KERNEL-BASED MAXIMUM CORRENTROPY CRITERION WITH GRADIENT DESCENT METHOD

TING HU
School of Mathematics and Statistics
Wuhan University, Wuhan, China

Abstract. In this paper, we study the convergence of the gradient descent method for the maximum correntropy criterion (MCC) associated with reproducing kernel Hilbert spaces (RKHSs). MCC is widely used in many real-world applications because of its robustness and ability to deal with non-Gaussian impulse noises. In the regression context, we show that the gradient descent iterates of MCC can approximate the target function and derive the capacity-dependent convergence rate by taking a suitable iteration number. Our result can nearly match the optimal convergence rate stated in the previous work, and in which we can see that the scaling parameter is crucial to MCC’s approximation ability and robustness property. The novelty of our work lies in a sharp estimate for the norms of the gradient descent iterates and the projection operation on the last iterate.

1. Introduction. Correntropy [19] is a widely used concept in information theoretical learning (ITL) [18], which is referred to as a robust nonlinear similarity measure between two random variables. Maximum correntropy criterion (MCC) is a principle induced by correntropy and provides a family of supervised learning algorithms. Recently, MCC has received increasing attention in domains of machine learning and signal processing for its effectiveness in the cases when data are disturbed by impulsive noise or containing large outliers [6, 7, 10, 15]. Therefore, it usually serves as a robust alternative of the traditional least squares method or other convex optimization algorithms in learning processes.

In the framework of non-parametric estimation, denote $X$ as the explanatory variable that takes values in a separable metric space $X$, $Y \in \mathcal{Y} \subset \mathbb{R}$ as a real valued response variable and let $\rho$ be the underlying distribution on $Z := X \times \mathcal{Y}$. Here we focus on the application of MCC for regression that is linked to the additive data model

$Y = f_\rho(X) + e, \quad \mathbb{E}(e|X) = 0,$

where $e$ is the noise and $f_\rho(x)$ is the regression function, that is the conditional mean $\mathbb{E}(Y|X = x)$ for given $X = x \in X$. The purpose of regression analysis is to study the functional relationship between $X$ and $Y$. It leads to estimate the target function $f_\rho$ according to a set of samples $z = \{(x_i, y_i)\}_{i=1}^m \subset Z := X \times \mathcal{Y}$ drawn independently from $\rho$. For the least squares method, the loss function $g^*(u) = u^2$, $u \in \mathbb{R}$ is employed to measure the quality of the estimator. This method aims to
minimize the variance of error variable and belongs to the second-order statistics. Its optimality relies heavily on the assumption of Gaussian distributions. However, non-Gaussian noise is ubiquitous in real world applications, for instance, artificial noise in electronic devices, atmospheric noises, and lighting spikes in natural phenomena [17, 25]. Thus, the least squares method is not a good choice from a robustness point of view when the occurrence of outliers presents in the data. This motivates the application of MCC into the regression problems and the reference therein [5, 6, 18].

For a hypothesis function \( f : X \rightarrow Y \), with the scaling parameter \( \sigma > 0 \), correntropy between \( f(X) \) and \( Y \) is defined by

\[
V_\sigma(f) := E \left[ G \left( \frac{(f(X) - Y)^2}{2\sigma^2} \right) \right]
\]

where \( G(u) \) is the Gaussian function \( \exp \left\{ -\frac{u^2}{2\sigma^2} \right\} \), \( u \in \mathbb{R} \). Given the sample set \( z \), the empirical form of \( V_\sigma \) is

\[
\hat{V}_\sigma(f) := \frac{1}{m} \sum_{i=1}^{m} G \left( \frac{(f(x_i) - y_i)^2}{2\sigma^2} \right).
\]

When applied to regression problems, MCC aims to maximize the empirical correntropy \( \hat{V}_\sigma \) over a hypothesis space \( H \), that is

\[
f_{z, H} := \arg \max_{f \in H} \hat{V}_\sigma(f). \tag{1.1}
\]

This scheme has succeeded in many real-world applications for its robustness to outliers or heavy-tailed distributions, e.g., wind power forecasting and pattern recognition [2, 9]. To better understand MCC in the statistical learning framework, we define the loss induced by correntropy \( \phi_\sigma : \mathbb{R} \rightarrow \mathbb{R}^+ \) as

\[
\phi_\sigma(u) := \sigma^2 \left( 1 - G \left( \frac{u^2}{2\sigma^2} \right) \right) = \sigma^2 \left( 1 - \exp \left\{ -\frac{u^2}{2\sigma^2} \right\} \right),
\]

where \( \sigma > 0 \) is the scaling parameter. It is a fact that the loss function is a variant of the Welsch function [11] and the estimator \( f_{z, H} \) of (1.1) is also the minimizer of the following empirical minimization risk scheme over \( H \)

\[
\min_{f \in H} \frac{1}{m} \sum_{i=1}^{m} \phi_\sigma(f(x_i) - y_i). \tag{1.2}
\]

It should be noted that \( \phi_\sigma \) is not convex and satisfies the redescending property, that is, \( \phi_\sigma' \) is non-decreasing near the origin, but decreasing toward 0 far from the origin. Various empirical and theoretical studies [4, 6, 8, 13, 24] have shown that non-convex losses with the redescending property can yield robust estimators. These work has further confirmed the effectiveness of MCC to deal with outliers or heavy-tailed distributions.

From a computational perspective, the optimization problem arising from MCC is non-convex and the gradient descent method is a usual way to implement it. The purpose of this paper is to study the convergence performance of the gradient descent estimator for solving the empirical risk minimization scheme (1.2), where an RKHS is taken as the hypothesis space \( H \). The recent work [8] also investigated the asymptotic behavior of the gradient descent algorithm generated by a family of robust loss functions. However, their analysis depends on the specific feature of the linear integral operator associated with the RKHS, that is restrictive in many situations. In our work, the rigorous error analysis is in terms of the covering...
number of the RKHS, which has been studied well in a lot of literatures. Here the convergence rate is capacity-dependent and nearly coincides with the optimal rate for regression problems. We also find that the projected last iterate can be a good estimator of the target function $f_ρ$ and as a byproduct, taking the iteration method, we have refined the uniform bound for norms of the gradient descent iterates in previous work [8, 12, 14].

The rest of paper is organized as follows. In Section 2, we introduce some necessary preliminaries and state our main results. Discussions and comparisons with related studies will be presented in Section 3. Section 4 is devoted to the proofs of our results.

2. Preliminaries and main results. Before giving our main results, we introduce some necessary preliminaries and notations. Throughout the paper, let $\rho_X$ be the marginal distribution of $\rho$ on $\mathcal{X}$ and $\rho(\cdot|x)$ be the conditional distribution on $\mathcal{Y}$ for given $x \in \mathcal{X}$. Denote $\|\cdot\|_\rho$ as the $L^2$-norm in $L^2_{\rho_X}$ space, which is defined by $\|\cdot\|_\rho = \|\cdot\|_{L^2_{\rho_X}} := \left(\int_{\mathcal{X}} \|\cdot\|^2 d\rho_X\right)^{\frac{1}{2}}$. Furthermore, assume that the response variable $Y$ is uniformly bounded, i.e. for some constant $B > 0$, $|Y| \leq B$.

Let $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Mercer kernel [1], i.e. a continuous, symmetric and positive semi-definite function. We say that $K$ is positive semi-definite, if for any finite set $\{u_1, \cdots, u_m\} \subset \mathcal{X}$ and $m \in \mathbb{N}$, the matrix $(K(u_i, u_j))_{i,j=1}^m$ is positive semi-definite. The RKHS $\mathcal{H}_K$ associated with the Mercer kernel $K$ is defined to be the completion of the linear span of the set of functions $\{K_x := K(x, \cdot), x \in \mathcal{X}\}$ with the inner product $\langle \cdot, \cdot \rangle_K$ given by $(K_x, K_u)_K = K(x, u)$. It has the reproducing property

$$f(x) = \langle f, K_x \rangle_K$$

for any $f \in \mathcal{H}_K$ and $x \in \mathcal{X}$. Denote $\kappa := \sup_{x \in \mathcal{X}} \sqrt{K(x, x)}$. By the property (2.1), we get that

$$\|f\|_\infty \leq \kappa \|f\|_K, \text{ for any } f \in \mathcal{H}_K. \quad (2.2)$$

We are now ready to state the gradient descent method for solving the scheme (1.2). Note that the loss induced by correntropy $\phi_\sigma$ is differentiable for any $u \in \mathbb{R}$. This together with the reproducing property (2.1), we can calculate the functional gradient of (1.2) at $\mathcal{H}_K$ space and define the gradient descent algorithm for kernel MCC as follows.

**Definition 1.** Given the i.i.d. sample $z = \{(x_i, y_i)\}_{i=1}^m \subset \mathcal{Z}$, the gradient descent algorithm for MCC regression starts with $f_1 = 0$, and for any $2 \leq t \leq T$,

$$f_{t+1} = f_t - \frac{\eta_t}{m} \sum_{i=1}^m \phi'_\sigma((f_t(x_i) - y_i))K_x,$$  

(2.3)

where $\{\eta_t > 0, t \in \mathbb{N}\}$ is the sequence of step sizes and $\phi'_\sigma((f_t(x_i) - y_i)) = G(\frac{(f_t(x_i) - y_i)^2}{2\sigma^2})(f_t(x_i) - y_i)$.

The purpose of this paper is to investigate the convergence of the iterates $\{f_t\}_{t=1}^{T+1}$ generated by (2.3) when the iteration time $T = T(m)$ is chosen properly according to the sample size $m$. A regression task is the concern of the algorithm (2.3), it is natural to use the mean squared error $\|\cdot - f_\rho\|_\rho^2$ to measure its effectiveness or equivalently the excess generalization error with the least squares loss $\phi^{ls}$. 

