Distributed Kernel Gradient Descent Algorithm for Minimum Error Entropy Principle

Ting Hu

School of Mathematics and Statistics, Wuhan University,
Wuhan 430072, China
Email: tinghu@whu.edu.cn

Qiang Wu

Department of Mathematical Sciences, Middle Tennessee State University,
Murfreesboro, TN 37132, USA
Email: qwu@mtsu.edu

Ding-Xuan Zhou

School of Data Science and Department of Mathematics, City University of Hong Kong,
Kowloon, Hong Kong, China
Email: mazhou@cityu.edu.hk

Abstract

Distributed learning based on the divide and conquer approach is a powerful too for big data processing. We introduce a distributed kernel gradient descent algorithm for the minimum error entropy principle and analyze its convergence. We show that the L^2 error decays at a minimax optimal rate under some mild conditions. As a tool we establish some concentration inequalities for U-statistics which play pivotal roles in our error analysis.

Keywords. Distributed learning, minimum error entropy, gradient descent algorithm, kernel method

1 Introduction

Distributed learning has received increasing attention in recent years for its power to handle big data. Among many strategies of distributed learning, the divide and conquer approach has been shown simple and effective. It starts with a data set that is stored distributively in local machines or dividing the whole data set into multiple subsets that are distributed to local machines, then applies a base algorithm to analyze each subset, and finally pools the information together by simple averaging. This approach is computationally efficient by enabling parallel computing in the second stage and can preserve data security and privacy by minimizing mutual information communications. Recently, it was also shown to be consistent for several base algorithms and sometimes achieve optimal learning rates. For instance, in [23], it was proved that the M-estimation of a fixed number of parameters by a distributed method is first order equivalent to the estimation using the whole data set and thus preserves statistical properties such as efficiency and robustness. In [33, 20], divide and conquer method for regression analysis with kernel ridge regression was shown to achieve optimal learning rates in a minimax sense provided that the number of subsets satisfies some constraints. Similar results were also extended to the spectral algorithm [13, 15], the gradient descent algorithm [21], and the bias correct regularization kernel network [14].

Minimum error entropy (MEE) was proposed as an alternative to the least square method in the literature of adaptive systems [8]. It was motivated to minimize the information contained in the prediction error to improve prediction accuracy. Recall that the least square regression is optimal for Gaussian noise but suboptimal for non-Gaussian noise. MEE shows to be robust to deal with heavy tailed or non-Gaussian impulse noises. Therefore, it has received considerable study in the literature and is widely used for many learning tasks; see [7, 9, 12, 25, 26, 5, 16, 17, 10, 3, 4, 28, 24] and a vast references therein. The MEE method is usually implemented by gradient descent algorithms and their convergence has been proved in [5, 18, 19]. In this paper we are interested in the implementation of MEE by a distributed gradient descent method in a big data setting. Note that the MEE loss function involves a pair of observations and is non-convex. So its analysis is essentially different from the least square method. Although existing works on distributed learning do shed some light on the understanding of distributed MEE, they do not apply directly. Rigorous analysis of distributed MEE is more involved and necessary to derive the consistency and learning rates.

The main contributions of this paper include the following. (i) We derive error bounds for the distributed kernel gradient descent MEE algorithm and show the algorithm can achieve the minimax optimal learning rate. This is a completely new result. (ii) As a byproduct, we prove that the kernel gradient MEE algorithm on a single data set can achieve the minimax optimal rate, which improves existing results in the literature. (iii) We establish some concentration inequalities for distributed U-statistics. They play pivotal roles in our analysis of the distributed MEE algorithm in this paper and could potentially apply to the analysis of other pairwise learning methods such as bipartite ranking, gradient learning, and AUC maximization.

The rest of this paper is arranged as follows. In Section 2 we give some notations and assumptions used throughout the paper and present our main results. In Section 3 we prove some useful concentration inequalities for distributed U-statistics. In Section 4 we present some useful lemmas for the proof. The proof of the main results is given in Section 5. Simulations are done in Section 6 to illustrate our theory. We close with some further discussions in Section 7.

To make it easy to follow our presentation below, in Table 1 we summarize some notations that are repeatedly used throughout this paper.

Table 1: List of notations used throughout the paper

Notation	Meaning of the notation
X	the input variable
Y	the response variable
$\mathcal{X}$	the sample space of X, a compact subset of a Euclidian space
$\mathcal{Y}$	the sample space of Y, a subset of $\mathbb{R}$
$\mathcal{Z}$	the product space $\mathcal{X} \times \mathcal{Y}$
ρ	an unknown probability measure on $\mathcal{Z}$
$\rho_{\mathcal{X}}$	marginal probability measure of ρ on $\mathcal{X}$
$\rho(\cdot\|x)$	conditional probability of Y give $X = x$
f	a function on $\mathcal{X}$
$\widetilde{f}$	a function on $\mathcal{X}^2$ induced from f by $\widetilde{f}(x,u) = f(x) - f(u)$
f_ρ	the mean regression function $f_\rho(x) = \mathbf{E}[Y\|X = x]$
$\widetilde{f}_\rho$	the function on $\mathcal{X}^2$ induced from f_ρ by $\widetilde{f}_\rho(x,u) = f_\rho(x) - f_\rho(u)$
x_i	the ith observation for the input variable X
y_i	the ith observation for the response variable Y
z_i	the paired observation (x_i, y_i)
N	the total number of observations
D	the collection of all observations $D = \{(x_1, y_1) \ldots, (x_N, y_N)\}$
k	the number of subsets that the whole data D is partitioned into
D_l	the lth subset of D
m	the sample size of each subset D_l, $m = \frac{N}{k}$ assuming N is divisible by k
G	loss function of MEE algorithm
K	a producing kernel on $\mathcal{X}$
$\widetilde{K}$	a reproducing kernel on $\mathcal{X}^2$ induced from K, defined in (2)
$\mathcal{H}_{\widetilde{K}}$	the reproducing kernel Hilbert space associated to $\widetilde{K}$
$L_{\widetilde{K}}$	integral operator associated to $\widetilde{K}$
$f_{t+1,D}$	the function output by the kernel gradient descent MEE algorithm with data D and kernel K after t iterations
$\widetilde{f}_{t+1,D}$	the function output by pairwise kernel gradient descent MEE algorithm with data D and kernel $\widetilde{K}$ after t iterations
$\widetilde{f}_{t+1,D_l}$	the function output by pairwise kernel gradient descent MEE algorithm with data D_l and kernel $\widetilde{K}$ after t iterations
$\overline{\widetilde{f}_{t+1,D}}$	the solution of distributed pairwise kernel gradient descent MEE algorithm after t iterations, equal to the average of $\widetilde{f}_{t+1,D_l}, l = 1, \ldots, k$
$\lfloor N/4 \rfloor$	the largest integer not exceeding $N/4$

2 Main results

Throughout this paper, let X be the input variable of predictors and Y the response variable. Assume they are linked by a regression model

$$Y = f^*(X) + \epsilon$$

with ϵ a noise variable having conditional mean zero given X. Assume the sample space of X is a compact subset $\mathcal{X}$ of $\mathbb{R}^n$ and the sample space of Y is a bounded subset $\mathcal{Y}$ of $\mathbb{R}$. Denote by ρ the joint probability measure of (X, Y) on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Let $\rho_{\mathcal{X}}$ be the marginal distribution of ρ on $\mathcal{X}$ and $\rho(\cdot|x)$ the conditional distribution of ρ for given $x \in \mathcal{X}$. The purpose of regression analysis is to infer an estimated model f_D from a sample $D = \{(x_i, y_i)\}_{i=1}^N$ of N observations drawn independently from ρ.

The MEE criterion was introduced for linear models in [8, 7] and extended to kernel models in [17, 19]. We focus on the kernel models in this paper because linear models can be regarded as special cases with linear kernels and are relatively easy to analyze. The Rényi quadratic entropy based MEE method minimizes the *empirical Rényi quadratic entropy*

$$\hat{H}_2(f) = -\log\left\{\frac{1}{N^2 h}\sum_{i=1}^N\sum_{j=1}^N G\left(\frac{[(y_i - f(x_i)) - (y_j - f(x_j))]^2}{2h^2}\right)\right\}$$

in a hypothesis space of functions, where $G : \mathbb{R}_+ \to \mathbb{R}$ is a loss function. Since the log function is monotone and does not affect the minimizer, the MEE method can be implemented by minimizing the transformed empirical risk

$$\mathcal{R}_D(f) = -\frac{h^2}{N^2}\sum_{i=1}^N\sum_{j=1}^N G\left(\frac{[(y_i - f(x_i)) - (y_j - f(x_j))]^2}{2h^2}\right).$$

The kernel MEE method minimizes $\mathcal{R}_D(f)$ in a reproducing kernel Hilbert space. A Mercer kernel is a continuous, symmetric, and positive semidefinite function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The linear span of the function set $\{K_x = K(x, \cdot), x \in \mathcal{X}\}$ with the inner product induced by $\langle K_x, K_y\rangle_K = K(x, y)$ forms a pre-Hilbert space. Its completion is a reproducing kernel Hilbert space $\mathcal{H}_K$. The reproducing property is given by $f(x) = \langle f, K_x\rangle_K$ and implies $\|f\|_\infty \le \sup_{x\in\mathcal{X}} \sqrt{K(x, x)}\|f\|_K$. The kernel MEE method is usually solved by the gradient descent algorithm.

Definition 2.1. *Given a sample $D = \{z_i = (x_i, y_i)\}_{i=1}^N$, the kernel gradient decent algorithm for MEE is defined by $f_{1,D} = 0$ and for $t \ge 2$,*

$$f_{t+1,D} = f_{t,D} - \eta_t \times \left\{\frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N -G'\Big(\frac{\xi^t(z_i, z_j)}{2h^2}\Big)\xi^t(z_i, z_j)(K_{x_j} - K_{x_i})\right\}, \tag{1}$$

where $\eta_t > 0$ is the step size and $\xi^t(z_i, z_j) = y_i - f_t(x_i) - (y_j - f_t(x_j))$.

Similar to other kernel based gradient descent algorithms, early stopping is required to avoid overfitting [29, 31]. Since the loss function of MEE is non-convex, the kernel gradient descent algorithm for MEE can be even more complicated. In an earlier work [19] we proved its consistency by a covering number based argument, which, however, did not provide the optimal learning rate.

In this paper we consider a setting that the data is big or arrives naturally in a distributed manner so that the kernel gradient descent algorithm cannot be done by a single processor and a distributed approach has to be used. We decompose the data set D into k disjoint subset $\{D_l\}_{l=1}^k$ of equal size so that each subset $D_l = \{z_i^{(l)} = (x_i^{(l)}, y_i^{(l)})\}_{i=1}^m$ has sample size $|D_l| = m = \frac{N}{k}$. Let f_{t,D_l} be the time t output of the kernel gradient descent algorithm (1) on D_l. The time t output of the distributed kernel gradient descent algorithm for MEE is $\bar{f}_{t+1,D} = \frac{1}{k}\sum_{l=1}^k f_{t+1,D_l}$. However, we will not analyze this scheme directly. Instead, we will analyze an equivalent scheme by using some advanced techniques that have been recently developed for pairwise learning [30]. For this purpose, define the pairwise kernel $\widetilde{K} : \mathcal{X}^2 \times \mathcal{X}^2 \to \mathbb{R}$ by

$$\begin{aligned} \widetilde{K}((x_1, x_2), (u_1, u_2)) &= K(x_1, u_1) + K(x_2, u_2) - K(x_2, u_1) - K(x_1, u_2) \\ &= \langle K_{x_1} - K_{x_2}, K_{u_1} - K_{u_2} \rangle_K. \end{aligned} \tag{2}$$

For each function f on $\mathcal{X}$, denote the function $f(x) - f(x')$ by $\tilde{f}(x, x') : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. It is verified that $\widetilde{K}$ defines a reproducing kernel and $\langle \widetilde{f}, \widetilde{K}_{(x,x')} \rangle_{\widetilde{K}} = \widetilde{f}(x, x')$ for all $\widetilde{f} \in \mathcal{H}_{\widetilde{K}}$. Define the integral operator $L_{\widetilde{K}} : \mathcal{H}_{\widetilde{K}} \to \mathcal{H}_{\widetilde{K}}$ by

$$L_{\widetilde{K}}(\widetilde{f}) = \int_{\mathcal{X}} \int_{\mathcal{X}} \left\langle \widetilde{f}, \widetilde{K}_{(x,x')} \right\rangle_{\widetilde{K}} \widetilde{K}_{(x,x')} d\rho_{\mathcal{X}} d\rho_{\mathcal{X}}, \qquad \widetilde{f} \in \mathcal{H}_{\widetilde{K}},$$

and the empirical operator $L_{\widetilde{K},D}$ on $\mathcal{H}_{\widetilde{K}}$ by

$$\begin{aligned} L_{\widetilde{K},D}(\widetilde{f}) &= \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left\langle \widetilde{f}, \widetilde{K}_{(x_i,x_j)} \right\rangle_{\widetilde{K}} \widetilde{K}_{(x_i,x_j)}, \\ &= \frac{1}{|D|^2} \sum_{x,x' \in D(x)} \left\langle \widetilde{f}, \widetilde{K}_{(x,x')} \right\rangle_{\widetilde{K}} \widetilde{K}_{(x,x')}, \qquad \widetilde{f} \in \mathcal{H}_{\widetilde{K}}. \end{aligned}$$

Here and in the following $|D| := N$ denotes the cardinal of the set D and $D(x) := \{x_i\}_{i=1}^N = \{x :$ there exists some y such that $(x, y) \in D\}$. With these notations, we see that the algorithm (1) is equivalent to

$$\begin{aligned} \widetilde{f}_{t+1,D} &= \widetilde{f}_{t,D} - \eta_t \times \left\{ \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} -G'\Big(\frac{\xi^t(z_i, z_j)}{2h^2}\Big) \xi^t(z_i, z_j) \widetilde{K}_{(x_i,x_j)} \right\} \\ &= \widetilde{f}_{t,D} - \frac{\eta_t}{|D|^2} \sum_{(x,y),(x',y') \in D} -G'\Big(\frac{\xi^t(z, z')}{2h^2}\Big) \xi^t(z, z') \widetilde{K}_{(x,x')}, \end{aligned} \tag{3}$$

where $\xi^t(z, z') = y - y' - \widetilde{f}_{t,D}(x, x')$, $z = (x, y)$ and $z' = (x', y')$. Correspondingly we have an equivalent scheme for the distributed kernel gradient descent algorithm with output

$$\overline{\widetilde{f}_{t+1,D}} = \frac{1}{k} \sum_{l=1}^{k} \widetilde{f}_{t+1,D_l}.$$

The goal of this paper is to estimate the learning error between $\overline{\widetilde{f}_{t+1,D}}$ and $\widetilde{f}_\rho$ in the $L^2_{\rho_\mathcal{X} \times \rho_\mathcal{X}}$-space. For simplicity, in the sequel we will use $\| \cdot \|$ to denote the norm $\| \cdot \|_{L^2_{\rho_\mathcal{X} \times \rho_\mathcal{X}}}$ with respect to the $L^2_{\rho_\mathcal{X} \times \rho_\mathcal{X}}$ space when the meaning is clear from the context.

