $\text{sgn}(f)$ can be bounded by means of the excess generalization error $\mathcal{E}(f) - \mathcal{E}(f_P)$ according to some comparison theorems. A comparison theorem was first proved in [37] for the hinge loss $V^{(h)}(t) = (1-t)_+$. For a general convex loss with zero point, the comparison theorems have been studied in [3, 5, 20, 29, 31].

Recently, [9] gives a complete study for the comparison theorems of the LUM loss functions stated in the following lemma.

Lemma 4. *(i) Let V be the LUM loss functions with $0 < p < \infty$ and $0 < q \le \infty$. For any probability measure P, any measurable function $f : X \to \mathbb{R}$, and some constant $C_p > 0$, it holds*

$$\mathcal{R}\big(sgn(f)\big) - \mathcal{R}(f_c) \le C_p\{\mathcal{E}(f) - \mathcal{E}(f_P)\}. \tag{2.6}$$

(ii) Let V be the LUM loss with $p = 0$. For any probability measure P, any measurable function $f : X \to \mathbb{R}$, and some constant $C_q > 0$, it holds

$$\mathcal{R}\big(sgn(f)\big) - \mathcal{R}(f_c) \le C_q\sqrt{\mathcal{E}(f) - \mathcal{E}(f_P)}. \tag{2.7}$$

Obviously, (2.7) for the case of $p = 0$ is worse than (2.6) for the case of $p > 0$. It may be improved by introducing the following Tsybakov noise condition [21] on the probability measure P.

Definition 2. *Let $0 \le \tau \le \infty$. We say that probability measure P satisfies a Tsybakov noise condition with exponent τ if there exists a constant C_τ such that*

$$P_X(\{x \in X : |2\eta(x) - 1| \le C_\tau t\}) \le t^\tau, \forall t > 0. \tag{2.8}$$

Lemma 5. *Let V be the LUM loss with $p = 0$. Under the assumption (2.8) with $0 \le \tau \le \infty$, the following comparison theorem holds true with some constant $C_{q,\tau} > 0$:*

$$\mathcal{R}\big(sgn(f)\big) - \mathcal{R}(f_c) \le C_{q,\tau}(\mathcal{E}(f) - \mathcal{E}(f_P))^{\frac{\tau+1}{\tau+2}}. \tag{2.9}$$

2.2 Projection Operator and Error Decomposition

We notice the fact that when $0 \le p < \infty$, the LUM loss V has no zero on $\mathbb{R}$, which leads to an unbounded target function f_P. It causes some difficulties in our analysis. In particular, the

projection technique and concentration inequality used for SVM cannot be directly applied here. To overcome this difficulty, we introduce a different projection operator π_M defined below:

Definition 3. *For any $M > 0$, the projection operator π_M on the space of function on X is defined by*

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

In [32], the similar projection operator was proposed to analyze the binary classification with logistic loss. Since the LUM loss without zero leads to an unbounded target function f_P, 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 paper differs from the original one introduced for classifying loss with zero (see e.g., [5, 34, 19, 33] for details).

Obviously, $\mathcal{R}\big(\operatorname{sgn}(f_{\mathbf{z}})\big) - \mathcal{R}(f_c) = \mathcal{R}\big(\operatorname{sgn}(\pi_M(f_{\mathbf{z}}))\big) - \mathcal{R}(f_c)$ since $\operatorname{sgn}(f_{\mathbf{z}}) = \operatorname{sgn}(\pi_M(f_{\mathbf{z}}))$. Therefore, we turn to estimate the excess misclassification error $\mathcal{R}\big(\operatorname{sgn}(\pi_M(f_{\mathbf{z}}))\big) - \mathcal{R}(f_c)$ with confidence as the sample size $m \to \infty$. This can be done by bounding the excess generalization error $\mathcal{E}(\pi_M(f_{\mathbf{z}})) - \mathcal{E}(f_P)$ because of the comparison theorems stated in (2.6) and (2.9). Define the empirical error associated with the loss V as

$$\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{m}\sum_{i=1}^{m} V(y_i f(x_i)) \quad \text{for } f : X \to \mathbb{R}.$$

Denote the sample free version of (1.3) by

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

where $\mathcal{E}(f)$ is defined in (2.2). Then we conduct an error decomposition on the excess generalization error $\mathcal{E}(\pi_M(f_{\mathbf{z}})) - \mathcal{E}(f_P)$ in the following lemma, which is a direct corollary of Propositions 5.4 and 5.6 in [26]. We include the proof in the Appendix for completeness.

Lemma 6. *Let V be the LUM loss with $0 \le p < \infty$ and $0 < q \le \infty$. Let $M > 0$, $f_\lambda \in \mathcal{H}_K$, and $f_{\mathbf{z}}$ be defined in (1.3). Then*

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

where the quantity $\mathcal{S}_{\mathbf{z}}(f)$ is defined for $f \in C(X)$ by

$$\mathcal{S}_{\mathbf{z}}(f) = [\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(\pi_M(f_P))] - [\mathcal{E}(f) - \mathcal{E}(\pi_M(f_P))],$$

and

$$\mathcal{D}(\lambda) := \mathcal{E}(f_\lambda) - \mathcal{E}(f_P) + \lambda \|f_\lambda\|_{\mathcal{H}_K}^2. \tag{2.10}$$

In Lemma 6, we call $\mathcal{S}_{\mathbf{z}}(f)$ sample error and $\mathcal{D}(\lambda)$ regularization error which is independent of sample. Notice that the main difference of Lemma 6 from the error decomposition in the literature [5, 34, 19, 33] is the appearance of $V(M)$. The reason is that $V(t)$ without zero is a strictly decreasing and positive function.

2.3 Main Results

To state our main results, we need the capacity of the hypothesis space measured by covering numbers and the bound of the regularization error.

