entropy

MDPI

Article

Semi-Supervised Minimum Error Entropy Principle with Distributed Method

Baobin Wang [1] and Ting Hu [2,*]

[1] School of Mathematics and Statistics, South-Central University for Nationalities, Wuhan 430074, China; wbb1818@126.com

[2] School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China

* Correspondence: tinghu@whu.edu.cn

Received: 26 October 2018; Accepted: 10 December 2018; Published: 14 December 2018

Abstract: The minimum error entropy principle (MEE) is an alternative of the classical least squares for its robustness to non-Gaussian noise. This paper studies the gradient descent algorithm for MEE with a semi-supervised approach and distributed method, and shows that using the additional information of unlabeled data can enhance the learning ability of the distributed MEE algorithm. Our result proves that the mean squared error of the distributed gradient descent MEE algorithm can be minimax optimal for regression if the number of local machines increases polynomially as the total datasize.

Keywords: information theoretical learning; distributed method; MEE algorithm; semi-supervised approach; gradient descent; reproducing kernel Hilbert spaces

1. Introduction

The minimum error entropy (MEE) principle is an important criterion proposed in information theoretical learning (ITL) [1] and was firstly addressed for adaptive system training by Erdogmus and Principe [2]. It has been applied to blind source separation, maximally informative subspace projections, clustering, feature selection, blind deconvolution, minimum cross-entropy for model selection, and some other topics [3–8]. Taking entropy as a measure of the error, the MEE principle can extract the information contained in data fully and produce robustness to outliers in the implementation of algorithms.

Let $X \in \mathcal{R}^n$ be an explanatory variable with values taken in a compact metric space $(\mathcal{X}, d)$, Y be a real response variable with $Y \in \mathcal{Y} \subset \mathcal{R}$, and $g : \mathcal{X} \to \mathcal{Y}$ be a prediction function. For a given set of labeled examples $D = \{(x_i, y_i)\}_{i=1}^N \subset \mathcal{X} \times \mathcal{Y}$ (N denotes the sample size) and a windowing function $G : \mathbb{R} \to \mathbb{R}_+$, the MEE principle is to find a minimizer of the empirical quadratic entropy:

$$\hat{H}(g) = -\log\left\{\frac{h^2}{N^2}\sum_{\substack{(x_i,y_i)\in D\\(x_j,y_j)\in D}} G\left(\frac{[(y_i - g(x_i)) - (y_j - g(x_j))]^2}{h^2}\right)\right\},$$

where $h > 0$ is the scaling parameter. Its goal is to solve the problem $y = g_\rho(x) + \varepsilon$, where ε is the noise and $g_\rho(x)$ is the target function. Taking a function $f(x_i, x_j) := g(x_i) - g(x_j)$, MEE belongs to pairwise learning problems, which involves with the intersections of example pairs. Since logarithmic function is monotonic, we only consider the empirical information error of MEE:

$$R(f) = -\frac{h^2}{N^2}\sum_{\substack{(x_i,y_i)\in D\\(x_j,y_j)\in D}} G\left(\frac{[y_i - y_j - f(x_i, x_j)]^2}{h^2}\right), \tag{1}$$

in the optimization process. Borrowing the idea from Reference [9], we introduced the Mercer kernel $K(\cdot,\cdot) : \mathcal{X}^2 \times \mathcal{X}^2 \to \mathbb{R}, (\mathcal{X}^2 := \mathcal{X} \times \mathcal{X})$ and employed the reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ as our hypothesis space. With K, $\mathcal{H}_K$ is defined as the linear span of the functions set $\{K_{(x,u)} := K\left((x,u),(\cdot,\cdot)\right), \forall (x,u) \in \mathcal{X}^2\}$, which is equipped with the inner product $\langle \cdot,\cdot \rangle_K$ and the reproducing property $\langle K_{(x,u)}, K_{(x',u')} \rangle_K = K((x,u),(x',u')), \forall (x,u),(x',u') \in \mathcal{X}^2$. For the G nonconvex, we usually solve Equation (1) using the kernel-based gradient descent method as follows. It starts with $f_{1,D} = 0$ and is updated by:

$$f_{t+1,D} = f_{t,D} - \eta \times \nabla\mathcal{R}(f_{t,D}), \tag{2}$$

in the t-th step, where $\eta > 0$ is a step size, ∇ is the gradient operator and:

$$\nabla\mathcal{R}(f_{t,D}) = -\frac{1}{N^2} \sum_{\substack{(x_i,y_i)\in D \\ (x_j,y_j)\in D}} G'\left(\frac{[y_i - y_j - f_{t,D}(x_i,x_j)]^2}{h^2}\right)[f_{t,D}(x_i,x_j) - y_i + y_j]K_{(x_i,x_j)},$$

as we know that the example pairs will grow quadratically with the increasing example size N, which will bring the computational burden in the MEE implementation. Thus, it is necessary to reduce the algorithmic complexity by the distributed method based on a divide-and-conquer strategy [10]. Semi-supervised learning (SSL) [11] has attracted extensive attention as an emerging field in machine learning research and data mining. Actually, in many practical problems, few data are given, but a large number of unlabeled data are available, since labeling data requires a lot of time, effort or money. In this paper, we study a distributed MEE algorithm in the framework of SSL and show that the learning ability of the MEE algorithm can be enhanced by the distributed method and the combination of labeled data with unlabeled data.

There are mainly three contributions in this paper. The first one is that we derive the explicit learning rate of the gradient descent method for distributed MEE in the context of SSL, which is comparable to the minimax optimal rate of the least squares in regression. This implies that the MEE algorithm can be an alternative of the least squares in SSL in the sense that both of them have the same prediction power. The second one is that we provide the theoretical upper bound for the number of local machines guaranteeing the optimal rate in the distributed computation. The last one is that we extend the range of the target function allowed in the distributed MEE algorithm.

In Table 1, we summarize some notations used in this paper.

Table 1. List of notations used throughout the paper.

Notation	Meaning of the Notation
X	the explanatory variable
Y	the response variable
$\mathcal{X}$	$X \in \mathcal{X}$, a compact subset of an Euclidian space $\mathbb{R}^n$
$\mathcal{Y}$	$Y \in \mathcal{Y}$, a subset of $\mathbb{R}$
$\rho(\cdot,\cdot)$	a Boreal measure on $\mathcal{X} \times \mathcal{Y}$
$\rho_{\mathcal{X}}$	the marginal probability measure of ρ on $\mathcal{X}$
$\rho(y\|x)$	the conditional probability measure of $y \in \mathcal{Y}$ given $X = x$
$g_\rho(x)$	the mean regression function $g_\rho(x) = \int_{\mathcal{Y}} y d\rho(y\|x)$
$f_\rho(x,u)$	the target function of MEE induced by $f_\rho(x,u) = g_\rho(x) - g_\rho(u)$
K	a reproducing kernel on $\mathcal{X} \times \mathcal{X}$
D	the labeled data set $D = \{(x_1,y_1),\ldots,(x_N,y_N)\}$
N	the size of labeled data set D
$\lceil N/4 \rceil$	the largest integer not exceeding $N/4$
$\|D\|$	the cardinality of D, $\|D\| = N$
D^*	the unlabeled data set $D^* = \{x_1,\ldots,x_S\}$
S	the size of unlabeled data set D^*
$\|D^*\|$	the cardinality of D^*, $\|D^*\| = S$

Table 1. *Cont.*

Notation	Meaning of the Notation
$\tilde{D}$	training data set used in the distributed MEE algorithm, consisting of D and D^*
$\vert\tilde{D}\vert$	the cardinality of $\tilde{D}$, $\vert\tilde{D}\vert = N + S$
m	the number of local machines
$\tilde{D}_l$	the lth subset of $\tilde{D}$, $1 \le l \le m$
G	the loss function of MEE algorithm
L_K	the integral operator associated with K
$L_{K,\tilde{D}}$	the empirical operator of L_K on $\tilde{D}$
$f_{t+1,D}$	the function output by the kernel gradient descent MEE algorithm with data D and kernel K after t iterations
f_{t+1,D_l}	the function output by the kernel gradient MEE algorithm with data D_l and kernel K after t iterations
$\overline{f}_{t+1,\tilde{D}}$	the global output averaging over local outputs $f_{t+1,\tilde{D}_l}, l = 1, \ldots, m$

2. Algorithms and Main Results

We considered MEE for the regression problem. To allow noise in sampling processes, we assumed that a Borel measure $\rho(\cdot,\cdot)$ is defined on the product space $\mathcal{X} \times \mathcal{Y}$. Let $\rho(y|x)$ be the conditional distribution of $y \in \mathcal{Y}$ for any given $x \in \mathcal{X}$, and $\rho_{\mathcal{X}}(\cdot)$ the marginal distribution on $\mathcal{X}$. For the semi-supervised MEE algorithm, our goal was to estimate the regression function $g_\rho(x) = \int_{\mathcal{Y}} y d\rho(y|x), x \in \mathcal{X}$, from labeled examples $D = \{(x_i, y_i)\}_{i=1}^N$ and unlabeled examples $D^* = \{x_j\}_{j=1}^S$ drawn from the distribution ρ and $\rho_{\mathcal{X}}$, respectively.

