ERROR ANALYSIS ON REGULARIZED REGRESSION BASED ON THE MAXIMUM CORRENTROPY CRITERION

BINGZHENG LI* AND ZHENGZHAN DAI

School of Mathematical Science, Zhejiang University
Hangzhou, 310027, China

(Communicated by Ding-Xuan Zhou)

Abstract. This paper aims at the regularized learning algorithm for regression associated with the correntropy induced losses in reproducing kernel Hilbert spaces. The main target is the error analysis for the regression problem in learning theory based on the maximum correntropy. Explicit learning rates are provided. From our analysis, when choosing a suitable parameter of the loss function, we obtain satisfactory learning rates. The rates depend on the regularization error and the covering numbers of the reproducing kernel Hilbert space.

1. Introduction. The regression problem can be traced back to the linear regression with least squares. We will give an example to explain what is the regression and lead to the mathematical model for regression in learning theory. Assume the law can be expressed as a function $f: \mathbb{R} \to \mathbb{R}$, and the function has a specific form $f_\alpha(x) = \sum_{i=1}^{N} \alpha_i \psi_i(x)$, where the $\psi_i$ are the elements of a basis of a specific function space. We need to learn the coefficients $\alpha := (\alpha_1, \alpha_2, \ldots, \alpha_N)$ from a set of data $\{(x_1, y_1), \ldots, (x_m, y_m)\}$. $y_i = f(x_i)$, $i = 1, \ldots, m$, if the measurements generating this set are exact. But in general the values $y_i$ to be affected by noise. So we computes the vector of coefficients $\alpha$ such that the value

$$\sum_{i=1}^{m} (f_\alpha(x_i) - y_i)^2,$$

with $f_\alpha(x) = \sum_{i=1}^{m} \alpha_i \psi_i(x)$,

is minimized. This is the typical least squares regression method which can going back to Gauss and Legendre (see [2]).

Generally, we assume that $\mathcal{X}$ is a compact metric space and $\mathcal{Y} \subset \mathbb{R}$. Let $\rho$ be a probability distribution on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, which is usually unknown. The generalization error ([2]) for $f: \mathcal{X} \to \mathcal{Y}$ is defined as

$$\mathcal{E}(f) = \int_{\mathcal{Z}} (f(x) - y)^2 \, dp.$$

$(f(x)-y)^2$ is the error suffered from the use of $f$ as a model for the process producing $y$ from $x$, for each $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

2010 Mathematics Subject Classification. 41A25, 68Q32.

Key words and phrases. Maximum correntropy, regularization error, covering number, learning theory, reproducing kernel Hilbert spaces.

* Corresponding author: Bingzheng Li.
Let $\rho(y|x)$ be the conditional probability distribution on $Y$ and $\rho_X$ be the marginal probability distribution of $\rho$ on $X$. Define $f_\rho : X \to Y$ by

$$f_\rho(x) = \int_Y y \, d\rho(y|x).$$

The function $f_\rho$ is called the regression function which minimizes the generalization error.

Let $z \in Z^m$, $z = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ be a sample in $Z^m$. The empirical error of $f$ is defined as

$$E_z(f) = \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2.$$

In this paper, we consider the general model of regression $Y = f^*(X) + \varepsilon$, $\mathbb{E}(\varepsilon|X = x) = 0$, (1.1)

where $X$ is the explanatory variable taking values in $X$ and $Y \in Y$ is the response variable. From (1.1), it is easy to see that almost surely there holds $f_\rho = f^*$.

The target of the regression problem is to find a good approximation of regression function from random samples. The study of regression problems can be found in a large literature of learning theory (see [20], [21] and [12] and references therein). As we know, the most employed methodology for quantifying the regression efficiency is the mean squared error. This is the classical tool that minimized the variance of $f(x) - y$ and belongs to the second-order statistics(see [2]). The drawback of the second-order statistics is that its optimality depends heavily on the assumption of Gaussian noise. However, in many practical applications, data may be contaminated by non-Gaussian noise or outliers. To solve this kind of problems, Hu et al ([10]and [11]) presented thorough studies on the minimum error entropy criterion for regression (MEECR) from a learning theory viewpoint. They presented the first results concerning the regression consistency and convergence rates by MEECR. This also motivates the introduction of the maximum correntropy criterion into the regression problems [5]. Recently, Feng and Ying([6]) investigated the behavior of correntropy based regression in the presence of outliers by using the Fourier analysis technique developed in Fan et al. ([3]) and by modeling outliers using Huber’s contamination model. It was also demonstrated in Feng et al. ([4]) that correntropy based regression essentially regresses towards the conditional mode function when the scale parameter diminishes towards zero. Meanwhile, Lv and Fan([14]) provided the optimal learning with Gaussians and correntropy loss.

A generalized correlation function named correntropy ([15]) is a generalized similarity measure between two scalar random variables $T_1$ and $T_2$, which is defined by

$$V^\sigma(T_1, T_2) = \mathbb{E}(K^\sigma(T_1, T_2)),$$

where $\mathbb{E}(\cdot)$ denotes mathematical expectation, $K^\sigma(t_1, t_2) = \exp\{-(t_1 - t_2)^2/\sigma^2\}$ is a Gaussian kernel with the scale parameter $\sigma > 0$, $(t_1, t_2)$ being a realization of $(T_1, T_2)$. We call it the maximum correntropy criterion for regression (MCCR), when the correntropy $V^\sigma$ is used in regression problems. The correntropy contains higher-order moments of the probability density function and can apply to non-Gaussian regression problems. Recently, this kind of generalized correlation function (correntropy) has drawn much attention in signal processing and machine
learning community. In [13], Liu et al. used the correntropy in non-Gaussian signal processing. The new method outperforms MSE (mean squared error) in the case of impulsive noise since correntropy is inherently insensitive to outliers. In [7, 8], He et al. presented a sparse correntropy framework for computing robust sparse representations of face images for recognition. The new method can improve both recognition accuracy and receiver operator characteristic curves. Feng et al. [5] gave a theoretical understanding on the maximum correntropy criterion for regression (MCCR) within the statistical learning framework. The maximum correntropy criterion has also succeeded in some real-world applications (see [1]).

We give the definition of correntropy loss function from [5] as follows.

**Definition 1.** The correntropy induced regression loss function \( l_\sigma : \mathbb{R} \times \mathbb{R} \to [0, +\infty) \) is defined as

\[
l_\sigma(y, t) = \sigma^2 \left( 1 - e^{-\frac{(y-t)^2}{\sigma^2}} \right), \quad y \in \mathcal{Y}, \ t \in \mathbb{R},
\]

with \( \sigma > 0 \) being a scale parameter.

**Definition 2.** The risk functional for MCCR algorithms is given by

\[
\mathcal{E}^\sigma(f) = \int \mathcal{Z} l_\sigma(y, f(x)) d\rho(x, y).
\]

The corresponding empirical risk on a set of observation \( \mathbf{z} \) is defined as

\[
\mathcal{E}^\sigma_{\mathbf{z}}(f) = \frac{1}{m} \sum_{i=1}^{m} l_\sigma(y_i, f(x_i)).
\]

We also need the following concepts about covering numbers [2]:

**Definition 3.** The covering number of the hypothesis space \( \mathcal{H} \), which is denoted as \( N(\mathcal{H}, \eta) \), with the radius \( \eta > 0 \), is defined as

\[
N(\mathcal{H}, \eta) := \inf \left\{ l \geq 1 : \text{there exist } f_1, \ldots, f_l \in \mathcal{H}, \text{such that } \mathcal{H} \subset \bigcup_{i=1}^{l} B(f_i, \eta) \right\},
\]

where \( B(f, \eta) = \{ g \in \mathcal{H} : ||f - g||_{\infty} \leq \eta \} \) denotes the closed ball in \( C(X) \) with center \( f \in \mathcal{H} \) and radius \( \eta \).

