Distributed Minimum Error Entropy Algorithms

Xin Guo
Department of Applied Mathematics
The Hong Kong Polytechnic University
Hong Kong, China, and
School of Mathematics and Physics
The University of Queensland
Brisbane, QLD 4072, Australia

Ting Hu
School of Mathematics and Statistics
Wuhan University
Wuhan, 430072, China

Qiang Wu
Department of Mathematical Sciences
Middle Tennessee State University
Murfreesboro, TN 37132, USA

Editor: Lorenzo Rosasco

Abstract

Minimum Error Entropy (MEE) principle is an important approach in Information Theoretical Learning (ITL). It is widely applied and studied in various fields for its robustness to noise. In this paper, we study a reproducing kernel-based distributed MEE algorithm, DMEE, which is designed to work with both fully supervised data and semi-supervised data. The divide-and-conquer approach is employed, so there is no inter-node communication overhead. Similar as other distributed algorithms, DMEE significantly reduces the computational complexity and memory requirement on single computing nodes. With fully supervised data, our proved learning rates equal the minimax optimal learning rates of the classical pointwise kernel-based regressions. Under the semi-supervised learning scenarios, we show that DMEE exploits unlabeled data effectively, in the sense that first, under the settings with weak regularity assumptions, additional unlabeled data significantly improves the learning rates of DMEE. Second, with sufficient unlabeled data, labeled data can be distributed to many more computing nodes, that each node takes only $O(1)$ labels, without spoiling the learning rates in terms of the number of labels. This conclusion overcomes the saturation phenomenon in unlabeled data size. It parallels a recent results for regularized least squares (Lin and Zhou, 2018), and suggests that an inflation of unlabeled data is a solution to the MEE learning problems with decentralized data source for the concerns of privacy protection. Our work refers to pairwise learning and non-convex loss. The theoretical analysis is achieved by distributed U-statistics and error decomposition techniques in integral operators.

Keywords: Information theoretic learning, minimum error entropy, distributed method, semi-supervised data, reproducing kernel Hilbert space

1. Introduction

Pioneered by the work of Principe and his collaborators in Erdogmus and Principe (2000), MEE principle has been playing an essential role in ITL (Principe, 2010). MEE principle is widely adopted as a powerful alternative to the traditional least squares method which is suboptimal in non-Gaussian situations (Erdogmus and Principe, 2000). The classical least square method minimizes the variance of the prediction error. Its optimality heavily depends on the assumption of Gaussianity.
due to the use of a second order statistics. When Gaussianity assumption is violated, high order 
methods are desired. Entropy is a functional of the probability density function of the error variable 
and measures the average information contained in the distribution. Minimizing entropy allows one 
to take into account high-order statistical behavior in the learning process and thus is advantageous 
in non-Gaussian scenarios. Given an error variable $E$, Renyi’s entropy and Shannon entropy are 
widely used to quantify the information contained in $E$. In this paper we focus on the quadratic 
Renyi’s entropy, which is defined by

$$H(E) = -\log \mathbb{E}(p_E) = -\log \int_E p_E^2(e)de,$$

where $p_E$ is the probability density function of $E$. MEE employs entropy as a new measurement 
of error to substitute the mean squared error $\int_E e^2 p_E(e)de$ in the least squares. Renyi’s entropy takes 
into consideration all higher moments rather than the variance used by the least squares. Hence, 
MEE is capable of dealing with outliers, heavy-tailed noise or skewed noise distribution. Because 
of its robustness to non-Gaussian noise, MEE performs well in a large number of applications such 
as signal processing, regression analysis, feature selection, and data clustering. See Erdogmus and 
Principe (2003); Chen et al. (2010); Gokcay and Principe (2002); Shen and Li (2015); Silva et al. 
(2010). Meanwhile, MEE has the nature of pairwise learning (Christmann and Zhou, 2016; Wang 
et al., 2012, Ying and Zhou, 2016), which focuses on approximating the difference of labels between 
each pair of sample points, incurring high computational complexity. The complexity restricts the 
application of MEE algorithms on the problems with large data size. Although there is a series of 
work on the theory and applications of MEE (Hu et al., 2015; Fan et al. 2016; Hu et al., 2013), few 
works have been done to reduce the computational complexity, which is one of the motivations of 
this paper.

In the recent decade, the growth of computing facility power falls way behind the growth of 
the scale of data, and the research and practice of privacy protection falls way behind the growing 
concerns of privacy. Distributed algorithms have drawn much attention of machine learning and 
opimization communities, and are widely implemented in industry. Distributed approaches reduce 
computational complexity and memory requirement for single computing nodes, and can also be 
applied to the scenarios where data have to be stored and analyzed locally for privacy concerns. In 
this paper, we study a distributed MEE algorithm without communication overhead. Specifically, 
one first divides a large data set into several subsets, then sends each subset as training sample to 
a computing node for a local output function, and finally averages these local output functions to 
synthesize the overall output function. Alternatively, different data subsets may directly be used 
locally to train local output functions, and the prediction is done by distributing the new instance, 
then collecting and averaging the local predictions. This scheme has been developed for a lot of 
classical learning algorithms, including kernel ridge regression (Lin et al., 2017; Zhang et al. 2015), 
stochastic gradient descent algorithm (Lin and Zhou, 2018, Zinkevich et al. 2010), spectral algorithm 
(Mücke and Blanchard 2018, Guo et al. 2017a), and bias correction (Guo et al., 2017b). For the 
applications of MEE algorithm that have privacy concerns, we adopt semi-supervised learning to our 
distributed scheme. Semi-supervised learning itself is an active research area, with one of the earliest 
ideas stemming from self-learning in classification, known as self-training, self-labeling, or decision-
directed learning (Chapelle et al., 2006), and is later extended in various forms to other applications, 
including co-training in text classification (Blum and Mitchell, 1998), graph-based methods (Wang 
et al., 2013), and manifold regularization (Belkin and Niyogi 2004). We focus on improving the 
distributed MEE algorithm performance by utilizing unlabeled data.

Most existing works on MEE methods study only linear models. The distributed MEE algorithms 
we study employ reproducing kernels and are able to fit nonlinear models. The non-convex and 
pairwise loss functions caused the main difficulties in analysis, which we overcome by employing 
some decomposition techniques in U-statistics.
This paper provides three main contributions. First, existing analysis of MEE algorithm in the literature has largely been improved, and extended losslessly to DMEE. Our obtained learning rates coincide with the minimax optimal rates of regularized least squares algorithms for pointwise learning. Second, we prove that unlabeled data can significantly improve learning rates under the setting with weak regularity assumptions. Third, we prove that with sufficient unlabeled data, the restriction on the maximum number of computing nodes that labeled data are distributed to is removed.

The paper is organized as follows. In Section 2, we review the background of MEE learning, define the DMEE algorithms, and present our main results on learning rates. In Section 3, we provide detailed discussions and comparisons. Mathematical analysis goes to Sections 4 and 5 for supervised and semi-supervised data respectively.

2. Backgrounds and main results

In this paper we study regression problems. We assume that the explanatory variable $X$ takes values in a compact domain $\mathcal{X}$ in an Euclidean space, the response variable $Y$ takes values in the output space $\mathcal{Y}$ which is a subset of the real line $\mathbb{R}$, and

$$Y = g^*(X) + \epsilon,$$

where $g^*$ is the target function and $\epsilon$ is the noise in the regression model. Let $\rho$ be a Borel probability measure on the product space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Let $\rho_X$ and $\rho(y|x)$ denote the marginal distribution of $\rho$ on $\mathcal{X}$, and the conditional distribution on $\mathcal{Y}$ given $x \in \mathcal{X}$, respectively. The purpose of regression is to estimate $g^*(X)$ according to a sample $D = \{(x_i, y_i)\}_{i=1}^{|D|}$ drawn independently from $\rho$, where $|D|$ is the sample size, the cardinality of $D$. Given a hypothesis space of functions $g : \mathcal{X} \to \mathcal{Y}$, MEE looks for a good approximation of $g^*$ by minimizing the entropy of the prediction error $E = g(X) - Y$. Here we consider Renyi’s quadratic entropy (1). Denote $e_i = g(x_i) - y_i$, $(x_i, y_i) \in D$ for $1 \leq i \leq |D|$ and $pE$ can be estimated by Parzen windowing (Parzen, 1962). Given a windowing function $G : (−\infty, +\infty) \to [0, +\infty)$ and a scaling parameter $h > 0$, one gets the density estimator

$$\hat{p}_E(e) = \frac{c}{|D|} \sum_{i=1}^{|D|} G_h(e - e_i) = \frac{c}{|D|h} \sum_{i=1}^{|D|} G\left(\frac{(e - e_i)^2}{h^2}\right),$$

where $c$ is a normalization constant so that $\int_{-\infty}^{\infty} cG_h(t)dt = 1$. A typical example is the windowing function $G(a) = \exp(-a)$ with $a \geq 0$, associated to which are the constant $c = \frac{1}{\sqrt{\pi}}$ and the Gaussian kernel $cG_h(t) = \frac{1}{\sqrt{\pi}h} \exp(-\frac{t^2}{2h^2})$. Then the empirical Renyi’s entropy of (1) is

$$H_D(g) = -\log \left\{ \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{1}{|D|h} \sum_{j=1}^{|D|} G\left(\frac{(e_i - e_j)^2}{h^2}\right) \right\} - \log c.$$

MEE algorithm searches over a suitable hypothesis space $\mathcal{H}$ for a minimizer of $H_D(g)$. Equivalently, one can just minimizes the factor

$$R_D(g) = -\frac{h^2}{|D|^2} \sum_{i=1}^{|D|} \sum_{j=1}^{|D|} G\left(\frac{(e_i - e_j)^2}{h^2}\right)$$

$$= -\frac{h^2}{|D|^2} \sum_{(x,y) \in D} \sum_{(u,v) \in D} G\left(\frac{[(g(x) - y) - (g(u) - v)]^2}{h^2}\right).$$

(2)
MEE algorithm outputs $g_D := \arg \min_{g \in \mathcal{H}} \mathcal{R}_D(g)$ as the estimator of $g^*$.

In this work, we study the kernel based MEE algorithm, which includes the linear models studied in the literature as a special case. The learning process of the kernel MEE method is associated with the pairwise empirical risk

$$\mathcal{E}_D(f) = \frac{h^2}{|D|^2} \sum_{(x,y) \in D} G\left(\frac{[(f(x,y) - y + v)]^2}{h^2}\right).$$

Our target function is now $f_D(x,u) := g^*(x) - g(u)$. It is pointed out in Ying and Zhou (2016) that by the restriction

$$K((x,u),(x',u')) = W(x,x') + W(u,u') - W(x,u') - W(u,x')$$

(3)

(where $W$ is a reproducing kernel on $X$), any pairwise function $f \in \mathcal{H}_K$ has the form of function difference $f(x,u) = g(x) - g(u)$ with $g \in \mathcal{H}_W$. However, here we do not impose such restriction and will give analysis for general pairwise Mercer kernels. To avoid overfitting, we consider the regularized MEE as follows.

**Definition 2.1** Given a labeled data set $D = \{(x_i,y_i)\}_{i=1}^{\left|D\right|}$, the regularized MEE algorithm with an RKHS $\mathcal{H}_K$ in supervised learning is defined by

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

(4)

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

The efficiency of the regularized MEE (4) in applications has been observed in considerable experimental results and theoretical analysis have been given in Hu et al. (2016); Fan et al. (2016). As a byproduct of our main results, we shall prove that the learning rates of (4) equal the minimax optimal learning rates of the classical pointwise regularized least squares. This greatly improves the results in the literature (we defer the detailed comparison to Section 3).

For simplicity and without loss of generality, we formulate the fully supervised data set $D$ for our distributed MEE algorithm as the union of $k$ independent and equal-sized subsets $D_1, \ldots, D_k$, all drawn independently from $(Z,\rho)$. So,

$$D = \bigcup_{i=1}^{k} D_i.$$

For technical simplicity, in this paper we assume

$$|D_1| = \ldots = |D_k| = \frac{|D|}{k} \geq 4.$$

A local predicted function $f_{D_i,\lambda}$ is obtained from (4) with $D_i$. The **distributed MEE algorithm** outputs its predicted function $\bar{f}_{D,\lambda}$ as the average of the local output functions

$$\bar{f}_{D,\lambda} = \frac{1}{k} \sum_{i=1}^{k} f_{D_i,\lambda}.$$
In this paper, we study the convergence of \( \tilde{f}_{D,\lambda} \) to \( f_0 \) in the square integrable space \( (L_{\rho_X}^2, \| \cdot \|_{\rho}) \), where

\[
L_{\rho_X}^2 \ := \ \left\{ f : \mathcal{X}^2 \rightarrow \mathbb{R} : \| f \|^2_{\rho} := \int_{\mathcal{X}^2} |f(x,u)|^2 d\rho_X(x)d\rho_X(u) < \infty \right\}.
\]

Below we elaborate three important assumptions to carry out the analysis. The first assumption (6) is about the regularity of the target function \( f_0 \). Define the integral operator \( L_K : L_{\rho_X}^2 \rightarrow L_{\rho_X}^2 \) associated with the pairwise kernel \( K \) by

\[
L_K f := \int_{\mathcal{X}} \int_{\mathcal{X}} f(x,u)K(x,u)d\rho_X(x)d\rho_X(u), \quad \forall f \in L_{\rho_X}^2.
\]

Since \( K \) is a Mercer kernel on the compact domain \( \mathcal{X}^2 \), \( L_K \) is of trace class (hence compact) and positive. So, we write \( L_K^r \) as the \( r \)-th power of \( L_K \) for \( r > 0 \). Our error bounds are stated in terms of the regularity of the target function \( f_0(x,u) \), given by

\[
f_0 = L_K^r(h_0), \quad \text{for some } r > 0 \text{ and } h_0 \in L_{\rho_X}^2.
\] (6)

The condition (6) characterizes the regularity of \( f_0 \) and is directly related to the smoothness of \( f_0 \) when \( H_K \) is a Sobolev space. If (6) holds with \( r \geq \frac{1}{2} \), \( f_0 \) lies in the space \( H_K \).