Based on the divide-and-conquer strategy, both D and D^* are partitioned equally into m subsets, $D = \sum_{l=1}^m \bigcup D_l$ and $D^* = \sum_{l=1}^m \bigcup D_l^*$. Here, we denote the size of subsets $|D_l| = n$ and $|D_l^*| = s$, $1 \le l \le m$, i.e., $N = mn, S = ms$. We construct a new dataset $\tilde{D} = \sum_{l=1}^m \bigcup \tilde{D}_l$ by:

$$\tilde{D}_l = D_l \cup D_l^* = \{(x_k, y_k)\}_{k=1}^{n+s},$$

where:

$$x_k = \begin{cases} x_k, & \text{if } (x_k, y_k) \in D_l, \\ x_k, & \text{if } x_k \in D_l^*, \end{cases} \quad \text{and} \quad y_k = \begin{cases} \frac{n+s}{n} y_k, & \text{if } (x_k, y_k) \in D_l, \\ 0, & \text{if } x_k \in D_l^*. \end{cases}$$

Based on the gradient descent algorithm (Equation (2)), we can get a set of local estimators $\{f_{t,\tilde{D}_l}\}$ for each subset $\tilde{D}_l, 1 \le l \le m$. Then, the global estimator averaging over these local estimators is given by:

$$\overline{f}_{t,\tilde{D}} = \frac{1}{m} \sum_{l=1}^m f_{t,\tilde{D}_l}. \tag{3}$$

In the pairwise setting, our target function $f_\rho(x, x') = g_\rho(x) - g_\rho(x'), x, x' \in \mathcal{X}$, which is the difference of the regression function g_ρ. Denote by $L^2_{\rho_{\mathcal{X}^2}}$ the space of square integrable functions on the product space $\mathcal{X}^2$:

$$L^2_{\rho_{\mathcal{X}^2}} := \left\{ f : \mathcal{X}^2 \to \mathcal{R} : \|f\|_{L^2} = \left(\int \int_{\mathcal{X}^2} |f(x, x')|^2 d\rho_{\mathcal{X}}(x) d\rho_{\mathcal{X}}(x') \right)^{\frac{1}{2}} < \infty \right\}.$$

The goodness of $\overline{f}_{t,\tilde{D}}$ is usually measured by the mean squared error $\|\overline{f}_{t,\tilde{D}} - f_\rho\|^2_{L^2}$.

Throughout the paper, we assumed that $\sup_{(x,x') \in \mathcal{X}^2} \sqrt{K((x,x'),(x,x'))} \le 1$ and for some constant $M > 0$, $|y| \le M$ almost surely. Without generality, windowing function G is assumed to be differentiable and satisfies $G'(0) = -1$, $G'(u) < 0$ for $u > 0$, $C_G := \sup_{u \in (0,\infty)} |G'(u)| < \infty$ and there exists some p such that $c_p > 0$ and:

$$|G'(u) - G'(0)| \leq c_p |u|^p, \ \forall u > 0. \tag{4}$$

It is easy to check that the Gaussian kernel $G(u) = \exp\{-u\}$ satisfies the assumptions above with $p = 1$.

Before we present our main results, define an integral operator $L_K : L^2_{\rho_{\mathcal{X}^2}} \longrightarrow L^2_{\rho_{\mathcal{X}^2}}$ associated with the kernel K by:

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

Our error analysis for the distributed MEE algorithm (Equation (3)) is stated in terms of the following regularity condition:

$$f_\rho = L_K^r(\phi) \quad for \quad some \quad r > 0, \quad \phi \in L^2_{\rho_{\mathcal{X}^2}}, \tag{5}$$

where L_K^r denotes the r-th power of L_K on $L^2_{\rho_{\mathcal{X}^2}}$ and is well defined, since the operator L_K is positive and compact with the Mercer kernel K. We use the effective dimension [12,13] $\mathcal{N}(\lambda)$ to measure the complexity of $\mathcal{H}_K$ with respect to $\rho_{\mathcal{X}}$, which is defined to be the trace of the operator $(\lambda I + L_K)^{-1} L_K$ as:

$$\mathcal{N}(\lambda) = Tr((\lambda I + L_K)^{-1} L_K), \lambda > 0.$$

To obtain optimal learning rates, we need to quantify $\mathcal{N}(\lambda)$ of $\mathcal{H}_K$. A suitable assumption is: that

$$\mathcal{N}(\lambda) \leq C_0 \lambda^{-\beta}, for\ some\ C_0 > 0\ and\ 0 < \beta \leq 1. \tag{6}$$

Remark 1. *When $\beta = 1$, Equation (6) always holds with $C_0 = Tr(L_K)$. For $0 < \beta < 1$, when $\mathcal{H}_K$ is a Sobolev space $W^\alpha(\mathcal{X})$ on $\mathcal{X} \subset \mathcal{R}^d$ with all derivative of order up to $\alpha > \frac{d}{2}$, then Equation (6) is satisfied with $\beta = \frac{d}{2\alpha}$ [14]. Moreover, if the eigenvalues $\{\gamma_i\}_{i=1}^{\infty}$ of the operator L_K decays as $\gamma_i = O(i^{-b})$ for some $b > 1$, then $\mathcal{N}(\lambda) = O(\lambda^{-\frac{1}{b}})$. The eigenvalues assumption is typical in the analysis of the performances of kernel methods estimators and recently used in References [13,15,16] to establish the optimal learning rate in the least square problems.*

The following theorem shows that the distributed gradient descent algorithm (Equation (3)) can achieve the optimal rate by providing the iteration time T and the maximal number of local machines, whose proof can be found in Section 3.

Theorem 1. (**Main Result**) *Assume Equations (5) and (6) hold for $r + \beta \geq \frac{1}{2}$. Let the iteration time $T = \lceil N/4 \rceil^{\frac{1}{2r+\beta}}$ and $S + N \geq N^{\frac{\beta+1}{2r+\beta}}$:*

$$m < \frac{\min\{(N+S)^{\frac{1}{2}} N^{-\frac{\beta+1}{4r+2\beta}}, (N+S)^{\frac{1}{3}} N^{-\frac{2-2r-\beta}{6r+3\beta}}\}}{\log^6 N}, \tag{7}$$

then for any $0 < \delta < 1$, with confidence at least $1 - \delta$:

$$\|\overline{f}_{T+1,\tilde{D}} - f_\rho\|_{L^2} \leq C' \max\left\{N^{-\frac{r}{2r+\beta}}, h^{-2p}(N+S)^{2p+1} N^{\frac{p+\frac{3}{2}}{2r+\beta} - (2p+1)}\right\} \log^4 \frac{24}{\delta}, \tag{8}$$

where C' is a constant independent of N, S, δ, h and $\lceil N/4 \rceil$ denotes the largest number not exceeding $N/4$.

Corollary 1. *Under the same conditions of Theorem 1, if the scaling parameter:*

$$h > (N+S)^{\frac{2p+1}{2p}} N^{\frac{r+p+\frac{3}{2}}{2p(2r+\beta)}} N^{-\frac{2p+1}{2p}},$$

then for any $0 < \delta < 1$, *with confidence at least* $1 - \delta$:

$$\|\overline{f}_{T+1,\tilde{D}} - f_\rho\|_{L^2} \le C' N^{-\frac{r}{2r+\beta}} \log^4 \frac{24}{\delta}. \tag{9}$$

Remark 2. *The rate* $O\left(N^{-\frac{r}{2r+\beta}}\right)$ *in Equation (9) is optimal in the minimax sense for kernel regression problems [13]. When* $m = 1$, *the result of Equation (9) shows that the kernel gradient descent MEE algorithm (Equation (2)) on a single big data set can achieve the minimax optimal rate for regression. Thus, MEE is a nice alternative of the classical least squares. Meanwhile, the upper bound (Equation (7)) for the number of local machines implies that the performance of the distributed MEE algorithm (Equation (3)) can be as good as the standard MEE algorithm (2) (acting on the whole data set* $\tilde{D}$*), provided that the subset* $\tilde{D}_l$*'s size* $n + s$ *is not too small.*

Remark 3. *If no unlabeled data is engaged in the algorithm (Equation (3)), then* $S = 0$ *and the upper bound (Equation (7)) for the number of local machines m that ensures the optimal rate is about* $O\left(N^{\frac{r-\frac{1}{2}}{2r+\beta}}\right)$. *So, when the regularity parameter r in Equation (5) is close to* $\frac{1}{2}$, *the upper bound* $O\left(N^{\frac{r-\frac{1}{2}}{2r+\beta}}\right)$ *reduces to a constant and then the distributed algorithm (Equation (3)) will not be feasible in real applications. A similar phenomenon is observed in various distributed algorithms [15–18]. When the size of unlabeled data* $S > 0$, *we see from Equation (7) that the upper bound of m keeps growing with the increase of S when the size of labeled data N is fixed. For example, let* $\beta > \frac{1}{2}$ *and* $S = N^{\frac{1}{2r+\beta}}$, *then the upper bound in Equation (7) is* $O\left(N^{\frac{r}{2r+\beta}}\right)$ *and will not be a constant when* $r \to \frac{1}{2}$. *Hence, with sufficient unlabeled data* D^*, *the distributed algorithm (Equation (3)) will allow more local machines in the distributed method.*

Remark 4. *A series of distributed works [15–19] were carried out when the target function* f_ρ *lies in the space* $\mathcal{H}_K$, *i.e., the regularization parameter* $r > \frac{1}{2}$. *As a byproduct, our work in Theorem 1 does not impose the restriction* $r > \frac{1}{2}$ *on the distributed algorithm (Equation (3)).*

3. Proof of Main Result

In this section we prove our main results in Theorem 1. To this end, we introduce the data-free gradient descent method in $\mathcal{H}_K$ for the least squares, defined as $f_1 = 0$ and:

$$f_{t+1} = f_t - \eta_t \int_{\mathcal{X}} \int_{\mathcal{X}} (f_t(x,x') - f_\rho(x,x')) K_{(x,x')} d\rho_{\mathcal{X}}(x) d\rho_{\mathcal{X}}(x'), \quad t \ge 1.$$

Recalling the definition of L_K, it can be written as:

$$f_{t+1} = f_t - \eta_t L_K(f_t - f_\rho) = (I - \eta_t L_K) f_t + \eta_t L_K(f_\rho), \quad t \ge 1. \tag{10}$$

Following the standard decomposition technique in leaning theory, we split the error $\bar{f}_{t+1,\tilde{D}} - f_\rho$ into the sample error $\bar{f}_{t+1,\tilde{D}} - f_{t+1}$ and the approximation error $f_{t+1} - f_\rho$.