Definition 4. (Uniform covering number) *For a subset S of $C(X)$ 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 numbers of unit balls of classical function spaces have been well studied in the literature (see e.g. [8, 2, 25, 39, 40]). Here we need the covering numbers of the balls of the RKHS $\mathcal{H}_K$. Denote $B_R = \{f : \|f\|_{\mathcal{H}_K} \leq R\}$. Estimating uniform convergence in terms of covering numbers has been well developed in learning theory, e.g. [7, 5, 27].

To derive the explicit convergence rates for the kernel based LUMs, we impose some assumptions on the covering number and the regularization error.

Assumption 2.1. *Assume for some $s > 0$ and $C_s > 0$ that*

$$\log \mathcal{N}(B_1, u) \leq C_s \Big(\frac{1}{u}\Big)^s, \forall u > 0. \tag{2.11}$$

Assumption 2.2. *Assume that for some constant $C_r > 0$, the regularization error satisfies*

$$\mathcal{D}(\lambda) \leq C_r \lambda^r, \ 0 < r \leq 1. \tag{2.12}$$

Let us illustrate our main result by the following special case which will be proved in Section 3.

Theorem 1. *Let V be the LUM loss with $0 \leq p < \infty$ and $0 < q < \infty$. Assume $X \subset \mathbb{R}^n$ and $K \in C^\infty(X \times X)$. Suppose Assumption 2.2 holds for $r = 1$. Take $\lambda = m^{-\frac{q}{2(q+1)}}$ and $M = m^{\frac{1}{2(q+1)}}$. Let $0 < \eta < \frac{q}{2(q+1)}$.*

(i) If $0 < p < \infty$, for any $0 < \delta < 1$, with confidence $1 - \delta$, there holds

$$\mathcal{R}\big(sgn(f_{\mathbf{z}})\big) - \mathcal{R}(f_c) \leq \tilde{C}_1 \log \frac{4}{\delta} m^{-(q/2(q+1)-\eta)}, \tag{2.13}$$

where $\tilde{C}_1$ is a constant independent of m or δ.

(*ii*) *If $p = 0$, for any probability measure P satisfying (2.8) with $0 < \tau \le \infty$, for any $0 < \delta < 1$, with confidence $1 - \delta$, there holds*

$$\mathcal{R}\big(sgn(f_{\mathbf{z}})\big) - \mathcal{R}(f_c) \le \tilde{C}_2 \log \frac{4}{\delta} m^{-\frac{\tau+1}{\tau+2}(q/2(q+1)-\eta)}, \tag{2.14}$$

where $\tilde{C}_2$ is a constant independent of m or δ.

Note that (2.12) holds with $r = 1$ if the target function $f_P \in \mathcal{H}_K$. See [17, 16, 35] and references therein for more discussions on the regularization error $\mathcal{D}(\lambda)$.

If $q > (1/2\eta) - 1$ for $0 < \eta < 1/2$, we see that the power index for the learning rates (2.13) is at least $(1/2) - 2\eta$. This index can be arbitrarily close to $1/2$ when η is small enough.

If we take $\tau = \infty$ leading to $\frac{\tau+1}{\tau+2} = 1$, Theorem 1 also provides the same learning rates $m^{-((1/2)-2\eta)}$ for the case of $p = 0$.

The next theorem gives a convergence result under general assumptions.

Theorem 2. *Let V be the LUM loss with $0 \le p < \infty$ and $0 < q < \infty$. Suppose Assumptions 2.1 and 2.2 hold for $s > 0$ and $0 < r \le 1$. Take $\lambda = m^{-\alpha}$ with $0 < \alpha \le 1$ and $\alpha < \frac{4q}{3s(q+1)}$. Set $M = m^{\beta}$ with $\beta = \frac{1}{2(q+1)}$. Let*

$$0 < \eta < \frac{(2+s)[2q - s\alpha(q+1)]}{s(4+s)(q+1)}. \tag{2.15}$$

(*i*) *If $0 < p < \infty$, for any $0 < \delta < 1$, with confidence $1 - \delta$, there holds*

$$\mathcal{R}\big(sgn(f_{\mathbf{z}})\big) - \mathcal{R}(f_c) \le \tilde{C}_1 \Big(\log \frac{2}{\eta} \Big)^2 \log \frac{4}{\delta} m^{-\vartheta}, \tag{2.16}$$

where $\tilde{C}_1$ is a constant independent of m or δ and the power index ϑ is given in terms of r, s, q, α and η by

$$\begin{aligned} \vartheta = \min\Big\{ & \alpha r,\ \frac{\alpha(r-1)+1}{2},\ \frac{q}{2(q+1)},\ \frac{q}{(2+s)(q+1)} - \frac{s\alpha(1-r)}{2(2+s)}, \\ & \frac{q}{(2+s)(q+1)} - \frac{s}{2+s}\Big(\frac{\alpha}{2} - \frac{q}{4(q+1)}\Big),\ \frac{q}{(2+s)(q+1)} - \frac{s[\alpha(3-r)-1]}{4(2+s)}, \\ & \frac{q}{(2+s)(q+1)} - \frac{s[\alpha(2+s)(q+1)-q]}{(2+s)(4+s)(q+1)} - \frac{s}{2+s}\eta \Big\}. \end{aligned} \tag{2.17}$$

(*ii*) *If $p = 0$, for any probability measure P satisfying (2.8) with $0 < \tau \le \infty$, for any $0 < \delta < 1$, with confidence $1 - \delta$, there holds*

$$\mathcal{R}\big(sgn(f_{\mathbf{z}})\big) - \mathcal{R}(f_c) \le \tilde{C}_2 \Big(\log \frac{2}{\eta} \Big)^2 \log \frac{4}{\delta} m^{-\frac{\tau+1}{\tau+2}\vartheta}, \tag{2.18}$$

where $\tilde{C}_2$ is a constant independent of m or δ and the power index ϑ is given in (2.17).

Note that the error bound obtained in Theorem 5.7 of [26] also applies to loss functions without zero under assumptions on the variance-expectation bound and $\|f_P\|_\infty < \infty$. In this paper we consider unbounded target function f_P and it will be of great interest to verify whether the variance-expectation bound holds for LUM loss functions in order to improve the learning rates.

