Distributed Learning with Random Features

Jian Li[1,2] **Yong Liu**[1*] **Weiping Wang**[1]

[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
{lijian9026, liuyong, wangweiping}@iie.ac.cn

Abstract

Distributed learning and random projections are the most common techniques in large scale nonparametric statistical learning. In this paper, we study the generalization properties of kernel ridge regression using both distributed methods and random features. Theoretical analysis shows the combination remarkably reduces computational cost while preserving the optimal generalization accuracy under standard assumptions. In a benign case, $\mathcal{O}(\sqrt{N})$ partitions and $\mathcal{O}(\sqrt{N})$ random features are sufficient to achieve $\mathcal{O}(1/N)$ learning rate, where N is the labeled sample size. Further, we derive more refined results by using additional unlabeled data to enlarge the number of partitions and by generating features in a data-dependent way to reduce the number of random features.

1 Introduction

A fundamental problem in machine learning is to reach a good tradeoff between statistical properties and computational cost [1]. While this challenge is more severe in kernel methods, despite excellent theoretical guarantee, kernel methods do not scale well in large scale settings because of high time and memory requirements, typically at least quadratic in the number of examples. To overcome the scalability issue, a variety of practical algorithms have been developed : distributed learning, which produces a global model after training disjoint subset on individual machines with necessary communications [2], random projections including Nyström [3] and random features [4] to overcome memory bottleneck and gradient methods, as well as stochastic and preconditioned extensions [5, 6], to improve computational efficiency.

From the theoretical perspective, many works studied the statistical learning of those large scale approaches together with kernel ridge regression (KRR) [7–9], achieving optimal learning rates by using integral operator techniques [10] and using the effective dimension to control the capability of the hypothesis space [11]. Recent statistical learning works demonstrate that KRR together with large scale approaches not only obtain great computational gains but also achieve optimal theoretical properties, such as KRR together with divide-and-conquer [2, 12], with random projections including random features [13] and Nyström [9] and with stochastic gradient descent (SGD) [8, 14]. Recently, combinations of those accelerated algorithms benefit a lot and attract much attention, of which learning properties have been explored including the combination of divide-and-conquer and multi-pass SGD [15] and the combination of random features and multi-pass SGD [16].

In this paper, we investigate the approach of combining divide-and-conquer and random features to deal with extremely large-scale applications, but still, our approach preserves the same optimal statistical properties. We begin with a general learning error bound by making use of the standard integral operator framework. Further, we introduce unlabeled data to enlarge the number of partitions in the same optimal learning rates by reducing label independent errors in error decomposition. The final result is given by exploring random features in a data-dependent generating way to reduce the

Preprint. Under review.

Table 1: Summary of the number of partitions, the number of random centers and computational costs for kernel ridge regression (KRR), KRR with Nyström (KRR-Nyström), KRR with random features (KRR-RF), KRR with divide-and-conquer (KRR-DC) and three theoretical results of the proposed `KRR-DC-RF`.

Methods	Partitions m	Random centers M	Space	Time
KRR [11]	/	/	$\mathcal{O}(N^2)$	$\mathcal{O}(N^3)$
KRR-Nyström [9]	/	$\mathcal{O}(N^{\frac{1}{2r+\gamma}})$	$\mathcal{O}(NM)$	$\mathcal{O}(NM^2)$
KRR-RF [13]	/	$\mathcal{O}(N^{\frac{(2r-1)\gamma+1}{2r+\gamma}})$	$\mathcal{O}(NM)$	$\mathcal{O}(NM^2)$
KRR-DC [12]	$\mathcal{O}(N^{\frac{2r-1}{2r+\gamma}})$	/	$\mathcal{O}(N^2/m^2)$	$\mathcal{O}(N^3/m^3)$
Theorem 1	$\mathcal{O}(N^{\frac{2r-1}{2r+\gamma}})$	$\mathcal{O}(N^{\frac{(2r-1)\gamma+1}{2r+\gamma}})$	$\mathcal{O}(NM/m)$	$\mathcal{O}(NM^2/m)$
Theorem 2	$\mathcal{O}(N^*N^{\frac{-\gamma-1}{2r+\gamma}})$	$\mathcal{O}(N^{\frac{(2r-1)\gamma+1}{2r+\gamma}})$	$\mathcal{O}(N^*M/m)$	$\mathcal{O}(N^*M^2/m)$
Theorem 3	$\mathcal{O}(N^*N^{\frac{-\gamma-1}{2r+\gamma}})$	$\mathcal{O}(N^{\frac{2r+\gamma-1}{2r+\gamma}})$	$\mathcal{O}(N^*M/m)$	$\mathcal{O}(N^*M^2/m)$

Note: All listed methods achieve the optimal learning rate $\mathcal{O}(N^{-2r/(2r+\gamma)})$, where N is the amount of labeled examples, N^* is the number of all samples including labeled and unlabeled examples, $\gamma \in [0, 1]$ is defined by Assumption 4 and $r \in [1/2, 1]$ is defined by Assumption 5. The results of Theorems 2 and 3 are simplified with $N^* \leq N^{(2r+3\gamma)/(2r+\gamma)}$ and $\alpha = \gamma$.

features needed in optimal statistical properties, of which a constant number of random features is sufficient to reach $\mathcal{O}(1/N)$ learning rate in some cases. In the steps of proof, we propose a novel error decomposition that decomposes the *excess risk* of `KRR-DC-RF` into variance, empirical error, distributed error, random feature error and approximation error. By this decomposition, we demonstrate how unlabeled data and data-dependent features reduce errors of some terms.

Related works and comparison. The proposed approach combining divide-and-conquer and random features (`KRR-DC-RF`) to reduce computational cost dramatically is very intuitive. The work in [17] has empirically validated high efficiency and favorable accuracy of `KRR-DC-RF`, while in this paper, we focus on its statistical learning to reach a good tradeoff between generalization performance and computation cost. The optimal learning rate for KRR with divide-and-conquer was firstly presented in [2, 18] under some eigenfunction assumptions and extended into feature space in [19]. Eigenfunction assumptions were removed in [12] by using traditional integral operator and extended to semi-supervised learning [20] and multi-pass SGD [15]. Rudi and Rosasco derived the optimal statistical error bounds of random features [13] by applying standard integral operator framework [10, 11] into feature space, and the result was further studied in [21] and [22]. Table 1 reports the statistical and computational properties of related approaches and our main results. The table demonstrates general result of `KRR-DC-RF` Theorem 1 improve computational efficiency and reduce memory requirement dramatically while preserving optimal statistical properties. For example, the learning rate achieves $\mathcal{O}(1/N)$ with $m = \mathcal{O}(\sqrt{N})$ and $M = \mathcal{O}(\sqrt{N})$ when $r = 1$ and $\gamma = 0$, corresponding $\mathcal{O}(N)$ in space and $\mathcal{O}(N^{1.5})$ in time. Theorem 2 employees additional unlabeled data to alleviate the dilemma of $\mathcal{O}(1)$ partitions when $r = 1/2$. Theorem 3 consider generating random features in a data-dependent way, dramatically reduce the number of features needed. For example, a constant number of random features is sufficient to achieve the optimal learning rate $\mathcal{O}(1/N)$ with $\mathcal{O}(N)$ space and $\mathcal{O}(N)$ time when $r = 1/2$ and $\gamma = 0$.

2 Distributed Learning with Random Feature

2.1 Kernel Ridge Regression (KRR)

In a standard framework of supervised learning, there is a probability space $\mathcal{X} \times \mathcal{Y}$ with a fixed but unknown distribution ρ, where $\mathcal{X} = \mathbb{R}^d$ is the input space and $\mathcal{Y}$ is the output space. The training set $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ is sampled identically and independently from $\mathcal{X} \times \mathcal{Y}$ with respect to ρ. Given a hypothesis space $\mathcal{H}$ of measurable functions from $\mathcal{X}$ to $\mathcal{Y}$, the goal of regression problem with squared loss and continuous output space $\mathcal{Y} = \mathbb{R}$ is to minimize the *expected risk*

$$\min_{f \in \mathcal{H}} \mathcal{E}(f), \quad \mathcal{E}(f) = \int_{\mathcal{X} \times \mathcal{Y}} (f(\mathbf{x}) - y)^2 d\rho(\mathbf{x}, y). \tag{1}$$

Kernel ridge regression (KRR) is a classical way to derive an empirical solution to (1), based on choosing a separable Reproducing Kernel Hilbert Space (RKHS) as hypothesis space $\mathcal{H}$, which is

induced by a Mercer kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Kernel ridge regression (KRR) can be state as

$$\widehat{f}_\lambda = \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{N} \sum_{i=1}^{N} (f(\mathbf{x}_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}. \tag{2}$$

With the represent theorem [23], the problem (2) exists a unique closed form solution

$$\widehat{f}_\lambda(\mathbf{x}) = \sum_{i=1}^{N} \widehat{\alpha}_i K(\mathbf{x}_i, \mathbf{x}), \quad \text{with} \quad \widehat{\alpha} = (\mathbf{K}_N + \lambda N I)^{-1} \mathbf{y}_N, \tag{3}$$

where $\lambda > 0, \mathbf{y}_N = (y_1, \cdots, \mathbf{y}_N)$ and $\mathbf{K}_N$ is the $N \times N$ kernel matrix with $\mathbf{K}_N(i,j) = K(\mathbf{x}_i, \mathbf{x}_j)$. Although KRR has has optimal statistical properties [10, 11], it becomes unfeasible as sample size n increases because of $\mathcal{O}(N^2)$ memory to store kernel matrix and $\mathcal{O}(N^3)$ time to solve the linear system (3) by matrix inversion.

To tackle those scalability issues but also keep the optimal learning rates, several speedup approaches have been studied: (1) Divide-and-conquer approaches [2, 7] which decompose a large scale problem into smaller ones and are processed in individual machines. (2) Random projections including Nyström methods [9] and random features [13**?**] to reduce data dimensionality. In this paper, we consider combining the benefits of both methods to deal with extremely large-scale applications but also obtain optimal statistical guarantees.

2.2 KRR with Distributed Learning (KRR-DC)

The paper focus on large scale setting where $N \gg d$. We use the divide-and-conquer scheme [18] due to its lowest communication rounds (only once). Let the training set D be randomly partitioned into m disjoint subsets $\{D_j\}_{j=1}^m$ with $|D_1| = \cdots = |D_m| = n$. Then those partitions are assigned to m disjoint local processors to produce a local estimator $\widehat{f}_{D_j,\lambda}$ by the solution KRR (3)

$$\widehat{f}_{D_j,\lambda}(\mathbf{x}) = \sum_{i=1}^{n} \widehat{\alpha}_{ij} K(\mathbf{x}_i, \mathbf{x}), \quad \text{with} \quad \widehat{\alpha}_j = (\mathbf{K}_n + \lambda n I)^{-1} \mathbf{y}_n, \tag{4}$$

where $\mathbf{K}_n$ is the empirical kernel matrix on subset D_j and $\mathbf{y}_n = (y_1, \cdots, y_n)$ on D_j. Finally, those local estimators are summarized to a central node and a global estimator $\widehat{f}_{D,\lambda}$ is computed by weighted average

$$\widehat{f}_{D,\lambda} = \frac{1}{m} \sum_{j=1}^{m} \widehat{f}_{D_j,\lambda}. \tag{5}$$

2.3 KRR with Divide-and-Conquer and Random Features (KRR-DC-RF)

The basic idea of random features [4, 24–26] is to approximate positive definite kernel by explicit feature mapping $\phi_M : \mathbb{R}^d \to \mathbb{R}^M$

$$K(\mathbf{x}, \mathbf{x}') \approx \langle \phi_M(\mathbf{x}), \phi_M(\mathbf{x}') \rangle. \tag{6}$$

We introduce a general strategy to produce random features to approximate kernel as the form (6) then. Assume that the kernel K have an integral representation

$$K(\mathbf{x}, \mathbf{x}') = \int_\Omega \psi(\mathbf{x}, \omega) \psi(\mathbf{x}', \omega) d\pi(\omega), \quad \forall \mathbf{x}, \mathbf{x}' \in \mathcal{X},$$

where (Ω, π) is a probability space and $\psi : \mathcal{X} \times \Omega \to \mathbb{R}$. Define analogous operators for the approximated kernel $K_M(\mathbf{x}, \mathbf{x}') = \phi_M(\mathbf{x})^\top \phi_M(\mathbf{x}')$ to approximate $K(\mathbf{x}, \mathbf{x}')$ in (6) with

$$\phi_M(\mathbf{x}) = \frac{1}{\sqrt{M}} \big(\psi(\mathbf{x}, \omega_1), \cdots, \psi(\mathbf{x}, \omega_M) \big),$$

where $\omega_1, \cdots, \omega_M$ are sampled independently with respect to π.

