mathematics

MDPI

Article

Maximum Correntropy Criterion with Distributed Method

Fan Xie [1], Ting Hu [2], Shixu Wang [1] and Baobin Wang [1,*]

[1] School of Mathematics and Statistics, South-Central University for Nationalities, Wuhan 430074, China; xf13971704143@126.com (F.X.); wsxc1100@163.com (S.W.)
[2] School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China; tinghu@whu.edu.cn
* Correspondence: wbb@scuec.edu.cn

Abstract: The Maximum Correntropy Criterion (MCC) has recently triggered enormous research activities in engineering and machine learning communities since it is robust when faced with heavy-tailed noise or outliers in practice. This work is interested in distributed MCC algorithms, based on a divide-and-conquer strategy, which can deal with big data efficiently. By establishing minmax optimal error bounds, our results show that the averaging output function of this distributed algorithm can achieve comparable convergence rates to the algorithm processing the total data in one single machine.

Keywords: correntropy; maximum correntropy criterion; distributed method; robustness; error analysis

Citation: Xie, F.; Hu, T.; Wang, S.; Wang, B. Maximum Correntropy Criterion with Distributed Method. *Mathematics* **2022**, *10*, 304. https://doi.org/10.3390/math10030304

Academic Editors: Andrea Prati, Carlos A. Iglesias, Luis Javier García Villalba and Vincent A. Cicirello

Received: 23 November 2021
Accepted: 14 January 2022
Published: 19 January 2022

1. Introduction

In the big data era, the rapid expansion of data generation brings data of prohibitive size and complexity. This brings challenges to many traditional learning algorithms requiring access to the whole data set. Distributed learning algorithms, based on the divide-and-conquer strategy, provide a simple and efficient way to address this issue and therefore have received increasing attention. Such a strategy starts with partitioning the big data set into multiple subsets that are distributed to local machines, then it obtains local estimators in each subset by using a base algorithm, and it finally pools the local estimators together by simple averaging. It can substantially cut the time and memory costs in the algorithm implementation, and in many practical applications its learning performance has shown to be as good as that of a big machine that can use all the data. This scheme has been developed in various learning contexts, including spectral algorithms [1,2], kernel ridge regression [3–5], gradient descent [6,7], a semi-supervised approach [8], minimum error entropy [9] and bias correction [10].

Regression estimation and inference play an important role in the fields of data mining and statistics. The traditional ordinary least squares (OLS) method provides an efficient estimator if the regression model error is normally distributed. However, heavy-tailed noise and outliers are common in the real world, which limits the application of OLS in practice. Various robust losses have been proposed to deal with the problem instead of least squares loss. The commonly used robust losses mainly include adaptive Huber loss [11], gain function [12], minimum error entropy [13], exponential squared loss [14], etc. Among them, the Maximum Correntropy Criterion (MCC) is widely employed as an efficient alternative to the ordinary least squares method which is suboptimal in the non-Gaussian and non-linear signal processing situations [15–19]. Recently, MCC has been studied extensively in the literature and is widely adopted for many learning tasks, e.g., wind power forecasting [20] and pattern recognition [19]. In this paper, we are interested in the implementation of MCC by a distributed gradient descent method in a big data setting. Note that the MCC loss function is non-convex, so its analysis is essentially different from

the least squares method. A rigorous analysis of distributed MCC is necessary to derive the consistency and learning rates.

Given a hypothesis function $f : \mathcal{X} \to \mathcal{Y}$ and the scaling parameter $\sigma > 0$, correntropy between $f(X)$ and Y is defined by

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

where $G(u)$ is the Gaussian function $\exp\{-u\}, u \in \mathbb{R}$. Given the sample set $D = \{(x_i, y_i)\}_{i=1}^N \subset \mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, the empirical form of V_σ is

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

The purpose of MCC is to maximize the empirical correntropy $\hat{V}_\sigma$ over a hypothesis space $\mathcal{H}$, that is

$$f_{\mathbf{z},\mathcal{H}} := \arg\max_{f\in\mathcal{H}} \hat{V}_\sigma(f). \tag{1}$$

In the statistical learning context, the loss induced by correntropy $\phi_\sigma : \mathbb{R} \to \mathbb{R}_+$ is defined 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. The loss function can be viewed as a variant of the Welsch function [21] and the estimator $f_{\mathbf{z},\mathcal{H}}$ of (1) is also the minimizer of the empirical minimization risk scheme over $\mathcal{H}$, that is

$$\min_{f\in\mathcal{H}} \frac{1}{N}\sum_{i=1}^N \phi_\sigma(f(x_i)-y_i). \tag{2}$$

This paper aims at rigorous analysis of distributed gradient descent MCC within the framework of reproducing kernel Hilbert spaces (RKHSs). Let $K : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ be a *Mercer kernel* [22], i.e., a continuous, symmetric and positive semi-definite function. A kernel K is said to be positive semi-definite, if the matrix $\left(K(u_i,u_j)\right)_{i,j=1}^m$ is positive semi-definite for any finite set $\{u_1,\cdots,u_m\} \subset \mathcal{X}$ and $m \in \mathbb{N}$. 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 $\langle K_x, K_u\rangle_K = K(x,u)$. It has the *reproducing property*

$$f(x) = \langle f, K_x\rangle_K \tag{3}$$

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 (3), we get that

$$\|f\|_\infty \le \kappa\|f\|_K, \text{ for any } f \in \mathcal{H}_K. \tag{4}$$

Definition 1. *Given the sample set $D = \{(x_i,y_i)\}_{i=1}^N \subset \mathcal{Z} := \mathcal{X}\times\mathcal{Y}$, the kernel gradient descent algorithm for solving (2) can be stated iteratively with $f_{1,D} = 0$ as*

$$f_{t+1,D} = f_{t,D} - \eta \times \frac{1}{N}\sum_{i=1}^N \phi'_\sigma((f_{t,D}(x_i)-y_i))K_{x_i},\ t \ge 2 \tag{5}$$

where η is the of step size and $\phi'_\sigma((f_{t,D}(x_i)-y_i)) = G\left(\frac{(f_{t,D}(x_i)-y_i)^2}{2\sigma^2}\right)(f_{t,D}(x_i)-y_i)$.

Divide-and-Conquer algorithm for the kernel gradient descent MCC (5) is easy to describe. Rather than performing on the whole N examples, the distributed algorithm executes the following three steps:

1. Partition the data set D evenly and uniformly into m disjoint subsets D_j, $1 \le j \le m$.
2. Perform algorithm (5) on each data set D_j, and get the local estimate f_{T+1,D_j} after T-th iteration.
3. Take an average $\bar{f}_{T+1,D} = \frac{1}{m}\sum_{j=1}^{m} f_{T+1,D_j}$ as a final output.

In the next section, we study the asymptotic behavior of the final estimator $\bar{f}_{T+1,D}$ and show that $\bar{f}_{T+1,D}$ can obtain the minimax optimal rates over all estimators using the total data set of N samples provided that the scaling parameter σ is chosen suitably.

2. Assumptions and Main Results

In the setting of non-parametric estimation, we denote X as the explanatory variable that takes values in a compact domain $\mathcal{X}$, $Y \in \mathcal{Y} \subset \mathbb{R}$ as a real-valued response variable. Let ρ be the underlying distribution on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$. Moreover, let $\rho_{\mathcal{X}}$ be the marginal distribution of ρ on $\mathcal{X}$ and $\rho(\cdot|x)$ be the conditional distribution on $\mathcal{Y}$ for given $x \in \mathcal{X}$.

This work focuses on the application of MCC in regression problems, which is linked to the additive noise 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, which is the conditional mean $\mathbb{E}(Y|X = x)$ for $X = x \in \mathcal{X}$. The goal of this paper is to estimate the mean square error between $\bar{f}_{T+1,D}$ and f_ρ in $L^2_{\rho_{\mathcal{X}}}$-metric, which is defined by $\|\cdot\|_{L^2_{\rho_{\mathcal{X}}}} := \left(\int_{\mathcal{X}} |\cdot|^2 d\rho_{\mathcal{X}}\right)^{\frac{1}{2}}$. For simplicity, we will use $\|\cdot\|$ to denote the norm $\|\cdot\|_{L^2_{\rho_{\mathcal{X}}}}$ when the meaning is clear from the context.

