Journal of Machine Learning Research 21 (2020) 1-31 Submitted 10/18; Revised 10/19; Published 7/20

Distributed Minimum Error Entropy Algorithms

Xin Guo XIN.GUO@UQ.EDU.AU
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 TINGHU@WHU.EDU.CN
School of Mathematics and Statistics
Wuhan University
Wuhan, 430072, China

Qiang Wu QWU@MTSU.EDU
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 the 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

$$H(E) = -\log \mathbb{E}(p_E) = -\log \int_E p_E^2(e)de, \tag{1}$$

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 optimization 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 ϵ is the noise in the regression model. Let ρ be a Borel probability measure on the product space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Let $\rho_{\mathcal{X}}$ and $\rho(y|x)$ denote the marginal distribution of ρ 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 ρ, 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 $1 \le i \le |D|$ and p_E 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 $a \ge 0$, associated to which are the constant $c = \frac{1}{\sqrt{\pi}}$ and the Gaussian kernel $cG_h(t) = \frac{1}{\sqrt{\pi}h} \exp(-\frac{t^2}{h^2})$. 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

$$\begin{aligned}\mathcal{R}_D(g) &= -\frac{h^2}{|D|^2} \sum_{i=1}^{|D|} \sum_{j=1}^{|D|} G\left(\frac{(e_i - e_j)^2}{h^2}\right) \\ &= -\frac{h^2}{|D|^2} \sum_{\substack{(x,y)\in D \\ (u,v)\in D}} G\left(\frac{[(g(x) - y) - (g(u) - v)]^2}{h^2}\right). \end{aligned} \tag{2}$$

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 a pairwise reproducing kernel Hilbert space (RKHS) (Ying and Zhou, 2016) $\mathcal{H}_K$ on $\mathcal{X}^2 := \mathcal{X}\times\mathcal{X}$. Denote by $K : \mathcal{X}^2\times\mathcal{X}^2 \to \mathbb{R}$ a pairwise Mercer kernel, which is continuous, symmetric and positive semi-definite. The pairwise RKHS $\mathcal{H}_K$ is defined to be the completion of the linear span of the set of functions $\{K_{(x,u)}(\cdot) := K((x,u),(\cdot,\cdot)) : (x,u)\in\mathcal{X}^2\}$ with respect to the inner product that satisfies $\langle K_{(x,u)}, K_{(x',u')}\rangle_K = K((x,u),(x',u'))$ for any $(x,u),(x',u')\in\mathcal{X}^2$.

We replace the hypothesis function difference $g(x)-g(u)$ in (2), by a pairwise function $f(x,u)$ in $\mathcal{H}_K$, and generalize the scheme (2) to the *pairwise empirical risk*

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

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

$$K((x,u),(x',u')) = W(x,x')+W(u,u')-W(x,u')-W(u,x') \tag{3}$$

(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 $D=\{(x_i,y_i)\}_{i=1}^{|D|}$, the regularized MEE algorithm with an RKHS $\mathcal{H}_K$ in supervised learning is defined by*

$$f_{D,\lambda} := \arg\min_{f\in\mathcal{H}_K} \mathcal{E}_D(f)+\lambda\|f\|_K^2, \tag{4}$$

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

$$\bar{f}_{D,\lambda} = \frac{1}{k}\sum_{l=1}^{k} f_{D_l,\lambda}. \tag{5}$$

In this paper we study the convergence of $\bar{f}_{D,\lambda}$ to f_ρ in the square integrable space $(L^2_{\rho_{\mathcal{X}^2}}, \|\cdot\|_\rho)$, where

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

Below we elaborate three important assumptions to carry out the analysis. The first assumption (6) is about the regularity of the target function f_ρ. Define the integral operator $L_K : L^2_{\rho_{\mathcal{X}^2}} \to L^2_{\rho_{\mathcal{X}^2}}$ associated with the pairwise kernel K by

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

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 $f_\rho(x,u)$, given by

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

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

The second assumption (7) is about the capacity of $\mathcal{H}_K$, measured by the *effective dimension* (Zhang, 2002; Caponnetto and Yao, 2010; Blanchard and Krämer, 2016)

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

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

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

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

The third assumption (8) is about the conditional probability distribution $\rho(y|x)$ 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 X, such that for any integer $q \geq 2$,

$$\mathbb{E}\left(|Y|^q | X\right) \leq \frac{1}{2} q! \sigma^2 M^{q-2}. \tag{8}$$

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

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

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

$$\sup_{(x,u) \in \mathcal{X}^2} \sqrt{K((x,u),(x,u))} = 1. \tag{10}$$

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,*

$$\begin{aligned}\|\bar{f}_{D,\lambda} - f_\rho\|_\rho \le \|h_\rho\|_\rho \lambda^{\min\{r,1\}} + (2\|f_\lambda\|_K + 8M + 8\sigma)\mathcal{A}_{D,\lambda,k} \log\frac{8}{\delta} \\ + 128(\|f_\lambda\|_K + M + \sigma)\lambda^{-\frac{1}{2}} \left(\log\frac{16k}{\delta}\right)^4 \max_{1\le l\le k}\left(\frac{\mathcal{A}^2_{D_l,\lambda}}{\lambda} + 1\right)\mathcal{A}^2_{D_l,\lambda} \\ + 16c_{p,\sigma,M}\left(1 + \lambda^{p+\frac{1}{2}}\right) h^{-2p}\lambda^{-p-1}\left(\log\frac{16k}{\delta}\right)^3 \left(\log\frac{16|D|}{\delta}\right)^{2p+1} \\ \times \left(1 + \lambda^{-\frac{1}{2}} \max_{1\le l\le k}\left(\frac{\mathcal{A}^2_{D_l,\lambda}}{\lambda} + 1\right)\mathcal{A}_{D_l,\lambda}\right) \end{aligned} \tag{11}$$

where the constant $c_{p,\sigma,M}$ is independent of D, δ, 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 $|D|/4$, $\mathcal{A}_{D,\lambda,k} := \frac{k}{|D|\sqrt{\lambda}} + \frac{1}{\lfloor |D|/4\rfloor\sqrt{\lambda}} + \sqrt{\frac{\mathcal{N}(\lambda)}{\lfloor |D|/4\rfloor}}$, $\mathcal{A}_{D,\lambda} := \mathcal{A}_{D,\lambda,1}$, and $\mathcal{A}_{D_l,\lambda} := \mathcal{A}_{D_l,\lambda,1}$.

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

$$\lambda = \begin{cases} |D|^{-\frac{1}{1+s}}, & \text{for } 0 < r \le \frac{1}{2}, \\ |D|^{-\frac{1}{s+2\min\{1,r\}}}, & \text{for } r > \frac{1}{2}. \end{cases}$$

and

$$k \begin{cases} = 1, & \text{for } 0 < r \le \frac{1}{2}, \\ \le \lambda^{-\min\left\{r-\frac{1}{2},\frac{1}{2}\right\}} \log^{-4}|D|, & \text{for } r > \frac{1}{2}. \end{cases} \tag{12}$$

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

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

where C_1 is a constant independent of D, δ, 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 $h \ge \left[\lambda^{-r-1-p}(\log|D|)^{2p+4}\right]^{\frac{1}{2p}}$, then with probability at least $1 - \delta$,*

$$\|\bar{f}_{D,\lambda} - f_\rho\|_\rho \le C_1 \begin{cases} |D|^{-\frac{r}{s+1}}\left(\log\frac{16}{\delta}\right)^{2p+4}, & \text{for } 0 < r \le \frac{1}{2}, \\ |D|^{-\min\left\{\frac{r}{s+2r},\frac{1}{s+2}\right\}}\left(\log\frac{16}{\delta}\right)^{2p+4}, & \text{for } r > \frac{1}{2}, \end{cases} \tag{14}$$

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

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

In particular, since the only assumption (12) on k permits the case $k = 1$, the above bounds (14) and (15) for $\bar{f}_{D,\lambda}$ 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 pariwise 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 $L^2_{\rho_{\mathcal{X}^2}}$ (note that $\rho_{\mathcal{X}^2}$ is a probability measure)

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

In particular, if $f(x,u) = g(x) - g(u)$ for some $g \in L^2_{\rho_{\mathcal{X}}}$, then $\mathcal{P}f = f$, $\mathcal{M}f = g - \int_{\mathcal{X}} g(u) d\rho_{\mathcal{X}}(u)$, and $\int_{\mathcal{X}} (\mathcal{M}f)(x) d\rho_X(x) = 0$. So $\left\|\bar{f}_{D,\lambda} - f_\rho\right\|_\rho^2 \geq \left\|\mathcal{P}\bar{f}_{D,\lambda} - f_\rho\right\|_\rho^2 = 2\left\|(\mathcal{M}\bar{f}_{D,\lambda} + g^* - \mathcal{M}f_\rho) - g^*\right\|^2_{L^2_{\rho_X}}$. With data D, one replaces $\mathcal{M}$ by $\mathcal{M}_D : C(\mathcal{X}^2) \to C(\mathcal{X})$,

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

and replaces the difference $g^* - \mathcal{M}f_\rho = \int_X g^*(x) d\rho_X(x)$ by its unbiased and efficient estimator $\frac{1}{|D|}\sum_{(x,y)\in D} y$.

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 $D = \cup_{l=1}^k D_l$ (with disjoint and equal-sized subsets $D_1, \ldots, D_k$), assume that we also have an unlabeled data set $\tilde{D} = \{\tilde{x}_i\}_{i=1}^{|\tilde{D}|}$. We assume that the input observations $\tilde{x}_i$ are drawn independently from $\rho_{\mathcal{X}}$, and $\tilde{D}$ is independent of D. For technical simplicity we assume that $\tilde{D}$ is also divided randomly into k disjoint and equal-sized subsets $\tilde{D} = \cup_{l=1}^k \tilde{D}_l$. We define the semi-supervised training data set by $D^* = \cup_{l=1}^k D_l^*$, where for each $1 \leq l \leq k$, we write $D_l = \{(x_i, y_i)\}_{i=1}^{|D_l|}$, $\tilde{D}_l = \{\tilde{x}_i\}_{i=1}^{|\tilde{D}_l|}$, and define $D_l^* = \{(x_i^*, y_i^*)\}_{i=1}^{|D_l^*|}$ with $|D_l^*| = |D_l| + |\tilde{D}_l|$ by

$$(x_i^*, y_i^*) = \begin{cases} (x_i, \frac{|D_l^*|}{|D_l|} y_i), & \text{for } 1 \leq i \leq |D_l|, \\ (\tilde{x}_{i-|D_l|}, 0), & \text{for } |D_l| + 1 \leq i \leq |D_l^*|. \end{cases}$$

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

$$\bar{f}_{D^*,\lambda} = \frac{1}{k}\sum_{l=1}^k f_{D_l^*,\lambda}. \tag{16}$$

In this subsection, we assume that K is antisymmetric. That is, we assume $K_{(x,u)} = -K_{(u,x)}$.