---

**Kernel MCC with Gradient Descent**

4161
Definition 2. With the least squares loss \( \phi^s(u) = u^2 \), \( u \in \mathbb{R} \), the generalization error for any \( f : \mathcal{X} \rightarrow \mathbb{R} \) is defined by

\[
\mathcal{E}(f) := \int_{\mathcal{Z}} \phi^s(f(x) - y)d\rho = \int_{\mathcal{Z}} (f(x) - y)^2d\rho.
\]

Obviously, the target function \( f_\rho \) is the minimizer of \( \mathcal{E}(f) \) over the measurable space, that is

\[
f_\rho(x) = \mathbb{E}(Y|X = x) = \int_{\mathcal{Y}} yd\rho(y|x), \quad x \in \mathcal{X}.
\]

The excess generalization error for any \( f : \mathcal{X} \rightarrow \mathbb{R} \) is defined by \( \mathcal{E}(f) - \mathcal{E}(f_\rho) \), and it is direct to calculate that

\[
\mathcal{E}(f) - \mathcal{E}(f_\rho) = \int_{\mathcal{Z}} (f(x) - f_\rho)^2d\rho_X = \|f - f_\rho\|_\rho^2. \tag{2.4}
\]

Recall that \( |Y| \leq B \). It implies that the value of \( f_\rho \) is supported on \([-B, B]\) almost surely. So, we will choose the last iterate \( f_{T+1} \) with \( |f_{T+1}| \leq B \) to approximate \( f_\rho \). Then we can make full use of the projection method introduced in [3].

Definition 3. For any \( f : \mathcal{X} \rightarrow \mathbb{R} \), the projected function \( \hat{f} \) is defined by

\[
\hat{f}(x) = \begin{cases} 
-B, & \text{if } f(x) < -B, \\
 f(x), & \text{if } |f(x)| \leq B, \\
 B, & \text{if } f(x) > B.
\end{cases}
\]

It is easy to see that for any fixed \( x \in \mathcal{X}, y \in \mathcal{Y}, |\hat{f}(x) - f_\rho(x)| \leq |f(x) - f_\rho(x)| \) and

\[
\phi^s(\hat{f}(x) - y) \leq \phi^s(f(x) - y). \tag{2.5}
\]

These confirm that the projection operation will lead to better estimators. The same idea can also be found in [23, 26]. Thus, in what follows, our approximation analysis aims at establishing bounds for \( \|\hat{f}_{T+1} - f_\rho\|_\rho^2 \) instead of \( \|f_{T+1} - f_\rho\|_\rho^2 \).

Our convergence results will be stated under some conditions on the triplet \((\mathcal{H}_K, f_\rho, \rho_\mathcal{X})\). The first condition is about the decay of the approximation error [20].

Assumption 1. Let \( \lambda > 0 \) and \( f_\lambda \) be the minimizer of the regularization error:

\[
f_\lambda := \arg \min_{f \in \mathcal{H}_K} \mathcal{E}(f) + \lambda\|f\|_K^2. \tag{2.6}
\]

The approximation error associated with the triplet \((\mathcal{H}_K, f_\rho, \rho_\mathcal{X})\) is defined by

\[
\mathcal{D}(\lambda) := \min_{f \in \mathcal{H}_K} \mathcal{E}(f) - \mathcal{E}(f_\rho) + \lambda\|f\|_K^2 = \mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho) + \lambda\|f_\lambda\|_K^2. \tag{2.7}
\]

We assume that for some \( \beta \in (0, 1] \) and \( c_\beta > 0 \), there holds

\[
\mathcal{D}(\lambda) \leq c_\beta \lambda^\beta, \quad \lambda > 0. \tag{2.8}
\]

The above assumption is common in the literature of learning theory. Assumption (2.8) always holds with \( \beta = 0 \). When the target function \( f_\rho \in \mathcal{H}_K \) and \( \mathcal{H}_K \) is dense in \( C(\mathcal{X}) \) which consists of bounded continuous functions on \( \mathcal{X} \), the approximation error \( \mathcal{D}(\lambda) \) goes to 0 as \( \lambda \rightarrow 0 \). Thus, the decay (2.8) is natural and can be illustrated in terms of interpolation spaces [20].

The second condition is related to the capacity of the hypothesis space \( \mathcal{H}_K \) which is measured by empirical covering numbers of balls in the RKHS \( \mathcal{B}_R = \).
\{f \in \mathcal{H}_K : \|f\|_K \leq R\}$. Recall that for a subset $\mathcal{G}$ of a metric space $(\mathcal{H}, d)$, the covering number $\mathcal{N}(\mathcal{G}, \varepsilon, d)$ is defined by

$$\mathcal{N}(\mathcal{G}, \varepsilon, d) = \inf \left\{ \ell \in \mathbb{N} : \exists f_1, \cdots, f_\ell \in \mathcal{H} \text{ such that } \mathcal{G} \subset \bigcup_{i=1}^{\ell} \{f \in \mathcal{G} : d(f, f_i) \leq \varepsilon\} \right\}.$$

**Assumption 2.** Let $\mathcal{G}$ be a set of functions on $\mathcal{X}$. The metric $d_{2, \mathcal{X}}$ on $\mathcal{G}$ is defined by

$$d_{2, \mathcal{X}}(f, g) := \left\{ \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - g(x_i))^2 \right\}^{\frac{1}{2}}, \quad \forall f, g \in \mathcal{G}.$$

We assume that for some $\zeta \in (0, 2)$, $c_{\zeta} > 0$, the covering number of the unit ball $B_1$ in $\mathcal{H}_K$ with respect to $d_{2, \mathcal{X}}$ satisfies

$$E_{\mathcal{X}} \left[ \log \mathcal{N}(\mathcal{G}, \varepsilon, d_{2, \mathcal{X}}) \right] \leq c_{\zeta} \left( \frac{1}{\varepsilon} \right)^{\zeta}, \quad \forall \varepsilon > 0. \quad (2.9)$$

Besides empirical covering numbers, there are other standard tools to measure the complexity of the RKHS: uniform covering numbers (covering numbers of $\mathcal{G}$ under the metric $\| \cdot \|_\infty$), entropy numbers and effective dimension associated the integral operator $L_K : L^2_{\rho_X} \rightarrow L^2_{\rho_X}$, given by $L_K(f) = \int_X f(x) K_x d\rho_X$, $\forall f \in L^2_{\rho_X}$. For the discussion of their equivalence, we can refer to the papers [23, 14]. The analysis of this paper is mainly achieved by a concentration technique which involves the empirical covering number and often leads to sharp error bounds [26].

Our first theorem below provides a novel bound for the iterates $\{f_t\}_{t=1}^{T+1}$ on $\mathcal{H}_K$-norm, which improves previous uniform bounds for the iterates generated by various gradient descent algorithms. Without losing generality, let $\kappa = 1$ for simplicity.

**Theorem 1.** Let $0 < \lambda \leq T^{-(1-\theta)}$ and $\lambda = m^{-\gamma}$, where

$$\gamma = \min \left\{ \frac{2}{\zeta + 2\beta}, 1 \right\}. \quad (2.10)$$

If the scaling parameter $\sigma^2 \geq T^{3(1-\theta)} \lambda^{1-\beta}$, then with confidence at least $1 - \delta$, there holds for any $\varepsilon > 0$,

$$\sup_{1 \leq t \leq T} \|f_{t+1}\|_K \leq c_5 \lambda^{\frac{\sigma^2}{\lambda} - \frac{\varepsilon}{\lambda}} \left( \log \frac{4\ell}{\delta} \right)^{\frac{1}{2}}, \quad (2.11)$$

where $c_5, \ell$ are constants independent of $\beta, \delta, \lambda, m, \sigma$ (given in the proof).

Taking $\lambda = T^{-(1-\theta)}$, we can see that the upper bound for $\sup_{1 \leq t \leq T} \|f_{t+1}\|_K$ is nearly $O(T^{(1-\theta)(1-2\beta)-\varepsilon})$ (ignoring the logarithmic factor). This bound is better than the uniform bound $O(T^{1-\beta})$ in the existing literature [8, 14, 12] and will help us get a shaper estimate for the gradient descent method.

**Theorem 2.** Suppose that Assumptions 1 and 2 hold with $\beta \in (0, 1]$ and $\zeta \in (0, 2)$. Taking the iteration number $T = m^\alpha$ with

$$\alpha = \min \left\{ \frac{2}{(1-\theta)(\zeta + 2\beta)} \cdot \frac{1}{1-\theta} \right\}. \quad (2.12)$$
If $\sigma^2 \geq T^{(2+\beta)(1-\theta)}$, then for any $\epsilon > 0$, $0 < \delta < 1$, with confidence at least $1 - \delta$, there holds
\[
\left\| \hat{f}_{T+1} - f_\rho \right\|^2_\rho \leq C_\epsilon \left( m^{-\min\{\frac{2\beta}{2+\beta}, \epsilon, \beta \}} \log \frac{12T}{\delta} + \frac{T^{3-2\theta}}{\sigma^2} \right)
\] (2.13)
where the constant $C_\epsilon$ independent of $m$ (given in the proof) and $\gamma$ is defined in (2.10).

From this theorem, we see that the value of the scaling parameter $\sigma$ will affect the convergence rate. The relatively large $\sigma$ will lead to the rate of order $m^{-\min\{\frac{2\beta}{2+\beta}, \epsilon, \beta \}}$, which nearly matches the best capacity-dependent rate for regression problems by ignoring arbitrarily small $\epsilon$ (([26] and p. 268, [23])). In particular, when $f_\rho \in H_K$, i.e. $\beta = 1$, the rate is $O(m^{-\frac{2\beta}{2+\beta}})$. To further illustrate our result, we give the following example when the regression function $f_\rho$ lies in some range space.

Example 1. Suppose that Assumption 2 holds with $\zeta \in (0, 2]$, and $f_\rho \in L^2_{\hat{\rho}_K}(L^2_{\rho_X})$ with $0 < \beta \leq 1$. Let $\eta_t = \eta t^{-\theta}$ with $0 \leq \theta < 1$. If the iteration time $T = m^{-\frac{1}{2+\beta}}$ and $\sigma^2 > m^{3-2\theta+(1-\theta)\beta} \alpha$, then with confidence at least $1 - \delta$, there holds
\[
\left\| \hat{f}_{T+1} - f_\rho \right\|^2_\rho = O \left( m^{-\min\{\frac{2\beta}{2+\beta}, \epsilon, \beta \}} \right),
\] (2.14)
for any $\epsilon > 0$.