Below, we present two important assumptions, which play a vital role in carrying out the analysis. The first assumption is about the regularity of the target function f_ρ. Define the integral operator $L_K : L^2_{\rho_{\mathcal{X}}} \to L^2_{\rho_{\mathcal{X}}}$ associated with K by

$$L_K f := \int_{\mathcal{X}} \int_{\mathcal{X}} f(x) K_x d\rho_{\mathcal{X}}(x), \quad \forall f \in L^2_{\rho_{\mathcal{X}}}.$$

As K is a Mercer kernel on the compact domain $\mathcal{X}$, the operator L_K is hence compact and positive. So, L_K^r as the r-th power of L_K for $r > 0$ is well defined. Our error bounds are stated in terms of the regularity of the target function f_ρ, given by [3,23]

$$f_\rho = L_K^r(h_\rho), \quad \text{for some } r > 0 \text{ and } h_\rho \in L^2_{\rho_{\mathcal{X}}}. \tag{6}$$

The condition (6) measures the regularity of f_ρ and is closely related to the smoothness of f_ρ when $\mathcal{H}_K$ is a Sobolev space. If (6) holds with $r \ge \frac{1}{2}$, f_ρ lies in the space $\mathcal{H}_K$.

The second assumption (7) is about the capacity of $\mathcal{H}_K$, measured by the effective dimension [24,25]

$$\mathcal{N}(\lambda) = \mathsf{Trace}((L_K + \lambda I)^{-1} L_K), \quad \text{for } \lambda > 0,$$

where I is the identity operator on $\mathcal{H}_K$. In this paper, we assume that

$$\mathcal{N}(\lambda) \le C\lambda^{-s} \quad \text{for some } C > 0 \text{ and } 0 < s \le 1. \tag{7}$$

Note that it always holds with $s = 1$. For $0 < s < 1$, it is almost equivalent to that the eigenvalues σ_i of L_K decay at a rate $i^{-\frac{1}{s}}$. The smoother the kernel function K is, the smaller s and the smaller function space $\mathcal{H}_K$. In particular, if K is a Gaussian kernel, then s can be arbitrarily close to 0, as $K \in C^\infty$.

Throughout the paper, we assume that $\kappa := \sup_{x\in\mathcal{X}}\sqrt{K(x,x)} \leq 1$ and $|y| \leq M$ for some $M > 0$. We denote $\lfloor a \rfloor$ as the smallest integer not less than a.

Theorem 1. *Assume that (6) and (7) hold for some $r > \frac{1}{2}$ and $0 < s \leq 1$. Taking $\eta_t = \eta t^{-\theta}$ with $0 < \eta \leq 1$ and $0 \leq \theta < 1$. If $T = \lfloor N^{\frac{1}{(2r+s)(1-\theta)}} \rfloor$ and the number of partition of the data set D*

$$m \leq \frac{N^{\frac{r-\frac{1}{2}}{2r+s}}}{(\log N)^5}, \tag{8}$$

then with confidence at least $1-\delta$,

$$\|\bar{f}_{T+1,D} - f_\rho\| \leq \tilde{C}\left\{N^{-\frac{r}{2r+s}} + N^{\frac{\frac{5}{2}}{2r+s}}\sigma^{-2}\right\}\left(\log\frac{12}{\delta}\right)^4,$$

where $\tilde{C}$ is a constant depending on θ.

Remark 1. *The above theorem, to be proved in Section 3, exhibits the concrete learning rates of the distributed estimator $\bar{f}_{T+1,D}$ (hence the standard estimator of (5) with $m = 1$). It implies that the kernel gradient descent for MCC on the single and distributed data set both achieves the learning rate $O\left(N^{-\frac{r}{2r+s}}\right)$ when σ is large enough. It equals the minimax optimal rates in the regression setting [24,26] in the case of $r > \frac{1}{2}$. This theorem suggests that the distributed MCC does not sacrifice the convergence rate provided that the partition number m satisfies the constraint (8). Thus, the distributed MCC estimator $\bar{f}_{T+1,D}$ enjoys both computational efficiency and statistical optimality.*

With the help of Theorem 1, we can easily deduce the following optimal learning rate in expectation.

Corollary 1. *Assume that (6) and (7) hold for some $r > \frac{1}{2}$ and $0 < s \leq 1$, taking $\eta_t = \eta t^{-\theta}$ with $0 < \eta \leq 1$ and $0 \leq \theta < 1$. If $T = \lfloor N^{\frac{1}{(2r+s)(1-\theta)}} \rfloor$, m satisfies (8) and $\sigma \geq N^{\frac{r/2+5/4}{2r+s}}$, then we have*

$$\mathbb{E}\left[\|\bar{f}_{T+1,D} - f_\rho\|\right] = O\left(N^{-\frac{r}{2r+s}}\right).$$

By the confidence-based error estimate in Theorem 1, we can obtain the following almost sure convergence of the distributed gradient descent algorithm for MCC.

Corollary 2. *Assume that (6) and (7) hold for some $r > \frac{1}{2}$ and $0 < s \leq 1$, taking $\eta_t = \eta t^{-\theta}$ with $0 < \eta \leq 1$ and $0 \leq \theta < 1$. If $T = \lfloor N^{\frac{1}{(2r+s)(1-\theta)}} \rfloor$, m satisfies (8) and $\sigma \geq N^{\frac{r/2+5/4}{2r+s}}$, and for arbitrary $\epsilon > 0$, we have*

$$\lim_{N\to\infty} N^{\frac{r}{2r+s}-\epsilon}\left[\|\bar{f}_{T+1,D} - f_\rho\|\right] = 0.$$

3. Discussion and Conclusions

In this work, we have studied the theoretical properties and convergence behaviors of a distributed kernel gradient descent MCC algorithm. As shown in Theorem 1, we derived minimax optimal error bounds for the distributed learning algorithm under the regularity condition on the regression function and capacity condition on RKHS. In the standard kernel gradient descent MCC algorithm ($m = 1$), the aggregate time complexity is $O(tN^2)$ after t iterations. However, in the distributed case ($m > 1$), the aggregate time complexity reduces to $O(tN^2/m)$ after t iterations. In conclusion, the kernel gradient descent MCC algorithm (5) with the distributed method can achieve fast convergence rates while successfully reducing algorithmic costs.

When the optimization problem arises from non-convex losses, the iteration sequence generated by the gradient descent algorithm is likely to only converge to a stationary point or a local minimizer. Note that the loss induced by correntropy ϕ_σ is not convex. Then, the convergence of the gradient descent method (5) to the global minimizer is not unconditionally guaranteed, which brings difficulties to the mathematical analysis of convergence. Our work on Theorem 1 addresses this issue, which shows that the iterative algorithm ensures the global optimality of its iterations in the theoretical analysis.

For regression problems, the distributed method has been introduced to the iteration algorithm in various learning paradigms and the minimax optimal rate has been obtained under different constraints on the partition number m. For distributed spectral algorithms [1], the lower bound of m that ensures the optimal rates is

$$m \leq N^{\min\{\frac{2}{2r+s}, \frac{2r-1}{2r+s}\}}. \tag{9}$$

We see from (9) that the restriction on m suffers from a saturation phenomenon in the sense that when $r \geq 3/2$ in the sense that the maximal m to guarantee the optimal learning rate does not improve as r is beyond $3/2$. Our restriction in (8) is worse than (9) when $r < 5/2$ but better when $r > 5/2$ as the upper bound in (8) increases with respect to r that overcomes the saturation effect in (9). For distributed kernel gradient descent algorithms with least squares method [6] and minimum error entropy (MEE) principle [9], the restrictions of m are improved as

$$m \leq \frac{N^{\frac{r-\frac{1}{2}}{2r+s}}}{(\log N)^4 + 1} \tag{10}$$

and

$$m \leq \frac{N^{\frac{r-\frac{1}{2}}{2r+s}}}{(\log N)^5}, \tag{11}$$

respectively. Our bound (8) for MCC differs with (10) for least squares only up to a logarithmic term, which has little impact on the upper bound of m ensuring optimal rates, but numerical experiments show that the distributed kernel gradient descent algorithm for least squares method is inferior to that for MCC in non-Gaussian noise models [15,27,28]. Our bound (8) is the same as (11) that is applied to the MEE principle. As we know, MEE also performs well in dealing with non-Gaussian noise or heavy-tail distribution [13,29]. However, MEE belongs to pairwise learning problems that work with pairs of samples rather than single sample in MCC. Hence, the distributed kernel gradient descent algorithm for MCC has an advantage over MEE in algorithmic complexity.