The second assumption (7) is about the capacity of \( H_K \), measured by the effective dimension (Zhang, 2002; Caponnetto and Yao, 2010; Blanchard and Krämer, 2016)

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

where \( I \) is the identity operator on \( H_K \). In this paper, we assume that

\[
\mathcal{N}(\lambda) \leq C_0 \lambda^{-s} \quad \text{for some } C_0 > 0 \text{ and } 0 < s \leq 1.
\] (7)

We postpone some discussions on (7) to Section 3.

The third assumption (8) is about the conditional probability distribution \( \rho(y|\mathcal{X}) \) on the output space \( \mathcal{Y} \). We only assume that the output variable \( Y \) satisfies the moment condition (van der Vaart and Wellner 1996 page 103): there exist two positive numbers \( \sigma, M > 0 \), both independent of \( \mathcal{X} \), such that for any integer \( q \geq 2 \),

\[
E(|Y|^q | \mathcal{X}) \leq \frac{1}{2} q! \sigma^2 M^{q-2}.
\] (8)

The assumption (8) covers many common distributions, for example, Gaussian and the distributions with compact support.

Throughout the paper, we assume that the windowing function \( G \) is differentiable, \( G'(0) = -1 \), and \( G(a) \leq G(0) \) for \( a > 0 \). We assume that

\[
C_G := \sup_{a \in (0,\infty)} |G'(a)| < \infty,
\]

and there exists some \( p, c_p > 0 \) such that

\[
|G'(a) - G'(0)| \leq c_p a^p, \quad \text{for any } a > 0.
\] (9)

For example, the windowing function \( G(a) = e^{-a} \) for Gaussian kernel satisfies the above assumptions with \( c_p = 1 \) and \( p = 1 \).

Since the convergence requires to select \( \lambda \to 0 \) as \( |D| \to \infty \), we assume \( \lambda \leq 1 \) in the sequel to simplify the notations. Without loss of generality, we also assume

\[
\sup_{(x,u) \in \mathcal{X}^2} \sqrt{K((x,u),(x,u))} = 1.
\] (10)
2.1. Convergence of DMEE with fully supervised data

The following theorem bounds the error of (5) with overwhelming probability.

**Theorem 2.2** Assume (6) for \( r > 0 \). For any \( 0 < \delta < 1 \), we have with probability at least \( 1 - \delta \) that,

\[
\| \tilde{f}_{D,\lambda} - f_\rho \|_p \leq \| h_\rho \|_\rho \lambda^{\min(r,1)} + (2\| f_\lambda \|_K + 8M + 8\sigma) A_{D,\lambda,k} \log \frac{8}{\delta} + 128(\| f_\lambda \|_K + M + \sigma) \lambda^{-\frac{1}{2}} \left( \frac{16k}{\delta} \right)^4 \max_{1 \leq i \leq k} \left( \frac{A_{D,\lambda,k}^2}{\lambda^2} + 1 \right) A_{D,\lambda,k}^2 \\
+ 16c_{p,\sigma,M} \left( 1 + \lambda^{p+\frac{1}{2}} \right) h^{-2p} \lambda^{-p-1} \left( \frac{16k}{\delta} \right)^3 \left( \frac{16|D|}{\delta} \right)^{2p+1} \\
\times \left( 1 + \lambda^{-\frac{1}{2}} \max_{1 \leq i \leq k} \left( \frac{A_{D,\lambda,k}^2}{\lambda^2} + 1 \right) A_{D,\lambda} \right) \tag{11}
\]

where the constant \( c_{p,\sigma,M} \) is independent of \( D, \delta, \) or \( h \), and it will be specified later after the bound (27). Here and in the sequel, \( \lfloor D/4 \rfloor \) denotes the largest integer not exceeding \( |D|/4 \), \( A_{D,\lambda,k} := \frac{k}{|D|\sqrt{\lambda}} + \frac{1}{\lfloor D/4 \rfloor \sqrt{\lambda}} + \sqrt{\frac{\lambda(A)}{|D|^2}} \), \( A_{D,\lambda} := A_{D,\lambda,1} \), and \( A_{D,\lambda,k} := A_{D,\lambda,1} \).

**Corollary 2.3** Assume (6) for \( r > 0 \) and (7). Let

\[
\lambda = \begin{cases} 
|D|^{-\frac{r+1}{2}}, & \text{for } 0 < r \leq \frac{1}{2}, \\
|D|^{-\frac{\min(r,1)}{2}}, & \text{for } r > \frac{1}{2},
\end{cases}
\]

and

\[
k = \begin{cases} 
1, & \text{for } 0 < r \leq \frac{1}{2}, \\
\lambda^{-\min\{r-\frac{1}{2},1\}} \log^{-1} |D|, & \text{for } r > \frac{1}{2}. \tag{12}
\end{cases}
\]

Then for any \( 0 < \delta < 1 \), with probability \( 1 - \delta \),

\[
\| \tilde{f}_{D,\lambda} - f_\rho \|_p \leq C_1 \max \left\{ \lambda^{\min(r,1)}, h^{-2p} \lambda^{-p-1} (\log |D|)^{2p+1} \right\} \left( \frac{16}{\delta} \right)^{2p+4}, \tag{13}
\]

where \( C_1 \) is a constant independent of \( D, \delta, k, \) or \( h \), and it will be specified in the proof.

As a direct corollary, the following theorem provides the learning rates for DMEE (and hence MEE) with large \( h \).

**Theorem 2.4** Under the same conditions of Corollary 2.3, if one further has \( h \geq \left[ \lambda^{-r-1-p}(\log |D|)^{2p+4} \right]^{\frac{1}{2p}} \), then with probability at least \( 1 - \delta \),

\[
\| \tilde{f}_{D,\lambda} - f_\rho \|_p \leq C_1 \left\{ \begin{array}{ll} 
|D|^{-\frac{r+1}{2}} \left( \frac{16}{\delta} \right)^{2p+4} & \text{for } 0 < r \leq \frac{1}{2}, \\
|D|^{-\min\left\{ \frac{r+1}{2p+4}, \frac{r-1}{2} \right\}} \left( \log \frac{16}{\delta} \right)^{2p+4} & \text{for } r > \frac{1}{2},
\end{array} \right. \tag{14}
\]

where \( C_1 \) is defined in Corollary 2.3 above. Furthermore, we employ Lemma B.1 in Appendix B to see that for every real number \( \mu > 0 \),

\[
[\mathbb{E}(\| \tilde{f}_{D,\lambda} - f_\rho \|_p)^\mu \leq [16\Gamma(2p+4)+1]^{1/\mu} C_1 \left\{ \begin{array}{ll} 
|D|^{-\frac{r+1}{2}}, & \text{for } 0 < r \leq \frac{1}{2}, \\
|D|^{-\min\left\{ \frac{r+1}{2p+4}, \frac{r-1}{2} \right\}} & \text{for } r > \frac{1}{2},
\end{array} \right. \tag{15}
\]

In particular, since the only assumption (12) on \( k \) permits the case \( k = 1 \), the above bounds (14) and (15) for \( \tilde{f}_{D,\lambda} \) also hold for \( f_{D,\lambda} \).
Remark 2.5 Theorem 2.4 suggests that as \( r \in (0, 1) \) increases, the learning rate is improved. However, further increasing of \( r \) beyond 1 may not help to improve the learning rate. This saturation phenomenon is widely observed in the literature; see e.g. Lo Gerfo et al. (2008); Lin et al. (2017).

Recall the goal of regression analysis is to get a good estimator of \( g^* \). In this work, we aim to learn the difference of the regression function \( f_\rho(x, u) = g^*(x) - g^*(u) \) based on the idea of pairwise learning (Ying and Zhou, 2016, 2017). For this purpose, it is natural to adopt pairwise reproducing kernels, and build the theory in a general way thereupon. To derive an estimator of \( g^* \), we first consider the following orthogonal projection \( \mathcal{P} \) on \( L^2_{\rho_X^2} \) (note that \( \rho_X^2 \) is a probability measure)

\[
(\mathcal{P}f)(x, u) = (\mathcal{M}f)(x) - (\mathcal{M}f)(u), \quad \text{where } (\mathcal{M}f)(x) = \frac{1}{2} \int_X [f(x, u) - f(u, x)] d\rho_X(u).
\]

In particular, if \( f(x, u) = g(x) - g(u) \) for some \( g \in L^2_{\rho_X} \), then \( \mathcal{P}f = f, \mathcal{M}f = g - \int_X g(u) d\rho_X(u) \), and

\[
\int_X (\mathcal{M}f)(x) d\rho_X(x) = 0.
\]

So \( \|\tilde{f}_{D,\lambda} - f_\rho\|_\rho^2 \geq \|\mathcal{P}\tilde{f}_{D,\lambda} - f_\rho\|_\rho^2 = 2 \|\mathcal{M}\tilde{f}_{D,\lambda} + g^* - \mathcal{M}f_\rho\|_2^2 \).

With data \( D \), one replaces \( \mathcal{M} \) by \( \mathcal{M}_D : C(\mathcal{X}^2) \rightarrow C(\mathcal{X}) \),

\[
(\mathcal{M}_D f)(x) := \frac{1}{2|D|} \sum_{(u,v) \in D} [f(x, u) - f(u, x)],
\]

and replaces the difference \( g^* - \mathcal{M}_D f_\rho = \int_X g^*(x) d\rho_X(x) \) by its unbiased and efficient estimator

\[
\frac{1}{|D_l|} \sum_{(x,y) \in D} y.
\]

2.2. Convergence of DMEE with semi-supervised data

We also study the influence of unlabeled data on the convergence of DMEE. Besides the labeled data \( D = \bigcup_{i=1}^k D_l \) (with disjoint and equal-sized subsets \( D_1, \ldots, D_k \)), assume that we also have an unlabeled data set \( \bar{D} = \{\hat{x}_i\}_{i=1}^{[\bar{D}]} \). We assume that the input observations \( \hat{x}_i \) are drawn independently from \( \rho_X \), and \( \bar{D} \) is independent of \( D \). For technical simplicity we assume that \( \bar{D} \) is also divided randomly into \( k \) disjoint and equal-sized subsets \( \bar{D} = \bigcup_{i=1}^k \bar{D}_l \). We define the semi-supervised training data set by \( D^*_l = \bigcup_{i=1}^k D^*_l \), where for each \( 1 \leq l \leq k \), we write \( D_l = \{(x_i, y_i)\}_{i=1}^{[D_l]}, \bar{D}_l = \{\hat{x}_i\}_{i=1}^{[\bar{D}_l]} \), and define \( D^*_l = \{(x^*_i, y^*_i)\}_{i=1}^{[D^*_l]} \) with \( |D^*_l| = |D_l| + |\bar{D}_l| \) by

\[
(x^*_i, y^*_i) = \begin{cases} (x_i, \frac{|D^*_l|}{|D_l|} y_i), & \text{for } 1 \leq i \leq |D_l|, \\ (\hat{x}_{i-|D_l|}, 0), & \text{for } |D_l| + 1 \leq i \leq |D^*_l|. \end{cases}
\]

Here the factor \( |D^*_l|/|D_l| \) for \( y_i \) is given to compensate the bias introduced by the "fake" labels 0 for the unlabeled data.

By "faking" the zero labels, there is no need to reform the algorithm itself. The output function \( f_{D^*,\lambda} \) of the regularized MEE algorithm with semi-supervised data is defined by (4) with \( D \) substituted by \( D^* \). The semi-supervised DMEE outputs the predictive function

\[
\bar{f}_{D^*,\lambda} = \frac{1}{k} \sum_{l=1}^k f_{D^*_l,\lambda}.
\]
**Theorem 2.6** The following bound holds with probability at least $1 - \delta$.

$$
\|\tilde{f}_{D^*,\lambda} - f_{\rho}\|_\rho \leq \|h_{\rho}\|_\rho \lambda^{\min\{r,1\}} + 256(1 + M + \sigma)\lambda^{-1/2} \left(\log \frac{16k}{\delta}\right)^4 \\
\times \max_{1 \leq \ell \leq k} \left(\frac{A_{D^*,\lambda}^2}{\lambda} + 1\right) A_{D^*,\lambda} \left( A_{D^*,\lambda} \|f_{\lambda}\|_K + A_{D_l, D^*,\lambda}\right) \\
+ 16\lambda^{-1/2}C_2h^{-2p} \left(\log \frac{16k}{\delta}\right)^3 \max_{1 \leq \ell \leq k} \left[ \left(\frac{A_{D^*,\lambda}^2}{\lambda} + 1\right) A_{D^*,\lambda}\lambda^{-1/2} + 1\right] \\
\times \Delta_{D_l, D^*,\lambda} \left(\log \frac{16|D|}{\delta}\right)^{2p+1} \\
+ (2A_{D^*,\lambda,K}\|f_{\lambda}\|_K + 8(M + \sigma)A_{D,D^*,\lambda,k}) \log \frac{16}{\delta},
$$

(17)

where $A_{D^*,\lambda}$ is defined in the same way as $A_{D,\lambda}$ in Theorem 2.2 by substituting $D$ with $D^*$, $A_{D,D^*,\lambda,k} = \frac{k}{|D^*|^3/|\lambda|} + \frac{1}{(|D^*|/|\lambda|)^3} + \sqrt{\frac{N(|\lambda)}{|D^*|^3/|\lambda|}}$, $A_{D,D^*,\lambda} = A_{D,D^*,\lambda,1}$, $\Delta_{D,D^*,\lambda} = \left(\frac{|D^*|}{|\lambda| |D|}\right)^{p+\frac{1}{2}} + \left(\frac{|D^*|}{|D|}\right)^{2p+1}$, and $C_2$ is a constant independent of $D$, $D^*$, $k$, $h$, or $\delta$, and it will be specified in the proof.

To demonstrate the idea of Theorem 2.6, we give the following corollary, which suggests that, with sufficient unlabeled data, the number $k$ of the computing nodes is only technically bounded from above by the assumption $|D_l| = \frac{|D|}{k} \geq 4$. Similar results for distributed regularized least squares are obtained in Lin and Zhou (2018). Note that this increase of computing nodes does not help to reduce single-node time or space complexity, but significantly improves the learning rates under the scenario that the regression function $f_{\rho}$ has low regularity $0 < r < \frac{1}{2}$, therefore our analysis suggests the semi-supervised scheme for learning less regular target functions. Another scenario of such learning rate improvement is when locally stored data can not be centralized due to privacy concerns, and therefore our analysis suggests an “inflation of unlabeled data” solution to DMEE for privacy-sensitive distributed learning. We will elaborate the details in Section 3.