Due to the compactness of $X$, the integral operator $L_K : L^2_{\rho_X} \to L^2_{\rho_X}$ induced by Mercer kernel $K$ is self-adjoint, positive and trace class on $L^2_{\rho_X}$, as well as restricting on $H_K$. By the spectral theorem, the range space $L^2_{\hat{\rho}_K}(L^2_{\rho_X})$ is well defined. The index $\beta$ measures the complexity of $f_\rho$. The bigger the index $\beta$, the higher the regularity of $f_\rho$. As we expect, the obtained convergence rate will improve if $\beta$ increases.

3. Discussions and comparison with related work. In this section, we discuss and compare our results with related work. In the paper [6], Feng et al. studied the consistency of the empirical scheme (1.2) over a compact subspace $H \subset C(X)$ and obtained the explicit convergence rate under some complexity conditions on $H$. In the case of $f_\rho \in H_K$, under Assumption 2, their obtained rate is of order $O(m^{-\frac{2\beta}{2+\beta}})$ with a suitable chosen $\sigma$, which nearly coincides our result in (2.13) by ignoring small $\epsilon$. When $f_\rho$ is out of $H_K$, by tracing their proof of Theorem 4 carefully, we find that with confidence $1 - \delta$, their rate is bounded by
\[
\inf_{f \in B_R} \| f - f_\rho \|^2_\rho + O \left( R^4(\sigma^{-2} + \sigma m^{-\frac{1}{2+\beta}}) \log \frac{2}{\delta} \right)
\]
where $p$ is a power index of the polynomial increasing condition for the logarithm of the uniform covering number $N(B_1, \epsilon, C(X))$. Let $R = ||f_\lambda||_K$, where the regularization parameter $\lambda$ is the tradeoff for the above error terms. So, the rate is up to $O(m^{-\frac{1}{2+\beta}})$ by taking $\sigma = m^{\frac{1}{2+\beta}}$ and $\lambda = m^{-\frac{1}{2+\beta}}$. Note that for any function set $F \subset C(X)$, the empirical covering number $N(B_1, \epsilon, C(X))$ is bounded by $N(B_1, \epsilon, C(X))$ since $d_{2, \lambda}(f, g) \leq ||f - g||_\infty$. Obviously, our rate $O(m^{-\min\{\frac{2\beta}{2+\beta}, \epsilon, \beta \}})$ in Theorem 1 is much better when $\sigma$ is chosen properly.

Denote the set of positive eigenvalues of $L_K : L^2_{\rho_X} \to L^2_{\rho_X}$ as $\{\lambda_i\}_i$, arranged in a decreasing order. Under the assumption that the eigenvalues $\lambda_i \leq ci^{-\frac{2}{\beta}}$ with
some $c > 0$ and $\zeta' \in (0, 2)$ and $|Y| \leq B$ almost surely, Guo et al. [8] investigated the performance of the last iterates generated by the gradient descent method associated with a family of robustness losses. Applying their results to the loss $\phi$ induced by correntropy, if $f_{\rho} \in L_2^\beta(L_2^\rho)$ with $0 < \beta \leq 1$, then with confidence at least $1 - \delta$, their rate is at most of order $O(m^{-\frac{\beta}{\beta + \zeta' \sigma}})$. According to the equivalence of the decay rate for the eigenvalues $\{\lambda_i\}$ and empirical covering numbers, as shown in [14, 22, 23], we know that Assumption 2 holds with $\zeta = \zeta'$ if $\lambda_i = O(i^{-\frac{2}{\beta}})$. So, their rate is worse than that in (2.14). In addition, it should be mentioned that their work was carried out under the assumption that the target function $f_0$ lies in the range of some power of the integral operator $L_K$. This assumption is strong for many commonly used RKHSs. For example, if $K$ is Gaussian, then it requires $f_\rho \in C^\infty$, that is restrictive for real problems. Our conditions in this paper are more general.

As mentioned in Section 1, MCC is usually taken as a good alternative of the least squares method. In what follows, we compare with the previous work on the least squares method. We first review the results on Tikhonov regularization with the least square loss function, i.e.

$$f_{z,\lambda} = \arg \min_{f \in \mathcal{H}_K} \left\{ e_n(f) + \lambda\|f\|_K^2 \right\}, \quad \lambda > 0.$$ 

For a subset $\mathcal{G}$ of a metric space $(\mathcal{H}, d)$, the $n$-th entropy number is defined by

$$e_n(\mathcal{G}, d) = \inf \left\{ \varepsilon > 0 : \exists f_1, \ldots, f_{2^n - 1} \text{ such that } \mathcal{G} \subset \bigcup_{i=1}^{2^n - 1} \{ f \in \mathcal{G} : d(f, f_i) \leq \varepsilon \} \right\}$$

Under the assumptions that the expectation of the random entropy number $E_n[\mathbb{E}_n(\mathcal{G}, \varepsilon, d_{2^n, z})] \leq c(\frac{1}{n})^{\frac{1}{2}}$ with some $c > 0$ and $\zeta' \in (0, 2)$, it has been proved in [23] (Page 268) that with high probability, there holds

$$\|\hat{f}_{z,\lambda} - f_\rho\|_\rho^2 = O\left( m^{-\min\left\{ \frac{2\beta}{\beta + \zeta' \sigma}, \frac{\beta}{\beta + \zeta' \sigma} \right\}} \right).$$

Note that the covering number and entropy number are equivalent (see [23], Lemma 6.21). Replacing the above decay assumption for $E_n[\mathbb{E}_n(\mathcal{G}, \varepsilon, d_{2^n, z})]$ with Assumption 2, the rate for $\|\hat{f}_{z,\lambda} - f_\rho\|_\rho^2$ still holds with $\zeta' = \zeta$, that is of order $O(m^{-\min\left\{ \frac{2\beta}{\beta + \zeta' \sigma}, \frac{\beta}{\beta + \zeta' \sigma} \right\}})$. This has been confirmed by the work [26], which studied the Tikhonov regularization with multi-kernels. To our knowledge, it is the best rate for the Tikhonov regularized empirical risk minimization with the least square loss without any extra regularity condition is imposed on $f_\rho$. Our result in (2.13) nearly matches it up to an arbitrarily small $\varepsilon$. Next we turn to iteration regularization scheme for the least squares loss [14], in which no constrain or penalization term is considered and the number of iterations serves as a regularization parameter. Like the algorithm (2.3) in our paper, it can be regarded as a gradient descent method to implement the empirical risk minimization for the least square. Under the Assumptions 1, 2 and $|Y| \leq B$, it has been proved in [14] that the error rate for the last iterative of iteration regularization scheme is of order $O(m^{-\min\left\{ \frac{2\beta}{\beta + \zeta' \sigma}, \frac{\beta}{\beta + \zeta' \sigma} \right\}})$ if $f_\rho \in L_2^\beta(L_2^\rho)$ with $\beta \in (0, 1]$. It is obviously inferior to our rate in (2.14).

At the end of this section, we would like to point out that, although our analysis was achieved by taking relatively large $\sigma$, the established convergence rates indicate that MCC’s implementation with the gradient descent method is feasible in
regression problems. It has been empirically reported that the robustness of the regression models induced by $\phi_\sigma$ can be enhanced with a decreasing value of $\sigma$. Thus, the scaling parameter $\sigma$ balances the robustness of the MCC algorithm (2.3) and its convergence rates, which plays a trade-off role in MCC’s applications for regression problems.

4. Error estimations and proofs. In this section, we prove the results presented in Section 2. Let us start with the error decomposition for the algorithm (2.3).

4.1. Error decomposition. By (2.4), to estimate $\|\hat{f}_{T+1} - f_\rho\|^2_{\rho}$, our goal is equivalent to bound the excess generalization error $\mathcal{E}(\hat{f}_{T+1}) - \mathcal{E}(f_\rho)$, which is split as

$$\mathcal{E}(\hat{f}_{T+1}) - \mathcal{E}(f_\rho) = \left\{ \mathcal{E}(\hat{f}_{T+1}) - \mathcal{E}(f_\rho) - \left[ \mathcal{E}_z(\hat{f}_{T+1}) - \mathcal{E}_z(f_\rho) \right] \right\} + \mathcal{E}_z(\hat{f}_{T+1}) - \mathcal{E}_z(f_\rho)$$

(4.1)

where $\mathcal{E}_z(f)$ is the empirical generalization error of $\mathcal{E}(f)$ according to the sample $z$, given as

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

The first term of the righthand side of (4.1) is referred to as a sample error and quantifies the discrepancy between the excess generalization error of the estimator and the empirical counterpart. It can be studied by taking empirical process theory and various uniform laws of large numbers are provided to deal with it. Noting that $\hat{f}_{T+1}$ relies on the sample $z$, we shall bound the sample error in a function set, that is

$$\mathcal{M}_z(R) := \sup_{\|f\|_\mathcal{K} \leq R} \left[ \mathcal{E}(\hat{f}) - \mathcal{E}(f_\rho) - \left[ \mathcal{E}_z(\hat{f}) - \mathcal{E}_z(f_\rho) \right] \right]$$

and $R$ denotes the upper bound for the $\mathcal{H}_K$-norm of $f_{T+1}$ such that $\|f_{T+1}\|_\mathcal{K} \leq R$. Here we are able to get the sharp sample error estimate by the projection operation and concentration inequality.

The second term $\mathcal{E}_z(\hat{f}_{T+1}) - \mathcal{E}_z(f_\rho)$ is called the computation error related to optimization, which plays an essential role in our analysis. To obtain a good computation error estimate, we introduce a reference function $f^* \in \mathcal{H}_K$ and get that

$$\mathcal{E}_z(\hat{f}_{T+1}) - \mathcal{E}_z(f_\rho) = \mathcal{E}_z(\hat{f}_{T+1}) - \mathcal{E}_z(f^*) \leq \left\{ \mathcal{E}_z(f^*) - \mathcal{E}_z(f_\rho) \right\} + \mathcal{E}(f^*) - \mathcal{E}(f_\rho).$$

(4.2)

To simplify the expression, we denote

$$\mathcal{F}_z(f^*) := \left| \mathcal{E}_z(f^*) - \mathcal{E}_z(f_\rho) \right|, \quad \mathcal{A}(f^*) := \left| \mathcal{E}(f^*) - \mathcal{E}(f_\rho) \right|.$$  

Collecting the above observations, we have the following lemma.