Several related questions are worthwhile for future research. First, our distributed result provides the optimal rates by requiring a large robust parameter σ. In practice, a moderate σ may be enough to ensure a good learning performance in robust estimation as shown by [17]. It is therefore of interest to investigate the convergence properties of distributed version of algorithm (5) when σ is chosen as a constant or $\sigma(N) \to 0$ as N approaches ∞.

Secondly, our algorithm is carried out in the framework of supervised learning; however, in numerous real-world applications, few labeled data are available, but a large amount of unlabeled data are given since the cost of labeling data is high such as time, money. Thus, we shall investigate how to enhance the learning performance of the MCC algorithm by the distributed method and the additional information given by unlabeled data.

Thirdly, as stated in Theorem 1, the choice of the last iteration T and the partition number m depends on the parameters r, s, which are usually unknown in advance. In practice, cross-validation is usually used to tune T and m adaptively. It would be interesting to know whether the kernel gradient descent MCC (5) with the distributed method can achieve the optimal convergence rate with adaptive T and m.

Last but not least, we should note that here that all the data $D = \{(x_i, y_i)\}_{i=1}^N$ are drawn independently according to the same distribution. In the distributed method, we partition D evenly and uniformly into m disjoint subsets. This means that $|D_1| = \cdots = |D_m| = \frac{N}{m}$ and each sample (x_i, y_i) is assigned to the subset D_j $(1 \le j \le m)$ with the same probability. In the context of uniform random sampling, such randomness splitting strategy should be reasonable and practical. So, our theoretical analysis is based on the uniform random splitting mechanism. However, for the theoretical analysis of other randomness or non-randomness splitting mechanisms, it is necessary to develop new mathematical tools for optimal performance. It is beyond the scope of this paper and will be left for our future work.

4. Proofs of Main Results

This section is devoted to proving main results in Section 2. Here and in the following, let the sample size of each subset $D_1, \cdots, D_m$ be n; that is, $D = D_1 \bigcup \cdots D_m$ and $N = mn$. Define the empirical operator $L_{K,D}$ on $\mathcal{H}_K$ as

$$L_{K,D}(f) = \frac{1}{N}\sum_{i=1}^{N}\langle f, K_{x_i}\rangle_K K_{x_i}, \qquad \forall\, f \in \mathcal{H}_K,$$

where $x_1 \cdots, x_N \in \{x : (x, y) \in D \text{ with some } y \in \mathcal{Y}\}$. Similarly, we can define the operator L_{K,D_j} on $\mathcal{H}_K$ for each subset $D_j, 1 \le j \le m$,

$$L_{K,D_j}(f) = \frac{1}{n}\sum_{i=1}^{n}\langle f, K_{x_i}\rangle_K K_{x_i}, \qquad \forall\, f \in \mathcal{H}_K,$$

where $x_1 \cdots, x_n \in \{x : (x, y) \in D_j \text{ with some } y \in \mathcal{Y}\}$.

4.1. Preliminaries

We first introduce some necessary lemmas in the proofs, which can be found in [3,6,9].

Lemma 1. *Let $g(z)$ be a measurable function defined on $\mathcal{Z}$ with $\|g\|_\infty \le M'$ almost definitely for some $M' > 0$. Let $0 < \delta < 1$; then, each of the following estimates holds with confidence at least $1 - \delta$,*

$$\left\|(L_K + \lambda I)^{-\frac{1}{2}}(L_K - L_{K,D})\right\| \le 2\mathcal{A}_{D,\lambda}\log\frac{2}{\delta},$$

$$\left\|(L_{K,D} + \lambda I)^{-1}(L_K + \lambda I)\right\| \le 2\Big(\frac{2\,A_{D,\lambda}\log\frac{2}{\delta}}{\sqrt{\lambda}}\Big)^2 + 2.$$

and

$$\left\|\frac{1}{N}\sum_{i=1}^{N}(L_K + \lambda I)^{-\frac{1}{2}}\Big[g(z_i)K_{x_i} - L_K g\Big]\right\| \le 2M'\mathcal{A}_{D,\lambda}\log\frac{2}{\delta}$$

where $\mathcal{A}_{D,\lambda} := \frac{1}{N\sqrt{\lambda}} + \sqrt{\frac{\mathcal{N}(\lambda)}{N}}$.

Let π_i^t denote the polynomial defined by $\pi_i^t(s) = \prod_{j=i}^t(1 - \eta_j x)$ if $i \le t$ and, for notation simplicity, let $\pi_{t+1}^t(s) = 1$ be the identity function. In our proof, we need to deal with the polynomial operators $\pi_i^t(L_K)$ and $\pi_i^t(L_{K,D})$. For this purpose we introduce the conventional notation $\sum_{j=T+1}^T := 1$ and the following preliminary lemmas.

Lemma 2. *If* $0 \le \alpha < 1, 0 \le \theta < 1$,*then for* $T \ge 3$,

$$\sum_{i=1}^{T} i^{-(\theta+\alpha)} \left(\sum_{j=i+1}^{T} j^{-\theta} \right)^{-1} \le C_{\theta,\alpha} T^{-\min\{\alpha, 1-\theta\}} \log T, \tag{12}$$

where $C_{\theta,\alpha}$ *is a constant depending only on* θ *and* α*, whose value is given in the proof. In particular, if* $\alpha = 0$*, we have*

$$\sum_{i=1}^{T} i^{-\theta} \left(\sum_{j=i+1}^{T} j^{-\theta} \right)^{-1} \le 15 \log T. \tag{13}$$

Lemma 3. *If* $\eta_t = \eta t^{-\theta}$ *with* $0 < \eta < 1$ *and* $0 \le \theta < 1$*, then for* $1 \le i \le T-1$,

$$\left\| \pi_i^t (L_{K,D}) \right\| \le 1 \tag{14}$$

$$\left\| \pi_i^t (L_K) \right\| \le 1 \tag{15}$$

$$\left\| L_{K,D} \pi_{i+1}^T (L_{K,D}) \right\| \le \left(e\eta \sum_{j=i+1}^{T} j^{-\theta} \right)^{-1}, \tag{16}$$

$$\left\| L_K \pi_{i+1}^T (L_K) \right\| \le \left(e\eta \sum_{j=i+1}^{T} j^{-\theta} \right)^{-1}, \tag{17}$$

$$\left\| \sum_{i=1}^{T} \eta_i \left[(L_{K,D} + \lambda I) \pi_{i+1}^T (L_{K,D}) \right] \right\| \le 1 + \frac{\eta\lambda}{1-\theta} T^{1-\theta}, \tag{18}$$

$$\left\| \sum_{i=1}^{T} \eta_i \left[(L_K + \lambda I) \pi_{i+1}^T (L_K) \right] \right\| \le 1 + \frac{\eta\lambda}{1-\theta} T^{1-\theta}. \tag{19}$$

Define a data-free gradient descent sequence for the least square method in $\mathcal{H}_K$ by $f_1 = 0$ and

$$f_{t+1} = f_t - \eta_t \int_{\mathcal{X}} \big(f_t(x) - f_\rho(x) \big) K_x d\rho_{\mathcal{X}} = (I - \eta_t L_K) f_t + \eta_t L_K f_\rho. \tag{20}$$

It has been well evidence in the literature [30] that under the assumption (6) with $r > \frac{1}{2}$, there are

$$\| f_t - f_\rho \| \le h_\rho t^{-r(1-\theta)} \tag{21}$$

and

$$\| f_t - f_\rho \|_K \le h_\rho t^{-(r-\frac{1}{2})(1-\theta)}, \tag{22}$$

where $h_\rho = \max \left\{ \|g\| (2r/e)^r, \|g\| [(2r-1)/e]^{r-\frac{1}{2}} \right\}$.

Lemma 4. *If* $\eta_t = \eta t^{-\theta}$ *with* $0 < \eta < 1$ *and* $0 \le \theta < 1$*, then there is a constant* $C_{\rho,\theta,r}$ *such that*

$$\sum_{i=1}^{T} \eta_i \| L_{K,D} \pi_{i+1}^T (L_{K,D}) \| \| f_i - f_\rho \|_K \le C_{\rho,\theta,r} \tag{23}$$

and

$$\sum_{i=1}^{T} \eta_i \| L_K \pi_{i+1}^T (L_K) \| \| f_i - f_\rho \|_K \le C_{\rho,\theta,r}. \tag{24}$$

Lemma 5. *If $\eta_t = \eta t^{-\theta}$ with $0 < \eta < 1$ and $0 \leq \theta < 1$, then there is a constant $D_{\rho,\theta,r}$ such that*

$$\sum_{i=1}^{T} \eta_i \|f_i - f_\rho\|_K \leq D_{\rho,\theta,r} T^{1-\theta}. \tag{25}$$

Recall that the isomorphism between $\mathcal{H}_K$ and $L^2_{\rho_\mathcal{X}}$, which yields in

$$\|f\| = \|L_K^{\frac{1}{2}} f\|_K \leq \|(L_K + \lambda I)^{\frac{1}{2}} f\|_K, \text{ for all } f \in \mathcal{H}_K. \tag{26}$$

4.2. Bound for the Learning Sequence

We will need the following bound for the learning sequence in the proof.