Using random features in (6), the approximate solution of a local estimator $\widehat{f}_{D_j,\lambda}$ in (4) is

$$\widehat{f}^M_{D_j,\lambda}(\mathbf{x}) = \phi_M(\mathbf{x})^\top \widehat{w}_j, \quad \text{with} \quad \widehat{w}_j = (\widehat{S}_M^\top \widehat{S}_M + \lambda I)^{-1} \widehat{S}_M^\top \widehat{\mathbf{y}}_n, \tag{7}$$

where $\lambda > 0$. Note that for j-th subset D_j, $\forall (\mathbf{x}, y) \in D_j$, $\widehat{S}_M^\top = \frac{1}{\sqrt{n}}(\phi_M(\mathbf{x}_1), \cdots, \phi_M(\mathbf{x}_n))$ and $\widehat{\mathbf{y}}_n = \frac{1}{\sqrt{n}}(y_1, \cdots, y_n)$.

The weighted average of approximate local estimators output a approximate global estimator

$$\widehat{f}^M_{D,\lambda} = \frac{1}{m} \sum_{j=1}^{m} \widehat{f}^M_{D_j,\lambda}. \tag{8}$$

3 Main Results

In this section, we present the theoretical analysis on generalization performance of kernel ridge regression with divide-and-conquer and random features. We firstly provide a general result with the optimal statistical properties under standard assumptions, the same as primal kernel ridge regression. Then, we consider additional unlabeled data to reduce distributed error and further increase the number of partitions with optimal learning rates. Finally, beyond uniform sampling, data-dependent features generating strategy is introduced to reduce the number of random features. The proofs of following results are given in the appendix.

In the beginning, we introduce the definition of the *excess risk* and three basic assumptions which are widely used in statistical learning of squared loss [10, 11]. To explore the generalization ability of `KRR-DC-RF` estimator $\widehat{f}^M_{D,\lambda}$, the *excess risk* is defined as

$$\mathbb{E}[\mathcal{E}(\widehat{f}^M_{D,\lambda})] - \mathcal{E}(f_{\mathcal{H}}). \tag{9}$$

To control basic properties of induced kernel which is continuous and bounded, we need the following assumption which is satisfied by popular Fourier random features to approximate shift-invariant kernels and other random features in [13, 27] and references therein.

Assumption 1 (Random features are continuous and bounded)**.** *Assume that ψ is continuous and there is a $\kappa \in [1, \infty)$, such that $|\psi(\mathbf{x}, \omega)| \leq \kappa, \forall \mathbf{x} \in \mathcal{X}, \omega \in \Omega$.*

Assumption 2 (Consistency assumption)**.** *Assume there exists the best solution $f_{\mathcal{H}} \in \mathcal{H}$, such that*

$$\mathcal{E}(f_{\mathcal{H}}) = \inf_{f \in \mathcal{H}} \mathcal{E}(f).$$

The above assumption is standard in kernel-based nonparametric regression [10, 11, 28]. We also need a basic assumption on data distribution to derive probabilistic results.

Assumption 3 (Moment assumption)**.** *Assume there exists $B > 0$ and $\sigma > 0$, such that for all $p \geq 2$ with $p \in \mathbb{N}$,*

$$\int_{\mathbb{R}} |y|^p d\rho(y|\mathbf{x}) \leq \frac{1}{2} p! B^{p-2} \sigma^2.$$

Typically, the above assumption on output y holds when y is bounded, sub-gaussian or sub-exponential. This assumption can be relaxed to $|y| \leq b, \forall b > 1$, then the assumption is satisfied with $\sigma = B = 2b$.

The above Assumptions 1, 2 and 3 are basic conditions in generalization analysis of kernel ridge regression, always leading $\mathcal{O}(1/\sqrt{N})$ learning rate in worst case.

3.1 General Result with Fast Rates

Using traditional integral operator techniques, we derive general results with fast rates under further favorable assumptions. Those two assumptions are common in kernel ridge regression and approximation theory [29], controlling the capacity of the hypothesis $\mathcal{H}$ and regularity of $f_{\mathcal{H}}$, respectively.

Definition 1 (Integral operator)**.** *Integral operator is defined as*

$$(Lg)(\mathbf{x}) = \int_{\mathcal{X}} K(\mathbf{x}, \mathbf{z}) g(\mathbf{z}) d\rho_X(\mathbf{z}), \quad \forall g \in L^2(\mathcal{X}, \rho_X),$$

where $L^2(\mathcal{X}, \rho_X) = \{f : \mathcal{X} \to \mathbb{R} \mid \|f\|_\rho^2 = \int |f(\mathbf{x})|^2 d\rho_X < \infty\}$, K is the induced kernel and ρ_X is the marginal distribution of ρ on $\mathcal{X}$.

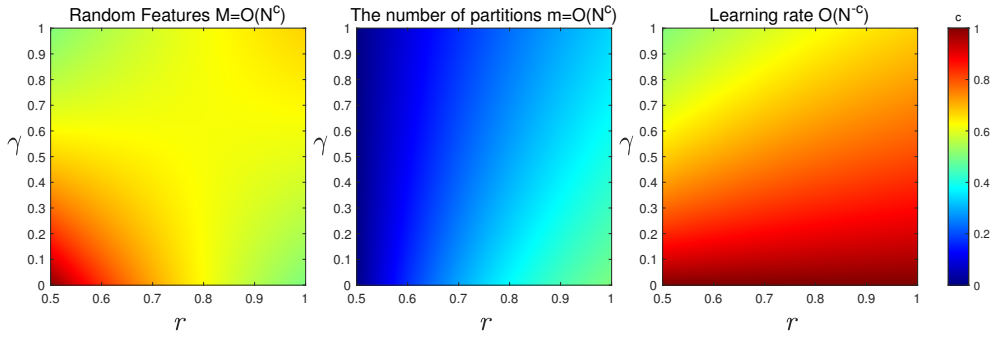

Figure 1: Influence of the capacity of RKHS and the regularity of $f_{\mathcal{H}}$. Bigger r : higher regularity. Smaller γ: smaller RKHS.

Since kernel function K is continuous, symmetric and positive definite, the integral operator L is a compact positive operator of trace class and $L + \lambda I$ is invertible. And integral operator is often used to measure the complexity of hypothesis $\mathcal{H}$ by the effective dimension.

Definition 2 (Effective dimension)**.** *The effective dimension is defined as*

$$\mathcal{N}(\lambda) = \operatorname{Tr}\left((L + \lambda I)^{-1} L\right), \quad \lambda > 0.$$

Assumption 4 (Capacity assumption)**.** *Assume there exists $Q > 0$ and $\gamma \in [0, 1]$, such that for any $\lambda > 0$*

$$\mathcal{N}(\lambda) \leq Q^2 \lambda^{-\gamma}.$$

Assumption 5 (Regularity assumption)**.** *Assume there exists $r \in [1/2, 1]$ and $g \in L^2(\mathcal{X}, \rho_X)$ such that*

$$f_{\mathcal{H}}(\mathbf{x}) = (L^r g)(\mathbf{x}).$$

Above two conditions are commonly used to prove the optimal statistical properties of combination of KRR and large scale algorithms including divide-and-conquer [7] and random features [13]. We provide some intuitive interpretation of the above assumptions and more details can be found in [11]. The effective dimension is often used to measure the complexity of the hypothesis space $\mathcal{H}$, thus Assumption 4 controls the variance of the estimator and is equivalent to the classic entropy and covering number conditions [30]. The value of γ inflects the size of RKHS $\mathcal{H}$. Thus, the more benign situation with smaller RKHS is obtained when $\gamma = 0$, while the worst case corresponds to $\gamma = 1$. Assumption 5 controls the bias of the estimator and is commonly used in approximation theory [10], which can be seen as regularity of $f_{\mathcal{H}}$. The case that $\gamma = 1$ and $r = 1/2$ corresponds making no assumptions on the kernel, reducing to the worst case.

Theorem 1. *Under Assumptions 1, 2, 3, 4 and 5, for the following condition $n \geq n_0, \lambda = N^{-\frac{1}{2r+\gamma}}$, and the number of random features M, the number of partitions m respectively corresponds to*

$$M \gtrsim N^{\frac{(2r-1)\gamma+1}{2r+\gamma}}, \quad m \lesssim N^{\frac{2r-1}{2r+\gamma}}$$

suffice to guarantee with high probability that

$$\mathbb{E}\left[\mathcal{E}(\widehat{f}^M_{D,\lambda})\right] - \mathcal{E}(f_{\mathcal{H}}) = \mathcal{O}\left(N^{-\frac{2r}{2r+\gamma}}\right).$$

Note that the optimal learning rates stated in Theorem 1 are the same as the bound obtained by primal KRR [11], KRR-DC [12] under the same restriction on the number of partitions that $m \lesssim N^{\frac{2r-1}{2r+\gamma}}$, and KRR-RF [13] under the same restriction on the number of random features that $M \gtrsim N^{\frac{(2r-1)\gamma+1}{2r+\gamma}}$. The result is optimal in a minimax sense [11] and lower bounds are proved in [11, 28]. Further, Figure 1 provides a visual representation of the number of random features needed, the number of local estimators allowed in terms of learning rates due to different settings of r and γ, where the direction of bottom-right leads higher regularity and smaller RKHS. In the best case $r = 1$ and $\gamma = 0$ (higher regularity and a smaller RKHS), a learning rate $\mathcal{O}(1/N)$ can be achieved by $\mathcal{O}(\sqrt{N})$ random features and $\mathcal{O}(\sqrt{N})$ partitions. Note that a smaller RKHS ($\gamma = 0$) provides optimal learning rates $\mathcal{O}(1/N)$ despite the value of r as shown in the right of Figure 1. Moreover, lower regularity ($r = 1/2$) leads to $\mathcal{O}(1)$ partitions as in the middle of Figure 1 that limits the applications of distributed learning.

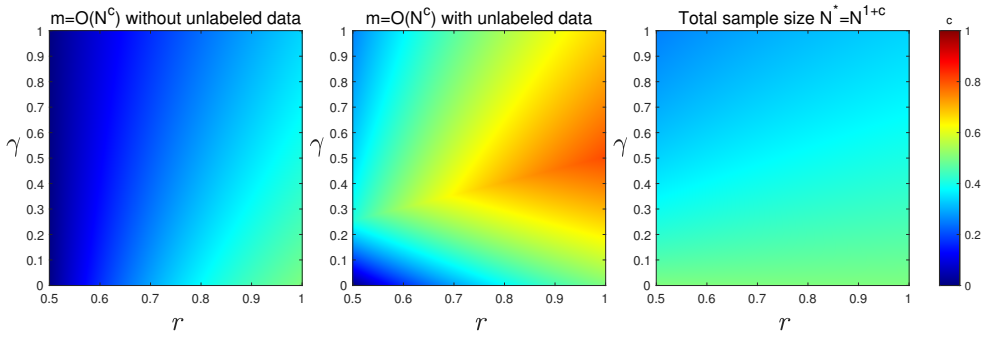

Figure 2: The number of local estimators m and total sample size N^* with $N^* = N^{1+\frac{r}{2r+\gamma}}$. Left: only use labeled data. Middle and Right: use additional unlabeled data.