**Lemma 1.** Suppose the bound $\|f_{T+1}\|_\mathcal{K} \leq R$ holds for some $R > 0$. For any given $f^* \in \mathcal{H}_K$, we have

$$\mathcal{E}_z(\hat{f}_{T+1}) - \mathcal{E}_z(f_\rho) \leq \mathcal{E}_z(\hat{f}_{T+1}) - \mathcal{E}_z(f^*) \mathcal{M}_z(R) + \mathcal{F}_z(f^*) + \mathcal{A}(f^*).$$

(4.3)

From this lemma, we see that the best iteration time and corresponding convergence rate are achieved by suitably balancing various error terms in the above decomposition. In next subsections, we will estimate them item by item.
4.2. Computation error for $f^\ast$. In this subsection, we estimate the computational error $\mathcal{E}_\ast(f_{T+1}^\ast) - \mathcal{E}_\ast(f^\ast)$ for an arbitrary $f^\ast \in \mathcal{H}_K$. To this end, we introduce the following lemma that gives an upper-bound of the iterates $\{f_t\}_{t=1}^{T+1}$ under the $\mathcal{H}_K$ norm. Its proof is similar to that in [8] for bounding the iterates of gradient descent algorithm with robust losses. Here we omit it for simplicity.

**Lemma 2.** Define $\{f_t\}_{t=1}^{T+1}$ by (2.3). If the step sizes $0 < \eta_t < 1$ for any $t \geq 1$, then we have

$$
\|f_{t+1}\|_K \leq B \sqrt{\sum_{j=1}^t \eta_j}, \quad t \geq 1. \tag{4.4}
$$

In particular, if $\eta_t = \eta t^{-\theta}$ with $0 < \eta \leq 1$ and $0 \leq \theta < 1$, then

$$
\|f_{t+1}\|_K \leq \frac{B}{\sqrt{1 - \theta}} t^{1-\theta}, \quad t \geq 1. \tag{4.5}
$$

The estimation of the computation error is based on some basic properties of quadratic functions, that will be shown in the proof.

**Lemma 3.** Define $\{f_t\}_{t=1}^{T+1}$ by (2.3). If $\eta_t = \eta t^{-\theta}$ with $0 < \eta \leq 1$ and $0 \leq \theta < 1$, then we have that

$$
\mathcal{E}_\ast(f_{T+1}) - \mathcal{E}_\ast(f^\ast) \leq C \left( \|f^\ast\|_K^2 T^{\theta - 1} + \frac{T^{2(1-\theta)} \|f^\ast\|_K + T^{3-2\theta}}{\sigma^2} \right), \tag{4.6}
$$

where $C$ is a constant independent of $T$, will be given in the proof.

**Proof.** Note that $b^2 = a^2 + 2a(b-a) + (b-a)^2$ for any $a, b \in \mathbb{R}$. Taking $b = f_{t+1}(x_i) - y_i$ and $a = f_t(x_i) - y_i$, then by the property (2.1),

$$
(f_{t+1}(x_i) - y_i)^2 = (f_t(x_i) - y_i)^2 + 2(f_{t+1} - f_t, (f_t(x_i) - y_i)K_{x_i})_K + (f_{t+1}(x_i) - f_t(x_i))^2.
$$

Averaging the above equality with $i = 1, \ldots, m$, by (2.2), we have

$$
\mathcal{E}_\ast(f_{t+1}) = \mathcal{E}_\ast(f_t) - 2\langle f_{t+1} - f_t, V_t \rangle_K + \frac{1}{m} \sum_{i=1}^m (f_{t+1}(x_i) - f_t(x_i))^2
$$

$$
\leq \mathcal{E}_\ast(f_t) - 2\langle f_{t+1} - f_t, V_t \rangle_K + \|f_{t+1} - f_t\|_K^2 \tag{4.7}
$$

where $V_t := \frac{1}{m} \sum_{i=1}^m (f_t(x_i) - y_i) K_{x_i}$. For simplicity, denote

$$
W_t := \frac{1}{m} \sum_{i=1}^m \phi'_\sigma((f_t(x_i) - y_i))K_{x_i} = \frac{1}{m} \sum_{i=1}^m G \left( \frac{(f_t(x_i) - y_i)^2}{2\sigma^2} \right) (f_t(x_i) - y_i)K_{x_i}.
$$

The algorithm (2.3) is rewritten as $f_{t+1} = f_t - \eta_t W_t$. Putting this into (4.7) yields that

$$
\mathcal{E}_\ast(f_{t+1}) \leq \mathcal{E}_\ast(f_t) - 2\langle f_{t+1} - f_t, W_t \rangle_K + \|W_t\|_K^2 + 2\langle f_{t+1} - f_t, W_t - V_t \rangle_K
$$

$$
\leq \mathcal{E}_\ast(f_t) - 2\eta \|W_t\|_K^2 + \eta^2 \|W_t\|_K^2 + 2\langle f_{t+1} - f_t, W_t - V_t \rangle_K
$$

$$
\leq \mathcal{E}_\ast(f_t) - \eta \|W_t\|_K^2 + 2\langle f_{t+1} - f_t, W_t - V_t \rangle_K,
$$

where the last inequality is derived by $0 < \eta \leq 1$.

Now we consider the term $\langle f_{t+1} - f_t, W_t - V_t \rangle_K$. Since the Gaussian function $G(u)$ is Lipschitz, then for any $u \in \mathbb{R}$,

$$
|G(u) - G(0)| = |G(u) - 1| \leq |u|.
$$
This together with (4.5) yields that
\[
(f_{t+1} - f_t, W_t - V_t)_K \leq \|f_{t+1} - f_t\|_K \|W_t - V_t\|_K
\]
\[
\leq (\|f_{t+1}\|_K + \|f_t\|_K) \frac{1}{m} \sum_{i=1}^{m} \frac{|f_i(x_i) - y_i|^3}{2\sigma^2}
\]
\[
\leq (\|f_{t+1}\|_K + \|f_t\|_K) \frac{\|f_\infty\|_K + B)^3}{2\sigma^2}
\]
\[
\leq \frac{8B^4}{(1 - \theta)^2\sigma^2} t^{2(1-\theta)}.
\]
Thus, we have that
\[
E_x(f_{t+1}) \leq E_x(f_t) - \eta_t \|W_t\|^2_K + \frac{16B^4}{(1 - \theta)^2\sigma^2} t^{2(1-\theta)}. \tag{4.8}
\]
According to the above inequality, it is not difficult to calculate that for any \( T > t \),
\[
E_x(f_{T+1}) \leq E_x(f_{t+1}) - \sum_{j=t+1}^{T} \eta_j \|W_j\|^2_K + \frac{16B^4}{(1 - \theta)^2\sigma^2} \sum_{j=t+1}^{T} j^{2(1-\theta)}
\]
\[
\leq E_x(f_{t+1}) + \frac{16B^4}{(1 - \theta)^2\sigma^2} \sum_{j=t+1}^{T} j^{2(1-\theta)}. \tag{4.9}
\]
Applying the elementary inequality \( b^2 \leq a^2 + 2(b-a)b \), \( a, b \in \mathbb{R} \) with \( b = (f_t(x_i) - y_i) \) and \( a = (f^*(x_i) - y_i) \), by (2.1) again, we get
\[
E_x(f_t) \leq E_x(f^*) + 2\langle f_t - f^*, V_t \rangle_K
\]
\[
= E_x(f^*) + 2\langle f_t - f^*, W_t \rangle_K + 2\langle f_t - f^*, V_t - W_t \rangle_K.
\]
Introducing this inequality into (4.8), one gets
\[
E_x(f_{t+1}) \leq E_x(f^*) \frac{1}{\eta_t} (2\eta_t \langle f_t - f^*, W_t \rangle_K - \eta_t^2 \|W_t\|^2_K) + 2\langle f_t - f^*, V_t - W_t \rangle_K
\]
\[
= E_x(f^*) + \frac{1}{\eta_t} (\|f_t - f^*\|^2_K - \|f_t - f^*\|^2_K - \eta_t W_t\|^2_K) + 2\langle f_t - f^*, V_t - W_t \rangle_K
\]
\[
= E_x(f^*) + \frac{1}{\eta_t} (\|f_t - f^*\|^2_K - \|f_{t+1} - f^*\|^2_K) + 2\langle f_t - f^*, V_t - W_t \rangle_K.
\]
It follows that
\[
\eta_t (E_x(f_{t+1}) - E_x(f^*)) \leq \|f_t - f^*\|^2_K - \|f_{t+1} - f^*\|^2_K + 2\eta_t \langle f_t - f^*, V_t - W_t \rangle_K. \tag{4.10}
\]
Summing up over \( t = 1, \ldots, T \) with \( f_1 = 0 \),
\[
\sum_{t=1}^{T} \eta_t (E_x(f_{t+1}) - E_x(f^*))
\]
\[
\leq \|f_1 - f^*\|_K - \|f_{T+1} - f^*\|_K + 2 \sum_{t=1}^{T} \eta_t \langle f_t - f^*, V_t - W_t \rangle_K
\]
\[
\leq \|f^*\|_K^2 + 2 \sum_{t=1}^{T} \eta_t \langle f_t - f^*, V_t - W_t \rangle_K. \tag{4.11}
\]
For the term \( \langle f_t - f^*, V_t - W_t \rangle_K \), using (4.5) and the Lipschitz property of \( G \) again, we find that
\[
\langle f_t - f^*, V_t - W_t \rangle_K \leq (\|f_t\|_K + \|f^*\|_K) \left( \frac{\|f_t\|_K + B}{2\sigma^2} \right)^3 \\
\leq \frac{8B^4}{(1-\theta)^2} t^2 (1-\theta) + \frac{t^{\frac{3(1-\theta)}{2}}}{\sigma^2} \|f^*\|_K := \Delta_t. \quad (4.12)
\]
So, we get
\[
\sum_{t=1}^{T} \eta_t (\mathcal{E}_t(f_{t+1}) - \mathcal{E}_t(f^*)) \leq \|f^*\|_K^2 + 2 \sum_{t=1}^{T} \eta_t \Delta_t.
\]
Plugging (4.9) into the above relation, we have
\[
\sum_{t=1}^{T} \eta_t (\mathcal{E}_t(f_{t+1}) - \mathcal{E}_t(f^*)) \\
\leq \sum_{t=1}^{T} \eta_t (\mathcal{E}_t(f_{t+1}) - \mathcal{E}_t(f^*)) + \frac{16B^4}{(1-\theta)^2\sigma^2} \sum_{t=1}^{T} \eta_t \sum_{j=t+1}^{T} j^{2(1-\theta)} \\
\leq \|f^*\|_K^2 + 2 \sum_{t=1}^{T} \eta_t \Delta_t + \frac{16B^4}{(1-\theta)^2\sigma^2} \sum_{t=1}^{T} \eta_t \sum_{j=t+1}^{T} j^{2(1-\theta)}. \quad (4.13)
\]
Dividing both sides of the above inequality by \( \sum_{t=1}^{T} \eta_t \), so
\[
\mathcal{E}_t(f_{t+1}) - \mathcal{E}_t(f^*) \leq \frac{\|f^*\|_K^2}{\sum_{t=1}^{T} \eta_t} + 2 \sup_{1 \leq t \leq T} \Delta_t + \frac{16B^4}{(1-\theta)^2\sigma^2} \sup_{1 \leq t \leq T} \sum_{j=t+1}^{T} j^{2(1-\theta)}.
\]
Noting the inequality \( \sum_{t=1}^{T} t^{-\theta} \geq \frac{T^{1-\theta}}{\epsilon} \), then
\[
\mathcal{E}_t(f_{t+1}) - \mathcal{E}_t(f^*) \leq C \left( \frac{T^{\frac{1-\theta}{1-(\hat{x}+\epsilon)}}}{T^{\frac{1-\theta}{1-(\hat{x}+\epsilon)}}} + \frac{T^{\frac{2(1-\theta)}{2(1-\theta)}}}{\sigma^2} \|f^*\|_K + \right) \\
\text{with } C = \frac{\epsilon}{\theta} + \frac{48B^4}{(1-\theta)^2\sigma^2}.
\]
Note by (2.5) that \( \mathcal{E}_t(f_{t+1}) - \mathcal{E}_t(f^*) \leq \mathcal{E}_t(f_{t+1}) - \mathcal{E}_t(f^*) \), then we can get the bound (4.6) based on the above estimate. The proof is completed. \( \square \)