The index ϑ can be viewed as a functions of parameters r, s, q, α, η. The restrictions $0 < \alpha < \frac{4q}{3s(q+1)}$ on α and (2.15) on η ensure that ϑ is positive, which verifies the valid learning rate in Theorem 2.

Assumption 2.1 is a measurement of regularity of the kernel K when X is a subset of $\mathbb{R}^n$. In particular, s can be arbitrarily small when K is smooth enough. In this case, the power index ϑ in (2.17) can be arbitrarily close to $\min\{\alpha r, \ \frac{\alpha(r-1)+1}{2}, \ \frac{q}{2(q+1)}\}$.

The convergence results can be extended to the case of $q = \infty$ by noting the exponential decay on the right tail of $V^{(he)}(t)$, which is beyond the scope of this paper.

3 Error Analysis

3.1 Sample Error

We are now in the position to estimate the sample error $\mathcal{S}_{\mathbf{z}}(f_\lambda)$ and $\mathcal{S}_{\mathbf{z}}(\pi_M(f_{\mathbf{z}}))$ defined in Lemma 6 by the following Hoeffding inequality and covering numbers.

Lemma 7. *Let ξ be a random variable on a probability space Z with mean $\mathbf{E}(\xi) = \mu$, and satisfy $|\xi - \mathbf{E}(\xi)| \le B$ for almost all $z \in Z$. Then for all $\epsilon > 0$,*

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\Big\{\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mu \ge \epsilon\Big\} \le \exp\Big\{-\frac{m\epsilon^2}{2B^2}\Big\}.$$

For $R \ge 1$, let $\mathcal{W}(R)$ be the subset of Z^m defined by

$$\mathcal{W}(R) = \{\mathbf{z} \in Z^m : \|f_{\mathbf{z}}\|_{\mathcal{H}_K} \le R\}. \tag{3.1}$$

It follows from (2.10) and (2.12) that

$$\|f_\lambda\|_{\mathcal{H}_K} \le \sqrt{\frac{\mathcal{D}(\lambda)}{\lambda}} \le \sqrt{C_r}\lambda^{\frac{r-1}{2}}, \quad r > 0. \tag{3.2}$$

Proposition 1. *Let V be the LUM loss with $0 \le p < \infty$ and $0 < q < \infty$. Suppose Assumptions 2.1 and 2.2 hold for $s > 0$ and $0 < r \le 1$. Let $R \ge 1, M \ge 1$, and $0 < \delta < 1$. Then there exists a*

subset $\mathcal{V}_R$ of Z^m with measure at most δ such that for any $\mathbf{z} \in \mathcal{W}(R)\backslash\mathcal{V}_R$,

$$\begin{aligned}\mathcal{E}(\pi_M(f_{\mathbf{z}})) - \mathcal{E}(f_P) + \lambda\|f_{\mathbf{z}}\|^2_{\mathcal{H}_K} \quad &\leq \quad C_r\lambda^r + C_{p,q}M^{-q} + (2\sqrt{2}+8)m^{-\frac{1}{2}}\log\frac{2}{\delta} \\ &\quad +(\sqrt{2}+8)Mm^{-\frac{1}{2}}\log\frac{2}{\delta} + \kappa\sqrt{2C_r}\lambda^{\frac{r-1}{2}}m^{-\frac{1}{2}}\log\frac{2}{\delta} \quad (3.3) \\ &\quad +8\sqrt{C_s}4^{\frac{1}{2+s}}M^{\frac{2}{2+s}}m^{-\frac{1}{2+s}}R^{\frac{s}{2+s}}.\end{aligned}$$

Proof. Step 1. Let us first estimate the quantity $\mathcal{S}_{\mathbf{z}}(f_\lambda)$ which can be expressed as $\frac{1}{m}\sum_{i=1}^m \xi(z_i) - \mathbf{E}(\xi)$ with a single random variable $\xi(z) = V(yf_\lambda(x)) - V(y\pi_M(f_P(x)))$ on (Z,P). It satisfies $-V(-M) \leq \xi \leq V(-\|f_\lambda\|_{L^\infty(X)})$. Hence $|\xi - \mathbf{E}(\xi)| \leq V(-\|f_\lambda\|_{L^\infty(X)}) + V(-M)$. Applying Lemma 7 to the above random variable, we know that

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\Big\{\frac{1}{m}\sum_{i=1}^m \xi(z_i) - \mathbf{E}(\xi) \geq \epsilon\Big\} \leq \exp\Big\{-\frac{m\epsilon^2}{2(V(-\|f_\lambda\|_{L^\infty(X)}) + V(-M))^2}\Big\} \qquad \forall \epsilon > 0.$$

Solving the equation for ϵ given by

$$\frac{m\epsilon^2}{2(V(-\|f_\lambda\|_{L^\infty(X)}) + V(-M))^2} = \log\frac{2}{\delta},$$

we have that there exists a subset $Z_{1,\delta}$ of Z^m with confidence at most $\delta/2$ such that

$$\frac{1}{m}\sum_{i=1}^m \xi(z_i) - \mathbf{E}(\xi) \leq \frac{\sqrt{2}\Big(V(-\|f_\lambda\|_{L^\infty(X)}) + V(-M)\Big)}{\sqrt{m}}\log\frac{2}{\delta}, \quad \forall \mathbf{z} \in Z^m\backslash Z_{1,\delta}.$$

Hence, for any $0 < \delta < 1$, with confidence $1 - \frac{\delta}{2}$, the term $\mathcal{S}_{\mathbf{z}}(f_\lambda)$ can be bounded as

$$\mathcal{S}_{\mathbf{z}}(f_\lambda) \leq \frac{\sqrt{2}\Big(V(-\|f_\lambda\|_{L^\infty(X)}) + V(-M)\Big)}{\sqrt{m}}\log\frac{2}{\delta}, \quad \forall \mathbf{z} \in Z^m\backslash Z_{1,\delta}.$$

Step 2. Next, we estimate the term $-\mathcal{S}_{\mathbf{z}}(\pi_M(f_{\mathbf{z}}))$ which involves the function $f_{\mathbf{z}}$. Here $f_{\mathbf{z}}$ runs over a set of functions since $\mathbf{z}$ is a random sample itself. To estimate it, we use a standard argument (see e.g. [7]) with Hoeffding inequality and covering numbers.