Theorem 2. *If the step size sequence $\eta_t = \eta t^{-\theta}$ with $0 < \eta \leq 1$ and $0 \leq \theta < 1$, then we have the following bound for the learning sequence $\{f_{t,D}\}$ by (5):*

$$\|f_{t,D}\|_K \leq M t^{\frac{1-\theta}{2}}. \tag{27}$$

Proof. We prove the statement by induction. First note the conclusion holds trivially for $t = 1$. Next, suppose that $\|f_{t,D}\|_K \leq M\sqrt{\sum_{i=1}^{t-1} \eta_i}$ holds. By the updating rule (5) and the reproducing property, we have

$$\begin{aligned}
\|f_{t+1,D}\|_K^2 =& \|f_{t,D}\|_K^2 - \frac{2\eta_t}{N} \sum_{i=1}^{N} \phi'_\sigma(f_{t,D}(x_i) - y_i) f_{t,D}(x_i) \\
&+ \frac{\eta_t^2}{N^2} \left\| \sum_{i=1}^{N} \phi'_\sigma(f_{t,D}(x_i) - y_i) K_{x_i} \right\|_K^2 \\
\leq& \|f_{t,D}\|_K^2 - \frac{2\eta_t}{N} \sum_{i=1}^{N} \phi'_\sigma(f_{t,D}(x_i) - y_i) f_{t,D}(x_i) \\
&+ \frac{\eta_t^2}{N} \sum_{i=1}^{N} \left| G\left(\frac{(f_{t,D}(x_i) - y_i)^2}{2\sigma^2} \right) \right|^2 (f_{t,D}(x_i) - y_i)^2 \\
=& \|f_{t,D}\|_K^2 + \frac{\eta_t}{N} \sum_{i=1}^{N} Q_i,
\end{aligned} \tag{28}$$

where

$$\begin{aligned}
Q_i =& \left[\eta_t \left| G\left(\frac{(f_{t,D}(x_i) - y_i)^2}{2\sigma^2} \right) \right|^2 - 2G\left(\frac{(f_{t,D}(x_i) - y_i)^2}{2\sigma^2} \right) \right] (f_{t,D}(x_i))^2 \\
&- 2\left(G\left(\frac{(f_{t,D}(x_i) - y_i)^2}{2\sigma^2} \right) + \eta_t \left| G\left(\frac{(f_{t,D}(x_i) - y_i)^2}{2\sigma^2} \right) \right|^2 \right) y_i f_{t,D}(x_i) \\
&+ \eta_t \left| G\left(\frac{(f_{t,D}(x_i) - y_i)^2}{2\sigma^2} \right) \right|^2 y_i^2.
\end{aligned}$$

The restriction $\eta_t \leq 1$ implies $\eta_t \left| G\left(\frac{(f_{t,D}(x_i)-y_i)^2}{2\sigma^2}\right)\right|^2 - 2G\left(\frac{(f_{t,D}(x_i)-y_i)^2}{2\sigma^2}\right) < 0$. By the property of quadratic function, we have

$$Q_i \leq \eta_t \left| G\left(\frac{(f_{t,D}(x_i)-y_i)^2}{2\sigma^2}\right)\right|^2 y_i^2 - \frac{\left(-G\left(\frac{(f_{t,D}(x_i)-y_i)^2}{2\sigma^2}\right) + \eta_t \left| G\left(\frac{(f_{t,D}(x_i)-y_i)^2}{2\sigma^2}\right)\right|^2\right) y_i^2}{\eta_t \left| G\left(\frac{(f_{t,D}(x_i)-y_i)^2}{2\sigma^2}\right)\right|^2 - 2G\left(\frac{(f_{t,D}(x_i)-y_i)^2}{2\sigma^2}\right)}$$
$$= \frac{G\left(\frac{(f_{t,D}(x_i)-y_i)^2}{2\sigma^2}\right) y_i^2}{2 - \eta_t G\left(\frac{(f_{t,D}(x_i)-y_i)^2}{2\sigma^2}\right)} \leq M^2.$$

Plugging it into (28), we obtain

$$\|f_{t+1,D}\|_K^2 \leq \|f_{t,D}\|_K^2 + M^2\eta_t \leq M^2 \sum_{i=1}^{t} \eta_i = M^2\eta \sum_{i=1}^{t} i^{-\theta} \leq M^2 t^{1-\theta}.$$

This completes the proof. □

4.3. Error Decomposition and Estimation of Error Bounds

Now we are in a position of bounding the error of the distributed kernel gradient descent MCC. For this purpose, we decompose the error $\|\bar{f}_{T+1,D} - f_\rho\|$ into two parts as

$$\|\bar{f}_{T+1,D} - f_\rho\| \leq \|f_{T+1} - f_\rho\| + \|\bar{f}_{T+1,D} - f_{T+1}\|. \tag{29}$$

As we have mentioned in the previous subsection, the first term can be bounded by (21) under the assumption (6) with $r > \frac{1}{2}$. Our key analysis is the second term, which can be bounded with the help of the following proposition.

Proposition 1. *Assume that* (6) *holds for some* $r > \frac{1}{2}$. *Let* $\eta_t = \eta t^{-\theta}$ *with* $0 < \eta \leq 1$ *and* $0 \leq \theta < 1$. *For* $\lambda > 0$, *there holds*

$$\|f_{T+1,D} - f_{T+1}\| \leq C'_{r,\theta}\left[\mathcal{B}_{D,\lambda}(\mathcal{C}_{D,\lambda} + \mathcal{G}_{D,\lambda})(1 + \lambda T^{1-\theta}) + T^{\frac{5(1-\theta)}{2}}\sigma^{-2}\right], \tag{30}$$

and

$$\|f_{T+1,D} - f_{T+1}\|_K \leq C'_{r,\theta}\left[\mathcal{B}_{D,\lambda}(\mathcal{C}_{D,\lambda} + \mathcal{G}_{D,\lambda})(1 + \lambda T^{1-\theta})/\sqrt{\lambda} + T^{\frac{5(1-\theta)}{2}}\sigma^{-2}\right], \tag{31}$$

where

$$\begin{aligned} \mathcal{B}_{D,\lambda} &= \|(L_{K,D} + \lambda I)^{-1}(L_K + \lambda I)\|, \\ \mathcal{C}_{D,\lambda} &= \|(L_K + \lambda I)^{-\frac{1}{2}}(L_K - L_{K,D})\|, \\ \mathcal{G}_{D,\lambda} &= \|(L_K + \lambda I)^{-\frac{1}{2}}(L_K f_\rho - \hat{f}_{\rho,D})\|_K, \\ \widehat{f}_{\rho,D} &= \frac{1}{N}\sum_{i=1}^{N} y_i K_{x_i} = \frac{1}{N}\sum_{(x,y)\in D} y K_x, \end{aligned} \tag{32}$$

and $C'_{r,\theta}$ *is given in the proof, depending on* r, θ.