**Definition 4.** Let \( \mathbf{x} = \{ x_1, x_2, \ldots, x_m \} \subset \mathcal{X}^m \). The \( l^2 \)-empirical covering number of the hypothesis space \( \mathcal{H} \), which is denoted as \( N_2(\mathcal{H}, \eta) \), with radius \( \eta > 0 \), is defined by

\[
N_2(\mathcal{H}, \eta) := \sup_{m \in \mathbb{N}} \sup_{\mathbf{x} \in \mathcal{X}^m} \inf \left\{ l \in \mathbb{N} : \exists \{ f_i \}_{i=1}^{l} \subset \mathcal{H} \text{ such that for each } f \in \mathcal{H}, \text{ there exists some } i \in \{ 1, 2, \ldots, l \} \text{ with } \frac{1}{m} \sum_{j=1}^{m} |f(x_j) - f_i(x_j)|^2 \leq \eta^2 \right\}.
\]

By minimizing the empirical risk over a set of continuous functions \( \mathcal{H} \), we can obtain a predictor \( f_{\mathbf{z}} \) as

\[
f_{\mathbf{z}} = \arg\min_{f \in \mathcal{H}} \mathcal{E}^\sigma_{\mathbf{z}}(f).
\]

(1.2)

Naturally, we want to know the performance of the predictor. For example, in the least squares regression the predictive power of a function \( f \) is measured by the mean squared error \( \mathcal{E}(f) = \int \mathcal{Z} (y - f(x))^2 d\rho \). From the definition of the regression function, we know that \( f_{\rho} \) minimized the mean squares error. So we also can estimate the excess mean squares error i.e. \( \mathcal{E}(f) - \mathcal{E}(f_{\rho}) = \|f - f_{\rho}\|_{L_2^\mathcal{X}}^2 \) to measure the goodness of a predictor \( f \).
However, existing theoretical results on understanding the loss $l_{\sigma}$ and MCCR model are very limited. The reason mainly lies in the non-convexity property of the loss function $l_{\sigma}$. Feng et al [5] presented the consistency properties and the convergence rates of the MCCR model (1.2) under the assumption of $\mathcal{H}$ being a compact subset of $C(\mathcal{X})$. By choosing the suitable parameters $\sigma$, they give the learning rate to bound the difference between $f_x$ and $f^*(X)$. They obtained the following main results [5].

**Theorem A.** Assume $\log \mathcal{N}(\mathcal{H}, \eta) \leq C_q \eta^{-q}$, $q > 0$ and $\int_{X} y^4 \, d\rho < \infty$. For any $0 < \delta < 1$, with confidence $1 - \delta$, there holds

$$\| f_x - f_\rho \|_{L^2_X}^2 \leq 3 \| f_\mathcal{H} - f_\rho \|_{L^2_X}^2 + C_{\mathcal{H}, \rho} \log(2/\delta) \left( \sigma^{-2} + \sigma m^{-1/(1+\eta)} \right),$$

where $f_\mathcal{H} = \arg\min_{f \in \mathcal{H}} \mathcal{E}(f)$, $C_{\mathcal{H}, \rho}$ is a positive constant independent of $m$, $\sigma$ or $\delta$.

**Theorem B.** Assume $\log \mathcal{N}(\mathcal{H}, \eta) \leq C_s \eta^{-s}$, $0 < s < 2$ and $|y| \leq M$ almost surely for some $M > 0$. Let $f_\rho \in \mathcal{H}$ with $\sigma = m^{1/(2+s)}$. For any $0 < \delta < 1$, with confidence $1 - \delta$, there holds

$$\| f_x - f_\rho \|_{L^2_X}^2 \leq C'_{\mathcal{H}, \rho} \log(2/\delta) m^{-s/2+s},$$

where $C'_{\mathcal{H}, \rho}$ is a positive constant independent of $m$, $\sigma$ or $\delta$.

Notice that the MCCR model (1.2) is a constrained optimization model since $\mathcal{H}$ is assumed to be a compact subset of a certain reproducing kernel Hilbert space $\mathcal{H}_K$ induced by some Mercer kernel $K$. In this paper, instead of evaluating the optimization model (1.2), we focus on the following unstrained version (1.3). That is to say we study the regularized MCCR algorithm based on minimizing $\mathcal{E}_\rho^x$. We want to emphasize that, in [5] the authors provided some numerical experiments by using the regularized regression model (1.3) (e.g., example of the noisy sinc function and example of the noisy Friedman’s benchmark function). The experiments verified that regularized MCCR model gave the best fitting results, especially at positions where data are corrupted by outliers. This indicates that it is very necessary to establish the error analysis theory of the regularized MCCR algorithm based on minimizing $\mathcal{E}_\rho^x$.

For this purpose, we need to introduce the definition of reproducing kernel Hilbert spaces. Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be continuous, symmetric and positive semi-definite, i.e., for any finite set of distinct points $x = \{x_1, \ldots, x_m\} \subset \mathcal{X}$ the matrix $(K(x_i, x_j))_{i,j=1}^m$ is positive semi-definite. Such a function is called a Mercer kernel. The reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ associated with the kernel $K$ is defined to be the closure of the span of $\{K_x = K(\cdot, x) : x \in \mathcal{X}\}$ with the inner product satisfying $\langle K_x, K_t \rangle_K = K(x, t)$. The reproducing property of the RKHS

$$f(x) = \langle f, K_x \rangle_K, \quad \forall x \in \mathcal{X}, \quad f \in \mathcal{H}_K.$$

**Definition 5.** The regularized MCCR algorithm in an RKHS $\mathcal{H}_K$ is defined as

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

where $\lambda > 0$ is a regularization parameter.

To obtain the main results, we need some assumptions.

Denote $B_1 = \{ f \in \mathcal{H}_K : \| f \|_K \leq R \}$. Throughout this paper, we assume that for some $q > 0$ and $C_q > 0$, the covering number of $B_1$ satisfies

$$\log \mathcal{N}(B_1, \eta) \leq C_q \eta^{-q}, \quad \forall \eta > 0.$$
This has been extensively studied in learning theory ([24, 25]). From that, we know for a $C^\infty$ kernel, (1.4) is true for any $q > 0$.

**Definition 6.** The regularization error $D(\lambda)$ is defined as

$$D(\lambda) = \inf_{f \in \mathcal{H}_K} \{||f - f_\rho||^2_{L^2_X} + \lambda||f||^2_K\}, \quad \lambda > 0,$$

where the regularizing function

$$f_\lambda = \arg\min_{f \in \mathcal{H}_K} \{||f - f_\rho||^2_{L^2_X} + \lambda||f||^2_K\}.$$

The regularization algorithm has been well understood in [16] and [12]. We shall assume that for some constants $0 < \beta \leq 1$ and $C_\beta > 0$,

$$D(\lambda) \leq C_\beta \lambda^{\beta}, \quad \forall \lambda > 0. \quad (1.5)$$

Throughout the paper we assume that for a constant $M > 0$, there holds $|Y| \leq M$ almost surely. It follows that $||f_\rho||_{L^2_X} \leq ||f_\rho||_\infty \leq M$. Denote $C_K = \sup_{x \in X} \sqrt{K(x,x)}$ which is always assumed to be finite. For any $R > 0$, we denote $W(R) = \{z \in Z^m : ||f_{z,\lambda}||_K \leq R\}$.