Remark 1. *The worst case $r = 1/2$ and $\gamma = 1$, in other word only under Assumptions 1, 2 and 3, shows that $\mathcal{O}(\sqrt{N})$ random features and a constant number of local estimators can guarantee $\mathcal{O}(1/\sqrt{N})$ learning rate. The number of local estimators in the worst case is $m = \mathcal{O}(1)$, independent on sample size N, which is very restrictive in large scale settings. In our follow-up work, we employee additional unlabeled samples to relax the restriction $m \lesssim N^{\frac{2r-1}{2r+\gamma}}$, as done in [20, 31].*

Remark 2. *The sampling scheme of random features is data-independent that discards a part of useful information [13]. In Section 3.3, We consider generating random features in a data-dependent way to reduce features needed for same learning rate [13, 32].*

3.2 More Partitions Using Unlabeled Data

The error decomposition in Lemma 1 of Section 4 demonstrates that additional unlabeled data plays a crucial role in deducing smaller empirical error and distributed error and thus relaxing heavily the restriction on m. Borrowing the distributed semi-supervised framework used in [20], additional unlabeled subsets $\{\widetilde{D}_j\}_{j=1}^m$ are drawn identically and independently from the conditional distribution ρ_X and are stored in local processors. Consider the merged dataset D^* on the j-th processor,

$$D_j^* = \{D_j \cup \widetilde{D}_j\}_{j=1}^m$$

with

$$\mathbf{x}_i^* = \begin{cases} \mathbf{x}_i, & \text{if}(\mathbf{x}_i, y_i) \in D_j, \\ \widetilde{\mathbf{x}}_i, & \text{otherwise}, \end{cases} \quad \text{and} \quad y_i^* = \begin{cases} \frac{|D_j^*|}{|D_j|} y_i, & \text{if}(\mathbf{x}_i, y_i) \in D_j, \\ 0, & \text{otherwise}. \end{cases}$$

Let $D^* = \bigcup_{j=1}^m D_{j=1}^*, |D^*| = N^*$ and $|D_1^*| = \cdots = |D_m^*| = n^*$. We define semi-supervised kernel ridge regression with divide-and-conquer and random features (`SKRR-DC-RF`) by

$$\widehat{f}_{D^*,\lambda}^M = \frac{1}{m} \sum_{j=1}^m \widehat{f}_{D_j^*,\lambda}^M. \tag{10}$$

According to error decomposition in Lemma 1 below, empirical error and distributed error are data-dependent but label-independent, thus additional unlabeled samples can reduce them to enlarge the number of local estimators under same optimal error bounds.

Theorem 2. *Under Assumptions 1, 2, 3, 4 and 5, if $n \geq n_0, \lambda = N^{-\frac{1}{2r+\gamma}}$, and the number of random features M, the number of partitions m corresponds to*

$$M \gtrsim N^{\frac{(2r-1)\gamma+1}{2r+\gamma}}, \qquad m \lesssim \min\left\{N^{\frac{2r+2\gamma-1}{2r+\gamma}},\ N^* N^{\frac{-\gamma-1}{2r+\gamma}}\right\}$$

then the following holds with high probability,

$$\mathbb{E}\big[\mathcal{E}(\widehat{f}_{D^*,\lambda}^M)\big] - \mathcal{E}(f_{\mathcal{H}}) = \mathcal{O}\Big(N^{-\frac{2r}{2r+\gamma}}\Big).$$

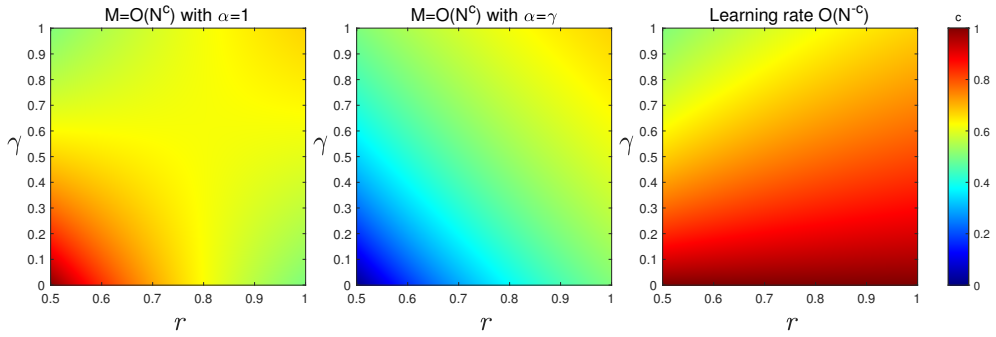

Figure 3: The number of random features needed for corresponding learning rates in different features generating ways. Left: data-independent. Middle: data-dependent.

When there is no unlabeled data that $N^* = N$, the result coincides with Theorem 1. Note that additional unlabeled data does not influence optimal learning rates. We consider $N^* = N^{1+\frac{r}{2r+\gamma}}$ thus $N^* \in [N^{1.25}, N^{1.5}]$ that is a common scene in large scale semi-supervised learning. Figure 2 shows the number of partitions increase a lot after taking into account unlabeled examples. Especially, the spacial cases of $\mathcal{O}(1)$ partitions are reduced from $r = 1/2$ to only one point $r = 1/2, \gamma = 1$.

Corollary 1 (The worst case after using unlabeled data)**.** *Under Assumptions 1, 2 and $y \leq |b|$ with $b > 0$, if $n \geq n_0, \lambda = N^{-1/2}$, and the number of random features M, the number of partitions m respectively corresponds to*

$$M \gtrsim \sqrt{N}, \quad m \lesssim \min\Big\{N, \frac{N^*}{N}\Big\}$$

is enough to guarantee with high probability, that

$$\mathbb{E}\big[\mathcal{E}(\widehat{f}^M_{D^*,\lambda})\big] - \mathcal{E}(f_{\mathcal{H}}) = \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big).$$

The learning rate $\mathcal{O}(1/\sqrt{N})$ of Corollary 1 in a worst case situation is the same prediction accuracy of the standard KRR. As long as there exists unlabeled data that $N^* = N^{1+\beta},\ \beta > 0$ then the estimator using $\mathcal{O}(N^\beta)$ partitions and $\mathcal{O}(\sqrt{N})$ random features has optimal generalization properties. That demonstrates more than a constant number of partitions are allowed as long as unlabeled data available, as well the number of partitions increases as the labeled sample size N does.

3.3 Fewer Random Features Using Data-dependent Sampling

Under the following assumption, we explore fewer random features to obtain optimal learning bounds by generating features in a data-dependent manner, which has been well studied in [13, 22, 27].

Assumption 6 (Compatibility assumption)**.** *Define the maximum dimension of random features as*

$$\mathcal{F}_\infty = \sup_{\omega\in\Omega} \|(L+\lambda I)^{-1/2}\psi(,\omega)\|^2_{\rho_X},$$

where $\lambda > 0$. Assume there exists $\alpha \in [0, 1]$ and $F > 0$, such that $\mathcal{F}_\infty \leq F\lambda^{-\alpha}$.

The above assumption bridges random features with data distribution by the operator L. It always holds when $F = \kappa^2$ and $\alpha = 1$ by Assumption 1 and the favorable case corresponds to $\alpha = \gamma$. Theoretical examples are given in [13, 27] and refined leverage score algorithms are stated in [22].

Theorem 3. *Under Assumption 6 and the same assumptions of Theorem 1, if $n \geq n_0, \lambda = N^{-\frac{1}{2r+\gamma}}$, and the number of random features M, the number of partitions m corresponds to*

$$M \gtrsim N^{\frac{(2r-1)(\gamma-\alpha+1)+\alpha}{2r+\gamma}}, \qquad m \lesssim \min\Big\{N^{\frac{2r+2\gamma-1}{2r+\gamma}},\ N^* N^{\frac{-\gamma-1}{2r+\gamma}}\Big\}$$

then the following holds with high probability,

$$\mathbb{E}\big[\mathcal{E}(\widehat{f}^M_{D^*,\lambda})\big] - \mathcal{E}(f_{\mathcal{H}}) = \mathcal{O}\Big(N^{-\frac{2r}{2r+\gamma}}\Big).$$

The above learning bound is the same as Theorems 1, 2. In Figure 3, we compare the number of features generating in data-independent way ($\alpha = 1$) and in data-dependent way ($\alpha = \gamma$). It shows that much fewer data-dependent features are needed than uniformly sampled ones for the same learning rates. Moreover, a constant number of data-dependent features are sufficient to guarantee $\mathcal{O}(1/N)$ learning rate when $r = 1/2$ and $\gamma = 0$. The above result shows the dramatic effect of problem dependent random features allowing computational gains without loss of accuracy.

4 Sketch of Proof

In this section, we introduce the sketch of proof while details are deferred to the appendix. The main idea of the proof is to decompose analytically *excess risk* $\mathbb{E}[\mathcal{E}(\widehat{f}^M_{D^*,\lambda})] - \mathcal{E}(f_{\mathcal{H}})$ in (9) into several errors, and then bound them by concentration inequalities. Different from error decomposition of the standard KRR, the proposed `SKRR-DC-RF` introduces two additional errors: distributed error and random features error, due to the using of divide-and-conquer and random features.

To explain the decomposition clearly, we provide some estimators at first. Firstly, we rewrite the `SKRR-DC-RF` estimator $\widehat{f}^M_{D^*_j,\lambda}$ in (10) in primal form and denote other useful estimators as follows

$$\widehat{f}^M_{D^*,\lambda} = \frac{1}{m}\sum_{j=1}^{m}\langle \widehat{w}_j, \phi_M(\cdot)\rangle,\ \widehat{w}_j = \mathop{\arg\min}_{w\in\mathbb{R}^M}\left\{\frac{1}{n^*}\sum_{i=1}^{n^*}(\langle w, \phi_M(\mathbf{x}^*_i)\rangle - y^*_i)^2 + \lambda\|w\|^2\right\},$$

$$\widetilde{f}^M_{D^*,\lambda} = \frac{1}{m}\sum_{j=1}^{m}\langle \widetilde{w}_j, \phi_M(\cdot)\rangle,\ \widetilde{w}_j = \mathop{\arg\min}_{w\in\mathbb{R}^M}\left\{\frac{1}{n^*}\sum_{i=1}^{n^*}\big(\langle w, \phi_M(\mathbf{x}^*_i)\rangle - f_{\mathcal{H}}(\mathbf{x}^*_i)\big)^2 + \lambda\|w\|^2\right\},$$

$$f^M_\lambda = \langle \widehat{u}, \phi_M(\cdot)\rangle,\ u = \mathop{\arg\min}_{u\in\mathbb{R}^M}\int_{\mathcal{X}}\big(\langle u, \phi_M(\mathbf{x})\rangle - f_{\mathcal{H}}(\mathbf{x})\big)^2 d\rho_X(\mathbf{x}) + \lambda\|u\|^2,$$

$$f_\lambda = \langle \widehat{u}, \phi(\cdot)\rangle,\ v = \mathop{\arg\min}_{v\in\mathcal{H}_K}\int_{\mathcal{X}}\big(\langle v, \phi(\mathbf{x})\rangle - f_{\mathcal{H}}(\mathbf{x})\big)^2 d\rho_X(\mathbf{x}) + \lambda\|v\|^2,$$

where $\phi : \mathbb{R}^d \to \mathcal{H}_K$ is feature map associated to the kernel K by $K(\mathbf{x},\mathbf{x}') = \langle\phi(\mathbf{x}), \phi(\mathbf{x}')\rangle$. The empirical estimator $\widetilde{f}^M_{D^*,\lambda}$ focuses on noise-free data. The last two vectors are both expected estimators defined by random features ϕ_M and implicit feature map ϕ. From [11, 33], there holds

$$\mathbb{E}\big[\mathcal{E}(\widehat{f}^M_{D^*,\lambda})\big] - \mathcal{E}(f_{\mathcal{H}}) = \mathbb{E}\|\widehat{f}^M_{D^*,\lambda} - f_{\mathcal{H}}\|^2_\rho. \tag{11}$$

Combining (11) and the identity $\widehat{f}^M_{D^*,\lambda} - f_{\mathcal{H}} = \widehat{f}^M_{D^*,\lambda} - f^M_\lambda + f^M_\lambda - f_\lambda + f_\lambda - f_{\mathcal{H}}$, we obtain the error decomposition in Lemma 1 and its proof is provided in appendix.