Throughout the paper, we assume, without loss of generality, that

$$\kappa := \sup_{(x,x') \in \mathcal{X}^2} \sqrt{\widetilde{K}((x, x'), (x, x'))} \leq 1$$

and $|y| \leq M_\rho$ for some $M_\rho > 0$. It is easy to get that $\|\widetilde{f}_\rho\|_\infty = \sup\limits_{x,x' \in \mathcal{X}} |f_\rho(x) - f_\rho(x')| \leq 2M_\rho$. We assume that there exist some $0 < s \leq 1$ and $C_0 > 0$ such that

$$\mathcal{N}(\lambda) = \mathrm{Tr}\big[L_{\widetilde{K}}(L_{\widetilde{K}} + \lambda I)^{-1}\big] \leq C_0 \lambda^{-s}, \qquad \forall\ \lambda > 0, \tag{4}$$

and

$$\widetilde{f}_\rho = L_{\widetilde{K}}^r g, \quad \text{for some } r > 0 \text{ and } g \in L^2_{\rho_{\mathcal{X}} \times \rho_{\mathcal{X}}}. \tag{5}$$

The assumption (4) measures the capacity of $\mathcal{H}_{\widetilde{K}}$ by the effective dimension, that is, the trace of the operator $L_{\widetilde{K}}(L_{\widetilde{K}} + \lambda I)^{-1}$. Note it always holds with $s = 1$. For $s < 1$, it is almost equivalent to that the eigenvalues σ_i of $L_{\widetilde{K}}$ decay at a rate $i^{-\frac{1}{s}}$. The smoother the kernel function $\widetilde{K}$ is, the smaller s and the smaller function space $\mathcal{H}_{\widetilde{K}}$. In particular, if $\widetilde{K} \in C^\infty$, then s can be arbitrarily small, as is the case for Gaussian kernels. The assumption (5) measures the regularity of the target function. It is the well known source condition. In general, $\widetilde{f}_\rho$ is smoother if r is larger. Both conditions (4) and (5) are widely used in the learning theory literature.

For the loss function used in MEE, we assume that $G'(0) = -1$, $G'(x) < 0$ for any $x > 0$, $\sup_{x\in\mathbb{R}} |G'(x)| \le C_G$, and there exists some $c_p > 0$, $p > 0$ such that $|G'(x) - G'(0)| \le c_p|x|^p$ for all $0 \le x \le 1$.

Theorem 2.2. *Assume that* (4) *and* (5) *hold for some* $r > \frac{1}{2}$ *and* $0 < s \le 1$. *Take* $\eta_t = \eta t^{-\theta}$ *with* $0 < \eta \le \min\{\frac{1}{C_G}, 1\}$ *and* $0 \le \theta < 1$. *If* $T = \lfloor N/4 \rfloor^{\frac{1}{(2r+s)(1-\theta)}}$ *and*

$$k \le \frac{N^{\frac{r-\frac{1}{2}}{2r+s}}}{(\log N)^5}, \tag{6}$$

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

$$\|\overline{\widetilde{f}_{T+1,D}} - \widetilde{f}_\rho\| \le C^* \left\{ N^{-\frac{r}{2r+s}} + N^{\frac{p+\frac{3}{2}}{2r+s}} h^{-2p} \right\} \left(\log \frac{12}{\delta} \right)^4.$$

where C^* *is a constant depending on* θ, r, p.

A direct implication of this theorem is that the kernel gradient descent MEE on a single data set (i.e. $k = 1$) can achieve a convergence rate $O(N^{-\frac{r}{2r+s}})$ if the bandwidth parameter h is chosen to be large enough. This is minimax optimal for the regularized least square regression [1] when $r \ge \frac{1}{2}$. The convergence analyses in the MEE literature [5, 18, 19] always present results that are worse than the regularized least square regression due to the pairwise feature of MEE algorithms. We in this paper overcome this difficulty and the result shows MEE can achieve the same minimax rate as the least square method.

In [30, 31] online pairwise learning with the least square loss or a general convex loss has been investigated. Online learning is different from gradient descent algorithm in two aspects. First, at each step t, online learning has only access to the sample $(x_i, y_i), i = 1, \ldots, t$, while gradient descent algorithm has access to all samples. Second, online learning has to complete $T = N$ iterations to go through all N samples while gradient descent algorithm can stop with $T \ll N$ iterations to avoid overfitting, which is the well known early stopping rule. To our best knowledge, in the analysis of online learning, it is difficult, if not impossible, to study the impact of the capacity of reproducing kernel Hilbert spaces. As a consequence, in [30, 31], under the assumption $r = \frac{1}{2}$ so that $\widetilde{f}_\rho \in \mathcal{H}_{\widetilde{K}}$ capacity independent rate $O(N^{-\frac{1}{6}})$ for the excess expected risk was obtained for online pairwise learning with the least square loss and $O(N^{-\frac{1}{5}})$ was obtained when the loss function is convex and has a bounded gradient. Our rate $O(N^{-\frac{1}{2(1+s)}})$ for the kernel gradient descent MEE, either in the single data set case or in the distributed case, is clearly faster.

For distributed regression with the regularized least square kernel method, the minimax optimal rate has been verified in [33, 20, 13] under different restrictions on the number of local machines. The most recent result in [13] states the restriction as

$$k \le N^{\min\{\frac{2}{2r+s}, \frac{2r-1}{2r+s}\}}. \tag{7}$$

When the gradient descent algorithm is used the restriction obtained in [21] is

$$k \leq \frac{N^{\frac{r-\frac{1}{2}}{2r+s}}}{(\log N)^4 + 1}. \tag{8}$$

Note (7) suffers a saturation effect that the number of local machines cannot increase when $r \geq \frac{3}{2}$. The restriction in (8) is worse than (7) when $r < \frac{5}{2}$ but better when $r > \frac{5}{2}$ as an award for overcoming the saturation effect. Our result for the distributed kernel gradient descent MEE algorithm is quite similar to (8). They differ only up to a logarithmic term which has minimal effect and is caused by the difficulty to handle the pairwise and non-convexity features of MEE algorithms.

3 Concentration inequalities for distributed U-statistics

In this section we prove some concentration inequalities for distributed U-statistics that will be used in the proof of our main results. We need the following lemma whose proof follows some standard techniques from [22] and will be given in the appendix.

Lemma 3.1. *Let $\{d_j\}_{j=1}^N$ be a sequence of martingale differences with values in a Hilbert space $(\mathcal{H}, \|\cdot\|)$ and $d_0 \equiv 0$. Set the conditional expectation $\mathbb{E}_{j-1}\|d_j\|^2 := \mathbb{E}(\|d_j\|^2 | d_1, \cdots, d_{j-1})$. If $\sum_{j=1}^N \mathbb{E}_{j-1}\|d_j\|^2 \leq \sigma^2 < \infty$ almost surely for some $\sigma^2 > 0$ and $\sup\limits_{1\leq j\leq N} \|d_j\|_\infty \leq M$ for some $M > 0$, then we have for any $c > 0$*

$$\mathbb{E}\cosh\left(c\left\|\sum_{j=1}^N d_j\right\|\right) \leq \exp\left\{\frac{(e^{cM} - 1 - cM)\sigma^2}{M^2}\right\}.$$

Furthermore, for any $\varepsilon > 0$, we also have that

$$\begin{aligned}\min_{c>0}\left\{\frac{\mathbb{E}\cosh\left(c\|\sum_{j=0}^N d_j\|\right)}{\cosh(c\varepsilon)}\right\} &\leq 2\exp\left\{-\frac{\varepsilon}{M}\left\{\left(1 + \frac{\sigma^2}{M\varepsilon}\right)\log\left(1 + \frac{M\varepsilon}{\sigma^2}\right) - 1\right\}\right\} \\ &\leq \exp\left\{-\frac{\varepsilon^2}{2(\sigma^2 + \frac{1}{3}M\varepsilon)}\right\}. \end{aligned} \tag{9}$$

Using Lemma 3.1 we can prove the following concentration inequality for distributed U-Statistics of a Hilbert space valued bivariate random variable.

Theorem 3.2. *Let $\xi(\cdot,\cdot)$ be a symmetric random variable defined on the probability space $(\mathcal{Z}\times\mathcal{Z}, \rho_{\mathcal{Z}}\times\rho_{\mathcal{Z}})$ with values in a Hilbert space $(\mathcal{H}, \|\cdot\|)$. Assume that $\|\xi\|_\infty \leq M$ almost surely and a sample $D = \{z_i\}_{i=1}^N$ is drawn independently from $(\mathcal{Z}, \rho)$. Let D be decomposed randomly into k disjoint subsets $\{D_l\}_{l=1}^k$ such that each subset $D_l = \{z_i^{(l)}\}_{i=1}^m$ has the same sample size $m = \frac{N}{k} \geq 2$. Then, with confidence at least $1-\delta$, we have*

$$\left\|\frac{1}{k}\sum_{l=1}^k \Big[\frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m \xi(z_i^{(l)}, z_j^{(l)})\Big] - \mathbb{E}\xi\right\| \leq \frac{2kM}{N} + \frac{2M\log(2/\delta)}{\lfloor N/4\rfloor} + \sqrt{\frac{2\mathbb{E}\|\xi\|^2\log(2/\delta)}{\lfloor N/4\rfloor}} \tag{10}$$

where $\lfloor N/4\rfloor$ denotes the largest integer not exceeding $N/4$. In particular, if $k = 1$, we have with confidence at least $1-\delta$,

$$\left\|\frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N \xi(z_i, z_j) - \mathbb{E}\xi\right\| \leq \frac{2M}{N} + \frac{2M\log(2/\delta)}{\lfloor N/4\rfloor} + \sqrt{\frac{2\mathbb{E}\|\xi\|^2\log(2/\delta)}{\lfloor N/4\rfloor}}. \tag{11}$$

Proof. We write

$$\begin{aligned}
&\left\|\frac{1}{k}\sum_{l=1}^{k}\Big[\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\xi(z_i^{(l)},z_j^{(l)})\Big]-\mathbb{E}\xi\right\| \\
\leq\ & \frac{1}{m}\left\|\Big[\frac{1}{k}\sum_{l=1}^{k}\frac{1}{m}\sum_{i=1}^{m}\xi(z_i^{(l)},z_i^{(l)})\Big]-\mathbb{E}\xi\right\|+\frac{m-1}{m}\left\|\Big[\frac{1}{k}\sum_{l=1}^{k}\frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}\xi(z_i^{(l)},z_j^{(l)})\Big]-\mathbb{E}\xi\right\| \\
\leq\ & \frac{2kM}{N}+\frac{m-1}{m}\left\|\Big[\frac{1}{k}\sum_{l=1}^{k}\frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}\xi(z_i^{(l)},z_j^{(l)})\Big]-\mathbb{E}\xi\right\|. \qquad (12)
\end{aligned}$$

Since ξ is symmetric on $\mathcal{Z}\times\mathcal{Z}$, we have

$$\frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}\xi(z_i^{(l)},z_j^{(l)})=\frac{1}{C_m^2}\sum_{m,2}\xi(z_{i_1}^{(l)},z_{i_2}^{(l)}),$$

where $C_m^2=\frac{m!}{(m-2)!2!}$ and the summation $\sum_{m,2}$ is taken over all two-tuples (i_1,i_2) of distinct positive integers not exceeding m. For each l, define

$$U_i^{(l)}:=\frac{1}{[m/2]}\Big[\xi(z_{i_1}^{(l)},z_{i_2}^{(l)})+\xi(z_{i_3}^{(l)},z_{i_4}^{(l)})+\cdots+\xi(z_{i_{2[m/2]-1}}^{(l)},z_{i_{2[m/2]}}^{(l)})\Big].$$