Theorem 2.6 *The following bound holds with probability at least* $1-\delta$.

$$\begin{aligned}\|\bar{f}_{D^*,\lambda}-f_\rho\|_\rho \leq & \|h_\rho\|_\rho \lambda^{\min\{r,1\}} + 256(1+M+\sigma)\lambda^{-1/2}\left(\log\frac{16k}{\delta}\right)^4 \\ & \times \max_{1\leq l\leq k}\left(\frac{\mathcal{A}^2_{D_l^*,\lambda}}{\lambda}+1\right)\mathcal{A}_{D_l^*,\lambda}\left(\mathcal{A}_{D_l^*,\lambda}\|f_\lambda\|_K+\mathcal{A}_{D_l,D_l^*,\lambda}\right) \\ & + 16\lambda^{-1/2}C_2 h^{-2p}\left(\log\frac{16k}{\delta}\right)^3 \max_{1\leq l\leq k}\left[\left(\frac{\mathcal{A}^2_{D_l^*,\lambda}}{\lambda}+1\right)\mathcal{A}_{D_l^*,\lambda}\lambda^{-1/2}+1\right] \\ & \times \Delta_{D_l,D_l^*,\lambda}\left(\log\frac{16|D|}{\delta}\right)^{2p+1} \\ & + (2\mathcal{A}_{D^*,\lambda,k}\|f_\lambda\|_K + 8(M+\sigma)\mathcal{A}_{D,D^*,\lambda,k})\log\frac{16}{\delta}, \end{aligned} \tag{17}$$

where $\mathcal{A}_{D_l^*,\lambda}$ *is defined in the same way as* $\mathcal{A}_{D,\lambda}$ *in Theorem 2.2 by substituting* D *with* D_l^*, $\mathcal{A}_{D,D^*,\lambda,k}=\frac{k}{|D^*|\sqrt{\lambda}}+\frac{1}{\lfloor |D|/4\rfloor\sqrt{\lambda}}+\sqrt{\frac{\mathcal{N}(\lambda)}{\lfloor |D|/4\rfloor}}$, $\mathcal{A}_{D,D^*,\lambda}=\mathcal{A}_{D,D^*,\lambda,1}$,

$$\Delta_{D,D^*,\lambda}=\left(\frac{|D^*|}{\lambda|D|}\right)^{p+\frac{1}{2}}+\left(\frac{|D^*|}{|D|}\right)^{2p+1},$$

and C_2 *is a constant independent of* D, D^*, k, h, *or* δ, *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_ρ 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* $|D^*|\geq\max\{|D|^{\frac{s+1}{2\min\{r,1\}+s}},2|D|\}$. *Let* $\lambda=|D|^{-\frac{1}{2\min\{r,1\}+s}}$, *and*

$$k\leq(\log|D|)^{-4}\min\left\{\sqrt{|D^*|\lambda^{1+s}},(|D^*|\lambda^{2-2\min\{r,1\}-s})^{1/3}\right\}. \tag{18}$$

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

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

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

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

$$h\geq\left[|D|^{\frac{3}{2+s}}\Delta_{D,D^*,\lambda}(\log|D|)^{2p+4}\right]^{\frac{1}{2p}}, \tag{19}$$

then with probability $1-\delta$,

$$\|\bar{f}_{D^*,\lambda} - f_\rho\|_\rho \leq C_3 |D|^{-\min\left\{\frac{r}{2r+s}, \frac{1}{2+s}\right\}} \left(\log \frac{16}{\delta}\right)^{2p+4}, \tag{20}$$

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* $\mu > 0$,

$$\left[\mathbb{E}\left(\|\bar{f}_{D^*,\lambda} - f_\rho\|_\rho^\mu\right)\right]^{1/\mu} \leq \left[16\Gamma((2p+4)\mu+1)\right]^{1/\mu} C_3 |D|^{-\min\left\{\frac{r}{2r+s}, \frac{1}{2+s}\right\}}. \tag{21}$$

In particular, since the only assumption (18) on k *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 $\gamma_i \leq C_0 i^{-\frac{1}{b}}$ for some $0 < b < 1$ and $C_0 > 0$, then

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

The condition (7) roughly measures the smoothness of K. For example, if $K \in C^\alpha(\mathcal{X}^2 \times \mathcal{X}^2)$ with some integer $\alpha \geq 1$, and $\mathcal{X}^2$ is locally the graph of a Lipschitz function, then (7) is satisfied with $s = \left(\frac{\alpha}{2\dim(\mathcal{X})} + \frac{1}{2}\right)^{-1}$ (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).

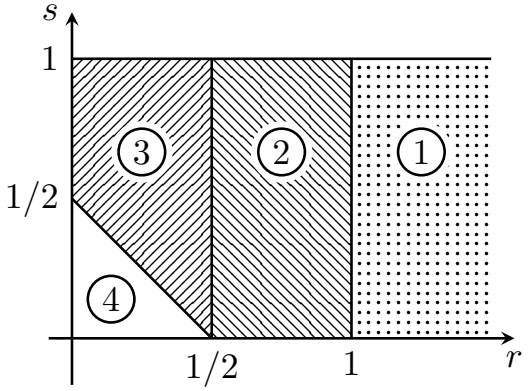

Figure 1: To organize the discussion of DMEE learning rates, we divide the space of regularity and capacity parameters into several parts. As we discussed at the beginning of Section 3, one always has $\mathcal{N}(\lambda) \leq C_0 \lambda^{-1}$, so we exclude the area $s > 1$.

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

squares (Lin et al., 2017). That is, when the regularity index r exceeds 1, its further increasing does not help to improve the algorithm convergence. This saturation is proved overcome by spectral algorithms (Lo Gerfo et al., 2008; Mücke and Blanchard, 2018; Blanchard and Mücke, 2018; Guo et al., 2017a), bias corrected approach (Guo et al., 2017b), or a gradient descent approach for MEE (Hu et al., 2020) without Tikhonov regularization.

In Area 2, DMEE achieves the learning rates which equal the minimax optimal rates $O(|D|^{-\frac{r}{2r+s}})$ 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 can not 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 $|D|^{1-\frac{r-(1/2)}{s+2r}}\log^4|D|$ for Area 2, and to $|D|^{1-\frac{1/2}{s+2}}\log^4|D|$ 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 $O(|D|^{-\frac{r}{2r+s}})$ for point kernel-based regression for $0 < r < \frac{1}{2}$ is derived in Steinwart et al. (2009) with the boundedness assumption of $L_K^r : L_2 \to L_\infty$. Fully supervised DMEE does not achieve this lower bound. Unlabeled data improves DMEE in two ways. First, semi-supervised DMEE achieves the learning rates $O(|D|^{-\frac{r}{2r+s}})$. 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 $O(|D|^{-\frac{r}{s+1}})$, and our analysis of semi-supervised DMEE fails to improve the rate to $O(|D|^{-\frac{r}{2r+s}})$. 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. 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 $\mathcal{H}$, $|y| \leq M$ and the logarithm of the covering number of the hypothesis space $\mathcal{H}$

by $C(\mathcal{X})$ balls of radius ϵ grows no faster than ϵ^{-p} for some index $p > 0$ when ϵ decays to zero, then with high probability the learning rate is of order $O\left(|D|^{-\frac{1}{2(1+p)}}\right)$. To elaborate it clearly, assume $\mathcal{H}$ to be an RKHS induced either by a pointwise kernel $W \in C^\alpha(\mathcal{X} \times \mathcal{X})$ or a pairwise kernel $K \in C^\alpha(\mathcal{X}^2 \times \mathcal{X}^2)$ and $\mathcal{X} \subset \mathbb{R}^d$ satisfied certain mild regularity conditions. By Cucker and Zhou (2007, page 72, Theorem 5.1) and Mendelson and Neeman (2010), the covering number index $p = 2d/\alpha$ while the effective dimension index $s = 2d/(d+\alpha)$. When target function lies in $\mathcal{H}_K$, i.e. $r \geq \frac{1}{2}$, the rate $O\left(|D|^{-\frac{1}{2(1+p)}}\right)$ in (Hu et al., 2015) is always inferior to $O\left(|D|^{-\frac{r}{2r+s}}\right)$ 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 $x \in \mathcal{X}$, write $\boldsymbol{W}(x) = (W(x, x_i))_{i=1}^{|D|}$ a column vector of dimension $|D|$. Write $\mathbb{C} = (c_{i,j})_{i,j=1}^{|D|}$ the coefficient matrix of the function

$$f_{\mathbb{C}}(x, u) = \sum_{i=1}^{|D|} \sum_{j=1}^{|D|} c_{i,j} K((x_i, x_j), (x, u)). \tag{22}$$

By the representer theorem, the solution to (4) takes the form of (22). Meanwhile, we write $\mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^{|D|}$ and use the kernel structure (3) to obtain that $f_{\mathbb{C}}(x, u) = \boldsymbol{c}^T(\boldsymbol{W}(x) - \boldsymbol{W}(u))$, where $\boldsymbol{c} = (\mathbb{C} - \mathbb{C}^T)\mathbf{1} \in \mathbb{R}^{|D|}$. We also have $\|f_{\mathbb{C}}\|_K^2 = \boldsymbol{c}^T \mathbb{W} \boldsymbol{c}$, 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 $\boldsymbol{c}$. Since $\mathbf{1}^T \boldsymbol{c} = 0$, the vector $\boldsymbol{c}$ 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 a 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 $L_{K,D} : \mathcal{H}_K \to \mathcal{H}_K$ by

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

so for any $f \in \mathcal{H}_K$, $L_{K,D} f = \dfrac{1}{|D|^2} \displaystyle\sum_{w,z \in D} f(x, u) K_{(x,u)}$. Then we have the following representation for $f_{D,\lambda}$.

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

$$f_{D,\lambda} = (L_{K,D} + \lambda I)^{-1} \hat{f}_{\rho,D} + (L_{K,D} + \lambda I)^{-1} E_{D,\lambda} \tag{23}$$

where

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

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. ■

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

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

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

$$\begin{aligned}
\lambda \|f_{D,\lambda}\|_K^2 \le & \mathcal{E}_D(0) - \mathcal{E}_D(f_{D,\lambda}) \le -\frac{h^2}{|D|^2} \sum_{w,z\in D} G\left(\frac{(y-v)^2}{h^2} \right) + h^2 G(0) \\
\le & \frac{C_G}{|D|^2} \sum_{w,z\in D} (y-v)^2 \le \frac{C_G}{|D|^2} \sum_{w,z\in D} 2(y^2 + v^2) \le 4 C_G \max_{w\in D} |y|^2.
\end{aligned}$$

It follows that

$$\|f_{D,\lambda}\|_K \le 2\sqrt{C_G} \lambda^{-\frac{1}{2}} \max_{w\in D} |y|. \tag{25}$$

By (9), we see that

$$\begin{aligned}
\|E_{D,\lambda}\|_K &\le c_p h^{-2p} \frac{1}{|D|^2} \sum_{w,z\in D} (\|f_{D,\lambda}\|_K + |y-v|)^{2p+1} \\
&\le 2^{2p} c_p h^{-2p} \frac{1}{|D|^2} \sum_{w,z\in D} \left(\|f_{D,\lambda}\|_K^{2p+1} + |y-v|^{2p+1} \right) \\
&\le 2^{2p} c_p h^{-2p} \left(\|f_{D,\lambda}\|_K^{2p+1} + 2^{2p+1} \max_{w\in D} |y|^{2p+1} \right).
\end{aligned} \tag{26}$$

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

$$\|E_{D,\lambda}\|_K \le c_{p,\sigma,M} \left(\lambda^{-(p+\frac{1}{2})} + 1 \right) h^{-2p} \left(\log \frac{|D|}{\delta} \right)^{2p+1} \tag{27}$$

where $c_{p,\sigma,M} := 2^{4p+1} c_p (C_G^{p+\frac{1}{2}} + 1)(4M + 5\sigma)^{2p+1}$.