4.3. **Estimations involving empirical process theory.** In this subsection, we proceed estimating the error terms \( \mathcal{M}_x(R) \) and \( \mathcal{F}_x(f^*) \) in (4.3), which depend on random sampling and can be studied by empirical process theory. To this end, we need some probability inequalities. The first one is the concentration inequality for the random variables in a family of functions, which can be found in [14, 26].

**Lemma 4.** Let \( \mathcal{G} \) be a set of measurable functions on \( Z \), and \( M, c > 0, \tau \in [0, 1] \) be constants such that each functions \( g \in \mathcal{G} \) satisfies \( \|g\|_\infty \leq M \) and \( E(g^2) \leq c (Eg)^{\tau} \). If for some \( a \geq M^2 \) and \( \zeta \in (0, 2) \),
\[
E_\mathcal{X} [\log N(\mathcal{G}, \epsilon, d_{2,1})] \leq a \left( \frac{1}{\epsilon} \right)^\zeta, \quad \forall \epsilon > 0.
\]
\[
(4.14)
\]
then there exists a positive $c'_\zeta$ depending only on $\zeta$ such that for any $b > 0$, with probability at least $1 - e^{-b}$, there holds
\[
\mathbb{E}_g - \frac{1}{m} \sum_{i=1}^{m} g(z_i) \leq \frac{1}{2} R^2 \mathbb{E}_g^\top + c'_\zeta \eta + 2 \left( \frac{c}{m} \right)^{1/2} \mathbb{E}_g^\top + \frac{18Mb}{m}, \quad \forall g \in \mathcal{G},
\]
where
\[
\eta := \max \left\{ \left( \frac{a}{m} \right)^{1/2}, \left( \frac{a}{m} \right)^{1/2 + \tau'}, M \right\}.
\]

The second probabilistic inequality is a variant of one-side Bernstein inequality for the single random variable, whose proof will be provided in Appendix.

**Lemma 5.** Let $\{\xi(z_i)\}_{i=1}^{m}$ be a set of functions defined on $\mathcal{Z}$ and $M, c > 0, \tau \in [0, 1]$ be the constant such that $\|\xi\|_{\infty} \leq M$ and $\mathbb{E}(\xi^2) \leq c(\mathbb{E}\xi)^\tau$, then for any $0 < \delta < 1$, with confidence at least $1 - \delta$, there holds
\[
\frac{1}{m} \sum_{i=1}^{m} \xi(z_i) - \mathbb{E}(\xi) \leq 2 \log \frac{1}{\delta} \max \left\{ \frac{M}{m}, \left( \frac{c}{m} \right)^{1/2 + \tau'}, \mathbb{E}\xi \right\}.
\]

With the above preparations, we are first ready to estimate $\mathcal{M}_\sharp(R)$.

**Lemma 6.** Suppose Assumptions 1 and 2 hold. For any $f \in \mathcal{B}_R, R \geq 1$, with confidence at least $1 - \delta$,
\[
\mathcal{M}_\sharp(R) \leq \frac{1}{2} \left[ \mathcal{E}(\hat{f}) - \mathcal{E}(f_{\rho}) \right] + C_1 \max \left\{ \left( \frac{R^\kappa}{m} \right)^{2 + \tau'}, \frac{1}{m} \log \frac{1}{\delta} \right\},
\]
where the constant $C_1$ is given in the proof.

**Proof.** We apply Lemma 4 to the function set
\[
\mathcal{G} = \left\{ g(x, y) := \phi(f(x) - y) - \phi(f_{\rho}(x) - y), f \in \mathcal{B}_R \right\}.
\]
Recall the fact $\|f_{\rho}\| \leq B$ and $|Y| \leq B$, we can get that
\[
\|g\|_{\infty} \leq \sup_{(x, y) \in Z} \left\{ |\hat{f}(x) - f_{\rho}(x)| \hat{f}(x) + f_{\rho}(x) - 2y| \right\} \leq 8B^2 := M,
\]
and
\[
\mathbb{E}(g^2) \leq \sup_{(x, y) \in Z} |\hat{f}(x) + f_{\rho}(x) - 2y|^2 \int_{\mathcal{X}} |\hat{f}(x) - f_{\rho}|^2 d\rho \leq 16B^2 \mathbb{E} g
\]
So, each $g \in \mathcal{G}$ satisfies that $\mathbb{E}(g^2) \leq c(\mathbb{E}g)^\tau$ with $\tau = 1$ and $c = 16B^2$. Since for any $f, h \in \mathcal{B}_R$,
\[
|\phi(y, \hat{f}(x)) - \phi(y, \hat{h}(x))| \leq \sup_{(x, y) \in Z} |\hat{f}(x) + \hat{h}(x) - 2y| |\hat{f}(x) - \hat{h}(x)| \leq 4B|f(x) - h(x)|,
\]
then
\[
\mathcal{N}(\mathcal{G}, \epsilon, d_{2, \sharp}) \leq \mathcal{N}(B_R, \frac{\epsilon}{4B}, d_{2, \sharp}) \leq \mathcal{N}(B_1, \frac{\epsilon}{4BR}, d_{2, \sharp}).
\]
It means the covering number condition holds with $a = c_\zeta(4BR)^{\kappa}$. Thus all the conditions in the Lemma 4 hold. We have that with confidence at least $1 - \delta$, for all $g \in \mathcal{G}$, there holds
\[
\mathbb{E} g - \frac{1}{m} \sum_{i=1}^{m} g(z_i) \leq \frac{1}{2} \mathbb{E} g + c_\zeta \eta + \frac{32B^2}{m} \log \frac{1}{\delta} + \frac{144B^2}{m} \log \frac{1}{\delta}
\]
where
\[ \eta = \max \left\{ (16B^2)^2 \left( \frac{c\xi(4BR)^{\xi}}{m} \right)^{\frac{2}{\lambda + 1}}, (8B^2)^2 \left( \frac{c\xi(4BR)^{\xi}}{m} \right)^{\frac{2}{\lambda + 1}} \right\} \leq C_1 \left( \frac{R^2}{m} \right)^{\frac{2}{\lambda + 1}}, \]
and \( C_1 = (16B^2)^2 \left( \frac{c\xi(4BR)^{\xi}}{m} \right)^{\frac{2}{\lambda + 1}} \). So, we get that with confidence at last \( 1 - \delta \), there holds for \( f \in B_R \),
\[ \left[ \mathcal{E}(\hat{f}) - \mathcal{E}(f_\rho) \right] - \left[ \mathcal{E}_\xi(\hat{f}) - \mathcal{E}_\xi(f_\rho) \right] \leq \frac{1}{2} \left[ \mathcal{E}(\hat{f}) - \mathcal{E}(f_\rho) \right] + c\xi\eta + \frac{176B^2}{m} \log \frac{1}{\delta}, \]
which yields the desired conclusion (4.15) with \( C_1 = c\xi C_1' + 176B^2 \). The proof is completed.

Next, we move on the estimation of the term \( \mathcal{F}_\xi(f^*) \). Observing the expression of \( \mathcal{F}_\xi(f^*) \), to get a better estimate by Lemma 5, we want to choose a reference function \( f^* \) in \( H_K \) that has a small generalization error. So, a natural choice of \( f^* \) can be \( f_\lambda \), the minimizer of the regularized generalization error with a free regularization parameter \( \lambda \). With this choice, the last error term in (4.3) can be controlled as
\[ A(f^*) = A(f_\lambda) \leq D(\lambda). \]
Keeping this in mind, we also give the estimation of \( \mathcal{F}_\xi(f_\lambda) \) for any \( \lambda > 0 \) as follows.

**Lemma 7.** Suppose Assumptions 1 holds. For any \( f^* \in H_K \) and \( \|f^*\|_K \leq \bar{R} \) with some constant \( \bar{R} \geq 0 \), then for any \( 0 < \delta < 1 \), with confidence at least \( 1 - \delta \), there holds
\[ \mathcal{F}_\xi(f^*) \leq C_2 \max \left\{ \frac{\bar{R}^2}{m}, \frac{1}{m}, A(f^*) \right\} \log \frac{2}{\delta}, \quad (4.16) \]
specially, let \( f^* = f_\lambda \), then we have with confidence at least \( 1 - \delta \),
\[ \mathcal{F}_\xi(f_\lambda) \leq C_2' \max \left\{ \frac{\lambda^{\beta-1}}{m}, \frac{1}{m}, \lambda^{\beta} \right\} \log \frac{2}{\delta}, \quad (4.17) \]
where \( C_2', C_2 \) are constants independent of \( m, \delta, \bar{R}, \lambda \) and will be given in the proof.