Proof. By the definition of $f_{t,D}$ in (5) and the definition of f_t in (20), we have

$$f_{t+1,D} - f_{t+1} = [I - \eta_t L_{K,D}](f_{t,D} - f_t) + \eta_t[L_K - L_{K,D}]f_t + \eta_t[\hat{f}_{\rho,D} - L_K(f_\rho)] + \eta_t E_{t,D}, \tag{33}$$

where $\widehat{f}_{\rho,D}$ is defined in (32) and

$$E_{t,D} = \frac{1}{N}\sum_{i=1}^{N}\left(1 - G\left(\frac{(f_{t,D}(x_i) - y_i)^2}{2\sigma^2}\right)\right)\left(f_{t,D}(x_i) - y_i\right)K_{x_i},$$

Applying (33) iteratively from $t = 1$ to T, we obtain

$$f_{T+1,D} - f_{T+1} = I_1 + I_2 + I_3 + I_4 \tag{34}$$

where

$$\begin{aligned} I_1 &= \sum_{i=1}^{T}\eta_i\pi_{i+1}^T(L_{K,D})[L_K - L_{K,D}](f_i - f_\rho), \\ I_2 &= \sum_{i=1}^{T}\eta_i\pi_{i+1}^T(L_{K,D})[L_K - L_{K,D}](f_\rho), \\ I_3 &= \sum_{i=1}^{T}\eta_i\pi_{i+1}^T(L_{K,D})[\hat{f}_{\rho,D} - L_K(f_\rho)], \\ I_4 &= \sum_{i=1}^{T}\eta_i\pi_{i+1}^T(L_{K,D})E_{i,D}. \end{aligned}$$

For I_1, by (26), Lemmas 4 and 5,

$$\begin{aligned} \|I_1\| &= \left\|\sum_{i=1}^{T}\eta_i(L_K + \lambda I)^{\frac{1}{2}}\pi_{i+1}^T(L_{K,D})[L_K - L_{K,D}](f_i - f_\rho)\right\|_K \\ &\le \sum_{i=1}^{T}\left\{\eta_i\left\|(L_K + \lambda I)^{\frac{1}{2}}(L_{K,D} + \lambda I)^{-\frac{1}{2}}\right\|\left\|(L_{K,D} + \lambda I)\pi_{i+1}^T(L_{K,D})\right\|\right. \\ &\qquad \left.\times\left\|(L_{K,D} + \lambda I)^{-\frac{1}{2}}(L_K + \lambda I)^{\frac{1}{2}}\right\|\left\|(L_K + \lambda I)^{-\frac{1}{2}}[L_K - L_{K,D}]\right\|\|f_i - f_\rho\|_K\right\} \\ &\le \mathcal{B}_{D,\lambda}\mathcal{C}_{D,\lambda}\left(\sum_{i=1}^{T}\eta_i\|L_{K,D}\pi_{i+1}^T(L_{K,D})\|\|f_i - f_\rho\|_K + \lambda\sum_{i=1}^{T}\eta_i\|f_i - f_\rho\|_K\right) \\ &\le \mathcal{B}_{D,\lambda}\mathcal{C}_{D,\lambda}\left(C_{\rho,\theta,r} + D_{\rho,\theta,r}\lambda T^{1-\theta}\right). \end{aligned} \tag{35}$$

For I_2, by (26), Lemma 3, and the fact $\|f_\rho\|_\infty \le M$, we have

$$\begin{aligned} \|I_2\| &= \left\|\sum_{i=1}^{T}\eta_i\pi_{i+1}^T(L_{K,D})[L_K - L_{K,D}](f_\rho)\right\| \\ &\le \left\|\sum_{i=1}^{T}\eta_i(L_K + \lambda I)^{\frac{1}{2}}\pi_{i+1}^T(L_{K,D})[L_K - L_{K,D}](f_\rho)\right\|_K \\ &\le \left\|(L_K + \lambda I)^{\frac{1}{2}}(L_{K,D} + \lambda I)^{-\frac{1}{2}}\right\|\left\|\sum_{i=1}^{T}\eta_i(L_{K,D} + \lambda I)\pi_{i+1}^T(L_{K,D})\right\| \\ &\qquad\times\left\|(L_{K,D} + \lambda I)^{-\frac{1}{2}}(L_K + \lambda I)^{\frac{1}{2}}\right\|\left\|(L_K + \lambda I)^{-\frac{1}{2}}[L_K - L_{K,D}]\right\|\|f_\rho\|_K \\ &\le M\left(1 + \frac{\lambda T^{1-\theta}}{1-\theta}\right)\mathcal{B}_{D,\lambda}\mathcal{C}_{D,\lambda}. \end{aligned} \tag{36}$$

Similarly, we can bound I_3 as

$$I_3 \le \left(1 + \frac{\lambda T^{1-\theta}}{1-\theta}\right)\mathcal{B}_{D,\lambda}\mathcal{G}_{D,\lambda}. \tag{37}$$

For I_4, first note that by the bound (27) of $\{f_{t,D}\}$, we see

$$\begin{aligned}&\left\|\left(G\left(\frac{(f_{t,D}(x_i)-y_i)^2}{2\sigma^2}\right)-1\right)(f_{t,D}(x_i)-y_i)K_{x_i}\right\|_K\\ &\le \frac{(M+\|f_{t,D}\|_K)^3}{2\sigma^2}\le\frac{2^2}{\sigma^2}\|f_{t,D}\|_K^3\\ &\le 2^2M^3t^{\frac{3(1-\theta)}{2}}\sigma^{-2}\end{aligned}$$

This implies that

$$\|E_{t,D}\|_K\le 2^2M^3t^{\frac{(1-\theta)(3)}{2}}\sigma^{-2}. \tag{38}$$

Thistogether with the estimate $\|\pi_{i+1}^t(L_{K,D})\|\le 1$ gives

$$\begin{aligned}\|I_4\|&\le\sum_{i=1}^T\eta_i\|E_{i,D}\|_K\le 2^2M^3\eta\sum_{i=1}^T i^{\frac{3(1-\theta)}{2}-\theta}\sigma^{-2}\\ &\le\frac{2^2M^3}{(1-\theta)(\frac{5}{2})}T^{\frac{5(1-\theta)}{2}}\sigma^{-2}.\end{aligned} \tag{39}$$

Combining the estimates in (36), (37), (39) and (35), we obtain (30) holds with

$$C'_{r,\theta}=C_{\rho,\theta,r}+D_{\rho,\theta,r}+\frac{2M}{1-\theta}+\frac{2^3M^3}{5(1-\theta)}.$$

Following a similar process we can obtain the bound in (31). □

The following theorem provides a bound for the second term in (29).

Theorem 3. *Take $\lambda=T^{-(1-\theta)}$. There is a constant $C''_{r,\theta}$ such that*

$$\begin{aligned}\|\bar{f}_{T+1,D}-f_{T+1}\|\le C''_{r,\theta}\Big[&\mathcal{G}_{D,\lambda}+\mathcal{C}_{D,\lambda}+\lambda^{-\frac{1}{2}}\log T\sup_{1\le l\le m}\mathcal{C}_{D_l,\lambda}\mathcal{B}_{D_l,\lambda}(\mathcal{C}_{D_l,\lambda}+\mathcal{G}_{D_l,\lambda})\\ &+\sigma^{-2}T^{\frac{5(1-\theta)}{2}}\Big(1+\log T\sup_{1\le l\le m}\mathcal{C}_{D_l,\lambda}\Big)\Big].\end{aligned} \tag{40}$$

Proof. For each subset D_l and each $1\le t\le T$, we have

$$f_{T+1,D_l}-f_{T+1}=[I-\eta_tL_K](f_{T,D_l}-f_t)+\eta_T[L_K-L_{K,D}]f_{T,D_l}+\eta_T[\hat{f}_{\rho,D_l}-L_K(f_\rho)]+\eta_TE_{T,D_l}.$$

This implies that

$$\begin{aligned}f_{T+1,D_l}-f_{T+1}=&\sum_{i=1}^T\eta_i\pi_{i+1}^T(L_K)[L_K-L_{K,D_l}]f_{i,D_l}\\ &+\sum_{i=1}^T\eta_i\pi_{i+1}^T(L_K)[\hat{f}_{\rho,D_l}-L_K(f_\rho)]+\sum_{i=1}^T\eta_i\pi_{i+1}^T(L_K)E_{i,D_l},\end{aligned}$$

and therefore

$$\begin{aligned}\|\bar{f}_{T+1,D} - f_{T+1}\| &= \left\| \frac{1}{m}\sum_{l=1}^{m}\left(f_{T+1,D_l} - f_{T+1}\right)\right\| \\ &\le \left\|\sum_{i=1}^{T}\eta_i\pi_{i+1}^{T}(L_K)\frac{1}{m}\sum_{l=1}^{m}[L_K - L_{K,D_l}]f_{i,D_l}\right\| \\ &\quad + \left\|\sum_{i=1}^{T}\eta_i\pi_{i+1}^{T}(L_K)\frac{1}{m}\sum_{l=1}^{m}[\hat{f}_{\rho,D_l} - L_K(f_\rho)]\right\| \\ &\quad + \left\|\frac{1}{m}\sum_{l=1}^{m}\sum_{i=1}^{T}\eta_i\pi_{i+1}^{T}(L_K)E_{i,D_l}\right\| \\ &:= J_1 + J_2 + J_3.\end{aligned}$$