Let $J = \mathcal{N}\Big(B_R, \frac{\epsilon}{4}\Big)$. Then there exists a set of $\{f_j\}_{j=1}^J \subset B_R$, such that B_R is covered by balls $B^{(j)}$ centered at f_j with radius $\frac{\epsilon}{4}$.

Let $j \in \{1, \cdots, J\}$. The random variable $\xi(z) = V(y\pi_M(f_j)(x)) - V(y\pi_M(f_P)(x))$ satisfies $-V(-M) \leq \xi \leq V(-M)$, which implies $|\xi - \mathbf{E}(\xi)| \leq 2V(-M)$. Therefore, the one-side Hoeffding inequality implies that

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\Big\{-\mathcal{S}_{\mathbf{z}}(\pi_M(f_j)) \geq \frac{\epsilon}{2}\Big\} = \mathrm{Prob}_{\mathbf{z}\in Z^m}\Big\{\mathcal{S}_{\mathbf{z}}(\pi_M(f_j)) \leq -\frac{\epsilon}{2}\Big\} \leq \exp\Big\{-\frac{m\epsilon^2}{32(V(-M))^2}\Big\}.$$

Note that, for each $f \in B^{(j)}$,

$$\begin{aligned}&\Big|-\mathcal{S}_{\mathbf{z}}(\pi_M(f)) + \mathcal{S}_{\mathbf{z}}(\pi_M(f_j))\Big| = \Big|\mathcal{E}(\pi_M(f)) - \mathcal{E}(\pi_M(f_j)) - [\mathcal{E}_{\mathbf{z}}(\pi_M(f)) - \mathcal{E}_{\mathbf{z}}(\pi_M(f_j))]\Big| \\ &\leq 2\|f - f_j\|_{C(X)} \leq 2 \times \frac{\epsilon}{4} = \frac{\epsilon}{2}.\end{aligned}$$

Hence

$$\sup_{f\in B^{(j)}}\{-\mathcal{S}_{\mathbf{z}}(\pi_M(f))\}\geq\epsilon \quad\Rightarrow\quad -\mathcal{S}_{\mathbf{z}}(\pi_M(f_j))\geq\frac{\epsilon}{2}.$$

Therefore, we conclude that

$$\begin{aligned}\mathrm{Prob}_{\mathbf{z}\in Z^m}\Big\{\sup_{f\in B_R}\{-\mathcal{S}_{\mathbf{z}}(\pi_M(f))\}\geq\epsilon\Big\} &\leq \sum_{j=1}^{J}\mathrm{Prob}_{\mathbf{z}\in Z^m}\Big\{\sup_{f\in B^{(j)}}\{-\mathcal{S}_{\mathbf{z}}(\pi_M(f))\}\geq\epsilon\Big\}\\ &\leq J\exp\Big\{-\frac{m\epsilon^2}{32(V(-M))^2}\Big\}.\end{aligned}$$

Denote $\epsilon^*(m,R,M,\delta/2)$ as the positive solution to the equation

$$C_s\Big(\frac{4R}{\epsilon}\Big)^s-\frac{m\epsilon^2}{32(V(-M))^2}=\log\frac{\delta}{2},$$

which can be expressed as

$$\epsilon^{2+s}-\frac{32(V(-M))^2}{m}\log\frac{2}{\delta}\epsilon^s-\frac{2^{2s+5}C_s(V(-M))^2}{m}R^s=0$$

By Lemme 7.2 in [7], the positive solution $\epsilon^*(m,R,M,\delta/2)$ of this equation can be bounded as

$$\begin{aligned}\epsilon^*(m,R,M,\delta/2) &\leq \max\Big\{\frac{8V(-M)}{\sqrt{m}}\log\frac{2}{\delta},8\sqrt{C_s}\Big(\frac{(V(-M))^2}{m}\Big)^{\frac{1}{2+s}}R^{\frac{s}{2+s}}\Big\}\\ &\leq \frac{8V(-M)}{\sqrt{m}}\log\frac{2}{\delta}+8\sqrt{C_s}\Big(\frac{(V(-M))^2}{m}\Big)^{\frac{1}{2+s}}R^{\frac{s}{2+s}}.\end{aligned}$$

Thus, there exists a second subset $Z_{2,\delta}$ of Z^m with measure at most $\frac{\delta}{2}$ such that

$$-\mathcal{S}_{\mathbf{z}}(\pi_M(f_{\mathbf{z}}))\leq\frac{8V(-M)}{\sqrt{m}}\log\frac{2}{\delta}+8\sqrt{C_s}\Big(\frac{(V(-M))^2}{m}\Big)^{\frac{1}{2+s}}R^{\frac{s}{2+s}},\quad\forall\mathbf{z}\in\mathcal{W}(R)\backslash Z_{2,\delta}.$$

Step 3. Combining above two steps, we estimate the total error $\mathcal{E}(\pi_M(f_{\mathbf{z}}))-\mathcal{E}(f_P)+\lambda\|f_{\mathbf{z}}\|^2_{\mathcal{H}_K}$. Notice the fact that the measure of the set $\mathcal{V}_R:=Z_{1,\delta}\cup Z_{2,\delta}$ is at most δ, hence, for $\mathbf{z}\in\mathcal{W}(R)\backslash\mathcal{V}_R$, we get that

$$\begin{aligned}\mathcal{E}(\pi_M(f_{\mathbf{z}}))-\mathcal{E}(f_P)+\lambda\|f_{\mathbf{z}}\|^2_{\mathcal{H}_K} &\leq \mathcal{D}(\lambda)+V(M)+\frac{\sqrt{2}\Big(V(-\|f_\lambda\|_{L^\infty(X)})+V(-M)\Big)}{\sqrt{m}}\log\frac{2}{\delta}\\ &\quad+\frac{8V(-M)}{\sqrt{m}}\log\frac{2}{\delta}+8\sqrt{C_s}\Big(\frac{(V(-M))^2}{m}\Big)^{\frac{1}{2+s}}R^{\frac{s}{2+s}}.\end{aligned}\tag{3.4}$$