**Proof.** Decompose \( \mathcal{F}_\xi(f^*) \) into
\[ \mathcal{F}_\xi(f^*) = \left\{ \mathcal{E}_\xi(f^*) - \mathcal{E}_\xi(\hat{f}) \right\} + \left\{ \mathcal{E}_\xi(\hat{f}) - \mathcal{E}_\xi(f_\rho) \right\} \]
\[ := S_1 + S_2. \]
Firstly, we consider \( S_1 \). Let \( \xi_1(z) = \phi(f^*(x) - y) - \phi(\hat{f}^*(x) - y) \). The relation (2.5) implies that \( \xi_1(z) \geq 0 \), then
\[ \|\xi_1\|_\infty \leq \sup_{(x,y) \in Z} |\phi(f^*(x) - y)| \leq (\|f^*\|_K + B)^2 \leq (1 + B^2)\bar{R}^2, \]
and \( \mathbf{E}\xi_1^2 \leq \|\xi_1\|_\infty \mathbf{E}\xi_1 \leq 2(1 + B^2)\bar{R}^2 \mathbf{E}\xi_1 \). Applying Lemma 5 with \( \xi = \xi_1, \ c = 2(1 + B^2)\bar{R}^2, \ M = 2(1 + B^2)\bar{R}^2 \) and \( \tau = 1 \), we know that there exists a subset \( Z^{\delta,1} \subset Z \) with the measure at least \( 1 - \delta \) such that \( z \in Z^{\delta,1} \), there holds
\[ S_1 \leq 4(1 + B^2) \log \frac{1}{\delta} \max \left\{ \frac{\bar{R}^2}{m}, \mathbf{E}\xi_1 \right\}. \quad (4.18) \]
Next, we estimate $S_2$. Let $\xi_2(z) = \phi(\hat{f}^*(x) - y) - \phi(f_\rho(x) - y)$. Recall the fact $\|f_\rho\| \leq B$ and $E(\hat{f}^*) - E(f_\rho) = \|\hat{f}^* - f_\rho\|^2$. Then,

$$\|\xi_2\|_\infty \leq \|\hat{f}^* - f_\rho\|_\infty (\|\hat{f}^*\|_\infty + \|f_\rho\|_\infty + 2B) \leq 8B^2$$

and

$$E\xi_2^2 \leq (\|\hat{f}^*\|_\infty + \|f_\rho\|_\infty + 2B)^2 \int_{X} |\hat{f}^*(x) - f_\rho(x)|^2 d\rho \leq 16B^2 E\xi_2.$$ 

Applying Lemma 5 with $\xi = \xi_2$ again, we know that there exists a subset $Z^{k,2} \subset Z$ with the measure at least $1 - \delta$ such that $z \in Z^{k,2}$, there holds

$$S_2 \leq 32B^2 \log \frac{1}{\delta} \max \left\{ \frac{1}{m}, E\xi_2 \right\}. \quad (4.19)$$

Note that $E\xi_1 \leq A(f^*)$ and $E\xi_2 \leq A(f^*)$. Combing (4.18) with (4.19), we get that there exists a subset $Z^2 \subset Z^{k,1} \cap Z^{k,2}$ with the measure at least $1 - 2\delta$ such that $z \in Z^2$, there holds

$$F_z(f^*) = S_1 + S_2 \leq (4 + 36B^2) \log \frac{1}{\delta} \max \left\{ \frac{\tilde{R}^2}{m}, \frac{1}{m}, A(f^*) \right\}. \quad (4.20)$$

Scaling $2\delta$ to $\delta$, then proof of (4.16) is completed with $C_2 = 4 + 36B^2$. For the estimation of (4.17), let $f^* = f_\lambda$, then $A(f_\lambda) \leq D(f_\lambda) \leq c_\beta \lambda^\beta$ and $\|f_\lambda\|_K \leq c_\beta^{\frac{1}{2}} \lambda^{\frac{\beta+1}{2}}$ by Assumption 1. Applying (4.16) with $\tilde{R} = c_\beta^{\frac{1}{2}} \lambda^{\frac{\beta+1}{2}}$, then the proof of (4.17) is completed with $C_2' = C_2 \beta$. 

4.4. Improving the bounds for iterates. According to the above estimations for the error terms in (4.3), with the latter choice $f^* = f_\lambda$, we know that a tight estimate for the $H_K$-norm of the last iterate $f_{T+1}$ will lead a better error bound for our algorithm (2.3). We shall prove the result in Theorem 1 and obtain the refined bound by using the following lemma iteratively.

**Proposition 1.** Assume that there exists a constant $R > 1$ such that $\|f_{t+1}\|_K \leq R$ for each $t = 1, \cdots, T$. If the step sizes $\eta_t$ takes the form as $\eta_t = \eta T^{1-\theta}$ with $\eta > 0$ and $0 \leq \theta < 1$, then for any $f^* \in H_K$ and all $1 \leq t \leq T$, with confidence at least $1 - \delta$,

$$\|f_{t+1}\|_K^2 \leq 4\|f^*\|_K^2 + \frac{2\eta}{1-\theta} \left[ F_z(f^*) + A(f^*) + C_1 \Delta_R + 2\Delta_t \right] \tilde{R}^{1-\theta}, \quad (4.20)$$

where the constant $C_1$ is given in (4.15) and $\Delta_R = \max \left\{ \left( \frac{R^2}{m} \right)^{\frac{1}{2\theta}}, \frac{1}{m} \log \frac{1}{\delta} \right\}$.

**Proof.** According to (4.10) and (4.12), by (2.5), we know that for any $1 \leq j \leq T$,

$$\|f_{j+1} - f^*\|_K^2 \leq \|f_j - f^*\|_K^2 - \eta_j \left[ E_z(f_{j+1}) - E_z(f_\rho) \right] + \eta_j [E_z(f^*) - E_z(f_\rho)] + 2\eta_j \Delta_j. \quad (4.21)$$

It follows from the above inequality that for $j = 1, \cdots, T$,

$$\|f_{j+1} - f^*\|_K^2 \leq \|f_j - f^*\|_K^2 - \eta_j \left[ E_z(f_{j+1}) - E_z(f_\rho) \right] + \eta_j [E_z(f^*) - E_z(f_\rho)] + 2\eta_j \Delta_j$$

$$\leq \|f_j - f_\rho\|_K^2 - \eta_j \left[ E_z(f_{j+1}) - E_z(f_\rho) \right] + \eta_j [E_z(f^*) - E_z(f_\rho)] + 2\eta_j \Delta_j$$

$$= \|f_j - f_\rho\|_K^2 + \eta_j \left[ E_z(f_{j+1}) - E(f_\rho) \right] - \left[ E_z(f_{j+1}) - E_z(f_\rho) \right]$$

$$+ \eta_j F_z(f^*) + \eta_j A(f^*) - \eta_j \left[ E_z(f_{j+1}) - E(f_\rho) \right] + 2\eta_j \Delta_j$$
Lemma 7, then with confidence at least $1 - \delta$ such that for $z \in Z^\delta$, there holds for each $j = 1, \cdots, T$,
\[
\|f_{j+1} - f^*\|_K^2 \leq \|f_j - f^*\|_K^2 + \eta_j \|\mathcal{M}_z(R) + \mathcal{F}_z(f^*) + \mathcal{A}(f^*) + 2\Delta_j\| + \frac{1}{2}\eta_j \left[\mathcal{E}(f_{j+1}) - \mathcal{E}(f_j)\right].
\]
Together with Lemma 6, we have that there exists a subset $Z^\delta \subset Z$ with the measure at least $1 - \delta$ such that for $z \in Z^\delta$, there holds for each $j = 1, \cdots, T$,
\[
\|f_{j+1} - f^*\|_K^2 \leq \|f_j - f^*\|_K^2 + \eta_j \|\mathcal{F}_z(f^*) + \mathcal{A}(f^*) + C_1\Delta_R + 2\Delta_j\|
\]
It follows from $\mathcal{E}(f_j) - \mathcal{E}(f_0) > 0$ that when $z \in Z^\delta$, there holds
\[
\|f_{j+1} - f^*\|_K^2 \leq \|f_j - f^*\|_K^2 + \eta_j = 1 \|\mathcal{F}_z(f^*) + \mathcal{A}(f^*) + C_1\Delta_R + 2\Delta_j\|, j = 1, \cdots, T.
\]
(4.22)
Summing up over $j = 1, \cdots, t$ for any $t \leq T$ with $f_1 = 0$,
\[
\|f_{t+1} - f^*\|_K^2 \leq \|f_t\|_K^2 + \eta \sum_{j=1}^t \|\mathcal{F}_z(f^*) + \mathcal{A}(f^*) + C_1\Delta_R + 2\Delta_t\|
\]
where the last inequality is derived by $\sum_{j=1}^t t^{-\theta} \leq \frac{t^{1-\theta}}{1-\theta}$ for any $0 \leq \theta < 1$. Notice the fact $\|f_{t+1}\|_K^2 \leq 2\|f_{t+1} - f^*\|_K^2 + 2\|f^*\|_K^2$, then we can obtain the bound (4.20).
The proof is completed.

We observe that the estimators of $\mathcal{F}_z(f^*), \mathcal{A}(f^*)$ and $\Delta_t$ in (4.20) heavily depend on the choice of $f^*$. Thus, this proposition asserts that for the suitable step sizes and scaling parameter, the sequence $\|f_{t+1}\|_K, t = 1, \cdots, T$ can be well bounded if there exists some $f^* = \hat{f}_\lambda$ such that the second term of the right hand side of (4.20) is small. So, we shall choose a suitable parameter $\lambda$ to prove Theorem 1.