Our main results are stated as follows.

**Theorem 1.** Assume the capacity condition (1.4) with $q > 0$. Let $\frac{M}{\sqrt{\lambda}} \geq 1$, $0 < \delta < 1$, $\frac{1}{m} \leq \lambda \leq 1$ and $\sigma > 1$, satisfy

$$\sigma \geq C_1 M \lambda^{-\frac{1}{2(1+q)(1+\beta)}} (\log 2/\delta)^{-\frac{1}{2}}. \quad (1.6)$$

Then with confidence $1 - \delta$, we have

$$||f_{z,\lambda} - f_\rho||^2_{L^2_X} + \lambda||f_{z,\lambda}||^2_K \leq C_2(\frac{M^2}{\lambda}m^{-\frac{1}{2(1+q)(1+\beta)}} + D(\lambda) + \frac{(D(\lambda))^2}{\lambda^2\sigma^2})log \frac{2}{\delta}.$$

Here $C_1$ and $C_2$ are constants independent of $m$, $\sigma$ or $\lambda$ and will be given explicitly in the proof.

From the Theorem, we can see that the decreasing value of the scale parameter $\sigma$ yields slower convergence rates. But the decreasing value of $\sigma$ enhance the robustness of the regression models. We realize that in the robustness literature, the scale parameter not only controls the robustness property of the regression model associated with the loss $l_\tau$ but also specifies its efficiency([9]). Hence $\sigma$ balance the convergence rates and the robustness of the model. In the following theorem, we discuss the influence of the scale parameter $\sigma$ on the convergence rates.

**Theorem 2.** Assume the capacity condition (1.4) with $q > 0$ and the approximation condition (1.5) with $0 < \beta < 1$. Take $\lambda = m^{-\frac{1}{(1+q)(1+\beta)}}$, then for $\sigma = m^{-\frac{2-\beta}{2(1+q)(1+\beta)}}$ and any $0 < \delta < 1$, with confidence $1 - \delta$ we have

$$||f_{z,\lambda} - f_\rho||^2_{L^2_X} = O(m^{-\frac{\beta}{(1+q)(1+\beta)}}).$$

**Theorem 3.** Assume the capacity condition (1.4) with $q > 0$ and the approximation condition (1.5) with $0 < \beta < 1$. Take $\lambda = m^{\frac{\alpha}{1+q}}$ with $0 < \alpha < \frac{1}{1+q}$. Then for $\sigma \geq m^{-\frac{\alpha}{1+q}}$ and any $0 < \delta < 1$, with confidence $1 - \delta$ we have

$$||f_{z,\lambda} - f_\rho||^2_{L^2_X} \leq C m^{\frac{\alpha}{1+q}} \left(\log \frac{2}{\delta} + \log \frac{2}{\alpha}\right)^\frac{2}{\alpha},$$

where $C$ is a constant independent of $m$ or $\delta$. 
Remark 1. When choose sufficiently large $\sigma$ and suitable regularization parameter $\lambda$, from Theorem 3 we can see that the convergence rate of the regularized MCCR algorithm is at least of order $O(m^{-2/(3+3q)})$ for an arbitrarily small $\alpha > 0$. When $H_\lambda$ contains $f_\rho$, $\beta = 1$ and the rate becomes $O(m^{-2/(3+3q)})$. When $f_\rho \in H$, let $\sigma = O(m^{1/(3+3q)})$, the rate in Theorem 1 becomes $O(m^{-2/(3+3q)})$. When $q \to 0$, the rate in Theorem A becomes $O(m^{-2/3})$, our result becomes $O(m^{-1})$ which is better than the rate in Theorem A. In Theorem B, when $s \to 0$, the rate becomes $O(m^{-1})$, our result is comparable to the rate in Theorem B.

The rest of this paper is organized as follows. In Section 2, we give the estimate and analysis for error bounds and prove Theorem 1 and Theorem 2. In Section 3, we improve the error rate by iteration method and give the proof of Theorem 3. In the last section, we give some conclusion and some ideas for the future work.

2. Estimate and analysis for error bounds. In this section we estimate and analyse some error bounds for the proof of our main results. We first show that the excess risk of $f$ associated with the $l_\sigma$ loss is a good approximation of the least squares loss when the scale parameter is large. In this paper, the excess risk of $f$ associated with the $l_\sigma$ loss refers to $E^\sigma(f) - E^\sigma(f_\rho)$ while the excess risk of $f$ associated with the least squares loss refers to $E(f) - E(f_\rho)$.

**Lemma 1.** For any essentially bounded measurable function $f$ on $X$, we have
\[
|E^\sigma(f) - E^\sigma(f_\rho) - \{E(f) - E(f_\rho)\}| \leq \sigma^{-2}(12M^4 + 4\|f\|_\infty^4).
\] (2.1)

**Proof.** From the Lemma 7 in [5] and the boundness of $Y$, we get the lemma. \qed

2.1. Error decomposition for excess MCCR risk. Lemma 1 enables us to analyse $E^\sigma(f_{x,\lambda}) - E^\sigma(f_\rho)$ instead of $E(f_{x,\lambda}) - E(f_\rho)$.

**Lemma 2.** For any $\lambda > 0$, there holds
\[
E^\sigma(f_{x,\lambda}) - E^\sigma(f_\rho) + \lambda\|f_{x,\lambda}\|_K^2 \leq S_1 + S_2 + D(\lambda) + \sigma^{-2}(12M^4 + 4C_1(\frac{D(\lambda)}{\lambda})^2),
\] (2.2)
where $S_1$ and $S_2$ are sample error terms defined as
\[
S_1 = \{E^\sigma_x(f_\lambda) - E^\sigma_x(f_\rho)\} - \{E^\sigma(f_\lambda) - E^\sigma(f_\rho)\},
\]
\[
S_2 = \{E^\sigma(f_{x,\lambda}) - E^\sigma(f_\rho)\} - \{E^\sigma_x(f_{x,\lambda}) - E^\sigma_x(f_\rho)\}.
\]

**Proof.** We can write $E^\sigma(f_{x,\lambda}) - E^\sigma(f_\rho) + \lambda\|f_{x,\lambda}\|_K^2$ as
\[
E^\sigma(f_{x,\lambda}) - E^\sigma_x(f_{x,\lambda}) + \{E^\sigma_x(f_{x,\lambda}) + \lambda\|f_{x,\lambda}\|_K^2\} - \{E^\sigma_x(f_\lambda) + \lambda\|f_\lambda\|_K^2\} + E^\sigma_x(f_\lambda) - E^\sigma(f_\lambda) + E^\sigma(f_\rho) + \lambda\|f_\lambda\|_K^2
\]
\[
\leq \{E^\sigma(f_{x,\lambda}) - E^\sigma_x(f_{x,\lambda})\} + \{E^\sigma_x(f_\lambda) - E^\sigma(f_\lambda)\} + \{E^\sigma(f_\rho) - E^\sigma(f_\rho)\} + \lambda\|f_\lambda\|_K^2
\]
\[
= \{E^\sigma(f_{x,\lambda}) - E^\sigma_x(f_{x,\lambda})\} + \{E^\sigma_x(f_\lambda) - E^\sigma(f_\lambda)\} + \lambda\|f_\lambda\|_K^2
\]
Because
\[
\{E^\sigma(f_{x,\lambda}) - E^\sigma_x(f_{x,\lambda})\} + \{E^\sigma_x(f_\lambda) - E^\sigma(f_\lambda)\}
\]
\[
= \{E^\sigma(f_{x,\lambda}) - E^\sigma(f_\rho)\} - \{E^\sigma_x(f_{x,\lambda}) - E^\sigma_x(f_\rho)\} + \{E^\sigma_x(f_\lambda) - E^\sigma_x(f_\rho)\}
\]
\[
= S_1 + S_2.
\]
Applying (2.1) of Lemma 1, we have
\[
E^\sigma(f_\lambda) - E^\sigma(f_\rho) + \lambda\|f_\lambda\|_K^2 \leq D(\lambda) + \sigma^{-2}(12M^4 + 4\|f_\lambda\|_\infty^4).
\]
Using the reproducing property and the Schwartz inequality, we have
\[ |f(x)| \leq C_K \|f\|_K, \quad \forall f \in \mathcal{H}_K. \]
So \( \|f_\lambda\|_\infty \leq C_K \|f_\lambda\|_K \leq C_K \sqrt{D(\lambda)/\lambda}. \) We finish the proof. \( \square \)