Plugging (2.12), (3.2) and Lemma 2 into (3.4), it follows that

$$\begin{aligned}\mathcal{E}(\pi_M(f_{\mathbf{z}}))-\mathcal{E}(f_P)+\lambda\|f_{\mathbf{z}}\|^2_{\mathcal{H}_K} &\leq C_r\lambda^r+C_{p,q}M^{-q}+(2\sqrt{2}+8)m^{-\frac{1}{2}}\log\frac{2}{\delta}\\ &\quad+(\sqrt{2}+8)Mm^{-\frac{1}{2}}\log\frac{2}{\delta}+\kappa\sqrt{2C_r}\lambda^{\frac{r-1}{2}}m^{-\frac{1}{2}}\log\frac{2}{\delta}\\ &\quad+8\sqrt{C_s}4^{\frac{1}{2+s}}M^{\frac{2}{2+s}}m^{-\frac{1}{2+s}}R^{\frac{s}{2+s}}.\end{aligned}$$

Here we have used the reproducing property (1.2) in $\mathcal{H}_K$ which yields

$$\|f\|_{L^\infty(X)}\leq\kappa\|f\|_{\mathcal{H}_K},\ \forall f\in\mathcal{H}_K.$$

This completes the proof of the theorem.

□

3.2 Strong Bound by Iteration

In Proposition 1, we need some $R \geq 1$ for $\mathbf{z} \in \mathcal{W}(R)$. We can choose $R = \lambda^{-1/2}$ according to

$$\|f_{\mathbf{z}}\|_{\mathcal{H}_K} \leq \lambda^{-1/2}, \quad \forall \mathbf{z} \in Z^m, \tag{3.5}$$

which comes immediately by taking $f = 0$ in (1.3). We observe that the bound in (3.2) is much better than this choice. This motivates us to get a similar tight bounds for $f_{\mathbf{z}}$. We will apply Proposition 1 iteratively to achieve this target which in turn improves learning rates. This iteration technique has been used in [28, 20].

Lemma 8. *Suppose Assumptions 2.1 and 2.2 hold for $s > 0$ and $0 < r \leq 1$. Take $\lambda = m^{-\alpha}$ with $0 < \alpha \leq 1$ and $M = m^{\beta}$ with $0 < \beta \leq \infty$. Let $0 < \eta < 1$. Then for any $0 < \delta < 1$, with confidence $1 - \delta$, there holds*

$$\|f_{\mathbf{z}}\|_{\mathcal{H}_K} \leq R^{(J)} \leq C_1' \Big(\log \frac{2}{\eta} \Big)^2 \sqrt{\log \frac{2}{\delta}} m^{\theta_\eta} \tag{3.6}$$

where θ_η is given by

$$\theta_\eta = \max\{ \frac{\alpha(1-r)}{2}, \frac{\alpha - \beta q}{2}, \frac{2(\alpha+\beta)-1}{4}, \frac{\alpha(3-r)-1}{4}, \frac{\alpha(2+s)+2\beta-1}{4+s} + \eta \} \geq 0$$

and

$$C_1' = 32\sqrt{C_s} 4^{\frac{1}{2+s}} \Big(1 + \sqrt{C_r} + \sqrt{C_{p,q}} + 2\sqrt{2\sqrt{2}+8} + \sqrt{\kappa \sqrt{2C_r}} \Big).$$

Proof. Putting $\lambda = m^{-\alpha}$ with $0 < \alpha \leq 1$ and $M = m^{\beta}$ with $0 < \beta \leq \infty$ into Proposition 1, we know that for any $R \geq 1$ there exists a subset V_R of Z^m with measure at most δ such that

$$\|f_{\mathbf{z}}\|_{\mathcal{H}_K} \leq a_m R^{\frac{s}{4+2s}} + b_m, \quad \forall \mathbf{z} \in \mathcal{W}(R) \setminus V_R,$$

where with $\zeta := \max\{\alpha(1-r)/2, (\alpha - \beta q)/2, (\alpha+\beta)/2 - 1/4, \alpha(3-r)/4 - 1/4\} \geq 0$, the constants are given by

$$a_m = \sqrt{8\sqrt{C_s} 4^{\frac{1}{2+s}}} m^{(\alpha/2) + ((2\beta-1)/(4+2s))},$$

$$b_m = \Big(\sqrt{C_r} + \sqrt{C_{p,q}} + 2\sqrt{2\sqrt{2}+8} \sqrt{\log \frac{2}{\delta}} + \sqrt{\kappa \sqrt{2C_r}} \sqrt{\log \frac{2}{\delta}} \Big) m^{\zeta}.$$

It follows that

$$\mathcal{W}(R) \subseteq \mathcal{W}(a_m R^{\frac{s}{4+2s}} + b_m) \cup V_R \tag{3.7}$$

Let us apply (3.7) iteratively to a sequence $\{R^{(j)}\}_{j=0}^{J}$ defined by $R^{(0)} = \lambda^{-1/2}$ and $R^{(j)} = a_m(R^{(j-1)})^{\frac{s}{4+2s}} + b_m$ where $J \in \mathbb{N}$ will be determined later. Then, $\mathcal{W}(R^{(j-1)}) \subseteq \mathcal{W}(R^{(j)}) \cup V_{R^{(j-1)}}$. (3.5) tells us that $\mathcal{W}(R^{(0)}) = Z^m$. So we have

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

As the measure of $V_{R^{(j)}}$ is at most δ, we know 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 $\Delta = \frac{s}{4+2s} < \frac{1}{2}$. The definition of sequence $\{R^{(j)}\}_{j=0}^{J}$ tells us that

$$R^{(J)} = a_m^{1+\Delta+\Delta^2+\cdots+\Delta^{J-1}}(R^{(0)})^{\Delta^J} + \sum_{j=1}^{J-1} a_m^{1+\Delta+\Delta^2+\cdots+\Delta^{j-1}} b_m^{\Delta^j} + b_m. \tag{3.8}$$

Let us bound the two terms on the right-hand side of (3.8).