**Corollary 2.7** Assume (8), (9), (6) for $r > 0$, $r + s \geq \frac{1}{2}$, and $|D^*| \geq \max\{|D|^{\min\{r,1\}}/s, 2|D|\}$. Let $\lambda = |D|^{-\frac{r + \frac{1}{2}}{\min\{r,1\} - s}}$, and

$$
k \leq (\log |D|)^{-4} \min \left\{ \sqrt{|D^*|^{1+s}}, (|D^*| \lambda^{2 - 2\min\{r,1\} - s})^{1/3} \right\}.
$$

(18)

Then with probability at least $1 - \delta$, one has

$$
\|\tilde{f}_{D^*,\lambda} - f_{\rho}\|_\rho \leq C_3 \max \left\{ |D|^{-\min\{r + \frac{1}{2}, \frac{1}{\min\{r,1\} - s}\}}, \frac{\Delta_{D,D^*,\lambda} (\log |D|)^{2p+4}}{\sqrt{l_h^{2p}}} \right\} \log \left(\frac{16}{\delta}\right)^{2p+4},
$$

where $C_3$ is a constant independent of $D$, $D^*$, $k$, $h$, or $\delta$, and will be specified in the proof.

**Theorem 2.8** Under the same conditions of Corollary 2.7, if one further has

$$
h \geq \left[ |D|^{\frac{1}{2p+4}} \Delta_{D,D^*,\lambda} (\log |D|)^{2p+4} \right]^\frac{1}{p+1},
$$

(19)
then with probability $1 - \delta$,

$$
\|\bar{f}_{D^*,\lambda} - f_\rho\|_{\rho} \leq C_3|D|^{-\min\{\frac{1}{2\rho\epsilon}, \frac{1}{\rho}\}} \left(\log \frac{16}{\delta}\right)^{2p+4},
$$

(20)

where $C_3$ is the constant defined in Corollary (2.7). Furthermore, we employ Lemma B.1 in Appendix B to see that for any real number $\mu > 0$,

$$
\left[\mathbb{E} \left(\|\bar{f}_{D^*,\lambda} - f_\rho\|_{\rho}\right)^{1/\mu}\right] \leq \left[16\Gamma((2p+4)\mu + 1)\right]^{1/\mu} C_3|D|^{-\min\{\frac{1}{2\rho\epsilon}, \frac{1}{\rho}\}}.
$$

(21)

In particular, since the only assumption (18) on $k_{perm}$ permits the case $k = 1$, the above bounds (20) and (21) for $\bar{f}_{D^*,\lambda}$ also holds for $f_{D^*,\lambda}$.

3. Discussion and comparison with other works

The condition (7) is widely adopted in the literature to characterize the capacity of $\mathcal{H}_K$ (Lin et al., 2017; Caponnetto and De Vito, 2007; Zhang, 2002; Blanchard and Krämer, 2010). We see that since $L_K$ is of trace class, the condition (7) always holds with $s = 1$. When the eigenvalues of $L_K$ decay faster, one can have (7) with a smaller $s$. In particular, if the eigenvalues $\{\gamma_i\}_{i=1}^\infty$ of the operator $L_K$ decay as $\gamma_i \leq C_0 i^{-b}$ for some $0 < b < 1$ and $C_0 > 0$, then

$$
N(\lambda) = \sum_{i=1}^\infty \frac{\gamma_i}{\gamma_i + \lambda} \leq \int_0^\infty \frac{C_0}{C_0 + \lambda t^{\frac{1}{b}}} dt \leq 2^\frac{1}{b} C_0^b \int_0^\infty \frac{1}{(1 + t)^{b}} dt = \frac{2^\frac{1}{b} C_0^b}{1 - b} \lambda^{-b}.
$$

The condition (7) roughly measures the smoothness of $K$. For example, if $K \in C^\alpha(X^2 \times X^2)$ with some integer $\alpha \geq 1$, and $X^2$ is locally the graph of a Lipschitz function, then (7) is satisfied with $s = \left(\frac{\alpha}{2\min(\lambda_1, \lambda_2)} + \frac{1}{2}\right)^{-1}$ (Mendelson and Neeman, 2010). There are some other capacity characteristics, for example covering numbers (Zhou, 2002) and entropy numbers (Steinwart et al., 2009). Compared to the regularity assumptions, capacity assumption is not necessary for deriving learning rates, and there are works on capacity independent analysis in the literature; see e.g. Smale and Zhou (2007).

We organize the following discussions around Figure 1. In Area 1, our analysis suggests the saturation phenomenon of DMEE with respect to regularity, as shared also by regularized least
apply to non-distributed MEE, which corresponds to the case $k = 1$, and they improve some existing results in the literature. Most studies on MEE algorithms in the literature are carried out empirically. Theoretical results are relatively sparse. In Chen et al. (2010), the consistency of MEE is proved in the local region and no explicit learning rate was given. In our earlier works (Hu et al., 2015; Fan et al., 2016), we studied MEE algorithms in the empirical risk minimization (ERM) and regularized ERM frameworks respectively. The main results include that if the target function lies in the hypothesis space $H$, the learning rate of fully supervised DMEE is $O(|D|^{-\frac{1}{2}})$ for pointwise regression learning (Steinwart et al., 2009; Caponnetto and De Vito, 2007; Bauer et al., 2007; Blanchard and Mücke, 2018). Without spoiling the learning rates, unlabeled data help to essentially remove the restriction of the maximum number of computing nodes DMEE can be distributed to. Note that, while allowing a “more distributed” computation, unlabeled data may increase the single-node computational complexity and memory requirement. This being said, the removal of the restriction on maximum computing nodes has significant impact on the applications where for privacy reasons, data cannot be centralized and must be processed locally. For these applications, computational complexity and memory requirement are usually not the concern, and our analysis suggests a solution of inflating data subsets with unlabeled data to improve generalization power. Similar observations are reported for pointwise regularized least squares (Lin and Zhou, 2018).

For fully supervised data, Theorem 2.4 suggests that without spoiling the learning rate, DMEE can at most reduce the single-node data size, from $|D|$ to $|D|^{1 - \frac{1}{2} \log^4 |D|}$ for Area 2, and to $|D|^{1 - \frac{1}{2} \log^4 |D|}$ for Area 1, respectively. However, for Area 3 and Area 4, Theorem 2.4 suggests that DMEE could not indeed reduce single-node computational complexity without sacrificing the learning rates. For the semi-supervised data, unlabeled data serve mainly for the purpose of relaxing the restriction of the maximum number of computing nodes for privacy protection, and improving the learning performance under weak regularity conditions. In fact, for Areas 1 and 2, our analysis suggests that to maintain the best possible learning rates that DMEE can achieve with fully supervised data, if one relaxes the restriction of the maximum number of computing nodes by adding unlabeled data, then the single-node sample size will be increased.

In Area 3, the optimal lower bound $O(|D|^{-\frac{1}{2}})$ for point kernel-based regression for $0 < r < \frac{1}{2}$ is derived in Steinwart et al. (2009) with the boundedness assumption of $L_K: L_2 \rightarrow L_\infty$. Fully supervised DMEE does not achieve this lower bound. Unlabeled data improves DMEE in two ways. First, semi-supervised DMEE achieves the learning rates $O(|D|^{-\frac{1}{2}})$. Second, again, semi-supervised DMEE has essentially no restriction on the maximum number of computing nodes. The coverage of Area 3 is also one of the important improvements we have in this paper, compared with Hu et al. (2020), and even Lin and Zhou (2018).

In Area 4, the learning rate of fully supervised DMEE is $O(|D|^{-\frac{1}{2}})$, and our analysis of semi-supervised DMEE fails to improve the rate to $O(|D|^{-\frac{1}{2}})$. The learning rates are only provided for fully supervised DMEE. Nevertheless, Area 4 seems to be the situation that one should avoid. In fact, with a less regular target function, typically one needs a larger hypothesis space, which corresponds to a larger $s$. It is unknown to us at this moment whether the suboptimal rates in Areas 3 for fully supervised DMEE and in Area 4 for semi-supervised DMEE are the inherent features of these algorithms or the consequences as limited by the analysis tools. It will be an interesting future research topic.

It is worth mentioning that all the error bounds and convergence rates obtained in this paper apply to non-distributed MEE, which corresponds to the case $k = 1$, and they improve some existing results in the literature.
by $C(X)$ balls of radius $\epsilon$ grows no faster than $\epsilon^{-p}$ for some index $p > 0$ when $\epsilon$ decays to zero, then with high probability the learning rate is of order $O\left(|D|^{-\frac{1}{2p}}\right)$. To elaborate it clearly, assume $H$ to be an RKHS induced either by a pointwise kernel $W \in C^\alpha(X \times X)$ or a pairwise kernel $K \in C^\alpha(X^2 \times X^2)$ and $X \subset \mathbb{R}^d$ satisfied certain mild regularity conditions. By Cucker and Zhou (2007, page 72, Theorem 5.1) and Mendelson and Neeman (2010), the covering number index $K$ is satisfied certain mild regularity conditions. By Cucker and Zhou (2007, page 72, Theorem 5.1) and Mendelson and Neeman (2010), the covering number index $K$ is always inferior to $C^{\alpha}$. When target function lies in $H_K$, i.e. $r \geq \frac{1}{2}$, the rate $O\left(|D|^{-\frac{1}{2p}}\right)$ in (Hu et al., 2015) is always inferior to $O\left(|D|^{-\frac{1}{2p}}\right)$ in (14).

To our best knowledge, various MEE algorithms in the existing literature are implemented by gradient descent-based methods. Note that in general, Algorithm (4) is not a convex optimization problem, and a comprehensive discussion about the global/local optimal, the dependence of convergence on the initial value of variables, and the convergence speed, go beyond the scope of this paper, and are interesting questions for future research. While our analysis is given for general pairwise reproducing kernels, the design (3) is usually adopted in practice (Ying and Zhou 2016, 2017). We point out that under the design (3), Algorithm (4) is reduced to a smooth (though not convex) optimization problem with $|D| - 1$ variables. In fact, for any $x \in X$, write $W(x) = (W(x, x_i))_{i=1}^{|D|}$ a column vector of dimension $|D|$. Write $C = (c_{i,j})_{i,j=1}^{D}$ the coefficient matrix of the function

$$f_C(x,u) = \sum_{i=1}^{D} \sum_{j=1}^{D} c_{i,j} K((x_i, x_j), (x, u)).$$

By the representor theorem, the solution to (4) takes the form of (22). Meanwhile, we write $1 = (1, \ldots, 1)^T \in \mathbb{R}^{D}$ and use the kernel structure (3) to obtain that $f_C(x,u) = c^T(W(x) - W(u))$, where $c = (C - CT)1 \in \mathbb{R}^{D}$. We also have $\|c\|_K^2 = c^T \mathbb{W} c$, where $\mathbb{W} = (W(x_i, x_j))_{i,j=1}^{D}$. So, Algorithm (3) is reduced to an optimization problem of a smooth function of $c$. Since $1^T c = 0$, the vector $c$ has only $|D| - 1$ free variables. The gradient vector and the Hessian matrix of the target function can be directly computed. The computational complexity can further be reduced for DMEE.

4. Estimates in supervised learning

Now we are in position to prove the consistency results stated in Section 2. First, we will estimate the bound of $f_{D,\lambda}$ defined by (4). In the sequel, for notational simplicity, write $w = (x, y)$ and $z = (u, v)$. Define the empirical operator $L_{K,D} : H_K \rightarrow H_K$ by

$$L_{K,D} := \frac{1}{|D|^2} \sum_{w, z \in D} \langle \cdot, K(w,u) \rangle K(w,u),$$

so for any $f \in H_K$, $L_{K,D} f = \frac{1}{|D|^2} \sum_{w, z \in D} f(w, u)K(w,u)$ . Then we have the following representation for $f_{D,\lambda}$.

**Lemma 4.1** Define $f_{D,\lambda}$ by (4). Then it satisfies

$$f_{D,\lambda} = (L_{K,D} + \lambda I)^{-1} \hat{f}_{p,D} + (L_{K,D} + \lambda I)^{-1} E_{D,\lambda}$$

where

$$\hat{f}_{p,D} = \frac{1}{|D|^2} \sum_{w, z \in D} (y - v) K(w,u)$$

and

$$E_{D,\lambda} = \frac{1}{|D|^2} \sum_{w, z \in D} \langle \cdot, K(w,u) \rangle K(w,u).$$
and

\[ E_{D,\lambda} = \frac{1}{|D|^2} \sum_{w,z \in D} \left[ G' \left( \frac{(f_{D,\lambda}(x,u) - y + v)^2}{h^2} \right) - G'(0) \right] (f_{D,\lambda}(x,u) - y + v)K(x,u). \]

**Proof.** Since \( f_{D,\lambda} \) is the minimizer of algorithm (4), we take the gradient of the regularized functional on \( \mathcal{H}_K \) in (4) to give

\[-\frac{1}{|D|^2} \sum_{w,z \in D} G' \left( \frac{(f_{D,\lambda}(x,u) - y + v)^2}{h^2} \right) (f_{D,\lambda}(x,u) - y + v)K(x,u) + \lambda f_{D,\lambda} = 0,\]

or equivalently (recall the assumption \( G'(0) = -1 \)),

\[ \frac{1}{|D|^2} \sum_{w,z \in D} (f_{D,\lambda}(x,u) - y + v)K(x,u) + \lambda f_{D,\lambda} - E_{D,\lambda} = 0, \]

which is \( (L_{K,D} + \lambda I) f_{D,\lambda} - \hat{f}_{\rho,D} - E_{D,\lambda} = 0 \). The proof is completed. \( \square \)

### 4.1. Bounds of \( f_{D,\lambda} \) and \( E_{D,\lambda} \)

Under the moment condition (8), similar to (Wang and Hu, 2019 Proposition 3) we can prove that, with probability at least \( 1 - \delta \), there holds

\[ \max \{|y| : \text{there exists an } x \in \mathcal{X}, \text{ such that } (x,y) \in D} \right) \leq (4M + 5\sigma) \log \frac{|D|}{\delta}. \] (24)