In next two subsections, we will estimate and analyse the sample error \( S_1 \) and \( S_2 \).

2.2. Bounding the sample error \( S_1 \). We need one-sided Bernstein’s concentration inequality [2].

**Lemma 3.** Let \( \xi \) be a random variable on a probability space \( \mathcal{Z} \) with variance \( \sigma^2 \), satisfying \( |\xi - \mathbb{E}(\xi)| \leq M_\xi \), almost surely for some constant \( M_\xi \) and for all \( z \in \mathcal{Z} \). Then

\[
\text{Prob}_{z \in \mathcal{Z}} \left\{ \frac{1}{m} \sum_{i=1}^{m} \xi(z_i) - \mathbb{E}(\xi) \geq \varepsilon \right\} \leq \exp \left\{ -\frac{m\varepsilon^2}{2(\sigma^2 + \frac{1}{3}M_\xi \varepsilon)} \right\}.
\]

We define a random variable \( \xi(z) \) with \( z \in \mathcal{Z} \) as
\[
\xi(z) := -\sigma^2 \exp\{-(y - f_\lambda(x))^2/\sigma^2\} + \sigma^2 \exp\{-(y - f_\rho(x))^2/\sigma^2\}.
\]

**Proposition 1.** For any \( 0 < \delta < 1 \), with confidence \( 1 - \delta \),
\[
S_1 \leq \left( \frac{4C_K^2 \log \frac{4}{\delta}}{m\lambda} + 3M + 1 \right) D(\lambda) + \frac{(24M^2 + 3M) \log \frac{4}{\delta}}{m}. \tag{2.3}
\]

**Proof.** We will use the one-sided Bernstein’s inequality to the random variable \( \xi \) to bound the sample error \( S_1 \). Hence, we need to verify conditions in Lemma 3.

Applying the mean value theorem to the auxiliary function \( k(s) = \exp\{-s\} \), \( s \in \mathbb{R} \) at \( s_1 = (y - f_\lambda(x))^2/\sigma^2 \), \( s_2 = (y - f_\rho(x))^2/\sigma^2 \) and it is easy to see that \( \|k'\|_\infty = 1 \), we get
\[
|\xi(z)| \leq |(y - f_\lambda(x))^2 - (y - f_\rho(x))^2| \leq (\|f_\lambda\|_\infty + M)(\|f_\lambda\|_\infty + 3M) \leq D_1,
\]

where \( D_1 = (\|f_\lambda\|_\infty + 3M)^2 \).

Consequently,
\[
|\xi - \mathbb{E}(\xi)| \leq 2\|\xi\|_\infty \leq D_2 := 2D_1.
\]

We are to bound the variance of the random variable \( \xi \), which is denoted by \( \text{var}(\xi) \).
\[
\text{var}(\xi) \leq \mathbb{E}(\xi^2) - \mathbb{E}(\xi)^2 \\
\leq (\|f_\lambda\|_\infty + 3M)^2 \|f_\lambda - f_\rho\|_{L^2,\mathcal{X}}^2 \leq D_1 D(\lambda).
\]

Now applying Lemma 3 to the random variable \( \xi \), for any \( \varepsilon > 0 \)
\[
S_1 = \frac{1}{m} \sum_{i=1}^{m} \xi(z_i) - \mathbb{E}(\xi) \leq \varepsilon,
\]

with confidence at least
\[
1 - \exp \left\{ -\frac{m\varepsilon^2}{2\text{var}(\xi) + \frac{2}{3}D_2 \varepsilon} \right\} \geq 1 - \exp \left\{ -\frac{m\varepsilon^2}{2D_1 (D(\lambda) + \frac{2}{3} \varepsilon)} \right\}.
\]

Choose \( \varepsilon^* \) to be the unique positive solution of the quadratic equation
\[
\frac{m\varepsilon^2}{2D_1 (D(\lambda) + \frac{2}{3} \varepsilon)} = \log \delta.
\]
Then with confidence $1 - \delta$, there holds

$$S_1 = \frac{1}{m} \sum_{i=1}^{m} \xi(z_i) - \mathbb{E}(\xi) \leq \varepsilon^*.$$ 

And

$$\varepsilon^* = \left( \frac{2D_1 \log(1/\delta)}{3m} + \sqrt{\frac{(2D_1 \log(1/\delta))^2 + 2D_1 m \log(1/\delta) D(\lambda)}{m}} \right) \leq \frac{4D_1 \log(1/\delta)}{3m} + \sqrt{\frac{2D_1 \log(1/\delta) D(\lambda)}{m}}.$$ 

Since $D_1 \leq \frac{2C^2 R D(\lambda)}{\lambda} + 18M^2$, it follows that

$$\sqrt{\frac{2D_1 \log(1/\delta) D(\lambda)}{m}} \leq \sqrt{\log(1/\delta)} \left( \frac{2C D(\lambda)}{\sqrt{m} \lambda} + 6M \sqrt{D(\lambda)} \right),$$ 

hence

$$\varepsilon^* \leq \frac{8C^2 R D(\lambda)}{3m \lambda} D(\lambda) + \frac{72M^2 \log(1/\delta)}{3m} + 2C \sqrt{\frac{\log(1/\delta)}{m \lambda}} D(\lambda) + \frac{3M \log(1/\delta)}{m} + 3M D(\lambda) \leq \left( \frac{8C^2 R D(\lambda)}{3m \lambda} + 2C \sqrt{\frac{\log(1/\delta)}{m \lambda}} + 3M \right) D(\lambda) + \frac{(24M^2 + 3M) \log(1/\delta)}{m}.$$ 

This proves the proposition.

2.3. Bounding the sample error $S_2$. For any $f \in B_R$, we define the random variable on $\mathcal{Z}$ as follows

$$\tilde{\xi}(z) = -\sigma^2 \exp\{- (y - f(x))^2 / \sigma^2 \} + \sigma^2 \exp\{- (y - f_\rho(x))^2 / \sigma^2 \}.$$ 

Proposition 2. Assume the capacity condition (1.4). Let $R \geq 1$, $0 < \delta < 1$, $\frac{1}{m} \leq \lambda \leq 1$ and $\sigma \geq 1$ satisfy

$$\sigma \geq C_1 R m^{\frac{m-1}{m}} (\log 2/\delta)^{-\frac{1}{2}}. \quad (2.4)$$

Then there exists a subset of $U_R$ of $\mathcal{Z}^m$ with measure at most $\delta$ such that for every $z \in W(R) \setminus U_R$ we have

$$\|f_{z,\lambda} - f_\rho\|_{L^2} + \lambda \|f_{z,\lambda}\|_{K} \leq C_2 (R^2 m^{-\frac{m-1}{m}} + D(\lambda) + \frac{(D(\lambda))^2}{\lambda^2 \sigma^2}) \log^2 \frac{2}{\delta},$$

where $C_1$ and $C_2$ are constants independent of $R$, $m$, $\sigma$ or $\lambda$.