The first term equals

$$(8\sqrt{C_s}4^{\frac{1}{2+s}})^{\frac{1-\Delta^J}{2(1-\Delta)}} m^{\frac{\alpha(2+s)+2\beta-1}{4+2s}\frac{1-\Delta^J}{1-\Delta}} m^{\frac{\alpha\Delta^J}{2}},$$

which can be bounded by

$$\begin{aligned} &8\sqrt{C_s}4^{\frac{1}{2+s}} m^{\frac{\alpha(2+s)+2\beta-1}{(4+2s)(1-\Delta)}} m^{\left(\frac{\alpha}{2} - \frac{\alpha(2+s)+2\beta-1}{(4+2s)(1-\Delta)}\right)\Delta^J} \\ &\le 8\sqrt{C_s}4^{\frac{1}{2+s}} m^{\frac{\alpha(2+s)+2\beta-1}{4+s}} m^{\frac{1}{4+s}2^{-J}}. \end{aligned}$$

Take J to be the smallest integer greater than or equal to $\log(1/\eta)/\log 2$. The above inequality can be bounded by $8\sqrt{C_s}4^{\frac{1}{2+s}} m^{\frac{\alpha(2+s)+2\beta-1}{4+s}+\eta}$.

The second term equals

$$\sum_{j=1}^{J-1} a_m^{1+\Delta+\Delta^2+\cdots+\Delta^{j-1}} b_m^{\Delta^j} + b_m \le \sum_{j=1}^{J-1} 8\sqrt{C_s}4^{\frac{1}{2+s}} m^{\frac{\alpha(2+s)+2\beta-1}{4+2s}\frac{1-\Delta^j}{1-\Delta}} b_1^{\Delta^j} m^{\zeta\Delta^j} + b_1 m^{\zeta},$$

where $b_1 := \sqrt{C_r} + \sqrt{C_{p,q}} + 2\sqrt{2\sqrt{2}+8}\sqrt{\log\frac{2}{\delta}} + \sqrt{\kappa\sqrt{2C_r}}\sqrt{\log\frac{2}{\delta}}$. It is bounded by

$$m^{\frac{\alpha(2+s)+2\beta-1}{4+s}} 8\sqrt{C_r}4^{\frac{1}{2+s}} b_1 \sum_{j=0}^{J-1} m^{\left(\zeta - \frac{\alpha(2+s)+2\beta-1}{4+s}\right)\frac{s^j}{(4+2s)^j}}.$$

When $\zeta \le \frac{\alpha(2+s)+2\beta-1}{4+s}$, the above expression is bounded by $8\sqrt{C_s}4^{\frac{1}{2+s}} b_1 J m^{\frac{\alpha(2+s)+2\beta-1}{4+s}}$. When $\zeta \ge \frac{\alpha(2+s)+2\beta-1}{4+s}$, it is bounded by $8\sqrt{C_s}4^{\frac{1}{2+s}} b_1 J m^{\zeta}$.

Based on the above discussion, we obtain

$$R^{(J)} \le (8\sqrt{C_s}4^{\frac{1}{2+s}} + 8\sqrt{C_s}4^{\frac{1}{2+s}} b_1 J) m^{\theta_\eta},$$

where $\theta_\eta = \max\{\frac{\alpha(2+s)+2\beta-1}{4+s} + \eta, \zeta\}$. Hence with confidence $1 - J\delta$, there holds

$$\|f_{\mathbf{z}}\|_{\mathcal{H}_K} \le R^{(J)} \le 8\sqrt{C_s}4^{\frac{1}{2+s}}\Big(1 + \sqrt{C_r} + \sqrt{C_{p,q}} + 2\sqrt{2\sqrt{2}+8} + \sqrt{\kappa\sqrt{2C_r}}\Big)J\sqrt{\log\frac{2}{\delta}}m^{\theta_\eta}.$$

Therefore, the desired result follows by replacing δ by δ/J and noting $J \le 2\log(2/\eta)$. □

Proof of Theorem 2. Take R to be the right side of (3.6). By Lemma 3.2, there exists a subset V_R' of Z^m with measure at most δ such that $Z^m \setminus V_R' \subseteq \mathcal{W}(R)$. Applying Proposition 1 to this R, we know that there exists another subset V_R of Z^m with measure at most δ such that for any $\mathbf{z} \in \mathcal{W}(R) \setminus V_R$,

$$\begin{aligned}\mathcal{E}(\pi_M(f_{\mathbf{z}})) - \mathcal{E}(f_P) \;\le\; & C_r m^{-\alpha r} + C_{p,q}m^{-\beta q} + 2(2\sqrt{2}+8)m^{\beta-\frac{1}{2}}\log\frac{2}{\delta} \\ & + \kappa\sqrt{2C_r}m^{-\frac{\alpha(r-1)}{2}-\frac{1}{2}}\log\frac{2}{\delta} + C_2'\Big(\log\frac{2}{\eta}\Big)^2\sqrt{\log\frac{2}{\delta}}m^{\frac{s\theta_\eta+2\beta-1}{2+s}}.\end{aligned}$$

where $C_2' = 8\sqrt{C_s}4^{\frac{1}{2+s}}(C_1')^{\frac{s}{2+s}}$. Since the set $V_R \cup V_R'$ has measure at most 2δ, after scaling 2δ to δ and setting the constant C_3' by