Lemma 1. *Let $\widehat{f}^M_{D^*,\lambda}, \widetilde{f}^M_{D^*,\lambda}, f^M_\lambda$ and f_λ be defined as the above, we have*

$$\mathbb{E}\big[\mathcal{E}(\widehat{f}^M_{D^*,\lambda})\big] - \mathcal{E}(f_{\mathcal{H}}) \tag{12}$$

$$\leq \frac{6}{m^2}\sum_{j=1}^{m}\mathbb{E}\|\widehat{f}^M_{D^*_j,\lambda} - \widetilde{f}^M_{D^*_j,\lambda}\|^2_\rho \quad (\textit{Variance}) \tag{13}$$

$$+ \frac{6}{m^2}\sum_{j=1}^{m}\mathbb{E}\|\widetilde{f}^M_{D^*_j,\lambda} - f^M_\lambda\|^2_\rho \quad (\textit{Empirical error}) \tag{14}$$

$$+ \frac{3}{m}\sum_{j=1}^{m}\mathbb{E}\|\widetilde{f}^M_{D^*_j,\lambda} - f^M_\lambda\|^2_\rho \quad (\textit{Distributed Error}) \tag{15}$$

$$+ \ 3\,\|f^M_\lambda - f_\lambda\|^2_\rho \quad (\textit{Random Features Error}) \tag{16}$$

$$+ \ 3\,\|f_\lambda - \ f_{\mathcal{H}}\|^2_\rho \quad (\textit{Approximation Error}). \tag{17}$$

Variance (13) is brought by noise on labels y thus output dependent. Empirical error (14) represents the gap between expected learning and empirical learning. Distributed error (15) measures the limitation of the distributed learning algorithm (10). Note that empirical error and distributed

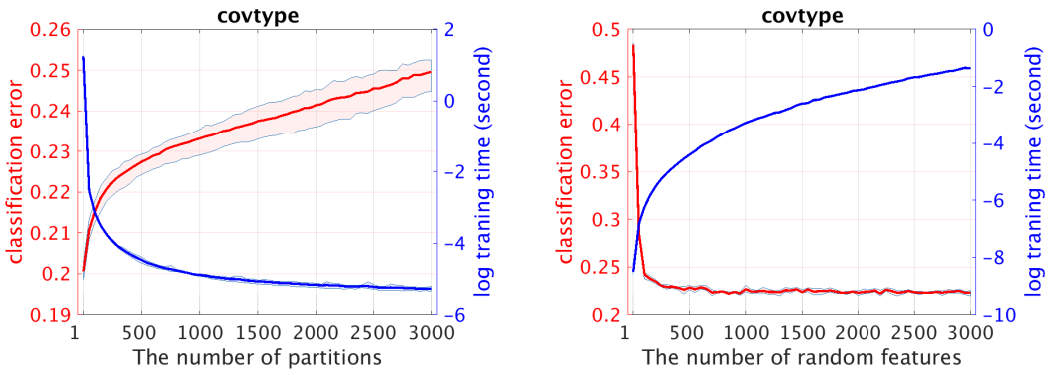

Figure 4: On dataset covtype.

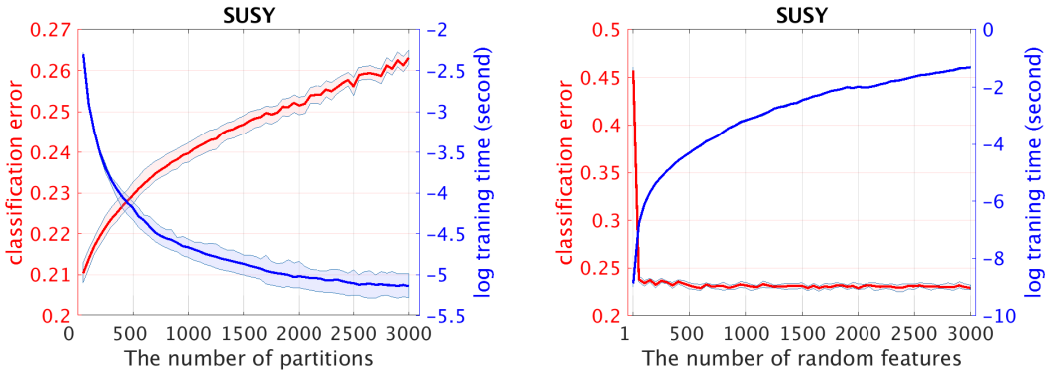

Figure 5: On dataset SUSY.

error focus on noise-free data, therefore, can be reduced by additional unlabeled data, resulting in Theorem 2. Independent on the sample, random features error (16) accounts for approximation capability of random features to the kernel and approximation error (17) reflects bias of the algorithm. Data-dependent generating features can reduce random features error (16) that motivates Theorem 3.

5 Experiments

We study the empirical performance of `KRR-DC-RF` algorithm on random sampled 2.5×10^5 data points on binary classification datasets covtype [2] and SUSY [3] and HIGGS [4], where $\sqrt{N} = 500$. We use random Fourier features to approximate Gaussian kernel $K(\mathbf{x}, \mathbf{x}') = \exp^{-\|\mathbf{x}-\mathbf{x}'\|^2/2\sigma^2}$ [4]. Random fourier features are in the form $\psi(\mathbf{x}, \omega) = \cos(\omega^T \mathbf{x} + b)$, where ω is drawn from the normal distribution and b is drawn from uniform distribution $[0, 2\pi]$. In the following experiments, we tune parameters σ and λ by 10-folds cross-validation for every dataset and report average over 10 repetitions of the algorithm.

Firstly, we explore how the number of partitions affect accuracy and training time of the algorithm. We use $\sqrt{N}$ random features and vary the number of partitions among $\{1, 50 \times \{1, 2, \cdots, 60\}\}$. Results in the left of Figures 4, 5 and 6 show that `KRR-DC-RF` can dramatically reduce training time but also not loss too much accuracy. Then, we study empirical performance in terms of different numbers of random features. esults in the right of Figures 5, 4 and 6 show that $\sqrt{N}$ random features provide favorable accuracy with high efficiency, which coincides to our analysis.

[2] https://archive.ics.uci.edu/ml/datasets/covertype

[3] https://archive.ics.uci.edu/ml/datasets/susy

[4] https://archive.ics.uci.edu/ml/datasets/higgs

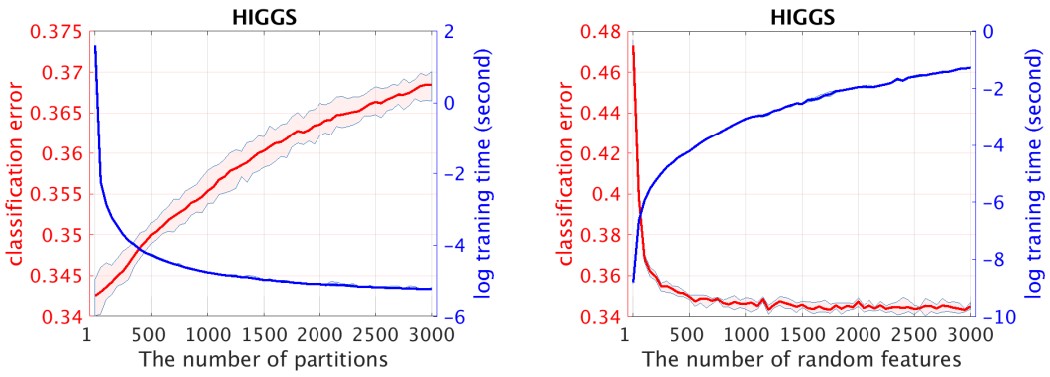

Figure 6: On dataset HIGGS.

6 Conclusion

In this paper, we explore the generalization performance of kernel ridge regression with commonly used efficient large scale techniques: divide-and-conquer and random features. Statistical learning shows the combination achieves a good tradeoff between statistical properties and computational requirements. We firstly present a general result for optimal statistical accuracy under standard assumptions. Further, we give refined results by using unlabeled data to increase the number of partitions and using data-dependent features, generating a way to reduce the number of random features. Moreover, we can extend the proposed work in several ways: (a) combine the approach with gradient algorithms such as multi-pass SGD [16] and preconditioned conjugate gradient [34]. (b) replace random features with other random projections (i.e. Nyström methods [9] or circulant[35]). (c) replace divide-and-conquer with asynchronous distributed methods [36, 37].

References

[1] Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems 21 (NIPS)*, pages 161–168, 2008.

[2] Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research*, 16(1):3299–3340, 2015.

[3] Christopher KI Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 14 (NIPS)*, pages 682–688, 2001.

[4] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 21 (NIPS)*, pages 1177–1184, 2007.

[5] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *The Journal of Machine Learning Research*, 15(1):335–366, 2014.

[6] Kurt Cutajar, Michael Osborne, John Cunningham, and Maurizio Filippone. Preconditioning kernel matrices. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 2529–2538, 2016.

[7] Shao-Bo Lin, Xin Guo, and Ding-Xuan Zhou. Distributed learning with regularized least squares. *The Journal of Machine Learning Research*, 18(1):3202–3232, 2017.

[8] Junhong Lin and Lorenzo Rosasco. Optimal learning for multi-pass stochastic gradient methods. In *Advances in Neural Information Processing Systems 29 (NIPS)*, pages 4556–4564, 2016.

[9] Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pages 1657–1665, 2015.

[10] Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172, 2007.

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

[12] Zheng-Chu Guo, Shao-Bo Lin, and Lei Shi. Distributed learning with multi-penalty regularization. *Applied and Computational Harmonic Analysis*, 2017.

[13] Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 3215–3225, 2017.

[14] Aymeric Dieuleveut, Francis Bach, et al. Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363–1399, 2016.

[15] Junhong Lin and Volkan Cevher. Optimal distributed learning with multi-pass stochastic gradient methods. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 3098–3107, 2018.

[16] Luigi Carratino, Alessandro Rudi, and Lorenzo Rosasco. Learning with sgd and random features. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 10192–10203, 2018.

[17] Brian McWilliams, Christina Heinze, Nicolai Meinshausen, Gabriel Krummenacher, and Hastagiri P Vanchinathan. Loco: Distributing ridge regression with random projections. *stat*, 1050:26, 2014.

[18] Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression. In *Conference on Learning Theory*, pages 592–617, 2013.

[19] Shusen Wang. A sharper generalization bound for divide-and-conquer ridge regression. 2019.

[20] Xiangyu Chang, Shao-Bo Lin, and Ding-Xuan Zhou. Distributed semi-supervised learning with kernel ridge regression. *Journal of Machine Learning Research*, 18(1):1493–1514, 2017.

[21] Junhong Lin and Volkan Cevher. Optimal rates of sketched-regularized algorithms for least-squares regression over hilbert spaces. *arXiv preprint arXiv:1803.04371*, 2018.

[22] Alessandro Rudi, Daniele Calandriello, Luigi Carratino, and Lorenzo Rosasco. On fast leverage score sampling and optimal learning. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 5672–5682, 2018.

[23] Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer, 2001.

[24] Quoc Le, Tamás Sarlós, and Alex Smola. Fastfood-approximating kernel expansions in loglinear time. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, volume 85, 2013.

[25] Jiyan Yang, Vikas Sindhwani, Haim Avron, and Michael Mahoney. Quasi-monte carlo feature maps for shift-invariant kernels. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 485–493, 2014.

[26] Felix Xinnan X Yu, Ananda Theertha Suresh, Krzysztof M Choromanski, Daniel N Holtmann-Rice, and Sanjiv Kumar. Orthogonal random features. In *Advances in Neural Information Processing Systems 29 (NIPS)*, pages 1975–1983, 2016.

[27] Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(21):1–38, 2017. 00054.

[28] Ingo Steinwart, Don Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In *Proceedings of the 22nd Conference on Learning Theory (COLT 2009)*, pages 79–93, 2009.

[29] Steve Smale and Ding-Xuan Zhou. Estimating the approximation error in learning theory. *Analysis and Applications*, 1(01):17–41, 2003.

[30] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Verlag, New York, 2008.

[31] Andrea Caponnetto and Yuan Yao. Cross-validation based adaptation for regularization operators in learning theory. *Analysis and Applications*, 8(02):161–183, 2010.

[32] Catalin Ionescu, Alin Popa, and Cristian Sminchisescu. Large-scale data-dependent kernel approximation. In *Artificial Intelligence and Statistics*, pages 19–27, 2017.

[33] Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bulletin of the American mathematical society*, 39(1):1–49, 2002.

[34] Haim Avron, Kenneth L Clarkson, and David P Woodruff. Faster kernel ridge regression using sketching and preconditioning. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1116–1138, 2017.

[35] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. Exploiting the circulant structure of tracking-by-detection with kernels. In *European conference on computer vision*, pages 702–715. Springer, 2012.

[36] Mu Li, David G Andersen, Jun Woo Park, Alexander J Smola, Amr Ahmed, Vanja Josifovski, James Long, Eugene J Shekita, and Bor-Yiing Su. Scaling distributed machine learning with the parameter server. In *Proceedings of the 11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14)*, volume 14, pages 583–598, 2014.

[37] Ananda Theertha Suresh, Felix X Yu, Sanjiv Kumar, and H Brendan McMahan. Distributed mean estimation with limited communication. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 3329–3337, 2017.

[38] Steve Smale and Ding-Xuan Zhou. Shannon sampling and function reconstruction from point values. *Bulletin of the American Mathematical Society*, 41(3):279–305, 2004.