3.1. Approximation Error

Firstly, we estimate the approximation error $\|f_{t+1} - f_\rho\|_{L^2}$. It has been proven in Reference [20] and shown in the lemmas as follows.

Lemma 1. *Define* $\{f_t\}$ *by Equation (10) with* $0 < \eta \le 1$. *If Equation (5) holds with* $r > 0$, *there are:*

$$\|f_t - f_\rho\|_{L^2} \le c_{\phi,r} t^{-r},$$

and when $r \geq \frac{1}{2}$:

$$\|f_t - f_\rho\|_K \leq c_{\phi,r} t^{-(r-\frac{1}{2})},$$

where $c_{\phi,r} = \max\left\{\|\phi\|_{L^2}(2r/e)^r, \|\phi\|_{L^2}[(2r-1)/e]^{r-\frac{1}{2}}\right\}$.

Moreover, we derive the uniform bound of the sequence $\{f_t\}$ by Equation (10) when $0 < r < \frac{1}{2}$, which is useful in our analysis. Here and in the sequel, denote $\pi_{i+1}^t(L)$ as the polynomial operator associated with an operator L defined by $\pi_{i+1}^t(L) := \prod_{j=i+1}^t (I - \eta L)$ and $\pi_{t+1}^t := I$. We use the conventional notation $\sum_{j=T+1}^T := 1$.

Lemma 2. *Define $\{f_t\}$ by Equation (10) with $0 < \eta \leq 1$. If Equation (5) holds with $0 < r < \frac{1}{2}$, there are:*

$$\|f_t\|_K \leq d_{\phi,\eta,r} t^{\frac{1}{2}-r}, \tag{11}$$

where $d_{\phi,\eta,r}$ is defined in the proof.

Proof. Using Equation (10) iteratively from t to 1, then we have that:

$$f_{t+1} = \sum_{i=1}^t \eta \pi_{i+1}^t(L_K) L_K(f_\rho), \quad for \quad all \quad t \geq 1.$$

With Equation (5):

$$\begin{aligned}\|f_{t+1}\|_K &= \left\|\sum_{i=1}^t \eta \pi_{i+1}^t(L_K) L_K(f_\rho)\right\|_K = \left\|\sum_{i=1}^t \eta \pi_{i+1}^t(L_K) L_K L_K^r(\phi)\right\|_K \\ &= \left\|\sum_{i=1}^t \eta \pi_{i+1}^t(L_K) L_K^{r+\frac{1}{2}}\right\| \|L_K^{\frac{1}{2}}\phi\|_K = \left\|\sum_{i=1}^t \eta \pi_{i+1}^t(L_K) L_K^{r+\frac{1}{2}}\right\| \|\phi\|_{L^2}.\end{aligned} \tag{12}$$

Let $\{\sigma_k\}_{k=1}^\infty$ be the eigenvalues of the operator L_K and $0 \leq \sigma_k \leq 1, k \geq 1$, since L_K is positive and $\|L_K\|_{\mathcal{H}_K \to \mathcal{H}_K} \leq 1$, then the norm:

$$\begin{aligned}&\left\|\sum_{i=1}^t \eta \pi_{i+1}^t(L_K) L_K^{r+\frac{1}{2}}\right\| = \sup_{k\geq 1}\left|\sum_{i=1}^t \eta \pi_{i+1}^t(\sigma_k)\sigma_k^{r+\frac{1}{2}}\right| \leq \eta \sup_{a>0}\left|\sum_{i=1}^{t-1} \pi_{i+1}^t(a) a^{r+\frac{1}{2}}\right| + \left\|\eta L_K^{r+\frac{1}{2}}\right\| \\ &\leq \eta \sup_{a>0}\left\{\sum_{i=1}^{t-1}\left[\exp\{-\eta a(t-i)\}\right] a^{r+\frac{1}{2}}\right\} + \eta \\ &\leq \sum_{i=1}^{t-1} \sup_{a>0}\left\{\eta\left[\exp\{-\eta a(t-i)\}\right] a^{r+\frac{1}{2}}\right\} + \eta.\end{aligned}$$

For each $i \leq t-1$, by a simple calculation, we have:

$$\begin{aligned}&\sup_{a>0}\left\{\left[\exp\{-\eta a(t-i)\}\right] a^{r+\frac{1}{2}}\right\} = \left\{\left[\exp\{-\eta a(t-i)\}\right] a^{r+\frac{1}{2}}\right\}\Big|_{a=(r+\frac{1}{2})(\eta(t-i))^{-1}} \\ &= \eta^{\frac{1}{2}-r}(r+\frac{1}{2})^{r+\frac{1}{2}} \exp\left\{-(r+\frac{1}{2})\right\}(t-i)^{-(r+\frac{1}{2})} \leq (t-i)^{-(r+\frac{1}{2})}.\end{aligned}$$

Thus, we have:

$$\left\|\sum_{i=1}^{t}\eta\pi_{i+1}^{t}(L_K)L_K^{r+\frac{1}{2}}\right\| \leq \eta\sum_{i=1}^{t-1}(t-i)^{-(r+\frac{1}{2})}+\eta = \eta\sum_{i=1}^{t-1}i^{-(r+\frac{1}{2})}+\eta.$$

By the elementary inequality $\sum_{i=1}^{t} t^{-\theta} \leq \frac{t^{1-\theta}}{1-\theta}$ with $0<\theta<1$, it follows that:

$$\left\|\sum_{i=1}^{t}\eta\pi_{i+1}^{t}(L_K)L_K^{r+\frac{1}{2}}\right\| \leq \eta\left(\frac{1}{1/2-r}+1\right)t^{\frac{1}{2}-r} = \eta\left(\frac{3/2-r}{1/2-r}\right)t^{\frac{1}{2}-r}.$$

Together with Equation (12), then the proof is completed by taking $d_{\phi,\eta,r} := \eta\left(\frac{3/2-r}{1/2-r}\right)\|\phi\|_{L^2}$. □

3.2. Sample Error

Define the empirical operator $L_{K,D}:\mathcal{H}_K\to\mathcal{H}_K$ by:

$$L_{K,D} := \frac{1}{N^2}\sum_{\substack{(x_i,y_i)\in D\\(x_j,y_j)\in D}}\langle\cdot,K_{(x_i,x_j)}\rangle_K K_{(x_i,x_j)},$$

and for any $f\in\mathcal{H}_K$:

$$L_{K,D}(f) = \frac{1}{N^2}\sum_{\substack{(x_i,y_i)\in D\\(x_j,y_j)\in D}}\langle f,K_{(x_i,x_j)}\rangle_K K_{(x_i,x_j)} = \frac{1}{N^2}\sum_{\substack{(x_i,y_i)\in D\\(x_j,y_j)\in D}} f(x_i,x_j)K_{(x_i,x_j)}.$$

Then, the MEE gradient descent algorithm (Equation (2)) on $\tilde{D}$ can be written as:

$$f_{t+1,\tilde{D}} = [I-\eta L_{K,\tilde{D}}](f_{t,\tilde{D}})+\eta f_{\rho,\tilde{D}}+\eta E_{t,\tilde{D}}, \tag{13}$$

where:

$$E_{t,\tilde{D}} = \frac{1}{(N+S)^2}\sum_{\substack{(x_i,y_i)\in\tilde{D}\\(x_j,y_j)\in\tilde{D}}}\left(G'\left(\frac{[y_i-y_j-f_{t,\tilde{D}}(x_i,x_j)]^2}{h^2}\right)-G'(0)\right)\left(f_{t,\tilde{D}}(x_i,x_j)-y_i+y_j\right)K_{(x_i,x_j)}, \tag{14}$$

and:

$$f_{\rho,\tilde{D}} = \frac{1}{(N+S)^2}\sum_{\substack{(x_i,y_i)\in\tilde{D}\\(x_j,y_j)\in\tilde{D}}}(y_i-y_j)K_{(x_i,x_j)}.$$

In the sequel, denote:

$$\begin{aligned}
\mathcal{B}_{\tilde{D},\lambda} &= \left\|(L_{K,\tilde{D}}+\lambda I)^{-1}(L_K+\lambda)\right\|,\\
\mathcal{C}_{\tilde{D},\lambda} &= \left\|(L_K+\lambda I)^{-\frac{1}{2}}(L_K-L_{K,\tilde{D}})\right\|,\\
\mathcal{D}_{\tilde{D},\lambda} &= \left\|\frac{1}{m}\sum_{l=1}^{m}(L_K+\lambda I)^{-\frac{1}{2}}(L_K-L_{K,\tilde{D}_l})\right\|,\\
\mathcal{F}_{\tilde{D},\lambda} &= \left\|\frac{1}{m}\sum_{l=1}^{m}(L_K+\lambda I)^{-\frac{1}{2}}[f_{\rho,\tilde{D}_l}-L_K(f_\rho)]\right\|_K,\\
\mathcal{G}_{\tilde{D},\lambda} &= \left\|(L_K+\lambda I)^{-\frac{1}{2}}(L_Kf_\rho-f_{\rho,\tilde{D}})\right\|_K.
\end{aligned}$$

With these preliminaries in place, we now turn to the estimates of the sample error $\bar{f}_{t+1,\tilde{D}} - f_{t+1}$ presented in the following Lemma, whose proof can be found in the Appendix. Here and in the sequel, we use the conventional notation $\sum_{i=1}^{t} (t-i)^{-1} := \sum_{i=1}^{t-1} (t-i)^{-1} + 1$.