$$C_3' = C_r + C_{p,q} + 2(2\sqrt{2}+8) + \kappa\sqrt{2C_r} + C_2',$$

we see that

$$\mathcal{E}(\pi_M(f_{\mathbf{z}})) - \mathcal{E}(f_P) \le C_3'\Big(\log\frac{2}{\eta}\Big)^2\log\frac{4}{\delta}m^{-\vartheta}$$

with confidence $1-\delta$ and the power index ϑ is given by

$$\vartheta = \min\Big\{\alpha r, \beta q, \frac{1-2\beta}{2}, \frac{\alpha(r-1)+1}{2}, \frac{1-2\beta-s\theta_\eta}{2+s}\Big\},$$

provided that

$$\theta_\eta < \frac{1-2\beta}{s}. \tag{3.9}$$

With the choice of $\beta = \frac{1}{2(q+1)} < \frac{1}{2}$, $\theta_\eta < \frac{q}{s(q+1)}$. By the restriction $0 < \alpha < \frac{4q}{3s(q+1)}$ on α, we find that $\frac{\alpha(1-r)}{2} < \frac{q}{s(q+1)}$, $\frac{\alpha-\beta q}{2} = \frac{2(\alpha+\beta)-1}{4} < \frac{q}{s(q+1)}$ and $\frac{\alpha(3-r)-1}{4} < \frac{q}{s(q+1)}$. Moreover, restriction (2.15) on η implies that $\frac{\alpha(2+s)+2\beta-1}{4+s} + \eta < \frac{q}{s(q+1)}$. Therefore, condition (3.9) is satisfied. The above restriction and the expression for ϑ tells us that the power index for the error bound can be exactly expressed by formular (2.17). Combining with (2.6) and (2.7), the proof of Theorem 2 is complete with $\tilde{C}_1 = C_p C_3'$ and $\tilde{C}_2 = C_{q,\tau}C_3'$ respectively.

Now we are in the position to prove Theorem 1.

Proof of Theorem 1. Since $X \subset \mathbb{R}^n$ and $K \in C^\infty(X \times X)$, we know from [40] that (2.11) holds true for any $s > 0$. With $0 < \eta < \frac{q}{2(q+1)}$, let us choose s to be a positive number satisfying the

following inequalities:

$$\begin{gathered}\frac{q}{2(q+1)} < \frac{4q}{3s(q+1)},\\ \frac{1}{3} < \frac{(2+s)[2q - s(q+1)q/(2(q+1))]}{s(4+s)(q+1)},\\ \frac{q}{2(q+1)} - \eta < \frac{q}{(2+s)(q+1)} - \frac{s[(2+s)(q+1)q/(2(q+1)) - q]}{(2+s)(4+s)(q+1)} - \frac{s}{3(2+s)},\\ \frac{q}{2(q+1)} - \eta < \frac{q}{(2+s)(q+1)}.\end{gathered} \tag{3.10}$$

The first inequality above tells us the restriction on α is satisfied by choosing $\alpha = \frac{q}{2(q+1)}$. The second inequality shows that condition (2.15) for the parameter η renamed now as η^* is also satisfied by taking $\eta^* = \frac{1}{3}$. Thus, we apply Theorem 2 and know that with confidence $1-\delta$, (2.16) and (2.18) hold with the power index ϑ given by (2.17). Since $r = 1$, $\alpha = \frac{q}{2(q+1)}$, and $\eta^* = \frac{1}{3}$, we can derive that

$$\vartheta = \min\{\alpha, \frac{q}{(2+s)(q+1)}, \frac{q}{(2+s)(q+1)} - \frac{s[\alpha(2+s)(q+1) - q]}{(2+s)(4+s)(q+1)} - \frac{s}{3(2+s)}\}.$$

The last two inequalities in (3.10) satisfied by s yield $\vartheta \geq \alpha - \eta$. So (2.16) and (2.18) verify (2.13) and (2.14). This completes the proof of the theorem.

4 Discussion

This paper established quantitative convergence analysis for a class of kernel based large-margin unified machines, which were proposed in order to solve the so-called "data piling" problem in the setting of high dimension and low sample size. We derived explicit learning rates for this kind of learning schemes under mild conditions on the regularization error and the capacity of RKHS measured by uniform covering number. Note that it is possible to improve the learning rates by considering some special Mercer kernels, such as Gaussian kernel [34, 31, 32]. It will also be interesting to study kernel based LUMs in the settings of multicategory learning [13] and pairwise learning [38, 23] within the framework of statistical learning theory. These will be our future work.

Appendix

Proof of Lemma 2. Rewrite the LUM loss as

$$V(t) = \frac{1}{1+p}\Big(\frac{q}{(1+p)t - p + q}\Big)^q = \frac{1}{1+p}\Big(\frac{q}{1+p}\Big)^q\Big(\frac{1}{t - \frac{p-q}{1+p}}\Big)^q.$$

If $p \leq q$, we observe that $t - \frac{p-q}{1+p} \geq t$. Notice that $t \geq \frac{p}{1+p}$. It follows that

$$V(t) \leq \frac{1}{1+p}\Big(\frac{q}{1+p}\Big)^q t^{-q}.$$

If $p > q$, since $t \geq \frac{p}{1+p}$, we observe that $t - \frac{p-q}{1+p} = t - \frac{p}{1+p}\frac{p-q}{p} \geq t - \frac{p-q}{p}t = \frac{q}{p}t$. It yields that

$$V(t) \leq \frac{1}{1+p}\Big(\frac{p}{1+p}\Big)^q t^{-q}.$$

Based on the above discussion, we obtain the desired result (2.1).