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_λ in $\mathcal{H}_K$, defined by

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

where $\mathcal{E}_{\mathsf{ls}}(f) = \int_{\mathcal{Z}^2}(f(x,u)-y+v)^2 d\rho(x,y)d\rho(u,v)$ is the expected risk associated with the pairwise square loss. Similar to the argument in Smale and Zhou (2007), we can verify that

$$f_\lambda = (L_K+\lambda I)^{-1}L_K f_\rho, \tag{28}$$

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 $L^2_{\rho_{\mathcal{X}^2}}$ is no greater than 1, thanks to the assumption (10). Under the regularity assumption (6) with $r>0$,

$$\|f_\lambda - f_\rho\|_\rho \le \begin{cases} \|h_\rho\|_\rho \lambda^r, & \text{when } 0<r\le 1,\\ \|h_\rho\|_\rho\lambda, & \text{when } r>1,\end{cases} \tag{29}$$

and

$$\|f_\lambda\|_K \le \begin{cases} \|h_\rho\|_\rho \lambda^{r-\frac12}, & \text{when } 0<r<1/2,\\ \|h_\rho\|_\rho, & \text{when } r\ge 1/2.\end{cases} \tag{30}$$

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

$$-f_\lambda = (L_{K,D}+\lambda I)^{-1}[(L_K - L_{K,D})f_\lambda - L_K f_\rho], \tag{31}$$

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

$$\begin{aligned} f_{D,\lambda}-f_\lambda =& (L_{K,D}+\lambda I)^{-1}(L_K-L_{K,D})f_\lambda + (L_{K,D}+\lambda I)^{-1}\left(\hat f_{\rho,D} - L_K f_\rho\right)\\ &+ (L_{K,D}+\lambda I)^{-1}E_{D,\lambda}. \end{aligned} \tag{32}$$

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

$$\begin{aligned} f_{D,\lambda}-f_\lambda &= f_{D,\lambda} - (L_K+\lambda I)^{-1}L_K f_\rho = (L_K+\lambda I)^{-1}[(L_K+\lambda I)f_{D,\lambda} - L_K f_\rho]\\ &= (L_K+\lambda I)^{-1}(L_K-L_{K,D})(f_{D,\lambda}-f_\lambda) + (L_K+\lambda I)^{-1}(L_K-L_{K,D})f_\lambda\\ &\quad + (L_K+\lambda I)^{-1}(\hat f_{\rho,D} - L_K f_\rho) + (L_K+\lambda I)^{-1}E_{D,\lambda}. \end{aligned} \tag{33}$$

In the sequel, we denote by $\|\cdot\|_{\mathsf{op}}$ the operator norm from $\mathcal{H}_K$ to itself, and

$$\begin{aligned} \mathcal{B}_{D,\lambda} &= \|(L_{K,D}+\lambda I)^{-1}(L_K+\lambda I)\|_{\mathsf{op}},\\ \mathcal{C}_{D,\lambda} &= \|(L_K+\lambda I)^{-\frac12}(L_K-L_{K,D})\|_{\mathsf{op}},\\ \mathcal{D}_{D,\lambda} &= \left\|\frac1k\sum_{l=1}^k (L_K+\lambda I)^{-\frac12}(L_K-L_{K,D_l})\right\|_{\mathsf{op}},\\ \mathcal{F}_{D,\lambda} &= \left\|\frac1k\sum_{l=1}^k (L_K+\lambda I)^{-\frac12}(\hat f_{\rho,D_l} - L_K f_\rho)\right\|_K,\\ \mathcal{G}_{D,\lambda} &= \|(L_K+\lambda I)^{-\frac12}(\hat f_{\rho,D} - L_K f_\rho)\|_K. \end{aligned}$$

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 $\|\cdot\|_{\mathsf{op}(\mathcal{H})}$ denote the operator norm. Then*

$$\|A^s B^s\|_{\mathsf{op}(\mathcal{H})} \le \|AB\|^s_{\mathsf{op}(\mathcal{H})}, \quad \text{for any } 0 \le s \le 1.$$

■

Noting that for any $f \in \mathcal{H}_K$,

$$\max\{\|f\|_\rho, \sqrt{\lambda}\|f\|_K\} \le \|(L_K + \lambda I)^{\frac{1}{2}} f\|_K \tag{34}$$

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

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

$$\|\bar{f}_{D,\lambda} - f_\lambda\|_\rho \le S_1 + S_2 + \mathcal{D}_{D,\lambda}\|f_\lambda\|_K + \mathcal{F}_{D,\lambda}, \tag{35}$$

where

$$S_1 = \max_{1\le l\le k} \left(\mathcal{B}_{D_l,\lambda}\mathcal{C}^2_{D_l,\lambda}\|f_\lambda\|_K \lambda^{-\frac{1}{2}} + \mathcal{B}_{D_l,\lambda}\mathcal{C}_{D_l,\lambda}\mathcal{G}_{D_l,\lambda}\lambda^{-\frac{1}{2}}\right)$$

and

$$S_2 = \max_{1\le l\le k} \left(\mathcal{B}_{D_l,\lambda}\mathcal{C}_{D_l,\lambda}\lambda^{-1}\|E_{D_l,\lambda}\|_K + \frac{1}{\sqrt{\lambda}}\|E_{D_l,\lambda}\|_K\right).$$

Proof. Let I_1, I_2, and I_3 denote the three terms on the right-hand side of (32), respectively. Consider the $\mathcal{H}_K$ 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,

$$\begin{aligned}
&\|(L_K + \lambda I)^{1/2} I_1\|_K \\
\le&\|(L_K + \lambda I)^{1/2}(L_{K,D} + \lambda I)^{-1/2}\|_{\mathsf{op}}\|(L_{K,D} + \lambda I)^{-1/2}(L_K + \lambda I)^{1/2}\|_{\mathsf{op}} \\
&\times \|(L_K + \lambda I)^{-1/2}(L_K - L_{K,D})\|_{\mathsf{op}}\|f_\lambda\|_K \\
\le&\mathcal{B}_{D,\lambda}\mathcal{C}_{D,\lambda}\|f_\lambda\|_K.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\|(L_K + \lambda I)^{1/2} I_2\|_K \\
\le&\|(L_K + \lambda I)^{1/2}(L_{K,D} + \lambda I)^{-1}(L_K + \lambda I)^{1/2}\|_{\mathsf{op}}\|(L_K + \lambda I)^{-1/2}(\hat{f}_{\rho,D} - L_K f_\rho)\|_K \\
\le&\mathcal{B}_{D,\lambda}\mathcal{G}_{D,\lambda},
\end{aligned}$$

and

$$\begin{aligned}
&\|(L_K + \lambda I)^{1/2} I_3\|_K \\
\le&\|(L_K + \lambda I)^{1/2}(L_{K,D} + \lambda I)^{-1}(L_K + \lambda I)^{1/2}\|_{\mathsf{op}}\frac{1}{\sqrt{\lambda}}\|E_{D,\lambda}\|_K \\
\le&\lambda^{-1/2}\mathcal{B}_{D,\lambda}\|E_{D,\lambda}\|_K.
\end{aligned}$$

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

$$\|f_{D,\lambda} - f_\lambda\|_\rho \le \mathcal{B}_{D,\lambda}\mathcal{C}_{D,\lambda}\|f_\lambda\|_K + \mathcal{B}_{D,\lambda}\mathcal{G}_{D,\lambda} + \lambda^{-\frac{1}{2}}\mathcal{B}_{D,\lambda}\|E_{D,\lambda}\|_K$$

and

$$\|f_{D,\lambda} - f_\lambda\|_K \leq \mathcal{B}_{D,\lambda}\mathcal{C}_{D,\lambda}\|f_\lambda\|_K\lambda^{-\frac{1}{2}} + \mathcal{B}_{D,\lambda}\mathcal{G}_{D,\lambda}\lambda^{-\frac{1}{2}} + \lambda^{-1}\mathcal{B}_{D,\lambda}\|E_{D,\lambda}\|_K. \tag{36}$$

By the fact $\bar{f}_{D,\lambda} - f_\lambda = \frac{1}{k}\sum_{l=1}^{k}(f_{D_l,\lambda} - f_\lambda)$ and the second decomposition (33), one obtains that

$$\begin{aligned}
\|\bar{f}_{D,\lambda} - f_\lambda\|_\rho &\leq \frac{1}{k}\sum_{l=1}^{k}\|(L_K+\lambda I)^{-\frac{1}{2}}(L_K - L_{K,D_l})(f_{D_l,\lambda} - f_\lambda)\|_K \\
&+ \left\|\frac{1}{k}\sum_{l=1}^{k}(L_K+\lambda I)^{-\frac{1}{2}}(L_K - L_{K,D_l})f_\lambda\right\|_K \\
&+ \left\|\frac{1}{k}\sum_{l=1}^{k}(L_K+\lambda I)^{-\frac{1}{2}}\left(\hat{f}_{\rho,D_l} - L_K f_\rho\right)\right\|_K + \lambda^{-\frac{1}{2}}\frac{1}{k}\sum_{l=1}^{k}\|E_{D_l,\lambda}\|_K \\
&\leq \mathcal{D}_{D,\lambda}\|f_\lambda\|_K + \mathcal{F}_{D,\lambda} + \max_{1\leq l\leq k}\left(\mathcal{C}_{D_l,\lambda}\|f_{D_l,\lambda} - f_\lambda\|_K + \lambda^{-\frac{1}{2}}\|E_{D_l,\lambda}\|_K\right).
\end{aligned}$$

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

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 $\mathcal{B}_{D,\lambda}, \mathcal{C}_{D,\lambda}, \mathcal{D}_{D,\lambda}, \mathcal{F}_{D,\lambda}$ and $\mathcal{G}_{D,\lambda}$. 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* $1-\delta$.

$$\mathcal{B}_{D,\lambda} \leq 2\left(\frac{2\mathcal{A}_{D,\lambda}\log\frac{2}{\delta}}{\sqrt{\lambda}}\right)^2 + 2, \qquad \mathcal{C}_{D,\lambda} \leq 2\mathcal{A}_{D,\lambda}\log\frac{2}{\delta}, \text{ and}$$
$$\mathcal{D}_{D,\lambda} \leq 2\mathcal{A}_{D,\lambda,k}\log\frac{2}{\delta}.$$

■

For other two quantities $\mathcal{F}_{D,\lambda}$ and $\mathcal{G}_{D,\lambda}$, 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* $1-\delta$.