Then

$$\frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}\xi(z_i^{(l)},z_j^{(l)})=\frac{1}{C_m^2}\sum_{m,2}\xi(z_{i_1}^{(l)},z_{i_2}^{(l)})=\frac{1}{m!}\sum_{m,m}U_i^{(l)}$$

and

$$\frac{1}{k}\sum_{l=1}^{k}\frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}\xi(z_i^{(l)},z_j^{(l)})=\frac{1}{k}\sum_{l=1}^{k}\Big[\frac{1}{m!}\sum_{m,m}U_i^{(l)}\Big]=\frac{1}{m!}\sum_{m,m}\Big[\frac{1}{k}\sum_{l=1}^{k}U_i^{(l)}\Big]$$

where the summation $\sum_{m,m}$ is taken over all permutations $(i_1,\cdots,i_m)$ of the integers $1,\cdots,m$. Since $U_i^{(1)},\cdots,U_i^{(k)}$ are independent and each $U_i^{(l)}, l=1,\cdots,k$ is a summation of $[m/2]$ independent random variables, we know that $\Big[\frac{1}{k}\sum_{l=1}^{k}U_i^{(l)}\Big]$ is a summation of $k[m/2]\geq\lfloor N/4\rfloor$ independent random variables. By Lemma 3.1 we have for any $\varepsilon\geq 0$

$$\begin{aligned}
\min_{c>0}\left\{\frac{\mathbb{E}\Big[\cosh\Big(c\Big\|\Big[\frac{1}{k}\sum_{l=1}^{k}U_i^{(l)}-\mathbb{E}\xi\Big]\Big\|\Big)\Big]}{\cosh(c\varepsilon)}\right\}&\leq 2\exp\left\{-\frac{\varepsilon^2}{2\left(\frac{\mathbb{E}\|\xi\|^2}{k[m/2]}+\frac{1}{3}\frac{2M}{k[m/2]}\varepsilon\right)}\right\}\\
&\leq 2\exp\left\{-\frac{\lfloor N/4\rfloor\varepsilon^2}{2(\mathbb{E}\|\xi\|^2+\frac{2}{3}M\varepsilon)}\right\}. \qquad (13)
\end{aligned}$$

By the convexity of cosh, we obtain

$$\begin{aligned}
&\mathbf{Prob}\left\{\Big\|\frac{1}{k}\sum_{l=1}^{k}\Big[\frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}\xi(z_i^{(l)},z_j^{(l)})\Big]-\mathbb{E}\xi\Big\|\geq\varepsilon\right\}\\
=\ &\mathbf{Prob}\left\{\Big\|\frac{1}{m!}\sum_{m,m}\Big[\frac{1}{k}\sum_{l=1}^{k}U_i^{(l)}-\mathbb{E}\xi\Big]\Big\|\geq\varepsilon\right\}
\end{aligned}$$

$$\begin{aligned}
&\leq \min_{c>0}\left\{\frac{\mathbb{E}\Big[\cosh\Big(c\Big\|\frac{1}{m!}\sum_{m,m}\Big[\frac{1}{k}\sum_{l=1}^{k}U_i^{(l)}-\mathbb{E}\xi\Big]\Big\|\Big)\Big]}{\cosh(c\varepsilon)}\right\}\\
&\leq \frac{1}{m!}\sum_{m,m}\min_{c>0}\left\{\frac{\mathbb{E}\Big[\cosh\Big(c\Big\|\Big[\frac{1}{k}\sum_{l=1}^{k}U_i^{(l)}-\mathbb{E}\xi\Big]\Big\|\Big)\Big]}{\cosh(c\varepsilon)}\right\}\\
&\leq 2\exp\left\{-\frac{\lfloor N/4\rfloor\varepsilon^2}{2(\mathbb{E}\|\xi\|^2+\frac{2}{3}M\varepsilon)}\right\}.
\end{aligned}$$

This implies that

$$\left\|\frac{1}{k}\sum_{l=1}^{k}\Big[\frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}\xi(z_i^{(l)},z_j^{(l)})\Big]-\mathbb{E}\xi\right\|\leq\frac{2M\log(2/\delta)}{\lfloor N/4\rfloor}+\sqrt{\frac{2\mathbb{E}\|\xi\|^2\log(2/\delta)}{\lfloor N/4\rfloor}}\tag{14}$$

with confidence at least $1-\delta$. Plugging the estimation (14) into (12) we obtain the the estimation in (10).

The estimation in (11) is a direct corollary of (10) with $k=1$. □

It is worth remarking that if $k=1$ we have $m=N$ and hence $k\lfloor m/2\rfloor=\lfloor N/2\rfloor$. So we see (13) still holds with $\lfloor N/4\rfloor$ replaced by $\lfloor N/2\rfloor$. As a consequence we can improve the inequality (11) a little bit by using $\lfloor N/2\rfloor$ instead of $\lfloor N/4\rfloor$ for the last two terms on the right hand side.

Recall that the definition of the operator $L_{\widetilde{K},D_l}$ on $\mathcal{H}_{\widetilde{K}}$ for each subset D_l is

$$\begin{aligned}
L_{\widetilde{K},D_l}(\widetilde{f})&=\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\langle\widetilde{f},\widetilde{K}_{(x_i,x_j)}\rangle_{\widetilde{K}}\widetilde{K}_{(x_i,x_j)}\\
&=\frac{1}{|D_l|^2}\sum_{x,x'\in D_l(x)}\left\langle\widetilde{f},\widetilde{K}_{(x,x')}\right\rangle_{\widetilde{K}}\widetilde{K}_{(x,x')},\quad\forall\widetilde{f}\in\mathcal{H}_{\widetilde{K}}.
\end{aligned}$$

Let I denote the identity operator. Consider the random variable

$$\xi(x,x')=(L_{\widetilde{K}}+\lambda I)^{-\frac{1}{2}}\left\langle\cdot,\widetilde{K}_{(x,x')}\right\rangle_{\widetilde{K}}\widetilde{K}_{(x,x')}.$$

Then $\mathbb{E}\xi(x,x')=(L_{\widetilde{K}}+\lambda I)^{-\frac{1}{2}}L_{\widetilde{K}}$ and the corresponding distributed sampling average is

$$\frac{1}{k}\sum_{l=1}^{k}\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\xi(x_i^{(l)},x_j^{(l)})=(L_{\widetilde{K}}+\lambda I)^{-\frac{1}{2}}\frac{1}{k}\sum_{l=1}^{k}L_{\widetilde{K},D_l}.$$

By a similar procedure to that in [20], we obtain the following lemmas.
Lemma 3.3. *With confidence at least* $1-\delta$, *we have*

$$\left\|(L_{\widetilde{K}}+\lambda I)^{-\frac{1}{2}}\left(L_{\widetilde{K}}-\frac{1}{k}\sum_{l=1}^{k}L_{\widetilde{K},D_l}\right)\right\|\leq 2\mathcal{A}_{D,\lambda,k}\log\frac{2}{\delta}$$

where

$$\mathcal{A}_{D,\lambda,k}=\frac{k}{|D|\sqrt{\lambda}}+\frac{1}{[|D|/4]\sqrt{\lambda}}+\sqrt{\frac{\mathcal{N}(\lambda)}{[|D|/4]}}.$$

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

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

and for each $l = 1, \cdots, k$, with confidence at least $1-\delta$,

$$\left\|(L_{\widetilde{K}} + \lambda I)^{-\frac{1}{2}}(L_{\widetilde{K}} - L_{\widetilde{K},D_l})\right\| \leq 2\mathcal{A}_{D_l,\lambda} \log \frac{2}{\delta}.$$

where $\mathcal{A}_{D,\lambda} = \mathcal{A}_{D,\lambda,1}$ and $\mathcal{A}_{D_l,\lambda} = \mathcal{A}_{D_l,\lambda,1}$.

Lemma 3.4. *With confidence at least with $1-\delta$,*

$$\left\|(L_{\widetilde{K},D} + \lambda I)^{-1}(L_{\widetilde{K}} + \lambda I)\right\| \leq 2\Big(\frac{2\mathcal{A}_{D,\lambda} \log \frac{2}{\delta}}{\sqrt{\lambda}}\Big)^2 + 2. \tag{15}$$

Lemma 3.5. *Assume that $g(z, z')$ is a measurable function defined on $\mathcal{Z} \times \mathcal{Z}$ with $\|g\|_\infty \leq M'$ almost surely for some $M' > 0$ and $D_l = \{z_i^{(l)}\}_{i=1}^m = \{(x_i^{(l)}, y_i^{(l)})\}_{i=1}^m, 1 \leq l \leq k$. With confidence at least $1-\delta$,*

$$\left\|\frac{1}{k}\sum_{l=1}^{k}\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}(L_{\widetilde{K}} + \lambda I)^{-\frac{1}{2}}\Big[g(z_i^{(l)}, z_j^{(l)})\widetilde{K}_{(x_i^{(l)}, x_j^{(l)})} - L_{\widetilde{K}}g\Big]\right\| \leq 2M'\mathcal{A}_{D,\lambda,k} \log \frac{2}{\delta}.$$

Specially, with confidence at least $1-\delta$, when $k = 1$, $D = \{z_i\}_{i=1}^N = \{(x_i, y_i)\}_{i=1}^N$,

$$\left\|\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(L_{\widetilde{K}} + \lambda I)^{-\frac{1}{2}}\Big[g(z_i, z_j)\widetilde{K}_{(x_i, x_j)} - L_{\widetilde{K}}g\Big]\right\| \leq 2M'\mathcal{A}_{D,\lambda} \log \frac{2}{\delta}.$$

4 Lemmas

We provide some useful lemmas before moving to the proof of our main results. Let π_i^t denote the polynomial defined by $\pi_i^t(s) = \prod_{j=i}^t (1 - \eta_j x)$ if $i \leq t$ 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 $\pi_i^t(L_{\widetilde{K}})$ and $\pi_i^t(L_{\widetilde{K},D})$. For this purpose we introduce the conventional notation $\sum_{j=T+1}^T := 1$ the following preliminary lemma.

Lemma 4.1. *If $0 \leq \alpha < 1, 0 \leq \theta < 1$,then for $T \geq 3$,*

$$\sum_{i=1}^{T} i^{-(\theta+\alpha)}\left(\sum_{j=i+1}^{T} j^{-\theta}\right)^{-1} \leq C_{\theta,\alpha} T^{-\min\{\alpha, 1-\theta\}} \log T, \tag{16}$$

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

$$\sum_{i=1}^{T} i^{-\theta}\left(\sum_{j=i+1}^{T} j^{-\theta}\right)^{-1} \leq 15 \log T. \tag{17}$$

The proof of Lemma 4.1 is given in the appendix. We now derive bounds for several operators that are used in the proof of our main theorem.

Lemma 4.2. *If $\eta_t = \eta t^{-\theta}$ with $0 < \eta < 1$ and $0 \leq \theta < 1$, then for $1 \leq i \leq T-1$,*

$$\left\|\pi_i^t(L_{\widetilde{K},D})\right\| \leq 1 \tag{18}$$

$$\left\|\pi_i^t(L_{\widetilde{K}})\right\| \leq 1 \tag{19}$$

$$\left\| L_{\widetilde{K},D}\pi_{i+1}^T(L_{\widetilde{K},D})\right\| \leq \left(e\eta \sum_{j=i+1}^{T} j^{-\theta}\right)^{-1}, \tag{20}$$

$$\left\| L_{\widetilde{K}}\pi_{i+1}^T(L_{\widetilde{K}})\right\| \leq \left(e\eta \sum_{j=i+1}^{T} j^{-\theta}\right)^{-1}, \tag{21}$$

$$\Big\| \sum_{i=1}^{T} \eta_i \big[(L_{\widetilde{K},D} + \lambda I)\pi_{i+1}^T(L_{\widetilde{K},D})\big]\Big\| \leq 1 + \frac{\eta\lambda}{1-\theta}T^{1-\theta}, \tag{22}$$

$$\Big\| \sum_{i=1}^{T} \eta_i \big[(L_{\widetilde{K}} + \lambda I)\pi_{i+1}^T(L_{\widetilde{K}})\big]\Big\| \leq 1 + \frac{\eta\lambda}{1-\theta}T^{1-\theta}. \tag{23}$$

Proof. Since $\|L_{\widetilde{K},D}\| \leq \kappa \leq 1$ and $0 < \eta < 1$, we have for each $1 \leq t \leq T$, the operator $I - \eta_t L_{\widetilde{K},D}$ is positive with a operator norm bounded by 1. This implies (18). The conclusion (19) follows analogously by noting $\|L_{\widetilde{K}}\| \leq 1$.

Denote by $\Lambda(L_{\widetilde{K},D})$ the eigenvalue set of the operator $L_{\widetilde{K},D}$ on $\mathcal{H}_{\widetilde{\mathcal{K}}}$. Since all eigenvalues of $L_{\widetilde{K},D}$ is bounded by $\|L_{\widetilde{K},D}\| \leq \kappa \leq 1$, we have

$$\begin{aligned}\left\|(L_{\widetilde{K},D})\pi_{i+1}^T(L_{\widetilde{K},D})\right\| &= \sup_{s\in\Lambda(L_{\widetilde{K},D})} \left|s\pi_{i+1}^T(s)\right| \leq \sup_{0<s\leq 1} |s\pi_{i+1}^T(s)| \\ &\leq \sup_{0<x\leq 1} s\exp\left\{ - s\sum_{j=i+1}^{T}\eta_j\right\} = \left(e\sum_{j=i+1}^{T}\eta_j\right)^{-1} \\ &= \left(e\eta\sum_{j=i+1}^{T} j^{-\theta}\right)^{-1}.\end{aligned}$$

This proves (20). The conclusion (21) follows similarly.