By the definition of \( f_{D,\lambda} \) in (4), we have that \( \mathcal{E}_D(f_{D,\lambda}) + \lambda \|f_{D,\lambda}\|_K^2 \leq \mathcal{E}_D(0) \). Recall that \( C_G = \sup_a |G'(a)| \). With the fact \( G(a) < G(0) \) for all \( a > 0 \) and Taylor expansion,

\[ \lambda \|f_{D,\lambda}\|_K^2 \leq \mathcal{E}_D(0) - \mathcal{E}_D(f_{D,\lambda}) \leq \frac{h^2}{|D|^2} \sum_{w,z \in D} G \left( \frac{(y - v)^2}{h^2} \right) + h^2G(0) \]

\[ \leq \frac{C_G}{|D|^2} \sum_{w,z \in D} (y - v)^2 \leq \frac{C_G}{|D|^2} \sum_{w,z \in D} 2(y^2 + v^2) \leq 4C_G \max_{w \in D} |y|^2. \]

It follows that

\[ \|f_{D,\lambda}\|_K \leq 2\sqrt{C_G} \lambda^{-\frac{1}{2}} \max_{w \in D} |y|. \] (25)

By (9), we see that

\[ \|E_{D,\lambda}\|_K \leq c_p h^{-2p} \frac{1}{|D|^2} \sum_{w,z \in D} (\|f_{D,\lambda}\|_K + |y - v|)^{2p+1} \]

\[ \leq 2^{2p}c_p h^{-2p} \frac{1}{|D|^2} \sum_{w,z \in D} \left( \|f_{D,\lambda}\|_K^{2p+1} + |y - v|^{2p+1} \right) \]

\[ \leq 2^{2p}c_p h^{-2p} \left( \|f_{D,\lambda}\|_K^{2p+1} + 2^{2p+1} \max_{w \in D} |y|^{2p+1} \right). \]

This in combination with the bounds (24) and (25) gives that, with probability at least \( 1 - \delta \),

\[ \|E_{D,\lambda}\|_K \leq c_{p,\sigma,\lambda} \left( \lambda^{-(p+\frac{1}{2})} + 1 \right) h^{-2p} \left( \log \frac{|D|}{\delta} \right)^{2p+1} \] (27)

where \( c_{p,\sigma,\lambda} := 2^{4p+1}c_p(C_G^{p+\frac{1}{2}} + 1)(4M + 5\sigma)^{2p+1} \).
4.2. Two error decompositions in MEE algorithms

To derive the explicit learning rate of the distributed algorithm (5) and (16), we introduce the regularization function \( f_\lambda \) in \( H_K \), defined by

\[
f_\lambda := \arg \min_{f \in H_K} \mathcal{E}_b(f) + \lambda \|f\|_K^2,
\]

where \( \mathcal{E}_b(f) = \int_{\mathbb{R}^2} (f(x,u) - y + v)^2 dp(x,y) dp(u,v) \) is the expected risk associated with the pairwise square loss. Similar to the argument in Smale and Zhou (2007), we can verify that

\[
f_\lambda = (L_K + \lambda I)^{-1} L_K f_\rho,
\]

so \( f_\lambda - f_\rho = -\lambda (L_K + \lambda I)^{-1} f_\rho \). We have used the property that the operator norm of \( L_K \) on \( L^2_{\rho} \) is no greater than 1, thanks to the assumption (10). Under the regularity assumption (6) with \( r > 0 \),

\[
\|f_\lambda - f_\rho\|_\rho \leq \begin{cases} \|h_\rho\|_\rho \lambda^r, & \text{when } 0 < r \leq 1, \\ \|h_\rho\|_\rho \lambda, & \text{when } r > 1,
\end{cases}
\]

and

\[
\|f_\lambda\|_K \leq \begin{cases} \|h_\rho\|_\rho \lambda^{r - \frac{1}{2}}, & \text{when } 0 < r < 1/2, \\ \|h_\rho\|_\rho, & \text{when } r \geq 1/2.
\end{cases}
\]

Now we state two error decompositions for \( f_{D,\lambda} - f_\lambda \). By (28), \(-L_{K,D} f_\lambda - \lambda f_\lambda = -L_{K,D} f_\lambda + L_K f_\lambda - L_K f_\rho \), so

\[
-f_\lambda = (L_{K,D} + \lambda I)^{-1} [(L_K - L_{K,D}) f_\lambda - L_K f_\rho],
\]

and we obtain the first decomposition by adding (23) and (31),

\[
f_{D,\lambda} - f_\lambda = (L_{K,D} + \lambda I)^{-1} [(L_K - L_{K,D}) f_\lambda + (L_{K,D} + \lambda I)^{-1} (\hat{f}_{\rho,D} - L_K f_\rho)] + (L_{K,D} + \lambda I)^{-1} E_{D,\lambda}.
\]

Recall that \( \lambda f_{D,\lambda} = -L_{K,D} f_{D,\lambda} + \hat{f}_{\rho,D} + E_{D,\lambda}, \) so

\[
(L_K + \lambda I) f_{D,\lambda} = (L_K - L_{K,D}) (f_{D,\lambda} - f_\lambda) + (L_K - L_{K,D}) f_\lambda + \hat{f}_{\rho,D} + E_{D,\lambda},
\]

and we obtain the second decomposition

\[
f_{D,\lambda} - f_\lambda = f_{D,\lambda} - (L_K + \lambda I)^{-1} L_K f_\rho = (L_K + \lambda I)^{-1} [(L_K + \lambda I) f_{D,\lambda} - L_K f_\rho] \\
= (L_K + \lambda I)^{-1} [(L_K - L_{K,D}) (f_{D,\lambda} - f_\lambda) + (L_K + \lambda I)^{-1} (L_K - L_{K,D}) f_\lambda + \hat{f}_{\rho,D} - L_K f_\rho] + (L_K + \lambda I)^{-1} E_{D,\lambda}.
\]

In the sequel, we denote by \( \| \cdot \|_{\text{op}} \) the operator norm from \( H_K \) to itself, and

\[
\mathcal{B}_{D,\lambda} = \|(L_{K,D} + \lambda I)^{-1} (L_K + \lambda I)\|_{\text{op}},
\]

\[
\mathcal{C}_{D,\lambda} = \|(L_K + \lambda I)^{-\frac{1}{2}} (L_K - L_{K,D})\|_{\text{op}},
\]

\[
\mathcal{D}_{D,\lambda} = \left\| \frac{1}{k} \sum_{i=1}^{k} (L_K + \lambda I)^{-\frac{1}{2}} (L_K - L_{K,D_i}) \right\|_{\text{op}},
\]

\[
\mathcal{F}_{D,\lambda} = \left\| \frac{1}{k} \sum_{i=1}^{k} (L_K + \lambda I)^{-\frac{1}{2}} (\hat{f}_{\rho,D_i} - L_K f_\rho) \right\|_{K},
\]

\[
\mathcal{G}_{D,\lambda} = \|(L_K + \lambda I)^{-\frac{1}{2}} (\hat{f}_{\rho,D} - L_K f_\rho)\|_{K}.
\]

We cite the following lemma from Blanchard and Krämer (2010 Lemma E.4), which was proved for positive definite matrices in Bhatia (1997, pages 255–256, Theorems IX.2.1-2).
Lemma 4.2  Let $A$ and $B$ be positive definite operators on a separable Hilbert space $\mathcal{H}$. Let $\| \cdot \|_{\text{op}(\mathcal{H})}$ denote the operator norm. Then

$$\|A^sB^s\|_{\text{op}(\mathcal{H})} \leq \|AB\|_{\text{op}(\mathcal{H})}^s, \quad \text{for any } 0 \leq s \leq 1.$$  

Noting that for any $f \in H_K$,

$$\max\{\|f\|, \sqrt{\lambda}\|f\|_K\} \leq \|(L_K + \lambda I)^{1/2}f\|_K$$

by the fact $\|f\|_\rho = \|L_K^{1/2}f\|_K$, one gets a bound for the sample error $\|f_{D,\lambda} - f_\lambda\|_\rho$ by the two decompositions (32) and (33) above.

Proposition 4.3  Define $\tilde{f}_{D,\lambda}$ by (5). Then there holds

$$\|\tilde{f}_{D,\lambda} - f_\lambda\|_\rho \leq S_1 + S_2 + D_{D,\lambda}\|f_\lambda\|_K + F_{D,\lambda},$$

where

$$S_1 = \max_{1 \leq l \leq k} \left( B_{D_l,\lambda}C_{D_l,\lambda}\|f_\lambda\|_K \lambda^{-\frac{1}{2}} + B_{D_l,\lambda}C_{D_l,\lambda}G_{D_l,\lambda}\lambda^{-\frac{3}{2}} \right)$$

and

$$S_2 = \max_{1 \leq l \leq k} \left( B_{D_l,\lambda}C_{D_l,\lambda}\lambda^{-1}\|E_{D_l,\lambda}\|_K + \frac{1}{\sqrt{\lambda}}\|E_{D_l,\lambda}\|_K \right).$$

Proof. Let $I_1$, $I_2$, and $I_3$ denote the three terms on the right-hand side of (32), respectively. Consider the $H_K$ norm of

$$(L_K + \lambda I)^{1/2}(f_{D,\lambda} - f_\lambda) = (L_K + \lambda I)^{1/2}(I_1 + I_2 + I_3).$$

By Lemma 4.2,

$$\|(L_K + \lambda I)^{1/2}I_1\|_K$$

$$\leq \|(L_K + \lambda I)^{1/2}(L_{K,D} + \lambda I)^{-1/2}\|_{\text{op}}\|(L_{K,D} + \lambda I)^{-1/2}(L_K + \lambda I)^{1/2}\|_{\text{op}}$$

$$\times \|(L_K + \lambda I)^{-1/2}(L_K - L_{K,D})\|_{\text{op}}\|f_\lambda\|_K$$

$$\leq B_{D,\lambda}C_{D,\lambda}\|f_\lambda\|_K.$$ 

Similarly,

$$\|(L_K + \lambda I)^{1/2}I_2\|_K$$

$$\leq \|(L_K + \lambda I)^{1/2}(L_{K,D} + \lambda I)^{-1}(L_K + \lambda I)^{1/2}\|_{\text{op}}\|(L_K + \lambda I)^{-1/2}(\tilde{f}_\rho,D - L_K f_\rho)\|_K$$

$$\leq B_{D,\lambda}G_{D,\lambda},$$

and

$$\|(L_K + \lambda I)^{1/2}I_3\|_K$$

$$\leq \|(L_K + \lambda I)^{1/2}(L_{K,D} + \lambda I)^{-1}(L_K + \lambda I)^{1/2}\|_{\text{op}}\frac{1}{\sqrt{\lambda}}\|E_{D,\lambda}\|_K$$

$$\leq \lambda^{-1/2}B_{D,\lambda}\|E_{D,\lambda}\|_K.$$ 

With the above bounds, we use (34) to obtain

$$\|f_{D,\lambda} - f_\lambda\|_\rho \leq B_{D,\lambda}C_{D,\lambda}\|f_\lambda\|_K + B_{D,\lambda}G_{D,\lambda} + \lambda^{-\frac{1}{2}}B_{D,\lambda}\|E_{D,\lambda}\|_K.$$
and
\[ \|f_{D,\lambda} - f_\lambda\|_K \leq B_{D,\lambda}C_{D,\lambda}\|f_\lambda\|_K\lambda^{-\frac{1}{2}} + B_{D,\lambda}G_{D,\lambda}\lambda^{-\frac{1}{2}} + \lambda^{-1}B_{D,\lambda}\|E_{D,\lambda}\|_K. \] (36)

By the fact \( \bar{f}_{D,\lambda} - f_\lambda = \frac{1}{k} \sum_{l=1}^{k} (f_{D,l} - f_\lambda) \) and the second decomposition (33), one obtains that
\[
\|\bar{f}_{D,\lambda} - f_\lambda\|_\rho \leq \frac{1}{k} \sum_{l=1}^{k} \|(L_K + \lambda I)^{-\frac{1}{2}}(L_K - L_{K,D})(f_{D,l} - f_\lambda)\|_K
\]
\[+ \frac{1}{k} \sum_{l=1}^{k} (L_K + \lambda I)^{-\frac{1}{2}}(L_K - L_{K,D})f_\lambda\|_K
\]
\[+ \frac{1}{k} \sum_{l=1}^{k} (L_K + \lambda I)^{-\frac{1}{2}}(\hat{f}_{\rho,D_l} - L_K f_\rho)\|_K + \lambda^{-\frac{1}{2}} \frac{1}{k} \sum_{l=1}^{k} \|E_{D,l}\|_K
\]
\[\leq D_{D,\lambda}\|f_\lambda\|_K + F_{D,\lambda} + \max_{1 \leq l \leq k} \left( C_{D,l}\|f_{D,l} - f_\lambda\|_K + \lambda^{-\frac{1}{2}}\|E_{D,l}\|_K \right).
\]

Plugging (36) into the above bounds (with substitution of \( D \) by \( D_l \)) completes the proof. \( \blacksquare \)

### 4.3. Estimates in distributed U-statistics

To present the learning power of the algorithm (5), we will make use of Proposition 4.3 that is related to the quantities \( B_{D,\lambda}, C_{D,\lambda}, D_{D,\lambda}, F_{D,\lambda} \) and \( G_{D,\lambda} \). By the work in Hu et al. (2020), we can see that the following bounds hold.

**Proposition 4.4** Each of the following three bounds holds with probability \( 1 - \delta \).
\[
B_{D,\lambda} \leq 2 \left( \frac{2A_{D,\lambda}\log \frac{2}{\delta}}{\sqrt{\lambda}} \right)^2 + 2, \quad C_{D,\lambda} \leq 2A_{D,\lambda}\log \frac{2}{\delta}, \quad \text{and}
\]
\[
D_{D,\lambda} \leq 2A_{D,\lambda,k}\log \frac{2}{\delta}.
\]

For other two quantities \( F_{D,\lambda} \) and \( G_{D,\lambda} \), they are both involved with the unbounded condition (8), which brings difficulties in pairwise distributed concentration inequalities. We will handle them by some decomposition techniques in U-statistics.