In order to prove proposition 2 we need following lemma. We use the method from [5] to prove this lemma.

Lemma 4. $\sigma \geq 1$, $R \geq 1$. If $\varepsilon$ satisfies

$$\varepsilon \geq \sigma^{-2} (12M^4 + 4C^4 R^4), \quad (2.5)$$
then we have

\[
\text{Prob} \left\{ \sup_{f \in B_R} \frac{\mathbb{E}(\xi) - \frac{1}{m} \sum_{i=1}^{m} \xi(z_i)}{\sqrt{\mathbb{E}(\xi) + 2 \varepsilon}} > 4 \sqrt{\varepsilon} \right\} \leq N \left( \frac{\varepsilon}{(C_K + 3M)R} \right) \exp \left\{ -\frac{3m \varepsilon}{10(C_K + 3M)^2 R^2} \right\}.
\] (2.6)

Proof. \( \mathbb{E}(\hat{\xi}) = \mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho) \). From Lemma 1, we have

\[
\mathbb{E}(\hat{\xi}) + 2 \varepsilon \geq \mathcal{E}(f) - \mathcal{E}(f_\rho) + \varepsilon, \quad \forall f \in B_R.
\] (2.7)

Following from the proof of Proposition 1, we know that

\[
|\hat{\xi}| \leq (C_K + 3M)^2 R^2, \quad |\hat{\xi} - \mathbb{E}(\hat{\xi})| \leq 2(C_K + 3M)^2 R^2,
\]

\[
\mathbb{E}(\hat{\xi}^2) \leq (C_K + 3M)^2 R^2 \|f - f_\rho\|^2_{L_2 X}.
\]

Taking \( \{f_j\}_{j=1}^{J} \) to be an \( \frac{\varepsilon}{(2C_K + 3M)R} \)-net of the set \( B_R \) with \( J \) being the covering number \( N(B_R, \frac{\varepsilon}{(2C_K + 3M)R}) \), let

\[
\mu = \sqrt{\mathcal{E}^\sigma(f_j) - \mathcal{E}^\sigma(f_\rho) + 2 \varepsilon}.
\]

Applying the one-sided Bernstein’s inequality to the following group of random variables

\[
\hat{\xi}_j(z) = -\sigma^2 \exp\{- (y - f_j(x))^2 / 2 \sigma^2\} + \sigma^2 \exp\{- (y - f_\rho(x))^2 / 2 \sigma^2\}, j = 1, \ldots, J,
\]

we have the following conclusion

\[
\text{Prob}_{z \in Z^m} \left\{ \frac{(\mathcal{E}^\sigma(f_j) - \mathcal{E}^\sigma(f_\rho)) - (\mathcal{E}^\sigma(f_j) - \mathcal{E}^\sigma(f_\rho))}{\sqrt{\mathcal{E}^\sigma(f_j) - \mathcal{E}^\sigma(f_\rho) + 2 \varepsilon}} > \sqrt{\varepsilon} \right\}
\]

\[
\leq \exp \left\{ -\frac{3m \varepsilon \mu^2}{4(C_K + 3M)^2 R^2 \sqrt{\varepsilon} \mu + 6(C_K + 3M)^2 R^2 \|f_j - f_\rho\|^2_{L_2 X}} \right\}
\]

\[
\leq \exp \left\{ -\frac{3m \varepsilon \mu^2}{4(C_K + 3M)^2 R^2 \sqrt{\varepsilon} \mu + 6(C_K + 3M)^2 R^2 \mu^2} \right\}
\]

\[
\leq \exp \left\{ -\frac{3m \varepsilon}{10(C_K + 3M)^2 R^2} \right\};
\]

note that \( \varepsilon \geq \sigma^{-2}(12M^4 + 4C_K^4)R^4 \), by Lemma 1

\[
\mu^2 = \mathcal{E}^\sigma(f_j) - \mathcal{E}^\sigma(f_\rho) + 2 \varepsilon \geq \mathcal{E}(f_j) - \mathcal{E}(f_\rho) + \varepsilon \geq \varepsilon.
\]

For any \( f \in B_R \), there exists some \( j \), such that \( \|f - f_j\|_{\infty} \leq \frac{\varepsilon}{(2C_K + 3M)R} \). Hence \( |\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_j)| \) and \( |\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho)| \) can be bounded by \( \varepsilon \), then we have following inequalities

\[
\frac{|(\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho)) - (\mathcal{E}^\sigma(f_j) - \mathcal{E}^\sigma(f_\rho))|}{\sqrt{\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho) + 2 \varepsilon}} \leq \sqrt{\varepsilon}, \quad (2.8)
\]

\[
\frac{|(\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho)) - (\mathcal{E}^\sigma(f_j) - \mathcal{E}^\sigma(f_\rho))|}{\sqrt{\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho) + 2 \varepsilon}} \leq \sqrt{\varepsilon}. \quad (2.9)
\]
Together with the fact that \( \varepsilon < \mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho) + 2\varepsilon \), we have
\[
\mathcal{E}^\sigma(f_j) - \mathcal{E}^\sigma(f_\rho) + 2\varepsilon = (\mathcal{E}^\sigma(f_j) - \mathcal{E}^\sigma(f_\rho)) - (\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho)) + \mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho) + 2\varepsilon \\
\leq \sqrt{\varepsilon \mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho)} + 2\varepsilon + \mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho) + 2\varepsilon \\
\leq 2(\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho) + 2\varepsilon).
\] (2.10)

For any \( f \in B_R \), if we have the following inequation
\[
\frac{(\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho)) - (\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho))}{\sqrt{\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho) + 2\varepsilon}} > 4\sqrt{\varepsilon},
\]
then together with (2.8), (2.9) and (2.10), we have
\[
\frac{(\mathcal{E}^\sigma(f_j) - \mathcal{E}^\sigma(f_\rho)) - (\mathcal{E}^\sigma(f_j) - \mathcal{E}^\sigma(f_\rho))}{\sqrt{\mathcal{E}^\sigma(f_j) - \mathcal{E}^\sigma(f_\rho) + 2\varepsilon}} > \sqrt{\varepsilon}.
\]

Using the above estimates, we have the following conclusion
\[
\mathbb{P}_{\varepsilon \in \mathbb{Z}^m} \left\{ \sup_{f \in B_R} \frac{(\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho)) - (\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho))}{\sqrt{\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho) + 2\varepsilon}} > 4\sqrt{\varepsilon} \right\}
\]
\[
\leq \sum_{j=1}^{N} \mathbb{P}_{\varepsilon \in \mathbb{Z}^m} \left\{ \frac{(\mathcal{E}^\sigma(f_j) - \mathcal{E}^\sigma(f_\rho)) - (\mathcal{E}^\sigma(f_j) - \mathcal{E}^\sigma(f_\rho))}{\sqrt{\mathcal{E}^\sigma(f_j) - \mathcal{E}^\sigma(f_\rho) + 2\varepsilon}} > \sqrt{\varepsilon} \right\}
\]
\[
\leq N \left( B_R; \frac{\varepsilon}{(C_K + 3M)R} \right) \exp \left\{ - \frac{3m\varepsilon}{10(C_K + 3M)^2R^2} \right\}.
\]

This proves the lemma. \( \square \)

Now we begin to prove the Proposition 2.