To prove (22), note that

$$\pi_i^T(L_{\widetilde{K},D}) = (I - \eta_i L_{\widetilde{K},D})\pi_{i+1}^T(L_{\widetilde{K},D}) = \pi_{i+1}^T(L_{\widetilde{K},D}) - \eta_i L_{\widetilde{K},D}\pi_{i+1}^T(L_{\widetilde{K},D}).$$

It follows that

$$\sum_{i=1}^{T}\eta_i L_{\widetilde{K},D}\pi_{i+1}^T(L_{\widetilde{K},D}) = \pi_{T+1}^T(L_{\widetilde{K},D}) - \pi_1^T(L_{\widetilde{K},D}) = I - \pi_1^T(L_{\widetilde{K},D})$$

and

$$\left\|\sum_{i=1}^{T}\eta_i L_{\widetilde{K},D}\pi_{i+1}^T(L_{\widetilde{K},D})\right\| = \left\|I - \pi_1^T(L_{\widetilde{K},D})\right\| \leq 1.$$

Therefore,

$$\begin{aligned}\left\|\sum_{i=1}^{T}\eta_i(L_{\widetilde{K},D} + \lambda)\pi_{i+1}^T(L_{\widetilde{K},D})\right\| &\leq \left\|\sum_{i=1}^{T}\eta_i L_{\widetilde{K},D}\pi_{i+1}^T(L_{\widetilde{K},D})\right\| + \lambda\left\|\sum_{i=1}^{T}\eta_i\pi_{i+1}^T(L_{\widetilde{K},D})\right\| \\ &\leq 1 + \lambda\sum_{i=1}^{T}\eta_i = 1 + \eta\lambda\sum_{i=1}^{T} i^{-\theta}.\end{aligned}$$

By Lemma 7.1 we obtain the desired bound in (22). The conclusion (23) follows in a similar way. □

Define a data-free gradient descent sequence for the least square method in $\mathcal{H}_{\widetilde{K}}$ by $\widetilde{f}_1 = 0$ and

$$\widetilde{f}_{t+1} = \widetilde{f}_t - \eta_t \int_X \int_X \left(\widetilde{f}_t(x,x') - \widetilde{f}_\rho(x,x')\right) \widetilde{K}(x,x') d\rho_{\mathcal{X}} \rho_{\mathcal{X}} = (I - \eta_t L_{\widetilde{K}})\widetilde{f}_t + \eta_t L_{\widetilde{K}} \widetilde{f}_\rho. \tag{24}$$

The difference $\widetilde{f}_{t+1} - \widetilde{f}_\rho$ measures the step t optimization error of the kernel gradient descent sequence for the least square method. It has been well investigated in the literature, e.g. [29]. Under the assumption (5) with $r > \frac{1}{2}$, there hold

$$\|\widetilde{f}_t - \widetilde{f}_\rho\| \le h_\rho t^{-r(1-\theta)} \tag{25}$$

and

$$\|\widetilde{f}_t - \widetilde{f}_\rho\|_{\widetilde{K}} \le h_\rho t^{-(r-\frac{1}{2})(1-\theta)}, \tag{26}$$

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

We will need the following estimations in the our proof.

Lemma 4.3. *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*

$$\sum_{i=1}^T \eta_i \|L_{\widetilde{K},D} \pi_{i+1}^T(L_{\widetilde{K},D})\| \|\widetilde{f}_i - \widetilde{f}_\rho\|_{\widetilde{K}} \le C_{\rho,\theta,r} \tag{27}$$

and

$$\sum_{i=1}^T \eta_i \|L_{\widetilde{K}} \pi_{i+1}^T(L_{\widetilde{K}})\| \|\widetilde{f}_i - \widetilde{f}_\rho\|_{\widetilde{K}} \le C_{\rho,\theta,r} \tag{28}$$

Proof. By (26), Lemma 4.1, and Lemma 4.2 we have

$$\begin{aligned}\sum_{i=1}^T \eta_i \|L_{\widetilde{K},D} \pi_{i+1}^T(L_{\widetilde{K},D})\| \|\widetilde{f}_i - \widetilde{f}_\rho\|_{\widetilde{K}} &\le h_\rho \sum_{i=1}^T i^{-\theta} \left(\sum_{j=i+1}^T j^{-\theta}\right)^{-1} i^{-(r-\frac{1}{2})(1-\theta)} \\ &\le h_\rho C_{\theta,(r-\frac{1}{2})(1-\theta)} T^{-\min\{1-\theta,(r-\frac{1}{2})(1-\theta)\}} \log T \\ &\le h_\rho C_{\theta,(r-\frac{1}{2})(1-\theta)} \left(e \min\{1-\theta, (r-\tfrac{1}{2})(1-\theta)\}\right)^{-1},\end{aligned}$$

where $C_{\theta,(r-\frac{1}{2})(1-\theta)}$ is defined in Lemma 4.1. This proves (27) with

$$C_{\rho,\theta,r} = h_\rho C_{\theta,(r-\frac{1}{2})(1-\theta)} \left(e \min\{1-\theta, (r-\tfrac{1}{2})(1-\theta)\}\right)^{-1}.$$

The estimate (28) follows analogously. □

Lemma 4.4. *If $\eta_t = \eta t^{-\theta}$ with $0 < \eta < 1$ and $0 \le \theta < 1$, then there is a constant $D_{\rho,\theta,r}$ such that*

$$\sum_{i=1}^T \eta_i \|\widetilde{f}_i - \widetilde{f}_\rho\|_{\widetilde{K}} \le D_{\rho,\theta,r} T^{1-\theta} \tag{29}$$

Proof. By (26) we obtain

$$\sum_{i=1}^T \eta_i \|\widetilde{f}_i - \widetilde{f}_\rho\|_{\widetilde{K}} \le h_\rho \eta \sum_{i=1}^T i^{-\theta-(r-\frac{1}{2})(1-\theta)}.$$

By Lemma 7.1, we know that

$$\sum_{i=1}^{T} i^{-\theta-(r-\frac{1}{2})(1-\theta)} \le \begin{cases} \frac{T^{(1-\theta)(\frac{3}{2}-r)}}{(1-\theta)(\frac{3}{2}-r)}, & \text{if} \quad r < \frac{3}{2}, \\ 2, & \text{if} \quad r = \frac{3}{2}, \\ \frac{\theta+(r-\frac{1}{2})(1-\theta)}{(1-\theta)(r-\frac{3}{2})}, & \text{if} \quad r > \frac{3}{2}. \end{cases}$$

The restriction $r > \frac{1}{2}$ implies that $(\frac{3}{2} - r)(1-\theta) \le (1-\theta)$. Then

$$\sum_{i=1}^{T} i^{-\theta-(r-\frac{1}{2})(1-\theta)} \le D_{\theta,r} T^{1-\theta}$$

where $D_{\theta,r}$ is defined by

$$D_{\theta,r} = \begin{cases} \frac{1}{(1-\theta)(\frac{3}{2}-r)}, & \text{if} \quad r < \frac{3}{2}, \\ 2, & \text{if} \quad r = \frac{3}{2}, \\ \frac{\theta+(r-\frac{1}{2})(1-\theta)}{(1-\theta)(r-\frac{3}{2})}, & \text{if} \quad r > \frac{3}{2}. \end{cases}$$

Therefore the desired conclusion holds with $D_{\rho,\theta,r} = h_\rho D_{\theta,r}$. □

The last fact we would recall in this section is the isomorphism between $\mathcal{H}_{\widetilde{K}}$ and $L^2_{\rho_\mathcal{X} \times \rho_\mathcal{X}}$, which states that

$$\|F\| = \|L_{\widetilde{K}}^{\frac{1}{2}} F\|_{\widetilde{K}} \le \|(L_{\widetilde{K}} + \lambda I)^{\frac{1}{2}} F\|_{\widetilde{K}}, \qquad \forall\ F \in \mathcal{H}_{\widetilde{K}}. \tag{30}$$

5 Proof

In this section we prove our main results. The estimation of the distributed solution path depends on the estimation of solution paths for the subsets. So we will first investigate the properties of the kernel gradient descent solution path on a single data set and then move to the analysis of the distributed solution path.

5.1 Bounding the solution path for a single data set

We first establish some upper bounds for the solution path on a single data set.

Theorem 5.1. *If the step size sequence satisfies $0 < \eta_t \le 1/C_G$, then we have the following bound for the learning sequence $\{\widetilde{f}_{t,D}\}$:*

$$\|\widetilde{f}_{t,D}\|_{\widetilde{K}} \le 2M_\rho \sqrt{C_G \sum_{i=1}^{t-1} \eta_i}, \quad t \in \mathbb{N}.$$

If $\eta_t = \eta t^{-\theta}$ with $0 < \eta \le 1/C_G$ and $0 \le \theta < 1$, then

$$\|\widetilde{f}_{t,D}\|_{\widetilde{K}} \le 2M_\rho t^{\frac{1-\theta}{2}}. \tag{31}$$

Proof. We prove the conclusion by induction. First note the conclusion holds trivially for $t = 1$. Next, suppose that $\|\widetilde{f}_{t,D}\|_{\widetilde{K}} \le 2M_\rho \sqrt{C_G \sum_{i=1}^{t-1} \eta_i}$ holds. By the updating rule (3) and the reproducing property, we have

$$\|\widetilde{f}_{t+1,D}\|^2_{\widetilde{K}} = \|\widetilde{f}_{t,D}\|^2_{\widetilde{K}} + \frac{2\eta_t}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} G'\Big(\frac{\xi^t(z_i,z_j)}{2h^2}\Big) \xi^t(z_i,z_j) \widetilde{f}_{t,D}(x_i,x_j)$$

$$
\begin{aligned}
&+\frac{\eta_t^2}{N^4}\left\|\sum_{i=1}^{N}\sum_{j=1}^{N}G'\Big(\frac{\xi^t(z_i,z_j)}{2h^2}\Big)\xi^t(z_i,z_j)\widetilde{K}_{(x_i,x_j)}\right\|_{\widetilde{K}}^2\\
\le&\|\widetilde{f}_{t,D}\|_{\widetilde{K}}^2+\frac{2\eta_t}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}G'\Big(\frac{\xi^t(z_i,z_j)}{2h^2}\Big)\xi^t(z_i,z_j)\widetilde{f}_{t,D}(x_i,x_j)\\
&+\frac{\eta_t^2}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|G'\Big(\frac{\xi^t(z_i,z_j)}{2h^2}\Big)\right|^2\big(\xi^t(z_i,z_j)\big)^2\\
=&\|\widetilde{f}_{t,D}\|_{\widetilde{K}}^2+\frac{\eta_t}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}Q_{ij}, \qquad (32)
\end{aligned}
$$

where

$$
\begin{aligned}
Q_{ij}=&\left[\eta_t\left|G'\Big(\frac{\xi^t(z_i,z_j)}{2h^2}\Big)\right|^2+2G'\Big(\frac{\xi^t(z_i,z_j)}{2h^2}\Big)\right]\Big(\widetilde{f}_{t,D}(x_i,x_j)\Big)^2\\
&-2\left(G'\Big(\frac{\xi^t(z_i,z_j)}{2h^2}\Big)+\eta_t\left|G'\Big(\frac{\xi^t(z_i,z_j)}{2h^2}\Big)\right|^2\right)(y_i-y_j)\widetilde{f}_{t,D}(x_i,x_j)\\
&+\eta_t\left|G'\Big(\frac{\xi^t(z_i,z_j)}{2h^2}\Big)\right|^2(y_i-y_j)^2.
\end{aligned}
$$

The restriction $\eta_t \le \frac{1}{C_G}$ implies $\eta_t|G'(\frac{\xi^t(i,j)}{2h^2})|^2+2G'(\frac{\xi^t(i,j)}{2h^2})<0$. By the property of quadratic function, we have

$$
\begin{aligned}
Q_{ij}\le&\ \eta_t\left|G'\Big(\frac{\xi^t(z_i,z_j)}{2h^2}\Big)\right|^2(y_i-y_j)^2-\frac{\Big(G'(\frac{\xi^t(z_i,z_j)}{2h^2})+\eta_t|G'(\frac{\xi^t(z_i,z_j)}{2h^2})|^2\Big)^2(y_i-y_j)^2}{\eta_t|G'(\frac{\xi^t(z_i,z_j)}{2h^2})|^2+2G'(\frac{\xi^t(z_i,z_j)}{2h^2})}\\
=&\ \frac{|G'(\frac{\xi^t(z_i,z_j)}{2h^2})|(y_i-y_j)^2}{2-\eta_t|G'(\frac{\xi^t(z_i,z_j)}{2h^2})|}\le 4M_\rho^2C_G.
\end{aligned}
$$

Plugging it into (32) we obtain

$$
\|\widetilde{f}_{t+1,D}\|_{\widetilde{K}}^2\le\|\widetilde{f}_{t,D}\|_{\widetilde{K}}^2+4M_\rho^2C_G\eta_t\le 4M_\rho^2C_G\sum_{i=1}^{t}\eta_i.
$$

This completes the proof. □

We remark that, if $\frac{1}{2}\le\theta<1$, a bound $\|f_{t,D}\|_K\le M_\rho t^{\frac{1-\theta}{2}}$ has been proved in [19] and (31) is an easy corollary. But the proof in [19] does not extend to $0\le\theta<\frac{1}{2}$.

5.2 Error bound for kernel gradient descent MEE on a single data

We bound the learning error of the kernel gradient descent MEE algorithm by decomposing it into two terms,

$$
\left\|\widetilde{f}_{t+1,D}-\widetilde{f}_\rho\right\|\le\left\|\widetilde{f}_{t+1,D}-\widetilde{f}_{t+1}\right\|+\|\widetilde{f}_{t+1}-\widetilde{f}_\rho\|. \qquad (33)
$$

As we have mentioned in Section 3, the second term is the step t optimization error of the kernel gradient descent sequence for the least square method and can be bounded by (25) under the assumption (5) with $r>\frac{1}{2}$.