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

Distributed Learning with Random features Supplementary Materials

We proves the main results based on traditional integral operator. The main novelties lie in : 1) Error decomposition for KRR with divide-and-conquer and random features, which indicates how additional unlabeled data and data-dependent random features effect errors of the *excess risk*. 2) In detailed proof, the norm of kernel space is replace by the norm of feature space, because estimators defined by random feature actually run in feature space.

We start with some useful definitions and rewrite estimators in closed form by integral operators. For the sake of simplification, the main process is based on $\widehat{f}_{D,\lambda}$ and *excess bound* of $\widehat{f}_{D^*,\lambda}$ is given in implicit bound in Theorem 4. Then, the error decomposition is derived and we use concentration inequalities bound the items in decompositions. Further, we propose an implicit *excess risk* bound in Theorem 4 defined by effective dimension $\mathcal{N}(\lambda)$ and maximum random feature dimension $\mathcal{F}_\infty$. Combining Assumptions 4 and 6, Theorem 3 is proved. Finally, other theorems are proved as special cases of Theorem 3.

A Preliminary definitions

In this section, we provide the notation, recall some useful facts and define some operators used in the rest of the appendix, part of which are given in [13]. In the rest of the paper we denote with $\|\cdot\|$ the operatorial norm and with $\|\cdot\|_{HS}$ the Hilbert-Schmidt norm. Let $\mathcal{L}$ be a Hilbert space, we denote with $\langle\cdot,\cdot\rangle_{\mathcal{L}}$ the associated inner product, with $\|\cdot\|_{\mathcal{L}}$ the norm and with $\mathrm{Tr}(\cdot)$ the trace. Let Q be a bounded self-adjoint linear operator on a separable Hilbert space $\mathcal{L}$, we denote with $\lambda_{\max}(Q)$ the biggest eigenvalue of Q, that is $\lambda_{\max}(Q) = \sup_{\|f\|_{\mathcal{L}}\leq 1} \langle f, Qf\rangle_{\mathcal{L}}$.

Definition 3. *For all $g \in L^2(X,\rho_X)$, $\beta \in \mathbb{R}^M$, $\alpha \in \mathbb{R}^n$ and for j-th subset D_j, we have*

- $S_M : \mathbb{R}^M \to L^2(X,\rho_X), \quad (S_M\beta)(\cdot) = \phi_M(\cdot)^\top\beta,$
- $S_M^* : L^2(X,\rho_X) \to \mathbb{R}^M, (S_M^* g)_i = \frac{1}{\sqrt{M}}\int_X \psi_{\omega_i}(x)g(x)d\rho_X(x)$, *where* $i \in \{1,\dots,M\}$,
- $L_M : L^2(X,\rho_X) \to L^2(X,\rho_X), \quad (L_M g)(\cdot) = \int_X K_M(\cdot,z)g(z)d\rho_X(z).$
- $\widehat{L}_M : L^2(X,\rho_X) \to L^2(X,\rho_X), \quad (\widehat{L}_M g)(\cdot) = \frac{1}{n}\sum_{i=1}^n K_M(\cdot,z)g(z).$
- $C_M : \mathbb{R}^M \to \mathbb{R}^M, \quad C_M = \int_X \phi_M(x)\phi_M(x)^\top d\rho_X(x),$
- $\widehat{C}_M : \mathbb{R}^M \to \mathbb{R}^M, \quad \widehat{C}_M = \frac{1}{n}\sum_{i=1}^n \phi_M(x_i)\phi_M(x_i)^\top.$

For any $\lambda > 0$ define the effective dimension $\mathcal{N}_M(\lambda)$ induced by the kernel K_M as follows,

$$\mathcal{N}_M(\lambda) = \mathrm{Tr}\big((L_M + \lambda I)^{-1}L_M\big).$$

Remark 3. *Under Assumption 1 the linear operators L is trace class and $L_M, C_M, S_M, \widehat{C}_M, \widehat{S}_M$ are finite dimensional. Moreover we have that $L = SS^*$, $L_M = S_M S_M^*$, $C_M = S_M^* S_M$ and $\widehat{C}_M = \widehat{S}_M^\top \widehat{S}_M$. Finally $L, L_M, C_M, \widehat{C}_M$ are self-adjoint and positive operators, with spectrum is $[0,\kappa^2]$. Moreover, we denote with Q_λ the operator $Q + \lambda I$, where Q is a linear operator, $\lambda \in \mathbb{R}$ and I the identity operator, so for example $\widehat{C}_{M,\lambda} := \widehat{C}_M + \lambda I$.*

Definition 4. *Let $f_\rho : \mathcal{X} \to \mathbb{R}$ be the regression function of ρ defined by*

$$f_\rho(\mathbf{x}) = \int_{\mathcal{Y}} y d\rho(y|\mathbf{x}).$$

where $\rho(\cdot|\mathbf{x})$ is the conditional distribution of ρ at $\mathbf{x} \in \mathcal{X}$. Note that $f_\rho(\mathbf{x})$ can be seen as the noise-free label of $\mathbf{x}$.

Remark 4. *Let $P : L^2(X,\rho_X) \to L^2(X,\rho_X)$ be the projection operator, ranging the closure of L. Under Assumptions 2, there holds [33]*

$$Pf_\rho = Sf_{\mathcal{H}}.$$

Then, Assumption 5 is equivalent to

$$Pf_\rho = L^r g, \tag{18}$$

where $r \in [1/2, 1], g \in L^2(X, \rho_X)$ *and* $R = \|f_\mathcal{H}\| = \|g\|_{L^2(X,\rho_X)}$.

Applying operators defined in Defintion. 3 and notations in Remark. 3 to estimators defined in Section 4, we can obtain the following equations on the subset D_j by traditional integral approach [10, 20]

$$\widehat{f}^M_{D_j,\lambda} = S_M \widehat{C}^{-1}_{M,\lambda} \widehat{S}^\top_M \widehat{y}, \tag{19}$$

$$\widetilde{f}^M_{D_j,\lambda} = S_M \widehat{C}^{-1}_{M,\lambda} S^*_M P f_\rho. \tag{20}$$

$$f^M_\lambda = L^{-1}_{M,\lambda} L_M P f_\rho, \tag{21}$$

$$f_\lambda = L^{-1}_\lambda L P f_\rho. \tag{22}$$

B Error decomposition

Applying the identity $\widehat{f}^M_{D,\lambda} - f_\mathcal{H} = \widehat{f}^M_{D,\lambda} - f^M_\lambda + f^M_\lambda - f_\lambda + f_\lambda - f_\mathcal{H}$, to the *excess risk* (11) and $(a+b+c)^2 \le 3a^2 + 3b^2 + 3c^2$, we have

$$\mathbb{E}\big[\mathcal{E}(\widehat{f}^M_{D,\lambda}) - \mathcal{E}(f_\mathcal{H})\big] \le 3\,\mathbb{E}\|\widehat{f}^M_{D,\lambda} - f^M_\lambda\|^2_\rho + 3\,\mathbb{E}\|f^M_\lambda - f_\lambda\|^2_\rho + 3\,\mathbb{E}\|f_\lambda - f_\mathcal{H}\|^2_\rho. \tag{23}$$

Note that the norm of $\widehat{f}^M_{D,\lambda} - f^M_\lambda$ contains variance, sample error and distributed error, which coincides to decompose it into three terms in the following Lemama 2. Consider that sample error consists two parts : label variance (noise data) and empirical learning (the difference between expected learning and empirical learning).

Lemma 2. *Let* $\widehat{f}^M_{D,\lambda}$ *be defined in Section 4, we have*

$$\begin{aligned}
&\mathbb{E}\|\widehat{f}^M_{D,\lambda} - f^M_\lambda\|^2_\rho \\
&\le \underbrace{\frac{1}{m^2}\sum_{j=1}^m \mathbb{E}\|\widehat{f}^M_{D_j,\lambda} - f^M_\lambda\|^2_\rho}_{\textit{Sample Error}} + \underbrace{\frac{1}{m}\sum_{j=1}^m \mathbb{E}\|\widetilde{f}^M_{D_j,\lambda} - f^M_\lambda\|^2_\rho}_{\textit{Distributed Error}}.
\end{aligned}$$

For further decomposition on sample error, there exists

$$\begin{aligned}
&\mathbb{E}\|\widehat{f}^M_{D,\lambda} - f^M_\lambda\|^2_\rho \\
&\le \underbrace{\frac{2}{m^2}\sum_{j=1}^m \mathbb{E}\|\widehat{f}^M_{D_j,\lambda} - \widetilde{f}^M_{D_j,\lambda}\|^2_\rho}_{\textit{Variance}} + \underbrace{\frac{2}{m^2}\sum_{j=1}^m \mathbb{E}\|\widetilde{f}^M_{D_j,\lambda} - f^M_\lambda\|^2_\rho}_{\textit{Empirical error}} + \underbrace{\frac{1}{m}\sum_{j=1}^m \mathbb{E}\|\widetilde{f}^M_{D_j,\lambda} - f^M_\lambda\|^2_\rho}_{\textit{Distributed Error}}.
\end{aligned}$$

Proof. Since

$$\begin{aligned}
\|\widehat{f}^M_{D,\lambda} - f^M_\lambda\|^2_\rho &= \left\| \frac{1}{m}\sum_{j=1}^m (\widehat{f}^M_{D_j,\lambda} - f^M_\lambda) \right\|^2_\rho \\
&= \frac{1}{m^2}\sum_{j=1}^m \|(\widehat{f}^M_{D_j,\lambda} - f^M_\lambda)\|^2_\rho + \frac{1}{m}\sum_{j=1}^m \left\langle \widehat{f}^M_{D_j,\lambda} - f^M_\lambda, \frac{1}{m}\sum_{k\ne j}(\widehat{f}^M_{D_k,\lambda} - f^M_\lambda) \right\rangle_\rho \\
&= \frac{1}{m^2}\sum_{j=1}^m \|(\widehat{f}^M_{D_j,\lambda} - f^M_\lambda)\|^2_\rho + \frac{1}{m}\sum_{j=1}^m \left\langle \widehat{f}^M_{D_j,\lambda} - f^M_\lambda, \widehat{f}^M_{D,\lambda} - f^M_\lambda - \frac{1}{m}(\widehat{f}^M_{D_j,\lambda} - f^M_\lambda) \right\rangle_\rho,
\end{aligned}$$

the expectation of $\|\widehat{f}_{D,\lambda}^{M} - f_{\lambda}^{M}\|_{\rho}^{2}$ becomes

$$\begin{aligned}
&\mathbb{E}\|\widehat{f}_{D,\lambda}^{M} - f_{\lambda}^{M}\|_{\rho}^{2} \\
=&\frac{1}{m^2}\sum_{j=1}^{m}\mathbb{E}\|(\widehat{f}_{D_j,\lambda}^{M} - f_{\lambda}^{M})\|_{\rho}^{2} + \frac{1}{m}\sum_{j=1}^{m}\left\langle \mathbb{E}[\widehat{f}_{D_j,\lambda}^{M}] - f_{\lambda}^{M}, \mathbb{E}[\widehat{f}_{D,\lambda}^{M}] - f_{\lambda}^{M} - \frac{1}{m}(\mathbb{E}[\widehat{f}_{D_j,\lambda}^{M}] - f_{\lambda}^{M})\right\rangle_{\rho}.
\end{aligned}$$

The second part equals

$$\begin{aligned}
&\frac{1}{m}\sum_{j=1}^{m}\left\langle \mathbb{E}[\widehat{f}_{D_j,\lambda}^{M}] - f_{\lambda}^{M}, \mathbb{E}[\widehat{f}_{D,\lambda}^{M}] - f_{\lambda}^{M}\right\rangle_{\rho} - \frac{1}{m}\sum_{j=1}^{m}\left\langle \mathbb{E}[\widehat{f}_{D_j,\lambda}^{M}] - f_{\lambda}^{M}, \frac{1}{m}(\mathbb{E}[\widehat{f}_{D_j,\lambda}^{M}] - f_{\lambda}^{M})\right\rangle_{\rho} \\
=&\|\mathbb{E}[\widehat{f}_{D,\lambda}^{M}] - f_{\lambda}^{M}\|_{\rho}^{2} - \frac{1}{m^2}\sum_{j=1}^{m}\|\mathbb{E}[\widehat{f}_{D_j,\lambda}^{M}] - f_{\lambda}^{M}\|_{\rho} \\
=&\left\|\frac{1}{m}\sum_{j=1}^{m}(\mathbb{E}[\widehat{f}_{D_j,\lambda}^{M}] - f_{\lambda}^{M})\right\|_{\rho}^{2} - \frac{1}{m^2}\sum_{j=1}^{m}\|\mathbb{E}[\widehat{f}_{D_j,\lambda}^{M}] - f_{\lambda}^{M}\|_{\rho}.
\end{aligned}$$