We first estimate J_2. By (26), Lemma 3, and the choice $\lambda = T^{-(1-\theta)}$, we obtain

$$\begin{aligned}J_2 &\le \left\|\sum_{i=1}^{T}\eta_i(L_K+\lambda)\pi_{i+1}^{T}(L_K)\frac{1}{m}\sum_{l=1}^{m}(L_K+\lambda)^{-\frac{1}{2}}[\hat{f}_{\rho,D_l} - L_K(f_\rho)]\right\|_K \\ &\le \left(1 + \frac{\lambda T^{1-\theta}}{1-\theta}\right)\left\|(L_K+\lambda)^{-\frac{1}{2}}[\hat{f}_{\rho,D} - L_K(f_\rho)]\right\|_K \\ &\le \frac{2M}{1-\theta}\left(1+\lambda T^{1-\theta}\right)\mathcal{G}_{D,\lambda} \\ &:= \frac{4M}{1-\theta}\mathcal{G}_{D,\lambda}. \end{aligned} \tag{41}$$

For J_3, by (39) we have

$$J_3 \le \sup_{1\le l\le m}\left\|\sum_{i=1}^{T}\eta_i\pi_{i+1}^{T}(L_K)E_{i,D_l}\right\| \le \frac{2^3M^3\eta}{5(1-\theta)}T^{5(1-\theta)/2}\sigma^{-2}. \tag{42}$$

The estimation of J_1 is much more complicated. We decompose it into three parts,

$$\begin{aligned}J_1 &\le \left\|\sum_{i=1}^{T}\eta_i(L_K+\lambda)\pi_{i+1}^{T}(L_K)\frac{1}{m}\sum_{l=1}^{m}(L_K+\lambda)^{-\frac{1}{2}}[L_K - L_{K,D_l}]f_{i,D_l}\right\|_K \\ &\le \left\|\sum_{i=1}^{T}\eta_i(L_K+\lambda)\pi_{i+1}^{T}(L_K)\frac{1}{m}\sum_{l=1}^{m}(L_K+\lambda)^{-\frac{1}{2}}[L_K - L_{K,D_l}](f_{i,D_l} - f_i)\right\|_K \\ &\quad + \left\|\sum_{i=1}^{T}\eta_i(L_K+\lambda)\pi_{i+1}^{T}(L_K)(L_K+\lambda)^{-\frac{1}{2}}[L_K - L_{K,D}](f_i - f_\rho)\right\|_K \\ &\quad + \left\|\sum_{i=1}^{T}\eta_i(L_K+\lambda)\pi_{i+1}^{T}(L_K)(L_K+\lambda)^{-\frac{1}{2}}[L_K - L_{K,D}](f_\rho)\right\|_K \\ &:= J_{11} + J_{12} + J_{13}.\end{aligned}$$

By Lemmas 4 and 5 and the fact $\lambda T^{1-\theta} = 1$, we obtain

$$\begin{aligned}J_{12} &\le \mathcal{C}_{D,\lambda}\left(\sum_{i=1}^{T}\left\|\eta_i L_K\pi_{i+1}^{T}(L_K)\right\|\|f_i - f_\rho\|_K + \lambda\sum_{i=1}^{T}\eta_i\|f_i - f_\rho\|_K\right) \\ &\le \mathcal{C}_{D,\lambda}\left(C_{\rho,\theta,r} + D_{\rho,\theta,r}\right).\end{aligned}$$

For J_{13}, by (19) we have

$$\begin{aligned} J_{13} &\leq \left\| \sum_{i=1}^{T} \eta_i (L_K + \lambda) \pi_{i+1}^T (L_K) \right\| \left\| (L_K + \lambda)^{-\frac{1}{2}} [L_K - L_{K,D}] \right\| \|f_\rho\|_K \\ &\leq M \left(1 + \frac{\lambda T^{1-\theta}}{1-\theta} \right) \mathcal{C}_{D,\lambda} = \frac{2M}{1-\theta} \mathcal{C}_{D,\lambda}. \end{aligned}$$

Now we turn to J_{11}. We have

$$\begin{aligned} J_{11} &\leq \sum_{i=1}^{T} \left\| \eta_i (L_K + \lambda) \pi_{i+1}^T (L_K) \right\| \left\| \frac{1}{m} \sum_{l=1}^{m} (L_K + \lambda)^{-\frac{1}{2}} [L_K - L_{K,D_l}] (f_{i,D_l} - f_i) \right\|_K \\ &\leq \sum_{i=1}^{T} \left\| \eta_i (L_K + \lambda) \pi_{i+1}^T (L_K) \right\| \sup_{1 \leq l \leq m} \left\| (L_K + \lambda)^{-\frac{1}{2}} [L_K - L_{K,D_l}] (f_{i,D_l} - f_i) \right\|_K \\ &\leq \sum_{i=1}^{T} \eta_i \Big[\Big(\sum_{j=i+1}^{T} \eta_j \Big)^{-1} + \lambda \Big] \sup_{1 \leq l \leq m} \mathcal{C}_{D_l,\lambda} \|f_{i,D_l} - f_i\|_K. \end{aligned} \tag{43}$$

By Theorem 1 and the choice $\lambda = T^{-(1-\theta)}$, for $1 \leq i \leq T$, there holds that $\lambda i^{(1-\theta)} \leq 1$ and

$$\begin{aligned} \|f_{i,D_l} - f_i\|_K &\leq C'_{r,\theta} \Big[\mathcal{B}_{D_l,\lambda} (\mathcal{C}_{D_l,\lambda} + \mathcal{G}_{D_l,\lambda}) (1 + \lambda i^{1-\theta}) / \sqrt{\lambda} + i^{\frac{5(1-\theta}{2})} \sigma^{-2} \Big] \\ &\leq C'_{r,\theta} \Big[2\mathcal{B}_{D_l,\lambda} (\mathcal{C}_{D_l,\lambda} + \mathcal{G}_{D_l,\lambda}) / \sqrt{\lambda} + T^{(\frac{5(1-\theta)}{2})} \sigma^{-2} \Big]. \end{aligned}$$

Plugging it into (43), we obtain

$$J_{11} \leq C'_{r,\theta} \sup_{1 \leq l \leq m} \mathcal{C}_{D_l,\lambda} \Big[2\mathcal{B}_{D_l,\lambda} (\mathcal{C}_{D_l,\lambda} + \mathcal{G}_{D_l,\lambda}) / \sqrt{\lambda} + T^{(\frac{5(1-\theta)}{2})} \sigma^{-2} \Big] \sum_{i=1}^{T} \eta_i \Big[\Big(\sum_{j=i+1}^{T} \eta_j \Big)^{-1} + \lambda \Big].$$

From Lemma 2, we see that

$$\sum_{i=1}^{T} \eta_i \Big[\Big(\sum_{j=i+1}^{T} \eta_j \Big)^{-1} + \lambda \Big] \leq 15 \log T + \frac{\eta \lambda T^{1-\theta}}{1-\theta} = 15 \log T + \frac{1}{1-\theta} \leq \left(15 + \frac{1}{1-\theta} \right) \log T.$$

So, we have

$$J_{11} \leq C'_{r,\theta} \left(15 + \frac{1}{1-\theta} \right) \log T \sup_{1 \leq l \leq m} \mathcal{C}_{D_l,\lambda} \Big[2\mathcal{B}_{D_l,\lambda} (\mathcal{C}_{D_l,\lambda} + \mathcal{G}_{D_l,\lambda}) / \sqrt{\lambda} + T^{(\frac{5(1-\theta)}{2})} \sigma^{-2} \Big].$$

Combining the estimations for J_{11}, J_{12} and J_{13}, we obtain

$$\begin{aligned} J_1 \leq & \Big(\frac{2M}{1-\theta} + C_{\rho,\theta,r} + D_{\rho,\theta,r} \Big) \mathcal{C}_{D,\lambda} \\ & + 2C'_{r,\theta} \left(15 + \frac{1}{1-\theta} \right) \lambda^{-\frac{1}{2}} \log T \sup_{1 \leq l \leq m} \mathcal{C}_{D_l,\lambda} \mathcal{B}_{D_l,\lambda} (\mathcal{C}_{D_l,\lambda} + \mathcal{G}_{D_l,\lambda}) \\ & + C'_{r,\theta} \left(15 + \frac{1}{1-\theta} \right) \sigma^{-2} T^{(\frac{5(1-\theta)}{2})} \log T \sup_{1 \leq l \leq m} \mathcal{C}_{D_l,\lambda}. \end{aligned} \tag{44}$$

Now the desired bound for $\|\bar{f}_{T+1,D} - f_{T+1}\|$ in (40) follows by combining the estimations for J_1, J_2, and J_3 and the constant is given by

$$C''_{r,\theta} := \Big(\frac{2M\theta}{1-\theta} + C_{\rho,\theta,r} + D_{\rho,\theta,r} \Big) + 3C'_{r,\theta} \left(15 + \frac{1}{1-\theta} \right) + \frac{2^3 M^3 \eta}{5(1-\theta)}.$$

This proves the theorem. □

4.4. Proofs

Now we can prove Theorem 1.