**Proof of Theorem 1.** By the fact $\mathcal{A}(f_\lambda) \leq \mathcal{D}(f_\lambda) \leq c_\beta \lambda^\beta$, and $\|f_\lambda\|^2_K \leq c_\beta \lambda^{\beta-1}$, applying Proposition 1 with $f^* = f_\lambda$ and $t^{1-\theta} \leq T^{1-\theta} \leq \lambda^{-1}$, we have that with confidence at least $1 - \delta$, for each $t = 1, \cdots, T$,
\[
\|f_{t+1}\|^2_K \leq \left(4 + \frac{2\eta}{1-\theta}\right) c_\beta \lambda^{\beta-1} + \frac{2\eta}{1-\theta} \left[\mathcal{F}_z(f_\lambda)\lambda^{-1} + C_1\Delta_R\lambda^{-1} + 2\Delta_{t^{1-\theta}}\right].
\]
Noting that $\sigma^2 \geq T^{3(1-\theta)} \lambda^{1-\beta}$ implies that $\Delta_{t^{1-\theta}} \leq \frac{16\beta^4}{(1-\theta)^2} \lambda^{\beta-1}$. Together with Lemma 7, then with confidence at least $1 - 2\delta$, for each $t = 1, \cdots, T$,
\[
\|f_{t+1}\|^2_K \leq \left[\left(4 + \frac{2\eta}{1-\theta}\right) c_\beta + \frac{2\eta}{1-\theta} \left(C_1 + C_2 + \frac{32\beta^4}{(1-\theta)^2}\right)\right] \lambda^{-1} \max \left\{\left(R^2 \lambda^{\beta-1}\right)^{\frac{2\gamma}{\gamma'}} \frac{1}{m} \log \frac{2}{\delta}, \frac{\lambda^{\beta-1}}{m} \log \frac{2}{\delta}, \frac{\lambda^{\beta-1} \log \frac{4}{\delta}}{\delta}\right\}.
\]
By the choice $\lambda = m^{-\gamma}$, scaling $2\delta$ to $\delta$, it follows that with confidence at least $1 - \delta$, for each $t = 1, \cdots, T$,
\[
\|f_{t+1}\|^2_K \leq \left[\left(4 + \frac{2\eta}{1-\theta}\right) c_\beta + \frac{2\eta}{1-\theta} \left(C_1 + C_2 + \frac{32\beta^4}{(1-\theta)^2}\right)\right] \lambda^{\beta-1} \max \left\{\left(R^2 \lambda^{\beta-1}\right)^{\frac{2\gamma}{\gamma'}} \log \frac{4}{\delta}\right\}.
\]
Thus, with confidence at least $1 - \delta$, for each $t = 1, \cdots, T$,
\[
\|f_{t+1}\|_K \leq c\lambda^{\frac{\sigma^2}{2\tau}} \left(\log \frac{4}{\delta}\right)^{\frac{1}{2}} + c\lambda^{\frac{\sigma^2}{2\tau}} R^{\frac{1}{\tau}}, \tag{4.23}
\]
where $c = \sqrt{\left((4 + \frac{2\eta}{1 - \delta})(C_1 + C_2 + \frac{\tau}{|\frac{\eta}{1 - \delta}|})\right)}$. It implies that with confidence at least $1 - \delta$ such that $\|f_{t+1}\|_K \leq W(R)$, $t = 1, \cdots, T$, where $W : \mathbb{R}_+ \to \mathbb{R}_+$ is defined by
\[
W(R) = c\lambda^{\frac{\sigma^2}{2\tau}} \left(\log \frac{4}{\delta}\right)^{\frac{1}{2}} + c\lambda^{\frac{\sigma^2}{2\tau}} R^{\frac{1}{\tau}}. \tag{4.24}
\]
Denote $W^{[0]}(R) = R$, $W^{[1]}(R) = W(R)$, $W^{[2]}(R) = W(W^{[1]}(R))$, $\cdots$, $W^{[l]}(R) = W(W^{[l-1]}(R))$ for $l \geq 2$.

Assume that there exists a constant such that $\|f_{t+1}\|_K \leq R_0$ for each $t = 1, \cdots, T$. So, we have that there exists a subset $Z^{\delta,1} \subset \mathcal{Z}$ with the measure at least $1 - \delta$ such that for $z = (x, y) \in Z^{\delta,1}$, there holds $\|f_{t+1}\|_K \leq W^{[1]}(R_0), t = 1, \cdots, T$.

Using (4.23) again, we have that there exists a subset $Z^{\delta,2} \subset \mathcal{Z}$ with the measure at least $1 - \delta$ such that for $z = (x, y) \in Z^{\delta,1} \cap Z^{\delta,2}$, there holds $\|f_{t+1}\|_K \leq W^{[2]}(R_0)$, $t = 1, \cdots, T, \cdots, \cdots$. When $l = 3, \cdots$, we use (4.23) iteratively, there exists a subset $Z^{\delta,l} \subset \mathcal{Z}$ with the measure at least $1 - \delta$ such that for $z = (x, y) \in Z^{\delta,1} \cap Z^{\delta,2} \cdots \cap Z^{\delta,l}$, there holds $\|f_{t+1}\|_K \leq W^{[l]}(R_0), t = 1, \cdots, T$.

Thus, we conclude that with confidence at $1 - l\delta$, for each $t = 1, \cdots, T$,
\[
\|f_{t+1}\|_K \leq W^{[l]}(R_0). \tag{4.25}
\]
Observe that $W(R_0) = w_0 + w_1 R_0^{\frac{\sigma^2}{2\tau}}$, where $w_0 = c\lambda^{\frac{\sigma^2}{2\tau}} \left(\log \frac{2}{\delta}\right)^{\frac{1}{2}}$ and $w_1 = c\lambda^{\frac{\sigma^2}{2\tau}}$ given in (4.24). It follows that
\[
W^{[2]}(R_0) = w_0 + w_1 W^{[1]}(R_0) R_0^{\frac{\sigma^2}{2\tau}} = w_0 + w_1 \left(w_0 + w_1 R_0^{\frac{\sigma^2}{2\tau}}\right) R_0^{\frac{\sigma^2}{2\tau}} \leq w_0 + w_1 w_0 R_0^{\frac{\sigma^2}{2\tau}} + w_1 R_0^{\frac{\sigma^2}{2\tau}},
\]
and whenever $l = 2, \cdots$,
\[
W^{[l]}(R_0) \leq w_1 \sum_{i=0}^{l-1} \left(\lambda^{\frac{\sigma^2}{2\tau}}\right)^i R_0^{\frac{\sigma^2}{2\tau}} + w_0 + w_1 w_0 R_0^{\frac{\sigma^2}{2\tau}} + w_1 w_0 R_0^{\frac{\sigma^2}{2\tau}} \leq w_0 \sum_{i=0}^{l-1} \lambda^{\frac{\sigma^2}{2\tau}} \left(\lambda^{\frac{\sigma^2}{2\tau}}\right)^i R_0^{\frac{\sigma^2}{2\tau}} + \sum_{i=0}^{l-1} w_1 \sum_{j=0}^{i-1} \lambda^{\frac{\sigma^2}{2\tau}} \left(\lambda^{\frac{\sigma^2}{2\tau}}\right)^j R_0^{\frac{\sigma^2}{2\tau}}.
\]
This in connection with the expressions for $w_0$ and $w_1$ gives
\[
W^{[l]}(R_0) \leq (c + 1)^{\frac{2\eta}{\delta}} \lambda^{\frac{\sigma^2}{2\tau}} \left(1 - \left(\lambda^{\frac{\sigma^2}{2\tau}}\right)^l\right)^l R_0^{\frac{\sigma^2}{2\tau}} + \sum_{i=1}^{l-1} (c + 1)^{\frac{2\eta}{\delta}} (c + 1) \left(\log \frac{4}{\delta}\right)^{\frac{1}{2}} \left(\lambda^{\frac{\sigma^2}{2\tau}}\right)^i \lambda^{\frac{\sigma^2}{2\tau}}.
\]
By the bound (4.5) and $\lambda \leq T^{-(1-\theta)}$, we take $R_0 = \lambda^{-\frac{1}{2}}$ and $W^{[0]}(R_0) := W(\lambda^{\frac{1}{2}})$, then

$$W^{[l]}(R_0) \leq (c + 1)^{\frac{\alpha + \zeta}{2}} \lambda^{\frac{\alpha - 1}{2}} \left\{ \lambda^{-\frac{\alpha}{2}} \left( \frac{\zeta}{2^{\alpha+1}} \right) + l \left( \log \frac{4}{\delta} \right)^{\frac{1}{2}} \right\}. $$

For any $\epsilon > 0$, we choose a smallest integer $l_\epsilon$ such that $l_\epsilon \geq \log \frac{1}{\epsilon} / \log \frac{2+\zeta}{\epsilon}$, then we have that

$$W^{[l_\epsilon]}(R_0) \leq (c + 1)^{\frac{\alpha + \zeta}{2}} \lambda^{\frac{\alpha - 1}{2}} \left\{ \lambda^{-\frac{\alpha}{2}} + l_\epsilon \left( \log \frac{4}{\delta} \right)^{\frac{1}{2}} \right\}. $$

Together with (4.25), then with confidence at least $1 - l_\epsilon \delta$, for each $t = 1, \cdots, T$

$$\|f_{t+1}\|_K \leq (c + 1)^{\frac{\alpha + \zeta}{2}} \lambda^{\frac{\alpha - 1}{2}} \left\{ \lambda^{-\frac{\alpha}{2}} + l_\epsilon \left( \log \frac{4}{\delta} \right)^{\frac{1}{2}} \right\} \leq 2(c + 1)^{\frac{\alpha + \zeta}{2}} l_\epsilon \lambda^{\frac{\alpha - 1}{2} - \frac{\alpha}{2}} \left( \log \frac{4}{\delta} \right)^{\frac{1}{2}}. $$

By scaling $l_\epsilon \delta$ to $\delta$, then we can get the desired conclusion with $c_\delta = 2(c + 1)^{\frac{\alpha + \zeta}{2}} l_\epsilon$. \hfill \qed

### 4.5. Bounding the total error.

**Proof of Theorem 2.** With the choice $f^* = f_\lambda$, by (4.3), we have that

$$\mathcal{E}(\hat{f}_{T+1}) - \mathcal{E}(f_\rho) \leq \mathcal{E}_\lambda(\hat{f}_{T+1}) - \mathcal{E}_\lambda(f_\lambda) + \mathcal{M}_\lambda(R) + \mathcal{F}_\lambda(f_\lambda) + \mathcal{A}(f_\lambda). \quad (4.26)$$