Lemma 3. *Let $\lambda > 0$ and $0 < \eta < \min\{C_G^{-1}, 1\}$, for any $f^* \in \mathcal{H}_K$, there holds:*

$$\|\bar{f}_{T+1,\tilde{D}} - f_{T+1}\|_{L^2} \le term\ 1 + term\ 2 + c_{p,M}|N+S|^{2p+1}N^{-(2p+1)}T^{p+3/2}h^{-2p}, \tag{15}$$

where the constant $c_{p,M} = 2^{4p+2}c_p C_G^{2p+1} M^{2p+1}$:

$$term\ 1 = \sup_{1 \le l \le m} \sum_{i=1}^{T} \Big((T-i)^{-1} + \eta\lambda \Big) \mathcal{C}_{\tilde{D}_l,\lambda} \times \Bigg\{ \sum_{s=1}^{i-1} \Big((i-s-1)^{-1} + \lambda\eta \Big) \|f_s - f^*\|_K \mathcal{B}_{\tilde{D}_l,\lambda} \mathcal{C}_{\tilde{D}_l,\lambda} \lambda^{-\frac{1}{2}}$$

$$+ (1+\lambda\eta i)\mathcal{B}_{\tilde{D}_l,\lambda}(\mathcal{C}_{\tilde{D}_l,\lambda}\|f^*\|_K + \mathcal{G}_{\tilde{D}_l,\lambda})\lambda^{-\frac{1}{2}} + c_{p,M}|N+S|^{2p+1}N^{-(2p+1)}i^{p+1/2}h^{-2p} \Bigg\},$$

and:

$$term\ 2 = \sum_{i=1}^{T} \Big((T-i)^{-1} + \eta\lambda \Big) \mathcal{D}_{\tilde{D},\lambda}\|f_i - f^*\|_K + (1+\lambda\eta T)(\mathcal{D}_{\tilde{D},\lambda}\|f^*\|_K + \mathcal{F}_{\tilde{D},\lambda}).$$

With the help of Lemma above, to bound the sample error $\|\bar{f}_{T+1,\tilde{D}} - f_{T+1}\|_{L^2}$, we first need to estimate the quantities the quantities $\mathcal{B}_{\tilde{D},\lambda}$, $\mathcal{C}_{\tilde{D},\lambda}$, $\mathcal{D}_{\tilde{D},\lambda}$ $\mathcal{F}_{\tilde{D}\lambda}$ and $\mathcal{G}_{\tilde{D},\lambda}$. Denote $\mathcal{A}_{D,\lambda} := \frac{1}{\lceil |D|/4 \rceil \sqrt{\lambda}} + \sqrt{\frac{\mathcal{N}(\lambda)}{\lceil |D|/4 \rceil}}$ ($|D|$ is the cardinality of D). In previous work [19,21–23], we have foundnd that each of the following inequality holds with confidence at least $1-\delta$:

$$\mathcal{B}_{\tilde{D},\lambda} \le 2\Big(\frac{2\mathcal{A}_{\tilde{D},\lambda}\log\frac{2}{\delta}}{\sqrt{\lambda}}\Big)^2 + 2, \quad \mathcal{C}_{\tilde{D},\lambda} \le 2\mathcal{A}_{\tilde{D},\lambda}\log\frac{2}{\delta}, \quad \mathcal{D}_{\tilde{D},\lambda} \le 2\mathcal{A}_{\tilde{D},\lambda}\log\frac{2}{\delta}$$

$$\mathcal{F}_{\tilde{D},\lambda} \le 16M\mathcal{A}_{D,\lambda}\log\frac{4}{\delta}, \qquad and \qquad \mathcal{G}_{\tilde{D},\lambda} \le 16M\mathcal{A}_{D,\lambda}\log\frac{4}{\delta}. \tag{16}$$

By Lemma 3, we also see that the function f^* is crucial to determine $\|\bar{f}_{T+1,\tilde{D}} - f_{T+1}\|$. To get a tight bound for the learning error, we should choose an appropriate $f^* \in \mathcal{H}_{\tilde{K}}$ according to the regularity of the target function. When $r \ge \frac{1}{2}$, $f_\rho \in \mathcal{H}_K$ and we take $f^* = f_\rho$. When $0 < r < \frac{1}{2}$, f_ρ is out of the space $\mathcal{H}_K$ and we let $f^* = 0$.

Now, we give the first main result when the target function f_ρ is out of $\mathcal{H}_K$ with $0 < r < \frac{1}{2}$.

Theorem 2. *Assume Equation (5) for $0 < r < \frac{1}{2}$. Let $0 < \eta < \min\{1, C_G^{-1}\}$, $T \in \mathbb{N}$ and $\lambda = T^{-1}$. Then, for any $0 < \delta < 1$, with probability at least $1-\delta$, there holds:*

$$\|\bar{f}_{T+1,\tilde{D}} - f_\rho\|_{L^2} \le C^* \Bigg\{ T^{-r} + \log^2(T)\mathcal{J}_{D,\tilde{D},\lambda}\log^4\frac{24m}{\delta} + \Big(\log(T)\mathcal{A}_{\tilde{D},\lambda}\lambda^{r-\frac{1}{2}} + \mathcal{A}_{D,\lambda} \Big)\log\frac{16}{\delta}$$

$$+ |N+S|^{2p+1}N^{-(2p+1)}h^{-2p}\left(T^{p+3/2} + T^{p+\frac{1}{2}}\log(T)\sup_{1\le l\le k}\mathcal{A}_{\tilde{D}_l,\lambda} \right)\log\frac{2}{\delta} \Bigg\}, \tag{17}$$

where C^* is a constant given in the proof, $\mathcal{J}_{D,\tilde{D},\lambda} = \sup_{1\le l\le m}\left(\Big(\frac{\mathcal{A}_{\tilde{D}_l,\lambda}}{\sqrt{\lambda}}\Big)^2 + 1 \right)(\mathcal{A}_{\tilde{D}_l,\lambda}^2\lambda^{r-1} + \mathcal{A}_{\tilde{D}_l,\lambda}\mathcal{A}_{D_l,\lambda}\lambda^{-\frac{1}{2}})$.

Proof. Decompose $\|\bar{f}_{T+1,\tilde{D}} - f_\rho\|_{L^2}$ into:

$$\|\bar{f}_{T+1,\tilde{D}} - f_\rho\|_{L^2} \le \|\bar{f}_{T+1,\tilde{D}} - f_{T+1}\|_{L^2} + \|f_{T+1} - f_\rho\|_{L^2}.$$

The estimate of $\|f_{T+1} - f_\rho\|_{L^2}$ is presented in Lemma 1. We only need to handle $\|\bar{f}_{T+1,\tilde{D}} - f_{T+1}\|_{L^2}$ by Lemma 3.

For any $0 < s \le T-1$ and $\lambda = T^{-1}$, by Equation (11), we have $\|f_s\|_K \le d_{\phi,\eta,r} s^{\frac{1}{2}-r} \le d_{\phi,\eta,r}\lambda^{r-\frac{1}{2}}$. Take $f^* = 0$ in Lemma 3, then:

$$\begin{aligned} term\ 1 \le (1 + d_{\phi,\eta,r} + c_{p,M}) \sup_{1\le l\le m} \sum_{i=1}^{T} \Big((T-i)^{-1} + \eta\lambda\Big)\mathcal{C}_{\tilde{D}_l,\lambda} \times \Bigg\{ \sum_{s=1}^{i-1} \Big((i-s-1)^{-1} + \lambda\eta\Big)\mathcal{B}_{\tilde{D}_l,\lambda}\mathcal{C}_{\tilde{D}_l,\lambda}\lambda^{r-1} \\ + (1+\lambda\eta i)\mathcal{B}_{\tilde{D}_l,\lambda}(\mathcal{C}_{\tilde{D}_l,\lambda}\lambda^{r-\frac{1}{2}} + \mathcal{G}_{\tilde{D}_l,\lambda})\lambda^{-\frac{1}{2}} + |N+S|^{2p+1}N^{-(2p+1)}i^{p+1/2}h^{-2p} \Bigg\}, \end{aligned}$$

and:

$$term\ 2 \le (1 + d_{\phi,\eta,r}) \left\{ \sum_{i=1}^{T} \Big((T-i)^{-1} + \eta\lambda\Big)\mathcal{D}_{\tilde{D},\lambda}\lambda^{r-\frac{1}{2}} + (1+\lambda\eta T)(\mathcal{D}_{\tilde{D},\lambda}\lambda^{r-\frac{1}{2}} + \mathcal{F}_{\tilde{D},\lambda}) \right\}.$$

Noticing the elementary inequality $\sum_{s=1}^{i} i^{-1} \le 2\log(i)$, then:

$$\sum_{s=1}^{i-1}\Big((i-s-1)^{-1} + \lambda\eta\Big) \le \sum_{s=1}^{i-1}\Big((i-s-1)^{-1} + T^{-1}\Big) \le 4\log(i),$$

$$\begin{aligned} &\sum_{i=1}^{T}\Big((T-i)^{-1} + \eta\lambda\Big)\sum_{s=1}^{i-1}\Big((i-s-1)^{-1} + \lambda\eta\Big) \le 4\sum_{i=1}^{T}\Big((T-i)^{-1} + T^{-1}\Big)\log(i) \\ &\le 16\sum_{i=1}^{T}\frac{\log(i)}{T-i} \le 16\log(T)\sum_{i=1}^{T}\frac{1}{T-i} = 16\log(T)\sum_{i=1}^{T-1} i^{-1} \le 32\log^2(T), \end{aligned}$$

$$\begin{aligned} &\sum_{i=1}^{T}\Big((T-i)^{-1} + \eta\lambda\Big)(1+\lambda\eta i) \le \sum_{i=1}^{T}\Big((T-i)^{-1} + \eta T^{-1}\Big)(1 + T^{-1}\eta i) \\ &\le 2\sum_{i=1}^{T}\Big((T-i)^{-1} + 1\Big) \le 8\log(T), \end{aligned}$$

and:

$$\sum_{i=1}^{T}\Big((T-i)^{-1} + \eta\lambda\Big) \le 4\log(T),$$

$$\sum_{i=1}^{T}\Big((T-i)^{-1} + \eta\lambda\Big)i^{p+1/2} \le \sum_{i=1}^{T}\Big((T-i)^{-1} + \eta\lambda\Big)T^{p+1/2} \le 4T^{p+1/2}\log(T).$$

Plugging the above inequalities into *term* 1 and *term* 2, then:

$$\begin{aligned} \mathit{term}\ 1 \leq \sup_{1\leq l\leq m} C_1\big(&\log^2(T)\mathcal{B}_{\tilde{D}_l,\lambda}\mathcal{C}^2_{\tilde{D}_l,\lambda}\lambda^{r-1} + \log(T)\mathcal{B}_{\tilde{D}_l,\lambda}\mathcal{C}^2_{\tilde{D}_l,\lambda}\lambda^{r-1} \\ &+ \log(T)\mathcal{B}_{\tilde{D}_l,\lambda}\mathcal{C}_{\tilde{D}_l,\lambda}\mathcal{G}_{\tilde{D}_l,\lambda}\lambda^{-\frac{1}{2}} + |N+S|^{2p+1}N^{-(2p+1)}T^{p+1/2}\log(T)\mathcal{C}_{\tilde{D}_l,\lambda}h^{-2p}\big), \end{aligned} \tag{18}$$

and:

$$\mathit{term}\ 2 \leq C_2(\log(T)\mathcal{D}_{\tilde{D},\lambda}\lambda^{r-\frac{1}{2}} + \mathcal{F}_{\tilde{D},\lambda}), \tag{19}$$

where $C_1 = 32(1 + d_{\phi,\eta,r} + c_{p,M})$ and $C_2 = 6(1 + d_{\phi,\eta,r})$.