**Proof of Proposition 2.** First we use the capacity condition (1.4) to bound the covering number in (2.6). We get the inequality
\[
C_q \left( \frac{(2C_K + 2M)R^2}{\varepsilon} \right)^q - \frac{3m\varepsilon}{10(C_K + 3M)^2R^2} \leq \log \frac{\delta}{2}.
\] (2.11)

The above inequality has the smallest positive solution can be bounded as
\[
\varepsilon^* := C_0 R^2 m^{\frac{2}{1+q}} \log^2 \frac{2}{\delta},
\]
where \( C_0 := \max\{ \frac{20}{3}(C_K + 3M)^2, (\frac{20C_0(C_K + 3M)^2}{3(2C_K + 2M)^4})^{\frac{1}{1+q}} \} \).

When \( \sigma \) satisfies
\[
\sigma \geq C_1 Rm^{\frac{4}{3(1+q)}} (\log 2/\delta)^{\frac{1}{2}}.
\] (2.12)

with
\[
C_1 = ((12M^4 + 4C_K)/C^0)^{\frac{1}{2}},
\]
the number \( \varepsilon^* \) satisfies the restriction (2.5) in Lemma 4, and with confidence \( 1 - \frac{\delta}{2} \), there holds
\[
\sup_{f \in B_R} \frac{(\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho)) - (\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho))}{\sqrt{\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f_\rho) + 2\varepsilon^*}} \leq 4\sqrt{\varepsilon^*},
\]
which implies that for every \( f \in B_R \)
\[
(E^\sigma(f) - E^\sigma(f_\rho)) - (E^\sigma_w(f) - E^\sigma_w(f_\rho)) \\
\leq 4\sqrt{\varepsilon^*} \sqrt{E^\sigma(f) - E^\sigma(f_\rho) + 2\varepsilon^*} \\
\leq \frac{1}{2}(E^\sigma(f) - E^\sigma(f_\rho)) + 9\varepsilon^*.
\]

By Proposition 1 and Lemma 2, we know that there exists a subset of \( U_R \) of \( Z^m \) with measure at most \( \delta \) such that for \( z \in W(R)\setminus U_R \), we have
\[
E^\sigma(f_{z,\lambda}) - E^\sigma(f_\rho) + \lambda\|f_{z,\lambda}\|_K^2 \\
\leq \frac{1}{2}(E^\sigma(f_{z,\lambda}) - E^\sigma(f_\rho)) + 9\varepsilon^* + \left(\frac{4C^2K\log^22}{m\lambda} + 3M + 1\right)D(\lambda) \\
+ \left(\frac{24M^2 + 3M}{m}\right)\log^22 + D(\lambda) + \sigma^{-2}(12M^4 + 4C^4K)\log^2\lambda.
\]
which yields
\[
E^\sigma(f_{z,\lambda}) - E^\sigma(f_\rho) + \lambda\|f_{z,\lambda}\|_K^2 \\
\leq 18\varepsilon^* + \left(\frac{8C^2K\log^22}{m\lambda} + 6M + 4\right)D(\lambda) \\
+ \left(\frac{48M^2 + 6M}{m}\right)\log^22 + \sigma^{-2}(24M^4 + 8C^4K)\log^2\lambda.
\]
Applying Lemma 1 to \( f = f_{z,\lambda} \), using \( \lambda \geq 1/m \) and the condition (2.12) for \( \sigma \), we see that for \( z \in W(R)\setminus U_R \)
\[
E(f_{z,\lambda}) - E(f_\rho) + \lambda\|f_{z,\lambda}\|_K^2 \\
\leq E^\sigma(f_{z,\lambda}) - E^\sigma(f_\rho) + \lambda\|f_{z,\lambda}\|_K^2 + \sigma^{-2}(12M^4 + 4C^4K)R^4 \\
\leq C_2(R^2m^{-\frac{1}{1+\rho}} + D(\lambda) + \frac{(D(\lambda))^2}{\lambda^2\sigma^2})\log^2\frac{2}{\delta},
\]
where the constant \( C_2 \) is given by
\[
C_2 = 18C_0 + 48M^2 + 6M + C_1^{-2}(36M^4 + 4C^4K) \\
+ 2(8C^2K + 6M + 4) + 16C^4K.
\]
This proves the proposition. \( \square \)

Note that for any \( z \in Z^m \), we have \( E^\sigma_w(f_{z,\lambda}) + \lambda\|f_{z,\lambda}\|_K^2 \leq E^\sigma_w(0) \).
\[
\lambda\|f_{z,\lambda}\|_K^2 \leq \frac{1}{m} \sum_{i=1}^{m} \sigma^2(1 - e^{-\frac{y^2}{\sigma^2}}) \\
\leq \frac{1}{m} \sum_{i=1}^{m} \sigma^2(\frac{y^2}{\sigma^2}) \leq M^2.
\]
This means \( \|f_{z,\lambda}\|_K \leq \frac{M}{\sqrt{\lambda}} \), \( W(\frac{M}{\sqrt{\lambda}}) = Z^m \).

**Proof of Theorem 1.** From Proposition 2, let \( R = \frac{M}{\sqrt{\lambda}} \), with confidence \( 1 - \delta \), there holds
\[
\|f_{z,\lambda} - f_\rho\|_{\rho_x}^2 + \lambda\|f_{z,\lambda}\|_K^2 \leq C_2(\frac{M^2}{\lambda}m^{-\frac{1}{1+\rho}} + D(\lambda) + \frac{(D(\lambda))^2}{\lambda^2\sigma^2})\log^2\frac{2}{\delta}.
\]
This proves the theorem. \( \square \)
Proof of Theorem 2. From Theorem 1 and (1.5), with confidence $1 - \delta$, there holds

$$\|f_{z,\lambda} - f_\rho\|_{L_2^2} \leq C_2 \left( \frac{M^2}{\lambda} m^{\frac{2-\beta}{1+q}} + C_\beta \lambda^{\beta} + C_\beta \lambda^{2\beta-2}\sigma^{-2} \right) \log^2 \frac{2}{\delta}. \quad (2.13)$$

Choose $\sigma \geq m^{\frac{2-\beta}{1+q}}$, we have

$$C_\beta \lambda^{2\beta-2}\sigma^{-2} \leq C_\beta \lambda^{\beta}.$$ 

Take $\lambda = m^{-\frac{1}{1+q}}$ in (2.13), then we have

$$\|f_{z,\lambda} - f_\rho\|_{L_2^2} = O(m^{-\frac{1}{1+q}}).$$

This proves the theorem.

3. Improving the error rate by iteration method. In this section we improve our error rate stated in Theorem 2. We will use the iteration technique which was introduced in [17].