Applying Lemma 6, we get that with confidence at least $1 - \delta$,

$$\mathcal{E}(\hat{f}_{T+1}) - \mathcal{E}(f_\rho) \leq \frac{1}{2}[\mathcal{E}(\hat{f}_{T+1}) - \mathcal{E}(f_\rho)] + C_1 \max \left\{ \left( \frac{R^c}{m} \right)^{\frac{2}{\alpha+1}}, \frac{1}{m} \log \frac{1}{\delta} \right\} \leq \mathcal{E}_\lambda(\hat{f}_{T+1}) - \mathcal{E}_\lambda(f_\lambda) + \mathcal{F}_\lambda(f_\lambda) + \mathcal{A}(f_\lambda).$$

Denote $R = c_\delta \lambda^{\frac{\alpha - 1}{2}} \left( \log \frac{4}{\delta} \right)^{\frac{1}{2}}$ and note that $\mathcal{A}(f_\lambda) \leq \mathcal{D}(\lambda) \leq c_\delta \lambda^2$. By (4.17) and (2.11), we get that with confidence at least $1 - 3\delta$, there holds

$$\mathcal{E}(\hat{f}_{T+1}) - \mathcal{E}(f_\rho) \leq (C_1 + C_2 + c_\delta \lambda^2) \max \left\{ \left( \frac{R^c}{m} \right)^{\frac{2}{\alpha+1}}, \frac{1}{m} \log \frac{2}{\delta}, \lambda^2 \log \frac{2}{\delta} \right\} \leq \frac{1}{2}[\mathcal{E}(\hat{f}_{T+1}) - \mathcal{E}(f_\rho)] + \mathcal{E}_\lambda(\hat{f}_{T+1}) - \mathcal{E}_\lambda(f_\lambda).$$

Scaling $3\delta$ to $\delta$ and subtracting $\frac{1}{2}[\mathcal{E}(\hat{f}_{T+1}) - \mathcal{E}(f_\rho)]$ from both sides of the above probability inequality, with $\lambda = m^{-\gamma}$, we get that with confidence at least $1 - \delta$,

$$\mathcal{E}(\hat{f}_{T+1}) - \mathcal{E}(f_\rho) \leq 2c_\delta \left( C_1 + C_2 + c_\delta \lambda^2 \right) \lambda^{\beta - \frac{\alpha \epsilon}{2}} \log \frac{12\lambda}{\delta} + 2 \left[ \mathcal{E}_\lambda(\hat{f}_{T+1}) - \mathcal{E}_\lambda(f_\lambda) \right].$$

Taking the regularization parameter $\lambda = T^{-(1-\theta)}$ and plugging the estimate (4.6) into the above inequality, with $\|f_\lambda\|_K \leq c_\delta \lambda^{\frac{\alpha - 1}{2}}$, we get

$$\mathcal{E}(\hat{f}_{T+1}) - \mathcal{E}(f_\rho) \leq 2c_\delta \left( C_1 + C_2 + c_\delta \lambda^2 \right) T^{-(1-\theta)\left( \beta - \frac{\alpha \epsilon}{2} \right)} \log \frac{12\lambda}{\delta} + 2C \left( C_\beta T^{1-\beta} \left( 1 + \frac{c_\delta}{c_\beta} \right)^{T^{3-2\theta}} \right).$$
Notice the choice $T = m^\alpha$ with (2.12), then we can get the bound (2.13) with
$$C_e = 2c_\epsilon \frac{2c}{\alpha} \left( C_1 + C_2'' + c_\beta \right) + 2c_\beta C + 2(1 + c_\beta^2)C.$$ The proof is completed. \qed

**Proof of Example 1.** It has been proved in [21] that $\mathcal{D}(\lambda) = O(\lambda^\beta)$ if $f_\rho \in L_\beta^2(L^2_{px})$ with $0 < \beta \leq 1$. Then we can get (2.14) by Theorem 2. The proof is completed. \qed

5. Appendix.

**Lemma 8** ([16]). Let $\{\xi(z_i)\}_{i=1}^m$ be a set of random variables defined on $\mathcal{Z}$ and $\tilde{M} > 0$ be the constant such that $\|\xi\|_\infty \leq \tilde{M}$ and the variance $\sigma^2(\xi) < \infty$, then for any $\epsilon > 0$,
$$\text{Prob} \left\{ \frac{1}{m} \sum_{i=1}^m \xi(z_i) - E(\xi) > \epsilon \right\} \leq \exp \left\{ -\frac{m\epsilon^2}{2(\sigma^2(\xi) + \frac{1}{3}\tilde{M}\epsilon)} \right\}.$$ 

**Proof of Lemma 5.** Set the right hand of Bernstein inequality $\exp \left\{ -\frac{m\epsilon^2}{2(\sigma^2(\xi) + \frac{1}{3}\tilde{M}\epsilon)} \right\} := \delta$. Solving it, we get that with confidence at least $1 - \delta$, there holds
$$\frac{1}{m} \sum_{i=1}^m \xi(z_i) - E(\xi) \leq \frac{2\tilde{M} \log \frac{1}{\delta}}{3m} + \sqrt{\frac{2 \log \frac{1}{\delta}}{m} \sigma^2(\xi)} \leq 2 \log \frac{1}{\delta} \max \left\{ \frac{\tilde{M}}{m} \sqrt{\frac{v(\xi)}{m}} \right\}^{-\frac{1}{2}}.$$ 

Applying the elementary inequality
$$a^{1-\tau}b^{1-\tau} \leq \tau b + (1-\tau)b, \quad a, b \in \mathbb{R},$$
we have
$$\left( \frac{c}{m} \right)^{1-\tau} (\xi)^{1-\tau} \leq \left( 1 - \frac{\tau}{2} \right) \left( \frac{c}{m} \right)^{1-\tau} + \frac{\tau}{2} \xi.$$ Plugging it into the inequality above, then the proof of Lemma 5 is completed. \qed

**Acknowledgments.** The author is grateful to the anonymous referees for many helpful comments and valuable suggestions on this paper.

**REFERENCES**

[1] N. Aronszajn, Theory of reproducing kernels, *Trans. Amer. Math. Soc.*, 68 (1950), 337–404.
[2] R. J. Bessa, V. Miranda and J. Gama, Entropy and correntropy against minimum square error in offline and online three-day ahead wind power forecasting, *IEEE Trans. Power Syst.*, 24 (2009), 1657–1666.
[3] D. R. Chen, Q. Wu, Y. Ying and D. X. Zhou, Support vector machine soft margin classifiers: Error analysis, *J. Mach. Learn. Res.*, 5 (2004), 1143–1175.
[4] M. Debruyne, A. Christmann, M. Hubert and J. A. K. Suykens, Robustness of reweighted least squares kernel based regression, *J. Multi. Anal.*, 101 (2010), 447–463.
[5] Y. Feng, J. Fan and J. A. Suykens, A statistical learning approach to modal regression, *J. Mach. Learn. Res.*, 21 (2020), 1–35.
[6] Y. Feng, X. Huang, S. Lei, Y. Yang and J. A. K. Suykens, Learning with the maximum correntropy criterion induced losses for regression, *J. Mach. Learn. Res.*, 16 (2015), 993–1034.
[7] Y. Feng and Y. Ying, Learning with correntropy-induced losses for regression with mixture of symmetric stable noise, *Appl. Comput. Harmon. Anal.*, 48 (2020), 795–810.
[8] Z. C. Guo, T. Hu and L. Shi, Gradient descent for robust kernel-based regression, *Inverse Probl.*, 34 (2018), Art. 065009.

[9] R. He, W. S. Zheng and B. G. Hu, Maximum correntropy criterion for robust face recognition, *IEEE Trans. Pattern Anal. Mach. Intell.*, 33 (2011), 1561–1576.

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

[11] P. W. Holland and R. E. Welsch, Robust regression using iteratively reweighted least-squares, *Commun. Statist.*, 6 (1977), 813–827.

[12] T. Hu, Q. Wu and D. X. Zhou, Distributed kernel gradient descent algorithm for minimum error entropy principle, *Appl. Comput. Harmon. Anal.*, 49 (2020), 229–256.

[13] P. J. Huber, *Robust Statistics*. Wiley, New York, 2004.

[14] J. Lin, L. Rosasco and D. X. Zhou, Iterative regularization for learning with convex loss functions, *J. Mach. Learn. Res.*, 17 (2016), 2718–2755.

[15] W. Liu, P. P. Pokharel and J. C. Principe, Correntropy: Properties and applications in non-gaussian signal processing, *IEEE Trans. Signal Process.*, 55 (2007), 5286–5298.

[16] I. Pinelis et al. Optimum bounds for the distributions of martingales in banach spaces, *Ann. Probab.*, 22 (1994), 1679–1706.

[17] K. N. Plataniotis, D. Androutsos and A. N. Venetsanopoulos, Nonlinear filtering of non-gaussian noise, *J. Intell. Robot. Syst.*, 19 (1997), 207–231.

[18] J. C. Principe, *Information Theoretic Learning: Renyi’s Entropy Entropy and Kernel Perspectives*, Springer, New York, 2010.

[19] I. Santamaria, P. P. Pokharel and J. C. Principe, Generalized correlation function: Definition, properties, and application to blind equalization, *IEEE Trans. Signal Process.*, 54 (2006), 2187–2197.

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

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

[22] I. Steinwart, Oracle inequalities for support vector machines that are based on random entropy numbers, *J. Complexity*, 25 (2009), 437–454.

[23] I. Steinwart and A. Christmann, *Support Vector Machines*, Springer Science & Business Media, 2008.

[24] X. Wang, Y. Jiang, M. Huang and H. Zhang, Robust variable selection with exponential squared loss, *J. Amer. Statist. Assoc.*, 108 (2013), 632–643.

[25] B. Weng and K. E. Barner, Nonlinear system identification in impulsive environments, *IEEE Trans. Signal Process.*, 53 (2005), 2588–2594.

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

Received September 2019; revised December 2019.

E-mail address: tinghu@whu.edu.cn