By Equation (16), for any fixed l, there exist three subsets with measure at least $1-\delta$ such that:

$$\mathcal{B}_{\tilde{D}_l,\lambda} \leq 2\Big(\frac{2\mathcal{A}_{\tilde{D}_l,\lambda}\log\frac{2}{\delta}}{\sqrt{\lambda}}\Big)^2 + 2, \quad \mathcal{C}_{\tilde{D}_l,\lambda} \leq 2\mathcal{A}_{\tilde{D}_l,\lambda}\log\frac{2}{\delta},$$

and:

$$\mathcal{G}_{\tilde{D}_l,\lambda} \leq 16M\mathcal{A}_{D_l,\lambda}\log\frac{4}{\delta}.$$

Thus, for any fixed l, with confidence at least $1-3\delta$, there holds:

$$\mathcal{B}_{\tilde{D}_l,\lambda}\mathcal{C}^2_{\tilde{D}_l,\lambda}\lambda^{r-1} \leq 32\left(\Big(\frac{\mathcal{A}_{\tilde{D}_l,\lambda}}{\sqrt{\lambda}}\Big)^2 + 1\right)\mathcal{A}^2_{\tilde{D}_l,\lambda}\lambda^{r-1}\log^4\frac{2}{\delta},$$

and:

$$\mathcal{B}_{\tilde{D}_l,\lambda}\mathcal{C}_{\tilde{D}_l,\lambda}\mathcal{G}_{\tilde{D}_l,\lambda}\lambda^{-\frac{1}{2}} \leq 256M\left(\Big(\frac{\mathcal{A}_{\tilde{D}_l,\lambda}}{\sqrt{\lambda}}\Big)^2 + 1\right)\mathcal{A}_{\tilde{D}_l,\lambda}\mathcal{A}_{D_l,\lambda}\lambda^{-\frac{1}{2}}\log^3\frac{2}{\delta}\log\frac{4}{\delta}.$$

Therefore, with confidence at least $1-3m\delta$, there holds:

$$\sup_{1\leq l\leq m}\mathcal{B}_{\tilde{D}_l,\lambda}\mathcal{C}^2_{\tilde{D}_l,\lambda}\lambda^{r-1} \leq 32\left(\Big(\frac{\mathcal{A}_{\tilde{D}_l,\lambda}}{\sqrt{\lambda}}\Big)^2 + 1\right)\mathcal{A}^2_{\tilde{D}_l,\lambda}\lambda^{r-1}\log^4\frac{2}{\delta},$$

and:

$$\sup_{1\leq l\leq m}\mathcal{B}_{\tilde{D}_l,\lambda}\mathcal{C}_{\tilde{D}_l,\lambda}\mathcal{G}_{\tilde{D}_l,\lambda}\lambda^{-\frac{1}{2}} \leq 256M\left(\Big(\frac{\mathcal{A}_{\tilde{D}_l,\lambda}}{\sqrt{\lambda}}\Big)^2 + 1\right)\mathcal{A}_{\tilde{D}_l,\lambda}\mathcal{A}_{D_l,\lambda}\lambda^{-\frac{1}{2}}\log^3\frac{2}{\delta}\log\frac{4}{\delta}.$$

Thus, by Equation (18), it follows that with confidence at least $1-\delta/2$ by scaling $3m\delta$ to $\delta/2$, there holds:

$$\begin{aligned} \mathit{term}\ 1 \leq C_3\sup_{1\leq l\leq m}\Bigg(&\log^2(T)\left(\Big(\frac{\mathcal{A}_{\tilde{D}_l,\lambda}}{\sqrt{\lambda}}\Big)^2 + 1\right)(\mathcal{A}^2_{\tilde{D}_l,\lambda}\lambda^{r-1} + \mathcal{A}_{\tilde{D}_l,\lambda}\mathcal{A}_{D_l,\lambda}\lambda^{-\frac{1}{2}})\log^4\frac{24m}{\delta} \\ &+ |N+S|^{2p+1}N^{-(2p+1)}T^{p+1/2}\log(T)\mathcal{A}_{\tilde{D}_l,\lambda}h^{-2p}\Bigg), \end{aligned}$$

where $C_3 = C_1(256M + 64)$.

Similarly, with confidence at least $1 - 2\delta$ such that:

$$\mathcal{D}_{\tilde{D},\lambda} \leq 2\mathcal{A}_{\tilde{D},\lambda} \log \frac{2}{\delta}, \quad \textit{and} \quad \mathcal{F}_{\tilde{D},\lambda} \leq 16M\mathcal{A}_{D,\lambda} \log \frac{4}{\delta}.$$

By Equation (19), it follows that with confidence at least $1 - \delta/2$ by scaling 2δ to $\delta/2$:

$$\textit{term } 2 \leq C_4 \Big(\log(T)\mathcal{A}_{\tilde{D},\lambda}\lambda^{r-\frac{1}{2}} + \mathcal{A}_{D,\lambda} \Big) \log \frac{16}{\delta},$$

where $C_4 = C_2(16M + 2)$. Together with Lemma 1, we obtain the desired bound (Equation (17)) with $C^* = c_{\phi,r} + C_3 + C_4 + c_{p,M}$. □

Next, we give the result when the target function f_ρ is in $\mathcal{H}_K$ with $r \geq \frac{1}{2}$.

Theorem 3. *Assume Equation (1) for $r \geq \frac{1}{2}$. Let $0 < \eta < \min\{1, C_G^{-1}\}$, $T \in \mathbb{N}$ and $\lambda = T^{-1}$. Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$, there holds:*

$$\|\bar{f}_{T+1,\tilde{D}} - f_\rho\|_{L^2} \leq C^* \Bigg\{ T^{-r} + \log^2(T)\mathcal{K}_{D,\tilde{D},\lambda} \log^4 \frac{24m}{\delta} + \Big(\log(T)\mathcal{A}_{\tilde{D},\lambda} + \mathcal{A}_{D,\lambda} \Big) \log \frac{16}{\delta}$$
$$+ |N+S|^{2p+1}N^{-(2p+1)}h^{-2p} \Bigg(T^{p+3/2} + T^{p+\frac{1}{2}} \log(T) \sup_{1 \leq l \leq k} \mathcal{A}_{\tilde{D}_l,\lambda} \Bigg) \log \frac{2}{\delta} \Bigg\}, \quad (20)$$

where $\mathcal{K}_{D,\tilde{D},\lambda} = \sup_{1 \leq l \leq m} \left(\left(\frac{\mathcal{A}_{\tilde{D}_l,\lambda}}{\sqrt{\lambda}} \right)^2 + 1 \right) (\mathcal{A}^2_{\tilde{D}_l,\lambda} + \mathcal{A}_{\tilde{D}_l,\lambda}\mathcal{A}_{D_l,\lambda})\lambda^{-\frac{1}{2}}$ *and C^* is a constant given in the proof.*

The proof is similar to that of Theorem 2. Here we omit it.

With these preliminaries in place, we can prove our main result in Theorem 1.

Proof of Theorem 1. We first prove Equation (8) by Theorem 2 when $0 < r < \frac{1}{2}$. Let $T = \lceil |D|/4 \rceil^{\frac{1}{2r+\beta}}$ and $\lambda = T^{-1}$. Notice that $|D| = N$, $|\tilde{D}| = N + S$ and $m|D_l| = |D|, m|\tilde{D}_l| = |\tilde{D}|$ for $1 \leq l \leq m$, with $r + \beta > \frac{1}{2}$ and Equation (7), we obtain that:

$$\begin{aligned}
\mathcal{A}_{D,\lambda} &= \lceil |D|/4 \rceil^{-1+\frac{1}{4r+2\beta}} + \sqrt{C_0} \lceil |D|/4 \rceil^{-\frac{1}{2}+\frac{\beta}{4r+2\beta}} \\
&\leq (\sqrt{C_0} + 1) \lceil |D|/4 \rceil^{-\frac{r}{2r+\beta}} \leq \sqrt{5}(\sqrt{C_0} + 1)|D|^{-\frac{r}{2r+\beta}} \\
&= \sqrt{5}(\sqrt{C_0} + 1)N^{-\frac{r}{2r+\beta}},
\end{aligned}$$