Proof of Theorem 3. From the assumption (1.5) and Proposition 2, for any $0 < \delta < 1$, there exists a subset $U_R$ of $\mathcal{Z}^m$ with measure at most $\delta$, such that for every $z \in W(R) \setminus U_R$, we have

$$\mathcal{E}(f_{z,\lambda}) - \mathcal{E}(f_\rho) + \lambda \|f_{z,\lambda}\|_K^2 \leq C_2 \left( R^2 m^{-\frac{2-\beta}{1+q}} + C_\beta \lambda^{\beta} + \frac{(C_\beta)^2 \lambda^{2\beta-2}}{\sigma^2} \right) \log^2 \frac{2}{\delta}. $$

Denote

$$u_{m,\delta} = \sqrt{C_2 m^{-\frac{2-\beta}{1+q}}} \lambda^{-\frac{1}{2}} \sqrt{\log^2 \frac{2}{\delta}},$$

and

$$v_{m,\delta} = \sqrt{C_2 C_\beta} + \sqrt{C_2 C_\beta^2} (\lambda^{\frac{\beta-1}{2}} + \sigma^{-1} \lambda^{\beta-\frac{3}{2}}) \sqrt{\log^2 \frac{2}{\delta}},$$

we have following inequality

$$\|f_{z,\lambda}\|_K \leq u_{m,\delta} R + v_{m,\delta}.$$ 

Hence

$$W(R) \subseteq W(u_{m,\delta} R + v_{m,\delta}) \bigcup U_R. \quad (3.1)$$

Now we define a sequence $\{R^{(k)}\}_{k=0}^{\infty}$ by $R^{(0)} = \frac{M}{\sqrt{\lambda}}$ and

$$R^{(k)} = u_{m,\delta} R^{(k-1)} + v_{m,\delta}, \quad k \in \mathbb{N}.$$ 

Because $\|f_{z,\lambda}\|_K \leq \frac{M}{\sqrt{\lambda}}$, $\mathcal{Z}^m = W(R^{(0)})$. Now we take $\lambda = m^{\alpha - \frac{1}{1+q}}$, $0 < \alpha < \frac{1}{1+q}$, in order to get

$$u_{m,\delta} \leq \sqrt{C_2} m^{-\frac{2-\beta}{2}} \sqrt{\log^2 \frac{2}{\delta}} \leq \frac{1}{2},$$

we restrict $m$ by

$$m \geq 2^{\frac{2}{\alpha}} \left( C_2 \log \frac{2}{\delta} \right)^{\frac{1}{2}}. \quad (3.2)$$

Since $\sigma \geq \lambda^{\beta-1}$, we get

$$v_{m,\delta} \leq 2 (\sqrt{C_2 C_\beta} + \sqrt{C_2 C_\beta^2}) m^{(1-\beta)(\frac{2-\beta}{1+q})} \sqrt{\log^2 \frac{2}{\delta}}. \quad (3.3)$$
It is easy to obtain the following equality from the definition of the \( \{R(k)\} \)
\[
R^{(k)} = u_{m, \delta}^{k} R^{(0)} + v_{m, \delta} \sum_{j=0}^{k-1} u_{m, \delta}^{j}.
\] (3.4)

For any \( k \in \mathbb{N} \), we have
\[
R^{(k)} \leq \frac{1}{2} R^{(0)} + 2 v_{m, \delta} \\
\leq \frac{M}{2 \sqrt{\lambda}} + 4(\sqrt{C_2C_\beta} + \sqrt{C_2C_\beta^2}) \alpha(1 - \beta)(\frac{1}{\alpha(1+q)} - \frac{2}{\alpha \beta(1+q)}) \sqrt{\log \frac{2}{\delta}} \\
\leq (M + 4(\sqrt{C_2C_\beta} + \sqrt{C_2C_\beta^2}) \alpha(1 - \beta)(\frac{1}{\alpha(1+q)} - \frac{2}{\alpha \beta(1+q)}) \sqrt{\log \frac{2}{\delta}} \\
=: C_3 \alpha(1 - \beta)(\frac{1}{\alpha(1+q)} - \frac{2}{\alpha \beta(1+q)}) \sqrt{\log \frac{2}{\delta}}.
\]

If we take
\[
m \geq C_1^2 C_3^2,
\] (3.5)
then
\[
\sigma \geq C_1 C_3 m^{\frac{1}{\alpha(1+q)} - \frac{2}{\alpha \beta(1+q)}},
\]
that means for any \( R = R^{(k)} \) with \( k \geq 0 \), the condition (2.4) for \( \sigma \) is satisfied and the inclusion (3.1) holds true. Hence we have \( W(R^{(k-1)}) \subset W(R^{(k)}) \cup U_{R^{(k-1)}} \) with the measure of the set \( U_{R^{(k-1)}} \) at most \( \delta \). Thus for \( k = 1, 2, \ldots, L \) with \( L \) satisfying
\[
\frac{1}{\alpha(1+q)} - 1 \geq \frac{1}{\alpha \beta(1+q)} - 1,
\] we have
\[
Z^m = W(R^{(0)}) \subseteq W(R^{(1)}) \cup U_{R^{(0)}} \subseteq \cdots \subseteq W(R^{(L)}) \cup \left\{ \bigcup_{k=0}^{L-1} U_{R^{(k)}} \right\}.
\]

It is easy to know that the set \( W(R^{(L)}) \) has measure at least \( 1 - L \delta \geq 1 - \frac{\delta}{\alpha(1+q)} \).

Now we need to bound \( R^{(L)} \). From (3.4) we have
\[
R^{(L)} \leq u_{m, \delta}^{L} R^{(0)} + 2 v_{m, \delta} \\
\leq C_2^L m^{-\frac{L \beta}{\alpha}} \alpha(1 - \beta)(\frac{1}{\alpha(1+q)} - \frac{2}{\alpha \beta(1+q)}) \sqrt{\log \frac{2}{\delta}} \\
+ 4(\sqrt{C_2C_\beta} + \sqrt{C_2C_\beta^2}) \alpha(1 - \beta)(\frac{1}{\alpha(1+q)} - \frac{2}{\alpha \beta(1+q)}) \sqrt{\log \frac{2}{\delta}} \\
\leq C_4 \alpha(1 - \beta)(\frac{1}{\alpha(1+q)} - \frac{2}{\alpha \beta(1+q)}) \sqrt{\log \frac{2}{\delta}},
\]
where
\[
C_4 = C_2^L M + 4(\sqrt{C_2C_\beta} + \sqrt{C_2C_\beta^2}).
\]
Because \( L \geq \frac{1}{\alpha(1+q)} - 1 \geq \frac{\beta}{\alpha(1+q)} - 1 \), we have
\[
m^{-\frac{L \beta}{\alpha}} m^{\frac{1}{\alpha(1+q)} - \frac{2}{\alpha \beta(1+q)}} \leq m^{(1 - \beta)(\frac{1}{\alpha(1+q)} - \frac{2}{\alpha \beta(1+q)})},
\]
which has been used in the last inequality. For
\[
m \geq m^* := \max \left\{ \left(4C_2 \log \frac{2}{\delta} \right)^{\frac{1}{\beta}}, C_1^2 C_3^2 \right\},
\]
we apply Theorem 1 with $R = R^{(L)}$, and with confidence at least $1 - (L + 1)\delta$,
\[
\|f_{\mathbf{z}, \lambda} - f_\rho\|^2_{L^2_{\mathbf{z}_X}} + \lambda \|f_{\mathbf{z}, \lambda}\|^2_K \leq C_2(C_4^2 + C_\beta + C_3^2) m^{\alpha - \frac{\alpha}{1+\gamma}} \left(\log \frac{2}{\delta}\right)^{\frac{2}{1+\gamma}}.
\]
For $m < m^*$, we simply apply the bound $\|f_{\mathbf{z}, \lambda}\|_K \leq \frac{M}{\sqrt{\alpha}}$ and know that
\[
\|f_{\mathbf{z}, \lambda} - f_\rho\|^2_{L^2_{\mathbf{z}_X}} \leq 2\|f_{\mathbf{z}, \lambda}\|^2_\infty + 2\|f_\rho\|^2_\infty \\
\leq \frac{2C_2^2 M^2}{\lambda} + 2M^2 \\
= 2C_2^2 M^2 m^{\frac{\alpha}{1+\gamma} - \alpha} + 2M^2 \\
\leq (2C_2^2 + 2) M^2 (m^*)^{\frac{\alpha}{1+\gamma}} m^{\alpha - \frac{\alpha}{1+\gamma}}.
\]
Hence, in both cases, with confidence at least $1 - (L + 1)\delta$, we have
\[
\|f_{\mathbf{z}, \lambda} - f_\rho\|^2_{L^2_{\mathbf{z}_X}} \leq C_5(C_6) m^{\frac{2}{1+\gamma}} m^{\alpha - \frac{\alpha}{1+\gamma}} \left(\log \frac{2}{\delta}\right)^{\frac{2}{1+\gamma}},
\]
where
\[
C_5 = C_2(C_4^2 + C_\beta + C_3^2) + (2C_2^2 + 2) M^2,
\]
and
\[
C_6 = \max\{4C_2, C_1^2 C_3^2\}.
\]
Scaling $(1 + L)\delta$ to $\delta$, we get the desired error bound. This complete the proof of Theorem 3.