Proof of Lemma 6. The regularized excess generalization error can be written as

$$\begin{aligned}\mathcal{E}(\pi_M(f_{\mathbf{z}})) - \mathcal{E}(f_P) + \lambda\|f_{\mathbf{z}}\|_{\mathcal{H}_K}^2 &= \big\{\mathcal{E}(\pi_M(f_{\mathbf{z}})) - \mathcal{E}_{\mathbf{z}}(\pi_M(f_{\mathbf{z}}))\big\} \\ &\quad + \big\{\mathcal{E}_{\mathbf{z}}(\pi_M(f_{\mathbf{z}})) + \lambda\|f_{\mathbf{z}}\|_{\mathcal{H}_K}^2 - (\mathcal{E}_{\mathbf{z}}(f_\lambda) + \lambda\|f_\lambda\|_{\mathcal{H}_K}^2)\big\} \quad (4.1) \\ &\quad + \big\{\mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}(f_\lambda)\big\} + \big\{\mathcal{E}(f_\lambda) - \mathcal{E}(f_P) + \lambda\|f_\lambda\|_{\mathcal{H}_K}^2\big\}.\end{aligned}$$

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

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

Our conclusion follows by subtracting and adding $\mathcal{E}(\pi_M(f_P))$ and $\mathcal{E}_{\mathbf{z}}(\pi_M(f_P))$ in the first and third term of (4.1).

References

[1] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68: 337–404, 1950.

[2] 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 Transactions on Information Theory*, 44: 525-536, 1998.

[3] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.*, 101: 138–156, 2006.

[4] B. E. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *Computational Learning Theory*, pp. 144–152, Madison, WI, 1992. ACM.

[5] D. R. Chen, Q. Wu, Y. Ying, and D. X. Zhou. Support vector machine soft margin classifiers: error analysis. *Journal of Machine Learning Research*, 5: 1143–1175, 2004.

[6] C. Cortes and V. Vapnik. Support vector networks. *Mach. Learn.*, 20: 273–297, 1995.

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

[8] D. E. Edmunds and H. Triebel. *Function Spaces, Entropy Numbers, Differential Operators*. Cambridge University Press, Cambridge, 1996.

[9] J. Fan and D. H. Xiang. Comparison theorems on large margin learning. Technique report, 2019.

[10] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55: 119–139, 1997.

[11] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting (with discussion). *Ann. Statist.*, 28: 337–407, 2000.

[12] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York, 2001.

[13] Y. W. Lei, U. Dogan, D. X. Zhou and M. Kloft. Data-dependent generalization bounds for multi-class classification. *IEEE Transactions on Information Theory*, 65: 2995–3021, 2019.

[14] Y. F. Liu, H. H. Zhang, and Y. C. Wu. Hard or soft classificaiton? large-margin unified machines. *Journal of the American Statistical Association*, 106: 166–177, 2011.

[15] J. S. Marron, M. Todd, and J. Ahn. Distance weighted discrimination. *Journal of the American Statistical Association*, 102: 1267–1271, 2007.

[16] S. Smale. and D. X. Zhou. Online learning with markov sampling. *Analysis and Applications*, 7: 87-113, 2009.

[17] S. Smale and D. X. Zhou. Learning theory estimates via integral operators and their approximations. *Constr. Approx.*, 26: 153–172, 2007.

[18] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *J. Mach. Learn. Res.*, 2: 67–93, 2001.

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

[20] I. Steinwart and C. Scovel. Fast rates for support vector machines using gaussian kernels. *Ann. Statist.*, 35: 575-607, 2007.

[21] A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32: 135–166, 2004.

[22] G. Wahba. Spline models for observational data. SIAM, 1990.

[23] C. Wang and T. Hu. Online minimum error entropy algorithm with unbounded sampling. *Anal. Appl.*,17: 293–322, 2019.

[24] B. X. Wang and H. Zou. Another look at distance-weighted discrimination. *Journal of the Royal Statistical Society*, 80: 177–198, 2018.

[25] R. Williamson, A. J. Smola, and B. Schölkopf. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. *IEEE Trans. Inform. Theory*, 47: 2516–2532, 2001.

[26] Q. Wu, Classification and Regularization in Learning Theory, VDM Verlag, 2009.

[27] Q. Wu, Y. M. Ying., and D. X. Zhou. Multi-kernel regularized classifiers. *Journal of Complexity*, 23: 108-134, 2007.

[28] Q. Wu, Y. Ying., and D. X. Zhou. Learning rates of least-square regularized regression. *Foundations of Computational Mathematics*, 6: 171-192, 2006.

[29] Q. Wu and D. X. Zhou. Svm soft margin classifiers: linear programming versus quadratic programming. *Neural Comput.*, 17: 1160-1187, 2005.

[30] Q. Wu and D. X. Zhou. Analysis of support vector machine classification. *Journal of Computational Analysis and Applications*, 8: 99-119, 2006.

[31] D. H. Xiang. Classification with gaussians and convex loss ii: improving error bounds by noise conditions. *Science China Mathematics*, 54: 165-171, 2011.

[32] D. H. Xiang. Logistic classification with varying gaussians. *Computers and Mathematics with Applications*, 61: 397-407, 2011.

[33] D. H. Xiang. Conditional quantiles with varying gaussians. *Adv. Comput. Math.*, 38: 723-735, 2013.

[34] D. H. Xiang and D. X. Zhou. Classification with gaussians and convex loss. *Journal of Machine Learning Research*, 10: 1447–1468, 2009.

[35] D. H. Xiang, T. Hu., and D. X. Zhou. Approximation analysis of learning algorithms for support vector regression and quantile regression. *Journal of Applied Mathematics*, 2012: 17 pages, 2012.

[36] Y. Ying and D. X. Zhou. Learnability of Gaussians with Flexible Variances. *J. Mach. Learn. Res.*, 8: 249–276, 2007.

[37] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Statist.*, 32: 56–85, 2004.

[38] Y. L. Zhao, J. Fan and L. Shi. Learing rates for regularized least squares ranking algorithm. *Anal. Appl.*,15: 815–836, 2017.

[39] D. X. Zhou. The covering number in learning theory. *J. Complexity*, 18: 739-767, 2002.

[40] D. X. Zhou. Capacity of reproducing kernel spaces in learning theory. *IEEE Transactions on Information Theory*, 49: 1743-1752, 2003.