The first term on the right hand side of (33) involves two errors: the sample error caused by approximating the population gradient by sample gradient and the error caused by the deviation of the non-convex MEE loss from the convex square loss. We bound it by the following theorem.

Theorem 5.2. *Define $\{\widetilde{f}_t\}$ by (24). Assume that (5) holds for some $r > \frac{1}{2}$. Let $\eta_t = \eta t^{-\theta}$ with $0 < \eta \leq \min\{\frac{1}{C_G}, 1\}$ and $0 \leq \theta < 1$. For $\lambda > 0$, there hold*

$$\|\widetilde{f}_{T+1,D} - \widetilde{f}_{T+1}\| \leq C'_{r,\theta,p}\Big[\mathcal{B}_{D,\lambda}(\mathcal{C}_{D,\lambda} + \mathcal{G}_{D,\lambda})(1 + \lambda T^{1-\theta}) + T^{(1-\theta)(p+\frac{3}{2})}h^{-2p}\Big], \tag{34}$$

and

$$\|\widetilde{f}_{T+1,D} - \widetilde{f}_{T+1}\|_{\widetilde{K}} \leq C'_{r,\theta,p}\Big[\mathcal{B}_{D,\lambda}(\mathcal{C}_{D,\lambda} + \mathcal{G}_{D,\lambda})(1 + \lambda T^{1-\theta})/\sqrt{\lambda} + T^{(1-\theta)(p+\frac{3}{2})}h^{-2p}\Big], \tag{35}$$

where

$$\begin{aligned}
\mathcal{B}_{D,\lambda} =& \|(L_{\widetilde{K},D} + \lambda I)^{-1}(L_{\widetilde{K}} + \lambda I)\|,\\
\mathcal{C}_{D,\lambda} =& \|(L_{\widetilde{K}} + \lambda I)^{-\frac{1}{2}}(L_{\widetilde{K}} - L_{\widetilde{K},D})\|,\\
\mathcal{G}_{D,\lambda} =& \|(L_{\widetilde{K}} + \lambda I)^{-\frac{1}{2}}(L_{\widetilde{K}}\widetilde{f}_\rho - \hat{f}_{\rho,D})\|_{\widetilde{K}},\\
\widehat{f}_{\rho,D} =& \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(y_i - y_j)\widetilde{K}_{(x_i,x_j)} = \frac{1}{|D|^2}\sum_{z,z'\in D}(y - y')\widetilde{K}_{(x,x')},
\end{aligned} \tag{36}$$

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

Proof. By the definition of $\widetilde{f}_{t,D}$ in (3) and the definition of $\widetilde{f}_t$ in (24), we have

$$\widetilde{f}_{t+1,D} - \widetilde{f}_{t+1} = [I - \eta_t L_{\widetilde{K},D}](\widetilde{f}_{t,D} - \widetilde{f}_t) + \eta_t[L_{\widetilde{K}} - L_{\widetilde{K},D}]\widetilde{f}_t + \eta_t[\hat{f}_{\rho,D} - L_{\widetilde{K}}(\widetilde{f}_\rho)] + \eta_t E_{t,D}, \tag{37}$$

where $\widehat{f}_{\rho,D}$ is defined in (36) and

$$E_{t,D} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(G'\Big(\frac{\xi_t(z_i,z_j)}{2h^2}\Big) - G'(0)\right)\left(\widetilde{f}_{t,D}(x_i,x_j) - y_i + y_j\right)\widetilde{K}(x_i,x_j),$$

Applying (37) iteratively from $t = 1$ to T, we obtain

$$\widetilde{f}_{T+1,D} - \widetilde{f}_{T+1} = I_1 + I_2 + I_3 + I_4 \tag{38}$$

where

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

For I_1, by (30), Lemma 4.3 and Lemma 4.4,

$$\begin{aligned}
\|I_1\| = &\left\|\sum_{i=1}^{T}\eta_i(L_{\widetilde{K}}+\lambda I)^{\frac{1}{2}}\pi_{i+1}^{T}(L_{\widetilde{K},D})[L_{\widetilde{K}}-L_{\widetilde{K},D}](\widetilde{f}_i-\widetilde{f}_\rho)\right\|_{\widetilde{K}} \\
\leq &\sum_{i=1}^{T}\left\{\eta_i\left\|(L_{\widetilde{K}}+\lambda I)^{\frac{1}{2}}(L_{\widetilde{K},D}+\lambda I)^{-\frac{1}{2}}\right\|\left\|(L_{\widetilde{K},D}+\lambda I)\pi_{i+1}^{T}(L_{\widetilde{K},D})\right\|\right. \\
&\quad\left.\times\left\|(L_{\widetilde{K},D}+\lambda I)^{-\frac{1}{2}}(L_{\widetilde{K}}+\lambda I)^{\frac{1}{2}}\right\|\left\|(L_{\widetilde{K}}+\lambda I)^{-\frac{1}{2}}[L_{\widetilde{K}}-L_{\widetilde{K},D}]\right\|\|\widetilde{f}_i-\widetilde{f}_\rho\|_{\widetilde{K}}\right\} \\
\leq &\,\mathcal{B}_{D,\lambda}\mathcal{C}_{D,\lambda}\left(\sum_{i=1}^{T}\eta_i\|L_{\widetilde{K},D}\pi_{i+1}^{T}(L_{\widetilde{K},D})\|\|\widetilde{f}_i-\widetilde{f}_\rho\|_{\widetilde{K}}+\lambda\sum_{i=1}^{T}\eta_i\|\widetilde{f}_i-\widetilde{f}_\rho\|_{\widetilde{K}}\right) \\
\leq &\,\mathcal{B}_{D,\lambda}\mathcal{C}_{D,\lambda}\Big(C_{\rho,\theta,r}+D_{\rho,\theta,r}\lambda T^{1-\theta}\Big). \qquad (39)
\end{aligned}$$

For I_2, by (30), Lemma 4.2, and the fact $\|\widetilde{f}_\rho\|_\infty \leq 2M_\rho$, we have

$$\begin{aligned}
\|I_2\| = &\left\|\sum_{i=1}^{T}\eta_i\pi_{i+1}^{T}(L_{\widetilde{K},D})[L_{\widetilde{K}}-L_{\widetilde{K},D}](\widetilde{f}_\rho)\right\| \\
\leq &\left\|\sum_{i=1}^{T}\eta_i(L_{\widetilde{K}}+\lambda I)^{\frac{1}{2}}\pi_{i+1}^{T}(L_{\widetilde{K},D})[L_{\widetilde{K}}-L_{\widetilde{K},D}](\widetilde{f}_\rho)\right\|_{\widetilde{K}} \\
\leq &\left\|(L_{\widetilde{K}}+\lambda I)^{\frac{1}{2}}(L_{\widetilde{K},D}+\lambda I)^{-\frac{1}{2}}\right\|\left\|\sum_{i=1}^{T}\eta_i(L_{\widetilde{K},D}+\lambda I)\pi_{i+1}^{T}(L_{\widetilde{K},D})\right\| \\
&\times\left\|(L_{\widetilde{K},D}+\lambda I)^{-\frac{1}{2}}(L_{\widetilde{K}}+\lambda I)^{\frac{1}{2}}\right\|\left\|(L_{\widetilde{K}}+\lambda I)^{-\frac{1}{2}}[L_{\widetilde{K}}-L_{\widetilde{K},D}]\right\|\left\|\widetilde{f}_\rho\right\|_{\widetilde{K}} \\
\leq &\,2M_\rho\left(1+\frac{\lambda T^{1-\theta}}{1-\theta}\right)\mathcal{B}_{D,\lambda}\mathcal{C}_{D,\lambda}. \qquad (40)
\end{aligned}$$

Similarly, we can bound I_3 as

$$I_3 \leq \left(1+\frac{\lambda T^{1-\theta}}{1-\theta}\right)\mathcal{B}_{D,\lambda}\mathcal{G}_{D,\lambda}. \qquad (41)$$

For I_4, first note that by the bound (31) of $\{f_{t,D}\}$, we see

$$\begin{aligned}
&\left\|\left(G'\Big(\frac{\xi_t(z_i,z_j)}{2h^2}\Big)-G'(0)\right)(\widetilde{f}_{t,D}(x_i,x_j)-y_i+y_j)\widetilde{K}(x_i,x_j)\right\|_{\widetilde{K}} \\
\leq &\, c_p\frac{(2M_\rho+2\|\widetilde{f}_{t,D}\|_{\widetilde{K}})^{2p+1}}{2^p h^{2p}} \leq c_p\frac{2^{3p+2}}{h^{2p}}\|\widetilde{f}_{t,D}\|_{\widetilde{K}}^{2p+1} \\
\leq &\, c_p 2^{5p+2}M_\rho^{2p+1}t^{\frac{(1-\theta)(2p+1)}{2}}h^{-2p}
\end{aligned}$$

This implies that

$$\|E_{t,D}\|_{\widetilde{K}} \leq c_p 2^{5p+2}M_\rho^{2p+1}t^{\frac{(1-\theta)(2p+1)}{2}}h^{-2p}. \qquad (42)$$

This together with the estimate $\|\pi_{i+1}^{t}(L_{\widetilde{K},D})\| \leq 1$ gives

$$\|I_4\| \leq \sum_{i=1}^{T}\eta_i\|E_{i,D}\|_{\widetilde{K}} \leq c_p 2^{5p+2}M_\rho^{2p+1}\eta\sum_{i=1}^{T}i^{\frac{(1-\theta)(2p+1)}{2}-\theta}h^{-2p}$$

$$\leq \frac{c_p 2^{5p+2} M_\rho^{2p+1}}{(1-\theta)(p+\frac{3}{2})} T^{(1-\theta)(p+\frac{3}{2})} h^{-2p}. \tag{43}$$

Combining the estimates in (40), (41), (43) and (39) we obtain (34) with

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

Following a similar process we can obtain the bound in (35). □

5.3 Error bound for distributed approach

Now we turn to bound the error of the distributed kernel gradient descent MEE. For this purpose we decompose the error $\|\overline{\widetilde{f}_{T+1,D}} - \widetilde{f}_\rho\|$ into two parts as

$$\|\overline{\widetilde{f}_{T+1,D}} - \widetilde{f}_\rho\| \leq \|\widetilde{f}_{T+1} - \widetilde{f}_\rho\| + \|\overline{\widetilde{f}_{T+1,D}} - \widetilde{f}_{T+1}\|.$$

The following theorem provides a bound for the first term.

Theorem 5.3. *Take $\lambda = T^{-(1-\theta)}$. There is a constant $C''_{r,\theta,p}$ such that*

$$\begin{aligned}\|\overline{\widetilde{f}_{T+1,D}} - \widetilde{f}_{T+1}\| \leq C''_{r,\theta,p}\Big[&\mathcal{F}_{D,\lambda} + \mathcal{D}_{D,\lambda} + \lambda^{-\frac{1}{2}} \log T \sup_{1\leq l\leq k} \mathcal{C}_{D_l,\lambda}\mathcal{B}_{D_l,\lambda}(\mathcal{C}_{D_l,\lambda} + \mathcal{G}_{D_l,\lambda}) \\ &+ h^{-2p} T^{(1-\theta)(p+\frac{3}{2})}\Big(1 + \log T \sup_{1\leq l\leq k} \mathcal{C}_{D_l,\lambda}\Big)\Big],\end{aligned} \tag{44}$$

where

$$\mathcal{D}_{D,\lambda} = \Big\|\frac{1}{k}\sum_{l=1}^{k}(L_{\widetilde{K}} + \lambda)^{-\frac{1}{2}}[L_{\widetilde{K}} - L_{\widetilde{K},D_l}]\Big\|,$$

$$\mathcal{F}_{D,\lambda} = \Big\|\frac{1}{k}\sum_{l=1}^{k}(L_{\widetilde{K}} + \lambda)^{-\frac{1}{2}}[\hat{f}_{\rho,D_l} - L_{\widetilde{K}}(\widetilde{f}_\rho)]\Big\|_{\widetilde{K}}.$$

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

$$\widetilde{f}_{T+1,D_l} - \widetilde{f}_{T+1} = [I - \eta_t L_{\widetilde{K}}](\widetilde{f}_{T,D_l} - \widetilde{f}_t) + \eta_T[L_{\widetilde{K}} - L_{\widetilde{K},D}]\widetilde{f}_{T,D_l} + \eta_T[\hat{f}_{\rho,D_l} - L_{\widetilde{K}}(\widetilde{f}_\rho)] + \eta_T E_{T,D_l}.$$

It implies that

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

and therefore

$$\|\overline{\widetilde{f}_{T+1,D}} - \widetilde{f}_{T+1}\| = \left\|\frac{1}{k}\sum_{l=1}^{k}\Big(\widetilde{f}_{T+1,D_l} - \widetilde{f}_{T+1}\Big)\right\|$$

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

We first estimate J_2. By (30), Lemma 4.2, and the choice $\lambda = T^{-(1-\theta),}$ we obtain

$$
\begin{aligned}
J_2 &\le \left\| \sum_{i=1}^{T} \eta_i (L_{\widetilde{K}} + \lambda) \pi_{i+1}^{T}(L_{\widetilde{K}}) \frac{1}{k} \sum_{l=1}^{k} (L_{\widetilde{K}} + \lambda)^{-\frac{1}{2}} [\hat{f}_{\rho,D_l} - L_{\widetilde{K}}(\widetilde{f}_\rho)] \right\|_{\widetilde{K}} \\
&\le \left(1 + \frac{\lambda T^{1-\theta}}{1-\theta}\right) \left\| \frac{1}{k} \sum_{l=1}^{k} (L_{\widetilde{K}} + \lambda)^{-\frac{1}{2}} [\hat{f}_{\rho,D_l} - L_{\widetilde{K}}(\widetilde{f}_\rho)] \right\|_{\widetilde{K}} \\
&\le \frac{2M_\rho}{1-\theta} \left(1 + \lambda T^{1-\theta}\right) \mathcal{F}_{D,\lambda} \\
&:= \frac{4M_\rho}{1-\theta} \mathcal{F}_{D,\lambda}.
\end{aligned} \tag{45}
$$