$$\mathcal{F}_{D,\lambda} \leq 8(M+\sigma)\mathcal{A}_{D,\lambda,k}\log\frac{2}{\delta}, \quad \text{and} \quad \mathcal{G}_{D,\lambda} \leq 8(M+\sigma)\mathcal{A}_{D,\lambda}\log\frac{2}{\delta}.$$

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,

$$\mathbb{E}\|\xi\|_K \leq 2\lambda^{-\frac{1}{2}}\mathbb{E}|Y| \leq 2\lambda^{-\frac{1}{2}}(\mathbb{E}|Y|^2)^{\frac{1}{2}} \leq 2\sigma\lambda^{-\frac{1}{2}}. \tag{37}$$

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

bound for any integer $q \geq 2$.

$$\begin{aligned}
\mathbb{E}[\|\xi - \mathbb{E}\xi\|_K^q] &\leq 2^{q-1}\mathbb{E}[\|\xi\|_K^q] + 2^{q-1}\|\mathbb{E}\xi\|_K^q \leq 2^q \mathbb{E}[\|\xi\|_K^q] \\
&\leq 2^{2q} \sup_{x',u' \in \mathcal{X}} \|(L_K + \lambda I)^{-1/2} K_{(x',u')}\|_K^{q-2} \mathbb{E}\left[|Y|^q \|(L_K + \lambda I)^{-1/2} K_{(x,u)}\|_K^2\right] \\
&\leq 2^{2q} \left(\frac{1}{\sqrt{\lambda}}\right)^{q-2} \mathbb{E}\left[\mathbb{E}(|Y|^q|X)\|(L_K + \lambda I)^{-1/2} K_{(x,u)}\|_K^2\right] \\
&\leq 2^{2q} \left(\frac{1}{\sqrt{\lambda}}\right)^{q-2} \frac{1}{2} q! \sigma^2 M^{q-2} \mathcal{N}(\lambda) \\
&= 8 q! \sigma^2 \mathcal{N}(\lambda) (4M/\sqrt{\lambda})^{q-2}.
\end{aligned}$$

Let π be a permutation of the set $\{1, \ldots, |D_l| = |D|/k\}$ of integers. Then $\{z^l_{\pi(1)}, \ldots, z^l_{\pi(|D_l|)}\}$ is the associate permutation of $D_l = \{z_i^l = (x_i^l, y_i^l)\}_{i=1}^{|D_l|}$. Let

$$U_\pi^l = \frac{1}{\lfloor |D_l|/2 \rfloor} \sum_{i=1}^{\lfloor |D_l|/2 \rfloor} \xi(z^l_{\pi(2i-1)}, z^l_{\pi(2i)}).$$

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

$$(L_K + \lambda I)^{-1/2} \hat{f}_{\rho, D_l} = \frac{|D_l| - 1}{|D_l|} \frac{1}{|D_l|(|D_l| - 1)} \sum_{w,z \in D_l} \xi(w,z) = \frac{|D_l| - 1}{|D_l|} \frac{1}{|D_l|!} \sum_\pi U_\pi^l,$$

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

$$\frac{1}{k} \sum_{l=1}^k (L_K + \lambda I)^{-1/2} \hat{f}_{\rho, D_l} = \frac{|D_1| - 1}{|D_1|} \frac{1}{|D_1|!} \sum_\pi \frac{1}{k} \sum_{l=1}^k U_\pi^l.$$

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

$$\begin{aligned}
\mathcal{F}_{D,\lambda} \leq & \frac{|D_1| - 1}{|D_1|} \left\| \frac{1}{|D_1|!} \sum_\pi \frac{1}{k} \sum_{l=1}^k U_\pi^l - \mathbb{E}\xi \right\|_K + \frac{1}{|D_1|} \|\mathbb{E}\xi\|_K \\
\leq & \left\| \frac{1}{|D_1|!} \sum_\pi \frac{1}{k} \sum_{l=1}^k U_\pi^l - \mathbb{E}\xi \right\|_K + \frac{2\sigma k}{|D| \sqrt{\lambda}}. \qquad (38)
\end{aligned}$$

We observe that for each π, $\frac{1}{k} \sum_{l=1}^k U_\pi^l$ is the average of $k\lfloor |D_1|/2 \rfloor$ independent copies of $\xi(z,z')$ with z and z' independently drawn from ρ, and $k\lfloor |D_1|/2 \rfloor \geq \lfloor |D|/4 \rfloor$ (in fact, recall our assumption $|D_1| \geq 4$ and $k|D_1| = |D|$, obviously when $|D_1|$ is even, $k\lfloor |D_1|/2 \rfloor = k|D_1|/2 = |D|/2 \geq \lfloor |D|/4 \rfloor$, and when $|D_1|$ is odd, one still has $k\lfloor |D_1|/2 \rfloor = k(|D_1| - 1)/2 \geq k|D_1|/4 \geq \lfloor |D|/4 \rfloor$). 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 $\epsilon > 0$,

$$\begin{aligned}
&\text{Prob}\left\{\left\|\frac{1}{|D_1|!}\sum_{\pi}\frac{1}{k}\sum_{l=1}^{k}U_{\pi}^{l}-\mathbb{E}\xi\right\|_K\geq\epsilon\right\}\\
&\leq\inf_{c>0}\mathbb{E}\left[\cosh\left(c\left\|\frac{1}{|D_1|!}\sum_{\pi}\frac{1}{k}\sum_{l=1}^{k}U_{\pi}^{l}-\mathbb{E}\xi\right\|_K\right)\right]/\cosh(c\epsilon)\\
&\leq\inf_{c>0}\frac{1}{|D_1|!}\sum_{\pi}\mathbb{E}\left[\cosh\left(c\left\|\frac{1}{k}\sum_{l=1}^{k}U_{\pi}^{l}-\mathbb{E}\xi\right\|_K\right)\right]/\cosh(c\epsilon)\\
&\leq 2\exp\left\{-\frac{\lfloor|D|/4\rfloor\epsilon^2}{2\left(16\mathcal{N}(\lambda)\sigma^2+4\lambda^{-\frac{1}{2}}M\epsilon\right)}\right\}. \qquad (39)
\end{aligned}$$

One takes the right-hand side of (39) as δ and recalls (38) to have that with probability $1-\delta$,

$$\mathcal{F}_{D,\lambda}\leq\frac{2\sigma k}{|D|\sqrt{\lambda}}+\frac{8M\log(2/\delta)}{\lfloor|D|/4\rfloor\sqrt{\lambda}}+\sqrt{\frac{32\mathcal{N}(\lambda)\sigma^2\log(2/\delta)}{\lfloor|D|/4\rfloor}}.$$

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 $\|\bar{f}_{D,\lambda}-f_\rho\|_\rho$ as the *sample error* $\|\bar{f}_{D,\lambda}-f_\lambda\|_\rho$ and the *approximation error* $\|f_\lambda-f_\rho\|_\rho$. As stated in (29), $\|f_\lambda-f_\rho\|_\rho\leq\lambda^r\|h_\rho\|_\rho$ for $0<r\leq 1$. Thus, we just estimate $\|\bar{f}_{D,\lambda}-f_\lambda\|_\rho$ 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,

$$\mathcal{B}_{D_l,\lambda}\mathcal{C}_{D_l,\lambda}^2\lambda^{-\frac{1}{2}}\leq 32\left(\frac{\mathcal{A}_{D_l,\lambda}^2}{\lambda}+1\right)\mathcal{A}_{D_l,\lambda}^2\lambda^{-\frac{1}{2}}\log^4\frac{2}{\delta},$$

$$\mathcal{B}_{D_l,\lambda}\mathcal{C}_{D_l,\lambda}\mathcal{G}_{D_l,\lambda}\lambda^{-\frac{1}{2}}\leq 128(M+\sigma)\left(\frac{\mathcal{A}_{D_l,\lambda}^2}{\lambda}+1\right)\mathcal{A}_{D_l,\lambda}^2\lambda^{-\frac{1}{2}}\log^4\frac{2}{\delta},$$

and

$$\begin{aligned}
&\mathcal{B}_{D_l,\lambda}\mathcal{C}_{D_l,\lambda}\lambda^{-1}\|E_{D_l,\lambda}\|_K\\
\leq&16c_{p,\sigma,M}(1+\lambda^{p+\frac{1}{2}})\left(\frac{\mathcal{A}_{D_l,\lambda}^2}{\lambda}+1\right)\mathcal{A}_{D_l,\lambda}\lambda^{-(p+\frac{3}{2})}h^{-2p}\left(\log\frac{2}{\delta}\right)^3\left(\log\frac{|D_l|}{\delta}\right)^{2p+1}.
\end{aligned}$$

With the notations in (35), it follows that with probability at least $1-4k\delta$, the following two bounds hold true simultaneously,

$$S_1\leq 128(\|f_\lambda\|_K+M+\sigma)\max_{1\leq l\leq k}\left(\frac{\mathcal{A}_{D_l,\lambda}^2}{\lambda}+1\right)\mathcal{A}_{D_l,\lambda}^2\lambda^{-\frac{1}{2}}\log^4\frac{2}{\delta} \qquad (40)$$

and

$$\begin{aligned}
S_2\leq&\,16c_{p,\sigma,M}(1+\lambda^{p+\frac{1}{2}})h^{-2p}\lambda^{-p-1}\left(\log^3\frac{2}{\delta}\right)\left(\log\frac{|D_1|}{\delta}\right)^{2p+1}\\
&\times\left[1+\lambda^{-\frac{1}{2}}\max_{1\leq l\leq k}\left(\frac{\mathcal{A}_{D_l,\lambda}^2}{\lambda}+1\right)\mathcal{A}_{D_l,\lambda}\right]. \qquad (41)
\end{aligned}$$

By Proposition 4.4 and 4.5 again, we see that with probability at least $1-\frac{\delta}{2}$, the following bounds hold simultaneously,

$$\mathcal{D}_{D,\lambda} \le 2\mathcal{A}_{D,\lambda,k}\log\frac{8}{\delta} \quad \text{and} \quad \mathcal{F}_{D,\lambda} \le 8(M+\sigma)\mathcal{A}_{D,\lambda,k}\log\frac{8}{\delta}.$$

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

$$\begin{aligned}
&\|\bar{f}_{D,\lambda}-f_\rho\|_\rho \le \|\bar{f}_{D,\lambda}-f_\lambda\|_\rho + \|f_\lambda - f_\rho\|_\rho \\
&\le \|h_\rho\|_\rho\lambda^r + S_1 + S_2 + \mathcal{D}_{D,\lambda}\|f_\lambda\|_K + \mathcal{F}_{D,\lambda} \\
&\le \|h_\rho\|_\rho\lambda^r + 128(\|f_\lambda\|_K + M + \sigma)\lambda^{-\frac{1}{2}}\left(\log^4\frac{16k}{\delta}\right)\max_{1\le l\le k}\left(\frac{\mathcal{A}^2_{D_l,\lambda}}{\lambda}+1\right)\mathcal{A}^2_{D_l,\lambda} \\
&\quad + 16c_{p,\sigma,M}(1+\lambda^{p+\frac{1}{2}})h^{-2p}\lambda^{-p-1}\left(\log^3\frac{16k}{\delta}\right)\left(\log\frac{16|D|}{\delta}\right)^{2p+1} \\
&\quad \times\left(1+\lambda^{-\frac{1}{2}}\max_{1\le l\le k}\left(\frac{\mathcal{A}^2_{D_l,\lambda}}{\lambda}+1\right)\mathcal{A}_{D_l,\lambda}\right) + (2\|f_\lambda\|_K + 8(M+\sigma))\mathcal{A}_{D,\lambda,k}\log\frac{8}{\delta}.
\end{aligned}$$