Proof. Firstly, note that with the choice $T = \lfloor N^{\frac{1}{(2r+s)(1-\theta)}} \rfloor$ and $\lambda = T^{-(1-\theta)}$, and under the restriction (8) on m, we have

$$\mathcal{A}_{D,\lambda} \leq N^{-1+\frac{1}{4r+2s}} + \sqrt{C}N^{-\frac{1}{2}+\frac{s}{4r+2s}} \leq (\sqrt{C}+1)N^{-\frac{r}{2r+s}}.$$

Therefore,

$$\begin{aligned}\mathcal{A}_{D_l,\lambda} &\leq mN^{-1}N^{\frac{1}{4r+2s}} + \sqrt{C}m^{\frac{1}{2}}N^{-\frac{1}{2}}N^{\frac{s}{4r+2s}} \\ &\leq (1+\sqrt{C})m^{\frac{1}{2}}N^{-\frac{r}{2r+s}}\end{aligned}$$

and

$$\frac{\mathcal{A}_{D_l,\lambda}}{\sqrt{\lambda}} \leq (1+\sqrt{C})m^{\frac{1}{2}}N^{-\frac{r}{2r+s}}N^{\frac{1}{4r+2s}} \leq (1+\sqrt{C}).$$

By applying Lemma 1, for any $1 \leq l \leq m$, we have with confidence at least $1-\frac{\delta}{6m}$,

$$\mathcal{B}_{D_l,\lambda} \leq 2\Big(\frac{2\mathcal{A}_{D_l,\lambda}\log\frac{12m}{\delta}}{\sqrt{\lambda}}\Big)^2 + 2, \qquad \mathcal{C}_{D_l,\lambda} \leq 2\mathcal{A}_{D_l,\lambda}\log\frac{12m}{\delta}, \qquad \mathcal{G}_{D_l,\lambda} \leq 4\mathcal{A}_{D_l,\lambda}M\log\frac{12m}{\delta}.$$

Consequently, these bounds hold simultaneously with confidence at least $1-\frac{\delta}{2}$. This implies that with confidence at least $1-\frac{\delta}{2}$, there holds

$$\begin{aligned}&\lambda^{-\frac{1}{2}}\log T \sup_{1\leq l\leq m} \mathcal{C}_{D_l,\lambda}\mathcal{B}_{D_l,\lambda}(\mathcal{C}_{D_l,\lambda}+\mathcal{G}_{D_l,\lambda}) \\ &\leq 2^6(M+1)\log T\Big[\Big(\frac{\mathcal{A}_{D_l,\lambda}}{\sqrt{\lambda}}\Big)^2+1\Big]\frac{\mathcal{A}^2_{D_l,\lambda}}{\sqrt{\lambda}}\left(\log\frac{12m}{\delta}\right)^4 \\ &\leq 2^6(M+1)\Big[\Big(1+\sqrt{C}\Big)^2+1\Big]^2 mN^{-\frac{2r-\frac{1}{2}}{2r+s}}\log T\left(\log\frac{12m}{\delta}\right)^4 \\ &\leq 2^{10}(M+1)\Big[\Big(1+\sqrt{C}\Big)^2+1\Big]^2 mN^{-\frac{2r-\frac{1}{2}}{2r+s}}\log T(\log m)^4\left(\log\frac{12}{\delta}\right)^4 \\ &\leq \frac{2^{10}(M+1)\Big[\Big(1+\sqrt{C}\Big)^2+1\Big]^2}{(2r+s)(1-\theta)} mN^{-\frac{2r-\frac{1}{2}}{2r+s}}(\log N)^5\left(\log\frac{12}{\delta}\right)^4 \\ &\leq \frac{2^{10}(M+1)\Big[\Big(1+\sqrt{C}\Big)^2+1\Big]^2}{(2r+s)(1-\theta)} N^{-\frac{r}{2r+s}}\left(\log\frac{12}{\delta}\right)^4\end{aligned} \tag{45}$$

and

$$
\begin{aligned}
&\sigma^{-2}T^{\frac{5(1-\theta)}{2}}\left(1+(\log T)\sup_{1\le l\le m}\mathcal{C}_{D_l,\lambda}\right)\\
&\le \sigma^{-2}T^{\frac{5(1-\theta)}{2}}\left(1+(\log T)\mathcal{A}_{D_l,\lambda}\log\frac{12m}{\delta}\right)\\
&\le 2\sigma^{-2}N^{\frac{\frac{5}{2}}{2r+s}}\left(1+\frac{2+2\sqrt{C}}{(2r+s)(1-\theta)}(\log N)m^{\frac{1}{2}}N^{-\frac{r}{2r+s}}\log m\log\frac{12}{\delta}\right)\\
&\le 2\sigma^{-2}N^{\frac{\frac{5}{2}}{2r+s}}\left(1+\frac{2+2\sqrt{C}}{(2r+s)(1-\theta)}m^{\frac{1}{2}}N^{-\frac{r}{2r+s}}(\log N)^2\log\frac{12}{\delta}\right)\\
&\le 2\sigma^{-2}N^{\frac{\frac{5}{2}}{2r+s}}\left(1+\frac{2+2\sqrt{C}}{(2r+s)(1-\theta)}\right)\log\frac{12}{\delta}.
\end{aligned}
\tag{46}
$$

By Lemma 1, we have with confidence at least $1-\frac{\delta}{4}$,

$$
\mathcal{C}_{D,\lambda}\le 2\mathcal{A}_{D,\lambda}\log\frac{8}{\delta}\le 2(\sqrt{C}+1)N^{-\frac{r}{2r+s}}\log\frac{12}{\delta} \tag{47}
$$

and

$$
\mathcal{G}_{D,\lambda}\le 2M\mathcal{A}_{D,\lambda}\log\frac{8}{\delta}\le 2M(\sqrt{C}+1)N^{-\frac{r}{2r+s}}\log\frac{12}{\delta}. \tag{48}
$$

Plugging the estimates (45)–(48) into (40), we obtain with confidence at least $1-\delta$,

$$
\|\hat{f}_{T+1,D}-f_{T+1}\|\le C\left(N^{-\frac{r}{2r+s}}+\sigma^{-2}N^{\frac{\frac{5}{2}}{2r+s}}\right)\left(\log\frac{12}{\delta}\right)^4
$$

where

$$
\begin{aligned}
C=C''_{r,\theta,p}\Bigg[&2M(\sqrt{C_1}+1)+2(\sqrt{C}+1)\\
&+\frac{2^{10}(M+1)\left[\left(1+\sqrt{C}\right)^2+1\right]^2}{(2r+s)(1-\theta)}+2\left(1+\frac{2+2\sqrt{C}}{(2r+s)(1-\theta)}\right)\Bigg].
\end{aligned}
$$

This, together with the bound

$$
\|f_{T+1}-f_\rho\|\le h_\rho T^{-r(1-\theta)}\le h_\rho N^{-\frac{r}{2r+s}},
$$

leads to the desired conclusion with $\tilde{C}=C+h_\rho$. □

Proof of Corollary 1. When $\sigma\ge N^{\frac{r/2+5/4}{2r+s}}$, by Theorem 1, we have that with confidence at least $1-\delta$, $\|\bar{f}_{T+1,D}-f_\rho\|\le 2\tilde{C}N^{-\frac{r}{2r+s}}\left(\log\frac{12}{\delta}\right)^4$. Replacing $2\tilde{C}N^{-\frac{r}{2r+s}}\left(\log\frac{12}{\delta}\right)^4$ by t, then

$$
\mathbf{Prob}\left\{D:\|\bar{f}_{T+1,D}-f_\rho\|\ge t\right\}\le 12\exp\left\{-(2\tilde{C})^{-\frac{1}{4}}N^{\frac{r}{4(2r+p)}}t^{\frac{1}{4}}\right\}.
$$

Using the probability to expectation formula

$$
\mathbb{E}[\xi]=\int_0^\infty \Pr\{\xi\ge t\}dt
$$

with $\xi=\|\bar{f}_{T+1,D}-f_\rho\|$, we have

$$\begin{aligned}\mathbb{E}\left[\|\bar{f}_{T+1,D}-f_\rho\|\right] &= \int_0^\infty \mathbf{Prob}\{D:\|\bar{f}_{T+1,D}-f_\rho\|\geq t\}dt \leq 12\int_0^\infty \exp\left\{-(2\tilde{C})^{-\frac{1}{4}}N^{\frac{r}{4(2r+p)}}t^{\frac{1}{4}}\right\}dt\\ &= 324\tilde{C}N^{-\frac{r}{2r+s}}\int_0^\infty u^3e^{-u}du = 324\tilde{C}\Gamma(4)N^{-\frac{r}{2r+s}},\end{aligned}$$

where $\Gamma(d)$ is the Gamma function defined for $u>0$ by $\Gamma(d)=\int_0^\infty u^{d-1}e^{-u}du$. The proof is complete. □

To prove Corollary 2, we need the following Borel-Cantelli Lemma which is provided in [31].