For J_3, by (43) we have

$$
J_3 \le \sup_{1 \le l \le k} \left\| \sum_{i=1}^{T} \eta_i \pi_{i+1}^{T}(L_{\widetilde{K}}) E_{i,D_l} \right\| \le \frac{c_p 2^{5p+2} M_\rho^{2p+1} \eta}{(1-\theta)(p+\frac{3}{2})} T^{(1-\theta)(p+\frac{3}{2})} h^{-2p}. \tag{46}
$$

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

$$
\begin{aligned}
J_1 &\le \left\| \sum_{i=1}^{T} \eta_i (L_{\widetilde{K}} + \lambda) \pi_{i+1}^{T}(L_{\widetilde{K}}) \frac{1}{k} \sum_{l=1}^{k} (L_{\widetilde{K}} + \lambda)^{-\frac{1}{2}} [L_{\widetilde{K}} - L_{\widetilde{K},D_l}] \widetilde{f}_{i,D_l} \right\|_{\widetilde{K}} \\
&\le \left\| \sum_{i=1}^{T} \eta_i (L_{\widetilde{K}} + \lambda) \pi_{i+1}^{T}(L_{\widetilde{K}}) \frac{1}{k} \sum_{l=1}^{k} (L_{\widetilde{K}} + \lambda)^{-\frac{1}{2}} [L_{\widetilde{K}} - L_{\widetilde{K},D_l}] (\widetilde{f}_{i,D_l} - \widetilde{f}_i) \right\|_{\widetilde{K}} \\
&\quad + \left\| \sum_{i=1}^{T} \eta_i (L_{\widetilde{K}} + \lambda) \pi_{i+1}^{T}(L_{\widetilde{K}}) \frac{1}{k} \sum_{l=1}^{k} (L_{\widetilde{K}} + \lambda)^{-\frac{1}{2}} [L_{\widetilde{K}} - L_{\widetilde{K},D_l}] (\widetilde{f}_i - \widetilde{f}_\rho) \right\|_{\widetilde{K}} \\
&\quad + \left\| \sum_{i=1}^{T} \eta_i (L_{\widetilde{K}} + \lambda) \pi_{i+1}^{T}(L_{\widetilde{K}}) \frac{1}{k} \sum_{l=1}^{k} (L_{\widetilde{K}} + \lambda)^{-\frac{1}{2}} [L_{\widetilde{K}} - L_{\widetilde{K},D_l}] (\widetilde{f}_\rho) \right\|_{\widetilde{K}} \\
&:= J_{11} + J_{12} + J_{13}.
\end{aligned}
$$

By Lemma 4.3, Lemma 4.4 and the fact $\lambda T^{1-\theta} = 1$, we obtain

$$
\begin{aligned}
J_{12} &\le \mathcal{D}_{D,\lambda} \left(\sum_{i=1}^{T} \left\| \eta_i L_{\widetilde{K}} \pi_{i+1}^{T}(L_{\widetilde{K}}) \right\| \|\widetilde{f}_i - \widetilde{f}_\rho\|_{\widetilde{K}} + \lambda \sum_{i=1}^{T} \eta_i \|\widetilde{f}_i - \widetilde{f}_\rho\|_{\widetilde{K}} \right) \\
&\le \mathcal{D}_{D,\lambda} \Big(C_{\rho,\theta,r} + D_{\rho,\theta,r} \Big).
\end{aligned}
$$

For J_{13}, by (23), we have

$$\begin{aligned} J_{13} &\leq \left\|\sum_{i=1}^{T}\eta_i(L_{\widetilde{K}}+\lambda)\pi_{i+1}^{T}(L_{\widetilde{K}})\right\|\left\|\frac{1}{k}\sum_{l=1}^{k}(L_{\widetilde{K}}+\lambda)^{-\frac{1}{2}}[L_{\widetilde{K}}-L_{\widetilde{K},D_l}]\right\|\left\|\widetilde{f}_{\rho}\right\|_{\widetilde{K}} \\ &\leq 2M_{\rho}\left(1+\frac{\lambda T^{1-\theta}}{1-\theta}\right)\mathcal{D}_{D,\lambda}=\frac{4M_{\rho}}{1-\theta}\mathcal{D}_{D,\lambda}. \end{aligned}$$

Now we turn to J_{11}. We have

$$\begin{aligned} J_{11} &\leq \sum_{i=1}^{T}\left\|\eta_i(L_{\widetilde{K}}+\lambda)\pi_{i+1}^{T}(L_{\widetilde{K}})\right\|\left\|\frac{1}{k}\sum_{l=1}^{k}(L_{\widetilde{K}}+\lambda)^{-\frac{1}{2}}[L_{\widetilde{K}}-L_{\widetilde{K},D_l}](\widetilde{f}_{i,D_l}-\widetilde{f}_i)\right\|_{\widetilde{K}} \\ &\leq \sum_{i=1}^{T}\left\|\eta_i(L_{\widetilde{K}}+\lambda)\pi_{i+1}^{T}(L_{\widetilde{K}})\right\|\sup_{1\leq l\leq k}\left\|(L_{\widetilde{K}}+\lambda)^{-\frac{1}{2}}[L_{\widetilde{K}}-L_{\widetilde{K},D_l}](\widetilde{f}_{i,D_l}-\widetilde{f}_i)\right\|_{\widetilde{K}} \\ &\leq \sum_{i=1}^{T}\eta_i\Big[\Big(\sum_{j=i+1}^{T}\eta_j\Big)^{-1}+\lambda\Big]\sup_{1\leq l\leq k}\mathcal{C}_{D_l,\lambda}\left\|\widetilde{f}_{i,D_l}-\widetilde{f}_i\right\|_{\widetilde{K}}. \end{aligned} \tag{47}$$

By Theorem 5.2 and the choice $\lambda=T^{-(1-\theta)}$, for $1\leq i\leq T$, there holds that $\lambda i^{(1-\theta)}\leq 1$ and

$$\begin{aligned} \left\|\widetilde{f}_{i,D_l}-\widetilde{f}_i\right\|_{\widetilde{K}} &\leq C'_{r,\theta,p}\Big[\mathcal{B}_{D_l,\lambda}(\mathcal{C}_{D_l,\lambda}+\mathcal{G}_{D_l,\lambda})(1+\lambda i^{1-\theta})/\sqrt{\lambda}+i^{(1-\theta)(p+\frac{3}{2})}h^{-2p}\Big] \\ &\leq C'_{r,\theta,p}\Big[2\mathcal{B}_{D_l,\lambda}(\mathcal{C}_{D_l,\lambda}+\mathcal{G}_{D_l,\lambda})/\sqrt{\lambda}+T^{(1-\theta)(p+\frac{3}{2})}h^{-2p}\Big]. \end{aligned}$$

Plugging it into (47) we obtain

$$J_{11}\leq C'_{r,\theta,p}\sup_{1\leq l\leq k}\mathcal{C}_{D_l,\lambda}\Big[2\mathcal{B}_{D_l,\lambda}(\mathcal{C}_{D_l,\lambda}+\mathcal{G}_{D_l,\lambda})/\sqrt{\lambda}+T^{(1-\theta)(p+\frac{3}{2})}h^{-2p}\Big]\sum_{i=1}^{T}\eta_i\Big[\Big(\sum_{j=i+1}^{T}\eta_j\Big)^{-1}+\lambda\Big]$$

By Lemma 7.1 and 4.1 , 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

$$J_{11}\leq C'_{r,\theta,p}\left(15+\frac{1}{1-\theta}\right)\log T\sup_{1\leq l\leq k}\mathcal{C}_{D_l,\lambda}\Big[2\mathcal{B}_{D_l,\lambda}(\mathcal{C}_{D_l,\lambda}+\mathcal{G}_{D_l,\lambda})/\sqrt{\lambda}+T^{(1-\theta)(p+\frac{3}{2})}h^{-2p}\Big]$$

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

$$\begin{aligned} J_1 \leq &\Big(\frac{4M_{\rho}}{1-\theta}+C_{\rho,\theta,r}+D_{\rho,\theta,r}\Big)\mathcal{D}_{D,\lambda} \\ &+2C'_{r,\theta,p}\left(15+\frac{1}{1-\theta}\right)\lambda^{-\frac{1}{2}}\log T\sup_{1\leq l\leq k}\mathcal{C}_{D_l,\lambda}\mathcal{B}_{D_l,\lambda}(\mathcal{C}_{D_l,\lambda}+\mathcal{G}_{D_l,\lambda}) \\ &+C'_{r,\theta,p}\left(15+\frac{1}{1-\theta}\right)h^{-2p}T^{(1-\theta)(p+\frac{3}{2})}\log T\sup_{1\leq l\leq k}\mathcal{C}_{D_l,\lambda}. \end{aligned} \tag{48}$$

Now the desired bound for $\|\bar{f}_{T+1,D}-\widetilde{f}_{T+1}\|$ in (44) follows by combining the estimations for J_1, J_2, and J_3 and the constant is given by

$$C''_{r,\theta,p}:=\Big(\frac{8M_{\rho}\theta}{1-\theta}+C_{\rho,\theta,r}+D_{\rho,\theta,r}\Big)+3C'_{r,\theta,p}\left(15+\frac{1}{1-\theta}\right)+\frac{c_p2^{5p+2}M_{\rho}^{2p+1}\eta}{(1-\theta)(p+\frac{3}{2})}.$$

This proves the theorem. □

5.4 Optimal rate analysis

Now we can prove Theorem 2.2.

Proof. Firstly, note that with the choice $T = \lfloor N/4 \rfloor^{\frac{1}{(2r+s)(1-\theta)}}$ and $\lambda = T^{-(1-\theta)}$, and under the restriction (6) on k, we have

$$\begin{aligned}\mathcal{A}_{D,\lambda,k} &\leq 5^{\frac{1}{4r+2s}} N^{-\frac{r+s}{2r+s}} + \lfloor N/4 \rfloor^{-1+\frac{1}{4r+2s}} + \sqrt{C_0} \lfloor N/4 \rfloor^{-\frac{1}{2}+\frac{s}{4r+2s}} \\ &\leq 5^{\frac{1}{4r+2s}} N^{-\frac{r+s}{2r+s}} + (\sqrt{C_0}+1) \lfloor N/4 \rfloor^{-\frac{r}{2r+s}} \\ &\leq \sqrt{5}(\sqrt{C_0}+2) N^{-\frac{r}{2r+s}},\end{aligned}$$

Therefore,

$$\begin{aligned}\mathcal{A}_{D_l,\lambda} = \mathcal{A}_{D_l,\lambda,1} &= m^{-1} \lfloor N/4 \rfloor^{\frac{1}{4r+2s}} + [m/4]^{-1} \lfloor N/4 \rfloor^{\frac{1}{4r+2s}} + \sqrt{C_0} [m/4]^{-\frac{1}{2}} \lfloor N/4 \rfloor^{\frac{s}{4r+2s}} \\ &\leq k N^{\frac{1}{4r+2s}-1} + 5k N^{\frac{1}{4r+2s}-1} + \sqrt{5C_0} k^{\frac{1}{2}} N^{-\frac{r}{2r+s}} \\ &\leq (6+\sqrt{5C_0}) k^{\frac{1}{2}} N^{-\frac{r}{2r+s}}\end{aligned}$$

and

$$\frac{\mathcal{A}_{D_l,\lambda}}{\sqrt{\lambda}} \leq (6+\sqrt{5C_0}) k^{\frac{1}{2}} N^{-\frac{r}{2r+s}} \lfloor N/4 \rfloor^{\frac{1}{4r+2s}} \leq (6+\sqrt{5C_0}).$$

Applying Lemma 3.3, Lemma 3.4 and Lemma 3.5, for any $1 \leq l \leq k$, we have with confidence at least $1-\frac{\delta}{6k}$,