4. Conclusion and future work. In this paper, we consider the regularized learning algorithm for regression associated with the correntropy induced losses in reproducing kernel Hilbert spaces. The main target is the error analysis for the regression problem in learning theory based on the maximum correntropy. From our analysis, we obtain satisfactory learning rates, when choosing a suitable parameter of the loss function. The rates depend on the regularization error and on the covering numbers of the reproducing kernel Hilbert space. We can see that the scale parameter in the loss function balances the convergence rates and the robustness of the model. From [9], there are many other robust loss functions which are not sensitive to outliers, such as:

- Huber’s loss: $\psi_\sigma(t) = t^2 \chi_{(|t| \leq \sigma)} + (2\sigma|t| - \sigma^2) \chi_{(|t| > \sigma)}$;
- Cauchy loss: $\psi_\sigma(t) = \sigma^2 \log(1 + t^2/\sigma^2)$;
- Tukey’s biweight loss: $\psi_\sigma(t) = (\sigma^2/6)(1 - (t/\sigma)^2)^3 \chi_{(|t| \leq \sigma)} + (\sigma^2/6) \chi_{(|t| > \sigma)}$,

where $\chi$ is an indicator function.

Naturally, we want to consider the following regularized learning algorithm for regression problem with robust loss functions
\[
f_{\mathbf{z}} = \arg\min_{f \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m \psi_\sigma(y_i - f(x_i)) + \lambda \|f\|^2_{L^2_{\mathbf{z}_X}} ,
\]
where $\psi_\sigma$ is any robust loss function.

However, for a robust loss function, there is no general method to get the learning rates for (4.1). This is our main future work.
Acknowledgement. The work described in this paper is supported by the Natural Science Foundation of Zhejiang Province [Project No.LY15A010003]. The authors would like to thank the anonymous reviewers for their valuable suggestions. The corresponding author is Bingzheng Li.

REFERENCES

[1] R. J. Bessa, V. Miranda and J. Gama, Entropy and correntropy against minimum square in offline and online three-day ahead wing power forecasting, *Power Systems, IEEE Transactions on*, 24 (2009), 1657–1666.

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

[3] J. Fan, T. Hu, Q. Wu and D. X. Zhou, Consistency analysis of an empirical minimum error entropy algorithm, *Applied and Computational Harmonic Analysis*, 41 (2016), 164–189.

[4] Y. L. Feng, J. Fang and J. A. K. Suykens, A statistical learning approach to modal regression, *Journal of Machine Learning Research*, 2020.

[5] Y. L. Feng, X. L. Huang, L. Shi and J. A. K. Suykens, Learning with the maximum correntropy criterion induced losses for regression, *Journal of Machine Learning Research*, 16 (2015), 993–1034.

[6] Y. L. Feng and Y. M. Ying, Learning with correntropy-induced losses for regression with mixture of symmetric stable noise, *Applied and Computational Harmonic Analysis*, 48 (2020), 795–810.

[7] R. He, W. S. Zheng and B. G. Hu, Maximum correntropy criterion for robust face recognition, *IEEE Transactions on pattern Analysis and Machine Intelligence*, 33 (2011), 1561–1576.

[8] R. He, W. S. Zheng, B. G. Hu and X. W. Kong, A regularized correntropy framework for robust pattern recognition, *Neural Computation*, 23 (2011), 2074–2100.

[9] P. J. Huber, *Robust Statistics*, John Wiley & Sons, 1981.

[10] T. Hu, J. Fan, Q. Wu and D. X. Zhou, Learning theory approach to minimum error entropy criterion, *Journal of Machine Learning Research*, 14 (2013), 377–397.

[11] T. Hu, J. Fan, Q. Wu and D. X. Zhou, Regularization schemes for minimum error entropy principle, *Analysis and Applications*, 13 (2015), 437–455.

[12] B. Z. Li, Approximation by multivariate Bernstein-Durrmeyer operators and learning rates of least-square regularized regression with multivariate polynomial kernels, *J. Approx. Theory*, 173 (2013), 33–55.

[13] W. Liu, P. P. Pokharel and J. C. Príncipe, Correntropy: Properties and application in non-gaussian signal processing, *IEEE Transactions on Signal Processing*, 55 (2007), 5286–5298.

[14] F. S. Lv and J. Fan, Optimal Learning with Gaussians and Correntropy Loss, *Analysis and Applications*, 2020.

[15] I. Santamaría, P. P. Pokharel and J. C. Príncipe, Generalized correlation function: Definition, properties, and application to blind equalization, *IEEE Transactions on Signal Processing*, 54 (2006), 2187–2197.

[16] S. Smale and D. X. Zhou, Estimating the approximation error in learning theory. *Analysis and Applications*, 1 (2003), 17–41.

[17] I. Steinwart and C. Scovel, Fast rates for support vector machines, *Lecture Notes in Computer Science*, 3559 (2005), 270–294.

[18] V. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, 1998.

[19] E. D. Vito, L. Rosasco, A. Caponnetto and U. D. Giovannini, Learning from Examples as an Inverse Problem, *Journal of Machine Learning Research*, 6 (2005), 883–904.

[20] C. Wang and D. X. Zhou, Optimal learning rates for least squares regularized regression with unbounded sampling, *Journal of Complexity*, 27 (2011), 55–67.

[21] Q. Wu, Y. Ying and D. X. Zhou, Learning rates of least-square regularized regression, *Foundations of Computation Mathematics*, 6 (2006), 171–192.

[22] Q. Wu, Y. Ying and D. X. Zhou, Multi-kernel regularized classifiers, *Journal of Complexity*, 23 (2007), 108–134.

[23] Y. M. Ying and D. X. Zhou, Learnability of Gaussians with flexible variances, *Journal of Machine Learning Research*, 8 (2007), 249–276.

[24] D. X. Zhou, The covering number in learning theory, *Journal of Complexity*, 18 (2002), 739–767.
[25] D. X. Zhou, Capacity of reproducing kernel spaces in learning theory, *IEEE Transactions on Information Theory*, 49 (2003), 1743–1752.

Received for publication December 2019.

E-mail address: libingzheng@zju.edu.cn
E-mail address: 382149444@qq.com