The proof is complete. ■

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

$$\begin{aligned}
\mathcal{A}_{D,\lambda,k} =& \frac{k}{|D|\sqrt{\lambda}} + \frac{1}{\lfloor |D|/4\rfloor\sqrt{\lambda}} + \sqrt{\frac{\mathcal{N}(\lambda)}{\lfloor |D|/4\rfloor}} \le \frac{8k}{|D|\sqrt{\lambda}} + \sqrt{\frac{7C_0\lambda^{-s}}{|D|}} \\
\le& (8+\sqrt{7C_0})\frac{\lambda^{-s/2}}{\sqrt{|D|}}\left(\frac{k}{\sqrt{|D|}}\lambda^{\frac{s-1}{2}}+1\right).
\end{aligned}$$

By (12),

$$\frac{k}{\sqrt{|D|}}\lambda^{\frac{s-1}{2}} \le \left\{\begin{array}{ll} \lambda^{\frac{s-1}{2}+\frac{1+s}{2}}, & \text{when } 0<r<\frac{1}{2} \\ \lambda^{\frac{1}{2}-r+\frac{s-1}{2}+r+\frac{s}{2}}, & \text{when } \frac{1}{2}\le r\le 1 \end{array}\right\} = \lambda^s \le 1. \tag{42}$$

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

$$\begin{aligned}
J_2 :=& (2\|f_\lambda\|_K + 8M + 8\sigma)\mathcal{A}_{D,\lambda,k}\log\frac{8}{\delta} \le C_1^1(1+\lambda^{r-\frac{1}{2}})\frac{\lambda^{-s/2}}{\sqrt{|D|}}\log\frac{8}{\delta} \\
=& C_1^1(1+\lambda^{r-\frac{1}{2}})\lambda^{-\frac{s}{2}+\frac{s}{2}+\max\{r,\frac{1}{2}\}}\log\frac{8}{\delta} \le 2C_1^1\lambda^r\log\frac{8}{\delta},
\end{aligned}$$

where $C_1^1 = (4\|h_\rho\|_\rho + 8M + 8\sigma)\times 2(8+\sqrt{7C_0})$. By definition and the assumption $|D_l|\ge 4$ (hence $\lfloor |D_l|/4\rfloor \ge |D_l|/7$),

$$\begin{aligned}
\mathcal{A}_{D_l,\lambda} =& \frac{1}{|D_l|\sqrt{\lambda}} + \frac{1}{\lfloor |D_l|/4\rfloor\sqrt{\lambda}} + \sqrt{\frac{\mathcal{N}(\lambda)}{\lfloor |D_l|/4\rfloor}} \le \frac{8k}{|D|\sqrt{\lambda}} + \sqrt{\frac{7C_0\lambda^{-s}k}{|D|}} \\
\le& (8+\sqrt{7C_0})\sqrt{\frac{k\lambda^{-s}}{|D|}}\left(\sqrt{\frac{k}{|D|}}\lambda^{\frac{s-1}{2}}+1\right) \le 2(8+\sqrt{7C_0})\sqrt{\frac{k\lambda^{-s}}{|D|}},
\end{aligned}$$

where the last step follows from (42). Since

$$\frac{k\lambda^{-s-1}}{|D|} \leq \left\{ \begin{array}{ll} 1\lambda^{-s-1}\lambda^{s+1} = 1, & \text{when } 0 < r < \frac{1}{2} \\ \lambda^{\frac{1}{2}-r}\lambda^{-1-s}\lambda^{2r+s} = \lambda^{r-\frac{1}{2}}, & \text{when } \frac{1}{2} \leq r \leq 1 \end{array} \right\} \leq 1, \tag{43}$$

we have $\mathcal{A}^2_{D_l,\lambda}/\lambda \leq 4(8+\sqrt{7C_0})^2 k\lambda^{-s-1}/|D| \leq 4(8+\sqrt{7C_0})^2$.

Now we bound the third term on the right-hand side of (11).

$$J_3 := 128(\|f_\lambda\|_K + M + \sigma)\lambda^{-\frac{1}{2}} \left(\log^4 \frac{16k}{\delta}\right) \max_{1\leq l\leq k} \left(\frac{\mathcal{A}^2_{D_l,\lambda}}{\lambda} + 1\right) \mathcal{A}^2_{D_l,\lambda}.$$

When $0 < r < \frac{1}{2}$, $k = 1$ and $\frac{1}{|D|} = \lambda^{1+s}$,

$$J_3 \leq C_1^2 \lambda^{r-1} \left(\log^4 \frac{16}{\delta}\right) \frac{k\lambda^{-s}}{|D|} = C_1^2 \lambda^r \log^4 \frac{16}{\delta}, \tag{44}$$

where $C_1^2 = 128(\|h_\rho\|_\rho + M + \sigma)[4(8+\sqrt{7C_0})^2 + 1] \times 4(8+\sqrt{7C_0})^2$. When $\frac{1}{2} \leq r \leq 1$, $k \leq \lambda^{\frac{1}{2}-r} \log^{-4}|D|$ and $\frac{1}{|D|} = \lambda^{2r+s}$. Recall that $|D| \geq 4$ which implies $\log \frac{16k}{\delta} \leq \log \frac{16|D|}{\delta} \leq 2(\log|D|)\log\frac{16}{\delta}$. So

$$J_3 \leq C_1^3 \lambda^{-\frac{1}{2}} \left(\log^4 |D|\right) \left(\log^4 \frac{16}{\delta}\right) \lambda^{\frac{1}{2}-r} \left(\log^{-4}|D|\right) \lambda^{-s}\lambda^{2r+s} = C_1^3 \lambda^r \log^4 \frac{16}{\delta}, \tag{45}$$

where $C_1^3 := 128(\|h_\rho\|_\rho + M + \sigma) \times 2^4[4(8+\sqrt{7C_0})^2 + 1] \times 4(8+\sqrt{7C_0})^2$.

The last term on the right-hand side of (11) is bounded as follows.

$$\begin{aligned} J_4 :=& 16c_{p,\sigma,M}(1+\lambda^{p+\frac{1}{2}})h^{-2p}\lambda^{-p-1} \left(\log^3 \frac{16k}{\delta}\right) \left(\log \frac{16|D|}{\delta}\right)^{2p+1} \\ &\times \left(1 + \lambda^{-\frac{1}{2}} \max_{1\leq l\leq k} \left(\frac{\mathcal{A}^2_{D_l,\lambda}}{\lambda} + 1\right) \mathcal{A}_{D_l,\lambda}\right) \\ \leq& C_1^4 h^{-2p}\lambda^{-p-1} (\log|D|)^{2p+4} \left(\log \frac{16}{\delta}\right)^{2p+4}, \end{aligned}$$

where $C_1^4 := 16c_{p,\sigma,M} \times 2^{2p+5} \times [1 + (4(8+\sqrt{7C_0})^2 + 1) \times 2(8+\sqrt{7C_0})]$. One completes the proof by letting $C_1 := \|h_\rho\|_\rho + 2C_1^1 + \max\{C_1^2, C_1^3\} + C_1^4$. ■

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 $\bar{f}_{D^*,\lambda}$ be defined in (16). One has*

$$\|\bar{f}_{D^*,\lambda} - f_\lambda\|_K \leq S_1^* + S_2^* + \mathcal{D}_{D^*,\lambda}\|f_\lambda\|_K + \mathcal{F}_{D^*,\lambda}, \tag{46}$$

where

$$S_1^* = \max_{1\leq l\leq k} \left(\mathcal{B}_{D_l^*,\lambda}\mathcal{C}^2_{D_l^*,\lambda}\|f_\lambda\|_K\lambda^{-\frac{1}{2}} + \mathcal{B}_{D_l^*,\lambda}\mathcal{C}_{D_l^*,\lambda}\mathcal{G}_{D_l^*,\lambda}\lambda^{-\frac{1}{2}}\right),$$

and

$$S_2^* = \max_{1\leq l\leq k} \left(\mathcal{B}_{D_l^*,\lambda}\mathcal{C}_{D_l^*,\lambda}\lambda^{-1} + \lambda^{-\frac{1}{2}}\right) \|E_{D_l^*,\lambda}\|_K.$$

■

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* $1-\delta$.

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

where $\mathcal{A}_{D^*,\lambda,k} = \frac{k}{|D^*|\sqrt{\lambda}} + \frac{1}{\lfloor |D^*|/4\rfloor\sqrt{\lambda}} + \sqrt{\frac{\mathcal{N}(\lambda)}{\lfloor |D^*|/4\rfloor}}$ *and* $\mathcal{A}_{D^*,\lambda} = \mathcal{A}_{D^*,\lambda,1}$, *which are consistent with the notations* $\mathcal{A}_{D,\lambda,k}$ *and* $\mathcal{A}_{D,\lambda}$ *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* $(\mathcal{W}\times\mathcal{Z}, \rho_{\mathcal{W}}\times\rho_{\mathcal{Z}})$. *Assume that there exist two constants* $\sigma>0$ *and* $M>0$, *such that for any integer* $q\geq 2$,

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

Suppose that from $(\mathcal{W},\rho_{\mathcal{W}})$, *one draws independently a sample* $D=\{w_1,\ldots,w_{|D|}\}$, *which is evenly divided to* k *disjoint subsets* $D=\cup_{l=1}^k D_l$ *with* $|D_1|=\cdots=|D_k|=|D|/k$. *Suppose that similarly, one divides another sample* $\tilde{D}=\cup_{l=1}^k \tilde{D}_l$ *independently drawn from* $(\mathcal{Z},\rho_{\mathcal{Z}})$ *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*

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

Proof. Let π be a permutation of the set $\{1,\ldots,|D_l|\}$ of integers, so $\{w^l_{\pi(1)},\ldots,w^l_{\pi(|D_l|)}\}$ is the associated permutation of $D_l=\{w_i^l\}_{i=1}^{|D_l|}$. Let ψ be a permutation of $\{1,\ldots,|\tilde{D}_l|\}$, so $\{z^l_{\psi(1)},\ldots,z^l_{\psi(|\tilde{D}_l|)}\}$ is the associated permutation of $\tilde{D}_l=\{z_i^l\}_{i=1}^{|\tilde{D}_l|}$. For any $1\leq l\leq k$, we write $U^l_{\pi,\psi}=\frac{1}{|D_l|}\sum_{i=1}^{|D_l|}\xi(w^l_{\pi(i)},z^l_{\psi(i)})$ to obtain

$$\frac{1}{|D_l||\tilde{D}_l|}\sum_{w\in D_l}\sum_{z\in\tilde{D}_l}\xi(w,z)=\frac{1}{|D_l|!|\tilde{D}_l|!}\sum_{\pi}\sum_{\psi}U^l_{\pi,\psi}, \tag{47}$$