**Proposition 4.5** Each of the following two bounds holds with probability \( 1 - \delta \).
\[
F_{D,\lambda} \leq 8(M + \sigma)A_{D,\lambda,k}\log \frac{2}{\delta}, \quad \text{and} \quad G_{D,\lambda} \leq 8(M + \sigma)A_{D,\lambda}\log \frac{2}{\delta}.
\]

**Proof.** Define a random variable
\[
\xi(w, z) = (y - v)(L_K + \lambda I)^{-\frac{1}{2}}K_{(x,u)}
\]
with \( w = (x, y) \) and \( z = (u, v) \). The moment condition (8) implies that,
\[
\mathbb{E}\|\xi\|_K \leq 2\lambda^{-\frac{1}{2}}\mathbb{E}|Y| \leq 2\lambda^{-\frac{1}{2}}(\mathbb{E}|Y|^2)^{\frac{1}{2}} \leq 2\sigma\lambda^{-\frac{1}{2}}.
\] (37)

One applies Hőlder’s inequality to have that for any \( q \geq 2 \), \( (\mathbb{E}\|\xi\|_K)^q \leq \mathbb{E}(\|\xi\|_K^q) \). Note the equation \( \mathbb{E}\|\bar{L}_K + \lambda I\|^{-1/2}K_{(x,u)}\|_K^2 = \mathcal{N}(\lambda) \) for \( \lambda > 0 \) (Lin et al., 2017 Lemma 18). We obtain the following
bound for any integer $q \geq 2$.

$$
\mathbb{E}[\|\xi - \mathbb{E}\|_K^q] \leq 2^{q-1}\mathbb{E}[\|\xi\|_K^q] + 2^{q-1}\mathbb{E}\|\xi\|_K^q \leq 2^q\mathbb{E}[\|\xi\|_K^q]
$$

$$
\leq 2^{2q} \sup_{x',u' \in X} \| (L_K + \lambda I)^{-1/2} K(x',u') \|_K^{q-2} \mathbb{E} \left[ |Y_1|^q \| (L_K + \lambda I)^{-1/2} K(x,u) \|_K^2 \right]
$$

$$
\leq 2^{2q} \left( \frac{1}{\sqrt{\lambda}} \right) \mathbb{E} \left[ \left( |Y_1|^q \right) \| (L_K + \lambda I)^{-1/2} K(x,u) \|_K^2 \right]
$$

$$
\leq 2^{2q} \left( \frac{1}{\sqrt{\lambda}} \right)^{q-2} \frac{1}{2}\sigma^2 M^{q-2} N(\lambda)
$$

$$
= 8q\sigma^2 N(\lambda)(4M/\sqrt{\lambda})^{q-2}.
$$

Let $\pi$ be a permutation of the set $\{1, \ldots, |D_i|\}$ of integers. Then $\{z_{i(1)}^l, \ldots, z_{i(|D_i|)}^l\}$ is the associate permutation of $D_i = \{z_i^l = (x_i^l, y_i^l)\}_{i=1}^{\|D_i\|}$. Let

$$
U_\pi^l = \frac{1}{|\{D_i\}/2|} \sum_{i=1}^{\|D_i\|/2} \xi(z_{\pi(2i-1)}^l, z_{\pi(2i)}^l).
$$

Since $\xi(z, z) = 0$, the average $(L_K + \lambda I)^{-1/2} \hat{J}_{\rho, D_i} = \frac{|D_i|}{|D_i|!} \sum_{w, z \in D_i} \xi(w, z)$ can be written as

$$
(L_K + \lambda I)^{-1/2} \hat{J}_{\rho, D_i} = \frac{|D_i| - 1}{|D_i|!} \left( \frac{1}{|D_i|(|D_i| - 1)} \sum_{u, z \in D_i} \xi(u, z) \right) + \frac{1}{|D_i|!} \sum_{\pi} U_\pi^l,
$$

where the last sum is taken over all the $|D_i|!$ permutations $\pi$ of $\{1, \ldots, |D_i|\}$. Since $|D_1| = \ldots = |D_k|$

$$
\frac{1}{k} \sum_{i=1}^{k} (L_K + \lambda I)^{-1/2} \hat{J}_{\rho, D_i} = \frac{|D_i| - 1}{|D_i|!} \frac{1}{|D_i|!} \sum_{\pi} \frac{1}{k} \sum_{i=1}^{k} U_\pi^l.
$$

By definition, it is easy to see that $\mathbb{E}[\xi] = (L_K + \lambda I)^{-1/2} L_K f_\rho$. One applies (37) to obtain

$$
\mathcal{F}_{D, \lambda} \leq \frac{|D_i| - 1}{|D_i|!} \left( \frac{1}{|D_i|!} \sum_{\pi} \frac{1}{k} \sum_{i=1}^{k} U_\pi^l - \mathbb{E}\xi \right) + \frac{1}{|D_i|!} \mathbb{E}\|\xi\|_K
$$

$$
\leq \frac{1}{|D_i|!} \sum_{\pi} \frac{1}{k} \sum_{i=1}^{k} U_\pi^l - \mathbb{E}\xi + \frac{2\sigma^2}{|D|/\sqrt{\lambda}}.
$$

(38)

We observe that for each $\pi$, $\frac{1}{k} \sum_{i=1}^{k} U_\pi^l$ is the average of $k|\{D_i\}/2|$ independent copies of $\xi(z, z')$ with $z$ and $z'$ independently drawn from $\rho$, and $k|\{D_i\}/2| \geq |\{D_i\}/4|$ (in fact, recall our assumption $|D_1| \geq 4$ and $k|\{D_i\}/2| = |\{D_i\}/4|$, obviously when $|D_i|$ is even, $k|\{D_i\}/2| = k|\{D_i\}/2| = |\{D_i\}/2| \geq \frac{|\{D_i\}/4|}{2}$, and when $|D_i|$ is odd, one still has $k|\{D_i\}/2| = k((|D_i| - 1)/2 \geq k|\{D_i\}/4| \geq \frac{|\{D_i\}/4|}{2}$). Note that the hyperbolic function $\cosh(x) = (e^x + e^{-x})/2$ is convex. To estimate the first term on the right-hand
side of (38), we apply Lemma A.2 to obtain that for any $\epsilon > 0$,
\[
\text{Prob} \left\{ \left\| \frac{1}{|D_i|} \sum_{i} \frac{1}{k} \sum_{i=1}^{k} U^l_i - \mathbb{E} \xi \right\|_K \geq \epsilon \right\} 
\]
\[
\leq \inf_{c > 0} \mathbb{E} \left[ \cosh \left( c \left\| \frac{1}{|D_i|} \sum_{i} \frac{1}{k} \sum_{i=1}^{k} U^l_i - \mathbb{E} \xi \right\|_K \right) \right] \cosh(\epsilon c)
\]
\[
\leq \inf_{c > 0} \frac{1}{|D_i|} \sum_{i} \mathbb{E} \left[ \cosh \left( c \left\| \frac{1}{k} \sum_{i=1}^{k} U^l_i - \mathbb{E} \xi \right\|_K \right) \right] \cosh(\epsilon c)
\]
\[
\leq 2 \exp \left\{ - \frac{|\|D_i\|/4\| \epsilon^2}{2 \left( 16N(\lambda)\sigma^2 + 4\lambda^{-\frac{\delta}{2}}M\epsilon \right)} \right\}.
\]
(39)

One takes the right-hand side of (39) as $\delta$ and recalls (38) to have that with probability $1 - \delta$,
\[
\mathcal{F}_{D,\lambda} \leq \frac{2\sigma k}{|D|\sqrt{\lambda}} + \frac{8M \log(2/\delta)}{|\|D_i\|/4\| \sqrt{\lambda}} + \sqrt{\frac{32N(\lambda)\sigma^2 \log(2/\delta)}{|\|D\|/4|}}.
\]

Note that $\mathcal{G}_{D,\lambda}$ equals $\mathcal{F}_{D,\lambda}$ when $k = 1$. The proof is complete.

4.4. Proof of learning rates in supervised learning

Proof of Theorem 2.2 We can decompose $\|f_{D,\lambda} - f_\rho\|$ as the sample error $\|f_{D,\lambda} - f_\rho\|$ and the approximation error $\|f_\rho - f_\rho\|$. As stated in (29), $\|f_\rho - f_\rho\| \leq \lambda^r \|h_\rho\|$ for $0 < r \leq 1$. Thus, we just estimate $\|f_{D,\lambda} - f_\rho\|$ by Proposition 4.3.

By Propositions 4.4 and 4.5, and the bound (27), we get that for any fixed $l$, with probability at least $1 - 4\delta$, the following three bounds hold simultaneously,
\[
B_{D,\lambda}^2 \left( \frac{A_{D,\lambda}^2}{\lambda} + 1 \right) \lambda^{-\frac{\delta}{2}} \leq 32 \left( \frac{A_{D,\lambda}^2}{\lambda} + 1 \right) \lambda^{-\frac{\delta}{2}} \log^4 \frac{2}{\delta},
\]
\[
B_{D,\lambda}^2 C_{D,\lambda} \mathcal{G}_{D,\lambda} \leq 128(M + \sigma) \left( \frac{A_{D,\lambda}^2}{\lambda} + 1 \right) \lambda^{-\frac{\delta}{2}} \log^4 \frac{2}{\delta},
\]
and
\[
B_{D,\lambda}^2 C_{D,\lambda} \lambda^{-1} \mathcal{E}_{D,\lambda} \leq 16 c_{p,\sigma,M} \left( 1 + \lambda^{p+\frac{3}{2}} \right) \left( \frac{A_{D,\lambda}^2}{\lambda} + 1 \right) \lambda^{-(p+\frac{3}{2})} h^{-2p} \left( \log \frac{2}{\delta} \right)^3 \left( \log \frac{|D_i|}{\delta} \right)^{2p+1}.
\]

With the notations in (35), it follows that with probability at least $1 - 4k\delta$, the following two bounds hold true simultaneously,
\[
S_1 \leq 128(\|\mathcal{F}_i\| + M + \sigma) \max_{1 \leq i \leq k} \left( \frac{A_{D,\lambda}^2}{\lambda} + 1 \right) \lambda^{-\frac{\delta}{2}} \log^4 \frac{2}{\delta}
\]
(40)
and
\[
S_2 \leq 16 c_{p,\sigma,M} \left( 1 + \lambda^{p+\frac{3}{2}} \right) h^{-2p} \lambda^{-p-1} \left( \log \frac{2}{\delta} \right) \left( \log \frac{|D_i|}{\delta} \right)^{2p+1}
\]
\[
\times \left[ 1 + \lambda^{-\frac{\delta}{2}} \max_{1 \leq i \leq k} \left( \frac{A_{D,\lambda}^2}{\lambda} + 1 \right) \lambda \right].
\]
(41)
By Proposition 4.4 and 4.5 again, we see that with probability at least $1 - \frac{\delta}{2}$, the following bounds hold simultaneously,

\[ D_{D,\lambda} \leq 2A_{D,\lambda,k} \log \frac{8}{\delta} \quad \text{and} \quad F_{D,\lambda} \leq 8(M + \sigma)A_{D,\lambda,k} \log \frac{8}{\delta}. \]

Substitute $\delta$ by $\frac{\delta}{8k}$ in (40) and (41), one has with probability at least $1 - \delta$ that

\[
\begin{align*}
\|\tilde{f}_{D,\lambda} - f_p\|_\rho & \leq \|\tilde{f}_{D,\lambda} - f_\lambda\|_\rho + \|f_\lambda - f_p\|_\rho \\
& \leq \|h_\rho\|_\rho \lambda^r + S_1 + S_2 + D_{D,\lambda}\|f_\lambda\|_K + F_{D,\lambda} \\
& \leq \|h_\rho\|_\rho \lambda^r + 128\|f_\lambda\|_K + M + \sigma)\lambda^{-\frac{1}{2}} \left( \log^4 \frac{16k}{\delta} \right) \max_{1 \leq t \leq k} \left( \frac{A_{D_{t},\lambda}}{\lambda} + 1 \right) A^2_{D_{t},\lambda} \\
& \quad + 16c_{p,\sigma,M}(1 + \lambda^{r-\frac{1}{2}})h^{-2p} \lambda^{p-1} \left( \log^3 \frac{16k}{\delta} \right) \left( \log \frac{16|M|}{\delta} \right)^{2p+1} \\
& \quad \times \left( 1 + \lambda^{-\frac{1}{2}} \max_{1 \leq t \leq k} \left( \frac{A_{D_{t},\lambda}}{\lambda} + 1 \right) A_{D_{t},\lambda} \right) + (2\|f_\lambda\|_K + 8(M + \sigma))A_{D,\lambda,k} \log \frac{8}{\delta}.
\end{align*}
\]

The proof is complete. \(
\square
\)

**Proof of Corollary 2.3.** Our assumption $|D| \geq 4$ implies $|D|/4 \geq |D|/7$ and $\log |D| > 1$. By (7),

\[
A_{D,\lambda,k} = \frac{k}{|D|^{\sqrt{\lambda}}} + \frac{1}{|D|/4} + \sqrt{\frac{N(\lambda)}{|D|/4}} \leq \frac{8k}{|D|^{\sqrt{\lambda}}} + \sqrt{\frac{7C_0\lambda^{-s}}{|D|}},
\]

\[
\leq (8 + \sqrt{7C_0}) \lambda^{-s/2} \left( \frac{k}{\sqrt{|D|}} \lambda^{s/2} + 1 \right).
\]

By (12),

\[
\frac{k}{\sqrt{|D|}} \lambda^{s/2} \leq \begin{cases} 
\lambda^{s/2 + 1/2} \lambda^{2/r + 1/2} & \text{when } 0 < r < \frac{1}{2} \\
\lambda^{s/2 + 1/2} \lambda^{2/r + 1/2} & \text{when } \frac{1}{2} \leq r \leq 1
\end{cases} \leq \lambda^s \leq 1.
\]