Due to Cauchy–Schwarz inequality, it holds

$$\left\|\frac{1}{m}\sum_{j=1}^{m}(\mathbb{E}[\widehat{f}_{D_j,\lambda}^{M}] - f_{\lambda}^{M})\right\|_{\rho}^{2} \leq \frac{1}{m}\sum_{j=1}^{m}\|(\mathbb{E}[\widehat{f}_{D_j,\lambda}^{M}] - f_{\lambda}^{M})\|_{\rho}^{2}. \tag{24}$$

According to Jensen's inequality, we have

$$\frac{1}{m}\sum_{j=1}^{m}\|(\mathbb{E}[\widehat{f}_{D_j,\lambda}^{M}] - f_{\lambda}^{M})\|_{\rho}^{2} \leq \frac{1}{m}\sum_{j=1}^{m}\mathbb{E}\|(\widetilde{f}_{D_j,\lambda}^{M} - f_{\lambda}^{M})\|_{\rho}^{2}.$$

Finally, combining the first part of (24), there holds

$$\mathbb{E}\|\widehat{f}_{D,\lambda}^{M} - f_{\lambda}^{M}\|_{\rho}^{2} \leq \frac{1}{m^2}\sum_{j=1}^{m}\mathbb{E}\|\widehat{f}_{D_j,\lambda}^{M} - f_{\lambda}^{M}\|_{\rho}^{2} + \frac{1}{m}\sum_{j=1}^{m}\mathbb{E}\|\widetilde{f}_{D_j,\lambda}^{M} - f_{\lambda}^{M}\|_{\rho}^{2}.$$

Then, we decompose $\|\widehat{f}_{D_j,\lambda}^{M} - f_{\lambda}^{M}\|_{\rho}^{2}$ as $\|\widehat{f}_{D_j,\lambda}^{M} - \widetilde{f}_{D_j,\lambda}^{M} + \widetilde{f}_{D_j,\lambda}^{M} - f_{\lambda}^{M}\|_{\rho}^{2}$ and the following holds according to $(a+b)^2 \leq 2a^2 + 2b^2$

$$\frac{1}{m^2}\sum_{j=1}^{m}\mathbb{E}\|\widehat{f}_{D_j,\lambda}^{M} - f_{\lambda}^{M}\|_{\rho}^{2} \leq \frac{2}{m^2}\sum_{j=1}^{m}\mathbb{E}\|\widehat{f}_{D_j,\lambda}^{M} - \widetilde{f}_{D_j,\lambda}^{M}\|_{\rho}^{2} + \frac{2}{m^2}\sum_{j=1}^{m}\mathbb{E}\|\widetilde{f}_{D_j,\lambda}^{M} - f_{\lambda}^{M}\|_{\rho}^{2}.$$

□

According to the above Lemma, the error decomposition in Lemma 1 can be easily proved. Compared with the sample error of a local estimator $\mathbb{E}\|\widehat{f}_{D_j^*,\lambda}^{M} - f_{\lambda}^{M}\|$, the sample error $\mathbb{E}\|\widehat{f}_{D^*,\lambda}^{M} - f_{\lambda}^{M}\|$ bounded by variance (13) and empirical error (14) has an additional $1/m$, demonstrating that distributed learning can reduce the sample error than local estimator. Moreover, the distributed error $\mathbb{E}\|\widetilde{f}_{D_j^*,\lambda}^{M} - f_{\lambda}^{M}\|_{\rho}^{2}$ in (15) focuses on noise-free data, therefore it is smaller than $\mathbb{E}\|\widehat{f}_{D_j^*,\lambda}^{M} - f_{\lambda}^{M}\|_{\rho}^{2}$. Then, the distributed error is possible to bounded in $\mathcal{O}(N^{-2r/(2r+\gamma)})$ with small m. But also, the distributed error can be reduce by unlabeled data, while the best convergence rate is hard to improve the number of partitions m can be reduced. Variance is dependent on labeled samples but also random feature error and approximation error are independent on dataset, so additional unlabeled data have no influence on those three kind of errors.

C Bound Terms

In this part, we combine the traditional integral operator approach [38, 11, 10] with a recently developed tool second order decomposition of operator inverses [7, 39] to propose an analytic result. There are four terms to bound $\mathbb{E}\|\widehat{f}^M_{D_j,\lambda} - \widetilde{f}^M_{D_j,\lambda}\|^2_\rho, \mathbb{E}\|\widetilde{f}^M_{D_j,\lambda} - f^M_\lambda\|^2_\rho, \|f^M_\lambda - f_\lambda\|^2_\rho$ and $\|f_\lambda - f_{\mathcal{H}}\|^2_\rho$.

Lemma 3. *Let $\delta \in (0, 1/2], N, M \in \mathbb{N}$ and $\lambda > 0$. Under Assumption 1, on the j-th local subset D_j the following holds with probability at least $1 - 2\delta$*

$$\mathbb{E}\|\widehat{f}^M_{D_j,\lambda} - \widetilde{f}^M_{D_j,\lambda}\|^2_\rho \leq c_0\Big(\frac{\mathcal{A}^2_{D_j,\lambda}}{\lambda} + 1\Big)^2 \mathcal{B}^2_{D_j,\lambda}$$

where $c_0 = 289(\kappa^2 + \kappa)^4 \log^6 \frac{2}{\delta}$ and

$$\mathcal{A}_{D_j,\lambda} = \frac{m}{N\sqrt{\lambda}} + \sqrt{\frac{m\mathcal{N}_M(\lambda)}{N}},$$
$$\mathcal{B}_{D_j,\lambda} = \frac{mB\kappa}{N\sqrt{\lambda}} + \sqrt{\frac{m\sigma^2\mathcal{N}_M(\lambda)}{N}}.$$

Proof. Let $\widehat{f}^M_{D_j,\lambda}$ and $\widetilde{f}^M_{D_j,\lambda}$ be defined as (19) and (20), we have

$$\begin{aligned}\|\widehat{f}^M_{D_j,\lambda} - \widetilde{f}^M_{D_j,\lambda}\| =& \|S_M\widehat{C}^{-1}_{M,\lambda}(\widehat{S}^\top_M\widehat{y} - S^*_M P f_\rho)\| \\ =& \|\big(S_M\widehat{C}^{-1}_{M,\lambda}C^{1/2}_{M,\lambda}\big)\big(C^{-1/2}_{M,\lambda}(\widehat{S}^\top_M\widehat{y} - S^*_M P f_\rho)\big)\| \\ \leq& \|S_M\widehat{C}^{-1}_{M,\lambda}C^{1/2}_{M,\lambda}\|\|C^{-1/2}_{M,\lambda}(\widehat{S}^\top_M\widehat{y} - S^*_M P f_\rho)\|.\end{aligned} \tag{25}$$

The last step is due to Cauchy–Schwarz inequality. For any X, T, bounded linear operators, with T positive, by multiplying and dividing for T_λ the following holds

$$\big\|XT\big\| \leq \big\|XT_\lambda\big\|\big\|T^{-1}_\lambda T\big\|,$$

and $\big\|T^{-1}_\lambda T\big\| \leq 1$, for any $\lambda > 0$. Thus we have $\|S_M\widehat{C}^{-1/2}_{M,\lambda}\| \leq \|C^{1/2}_{M,\lambda}\widehat{C}^{-1/2}_{M,\lambda}\|$, it holds

$$\|S_M\widehat{C}^{-1}_{M,\lambda}C^{1/2}_{M,\lambda}\| \leq \|C^{1/2}_{M,\lambda}\widehat{C}^{-1/2}_{M,\lambda}\|^2. \tag{26}$$

The estimate of $\|C^{1/2}_{M,\lambda}\widehat{C}^{-1/2}_{M,\lambda}\|^2$ was given in Lemma 9 of [20] or [12], that holds with probability at least $1 - \delta$

$$\begin{aligned}\|C^{1/2}_{M,\lambda}\widehat{C}^{-1/2}_{M,\lambda}\|^2 \leq& 8(\kappa^2 + \kappa)^2 \log^2\frac{2}{\delta}\,\frac{\mathcal{A}^2_{D_j,\lambda}}{\lambda} + 2 \\ \leq& 8.5(\kappa^2 + \kappa)^2 \log^2\frac{2}{\delta}\,\Big(\frac{\mathcal{A}^2_{D_j,\lambda}}{\lambda} + 1\Big),\end{aligned} \tag{27}$$

where the last step is due to $\kappa \in [1, \infty)$ defined in Assumption 1. Under Assumption 3, applying Bernstein inequality for vector-valued random variables as in Lemma 2 of [10] or Lemma 6 of [13], for a local subset D_j we have with probability at least $1 - \delta$

$$\|C^{-1/2}_{M,\lambda}(\widehat{S}^\top_M\widehat{y} - S^*_M P f_\rho)\| \leq 2\left(\frac{B\kappa}{n\sqrt{\lambda}} + \sqrt{\frac{\sigma^2\mathcal{N}_M(\lambda)}{n}}\right)\log\frac{2}{\delta}. \tag{28}$$

Combining the above results (26), (27) and (28) to (25), with $n = N/m$ we prove the lemma. □

Lemma 4. *Let $\delta \in (0, 1/2], N, M \in \mathbb{N}, 0 < \lambda \leq \frac{3}{4}\|L\|$ and $M \geq 32\big(\frac{\kappa^2}{\|L\| + \kappa^2}\big)\log\frac{2}{\delta}$. Under Assumptions 1, 2, 3 and 5 on the j-th local subset D_j the following holds with probability at least $1 - 2\delta$*

$$\mathbb{E}\|\widetilde{f}^M_{D_j,\lambda} - f^M_\lambda\|^2_\rho \leq c_1\left(\frac{\mathcal{A}^2_{D_j,\lambda}}{\lambda} + 1\right)^2 \mathcal{A}^2_{D_j,\lambda}\lambda^{2r-1}$$

where $c_1 = 676\kappa^4(\kappa^2 + \kappa)^6 R^2 \log^6\frac{2}{\delta}$ and

$$\mathcal{A}_{D_j,\lambda} = \frac{m}{N\sqrt{\lambda}} + \sqrt{\frac{m\mathcal{N}_M(\lambda)}{N}},$$

Proof. Under definitions in (20) and (21), using the identity $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$ for positive operators A, B, we have

$$\begin{aligned}
\|\widetilde{f}^M_{D_j,\lambda} - f^M_\lambda\| &= \|S_M \widehat{C}^{-1}_{M,\lambda} S^*_M P f_\rho - L^{-1}_{M,\lambda} L_M P f_\rho\| \\
&= \|\widehat{C}^{-1}_{M,\lambda}(\widehat{C}_M - L_M) P f_\rho + (\widehat{C}^{-1}_{M,\lambda} - L^{-1}_{M,\lambda}) L_M P f_\rho\| \\
&= \|\widehat{C}^{-1}_{M,\lambda}(\widehat{C}_M - L_M) P f_\rho + \widehat{C}^{-1}_{M,\lambda}(L_M - \widehat{C}_M) L^{-1}_{M,\lambda} L_M P f_\rho\| \\
&= \|\widehat{C}^{-1}_{M,\lambda}(\widehat{C}_M - L_M)(P f_\rho - f^M_\lambda)\| \\
&= \|\big(\widehat{C}^{-1}_{M,\lambda} C^{1/2}_{M,\lambda}\big)\big(C^{-1/2}_{M,\lambda}(\widehat{C}_M - L_M)\big)\big(P f_\rho - f^M_\lambda\big)\| \\
&= \|\widehat{C}^{-1}_{M,\lambda} C^{1/2}_{M,\lambda}\| \|C^{-1/2}_{M,\lambda}(\widehat{C}_M - L_M)\| \|P f_\rho - f^M_\lambda\|
\end{aligned} \tag{29}$$