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

$$\begin{aligned}\frac{1}{k}\sum_{l=1}^k\frac{1}{|D_l||\tilde{D}_l|}\sum_{w\in D_l}\sum_{z\in\tilde{D}_l}\xi(w,z)=&\frac{1}{k}\sum_{l=1}^k\frac{1}{|D_l|!|\tilde{D}_l|!}\sum_{\pi,\psi}U^l_{\pi,\psi}\\ =&\frac{1}{|D_1|!|\tilde{D}_1|!}\sum_{\pi,\psi}\frac{1}{k}\sum_{l=1}^k U^l_{\pi,\psi},\end{aligned}$$

where

$$\frac{1}{k}\sum_{l=1}^{k}U^l_{\pi,\psi}=\frac{1}{k}\sum_{l=1}^{k}\frac{1}{|D_l|}\sum_{i=1}^{|D_l|}\xi(w^l_{\pi(i)},z^l_{\psi(i)})=\frac{1}{|D|}\sum_{l=1}^{k}\sum_{i=1}^{|D_l|}\xi(w^l_{\pi(i)},z^l_{\psi(i)}).$$

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

$$\begin{aligned}
&\text{Prob}\left\{\left\|\frac{1}{k}\sum_{l=1}^{k}\frac{1}{|D_l||\tilde{D}_l|}\sum_{w\in D_l}\sum_{z\in\tilde{D}_l}\xi(w,z)-\mathbb{E}\xi\right\|\geq\epsilon\right\}\\
\leq&\mathbb{E}\cosh\left(c\left\|\frac{1}{k}\sum_{l=1}^{k}\frac{1}{|D_l|!|\tilde{D}_l|!}\sum_{\pi}\sum_{\psi}U^l_{\pi,\psi}-\mathbb{E}\xi\right\|\right)/\cosh(c\epsilon)\\
=&\mathbb{E}\cosh\left(c\left\|\frac{1}{|D_1|!|\tilde{D}_1|!}\sum_{\pi,\psi}\left[\frac{1}{k}\sum_{l=1}^{k}U^l_{\pi,\psi}-\mathbb{E}\xi\right]\right\|\right)/\cosh(c\epsilon)\\
\leq&\frac{1}{|D_1|!|\tilde{D}_1|!}\sum_{\pi,\psi}\mathbb{E}\cosh\left(c\left\|\frac{1}{|D|}\sum_{l=1}^{k}\sum_{i=1}^{|D_l|}\xi(w^l_{\pi(i)},z^l_{\psi(i)})-\mathbb{E}\xi\right\|\right)/\cosh(c\epsilon).
\end{aligned}$$

One employs Lemma A.2 in Appendix to obtain

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

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* $1-\delta$.

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

Proof. Recall that K is antisymmetric. So

$$\begin{aligned}
L_Kf_\rho&=\int_{\mathcal{Z}}\int_{\mathcal{Z}}(y-v)K_{(x,u)}d\rho(x,y)d\rho(u,v)\\
&=\int_{\mathcal{Z}}\int_{\mathcal{Z}}yK_{(x,u)}d\rho(x,y)d\rho(u,v)-\int_{\mathcal{Z}}\int_{\mathcal{Z}}vK_{(x,u)}d\rho(x,y)d\rho(u,v)\\
&=\int_{\mathcal{Z}}\int_{\mathcal{Z}}2yK_{(x,u)}d\rho(x,y)d\rho(u,v).
\end{aligned}$$

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

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

where $\hat{f}_{\rho,D}$ is defined in Lemma 4.1 and $\hat{f}_{\rho,D,\tilde{D}} = \frac{1}{|D||\tilde{D}|}\sum_{w\in D, z\in\tilde{D}} 2yK_{(x,u)}$. Below, $\hat{f}_{\rho,D_l}$ and $\hat{f}_{\rho,D_l,\tilde{D}_l}$ are similarly defined by substituting D and $\tilde{D}$ with D_l and $\tilde{D}_l$, respectively. Note that both $\hat{f}_{\rho,D}$ and $\hat{f}_{\rho,D,\tilde{D}}$ are empirical analogs of $L_K f_\rho$. We have

$$\begin{aligned}\mathcal{F}_{D^*,\lambda} \leq & \frac{|D|}{|D^*|}\left\|\frac{1}{k}\sum_{l=1}^{k}(L_K+\lambda I)^{-\frac{1}{2}}(\hat{f}_{\rho,D_l}-L_K f_\rho)\right\|_K \\ & + \frac{|\tilde{D}|}{|D^*|}\left\|\frac{1}{k}\sum_{l=1}^{k}(L_K+\lambda I)^{-\frac{1}{2}}(\hat{f}_{\rho,D_l,\tilde{D}_l}-L_K f_\rho)\right\|_K.\end{aligned}$$

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

$$\mathcal{F}_{D,\lambda} = \left\|\frac{1}{k}\sum_{l=1}^{k}(L_K+\lambda I)^{-\frac{1}{2}}(\hat{f}_{\rho,D_l}-L_K f_\rho)\right\|_K \leq 8(M+\sigma)\mathcal{A}_{D,\lambda,k}\log\frac{4}{\delta}.$$

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

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

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

$$\begin{aligned}\mathbb{E}[\|\xi-\mathbb{E}\xi\|_K^q] \leq & 2^{q-1}\mathbb{E}[\|\xi\|_K^q] + 2^{q-1}\|\mathbb{E}\xi\|_K^q \leq 2^q\mathbb{E}[\|\xi\|_K^q] \\ \leq & 2^{2q}\sup_{x',u'\in\mathcal{X}}\|(L_K+\lambda I)^{-1/2}K_{(x',u')}\|_K^{q-2}\mathbb{E}\left[|Y|^q\|(L_K+\lambda I)^{-1/2}K_{(x,u)}\|_K^2\right] \\ \leq & 2^{2q}\left(\frac{1}{\sqrt{\lambda}}\right)^{q-2}\mathbb{E}\left[\mathbb{E}(|Y|^q|X)\|(L_K+\lambda I)^{-1/2}K_{(x,u)}\|_K^2\right] \\ \leq & 2^{2q}\left(\frac{1}{\sqrt{\lambda}}\right)^{q-2}\frac{1}{2}q!\sigma^2M^{q-2}\mathcal{N}(\lambda) = \frac{1}{2}q![16\sigma^2\mathcal{N}(\lambda)](4M/\sqrt{\lambda})^{q-2}.\end{aligned}$$

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

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

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

$$\begin{aligned}& \frac{|D|}{|D^*|}\mathcal{A}_{D,\lambda,k} + \frac{|\tilde{D}|}{|D^*|}\left(\frac{1}{|D|\sqrt{\lambda}}+\sqrt{\frac{\mathcal{N}(\lambda)}{|D|}}\right) \\ \leq & \frac{|D|}{|D^*|}\frac{k}{|D|\sqrt{\lambda}} + \left(\frac{1}{\lfloor|D|/4\rfloor\sqrt{\lambda}}+\sqrt{\frac{\mathcal{N}(\lambda)}{\lfloor|D|/4\rfloor}}\right)\left(\frac{|D|}{|D^*|}+\frac{|\tilde{D}|}{|D^*|}\right) \\ = & \frac{k}{|D^*|\sqrt{\lambda}} + \frac{1}{\lfloor|D|/4\rfloor\sqrt{\lambda}} + \sqrt{\frac{\mathcal{N}(\lambda)}{\lfloor|D|/4\rfloor}} = \mathcal{A}_{D,D^*,\lambda,k}.\end{aligned}$$

One completes the proof by observing that $\mathcal{G}_{D^*,\lambda}$ is a special case of $\mathcal{F}_{D^*,\lambda}$ with $k=1$. ■

5.2. Proof of learning rates in semi-supervised learning

Proof of Theorem 2.6. We can decompose $\|\bar{f}_{D^*,\lambda} - f_\rho\|_\rho$ as the *sample error* $\|\bar{f}_{D^*,\lambda} - f_\lambda\|_\rho$ and the *approximation error* $\|f_\lambda - f_\rho\|_\rho$. As stated in (29), $\|f_\lambda - f_\rho\|_\rho \le \lambda^r \|h_\rho\|_\rho$ for $0 < r \le 1$. 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. One has

$$\begin{aligned}
\lambda\|f_{D^*,\lambda}\|_K^2 &\le \mathcal{E}_{D^*}(0) - \mathcal{E}_{D^*}(f_{D^*,\lambda}) \le -\frac{h^2}{|D^*|^2}\sum_{w,z\in D^*} G\left(\frac{(y-v)^2}{h^2}\right) + h^2 G(0) \\
&\le \frac{C_G}{|D^*|^2}\sum_{w,z\in D^*}(y-v)^2 = \frac{C_G}{|D^*|^2}\left[\sum_{w,z\in D}\left(\frac{|D^*|}{|D|}y - \frac{|D^*|}{|D|}v\right)^2 + 2\sum_{w\in D, z\in\tilde{D}}\left(\frac{|D^*|}{|D|}y\right)^2\right] \\
&\le 4C_G\left(\frac{1}{|D|}\sum_{w\in D} y^2 + \frac{|\tilde{D}|}{|D|}\cdot\frac{1}{|D|}\sum_{w\in D} y^2\right) \\
&\le 4C_G\left(1 + \frac{|\tilde{D}|}{|D|}\right)\max_{w\in D} y^2 = 4C_G\frac{|D^*|}{|D|}\max_{w\in D} y^2.
\end{aligned}$$

Thus, $\|f_{D^*,\lambda}\|_K \le 2\left(\frac{C_G|D^*|}{\lambda|D|}\right)^{1/2}\max_{w\in D}|y|$. Similar to the estimation (26),

$$\begin{aligned}
&\|E_{D^*,\lambda}\|_K \\
\le& c_p h^{-2p}\frac{1}{|D^*|^2}\sum_{w,z\in D^*}\left(\|f_{D^*,\lambda}\|_K + |y-v|\right)^{2p+1} \\
\le& 2^{2p}c_p h^{-2p}\left(\|f_{D^*,\lambda}\|_K^{2p+1} + 2^{2p+1}\max_{w\in D^*}|y|^{2p+1}\right) \\
\le& 2^{2p}c_p h^{-2p}\left(2^{2p+1}\left(\frac{C_G|D^*|}{\lambda|D|}\right)^{p+\frac{1}{2}}\max_{w\in D}|y|^{2p+1} + 2^{2p+1}\left(\frac{|D^*|}{|D|}\right)^{2p+1}\max_{w\in D}|y|^{2p+1}\right) \\
\le& 2^{4p+1}c_p h^{-2p}(C_G^{p+\frac{1}{2}}+1)\Delta_{D,D^*,\lambda}(4M+5\sigma)^{2p+1}\log^{2p+1}\frac{|D|}{\delta},
\end{aligned}$$