By (30), $\|f_\lambda\|_K \leq \|h_\rho\|_\rho (1 + \lambda^{s-\frac{1}{2}})$ for $0 < r \leq 1$. Since $\frac{1}{|D|} = \lambda^{s+\max\{2r,1\}}$, the second term on the right-hand side of (11) is bounded by

\[
J_2 := (2\|f_\lambda\|_K + 8M + 8\sigma)A_{D,\lambda,k} \log \frac{8}{\delta} \leq C^1_1 (1 + \lambda^{s-\frac{1}{2}}) \lambda^{-s/2} \sqrt{|D|} \log \frac{8}{\delta}
\]

\[
= C^1_1 (1 + \lambda^{s-\frac{1}{2}}) \lambda^{-s/2} \log \frac{8}{\delta} \leq \frac{2C^1_1}{\lambda^s} \log \frac{8}{\delta},
\]

where $C^1_1 = (4\|h_\rho\|_\rho + 8M + 8\sigma) \times 2(8 + \sqrt{7C_0})$. By definition and the assumption $|D| \geq 4$ (hence $|D|/4 \geq |D|/7$),

\[
A_{D_{t},\lambda} = \frac{1}{|D_{t}|^{\sqrt{\lambda}}} + \frac{1}{|D_{t}|/4} + \sqrt{\frac{N(\lambda)}{|D_{t}|/4}} \leq \frac{8k}{|D_{t}|^{\sqrt{\lambda}}} + \sqrt{\frac{7C_0\lambda^{-s}k}{|D|}}
\]

\[
\leq (8 + \sqrt{7C_0}) \frac{k\lambda^{-s}}{|D|} \left( \frac{k}{|D|} \lambda^{s/2} + 1 \right) \leq 2(8 + \sqrt{7C_0}) \frac{k\lambda^{-s}}{|D|},
\]

18
where the last step follows from (42). Since
\[
\frac{k\lambda^{-s-1}}{|D|} \leq \left\{ \begin{array}{ll}
1\lambda^{-s-1}\lambda^{s+1} = 1, & \text{when } 0 < r < \frac{1}{2} \\
\lambda^{-r}\lambda^{-1-s}\lambda^{2r+s} = \lambda^{-\frac{r}{2}} & \text{when } \frac{1}{2} \leq r \leq 1
\end{array} \right. \leq 1,
\]
we have \( A_{D_1}^2 \lambda / \lambda \leq 4(8 + \sqrt{T_0})^2 k\lambda^{-s-1}/|D| \leq 4(8 + \sqrt{T_0})^2. \)

Now we bound the third term on the right-hand side of (11).
\[
J_3 := 128\|f_\lambda\|_K + M + \sigma \lambda^{-\frac{1}{2}} \left( \log^4 \frac{16k}{\delta} \right) \max_{1 \leq t \leq k} \left( \frac{A_{D_1}^2 \lambda}{\lambda} + 1 \right) A_{D_1}^2.
\]
When \( 0 < r < \frac{1}{2}, k = 1 \) and \( \frac{1}{|D|} = \lambda^{1+s}, \)
\[
J_3 \leq C^2_1 \lambda^{r-1} \left( \log^4 \frac{16}{\delta} \right) \frac{k\lambda^{-s}}{|D|} = C^2_1 \lambda^r \log^4 \frac{16}{\delta},
\]
where \( C^2_1 = 128(\|h_\rho\|_\rho + M + \sigma)[4(8 + \sqrt{T_0})^2 + 1] \times 4(8 + \sqrt{T_0})^2. \) When \( \frac{1}{2} \leq r \leq 1, k \leq \lambda^2 \lambda^{-r} \log^{-4} |D| \) and \( \frac{1}{|D|} = \lambda^{2r+s} \). Recall that \( |D| \geq 4 \) which implies \( \log^4 \frac{16}{\delta} \leq \log^4 \frac{16|D|}{\delta} \leq \log^4 \frac{16|D|}{\delta} \leq \log^4 \frac{16|D|}{\delta}. \)
So
\[
J_3 \leq C^1_1 \lambda^{-\frac{1}{2}} \left( \log^4 |D| \right) \left( \log^4 \frac{16}{\delta} \right) \lambda^r \left( \log^{-4} |D| \right) \lambda^{-s} \lambda^{2r+s} = C^1_1 \lambda^r \log^4 \frac{16}{\delta},
\]
where \( C^1_1 := 128(\|h_\rho\|_\rho + M + \sigma) \times 2^4[4(8 + \sqrt{T_0})^2 + 1] \times 4(8 + \sqrt{T_0})^2. \)

The last term on the right-hand side of (11) is bounded as follows.
\[
J_4 := 16c_{p,\sigma,M} (1 + \lambda^{p+\frac{1}{2}}) h^{-2p} \lambda^{-p-1} \left( \log^3 \frac{16k}{\delta} \right) \left( \log \frac{16|D|}{\delta} \right)^{2p+1} \times \left( 1 + \lambda^{-\frac{1}{2}} \max_{1 \leq t \leq k} \left( \frac{A_{D_1}^2 \lambda}{\lambda} + 1 \right) A_{D_1} \right)
\leq C^4_1 h^{-2p} \lambda^{-p-1} \left( \log |D| \right)^{2p+4} \left( \log \frac{16}{\delta} \right)^{2p+4},
\]
where \( C^4_1 := 16c_{p,\sigma,M} \times 2^{2p+5} \times \left[ 1 + (4(8 + \sqrt{T_0})^2 + 1) \times 2(8 + \sqrt{T_0}) \right]. \) One completes the proof by letting \( C := \|h_\rho\|_\rho + 2C^1_1 + \max\{C^2_1, C^3_1, C^4_1\}. \)

5. Estimates in semi-supervised learning

To derive the optimal learning rate in semi-supervised learning, we give the following proposition by taking similar procedures in the proof of Proposition 4.3.

**Corollary 5.1** Let \( \tilde{f}_{D^*}^{\cdot,\lambda} \) be defined in (16). One has
\[
\|\tilde{f}_{D^*}^{\cdot,\lambda} - f_\lambda\|_K \leq S^*_1 + S^*_2 + D_{D^*}^{\cdot,\lambda}\|f_\lambda\|_K + F_{D^*}^{\cdot,\lambda},
\]
where
\[
S^*_1 = \max_{1 \leq t \leq k} \left( B_{D^*_t}^{\cdot,\lambda} C_{D^*_t}^{2,\cdot,\lambda} \|f_\lambda\|_K \lambda^{-\frac{1}{2}} + B_{D^*_t}^{\cdot,\lambda} C_{D^*_t}^{\cdot,\lambda} G_{D^*_t}^{\cdot,\lambda} \lambda^{-\frac{1}{2}} \right),
\]
and
\[
S^*_2 = \max_{1 \leq t \leq k} \left( B_{D^*_t}^{\cdot,\lambda} C_{D^*_t}^{\cdot,\lambda} \lambda^{-1} + \lambda^{-\frac{1}{2}} \right) \|E_{D^*_t}^{\cdot,\lambda}\|_K.
\]

\[\square\]
To quantify the above bounds, we get the following probability inequalities about $\mathcal{B}_{D^*,\lambda}, \mathcal{C}_{D^*,\lambda}, \mathcal{D}_{D^*,\lambda}$, whose proofs can be found in the earlier work in Hu et al. (2020).

**Corollary 5.2** Each of the following three bounds holds with probability $1 - \delta$.

$$\mathcal{B}_{D^*,\lambda} \leq 2 \left( \frac{2A_{D^*,\lambda} \log \frac{2}{3}}{\sqrt{\lambda}} \right)^2 + 2, \quad \mathcal{C}_{D^*,\lambda} \leq 2A_{D^*,\lambda} \log \frac{2}{\delta},$$

and

$$\mathcal{D}_{D^*,\lambda} \leq 2A_{D^*,\lambda,k} \log \frac{2}{\delta},$$

where $A_{D^*,\lambda,k} = \frac{k}{|D||\sqrt{\lambda}|} + \frac{1}{\|D^*|\sqrt{\lambda|A\lambda|}} + \sqrt{\frac{N(\lambda)}{|D^*|\lambda}}$ and $A_{D^*,\lambda} = A_{D^*,\lambda,1}$, which are consistent with the notations $A_{D^*,\lambda,k}$ and $A_{D^*,\lambda}$ we defined in Theorem 2.2. 

5.1. Distributed U-statistics with unlabeled data and unbounded sampling

Next we turn to the bounds of $\mathcal{F}_{D^*,\lambda}$ and $\mathcal{G}_{D^*,\lambda}$, which are more involved in semi-supervised learning. To this end, we need the following lemma.

**Lemma 5.3** Let $(\mathcal{H}, \| \cdot \|)$ be a separable Hilbert space. Let $\xi(w, z)$ be an $\mathcal{H}$-valued random variable defined on $(W \times Z, \rho_W \times \rho_Z)$. Assume that there exist two constants $\sigma > 0$ and $M > 0$, such that for any integer $q \geq 2$,

$$\mathbb{E} \|\xi - \mathbb{E}\xi\|^q \leq \frac{1}{2} q! \sigma^2 M^{q-2}.$$

Suppose that from $(W, \rho_W)$, one draws independently a sample $D = \{w_1, \ldots, w_{|D|}\}$, which is evenly divided to $k$ disjoint subsets $D = \bigcup_{l=1}^k D_l$ with $|D_1| = \cdots = |D_k| = |D|/k$. Suppose that similarly, one divides another sample $\tilde{D} = \bigcup_{l=1}^k \tilde{D}_l$ independently drawn from $(Z, \rho_Z)$ such that the subsets $\tilde{D}_l$’s are disjoint and $|\tilde{D}_1| = \cdots = |\tilde{D}_k| = |\tilde{D}|/k$. Assume that $|D| \leq |\tilde{D}|$. Then with probability at least $1 - \delta$, there holds

$$\left\| \frac{1}{k} \sum_{l=1}^k \frac{1}{|D_l||D_l|} \sum_{w \in D_l} \sum_{z \in \tilde{D}_l} \xi(w, z) - \mathbb{E}\xi \right\| \leq \frac{2M \log(2/\delta)}{|D|} + \sqrt{\frac{2\sigma^2 \log(2/\delta)}{|\tilde{D}|} \mathbb{E} \|\xi - \mathbb{E}\xi\|^4}.$$

**Proof.** Let $\pi$ be a permutation of the set $\{1, \ldots, |D_l|\}$ of integers, so $\{w_{\pi(1)}, \ldots, w_{\pi(|D_l|)}\}$ is the associated permutation of $D_l = \{w_l\}_{i=1}^{|D_l|}$. Let $\psi$ be a permutation of $\{1, \ldots, |\tilde{D}_l|\}$, so $\{z_{\psi(1)}, \ldots, z_{\psi(|\tilde{D}_l|)}\}$ is the associated permutation of $\tilde{D}_l = \{z_l\}_{i=1}^{|\tilde{D}_l|}$. For any $1 \leq l \leq k$, we write $U_{\pi,\psi}^l = \frac{1}{|D_l|} \sum_{i=1}^{|D_l|} \xi(w_{\pi(i)}, z_{\psi(i)})$ to obtain

$$\frac{1}{|D_l||D_l|} \sum_{w \in D_l} \sum_{z \in \tilde{D}_l} \xi(w, z) = \frac{1}{|D_l||D_l|} \sum_{\pi} \sum_{\psi} U_{\pi,\psi}^l,$$

(47)

where the last two sums are taken over all the $|D_l|!$ permutations $\pi$ of $\{1, \ldots, |D_l|\}$ and all the $|\tilde{D}_l|!$ permutations $\psi$ of $\{1, \ldots, |\tilde{D}_l|\}$, respectively. Note that $|D_1| = \cdots = |D_l| = |D|/k$ and $|\tilde{D}_1| = \cdots = |\tilde{D}_l| = |\tilde{D}|/k$. One takes the average of (47) over $1 \leq l \leq k$, to give

$$\frac{1}{k} \sum_{l=1}^k \frac{1}{|D_l||D_l|} \sum_{w \in D_l} \sum_{z \in \tilde{D}_l} \xi(w, z) = \frac{1}{|D_l||D_l|} \sum_{\pi,\psi} U_{\pi,\psi}^l = \frac{1}{|D_l||D_l|} \sum_{\pi,\psi} \frac{1}{k} \sum_{l=1}^k U_{\pi,\psi}^l,$$

$$= \frac{1}{|D_l||D_l|} \sum_{\pi,\psi} \frac{1}{k} \sum_{l=1}^k U_{\pi,\psi}^l.$$
where

$$\frac{1}{k} \sum_{l=1}^{k} U_{\pi,\psi}^l = \frac{1}{k} \sum_{l=1}^{k} \frac{|D_l|}{|D|} \sum_{i=1}^{D_l} \xi(u_{\pi(i)}, z_{\psi(i)}^l) = \frac{1}{|D|} \sum_{l=1}^{k} \sum_{i=1}^{D_l} \xi(u_{\pi(i)}, z_{\psi(i)}^l).$$

By the convexity of the function $\cosh(t)$, we obtain that for any $c, \epsilon > 0$,

$$\Pr \left\{ \left\| \frac{1}{k} \sum_{l=1}^{k} \frac{|D_l|}{|D|} \sum_{w \in D_l} \sum_{z \in D_l} \xi(w, z) - \mathbb{E} \xi \right\| \geq \epsilon \right\} \leq \mathbb{E} \cosh \left( c \left\| \frac{1}{k} \sum_{l=1}^{k} \frac{|D_l|}{|D|} \sum_{\pi,\psi} U_{\pi,\psi}^l \right\| \right) / \cosh(c\epsilon)$$

$$= \mathbb{E} \cosh \left( c \left\| \frac{1}{|D|} \sum_{l=1}^{k} \sum_{\pi,\psi} \left[ \frac{1}{k} \sum_{l=1}^{k} U_{\pi,\psi}^l \right] \right\| \right) / \cosh(c\epsilon)$$

$$\leq \frac{1}{|D|} \sum_{\pi,\psi} \mathbb{E} \cosh \left( c \left\| \frac{1}{|D|} \sum_{l=1}^{k} \sum_{i=1}^{D_l} \xi(u_{\pi(i)}, z_{\psi(i)}^l) - \mathbb{E} \xi \right\| \right) / \cosh(c\epsilon).$$

One employs Lemma A.2 in Appendix to obtain

$$\Pr \left\{ \left\| \frac{1}{k} \sum_{l=1}^{k} \frac{|D_l|}{|D|} \sum_{w \in D_l} \sum_{z \in D_l} \xi(w, z) - \mathbb{E} \xi \right\| \geq \epsilon \right\} \leq 2 \exp \left\{ -\frac{|D|\epsilon^2}{2(\sigma^2 + M\epsilon)} \right\}.\quad \blacksquare$$

We take $\delta = 2 \exp \left\{ -\frac{|D|\epsilon^2}{2(\sigma^2 + M\epsilon)} \right\}$ to complete the proof.