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

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 hold

$$\begin{aligned}&\lambda^{-\frac{1}{2}} \log T \sup_{1 \leq l \leq k} \mathcal{C}_{D_l,\lambda} \mathcal{B}_{D_l,\lambda} (\mathcal{C}_{D_l,\lambda} + \mathcal{G}_{D_l,\lambda}) \\ &\leq 2^6 (M_\rho+1) \log T \Big[\Big(\frac{\mathcal{A}_{D_l,\lambda}}{\sqrt{\lambda}}\Big)^2 + 1\Big] \frac{\mathcal{A}^2_{D_l,\lambda}}{\sqrt{\lambda}} \left(\log \frac{12k}{\delta}\right)^4 \\ &\leq 2^6 (M_\rho+1) \Big[\Big(6+\sqrt{5C_0}\Big)^2 + 1\Big]^2 k N^{-\frac{2r-\frac{1}{2}}{2r+s}} \log T \left(\log \frac{12k}{\delta}\right)^4 \\ &\leq 2^{10} (M_\rho+1) \Big[\Big(6+\sqrt{5C_0}\Big)^2 + 1\Big]^2 k N^{-\frac{2r-\frac{1}{2}}{2r+s}} \log T \, (\log k)^4 \left(\log \frac{12}{\delta}\right)^4 \\ &\leq \frac{2^{10} (M_\rho+1) \Big[\Big(6+\sqrt{5C_0}\Big)^2 + 1\Big]^2}{(2r+s)(1-\theta)} k N^{-\frac{2r-\frac{1}{2}}{2r+s}} (\log N)^5 \left(\log \frac{12}{\delta}\right)^4 \\ &\leq \frac{2^{10} (M_\rho+1) \Big[\Big(6+\sqrt{5C_0}\Big)^2 + 1\Big]^2}{(2r+s)(1-\theta)} N^{-\frac{r}{2r+s}} \left(\log \frac{12}{\delta}\right)^4 \qquad (49)\end{aligned}$$

and

$$h^{-2p} T^{(1-\theta)(p+\frac{3}{2})} \left(1 + (\log T) \sup_{1 \leq l \leq k} \mathcal{C}_{D_l,\lambda}\right)$$

$$
\begin{aligned}
&\leq 2h^{-2p}T^{(1-\theta)(p+\frac{3}{2})}\left(1+(\log T)\mathcal{A}_{D_l,\lambda}\log\frac{12k}{\delta}\right)\\
&\leq 2h^{-2p}N^{\frac{p+\frac{3}{2}}{2r+s}}\left(1+\frac{12+2\sqrt{5C_0}}{(2r+s)(1-\theta)}(\log N)k^{\frac{1}{2}}N^{-\frac{r}{2r+s}}\log k\log\frac{12}{\delta}\right)\\
&\leq 2h^{-2p}N^{\frac{p+\frac{3}{2}}{2r+s}}\left(1+\frac{12+2\sqrt{5C_0}}{(2r+s)(1-\theta)}k^{\frac{1}{2}}N^{-\frac{r}{2r+s}}(\log N)^2\log\frac{12}{\delta}\right)\\
&\leq 2h^{-2p}N^{\frac{p+\frac{3}{2}}{2r+s}}\left(1+\frac{12+2\sqrt{5C_0}}{(2r+s)(1-\theta)}\right)\log\frac{12}{\delta}. \qquad (50)
\end{aligned}
$$

By Lemma 3.3, we have with confidence at least $1-\frac{\delta}{4}$,

$$
\mathcal{D}_{D,\lambda}\leq 2\mathcal{A}_{D,\lambda,k}\log\frac{8}{\delta}\leq 2\sqrt{5}(\sqrt{C_0}+2)N^{-\frac{r}{2r+s}}\log\frac{12}{\delta}. \qquad (51)
$$

By Lemma 3.5 with $g(z,z')=y-y'$ we have with confidence at least $1-\frac{\delta}{4}$,

$$
\mathcal{F}_{D,\lambda}\leq 4M_\rho\mathcal{A}_{D,\lambda,k}\log\frac{8}{\delta}\leq 4\sqrt{5}M_\rho(\sqrt{C_0}+2)N^{-\frac{r}{2r+s}}\log\frac{12}{\delta}. \qquad (52)
$$

Plugging the estimates (49), (50), (51)) and (52) into (44), we obtain with confidence at least $1-\delta$,

$$
\|\overline{\widetilde{f}_{T+1,D}}-\widetilde{f}_{T+1}\|\leq C\left(N^{-\frac{r}{2r+s}}+h^{-2p}N^{\frac{p+\frac{3}{2}}{2r+s}}\right)\left(\log\frac{12}{\delta}\right)^4
$$

where

$$
\begin{aligned}
C = C''_{r,\theta,p}\Bigg[&4M_\rho(5^{\frac{r}{2r+s}})(\sqrt{C_0}+2)+2\sqrt{5}(\sqrt{C_0}+2)\\
&+\frac{2^{10}(M_\rho+1)\left[\left(6+\sqrt{5C_0}\right)^2+1\right]^2}{(2r+s)(1-\theta)}+2\left(1+\frac{12+2\sqrt{5C_0}}{(2r+s)(1-\theta)}\right)\Bigg].
\end{aligned}
$$

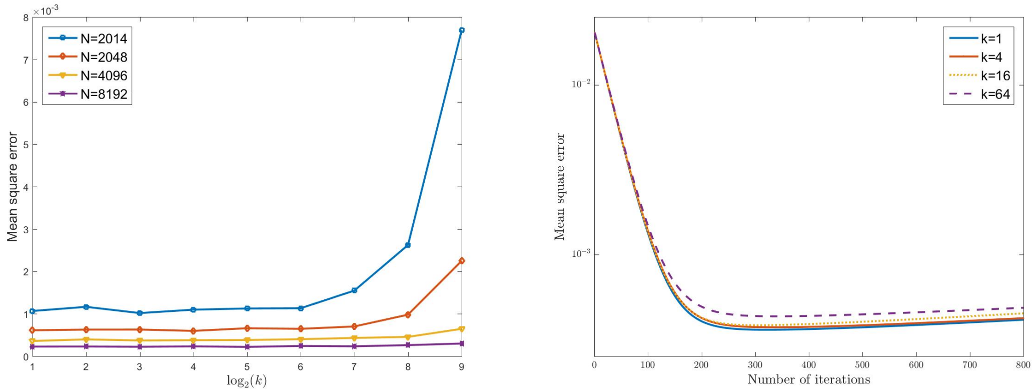
This together with the bound

$$
\|\widetilde{f}_{T+1}-\widetilde{f}_\rho\|\leq h_\rho T^{-r(1-\theta)}\leq\sqrt{5}h_\rho N^{-\frac{r}{2r+s}}.
$$

leads to the desired conclusion with $C^*=C+\sqrt{5}h_\rho$. □

6 Simulations

We further discuss and demonstrate our theory by an illustrative example. Consider the model $f^*(x)=\min\{x,1-x\}$ with $x\sim\text{Uniform}[0,1]$ and the noise $\epsilon\sim N(0,\sigma^2)$ with $\sigma^2=\frac{1}{5}$. This model was used to illustrate distributed kernel ridge regression in [33]. Let $K(x,t)=1+\min\{x,t\}$. Then $f^*\in\mathcal{H}_K$ and $\|f^*\|_K=1$. We apply the distributed kernel gradient descent MEE with $N\in\{1024,2048,4096,8192\}$ and $k\in\{1,2,4,8,16,32,64,128,256,512\}$. Note that $m=1$ corresponds to the kernel gradient descent MEE without using distributed techniques. It serves as a baseline for the comparison. For each value N and each value k we run the experiment 20 times. The mean square errors with an optimal number T of iterations are plotted in Figure 1 (a). We see when $k\leq 32$ the distributed methods are comparable with the baseline. When $k\geq 64$, the performance of distributed method for $N=1024$ and $N=2048$ becomes poor. But the performance for $N=4096$ and $N=8192$ is still good. This coincides with our theory that the distributed kernel gradient



(a) (b)

Figure 1: (a) Optimal mean square errors for $N \in \{1024, 2048, 4096, 8192\}$ as k varies. (b) Mean square errors for $N = 4096$ and $k \in \{1, 4, 16, 64\}$ as T varies.

descent MEE is asymptotically rate equivalent to learning with the whole data set directly if k does not increase too fast as a function of N.

We next discuss how the performance changes as the number of iterations increases. We plot in Figure 1 (b) the mean square errors for $N = 4096$ and different values of k when T varies. We see that the optimal T for different k values are quite similar. This phenomenon is also observed for other N values. It indicates that k is irrelevant to the choice of T and is consistent with our theory that the choice of T depends on the size of the whole data set, not on the size of each subset.

7 Conclusions and Discussions

We have studied the convergence of a distributed kernel gradient descent MEE algorithm. We first derived error bounds of the kernel gradient descent MEE algorithm for the single data case. Then, by the aid of a concentration inequality for distributed U-statistics, we derived error bounds and minimax optimal rates for the distributed case under certain regularity condition for the target regression function and capacity condition for the reproducing kernel. Our analysis shows that the error bound for the distributed case is larger due to additional error caused by loss of information from interactions between samples of different subsets. But this additional error is of the same magnitude as the error for single data set case if the parameters are selected appropriately. As a consequence, the distributed kernel gradient MEE algorithm has the same minimax optimal rates as the single data set case. This phenomenon has also been observed for other distributed learning algorithms.

Several related questions are worthwhile for future research. First, our analysis provides very useful insights on the application of distributed kernel gradient MEE algorithms. It tells that the number of iterations should be selected according to the total sample size instead of the sample size on each local machine. The simulation is also consistent to our theory. However, the optimal parameters should be selected according to capacity of the kernel space and the regularity of the target function which are unknown in practice. It is necessary to develop empirically applicable parameter selection strategies for optimal performance.

Second, non-convexity of the MEE loss function is one of the major challenges we need to overcome in the error analysis. In the literature the study of MEE algorithms has focused on the use of Gaussian kernel density estimators. For this special situation, it was proved in [27] that the loss function is invex and may lead to better optimization properties than general nonconvex loss functions. In this paper we allowed a general density estimator to be used in the distributed kernel gradient MEE algorithm. Invexity is not necessarily true. Furthermore, our error bounds already imply minimax optimal rates. It is yet unknown to us whether invexity can be incorporated in the error analysis and if yes, it is interesting to know what benefits it can bring in.

Third, our distributed kernel gradient MEE algorithms are based on the divide and conquer approach. No information communications are needed between local machines. This not only makes the algorithm easy to implement but also is particularly important to the scenarios where data have to be stored and analyzed locally for privacy concerns. Recently, averaging stochastic gradient descent algorithms were developed in the context of deep machine learning where mutual information communications are used between local machines and a master machine to speed up the convergence [32, 2]. The techniques may also be applied to the distributed MEE for better performance if data privacy is not a concern.

Last, it is worth mentioning that a large bandwidth parameter is necessary for our convergence analysis. In practice, however, a moderate choice of the bandwidth parameter may be sufficient for the algorithm to have tolerable small error. But a too small bandwidth parameter may lead to theoretical inconsistency. In [10] a counterexample is given to show that mean regression function is not necessarily the minimizer of the Renyi quadratic entropy and hence MEE fails to converge to the mean regression function if $h \to 0$. In a recent work [11] learning with correntropy loss was related to modal regression. It would be interesting to investigate whether MEE with $h \to 0$ can also be interpreted by modal regression. This is beyond the scope of this paper and will be left for our future research.

Acknowledgement

The work is partially supported by and National Natural Science Foundation of China [Grant Numbers 11671307, 11501078, 11671171, 11461161006 and 11471292] and the Research Grants Council of Hong Kong [Project No. CityU 11338616].

Appendix

We now prove Lemma 3.1 and Lemma 4.1.

Proof of Lemma 3.1. For any two elements $f, g \in \mathcal{H}$, define a function H on $[0,1]$ by $H(t) := \|f + tg\|$. As a function of $t \in [0,1]$, H is differentiable, $|H'(t)| \leq \|g\|$ and $(H^2(t))'' \leq 2\|g\|^2$, and $(\cosh H)'' = (H')^2 \cosh H + H'' \sinh H$. If $H'' > 0$, then we see that

$$(\cosh H)'' = (H')^2 \cosh H + H'' \sinh H \leq ((H')^2 + H''H) \cosh H = \frac{1}{2}(H^2)'' \cosh H \leq \|g\|^2 \cosh H.$$

If $H'' \leq 0$, then $(\cosh H)'' \leq (H')^2 \cosh H \leq \|g\|^2 \cosh H$. So, the inequality $(\cosh H)'' \leq \|g\|^2 \cosh H$ always holds.

Let $W(t) = \mathbb{E}_{j-1} \cosh(H(t))$ with $c > 0$, $f = c \sum_{k=1}^{j-1} d_k$ and $g = cd_j$, that is,

$$W(t) = \mathbb{E}_{j-1} \cosh\left(c \left\| \sum_{k=1}^{j-1} d_k + td_j \right\|\right), \quad t \in \mathbb{R}.$$

We have for all $t \in [0,1]$,

$$W''(t) \leq c^2 \mathbb{E}_{j-1} \|d_j\|^2 \cosh\left(c \left\| \sum_{k=1}^{j-1} d_k + td_j \right\|\right)$$

$$\leq c^2\mathbb{E}_{j-1}\|d_j\|^2\cosh\left(c\left\|\sum_{k=1}^{j-1}d_k\right\|+ct\|d_j\|\right)$$
$$\leq c^2\mathbb{E}_{j-1}\|d_j\|^2e^{ct\|d_j\|}\cosh\left(c\left\|\sum_{k=1}^{j-1}d_k\right\|\right),$$

where the last inequality follows from the the elementary inequality $\cosh(a+b)\leq e^a\cosh(b)$ for all a, $b>0$. Since $\{d_i\}$ is a sequence of martingale differences, then $\mathbb{E}_{j-1}d_j=0$ and $W'(0)=0$. Therefore,

$$\begin{aligned}\mathbb{E}_{j-1}\cosh\left(c\left\|\sum_{k=1}^{j}d_k\right\|\right)&=W(1)=W(0)+\int_0^1(1-t)W''(t)dt\\&=\cosh\left(c\left\|\sum_{k=1}^{j-1}d_k\right\|\right)+\int_0^1(1-t)W''(t)dt\\&\leq(1+e_j)\cosh\left(c\left\|\sum_{k=1}^{j-1}d_k\right\|\right),\end{aligned}\tag{53}$$

where $e_j=\mathbb{E}_{j-1}(e^{c\|d_j\|}-1-c\|d_j\|)$.