$$\begin{aligned}
\mathcal{A}_{\tilde{D},\lambda} &= \lceil |\tilde{D}|/4 \rceil^{-1} [|D|/4]^{\frac{1}{4r+2\beta}} + \sqrt{C_0} \lceil |\tilde{D}|/4 \rceil^{-\frac{1}{2}} \lceil |D|/4 \rceil^{\frac{\beta}{4r+2\beta}} \\
&\leq \sqrt{5}(\sqrt{C_0} + 1)(|\tilde{D}|^{-1}|D|^{\frac{1}{4r+2\beta}} + |\tilde{D}|^{-\frac{1}{2}}|D|^{\frac{\beta}{4r+2\beta}}) \\
&= \sqrt{5}(\sqrt{C_0} + 1)(|N+S|^{-1}N^{\frac{1}{4r+2\beta}} + |N+S|^{-\frac{1}{2}}N^{\frac{\beta}{4r+2\beta}}),
\end{aligned}$$

$$\begin{aligned}
\mathcal{A}_{D_l,\lambda} &= \lceil |D_l|/4 \rceil^{-1} \lceil |D|/4 \rceil^{\frac{1}{4r+2\beta}} + \sqrt{C_0} \lceil |D_l|/4 \rceil^{-\frac{1}{2}} \lceil |D|/4 \rceil^{\frac{\beta}{4r+2\beta}} \\
&\leq \sqrt{5}(\sqrt{C_0} + 1)(m|D|^{-1+\frac{1}{4r+2\beta}} + m^{\frac{1}{2}}|D|^{-\frac{1}{2}+\frac{\beta}{4r+2\beta}}) \\
&= \sqrt{5}(\sqrt{C_0} + 1)(mN^{-1+\frac{1}{4r+2\beta}} + m^{\frac{1}{2}}N^{-\frac{1}{2}+\frac{\beta}{4r+2\beta}}),
\end{aligned}$$

and:

$$\begin{aligned}\mathcal{A}_{\tilde{D}_l,\lambda} &\le \sqrt{5}(\sqrt{C_0}+1)(|\tilde{D}_l|^{-1}|D|^{\frac{1}{4r+2\beta}} + |\tilde{D}_l|^{-\frac{1}{2}}|D|^{\frac{\beta}{4r+2\beta}})\\ &= \sqrt{5}(\sqrt{C_0}+1)(m|\tilde{D}|^{-1}|D|^{\frac{1}{4r+2\beta}} + m^{\frac{1}{2}}|\tilde{D}|^{-\frac{1}{2}}|D|^{\frac{\beta}{4r+2\beta}})\\ &\le 2\sqrt{5}(\sqrt{C_0}+1)m^{\frac{1}{2}}|\tilde{D}|^{-\frac{1}{2}}|D|^{\frac{\beta}{4r+2\beta}}.\end{aligned}$$

Thus:

$$\begin{aligned}\frac{\mathcal{A}_{\tilde{D}_l,\lambda}}{\sqrt{\lambda}} &\le \sqrt{5}(\sqrt{C_0}+1)(m|\tilde{D}|^{-1}|D|^{\frac{1}{4r+2\beta}} + m^{\frac{1}{2}}|\tilde{D}|^{-\frac{1}{2}}|D|^{\frac{\beta}{4r+2\beta}})\lceil |D|/4\rceil^{\frac{1}{2(2r+\beta)}}\\ &\le 2\sqrt{5}(\sqrt{C_0}+1)m^{\frac{1}{2}}|\tilde{D}|^{-\frac{1}{2}}|D|^{\frac{\beta+1}{4r+2\beta}}\\ &= 2\sqrt{5}(\sqrt{C_0}+1)m^{\frac{1}{2}}|N+S|^{-\frac{1}{2}}N^{\frac{\beta+1}{4r+2\beta}} \le 2\sqrt{5}(\sqrt{C_0}+1).\end{aligned}$$

It follows for $l = 1, \cdots, m$:

$$\Big(\frac{\mathcal{A}_{\tilde{D}_l,\lambda}}{\sqrt{\lambda}}\Big)^2 + 1 \le 20(\sqrt{C_0}+1)^2 + 1,$$

$$\begin{aligned}\mathcal{A}^2_{\tilde{D}_l,\lambda}\lambda^{r-1} &\le 10(\sqrt{C_0}+1)^2(m^2|\tilde{D}|^{-2}|D|^{\frac{1}{2r+\beta}} + m|\tilde{D}|^{-1}|D|^{\frac{s}{2r+\beta}})|D|^{\frac{1-r}{2r+\beta}}\\ &= 10(\sqrt{C_0}+1)^2(m^2|\tilde{D}|^{-2}|D|^{\frac{2}{2r+\beta}} + m|\tilde{D}|^{-1}|D|^{\frac{1+\beta}{2r+\beta}})|D|^{-\frac{r}{2r+\beta}}\\ &\le 20(\sqrt{C_0}+1)^2|D|^{-\frac{r}{2r+s}} / \log^6|D|\\ &= 20(\sqrt{C_0}+1)^2N^{-\frac{r}{2r+s}} / \log^6 N,\end{aligned}$$

$$\begin{aligned}\mathcal{A}_{\tilde{D}_l,\lambda}\mathcal{A}_{D_l,\lambda}\lambda^{-\frac{1}{2}} &\le 10(\sqrt{C_0}+1)^2m^{\frac{1}{2}}|\tilde{D}|^{-\frac{1}{2}}|D|^{\frac{\beta}{4r+2\beta}}(m|D|^{-1+\frac{1}{4r+2\beta}} + m^{\frac{1}{2}}|D|^{-\frac{1}{2}+\frac{\beta}{4r+2\beta}})|D|^{\frac{1}{4r+2\beta}}\\ &\le 10(\sqrt{C_0}+1)^2(m^{\frac{3}{2}}|\tilde{D}|^{-\frac{1}{2}}|D|^{-\frac{\beta+4r-2}{4r+2\beta}} + m|\tilde{D}|^{-\frac{1}{2}}|D|^{\frac{\beta+1-2r}{4r+2\beta}})\\ &\le 20(\sqrt{C_0}+1)^2|D|^{-\frac{r}{2r+\beta}} / \log^6|D|\\ &= 20(\sqrt{C_0}+1)^2N^{-\frac{r}{2r+\beta}} / \log^6 N.\end{aligned}$$

Thus, by the above estimates:

$$\mathcal{J}_{D,\tilde{D},\lambda} \le 40(20(\sqrt{C_0}+1)^2+1)(\sqrt{C_0}+1)^2N^{-\frac{r}{2r+\beta}} / \log^6 N$$

Thus:

$$\begin{aligned}&\log^2(T)\mathcal{J}_{D,\tilde{D},\lambda}\log^4\frac{24m}{\delta} \le 2^4\log^2(T)\mathcal{J}_{D,\tilde{D},\lambda}(\log^4 m)\log^4\frac{24}{\delta}\\ &\le 2^4(2r+\beta)^{-2}(\log^2 N)\mathcal{J}_{D,\tilde{D},\lambda}(\log^4 m)\log^4\frac{24}{\delta}\\ &\le 2^4(2r+\beta)^{-2}(\log^6 N)\mathcal{J}_{D,\tilde{D},\lambda}\log^4\frac{24}{\delta}\\ &\le 2^{10}(2r+\beta)^{-2}(20(\sqrt{C_0}+1)^2+1)(\sqrt{C_0}+1)^2|D|^{-\frac{r}{2r+\beta}}\log^4\frac{24}{\delta}\\ &= 2^{10}(2r+\beta)^{-2}(20(\sqrt{C_0}+1)^2+1)(\sqrt{C_0}+1)^2N^{-\frac{r}{2r+\beta}}\log^4\frac{24}{\delta},\end{aligned}$$

and:

$$\log(T)\mathcal{A}_{\tilde{D},\lambda}\lambda^{r-\frac{1}{2}} \leq (2r+\beta)^{-1}\sqrt{5}(\sqrt{C_0}+1)\log|D|(|\tilde{D}|^{-1}|D|^{\frac{1}{2r+\beta}}+|\tilde{D}|^{-\frac{1}{2}}|D|^{\frac{\beta+1}{4r+2\beta}})|D|^{-\frac{r}{2r+\beta}}$$
$$\leq 2\sqrt{5}(2r+\beta)^{-1}(\sqrt{C_0}+1)|D|^{-\frac{r}{2r+\beta}} = 2\sqrt{5}(2r+\beta)^{-1}(\sqrt{C_0}+1)N^{-\frac{r}{2r+\beta}}.$$

Putting the above estimates into Theorem 2, we have the desired conclusion (Equation (8)) with:

$$C' = C^*\left(2^{10}(2r+\beta)^{-2}(20(\sqrt{C_0}+1)^2+1)(\sqrt{C_0}+1)^2+2\sqrt{5}(2r+\beta)^{-1}(\sqrt{C_0}+1)+\sqrt{5}(\sqrt{C_0}+2)\right).$$

When $r \geq \frac{1}{2}$, we apply Theorem 3 and take the same proof procedure a above. Then, the conclusion (Equation (8)) can be obtained. The proof is completed. □

4. Simulation and Conclusions

In this section, we provide the simulation to verify our theoretical statements. We assume that the inputs $\{x_i\}$ are independently drawn according to the uniform distribution on $[0,1]$. Consider the regression model $y_i = g_\rho(x_i)+\varepsilon_i, i = 1,\cdots,N$, where ε_i is the independent Gaussian noise $\mathcal{N}(0,0.1^2)$ and:

$$g_\rho(x) = \begin{cases} x, & \text{if } 0 < x \leq 0.5, \\ 1-x, & \text{if } 0.5 < x \leq 1. \end{cases}$$

Define the pairwise kernel $K : \mathcal{X}^2 \times \mathcal{X}^2 \to \mathbb{R}$ by $K((x,u),(x',u')) := K_1(x,x')+K_1(u,u')-K_1(x,u')-K_1(x',u)$ where:

$$K_1(x,x') = 1+\min\{x,x'\}.$$

We apply the kernel K to the distributed algorithm (Equation (3)). In Figure 1, we plot the mean squared error of Equation (3) for $N = 600$ and $S = 0, 300, 600$ when the number of local machines m varies. Note that $S = 0$, and it is a standard distributed MEE algorithm without unlabeled data. When m becomes large, the red curve increases dramatically. However, when we add 300 or 600 unlabeled data, the error curves begin to increase very slowly. This coincides with our theory that using unlabeled data can enlarge the range of m in the distributed method.