Under Assumption 1 and $M \geq 32\big(\frac{\kappa^2}{\|L\|+\kappa^2}\big)\log\frac{2}{\delta}$, we apply Lemma 9 of [13], that holds $\|C_M\| \geq \frac{3}{4}\|L\|$. Thus, it holds that $\|C_M^{-1/2}\| \leq \frac{\sqrt{3}}{2}\sqrt{\|L\|} \leq \frac{\sqrt{3}}{2}\kappa$. For the first term, we have

$$\|\widehat{C}^{-1}_{M,\lambda} C^{1/2}_{M,\lambda}\| \leq \|C^{-1/2}_{M,\lambda}\| \|C^{1/2}_{M,\lambda}\widehat{C}^{-1/2}_{M,\lambda}\|^2 \leq \|C^{-1/2}_M\| \|C^{1/2}_{M,\lambda}\widehat{C}^{-1/2}_{M,\lambda}\|^2 \leq \frac{\sqrt{3}}{2}\kappa \|C^{1/2}_{M,\lambda}\widehat{C}^{-1/2}_{M,\lambda}\|^2.$$

As we known, $\|C^{1/2}_{M,\lambda}\widehat{C}^{-1/2}_{M,\lambda}\|^2$ is also used in Lemma 3, it was given in [12], thus we have with probability at least $1 - \delta$

$$\|\widehat{C}^{-1}_{M,\lambda} C^{1/2}_{M,\lambda}\| \leq \frac{\sqrt{3}}{2}\kappa \|C^{1/2}_{M,\lambda}\widehat{C}^{-1/2}_{M,\lambda}\|^2 \leq \frac{17\sqrt{3}}{4}\kappa(\kappa^2+\kappa)^2 \log^2\frac{2}{\delta}\left(\frac{\mathcal{A}^2_{D_j,\lambda}}{\lambda} + 1\right). \tag{30}$$

Using Bennett inequality, $\|C^{-1/2}_{M,\lambda}(\widehat{C}_M - L_M)\|$ is bounded in Lemma 7 of [13] with probability at least $1 - \delta$

$$\|C^{-1/2}_{M,\lambda}(\widehat{C}_M - L_M)\| \leq 2(2\kappa^2+\kappa)\log\frac{2}{\delta}\,\mathcal{A}_{D_j,\lambda} \leq 4(\kappa^2+\kappa)\log\frac{2}{\delta}\,\mathcal{A}_{D_j,\lambda}. \tag{31}$$

Under Assumptions 2 and 5, applying Lemma 8 of [20], there holds

$$\|P f_\rho - f^M_\lambda\| \leq \lambda^r \|g\|$$

Note that $\lambda^r = \lambda^{r-1/2}\lambda^{1/2} \leq \frac{\sqrt{3}}{2}\sqrt{\|L\|}\lambda^{r-1/2} \leq \frac{\sqrt{3}}{2}\kappa\lambda^{r-1/2}$ due to $\lambda \leq \frac{3}{4}\|L\|$. Meanwhile $R = \|g\|_{\rho_X}$ according to Remark. 4. Such that we have

$$\|P f_\rho - f^M_\lambda\| \leq \frac{\sqrt{3}}{2}\kappa\lambda^{r-1/2}R. \tag{32}$$

Combing (29), (30), (31) and (32), the proof is completed. □

The next Lemma bounds the distance between the Tikhonov solution with RF and the Tikhonov solution without RF, reflecting the approximation ability of random features.

Lemma 5. *Under Assumptions 1 and 2 for $\delta \in (0, 1/2]$ and $\lambda > 0$, when*

$$M \geq 4\kappa^2\left(\frac{\mathcal{N}(\lambda)}{\lambda}\right)^{2r-1}\left(\mathcal{F}_\infty \log\frac{11\kappa^2}{\lambda}\right)^{2-2r} \vee (4 + 18\mathcal{F}_\infty(\lambda))\log\frac{8\kappa^2}{\lambda\delta},$$

the following holds with probability at least $1 - 2\delta$

$$\|f^M_\lambda - f_\lambda\|^2_\rho \leq 16R^2\lambda^{2r},$$

where $t := \log\frac{11\kappa^2}{\lambda}$.

Proof. Combining Lemma 4 and Lemma 8 of [13], when $M \geq (4 + 18\mathcal{F}_\infty(\lambda))\log\frac{8\kappa^2}{\lambda\delta}$ there exists

$$\|f^M_\lambda - f_\lambda\| \leq 4\mathfrak{C}(\lambda, M), \tag{33}$$

where $\mathfrak{C}(\lambda, M) = R\kappa^{2r-1}\left(\frac{\sqrt{\lambda\mathcal{F}_\infty(\lambda)}\log\frac{2}{\delta}}{M^r} + \sqrt{\frac{\lambda\mathcal{N}(\lambda)^{2r-1}\mathcal{F}_\infty(\lambda)^{2-2r}\log\frac{2}{\delta}}{M}}\right)t^{1-r}$ and $t := \log\frac{11\kappa^2}{\lambda}$.
Proof details in Theorem 6 of [13] shows that under the condition

$$M \geq 4\kappa^2\lambda^{1-2v}\mathcal{N}(\lambda)^{2v-1}\mathcal{F}_\infty(\lambda)^{2-2v}t^{2-2r}, \tag{34}$$

we have

$$\mathfrak{C}(\lambda, M) \leq R\lambda^r. \tag{35}$$

Then, we complete the proof by applying by combining (33) and (35). □

The last term we need to estimate is approximation error $\|f_\lambda - f_\mathcal{H}\|_\rho^2$, which is standard [10, 11].

Lemma 6. *Under Assumption 1 the following holds for any $\lambda > 0$,*

$$\|f_\lambda - f_\mathcal{H}\|_\rho^2 \leq R^2\lambda^{2r}.$$

Proof. Using Remark. 3, we have $Pf_\rho = L^r g$. By the identity $A(A+\lambda I)^{-1} = I - \lambda(A+\lambda)^{-1}$ for $\lambda > 0$, there holds $LL_\lambda^{-1}Pf_\rho - Pf_\rho = (I - LL_\lambda^{-1})Pf_\rho$. And then by definitions in (21) and (22)

$$\begin{aligned}\|f_\lambda - f_\mathcal{H}\| =& \left\|LL_\lambda^{-1}Pf_\rho - Pf_\rho\right\| = \|\lambda L_\lambda^{-1}Pf_\rho = \lambda L_\lambda^{-1}L^r g\| \\ =& \|\lambda^r(\lambda^{1-r}L_\lambda^{-(1-r)})(L_\lambda^{-r}L^r)g\| \\ \leq& \|\lambda^r\|\|\lambda^{1-r}L_\lambda^{-(1-r)}\|\|L_\lambda^{-r}L^r\|\|g\|\end{aligned}$$

Note that $\left\|\lambda^{1-r}L_\lambda^{-(1-r)}\right\| \leq 1$ and $\left\|L_\lambda^{-r}L^r\right\| \leq 1$, while $R := \left\|g\right\|_{\rho_X}$ according to Remark. 3. The proof is completed. □

D Proofs of Main Results

Theorem 4 (Implicit *excess risk* bound)**.** *Let $\delta \in (0, 1]$ and $\widehat{f}_{D_j^*,\lambda}^M$ be defined by (10). Under Assumptions 1, 2, 3 and 5, when $0 < \lambda \leq \frac{3}{4}\|L\|$ and*

$$M \geq 4\kappa^2\left(\frac{\mathcal{N}(\lambda)}{\lambda}\right)^{2r-1}\left(\mathcal{F}_\infty\log\frac{11\kappa^2}{\lambda}\right)^{2-2r} \vee (4 + 18\mathcal{F}_\infty(\lambda))\log\frac{48\kappa^2}{\lambda\delta}$$

then the following holds with probability at least $1-\delta$,

$$\mathbb{E}\left[\mathcal{E}(\widehat{f}_{D^*,\lambda}^M)\right] - \mathcal{E}(f_\mathcal{H}) \leq c_2\left[\left(\frac{\mathcal{A}_{D_j^*,\lambda}^2}{\lambda} + 1\right)^2\left(\frac{1}{m}\mathcal{B}_{D_j,\lambda}^2 + \mathcal{A}_{D_j^*,\lambda}^2\lambda^{2r-1}\right) + \lambda^{2r}\right],$$

where c_2 is a constant independent on m, N, N^* that

$$c_2 = 6(\kappa^2+\kappa)^4\left[289 + 677\kappa^4(\kappa^2+\kappa)^2R^2\right]\log^6\frac{12}{\delta},$$

and

$$\mathcal{A}_{D_j^*,\lambda} = \frac{m}{N^*\sqrt{\lambda}} + \sqrt{\frac{m\mathcal{N}_M(\lambda)}{N^*}}, \quad \mathcal{B}_{D_j,\lambda} = \frac{mB\kappa}{N\sqrt{\lambda}} + \sqrt{\frac{m\sigma^2\mathcal{N}_M(\lambda)}{N}}.$$

Proof. For SKRR-DC-RF (10), a similar error decomposition holds

$$\begin{aligned}&\mathbb{E}\left[\mathcal{E}(\widehat{f}_{D^*,\lambda}^M)\right] - \mathcal{E}(f_\mathcal{H}) \\ \leq& \frac{6}{m^2}\sum_{j=1}^m \mathbb{E}\|\widehat{f}_{D_j^*,\lambda}^M - \widetilde{f}_{D_j^*,\lambda}^M\|_\rho^2 \quad \text{(Variance)} \\ +& \frac{6}{m^2}\sum_{j=1}^m \mathbb{E}\|\widetilde{f}_{D_j^*,\lambda}^M - f_\lambda^M\|_\rho^2 \quad \text{(Empirical error)} \\ +& \frac{3}{m}\sum_{j=1}^m \mathbb{E}\|\widetilde{f}_{D_j^*,\lambda}^M - f_\lambda^M\|_\rho^2 \quad \text{(Distributed Error)} \\ +& 3\|f_\lambda^M - f_\lambda\|_\rho^2 \quad \text{(Random Feature Error)} \\ +& 3\|f_\lambda - f_\mathcal{H}\|_\rho^2 \quad \text{(Approximation Error)}.\end{aligned} \tag{36}$$

We can see that variance is dependent on labeled samples but also random feature error and approximation error are independent on dataset, so additional unlabeled data have no influence on those three kind of errors. However, unlabeled samples can reduce empirical error and distributed error because they are data dependent but output independent. For distributed learning, we usually have $m \geq 2$ such that the empirical error is smaller than distributed error.

Let $\tau = \delta/6, \tau \in (0, 1]$ and replace the probability value δ with τ, such that both Lemma 3, Lemma 4 and Lemma 5 hold with probability at least $1 - 2\tau$ with

$$c_0 = 289(\kappa^2 + \kappa)^4 \log^6 \frac{2}{\tau}, \quad c_1 = 676\kappa^4(\kappa^2 + \kappa)^6 R^2 \log^6 \frac{2}{\tau}$$
$$M \geq 4\kappa^2 \left(\frac{\mathcal{N}(\lambda)}{\lambda}\right)^{2r-1} \left(\mathcal{F}_\infty \log \frac{11\kappa^2}{\lambda}\right)^{2-2r} \vee (4 + 18\mathcal{F}_\infty(\lambda)) \log \frac{8\kappa^2}{\lambda\tau}.$$

Specifically, for estimates of $\mathbb{E}\|\widehat{f}^M_{D^*_j,\lambda} - \widetilde{f}^M_{D^*_j,\lambda}\|$ and $\mathbb{E}\|\widetilde{f}^M_{D^*_j,\lambda} - f^M_\lambda\|^2_\rho$, additional unlabeled samples have no influence on $\mathcal{A}_{D_j,\lambda}$ but $\mathcal{B}_{D_j,\lambda}$ is dependent on labels of dataset, which need to be replaced by

$$\mathcal{A}_{D^*_j,\lambda} = \frac{m}{N^*\sqrt{\lambda}} + \sqrt{\frac{m\mathcal{N}_M(\lambda)}{N^*}},$$

where N^* is the number of all examples including labeled and unlabeled ones. Combining error decomposition (36) and Lemma 3, 4, 5 and 6, we have

$$\begin{aligned}
&\mathbb{E}\big[\mathcal{E}(\widehat{f}^M_{D^*,\lambda})\big] - \mathcal{E}(f_\mathcal{H}) \\
\leq &\frac{6}{m} c_0 \left(\frac{\mathcal{A}^2_{D^*_j,\lambda}}{\lambda} + 1\right)^2 \mathcal{B}^2_{D_j,\lambda} + \left(\frac{6}{m} + 3\right) c_1 \left(\frac{\mathcal{A}^2_{D^*_j,\lambda}}{\lambda} + 1\right)^2 \mathcal{A}^2_{D^*_j,\lambda}\lambda^{2r-1} + 48R^2\lambda^{2r} + 3R^2\lambda^{2r} \\
\leq &6\left(\frac{\mathcal{A}^2_{D^*_j,\lambda}}{\lambda} + 1\right)^2 \Big(\frac{c_0}{m}\mathcal{B}^2_{D_j,\lambda} + c_1\mathcal{A}^2_{D^*_j,\lambda}\lambda^{2r-1}\Big) + 51R^2\lambda^{2r} \\
\leq &6\Big(c_0 + c_1 + 9R^2\Big) \left[\left(\frac{\mathcal{A}^2_{D^*_j,\lambda}}{\lambda} + 1\right)^2 \left(\frac{1}{m}\mathcal{B}^2_{D_j,\lambda} + \mathcal{A}^2_{D^*_j,\lambda}\lambda^{2r-1}\right) + \lambda^{2r}\right].
\end{aligned}$$