Lemma 6. *Let $\{a_N\}$ be a sequence of events in some probability space and $\{\xi_N\}$ be a sequence of positive numbers satisfying $\lim_{N\to\infty}\xi_N=0$. If*

$$\sum_{N=1}^{\infty}\mathbf{Prob}\{|a_N-a|>\xi_N\}<\infty,$$

then a_N will almost certainly converge to a.

Proof of Corollary 2. Let $\delta=N^{-2}$ in Theorem 1; then we have

$$\mathbf{Prob}\left\{N^{\frac{r}{2r+s}}\|\bar{f}_{T+1,D}-f_\rho\|_\rho\geq 2^5\tilde{C}(\log 12N)\right\}<N^{-2}.$$

Thus, for any $\epsilon>0$,

$$\mathbf{Prob}\left\{N^{\frac{r}{2r+s}-\epsilon}\|\bar{f}_{T+1,D}-f_\rho\|\geq 2^5\tilde{C}(\log 12N)N^{-\epsilon}\right\}<N^{-2}.$$

Applying Lemma 6 with $a_N=N^{\frac{r}{2r+s}-\epsilon}\|\bar{f}_{T+1,D}-f_\rho\|$, $a=0$ and $\xi_N=2^5\tilde{C}(\log 12N)N^{-\epsilon}$, we can obtain the conclusion of Corollary 2 by noting $\lim_{N\to\infty}\xi_N=0$ and $\sum_{N=1}^{\infty}N^{-2}<\infty$. The proof is finished. □

Author Contributions: Validation, F.X. and S.W.; Writing (original draft), B.W.; Writing (review and editing), T.H. All authors have read and agreed to the published version of the manuscript.

Funding: The work is supported partially by the National Key Research and Development Program of China (Grant No. 2021YFA1000600) and the National Natural Science Foundation of China (Grant No.12071356).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Guo, Z.C.; Lin, S.B.; Zhou, D.X. Learning theory of distributed spectral algorithms. *Inverse Probl.* **2017**, *33*, 074009. [CrossRef]
2. Mücke, N.; Blanchard, G. Parallelizing spectrally regularized kernel algorithms. *J. Mach. Learn. Res.* **2018**, *19*, 1069–1097.
3. Lin, S.B.; Guo, X.; Zhou, D.X. Distributed learning with regularized least squares. *J. Mach. Learn. Res.* **2017**, *18*, 3202–3232.
4. Hu, T.; Zhou, D.X. Distributed regularized least squares with flexible Gaussian kernels. *Appl. Comput. Harmon. Anal.* **2021**, *53*, 349–377. [CrossRef]
5. Zhang, Y.; Duchi, J.; Wainwright, M. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *J. Mach. Learn. Res.* **2015**, *16*, 3299–3340.
6. Lin, S.B.; Zhou, D.X. Distributed kernel-based gradient descent algorithms. *Constr. Approx.* **2018**, *47*, 249–276. [CrossRef]
7. Shamir, O.; Srebro, N. Distributed stochastic optimization and learning. In Proceedings of the 2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 30 September–3 October 2014; pp. 850–857.
8. Chang, X.; Lin, S.B.; Zhou, D.X. Distributed semi-supervised learning with kernel ridge regression. *J. Mach. Learn. Res.* **2017**, *18*, 1493–1514.
9. Hu, T.; Wu, Q.; Zhou, D.X. Distributed kernel gradient descent algorithm for minimum error entropy principle. *Appl. Comput. Harmon. Anal.* **2020**, *49*, 229–256. [CrossRef]

10. Sun, H.; Wu, Q. Optimal Rates of Distributed Regression with Imperfect Kernels. *J. Mach. Learn. Res.* **2021**, *22*, 1–34.
11. Sun, Q.; Zhou, W.X.; Fan, J. Adaptive Huber regression. *J. Am. Stat. Assoc.* **2020**, *115*, 254–265. [CrossRef]
12. Feng, Y.; Wu, Q. A Framework of Learning Through Empirical Gain Maximization. *Neural Comput.* **2021**, *33*, 1656–1697. [CrossRef]
13. Erdogmus, D.; Principe, J.C. Comparison of entropy and mean square error criteria in adaptive system training using higher order statistics. *Proc. ICA* **2000**, *5*, 6.
14. Song, Y.; Liang, X.; Zhu, Y.; Lin, L. Robust variable selection with exponential squared loss for the spatial autoregressive model. *Comput. Stat. Data Anal.* **2021**, *155*, 107094. [CrossRef]
15. Feng, Y.; Fan, J.; Suykens, J.A. A Statistical Learning Approach to Modal Regression. *J. Mach. Learn. Res.* **2020**, *21*, 1–35.
16. Feng, Y.; Huang, X.; Shi, L.; Yang, Y.; Suykens, J.A. Learning with the maximum correntropy criterion induced losses for regression. *J. Mach. Learn. Res.* **2015**, *16*, 993–1034.
17. Feng, Y.; Ying, Y. Learning with correntropy-induced losses for regression with mixture of symmetric stable noise. *Appl. Comput. Harmon. Anal.* **2020**, *48*, 795–810. [CrossRef]
18. Gunduz, A.; Principe, J.C. Correntropy as a novel measure for nonlinearity tests. *Signal Process.* **2009**, *89*, 14–23. [CrossRef]
19. He, R.; Zheng, W.S.; Hu, B.G.; Kong, X.W. A regularized correntropy framework for robust pattern recognition. *Neural Comput.* **2011**, *23*, 2074–2100. [CrossRef]
20. Bessa, R.J.; Miranda, V.; Gama, J. Entropy and correntropy against minimum square error in offline and online three-day ahead wind power forecasting. *IEEE Trans. Power Syst.* **2009**, *24*, 1657–1666. [CrossRef]
21. Holland, P.W.; Welsch, R.E. Robust regression using iteratively reweighted least-squares. *Commun. Stat.-Theory Methods* **1977**, *6*, 813–827. [CrossRef]
22. Aronszajn, N. Theory of reproducing kernels. *Trans. Am. Math. Soc.* **1950**, *68*, 337–404. [CrossRef]
23. Smale, S.; Zhou, D.X. Learning theory estimates via integral operators and their approximations. *Constr. Approx.* **2007**, *26*, 153–172. [CrossRef]
24. Caponnetto, A.; De Vito, E. Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.* **2007**, *7*, 331–368. [CrossRef]
25. Steinwart, I.; Christmann, A. *Support Vector Machines*; Springer Science & Business Media: Berlin, Germany, 2008.
26. Blanchard, G.; Mücke, N. Optimal rates for regularization of statistical inverse learning problems. *Found. Comput. Math.* **2018**, *18*, 971–1013. [CrossRef]
27. Santamaría, I.; Pokharel, P.P.; Principe, J.C. Generalized correlation function: Definition, properties, and application to blind equalization. *IEEE Trans. Signal Process.* **2006**, *54*, 2187–2197. [CrossRef]
28. Liu, W.; Pokharel, P.P.; Principe, J.C. Correntropy: Properties and applications in non-Gaussian signal processing. *IEEE Trans. Signal Process.* **2007**, *55*, 5286–5298. [CrossRef]
29. Principe, J.C. *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*; Springer: New York, NY, USA, 2010.
30. Yao, Y.; Rosasco, L.; Caponnetto, A. On early stopping in gradient descent learning. *Constr. Approx.* **2007**, *26*, 289–315. [CrossRef]
31. Durrett, R. *Probability: Theory and Examples*; Cambridge University Press: Cambridge, UK, 2017.