Noting that if $f(x, u) = -f(u, x)$ for any $(x, u) \in \mathcal{X}^2$, then $K_{(x,u)}$ is antisymmetric, i.e., $K_{(x,u)} = -K_{(u,x)}$ for any $(x, u) \in \mathcal{X}^2$. The quantities $\mathcal{F}_{D^*,\lambda}$ and $\mathcal{G}_{D^*,\lambda}$ are involved with unlabeled data, and we will handle them by the feature of antisymmetry of $K$ and get the bounds as follows.

**Proposition 5.4** Each of the following two bounds holds with probability at least $1 - \delta$.

$$\mathcal{F}_{D^*,\lambda} \leq 8(M + \sigma)A_{D,D^*,\lambda,k} \log \frac{4}{\delta}, \quad \text{and} \quad \mathcal{G}_{D^*,\lambda} \leq 8(M + \sigma)A_{D,D^*,\lambda} \log \frac{4}{\delta}.$$

**Proof.** Recall that $K$ is antisymmetric. So

$$L_K f_\rho = \int_Z \int_Z (y - v)K_{(x,u)}d\rho(x,y)d\rho(u,v)$$

$$= \int_Z \int_Z yK_{(x,u)}d\rho(x,y)d\rho(u,v) - \int_Z \int_Z vK_{(x,u)}d\rho(x,y)d\rho(u,v)$$

$$= \int_Z \int_Z 2yK_{(x,u)}d\rho(x,y)d\rho(u,v).$$

Recall that we have the relation $(x, y) \mapsto (x, \frac{|D^*|}{|D|} y)$ when we embed $D$ to $D^*$. Write $w = (x, y)$ and $z = (u, v)$. We have the following decomposition

$$\hat{f}_{\rho,D^*} := \frac{1}{|D^*|^2} \sum_{w,z \in D^*} (y - v)K_{(x,u)}$$

$$= \frac{1}{|D^*|^2} \sum_{w,z \in D} \frac{|D^*|}{|D|} (y - v)K_{(x,u)} + \frac{2}{|D^*|^2} \sum_{w \in D, z \in \hat{D}} \frac{|D^*|}{|D|} yK_{(x,u)}$$

$$= \frac{|D|}{|D^*|} f_{\rho,D} + \frac{\hat{D}}{|D^*|} \hat{f}_{\rho,D,D^*}.$$
where $\hat{f}_{\rho,D}$ is defined in Lemma 4.1 and $\hat{f}_{\rho,Dt,\hat{D}} = \frac{1}{|D||\hat{D}|} \sum_{w \in D, z \in \hat{D}} 2yK(x,u)$. Below, $\hat{f}_{\rho,Dt}$ and $\hat{f}_{\rho,Dt,\hat{D}_t}$ are similarly defined by substituting $D$ and $\hat{D}$ with $D_t$ and $\hat{D}_t$, respectively. Note that both $\hat{f}_{\rho,D}$ and $\hat{f}_{\rho,D,\hat{D}}$ are empirical analogs of $L_Kf_\rho$. We have

$$F_{D,\lambda} = \left\| \frac{1}{k} \sum_{l=1}^{k} (L_K + 2\lambda I)^{-\frac{1}{2}} (\hat{f}_{\rho,Dt} - L_Kf_\rho) \right\|_{L_K} \leq 8(M + \sigma)A_{D,\lambda,k} \log \frac{4}{\delta}.$$  

For the first term, it has been proved that with probability at least $1 - \frac{\delta}{2}$, there holds

$$F_{D,\lambda} = \left\| \frac{1}{k} \sum_{l=1}^{k} (L_K + 2\lambda I)^{-\frac{1}{2}} (\hat{f}_{\rho,Dt} - L_Kf_\rho) \right\|_{L_K} \leq 8(M + \sigma)A_{D,\lambda,k} \log \frac{4}{\delta}.$$  

For the second term, let $\xi(w,z) = (L_K + 2\lambda I)^{-\frac{1}{2}} 2yK(x,u)$ with $w = (x,y) \in D$ and $z = (u,v) \in \hat{D}$, so

$$(L_K + 2\lambda I)^{-1/2} \hat{f}_{\rho,Dt,\hat{D}_t} = \frac{1}{|D||\hat{D}|} \sum_{w \in D} \sum_{z \in \hat{D}} \xi(w,z),$$

and $(L_K + 2\lambda I)^{-1/2} L_Kf_\rho = \mathbb{E}\xi$. With (8), for any integer $q \geq 2$,

$$\mathbb{E}[\xi - \mathbb{E}\xi]\|K\| \leq 2^{q-1} \mathbb{E}[\|\xi\|_q^q] + 2^q \|\mathbb{E}\xi\|_q^q \leq 2^q \mathbb{E}[\|\xi\|_q^q]$$

$$\leq 2^q \left( \frac{1}{\sqrt{\lambda}} \right)^{q-2} \mathbb{E}[\|Y\|_q\|X\|_q\|K\|]$$

$$\leq 2^{q-2} \left( \frac{1}{\sqrt{\lambda}} \right)^{q-2} \frac{1}{2} q! \sigma^2 M^{q-2} N(\lambda) = \frac{1}{2} q! (16\sigma^2 N(\lambda))(4M/\sqrt{\lambda})^{q-2}.$$  

Applying Lemma 5.3 one gets that with probability at least $1 - \frac{\delta}{2}$,

$$\left\| \frac{1}{k} \sum_{l=1}^{k} (L_K + 2\lambda I)^{-\frac{1}{2}} (\hat{f}_{\rho,Dt,\hat{D}_t} - L_Kf_\rho) \right\|_{L_K} \leq 8(M + \sigma) \left( \frac{1}{|D|\sqrt{\lambda}} + \frac{\sqrt{N(\lambda)}}{|D|} \right) \log \frac{4}{\delta}.$$  

Recall that $|D| + |\hat{D}| = |D^*|$. We combine the analysis above by

$$\frac{|D|}{|D^*|} A_{D,\lambda,k} + \frac{|\hat{D}|}{|D^*|} \left( \frac{1}{|D|\sqrt{\lambda}} + \frac{\sqrt{N(\lambda)}}{|D|} \right)$$

$$\leq \frac{|D|}{|D^*|} \frac{k}{|D|\sqrt{\lambda}} \left( \frac{1}{|D|\sqrt{\lambda}} + \frac{\sqrt{N(\lambda)}}{|D^*|} \right) \left( \frac{|D|}{|D^*|} + \frac{|\hat{D}|}{|D^*|} \right)$$

$$= \frac{k}{|D^*|\sqrt{\lambda}} + \frac{1}{|D^*|\sqrt{\lambda}} + \frac{\sqrt{N(\lambda)}}{|D^*|} = A_{D,D^*,\lambda,k}.$$  

One completes the proof by observing that $G_{D^*,\lambda}$ is a special case of $F_{D^*,\lambda}$ with $k = 1$.}

22
5.2. Proof of learning rates in semi-supervised learning

Proof of Theorem 2.6. We can decompose \( \| \hat{f}_{D^*, \lambda} - f_\rho \|_p \) as the sample error \( \| \hat{f}_{D^*, \lambda} - f_\lambda \|_p \) and the approximation error \( \| f_\lambda - f_\rho \|_p \). As stated in (29), \( \| f_\lambda - f_\rho \|_p \leq \lambda^r \| h_\rho \|_p \) for \( 0 < r \leq 1 \). Thus, we just estimate the sample error by Corollary 5.1 and bound the right-hand side of (46) term by term.

Recall the definition of \( D^* \) in Section 2.2. We include it for the sake of completeness.

Thus, \( E_{D^*, \lambda} \leq 2 \left( \frac{c_0 |D^*|}{\lambda |D^*|} \right)^{1/2} \max_{w \in D} [y]. \) Similar to the estimation (26),

\[
\| E_{D^*, \lambda} \|_K \leq c_2 \| f_{D^*, \lambda} \|_K \leq 2 \left( \frac{c_0 |D^*|}{\lambda |D^*|} \right)^{1/2} \max_{w \in D} [y].
\]

where \( \Delta_{D, D^*, \lambda} = \left( \frac{|D^*|}{\lambda |D^*|} \right)^{1/2} + \left( \frac{|D^*|}{\lambda |D^*|} \right)^{2p+1}. \) Therefore with probability \( 1 - \delta \),

\[
\| E_{D^*, \lambda} \|_K \leq C_2 \Delta_{D, D^*, \lambda} h^{-2p} \log^{2p+1} \left[ \frac{|D|}{\delta} \right],
\]

where \( C_2 = 2^{2p+1} c_2 (C_0^{p+1} + 1) (4M + 5\sigma)^{2p+1}. \) The following part of the proof is similar to the proof of Theorem 2.2. We include it for the sake of completeness.

Recall Corollary 5.2 and Proposition 5.4. For each fixed \( l = 1, \ldots, k \), with probability at least \( 1 - 4\delta \), the following three bounds hold true simultaneously.

\[
B_{D^*, \lambda} C_{D^*, \lambda}^{2} \lambda^{-1/2} \leq 32 \left( \frac{A_{D^*, \lambda}^{2}}{\lambda} + 1 \right) \left( \frac{\log^{2} \frac{2}{\delta}}{\delta} \right) \lambda^{-1/2}.
\]

23
\[ B_{D^*_l, \lambda} C_{D^*_l, \lambda} C_{D^*_l, \lambda} \leq |E_{D^*_l, \lambda}| \]
\[ \leq 8 \left( \frac{A_{D^*_l, \lambda}^2}{\lambda} + 1 \right) \left( \log^2 \frac{2}{\delta} \right) \times 2 A_{D^*_l, \lambda} \left( \log^2 \frac{2}{\delta} \right) \lambda^{-1} C_2 h^{-2p} \Delta_{D^*_l, \lambda} \log^{2p+1} |D^*_l| \]
\[ = 16 C_2 \Delta_{D^*_l, \lambda} h^{-2p} \lambda^{-1} \left( \frac{A_{D^*_l, \lambda}^2}{\lambda} + 1 \right) A_{D^*_l, \lambda} \left( \log^3 \frac{2}{\delta} \right) \log^{2p+1} |D^*_l|, \]

and
\[ B_{D^*_l, \lambda} \leq 8 \left( \frac{A_{D^*_l, \lambda}^2}{\lambda} + 1 \right) \left( \log^2 \frac{2}{\delta} \right) \times 2 A_{D^*_l, \lambda} \left( \log^2 \frac{2}{\delta} \right) \lambda^{-1/2} \]
\[ \leq 256 (M + \sigma) \left( \frac{A_{D^*_l, \lambda}^2}{\lambda} + 1 \right) A_{D^*_l, \lambda} \lambda^{-1/2} \log^4 \frac{2}{\delta}, \]

where we used \( \log \frac{\delta}{M} \leq 2 \log \frac{2}{\delta}, \) which follows from \( 4/\delta \leq 4\delta^2. \) Therefore, with probability at least \( 1 - 4k \delta, \) the following two bounds hold true simultaneously.

\[ S_1^* \leq 256 (1 + M + \sigma) \lambda^{-1/2} \left( \log^4 \frac{2}{\delta} \right) \]
\[ \times \max_{1 \leq l \leq k} \left[ \left( \frac{A_{D^*_l, \lambda}^2}{\lambda} + 1 \right) A_{D^*_l, \lambda} \left( A_{D^*_l, \lambda} \| f_\lambda \|_K + A_{D^*_l, \lambda} \right), \right. \] (48)
\[ S_2^* \leq 16 \lambda^{-1/2} C_2 \Delta_{D^*_l, \lambda} h^{-2p} \left( \log^3 \frac{2}{\delta} \right) \]
\[ \times \max_{1 \leq l \leq k} \left[ \left( \frac{A_{D^*_l, \lambda}^2}{\lambda} + 1 \right) A_{D^*_l, \lambda} \lambda^{-1/2} + 1 \right] \log^{2p+1} |D^*_l|. \] (49)

By Corollary 5.2 and Proposition 5.4 we see that with probability at least \( 1 - \frac{\delta}{2}, \) the following two bounds hold simultaneously.

\[ D_{D^*_l, \lambda} \leq 2 A_{D^*_l, \lambda} \| K \|_K \log \frac{8}{\delta}, \] (50)
\[ F_{D^*_l, \lambda} \leq 8 (M + \sigma) A_{D^*_l, \lambda} \log \frac{16}{\delta}. \] (51)

The proof is completed by scaling \( \delta \) to \( \frac{\delta}{8k} \) in (48) and (49), and combining them with (50) and (51).

**Proof of Corollary 2.7.** We have \(|D^*| \geq |D|^{\frac{2k - 1}{2k}} = \lambda^{-s - 1}\) when \( 0 < r < \frac{1}{2}, \) and \(|D^*| \geq |D| = \lambda^{-2r-s}\) when \( \frac{1}{2} \leq r \leq 1. \) By (30),

\[ \frac{\| f_\lambda \|_K}{|D^*|} \leq \begin{cases} \| h_\rho \|, \lambda^{-r-s}, & \text{when } 0 < r < \frac{1}{2}, \\ \| h_\rho \|_p \sqrt{|D^*|}, & \text{when } \frac{1}{2} \leq r \leq 1 \end{cases} \]

\[ \| f_\lambda \|_K \leq \| h_\rho \|_p \lambda^{r-s}. \] (52)

For any \( 1 \leq l \leq k, \) the assumption \(|D^*_l| \geq |D_l| \geq 4\) implies \(|D^*_l| / 4 \geq |D^*_l| / 7. \) The assumption \( k \leq \sqrt{|D^*| \lambda^{1+s}} \) implies \( k \leq \sqrt{|D^*| \lambda^{1+s}} \) \leq \( \frac{k}{|D^*|} \lambda^{-(1+s)/2} \leq 1. \) Recall \( N(\lambda) \leq C\lambda^{-s}. \) We have

\[ A_{D^*_l, \lambda} = \frac{1}{|D^*_l|^{1/\lambda}} + \frac{1}{|D^*_l| / 4} \sqrt{\lambda} + \sqrt{\frac{N(\lambda)}{|D^*_l| / 4}} \leq \frac{8k}{|D^*|^{1/\lambda}} + \sqrt{7C_0 \lambda^{-s} k} \]
\[ \leq (8 + \sqrt{7C_0}) \sqrt{\frac{k}{|D^*|^{1/\lambda}}} \left( \sqrt{\frac{k}{|D^*|^{\lambda^2 - s} + 1}} \right) \leq 2 (8 + \sqrt{7C_0}) \sqrt{\frac{k}{|D^*|^{\lambda^s}}}. \]
So,
\[
\frac{A_{D_l,\lambda}^2}{\lambda} + 1 \leq 4(8 + \sqrt{7C_0})^2 \frac{k}{|D^*|} \lambda^{-s-1} + 1 \leq C_3^1,
\]
where \( C_3^1 := 4(8 + \sqrt{7C_0})^2 + 1 \). From definition,
\[
A_{D_l, D^*_l, \lambda} = \frac{1}{|D_l|^{\sqrt{\lambda}}} + \frac{1}{|D_l|\sqrt{\lambda}} + \sqrt{\frac{\mathcal{N}(\lambda)}{|D_l|/4}}
\leq \frac{8k}{|D|^{\sqrt{\lambda}}} + \frac{\sqrt{7C_0\lambda^{-s-1}}}{|D|} = 8k\lambda^{2r+s-\frac{1}{2}} + \sqrt{7C_0k\lambda^r}.
\]
So, for any \( 1 \leq l \leq k \),
\[
A_{D_l^*, \lambda} \|f\|_K + A_{D_l, D^*_l, \lambda} \leq 2(8 + \sqrt{7C_0})\sqrt{k\lambda^{-s}}\|f\|_K + A_{D_l, D^*_l, \lambda}
\leq 2(8 + \sqrt{7C_0})\|h\rho\|_\rho\sqrt{k\lambda^r} + 8k\lambda^{2r+s-\frac{1}{2}} + \sqrt{7C_0k\lambda^r}
\leq C_3^2 \left( \sqrt{k\lambda^r} + k\lambda^{2r+s-\frac{1}{2}} \right),
\]
where \( C_3^2 := 2(8 + \sqrt{7C_0})\|h\rho\|_\rho + 8 + \sqrt{7C_0} \).