We only consider the case which has more than one partitions that is $m \geq 2$ for distributed learning. With at least $1 - \delta$ probability, we use

$$\begin{aligned}
c_2 = 6\big(c_0 + c_1 + 9R^2\big) =&6\Big[289(\kappa^2 + \kappa)^4 + 676\kappa^4(\kappa^2 + \kappa)^6 R^2 + 9R^2\Big] \log^6 \frac{12}{\delta} \\
\leq &6(\kappa^2 + \kappa)^4 \Big[289 + 677\kappa^4(\kappa^2 + \kappa)^2 R^2\Big] \log^6 \frac{12}{\delta}
\end{aligned}$$

and then complete the proof. □

Theorem 5 (General *excess risk* bound)**.** *Let $\delta \in (0, 1]$. Under Assumptions 1, 2,3, 4, 5, and*

$$n \geq \left(\frac{4}{3\|L\|}\right)^{2r+\gamma}, \quad \lambda = N^{-\frac{1}{2r+\gamma}},$$
$$M \geq c_3 N^{\frac{(2r-1)(\gamma-\alpha+1)+\alpha}{2r+\gamma}} \log \frac{56\kappa^2}{\lambda\delta},$$
$$m \leq \min\left\{N^{\frac{2r+2\gamma-1}{2r+\gamma}}, N^* N^{\frac{-\gamma-1}{2r+\gamma}}\right\},$$

with $c_3 = 4(\kappa^2 Q^{4r-2} F^{2-2r} + 4 + 18F)$, then the following holds with probability at least $1 - \delta$,

$$\mathbb{E}\big[\mathcal{E}(\widehat{f}^M_{D^*,\lambda})\big] - \mathcal{E}(f_\mathcal{H}) \leq c_4\, N^{-\frac{2r}{2r+\gamma}},$$

where $c_4 = 21(\kappa^2 + \kappa)^4 \Big[289 + 677\kappa^4(\kappa^2 + \kappa)^2 R^2\Big] \log^6 \frac{14}{\delta} + B^2\kappa^2 + \sigma^2 + 8\big(Q + 2\big)^6$.

Proof. By Proposition 10 of [13], there exists $\mathcal{N}_M(\lambda) \leq 1.5\mathcal{N}(\lambda)$ with probability at least $1-\delta$. Theorem 4 can be further write as with at least $1-\delta$

$$\begin{aligned}\mathbb{E}\big[\mathcal{E}(\widehat{f}^M_{D^*,\lambda})\big] - \mathcal{E}(f_{\mathcal{H}}) \leq & c_2 \left[\left(\frac{\mathcal{A}^2_{D_j^*,\lambda}}{\lambda}+1\right)^2 \left(\frac{1}{m}\mathcal{B}^2_{D_j,\lambda} + \mathcal{A}^2_{D_j^*,\lambda}\lambda^{2r-1}\right) + \lambda^{2r}\right], \\ \leq & c_5 \left[\left(\frac{\mathcal{A}^2_{D_j^*,\lambda}}{\lambda}+1\right)^2 \left(\frac{1}{m}\mathcal{A}^2_{D_j,\lambda} + \mathcal{A}^2_{D_j^*,\lambda}\lambda^{2r-1}\right) + \lambda^{2r}\right],\end{aligned}$$

with $c_5 = 21(\kappa^2+\kappa)^4\Big[289 + 677\kappa^4(\kappa^2+\kappa)^2R^2\Big]\log^6\frac{14}{\delta} + B^2\kappa^2 + \sigma^2$ and

$$\begin{aligned}&\mathcal{A}_{D_j,\lambda} = \frac{m}{N\sqrt{\lambda}} + \sqrt{\frac{m\mathcal{N}(\lambda)}{N}}, \quad \mathcal{A}_{D_j^*,\lambda} = \frac{m}{N^*\sqrt{\lambda}} + \sqrt{\frac{m\mathcal{N}(\lambda)}{N^*}} \\ &M \geq 4\kappa^2\left(\frac{\mathcal{N}(\lambda)}{\lambda}\right)^{2r-1}\left(\mathcal{F}_\infty \log\frac{11\kappa^2}{\lambda}\right)^{2-2r} \vee (4+18\mathcal{F}_\infty(\lambda))\log\frac{56\kappa^2}{\lambda\delta}.\end{aligned}$$

Let $\lambda = N^{-\frac{1}{2r+\gamma}}$, $|D_1^*| = \cdots = |D_m^*|$ and $|D_1| = \cdots = |D_m| = N/m$, under Assumption 4 with the fact $r+s \geq r \geq 1/2$ we have

$$\mathcal{A}_{D_j,\lambda} \leq mN^{-\frac{2r+\gamma-1/2}{2r+\gamma}} + Q\sqrt{m}N^{-\frac{r}{2r+\gamma}}, \tag{37}$$

$$\mathcal{A}_{D_j^*,\lambda} \leq \frac{mN^{\frac{1}{4r+2\gamma}}}{N^*} + \frac{Q\sqrt{m}N^{\frac{\gamma}{4r+2\gamma}}}{\sqrt{N^*}}. \tag{38}$$

Then with $m \leq \min\left\{N^{\frac{2r+2\gamma-1}{2r+\gamma}}, N^*N^{\frac{-\gamma-1}{2r+\gamma}}\right\}$, we have

$$\lambda^{-1/2}\mathcal{A}_{D_j^*,\lambda} \leq Q+1. \tag{39}$$

Combing (37), (38), (39) with Theorem 4, we have

$$\mathbb{E}\big[\mathcal{E}(\widehat{f}^M_{D^*,\lambda})\big] - \mathcal{E}(f_{\mathcal{H}}) \leq c_5N^{-\frac{2r}{2r+\gamma}} + 8c_5\big(Q+2\big)^6N^{-\frac{2r}{2r+\gamma}}.$$

Note that n need to satisfy the associated constraint with respect to λ that $\lambda \in (0, \frac{3}{4}\|L\|]$, such that we need $n \geq \left(\frac{4}{3\|L\|}\right)^{2r+\gamma}$. According to Assumptions 4 and 6, we have

$$\mathcal{N}(\lambda) \leq Q^2\lambda^{-\gamma}, \quad \mathcal{F}_\infty \leq F\lambda^{-\alpha}.$$

Combing them with

$$M \geq 4\kappa^2\left(\frac{\mathcal{N}(\lambda)}{\lambda}\right)^{2r-1}\left(\mathcal{F}_\infty \log\frac{11\kappa^2}{\lambda}\right)^{2-2r} \vee (4+18\mathcal{F}_\infty(\lambda))\log\frac{56\kappa^2}{\lambda\delta},$$

we get

$$M \geq c_3N^{\frac{(2r-1)(\gamma-\alpha+1)+\alpha}{2r+\gamma}}\log\frac{56\kappa^2}{\lambda\delta}$$

with $c_3 = 4(\kappa^2Q^{4r-2}F^{2-2r} + 4 + 18F)$.

□

Proof of Theorem 3 Theorem 5 is the detailed version of Theorem 3.

Proof of Theorem 2 This theorem is a special case of Theorem 5. The special case $F = \kappa^2$ and $\alpha = 1$ equals to the condition without Assumption 6. Setting $F = \kappa^2$ and $\alpha = 1$, we have

$$\begin{aligned}&n \geq \left(\frac{4}{3\|L\|}\right)^{2r+\gamma}, \quad \lambda = N^{-\frac{1}{2r+\gamma}}, \\ &M \geq c_6N^{\frac{(2r-1)\gamma+1}{2r+\gamma}}\log\frac{56\kappa^2}{\lambda\delta}, \\ &m \leq \min\left\{N^{\frac{2r+2\gamma-1}{2r+\gamma}}, N^*N^{\frac{-\gamma-1}{2r+\gamma}}\right\},\end{aligned}$$

with $c_6 = 4(\kappa^2 Q^{4r-2}\kappa^{4-4r} + 4 + 18\kappa^2)$, then the following holds with probability at least $1-\delta$,

$$\mathbb{E}\big[\mathcal{E}(\widehat{f}^{M}_{D^*,\lambda})\big] - \mathcal{E}(f_{\mathcal{H}}) \leq c_4\, N^{-\frac{2r}{2r+\gamma}},$$

where $c_4 = 21(\kappa^2+\kappa)^4\Big[289 + 677\kappa^4(\kappa^2+\kappa)^2R^2\Big]\log^6\frac{14}{\delta} + B^2\kappa^2 + \sigma^2 + 8\big(Q+2\big)^6$.

Proof of Corollary 1 Assumption 2 can be relaxed to $|y| \leq b, \forall b > 1$, then the assumption is satisfied with $\sigma = B = 2b$. Assumption 4 is always satisfied with $\gamma = 1$ and Assumption 5 is always satisfied with $r = 1/2$. Then, setting $\sigma = B = 2b$, $\gamma = 1, Q = \kappa$ and $r = 1/2$, we get the worst case, applying them to Theorem 2, we get error bounds in worst case

$$\begin{aligned} n &\geq \Big(\frac{4}{3\|L\|}\Big)^2, \quad \lambda = \frac{1}{\sqrt{N}}, \\ M &\geq c_6\sqrt{N}\log\frac{56\kappa^2}{\lambda\delta}, \\ m &\leq \min\Big\{N,\ \frac{N^*}{N}\Big\}, \end{aligned}$$

with $c_6 = 4(\kappa^4 + 4 + 18\kappa^2)$, then the following holds with probability at least $1-\delta$,

$$\mathbb{E}\big[\mathcal{E}(\widehat{f}^{M}_{D^*,\lambda})\big] - \mathcal{E}(f_{\mathcal{H}}) \leq c_4\,\frac{1}{\sqrt{N}},$$

where $c_4 = 21(\kappa^2+\kappa)^4\Big[289 + 677\kappa^4(\kappa^2+\kappa)^2R^2\Big]\log^6\frac{14}{\delta} + 4b^2\kappa^2 + 4b^2 + 8\big(\kappa+2\big)^6$.

Proof of Theorem 1 This theorem is a special case of Theorem 5 without unlabeled data. When there is no unlabeled samples available that is $N^* = N$, we have

$$\begin{aligned} n &\geq \Big(\frac{4}{3\|L\|}\Big)^{2r+\gamma}, \quad \lambda = N^{-\frac{1}{2r+\gamma}}, \\ M &\geq c_6 N^{\frac{(2r-1)\gamma+1}{2r+\gamma}}\log\frac{56\kappa^2}{\lambda\delta}, \\ m &\leq N^{\frac{2r-1}{2r+\gamma}}, \end{aligned}$$

with $c_6 = 4(\kappa^2 Q^{4r-2}\kappa^{4-4r} + 4 + 18\kappa^2)$, then the following holds with probability at least $1-\delta$,

$$\mathbb{E}\big[\mathcal{E}(\widehat{f}^{M}_{D^*,\lambda})\big] - \mathcal{E}(f_{\mathcal{H}}) \leq c_4\, N^{-\frac{2r}{2r+\gamma}},$$

where $c_4 = 21(\kappa^2+\kappa)^4\Big[289 + 677\kappa^4(\kappa^2+\kappa)^2R^2\Big]\log^6\frac{14}{\delta} + B^2\kappa^2 + \sigma^2 + 8\big(Q+2\big)^6$.

Proof of Lemma 1 Combing (23) and Lemma 2, we prove the result.