Define a sequence $\{G_j\}_{j=0}^N$ with $G_0=1$ and

$$G_j=\frac{\cosh\left(c\|\sum_{k=1}^j d_k\|\right)}{\prod_{i=1}^j(1+e_i)},\quad j\geq 1.$$

The inequality (53) implies that $\{G_j\}_{j=0}^N$ is a positive supermartingale and

$$\mathbb{E}G_N=\mathbb{E}\left[\frac{\cosh\left(c\|\sum_{k=1}^N d_k\|\right)}{\prod_{i=1}^N(1+e_i)}\right]\leq 1=\mathbb{E}G_0.\tag{54}$$

For each e_j, by the bound $\sup\limits_{1\leq j\leq N}\|d_j\|\leq M$ and Taylor expansion,

$$\begin{aligned}e_j=&\mathbb{E}_{j-1}\left(e^{c\|d_j\|}-1-c\|d_j\|\right)=1+c\mathbb{E}_{j-1}\|d_j\|+\sum_{l=2}^{\infty}\frac{c^l\mathbb{E}_{j-1}(\|d_j\|^l)}{l!}-1-c\mathbb{E}_{j-1}\|d_j\|\\&\leq\sum_{l=2}^{\infty}\frac{c^lM^{l-2}\mathbb{E}_{j-1}(\|d_j\|^2)}{l!}\leq\frac{\mathbb{E}_{j-1}(\|d_j\|^2)}{M^2}(e^{cM}-1-cM).\end{aligned}$$

This together with the assumption $\sum_{j=1}^N\mathbb{E}_{j-1}\|d_j\|^2\leq\sigma^2$ gives

$$\prod_{j=1}^N(1+e_j)\leq\exp\left(\sum_{j=1}^N e_j\right)\leq\exp\left\{\frac{\sigma^2}{M^2}\left(e^{cM}-1-cM\right)\right\}.$$

Plugging it into (54), we obtain

$$\mathbb{E}\cosh\left(c\left\|\sum_{j=1}^N d_j\right\|\right)\leq\exp\left\{\frac{\sigma^2}{M^2}\left(e^{cM}-1-cM\right)\right\}.$$

This prove the first part of the lemma.

Since $\cosh t \geq \frac{e^t}{2}$, for any $c > 0$ and $\varepsilon > 0$, we have that

$$\frac{\mathbb{E}\cosh\left(c\|\sum_{j=1}^{N} d_j\|\right)}{\cosh(c\varepsilon)} \leq 2\exp\left\{-c\varepsilon + \frac{\sigma^2}{M^2}\left(e^{cM} - 1 - cM\right)\right\}.$$

Taking $c = \frac{1}{M}\log(1+\frac{M\varepsilon}{\sigma^2})$, the minimizer of the bound on the right hand side, we get the first desired inequality in (9). The second one can be deduced easily by some basic mathematical analysis, which can be found in some classical books, see e.g. [6]. □

To prove Lemma 4.1 we need the following lemma whose proof is trivial and hence is omitted.

Lemma 7.1. *For any* $0 \leq \theta < 1$ *and* $j \geq 1$,

$$\frac{(T+1)^{1-\theta} - j^{1-\theta}}{1-\theta} \leq \sum_{t=j}^{T} t^{-\theta} \leq \frac{T^{1-\theta} - (j-1)^{1-\theta}}{1-\theta} \leq \frac{T^{1-\theta}}{1-\theta}.$$

For $\theta = 1$, *if* $T \geq 3$, *then*

$$\sum_{t=1}^{T} t^{-1} \leq 2\log T.$$

For $\theta > 1$,

$$\sum_{t=1}^{T} t^{-\theta} \leq \frac{\theta}{\theta - 1}.$$

Proof of Lemma 4.1. We decompose the summation on the left of (16) into three parts,

$$\Upsilon_1 = \sum_{1 \leq i < \frac{T}{2}} i^{-(\theta+\alpha)}\left(\sum_{j=i+1}^{T} j^{-\theta}\right)^{-1},$$

$$\Upsilon_2 = \sum_{\frac{T}{2} \leq i < T-1} i^{-(\theta+\alpha)}\left(\sum_{j=i+1}^{T} j^{-\theta}\right)^{-1},$$

$$\Upsilon_3 = \sum_{i=T-1}^{T} i^{-(\theta+\alpha)}\left(\sum_{j=i+1}^{T} j^{-\theta}\right)^{-1}.$$

For Υ_1, when $\alpha + \theta < 1$, by Lemma 7.1, we obtain

$$\begin{aligned}\Upsilon_1 &\leq (1-\theta)\sum_{1 \leq i < \frac{T}{2}} i^{-(\theta+\alpha)}[(T+1)^{1-\theta} - (i+1)^{1-\theta}]^{-1} \\ &\leq [1 - 2^{-(1-\theta)}]^{-1}(1-\theta)T^{-(1-\theta)}\sum_{1 \leq i < \frac{T}{2}} i^{-(\theta+\alpha)} \\ &\leq \frac{2^{-\alpha}[1 - 2^{-(1-\theta)}]^{-1}}{1-\alpha-\theta}T^{-\alpha}.\end{aligned}$$

When $\alpha + \theta = 1$,

$$\Upsilon_1 \leq [1 - 2^{-(1-\theta)}]^{-1}(1-\theta)T^{-(1-\theta)}\sum_{1 \leq i < \frac{T}{2}} i^{-1} \leq 2[1 - 2^{-(1-\theta)}]^{-1}(1-\theta)T^{-\alpha}\log T.$$

When $\alpha + \theta > 1$,

$$\Upsilon_1 \leq [1 - 2^{-(1-\theta)}]^{-1}(1-\theta)T^{-(1-\theta)} \sum_{1 \leq i < \frac{T}{2}} i^{-(\theta+\alpha)} \leq \frac{[1 - 2^{-(1-\theta)}]^{-1}(1-\theta)(\alpha+\theta)}{\theta+\alpha-1} T^{-(1-\theta)}$$

Thus, $\Upsilon_1 \leq C'_{\theta,\alpha} T^{-\min\{1-\theta,\alpha\}} \log T$ with

$$C'_{\theta,\alpha} = \begin{cases} \frac{2^{-\alpha}[1-2^{-(1-\theta)}]^{-1}(1-\theta)}{1-\alpha-\theta}, & \text{if} \quad \alpha+\theta < 1, \\ 2[1 - 2^{-(1-\theta)}]^{-1}(1-\theta), & \text{if} \quad \alpha+\theta = 1, \\ \frac{[1-2^{-(1-\theta)}]^{-1}(1-\theta)(\alpha+\theta)}{\theta+\alpha-1}, & \text{if} \quad \alpha+\theta > 1. \end{cases}$$

For Υ_2, by Lemma 7.1 again,

$$\begin{aligned} \Upsilon_2 &\leq (1-\theta)2^{\alpha}T^{-\alpha} \sum_{\frac{T}{2} \leq i < T-1} i^{-\theta}[(T+1)^{1-\theta} - (i+1)^{1-\theta}]^{-1} \\ &= (1-\theta)2^{\alpha}T^{-\alpha} \sum_{\frac{T}{2} \leq i < T-1} \int_{i-1}^{i} i^{-\theta}[(T+1)^{1-\theta} - (i+1)^{1-\theta}]^{-1} dx \\ &\leq (1-\theta)2^{\alpha}T^{-\alpha} \sum_{\frac{T}{2} \leq i < T-1} \int_{i-1}^{i} x^{-\theta}[(T+1)^{1-\theta} - (x+2)^{1-\theta}]^{-1} dx \\ &\leq 3^{\theta}(1-\theta)2^{\alpha}T^{-\alpha} \int_{\frac{T}{2}-1}^{T-2} (x+2)^{-\theta}[(T+1)^{1-\theta} - (x+2)^{1-\theta}]^{-1} dx \\ &= 3^{\theta}(1-\theta)2^{\alpha}T^{-\alpha} \int_{\frac{T}{2}-1}^{T-2} [(T+1)^{1-\theta} - (x+2)^{1-\theta}]^{-1} d(x+2)^{1-\theta} \\ &\leq 3^{\theta}(1-\theta)2^{\alpha+1}T^{-\alpha} \log[(T+1)^{1-\theta} - (\frac{T}{2}+1)^{1-\theta}] \\ &\leq 3^{\theta}2^{\alpha+2}(1-\theta)^2 T^{-\alpha} \log T. \end{aligned}$$

For Υ_3, it is easy to check that

$$\Upsilon_3 = (T-1)^{-(\alpha+\theta)}T^{-\alpha} + T^{-(\alpha+\theta)} \leq T^{-\alpha}.$$

Combining the above bounds of Υ_1, Υ_2 and Υ_3, we get the desired conclusion (16). The inequality (17) can be deduced from (16) easily. □

References

[1] A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[2] P. Chaudhari, A. Oberman, S. Osher, S. Soatto, and G. Carlier. Deep relaxation: partial differential equations for optimizing deep neural networks. *Research in the Mathematical Sciences*, 5(3):30, 2018.

[3] B. Chen and J. C. Principe. Stochastic gradient algorithm under (h, ϕ)-entropy criterion. *Circuits, Systems, and Signal Processing*, 26(6):941–960, 2007.

[4] B. Chen, P. Zhu, and J. C. Principe. Survival information potential: a new criterion for adaptive system training. *IEEE Transactions on Signal Processing*, 60(3):1184–1194, 2012.

[5] B. Chen, Y. Zhu, and J. Hu. Mean-square convergence analysis of adaline training with minimum error entropy criterion. *IEEE Transactions on Neural Networks*, 21(7):1168–1179, 2010.

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

[7] D. Erdogmus, K. Hild II, and J. C. Principe. Blind source separation using Rényi's α-marginal entropies. *Neurocomputing*, 49:25–38, 2002.

[8] D. Erdogmus and J. C. Principe. Comparison of entropy and mean square error criteria in adaptive system training using higher order statistics. In *Proceedings of the Intl. Conf. on ICA and Signal Separation*, pages 75–90. Berlin: Springer-Verlag, 2000.

[9] D. Erdogmus and J. C. Principe. Convergence properties and data efficiency of the minimum error entropy criterion in adaline training. *IEEE Transactions on Signal Processing*, 51:1966–1978, 2003.

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

[11] Y. Feng, J. Fan, and J. A. Suykens. A statistical learning approach to modal regression. *arXiv preprint arXiv:1702.05960*, 2017.

[12] E. Gokcay and J. C. Principe. Information theoretic clustering. *IEEE Transactions on Pattern Analysis and Machine Learning*, 24(2):158–171, 2002.

[13] Z.-C. Guo, S.-B. Lin, and D.-X. Zhou. Learning theory of distributed spectral algorithms. *Inverse Problems*, 33(7):074009 (29 pages), 2017.

[14] Z.-C. Guo, L. Shi, and Q. Wu. Learning theory of distributed regression with bias corrected regularization kernel network. *Journal of Machine Learning Research*, 18(118):1–25, 2017.

[15] Z.-C. Guo, D.-H. Xiang, X. Guo, and D.-X. Zhou. Thresholded spectral algorithms for sparse approximations. *Analysis and Applications*, 15(3):433–455, 2017.

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

[17] T. Hu, J. Fan, Q. Wu, and D.-X. Zhou. Regularization schemes for minimum error entropy principle. *Analysis and Applications*, 13(4):437–455, 2015.

[18] T. Hu, Q. Wu, and D.-X. Zhou. Convergence of gradient descent method for minimum error entropy principle in linear regression. *IEEE Transactions on Signal Processing*, 64(24):6571–6579, 2016.

[19] T. Hu, Q. Wu, and D.-X. Zhou. Kernel gradient descent algorithm for information theoretic learning. 2016. preprint.

[20] S.-B. Lin, X. Guo, and D.-X. Zhou. Distributed learning with regularized least squares. *Journal of Machine Learning Research*, 18(92):1–31, 2017.

[21] S.-B. Lin and D.-X. Zhou. Distributed kernel gradient descent algorithms. *Constructive Approximation*, 47:249–276, 2018.

[22] I. Pinelis. Optimum bounds for the distributions of martingales in banach space. *Ann. Probab.*, 22(4):1679–1706, 1994.

[23] J. D. Rosenblatt and B. Nadler. On the optimality of averaging in distributed statistical learning. *Information and Inference: A Journal of the IMA*, 5(4):379–404, 2016.

[24] P. Shen and C. Li. Minimum total error entropy method for parameter estimation. *IEEE Transactions on Signal Processing*, 63(15):4079–4090, 2015.

[25] L. M. Silva, J. Marques de Sá, and L. A. Alexandre. Neural network classification using Shannon's entropy. In *Proceedings of the European Symposium on Artificial Neural Networks*, pages 217–222. Bruges: d-side, 2005.

[26] L. M. Silva, J. Marques de Sá, and L. A. Alexandre. The MEE principle in data classification: A perceptron-based analysis. *Neural Computation*, 22:2698–2728, 2010.

[27] M. Syed, P. Pardalos, and J. Principe. Invexity of the minimum error entropy criterion. *IEEE Signal Processing Letters*, 20(12):1159–1162, 2013.

[28] Z. Wu, S. Peng, W. Ma, B. Chen, and J. C. Principe. Minimum error entropy algorithms with sparsity penalty constraints. *Entropy*, 17(5):3419–3437, 2015.

[29] Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

[30] Y. Ying and D.-X. Zhou. Online pairwise learning algorithms. *Neural Computation*, 28(4):743–777, 2016.

[31] Y. Ying and D.-X. Zhou. Unregularized online learning algorithms with general loss functions. *Applied and Computational Harmonic Analysis*, 42(2):224–244, 2017.

[32] S. Zhang, A. E. Choromanska, and Y. LeCun. Deep learning with elastic averaging sgd. In *Advances in Neural Information Processing Systems*, pages 685–693, 2015.

[33] Y. Zhang, J. C. Duchi, and M. J. Wainwright. Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research*, 16:3299–3340, 2015.