Recall \( \log \frac{16k}{\delta} \leq \log \frac{16|D|}{\delta} \leq 2(\log |D|) \log \frac{16}{\delta} \). We summarize the analysis above to give the bound of the second term on the right-hand side of (17).

\[
J_2^* := 256(1 + M + \sigma)\lambda^{-1/2} \left( \log^4 \frac{16k}{\delta} \right)
\times \max_{1 \leq l \leq k} \left( \frac{A_{D_l^*, \lambda}^2}{\lambda} + 1 \right) A_{D_l^*, \lambda} (A_{D_l^*, \lambda} \|f\|_K + A_{D_l, D^*_l, \lambda})
\leq C_3^3 \left( \log^4 |D| \right) \left( \log^4 \frac{16}{\delta} \right) \sqrt{\frac{k}{|D^*|^{\sqrt{\lambda}}} \left( \sqrt{k\lambda^r} + k\lambda^{2r+s-\frac{1}{2}} \right)},
\]
where \( C_3^3 := 256(1 + M + \sigma) \times 2^4C_3^1 \times 2(8 + \sqrt{7C_0})C_3^2 \). We use the assumption (18) to obtain
\[
\frac{k\lambda^r \log^4 |D|}{\sqrt{|D^*|^{\lambda^{-s+1}}}} \leq \lambda^r,
\]
\[
\frac{k^{3/2}\lambda^{2r+s-1}}{\sqrt{|D^*|^{\lambda^s}}} \log^4 |D| \leq \left( \frac{|D^*|^3\lambda^{2r-s-\frac{3}{2}}}{|D^*|^{\lambda^s}} \right)^{3/2} \leq \lambda^{2r-s-1} \leq \lambda^r.
\]
So \( J_2^* \leq 2C_3^3\lambda^r \log^4 \frac{16}{\delta} \).

25
Next, it is easy to see that for $1 \leq l \leq k$, $\Delta_{D_l, D^*_l, 3} = \Delta_{D, D^*, 3}$. Note that $a \leq a^2 + 1$ for $a \geq 0$. We have
\[
J^*_3 := 16\lambda^{-1/2}C_2h^{-2p} \left( \log^3 \frac{16k}{\delta} \right) \times \max_{1 \leq i \leq k} \left[ \left( \frac{\lambda}{A_{D^*_i, \lambda}} + 1 \right) \frac{A_{D^*_i, \lambda}}{\sqrt{\lambda}} + 1 \right] \Delta_{D, D^*, \lambda} \log^{2p+1} \frac{16|D|}{\delta} \leq 16C_2h^{-2p} \times 2^3 \left( \log^3 |D| \right) \left( \log^3 \frac{16}{\delta} \right) \lambda^{-1/2} ((C_3^1)^2 + 1) \times \Delta_{D, D^*, \lambda} \times 2^{2p+1} (\log |D|)^{2p+1} \left( \log \frac{16}{\delta} \right)^{2p+1} = C_3^2h^{-2p}\lambda^{-1/2} \Delta_{D, D^*, \lambda} (\log |D|)^{2p+4} \left( \log \frac{16}{\delta} \right)^{2p+4},
\]
where $C_3^2 := 16C_2 \times 2^{2p+4} ((C_3^1)^2 + 1)$.

Now we bound the last term on the right-hand side of (17). By definition and the bound (52) above,
\[
A_{D^*, \lambda, k} \|f_\lambda\|_K = \|f_\lambda\|_K \left( \frac{k}{|D^*|\sqrt{\lambda}} + \frac{1}{(|D^*|/4)\sqrt{\lambda}} + \sqrt{\frac{\mathcal{N}(\lambda)}{|D^*|/4}} \right) \leq \|h_\rho\|_{\rho} \lambda^{r + \frac{s}{2}} \left( \frac{8k}{\sqrt{|D^*|\lambda}} + \sqrt{7C_0\lambda^{-s}} \right).
\]
The assumption (18) of $k$ implies $\sqrt{\frac{k}{|D^*|\lambda}} \leq \lambda^{s/2} \leq 1$. So $A_{D^*, \lambda, k} \|f_\lambda\|_K \leq \|h_\rho\|_{\rho} (8 + \sqrt{7C_0}) \lambda^r$. By the assumption $|D^*| \geq \lambda^{-1-s}$, $k \leq \sqrt{|D^*|}\lambda^{1+s}$, and $r + s \geq \frac{1}{2}$,
\[
A_{D, D^*, \lambda, k} = \frac{k}{|D^*|\sqrt{\lambda}} + \frac{1}{(|D^*|/4)\sqrt{\lambda}} + \sqrt{\frac{\mathcal{N}(\lambda)}{|D^*|/4}} \leq \sqrt{\frac{\lambda^s}{|D^*|}} + 7\lambda^{2r+s-r} + \sqrt{7C_0\lambda^{-s+r+\frac{s}{2}}} \leq (8 + \sqrt{7C_0}) \lambda^r.
\]
We summarize the above analysis and use (17) to obtain that with probability $1 - \delta$,
\[
\|\tilde{f}_{D^*, \lambda} - f_\rho\|_\rho \leq \|h_\rho\|_{\rho} \lambda^{r} + 2C_3^3 \lambda^{r} \log^4 \frac{16}{\delta} + C_3^4h^{-2p}\lambda^{-1/2} \Delta_{D, D^*, \lambda} (\log |D|)^{2p+4} \left( \log \frac{16}{\delta} \right)^{2p+4} + (2\|h_\rho\|_{\rho} + 8M + 8\sigma)(8 + \sqrt{7C_0}) \lambda^r \log^4 \frac{16}{\delta},
\]
which yields the conclusion with $C_3 := \|h_\rho\|_{\rho} + 2C_3^3 + C_3^4 + (2\|h_\rho\|_{\rho} + 8M + 8\sigma)(8 + \sqrt{7C_0})$. \qed

**Acknowledgments**

The work described in this paper was partially supported by National Natural Science Foundation of China (Projects 11671307, 11571078, 11671171) and the Research Grants Council of the Hong Kong
Appendix A. Concentration inequalities

Lemma A.1 (Pinelis and Sakhanenko, 1986, Theorem 3) Let \( \{\eta_i\}_{i=1}^n \) be a sequence of independent random variables with values in a separable Hilbert space \( (H, \| \cdot \|) \) and \( \mathbb{E}(\eta_i) = 0 \) for each \( i = 1 \cdots n \). Then for any \( c > 0 \), there holds

\[
\mathbb{E}\left[ \cosh\left( c \left\| \sum_{i=1}^n \eta_i \right\| \right) \right] \leq \prod_{i=1}^n \mathbb{E}\left( e^{c \left\| \eta_i \right\|} - c \left\| \eta_i \right\| \right).
\]

\[\blacksquare\]

Lemma A.2 Let \( \xi(z) \) be a random variable defined on \( (Z, \rho) \) with values in a separable Hilbert space \( (H, \| \cdot \|) \). Assume that there are two positive constants \( \sigma > 0 \) and \( M > 0 \) such that for any integer \( q \geq 2 \),

\[
\mathbb{E}(\|\xi(z) - \mathbb{E}(\xi)\|^q) \leq \frac{1}{2} q! \sigma^2 M^{q-2},
\]

then for any \( \epsilon > 0 \) and any positive integer \( n \), there holds

\[
\inf_{c>0} \mathbb{E}\left[ \cosh\left( c \left\| \frac{1}{n} \sum_{i=1}^n \xi(z_i) - \mathbb{E}(\xi) \right\| \right) \right] / \cosh(ce) \leq 2 \exp\left\{ - \frac{nc^2}{2(\sigma^2 + Mc)} \right\}.
\]

Proof. Applying Lemma A.1 with \( \eta_i = \frac{1}{n} [\xi(z_i) - \mathbb{E}(\xi)] \), we get that

\[
\mathbb{E}\left[ \cosh\left( c \left\| \frac{1}{n} \sum_{i=1}^n \xi(z_i) - \mathbb{E}(\xi) \right\| \right) \right] \leq \prod_{i=1}^n \mathbb{E}\left( e^{c \|\xi(z_i) - \mathbb{E}(\xi)\|} - c \frac{\|\xi(z_i) - \mathbb{E}(\xi)\|}{n} \right).
\]

For each \( 1 \leq i \leq n \) and any \( c > 0 \), by Taylor expansion and the elementary inequality \( 1 + a \leq e^a \) for any \( a \in \mathbb{R} \), we have

\[
\mathbb{E}\left( e^{c \|\xi(z_i) - \mathbb{E}(\xi)\|} - c \frac{\|\xi(z_i) - \mathbb{E}(\xi)\|}{n} \right) = \sum_{q=2}^{\infty} c^q \mathbb{E}(\|\xi(z) - \mathbb{E}(\xi)\|^q) \frac{1}{n^q q!} \leq \exp\left\{ \sum_{q=2}^{\infty} c^q \mathbb{E}(\|\xi(z) - \mathbb{E}(\xi)\|^q) \frac{1}{n^q q!} \right\}.
\]

We set \( 0 < c < n/M \) and recall (53) to obtain,

\[
\mathbb{E}\left( e^{c \|\xi(z_i) - \mathbb{E}(\xi)\|} - c \frac{\|\xi(z_i) - \mathbb{E}(\xi)\|}{n} \right) \leq \exp\left\{ \frac{\sigma^2}{2} \sum_{q=2}^{\infty} c^q M^{q-2} \frac{q^q}{n^q q!} \right\} = \exp\left\{ \frac{\sigma^2 c^2}{2(n - cM)}\right\}.
\]

Therefore,

\[
\mathbb{E}\left[ \cosh\left( c \left\| \frac{1}{n} \sum_{i=1}^n \xi(z_i) - \mathbb{E}(\xi) \right\| \right) \right] \leq \exp\left\{ \frac{\sigma^2 c^2}{2(n - cM)}\right\}.
\]

Taking \( c = \frac{nc}{\sigma^2 + Mc} \in (0, n/M) \) and noting that \( \cosh(ce) \geq e^c \), the conclusion is obtained. \(\blacksquare\)
Appendix B. A technical lemma

The following lemma is straightforward, and is used in the literature to yield learning rates in expectation, from the learning rates with probability. We include it just for completeness.

**Lemma B.1** Let $R$ be a non-negative random variable. Let $\alpha, \gamma > 0$ and $\beta \geq 1$. If for any $0 < \delta < 1$, with probability at least $1 - \delta$,

$$R \leq \alpha \log^\gamma \frac{\beta}{\delta},$$

then for any real number $\mu > 0$,

$$[\mathbb{E}(R^\mu)]^{1/\mu} \leq \alpha \left[ \beta \Gamma(\mu \gamma + 1) \right]^{1/\mu},$$

where $\Gamma(t) = \int_0^\infty e^{-u} u^{t-1} du$ is the Gamma function.

**Proof.** For any real number $\mu > 0$ and $0 < \delta < 1$, we have $\text{Prob}(R^\mu > \alpha^\mu \log^{\mu \gamma} \frac{\beta}{\delta}) \leq \delta$. Let

\[t = \alpha^\mu \log^{\mu \gamma} \frac{\beta}{\delta} > \alpha^\mu \log^{\mu \gamma} \beta\]

to give $\delta = \beta \exp\left\{- \left(\frac{t}{\alpha^\mu}\right)^{\frac{1}{\mu \gamma}}\right\}$. So, when $t > \alpha^\mu \log^{\mu \gamma} \beta$,

$$\text{Prob}\{R^\mu > t\} \leq \beta \exp\left\{- \left(\frac{t}{\alpha^\mu}\right)^{\frac{1}{\mu \gamma}}\right\}. \quad (55)$$

When $0 < t < \alpha^\mu \log^{\mu \gamma} \beta$, the bound (55) also holds true because its right-hand side is greater than 1,

$$\beta \exp\left\{- \left(\frac{t}{\alpha^\mu}\right)^{\frac{1}{\mu \gamma}}\right\} > \beta \exp\{- \log \beta\} = 1.$$

Therefore

\[
\mathbb{E}[R^\mu] = \int_0^\infty \text{Prob}(R^\mu > t) dt \leq \beta \int_0^\infty \exp\left\{- \left(\frac{t}{\alpha^\mu}\right)^{\frac{1}{\mu \gamma}}\right\} dt = \beta \alpha^\mu \mu \gamma \int_0^\infty e^{-u} u^{\mu \gamma - 1} du = \alpha^\mu \beta \Gamma(\mu \gamma + 1) \cdot _u^{u^{\mu \gamma} = t/\alpha^\mu} .
\]

The proof is complete. ■

References


Distributed Minimum Error Entropy Algorithms