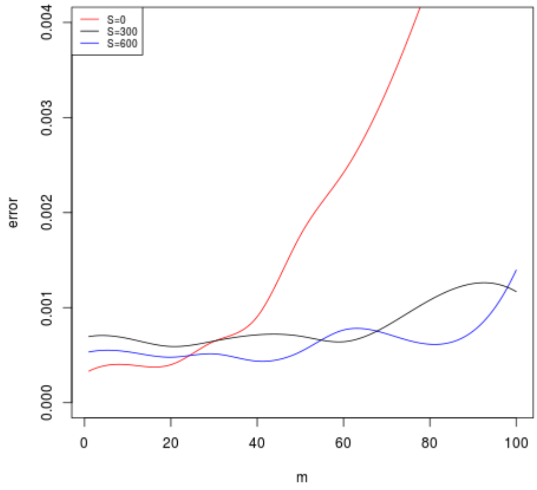

Figure 1. The mean square errors for the size of unlabeled data $S \in \{0, 300, 600\}$ as the number of local machines m varies.

This paper studied the convergence rate of the distribute gradient descent MEE algorithm in a semi-supervised setting. Our results demonstrated that using additional unlabeled data can improve the learning performance of the distributed MEE algorithm, especially in enlarging the range of m to guarantee the learning rate. As we know, there are many gaps between theory and empirical studies. We regard this paper as mainly a theoretical paper and expect that the theoretical analysis give some guidance to real applications.

Author Contributions: B.W. conceived the presented idea. T.H. developed the theory and performed the computations. All authors discussed the results and contributed to the final manuscript.

Funding: The work described in this paper is partially supported by the National Natural Science Foundation of China [Nos. 11671307 and 11571078], the Natural Science Foundation of Hubei Province in China [No. 2017CFB523], and the Fundamental Research Funds for the Central Universities, South-Central University for Nationalities [No. CZY18033].

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

Appendix A. Proof of Lemma 3

We state two useful lemmas as follows, whose proof can be found in Reference [22].

Lemma A1. *For $\lambda > 0$, $0 < \eta < 1$ and $j = 1, \cdots, t-1$, we have:*

$$\max\{\|\eta(L_K+\lambda)\pi_{j+1}^t(L_K)\|, \|\eta(L_{K,\tilde{D}}+\lambda)\pi_{j+1}^t(L_{K,\tilde{D}})\|\} \le \frac{1}{t-j} + \eta\lambda,$$

$$\max\{\|\sum_{j=1}^t \eta(L_K+\lambda)\pi_{j+1}^t(L_K)\|, \|\sum_{j=1}^t \eta(L_{K,\tilde{D}}+\lambda)\pi_{j+1}^t(L_{K,\tilde{D}})\|\} \le 1 + \eta\lambda t.$$

Lemma A2. *For any $\lambda > 0$ and $f^* \in \mathcal{H}_K$, there holds:*

$$\begin{aligned}\|f_{t+1,\tilde{D}} - f_{t+1}\|_{L^2} \le\ & \mathcal{B}_{\tilde{D},\lambda}\mathcal{C}_{\tilde{D},\lambda}\sum_{i=1}^t \left[(t-i)^{-1} + \eta\lambda\right] \|f_i - f^*\|_K \\ & + \mathcal{B}_{\tilde{D},\lambda}\,(1+\eta\lambda t)\left(\mathcal{C}_{\tilde{D},\lambda}\|f^*\|_K + \mathcal{G}_{\tilde{D},\lambda}\right) + c_{p,M}(N+S)^{2p+1}N^{-(2p+1)}t^{p+1/2}h^{-2p},\end{aligned} \quad \text{(A1)}$$

and:

$$\begin{aligned}\|f_{t+1,\tilde{D}} - f_{t+1}\|_K \le\ & \mathcal{B}_{\tilde{D},\lambda}\mathcal{C}_{\tilde{D},\lambda}\sum_{i=1}^t \left[(t-i)^{-1} + \eta\lambda\right] \|f_i - f^*\|_K/\sqrt{\lambda} \\ & + \mathcal{B}_{\tilde{D},\lambda}\,(1+\eta\lambda t)\left(\mathcal{C}_{\tilde{D},\lambda}\|f^*\|_K + \mathcal{G}_{\tilde{D},\lambda}\right)/\sqrt{\lambda} + c_{p,M}(N+S)^{2p+1}N^{-(2p+1)}t^{p+1/2}h^{-2p},\end{aligned} \quad \text{(A2)}$$

where the constant $c_{p,M} = 2^{4p+2}c_pC_G^{2p+1}M^{2p+1}$.

Proof of Lemma A2. By Equations (10) and (13), we get a two error decomposition for $f_{t+1,D} - f_{t+1}$.

The first one is:

$$\begin{aligned}f_{t+1,\tilde{D}} - f_{t+1} =\ & \eta\sum_{i=1}^t \pi_{i+1}^t(L_{K,\tilde{D}})[L_K - L_{K,\tilde{D}}](f_i) \\ & + \eta\sum_{i=1}^t \pi_{i+1}^t(L_{K,\tilde{D}})[f_{\rho,\tilde{D}} - L_K(f_\rho)] + \eta\sum_{i=1}^t \pi_{i+1}^t(L_{K,\tilde{D}})E_{i,\tilde{D}},\end{aligned} \quad \text{(A3)}$$

and the second one is:

$$f_{t+1,\tilde{D}} - f_{t+1} = \eta \sum_{i=1}^{t} \pi_{i+1}^{t}(L_K)[L_K - L_{K,\tilde{D}}](f_{i,\tilde{D}}) \\ + \eta \sum_{i=1}^{t} \pi_{i+1}^{t}(L_K)[f_{\rho,\tilde{D}} - L_K(f_\rho)] + \eta \sum_{i=1}^{t} \pi_{i+1}^{t}(L_K) E_{i,\tilde{D}}. \tag{A4}$$

It has been proven in Reference [22] that $\{f_{t,\tilde{D}}\}$ as $\|f_{t,\tilde{D}}\|_K \leq 2C_G \frac{|\tilde{D}|}{|D|} M t^{\frac{1}{2}} = 2C_G \frac{|N+S|}{N} M t^{\frac{1}{2}}$. It follows from Equation (4) that:

$$\|E_{t,\tilde{D}}\|_K \leq \frac{1}{|N+S|^2} \sum_{\substack{(x_i,y_i) \in \tilde{D}, \\ (x_j,y_j) \in \tilde{D}}} \left\| \left[G'\left(\frac{(f_{t,D^*}(x_i,x_j) - y_i + y_j)^2}{h^2} \right) - G'(0) \right] (f_{t,\tilde{D}}(x_i,x_j) - y_i + y_j) K_{(x_i,x_j)} \right\|_K \\ \leq c_p \frac{\left(2\frac{|N+S|}{N} C_G M + 2\right)^{2p+1}}{h^{2p}} \|f_{t,\tilde{D}}\|_K^{2p+1} \leq 2^{4p+2} c_p C_G^{2p+1} M^{2p+1} \frac{|N+S|^{2p+1}}{N^{2p+1}} t^{p+1/2} h^{-2p}. \tag{A5}$$

Then, we can follow the proof procedure in Proposition 1 of Reference [24] to prove Equations (A2) and (A1). □

With the help of the lemmas above, we can prove Lemma 3.

Proof of Lemma 3. Applying Equation (A4) with $\tilde{D} = \tilde{D}_l$ for $l = 1, \cdots, m$, we have that:

$$\|\bar{f}_{T+1,\tilde{D}_l} - f_{T+1}\|_{L^2} = \left\| \frac{1}{m} \sum_{l=1}^{m} \left(\bar{f}_{T+1,\tilde{D}_l} - f_{T+1} \right) \right\|_{L^2} \\ \leq \|\eta \sum_{i=1}^{T} \pi_{i+1}^{T}(L_K) \frac{1}{m} \sum_{l=1}^{m} [L_K - L_{K,\tilde{D}_l}](f_{i,\tilde{D}_l})\|_{L^2} \\ + \|\eta \sum_{i=1}^{T} \pi_{i+1}^{T}(L_K) \frac{1}{m} \sum_{l=1}^{m} [\hat{f}_{\rho,\tilde{D}_l} - L_K(f_\rho)]\|_{L^2} + \|\eta \sum_{i=1}^{T} \pi_{i+1}^{T}(L_K) \frac{1}{m} \sum_{l=1}^{m} E_{i,\tilde{D}_l}\|_{L^2} \\ := I_1 + I_2 + I_3.$$

Firstly, we will bound I_1, which is most difficult to handle. It can be decomposed as:

$$I_1 \leq \left\| \eta \sum_{i=1}^{T} \pi_{i+1}^{T}(L_K) \frac{1}{m} \sum_{l=1}^{m} (L_K + \lambda)^{\frac{1}{2}} [L_K - L_{K,\tilde{D}_l}](f_{i,\tilde{D}_l} - f_i) \right\|_K \\ + \left\| \eta \sum_{i=1}^{T} \pi_{i+1}^{T}(L_K) \frac{1}{m} \sum_{l=1}^{m} (L_K + \lambda)^{\frac{1}{2}} [L_K - L_{K,\tilde{D}_l}](f_i - f^*) \right\|_K \\ + \left\| \eta \sum_{i=1}^{T} \pi_{i+1}^{T}(L_K) \frac{1}{m} \sum_{l=1}^{m} (L_K + \lambda)^{\frac{1}{2}} [L_K - L_{K,\tilde{D}_l}](f^*) \right\|_K \\ := I_{11} + I_{12} + I_{13}.$$