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

$$\|E_{D^*,\lambda}\|_K \le C_2\Delta_{D,D^*,\lambda}h^{-2p}\log^{2p+1}\frac{|D|}{\delta},$$

where $C_2 = 2^{4p+1}c_p(C_G^{p+\frac{1}{2}}+1)(4M+5\sigma)^{2p+1}$. 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.

$$\begin{aligned}
\mathcal{B}_{D_l^*,\lambda}\mathcal{C}_{D_l^*,\lambda}^2\lambda^{-1/2} \le& 8\left(\frac{\mathcal{A}_{D_l^*,\lambda}^2}{\lambda}+1\right)\left(\log^2\frac{2}{\delta}\right)\times 4\mathcal{A}_{D_l^*,\lambda}^2\left(\log^2\frac{2}{\delta}\right)\lambda^{-1/2} \\
=& 32\left(\frac{\mathcal{A}_{D_l^*,\lambda}^2}{\lambda}+1\right)\mathcal{A}_{D_l^*,\lambda}^2\lambda^{-1/2}\log^4\frac{2}{\delta},
\end{aligned}$$

$$
\begin{aligned}
&\mathcal{B}_{D_l^*,\lambda}\mathcal{C}_{D_l^*,\lambda}\lambda^{-1}\|E_{D_l^*,\lambda}\|_K \\
\leq& 8\left(\frac{\mathcal{A}^2_{D_l^*,\lambda}}{\lambda}+1\right)\left(\log^2\frac{2}{\delta}\right)\times 2\mathcal{A}_{D_l^*,\lambda}\left(\log\frac{2}{\delta}\right)\lambda^{-1}C_2h^{-2p}\Delta_{D_l,D_l^*,\lambda}\log^{2p+1}\frac{|D_l|}{\delta} \\
=& 16C_2\Delta_{D_l,D_l^*,\lambda}h^{-2p}\lambda^{-1}\left(\frac{\mathcal{A}^2_{D_l^*,\lambda}}{\lambda}+1\right)\mathcal{A}_{D_l^*,\lambda}\left(\log^3\frac{2}{\delta}\right)\log^{2p+1}\frac{|D_l|}{\delta},
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathcal{B}_{D_l^*,\lambda}\mathcal{C}_{D_l^*,\lambda}\mathcal{G}_{D_l^*,\lambda}\lambda^{-1/2} \\
\leq& 8\left(\frac{\mathcal{A}^2_{D_l^*,\lambda}}{\lambda}+1\right)\left(\log^2\frac{2}{\delta}\right)\times 2\mathcal{A}_{D_l^*,\lambda}\left(\log\frac{2}{\delta}\right)\times 8(M+\sigma)\mathcal{A}_{D_l,D_l^*,\lambda}\left(\log\frac{4}{\delta}\right)\lambda^{-1/2} \\
\leq& 256(M+\sigma)\left(\frac{\mathcal{A}^2_{D_l^*,\lambda}}{\lambda}+1\right)\mathcal{A}_{D_l^*,\lambda}\mathcal{A}_{D_l,D_l^*,\lambda}\lambda^{-1/2}\log^4\frac{2}{\delta},
\end{aligned}
$$

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

$$
\begin{aligned}
S_1^* \leq& 256(1+M+\sigma)\lambda^{-1/2}\left(\log^4\frac{2}{\delta}\right) \\
&\times\max_{1\leq l\leq k}\left(\frac{\mathcal{A}^2_{D_l^*,\lambda}}{\lambda}+1\right)\mathcal{A}_{D_l^*,\lambda}\left(\mathcal{A}_{D_l^*,\lambda}\|f_\lambda\|_K+\mathcal{A}_{D_l,D_l^*,\lambda}\right), \qquad (48)\\
S_2^* \leq& 16\lambda^{-1/2}C_2\Delta_{D_l,D_l^*,\lambda}h^{-2p}\left(\log^3\frac{2}{\delta}\right) \\
&\times\max_{1\leq l\leq k}\left[\left(\frac{\mathcal{A}^2_{D_l^*,\lambda}}{\lambda}+1\right)\mathcal{A}_{D_l^*,\lambda}\lambda^{-1/2}+1\right]\log^{2p+1}\frac{|D_l|}{\delta}. \qquad (49)
\end{aligned}
$$

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.

$$
\begin{aligned}
\mathcal{D}_{D^*,\lambda} \leq& 2\mathcal{A}_{D^*,\lambda,k}\log\frac{8}{\delta}, \qquad (50)\\
\mathcal{F}_{D^*,\lambda} \leq& 8(M+\sigma)\mathcal{A}_{D,D^*,\lambda,k}\log\frac{16}{\delta}. \qquad (51)
\end{aligned}
$$

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

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

$$
\frac{\|f_\lambda\|_K}{\sqrt{|D^*|}}\leq\left\{\begin{array}{ll}\|h_\rho\|_\rho\lambda^{r-\frac{1}{2}+\frac{s}{2}+\frac{1}{2}}, & \text{when } 0<r<\frac{1}{2}\\ \|h_\rho\|_\rho/\sqrt{|D|}, & \text{when } \frac{1}{2}\leq r\leq 1\end{array}\right\}=\|h_\rho\|_\rho\lambda^{r+\frac{s}{2}}. \qquad (52)
$$

For any $1\leq l\leq k$, the assumption $|D_l^*|\geq|D_l|\geq 4$ implies $\lfloor|D_l^*|/4\rfloor\geq|D_l^*|/7$. The assumption $k\leq\sqrt{|D^*|\lambda^{1+s}}$ implies $\frac{k}{\sqrt{|D^*|}}\lambda^{-(1+s)/2}\leq 1$. Recall $\mathcal{N}(\lambda)\leq C_0\lambda^{-s}$. We have

$$
\begin{aligned}
\mathcal{A}_{D_l^*,\lambda} =& \frac{1}{|D_l^*|\sqrt{\lambda}}+\frac{1}{\lfloor|D_l^*|/4\rfloor\sqrt{\lambda}}+\sqrt{\frac{\mathcal{N}(\lambda)}{\lfloor|D_l^*|/4\rfloor}}\leq\frac{8k}{|D^*|\sqrt{\lambda}}+\sqrt{\frac{7C_0\lambda^{-s}k}{|D^*|}} \\
\leq& (8+\sqrt{7C_0})\sqrt{\frac{k}{|D^*|\lambda^s}}\left(\sqrt{\frac{k}{|D^*|}}\lambda^{\frac{s}{2}-\frac{1}{2}}+1\right)\leq 2(8+\sqrt{7C_0})\sqrt{\frac{k}{|D^*|\lambda^s}}.
\end{aligned}
$$

So,

$$\frac{\mathcal{A}^2_{D_l^*,\lambda}}{\lambda}+1\leq 4(8+\sqrt{7C_0})^2\frac{k}{|D^*|}\lambda^{-s-1}+1\leq C_3^1,$$

where $C_3^1:=4(8+\sqrt{7C_0})^2+1$. From definition,

$$\begin{aligned}\mathcal{A}_{D_l,D_l^*,\lambda}=&\frac{1}{|D_l^*|\sqrt{\lambda}}+\frac{1}{\lfloor|D_l|/4\rfloor\sqrt{\lambda}}+\sqrt{\frac{\mathcal{N}(\lambda)}{\lfloor|D_l|/4\rfloor}}\\ \leq&\frac{8k}{|D|\sqrt{\lambda}}+\sqrt{\frac{7C_0\lambda^{-s}k}{|D|}}=8k\lambda^{2r+s-\frac{1}{2}}+\sqrt{7C_0k}\lambda^r.\end{aligned}$$

So, for any $1\leq l\leq k$,

$$\begin{aligned}\mathcal{A}_{D_l^*,\lambda}\|f_\lambda\|_K+\mathcal{A}_{D_l,D_l^*,\lambda}\leq&2(8+\sqrt{7C_0})\sqrt{k\lambda^{-s}}\frac{\|f_\lambda\|_K}{\sqrt{|D^*|}}+\mathcal{A}_{D_l,D_l^*,\lambda}\\ \leq&2(8+\sqrt{7C_0})\|h_\rho\|_\rho\sqrt{k}\lambda^r+8k\lambda^{2r+s-\frac{1}{2}}+\sqrt{7C_0k}\lambda^r\\ \leq&C_3^2\left(\sqrt{k}\lambda^r+k\lambda^{2r+s-\frac{1}{2}}\right),\end{aligned}$$

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).

$$\begin{aligned}J_2^*:=&256(1+M+\sigma)\lambda^{-1/2}\left(\log^4\frac{16k}{\delta}\right)\\ &\times\max_{1\leq l\leq k}\left(\frac{\mathcal{A}^2_{D_l^*,\lambda}}{\lambda}+1\right)\mathcal{A}_{D_l^*,\lambda}\left(\mathcal{A}_{D_l^*,\lambda}\|f_\lambda\|_K+\mathcal{A}_{D_l,D_l^*,\lambda}\right)\\ =&\frac{C_3^3}{\sqrt{\lambda}}\left(\log^4|D|\right)\left(\log^4\frac{16}{\delta}\right)\sqrt{\frac{k}{|D^*|\lambda^s}}\left(\sqrt{k}\lambda^r+k\lambda^{2r+s-\frac{1}{2}}\right),\end{aligned}$$

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

$$\frac{k\lambda^r\log^4|D|}{\sqrt{|D^*|\lambda^{s+1}}}\leq\lambda^r,$$

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

So $J_2^*\leq 2C_3^3\lambda^r\log^4\frac{16}{\delta}$.

Next, it is easy to see that for $1 \le l \le k$, $\Delta_{D_l,D_l^*,\lambda} = \Delta_{D,D^*,\lambda}$. Note that $a \le a^2 + 1$ for $a \ge 0$. We have

$$\begin{aligned}
J_3^* :=& 16\lambda^{-1/2} C_2 h^{-2p} \left(\log^3 \frac{16k}{\delta}\right) \\
&\times \max_{1\le l\le k} \left[\left(\frac{\mathcal{A}^2_{D_l^*,\lambda}}{\lambda} + 1\right) \frac{\mathcal{A}_{D_l^*,\lambda}}{\sqrt{\lambda}} + 1\right] \Delta_{D_l,D_l^*,\lambda} \log^{2p+1} \frac{16|D|}{\delta} \\
\le& 16 C_2 h^{-2p} \times 2^3 \left(\log^3 |D|\right) \left(\log^3 \frac{16}{\delta}\right) \lambda^{-1/2} \left((C_3^1)^2 + 1\right) \\
&\times \Delta_{D,D^*,\lambda} \times 2^{2p+1} \left(\log |D|\right)^{2p+1} \left(\log \frac{16}{\delta}\right)^{2p+1} \\
=& C_3^4 h^{-2p} \lambda^{-1/2} \Delta_{D,D^*,\lambda} \left(\log |D|\right)^{2p+4} \left(\log \frac{16}{\delta}\right)^{2p+4},
\end{aligned}$$

where $C_3^4 := 16 C_2 \times 2^{2p+4} \left((C_3^1)^2 + 1\right)$.