Then, it is easy to get that by Lemma A1 and $f_{1,\tilde{D}_l} = f_1 = 0$:

$$\begin{aligned} I_{11} &= \left\| \sum_{i=1}^{T} \eta(L_K+\lambda)\pi_{i+1}^T(L_K)\frac{1}{m}\sum_{l=1}^{m}(L_K+\lambda)^{-\frac{1}{2}}[L_K - L_{K,\tilde{D}_l}](f_{i,\tilde{D}_l}-f_i)\right\|_K \\ &\le \sum_{i=1}^T \left\|\eta(L_K+\lambda)\pi_{i+1}^T(L_K)\right\| \left\|\frac{1}{m}\sum_{l=1}^m (L_K+\lambda)^{-\frac12}[L_K-L_{K,\tilde D_l}](f_{i,\tilde D_l}-f_i)\right\|_K \\ &\le \sum_{i=1}^T(\eta\lambda+(T-i)^{-1})\sup_{1\le l\le m}\left\|(L_K+\lambda)^{-\frac12}[L_K-L_{K,\tilde D_l}](f_{i,\tilde D_l}-f_i)\right\|_K \\ &\le \sup_{1\le l\le m}\sum_{i=1}^T(\eta\lambda+(T-i)^{-1})\|f_{i,\tilde D_l}-f_i\|_K\mathcal{C}_{\tilde D_l,\lambda}. \end{aligned}$$

Applying Proposition A2 with $\tilde{D} = \tilde{D}_l$ and $t+1=i$, we have that:

$$\begin{aligned} \|f_{i,\tilde D_l}-f_i\|_K &\le \sum_{s=1}^{i-1}\Big((i-s-1)^{-1}+\lambda\eta\Big)\|f_s-f^*\|_K\mathcal{B}_{\tilde D_l,\lambda}\mathcal{C}_{\tilde D_l,\lambda}\lambda^{-\frac12} \\ &+ (1+\lambda\eta i)\mathcal{B}_{\tilde D_l,\lambda}(\mathcal{C}_{\tilde D_l,\lambda}\|f^*\|_K+\mathcal{G}_{\tilde D_l,\lambda})\lambda^{-\frac12}+c_{p,M}\frac{|N+S|^{2p+1}}{N^{2p+1}}i^{p+1/2}h^{-2p}. \end{aligned}$$

Thus:

$$\begin{aligned} I_{11} \le \sup_{1\le l\le m}\sum_{i=1}^T(\eta\lambda+(T-i)^{-1})\mathcal{C}_{\tilde D_l,\lambda}\times\Bigg\{&\sum_{s=1}^{i-1}\Big((i-s-1)^{-1}+\lambda\eta\Big)\|f_s-f^*\|_K\mathcal{B}_{\tilde D_l,\lambda}\mathcal{C}_{\tilde D_l,\lambda}\lambda^{-\frac12} \\ +(1+\lambda\eta i)\mathcal{B}_{\tilde D_l,\lambda}(\mathcal{C}_{\tilde D_l,\lambda}\|f^*\|_K+\mathcal{G}_{\tilde D_l,\lambda})\lambda^{-\frac12}+c_{p,M}&\frac{|N+S|^{2p+1}}{N^{2p+1}}i^{p+1/2}h^{-2p}\Bigg\}. \end{aligned} \tag{A6}$$

By Lemma A1 again, we have:

$$\begin{aligned} I_{12} &\le \left\|\sum_{i=1}^T\eta(L_K+\lambda)\pi_{i+1}^T(L_K)\frac1m\sum_{l=1}^m(L_K+\lambda)^{-\frac12}[L_K-L_{K,\tilde D_l}](f_i-f^*)\right\|_K \\ &\le \sum_{i=1}^T\left\|\eta(L_K+\lambda)\pi_{i+1}^T(L_K)\right\|\left\|\frac1m\sum_{l=1}^m(L_K+\lambda)^{-\frac12}[L_K-L_{K,\tilde D_l}]\|\|f_i-f^*\right\|_K \\ &\le \sum_{i=1}^T(\eta\lambda+(T-i)^{-1})\mathcal{D}_{\tilde D,\lambda}\|f_i-f^*\|_K, \end{aligned} \tag{A7}$$

and:

$$\begin{aligned} I_{13} &\le \left\|\sum_{i=1}^T\eta(L_K+\lambda)\pi_{i+1}^T(L_K)\right\|\left\|\frac1m\sum_{l=1}^m(L_K+\lambda)^{-\frac12}[L_K-L_{K,\tilde D_l}]\right\|\|f^*\|_K \\ &\le (1+\lambda\eta T)\mathcal{D}_{\tilde D,\lambda}\|f^*\|_K. \end{aligned} \tag{A8}$$

This completes the estimate of I_1 with Equations (A6), (A7), and (A8).

Now we turn to bound I_2. Then, by the definition of $\mathcal{F}_{\tilde D,\lambda}$, the bound (Equation (A5)) of $E_{t,\tilde D}$ and Lemma A1, we obtain that:

$$I_2 \le (1+\eta\lambda T)\mathcal{F}_{\tilde D,\lambda},$$

and:

$$I_3 \le c_{p,M}\frac{|\tilde D|^{2p+1}}{|D|^{2p+1}}T^{p+3/2}h^{-2p} = c_{p,M}\frac{|N+S|^{2p+1}}{N^{2p+1}}T^{p+3/2}h^{-2p}.$$

Together with the bound of Equations (A6), (A7), and (A8), we can get the desired conclusion (Equation (15)). □

References

1. Principe, J.C. Renyi's entropy and Kernel perspectives. In *Information Theoretic Learning*; Springer: New York, NY, USA, 2010.
2. Erdogmus, D.; Principe, J.C. Comparison of entropy and mean square error criteria in adaptive system training using higher order statistics. In *Proceedings of the International Conference on ICA and Signal Separation*; Springer: Berlin, Germany, 2000; pp. 75–90.
3. Erdogmus, D.; Hild, K.; Principe, J.C. Blind source separation using Renyi's α-marginal entropies. *Neurocomputing* **2002**, *49*, 25–38. [CrossRef]
4. Erdogmus, D.; Principe, J.C. Convergence properties and data efficiency of the minimum error entropy criterion in adaline training. *IEEE Trans. Signal Process.* **2003**, *51*, 1966–1978. [CrossRef]
5. Gokcay, E.; Principe, J.C. Information theoretic clustering. *IEEE Trans. Pattern Anal. Mach. Learn.* **2002**, *24*, 158–171. [CrossRef]
6. Silva, L.M.; Marques, J.; Alexandre, L.A. Neural network classification using Shannon's entropy. In *Proceedings of the European Symposium on Artificial Neural Networks*; D-Side: Bruges, Belgium, 2005; pp. 217–222.
7. Silva, L.M.; Marques, J.; Alexandre, L.A. The MEE principle in data classification: A perceptron-based analysis. *Neural Comput.* **2010**, *22*, 2698–2728. [CrossRef]
8. Choe, Y. Information criterion for minimum cross-entropy model selection. *arXiv* **2017**, arXiv:1704.04315.
9. Ying, Y.; Zhou, D.X. Online pairwise learning algorithms. *Neural Comput.* **2016**, *28*, 743–777. [CrossRef] [PubMed]
10. Zhang, Y.; Duchi, J.C.; Wainwright, M.J. Divide and conquer kernel ridge regression: A dis tributed algorithm with minimax optimal rates. *J. Mach. Learn. Res.* **2013**, *30*, 592–617.
11. Chapelle, O.; Zien, A. *Semi-Supervised Learning (Adaptive Computation and Machine Learning)*; The MIT Press: Cambridge, MA, USA, 2006.
12. Zhang, T. Learning bounds for kernel regression using effective data dimensionality. *Neural Comput.* **2005**, *17*, 2077–2098. [CrossRef] [PubMed]
13. Caponnetto, A.; Vito, E.D. Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.* **2007**, *7*, 331–368. [CrossRef]
14. Steinwart, I.; Hush, D.R.; Scovel, C. Optimal rates for regularized least squares regression. In Proceedings of the COLT 2009—the Conference on Learning Theory, Montreal, QC, Canada, 18–21 June 2009.
15. Lin, S.B.; Guo X.; Zhou, D.X. Distributed learning with regularized least squares. *J. Mach. Learn. Res.* **2017**, *18*, 3202–3232.
16. Guo, Z.C.; Lin, S.B.; Zhou, D.X. Learning theory of distributed spectral algorithms. *Inverse Prob.* **2017**, *33*, 074009. [CrossRef]
17. Guo, Z.C.; Shi, L.; Wu, Q. Learning theory of distributed regression with bias corrected regu-larization kernel network. *J. Mach. Learn. Res.* **2017**, *18*, 4237–4261.
18. M*ü*cke, N.; Blanchard, G. Parallelizing spectrally regularized kernel algorithms. *J. Mach. Learn. Res.* **2018**, *19*, 1069–1097.
19. Lin, S.B.; Zhou, D.X. Distributed kernel-based gradient descent algorithms. *Constr. Approx.* **2018**, *47*, 249–276. [CrossRef]
20. Yao, Y.; Rosasco, L.; Caponnetto, A. On early stopping in gradient descent learning. *Constr. Approx.* **2007**, *26*, 289–315. [CrossRef]
21. Chang, X.; Lin, S.B.; Zhou, D.X. Distributed semi-supervised learning with kernel ridge re-gression. *J. Mach. Learn. Res.* **2017**, *18*, 1493–1514.
22. Hu, T.; Wu, Q.; Zhou, D.X. Distributed kernel gradient descent algorithm for minimum error entropy principle. Unpublished work, 2018.

23. Guo, X.; Hu, T.; Wu, Q. Distributed minimum error entropy algorithms. Unpublished work, 2018.
24. Wang, B.; Hu, T. Distributed pairwise algorithms with gradient descent methods. Unpublished work, 2018.