Now we bound the last term on the right-hand side of (17). By definition and the bound (52) above,

$$\begin{aligned}
\mathcal{A}_{D^*,\lambda,k} \|f_\lambda\|_K =& \|f_\lambda\|_K \left(\frac{k}{|D^*|\sqrt{\lambda}} + \frac{1}{\lfloor |D^*|/4 \rfloor \sqrt{\lambda}} + \sqrt{\frac{\mathcal{N}(\lambda)}{\lfloor |D^*|/4 \rfloor}}\right) \\
\le& \|h_\rho\|_\rho \lambda^{r+\frac{s}{2}} \left(\frac{8k}{\sqrt{\lambda |D^*|}} + \sqrt{7C_0 \lambda^{-s}}\right).
\end{aligned}$$

The assumption (18) of k implies $\frac{k}{\sqrt{\lambda|D^*|}} \le \lambda^{s/2} \le 1$. So $\mathcal{A}_{D^*,\lambda,k} \|f_\lambda\|_K \le \|h_\rho\|_\rho (8 + \sqrt{7C_0}) \lambda^r$. By the assumption $|D^*| \ge \lambda^{-1-s}$, $k \le \sqrt{|D^*| \lambda^{1+s}}$, and $r + s \ge \frac{1}{2}$,

$$\begin{aligned}
\mathcal{A}_{D,D^*,\lambda,k} =& \frac{k}{|D^*|\sqrt{\lambda}} + \frac{1}{\lfloor |D|/4 \rfloor \sqrt{\lambda}} + \sqrt{\frac{\mathcal{N}(\lambda)}{\lfloor |D|/4 \rfloor}} \\
\le& \sqrt{\frac{\lambda^s}{|D|}} + 7\lambda^{2r+s-\frac{1}{2}} + \sqrt{7C_0} \lambda^{-\frac{s}{2}+r+\frac{s}{2}} \le (8 + \sqrt{7C_0}) \lambda^r.
\end{aligned}$$

We summarize the above analysis and use (17) to obtain that with probability $1 - \delta$,

$$\begin{aligned}
\|\bar{f}_{D^*,\lambda} - f_\rho\|_\rho \le& \|h_\rho\|_\rho \lambda^r + 2C_3^3 \lambda^r \log^4 \frac{16}{\delta} \\
&+ C_3^4 h^{-2p} \lambda^{-1/2} \Delta_{D,D^*,\lambda} \left(\log |D|\right)^{2p+4} \left(\log \frac{16}{\delta}\right)^{2p+4} \\
&+ (2\|h_\rho\|_\rho + 8M + 8\sigma)(8 + \sqrt{7C_0}) \lambda^r \log \frac{16}{\delta},
\end{aligned}$$

which yields the conclusion with $C_3 := \|h_\rho\|_\rho + 2C_3^3 + C_3^4 + (2\|h_\rho\|_\rho + 8M + 8\sigma)(8 + \sqrt{7C_0})$. ■

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

Special Administrative Region, China (Project No. PolyU 25301115). We thank the anonymous reviewers for their constructive comments. All the three authors contributed equally to the paper. The corresponding author is Ting Hu.

Appendix A. Concentration inequalities

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

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

■

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

$$\mathbb{E}[\|\xi(z) - \mathbb{E}\xi\|^q] \leq \frac{1}{2} q! \sigma^2 M^{q-2}, \tag{53}$$

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

$$\inf_{c>0} \mathbb{E}\left[\cosh\left(c\left\|\frac{1}{n}\sum_{i=1}^n \xi(z_i) - \mathbb{E}(\xi)\right\|\right)\right] / \cosh(c\epsilon) \leq 2\exp\left\{-\frac{n\epsilon^2}{2(\sigma^2 + M\epsilon)}\right\}. \tag{54}$$

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

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

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

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

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

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

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 $c = \frac{n\epsilon}{\sigma^2 + M\epsilon} \in (0, n/M)$ and noting that $\cosh(c\epsilon) \geq \frac{e^{c\epsilon}}{2}$, the conclusion is obtained. ■

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

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

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

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

Therefore

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

The proof is complete. ■

References

Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *J. Complexity*, 23(1):52–72, 2007.

Mikhail Belkin and Partha Niyogi. Semi-supervised learning on Riemannian manifolds. *Machine Learning*, 56(1):209–239, 2004.

Rajendra Bhatia. *Matrix analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1997.

Gilles Blanchard and Nicole Krämer. Optimal learning rates for kernel conjugate gradient regression. In *Advances in Neural Information Processing Systems 23*, pages 226–234. Curran Associates, Inc., 2010.

Gilles Blanchard and Nicole Krämer. Convergence rates of kernel conjugate gradient for random design regression. *Anal. Appl. (Singap.)*, 14(6):763–794, 2016.

Gilles Blanchard and Nicole Mücke. Optimal rates for regularization of statistical inverse learning problems. *Found. Comput. Math.*, 18(4):971–1013, 2018.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory (Madison, WI, 1998)*, pages 92–100. ACM, New York, 1998.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.*, 7(3):331–368, 2007.

Andrea Caponnetto and Yuan Yao. Cross-validation based adaptation for regularization operators in learning theory. *Anal. Appl. (Singap.)*, 8(2):161–183, 2010.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. MIT Press, 2006.

Badong Chen, Yu Zhu, and Jinchun Hu. Mean-square convergence analysis of ADALINE training with minimum error entropy criterion. *IEEE Transactions on Neural Networks*, 21(7):1168–1179, 2010.

Andreas Christmann and Ding-Xuan Zhou. On the robustness of regularized pairwise learning methods based on kernels. *J. Complexity*, 37:1–33, 2016.

Felipe Cucker and Ding-Xuan Zhou. *Learning theory: an approximation theory viewpoint*, volume 24 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 2007. With a foreword by Stephen Smale.

Deniz Erdogmus and Jose C. Principe. Comparison of entropy and mean square error criteria in adaptive system training using higher order statistics. In *Proceedings of the Second International Workshop on Independent Component Analysis and Blind Signal*, pages 75–90. Berlin: Springer-Verlag, 2000.

Deniz Erdogmus and Jose C. Principe. Convergence properties and data efficiency of the minimum error entropy criterion in ADALINE training. *IEEE Transactions on Signal Processing*, 51(7): 1966–1978, 2003.

Jun Fan, Ting Hu, Qiang Wu, and Ding-Xuan Zhou. Consistency analysis of an empirical minimum error entropy algorithm. *Appl. Comput. Harmon. Anal.*, 41(1):164–189, 2016.

Erhan Gokcay and Jose C. Principe. Information theoretic clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):158–171, 2002.

Zheng-Chu Guo, Shao-Bo Lin, and Ding-Xuan Zhou. Learning theory of distributed spectral algorithms. *Inverse Problems*, 33(7):074009, 29, 2017a.

Zheng-Chu Guo, Lei Shi, and Qiang Wu. Learning theory of distributed regression with bias corrected regularization kernel network. *J. Mach. Learn. Res.*, 18:Paper No. 118, 25, 2017b.

Ting Hu, Jun Fan, Qiang Wu, and Ding-Xuan Zhou. Learning theory approach to minimum error entropy criterion. *J. Mach. Learn. Res.*, 14:377–397, 2013.

Ting Hu, Jun Fan, Qiang Wu, and Ding-Xuan Zhou. Regularization schemes for minimum error entropy principle. *Anal. Appl. (Singap.)*, 13(4):437–455, 2015.

Ting Hu, Qiang Wu, and Ding-Xuan Zhou. Convergence of gradient descent for minimum error entropy principle in linear regression. *IEEE Trans. Signal Process.*, 64(24):6571–6579, 2016.

Ting Hu, Qiang Wu, and Ding-Xuan Zhou. Distributed kernel gradient descent algorithm for minimum error entropy principle. *Appl. Comput. Harmon. Anal.*, 49(1):229–256, 2020.

Shao-Bo Lin and Ding-Xuan Zhou. Distributed kernel-based gradient descent algorithms. *Constr. Approx.*, 47(2):249–276, 2018.

Shao-Bo Lin, Xin Guo, and Ding-Xuan Zhou. Distributed learning with regularized least squares. *J. Mach. Learn. Res.*, 18:Paper No. 92, 31, 2017.

L. Lo Gerfo, L. Rosasco, F. Odone, E. De Vito, and A. Verri. Spectral algorithms for supervised learning. *Neural Comput.*, 20(7):1873–1897, 2008.

Shahar Mendelson and Joseph Neeman. Regularization in kernel learning. *Ann. Statist.*, 38(1): 526–565, 2010.

Nicole Mücke and Gilles Blanchard. Parallelizing spectrally regularized kernel algorithms. *J. Mach. Learn. Res.*, 19:Paper No. 30, 29, 2018.

Emanuel Parzen. On estimation of a probability density function and mode. *Ann. Math. Statist.*, 33:1065–1076, 1962.

I. F. Pinelis and A. I. Sakhanenko. Remarks on inequalities for large deviation probabilities. *Theory of Probability & Its Applications*, 30(1):143–148, 1986.

Jose C. Principe. *Information theoretic learning*. Information Science and Statistics. Springer, New York, 2010. Renyi's entropy and kernel perspectives.

Pengcheng Shen and Chunguang Li. Minimum total error entropy method for parameter estimation. *IEEE Trans. Signal Process.*, 63(15):4079–4090, 2015.

Luís M. Silva, J. Marques de Sá, and Luís A. Alexandre. The MEE principle in data classification: a perceptron-based analysis. *Neural Comput.*, 22(10):2698–2728, 2010.

Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constr. Approx.*, 26(2):153–172, 2007.

I. Steinwart, D. R. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21*, 2009.

Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

Cheng Wang and Ting Hu. Online minimum error entropy algorithm with unbounded sampling. *Anal. Appl. (Singap.)*, 17(2):293–322, 2019.

Jun Wang, Tony Jebara, and Shih-Fu Chang. Semi-supervised learning using greedy max-cut. *J. Mach. Learn. Res.*, 14:771–800, 2013.

Y. Wang, R. Khardon, D. Pechyony, and R. Jones. Generalization bounds for online learning algorithms with pairwise loss functions. In *COLT 2012 - The 25th Annual Conference on Learning Theory, June 25-27, 2012, Edinburgh, Scotland*, pages 13.1–13.22, 2012.

Yiming Ying and Ding-Xuan Zhou. Online pairwise learning algorithms. *Neural Comput.*, 28(4): 743–777, 2016.

Yiming Ying and Ding-Xuan Zhou. Unregularized online learning algorithms with general loss functions. *Appl. Comput. Harmon. Anal.*, 42(2):224–244, 2017.

T. Zhang. Effective dimension and generalization of kernel learning. In *Advances in Neural Information Processing Systems*, pages 454–461, 2002.

Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates. *J. Mach. Learn. Res.*, 16:3299–3340, 2015.

Ding-Xuan Zhou. The covering number in learning theory. *J. Complexity*, 18(3):739–767, 2002.

M. Zinkevich, M. Weimer, L. Li, and A. J. Smola. Parallelized stochastic gradient descent. In *Advances in Neural Information Processing Systems 23*, pages 2595–2603. Curran Associates, Inc., 2010.