Distributed Learning with Random Features

Jian Li\textsuperscript{1,2}, Yong Liu\textsuperscript{1*}, Weiping Wang\textsuperscript{1}

\textsuperscript{1}Institute of Information Engineering, Chinese Academy of Sciences
\textsuperscript{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, $O(\sqrt{N})$ partitions and $O(\sqrt{N})$ random features are sufficient to achieve $O(1/N)$ 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 \cite{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 \cite{2}, random projections including Nyström \cite{3} and random features \cite{4} to overcome memory bottleneck and gradient methods, as well as stochastic and preconditioned extensions \cite{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) \cite{7,9}, achieving optimal learning rates by using integral operator techniques \cite{10} and using the effective dimension to control the capability of the hypothesis space \cite{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 \cite{2,12}, with random projections including random features \cite{13} and Nyström \cite{9} and with stochastic gradient descent (SGD) \cite{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 \cite{15} and the combination of random features and multi-pass SGD \cite{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
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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Partitions $m$</th>
<th>Random centers $M$</th>
<th>Space</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>KRR</td>
<td>$\mathcal{O}(1)$</td>
<td>$\mathcal{O}(N^{2r/(2r+\gamma)})$</td>
<td>$\mathcal{O}(N^2)$</td>
<td>$\mathcal{O}(N^{3r})$</td>
</tr>
<tr>
<td>KRR-Nyström</td>
<td>$\mathcal{O}(N^{2r/(2r+\gamma)})$</td>
<td>$\mathcal{O}(N^{(2r-1)/2r+\gamma})$</td>
<td>$\mathcal{O}(NM)$</td>
<td>$\mathcal{O}(NM^2)$</td>
</tr>
<tr>
<td>KRR-RF</td>
<td>$\mathcal{O}(N^{2r/(2r+\gamma)})$</td>
<td>$\mathcal{O}(N^{(2r-1)/2r+\gamma})$</td>
<td>$\mathcal{O}(NM)$</td>
<td>$\mathcal{O}(NM^2)$</td>
</tr>
<tr>
<td>KRR-DC</td>
<td>$\mathcal{O}(N^{2r/(2r+\gamma)})$</td>
<td>$\mathcal{O}(N^{2r/(2r+\gamma)})$</td>
<td>$\mathcal{O}(N\sqrt{m})$</td>
<td>$\mathcal{O}(N^3/m^3)$</td>
</tr>
<tr>
<td>Theorem 1 (1)</td>
<td>$\mathcal{O}(N^{2r/(2r+\gamma)})$</td>
<td>$\mathcal{O}(N^{2r/(2r+\gamma)})$</td>
<td>$\mathcal{O}(N\sqrt{m})$</td>
<td>$\mathcal{O}(N^3/m^3)$</td>
</tr>
<tr>
<td>Theorem 2 (1)</td>
<td>$\mathcal{O}(N^{2r/(2r+\gamma)})$</td>
<td>$\mathcal{O}(N^{2r/(2r+\gamma)})$</td>
<td>$\mathcal{O}(N\sqrt{m})$</td>
<td>$\mathcal{O}(N^3/m^3)$</td>
</tr>
<tr>
<td>Theorem 3 (2)</td>
<td>$\mathcal{O}(N^{2r/(2r+\gamma)})$</td>
<td>$\mathcal{O}(N^{2r/(2r+\gamma)})$</td>
<td>$\mathcal{O}(N\sqrt{m})$</td>
<td>$\mathcal{O}(N^3/m^3)$</td>
</tr>
</tbody>
</table>

Note: All listed methods achieve the optimal learning rate $\mathcal{O}(N^{-2r/(2r+\gamma)})$, where $N$ 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 1 and 2 are simplified with $N^* \leq N^{2r/(2r+\gamma)}$ 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 improves 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 [12] employees additional unlabeled data to alleviate the dilemma of $\mathcal{O}(1)$ partitions when $r = 1/2$. Theorem 3 [15] 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 $\rho$, where $\mathcal{X} = \mathbb{R}^d$ is the input space and $\mathcal{Y}$ is the output space. The training set $D = \{(x_i, y_i)\}_{i=1}^N$ is sampled identically and independently from $\mathcal{X} \times \mathcal{Y}$ with respect to $\rho$. 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

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

(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 $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$. Kernel ridge regression (KRR) can be state as
\[
\hat{f}_\lambda = \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{N} \sum_{i=1}^{N} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}.
\] (2)

With the represent theorem [23], the problem (2) exists a unique closed form solution
\[
\hat{f}_\lambda(x) = \sum_{i=1}^{N} \hat{\alpha}_i K(x_i, x), \quad \text{with} \quad \hat{\alpha}_i = (K_N + \lambda N I)^{-1} y_N,
\] (3)

where $\lambda > 0$, $y_N = (y_1, \cdots, y_N)$ and $K_N$ is the $N \times N$ kernel matrix with $K_N(i, j) = K(x_i, x_j)$. Although KRR has has optimal statistical properties [10, 11], it becomes unfeasible as sample size $n$ increases because of $O(N^2)$ memory to store kernel matrix and $O(N^3)$ 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, 14] 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 $\hat{f}_{D_j, \lambda}$ by the solution KRR (3)
\[
\hat{f}_{D_j, \lambda}(x) = \sum_{i=1}^{n} \hat{\alpha}_{ij} K(x_i, x), \quad \text{with} \quad \hat{\alpha}_{ij} = (K_n + \lambda n I)^{-1} y_n,
\] (4)

where $K_n$ is the empirical kernel matrix on subset $D_j$ and $y_n = (y_1, \cdots, y_n)$ on $D_j$. Finally, those local estimators are summarized to a central node and a global estimator $\hat{f}_{D, \lambda}$ is computed by weighted average
\[
\hat{f}_{D, \lambda} = \frac{1}{m} \sum_{j=1}^{m} \hat{f}_{D_j, \lambda}.
\] (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 $\phi_M : \mathbb{R}^d \rightarrow \mathbb{R}^M$
\[
K(x, x') \approx \langle \phi_M(x), \phi_M(x') \rangle.
\] (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
\[
K(x, x') = \int_{\Omega} \psi(x, \omega) \psi(x', \omega) d\pi(\omega), \quad \forall x, x' \in \mathcal{X},
\]
where $(\Omega, \pi)$ is a probability space and $\psi : \mathcal{X} \times \Omega \rightarrow \mathbb{R}$. Define analogous operators for the approximated kernel $K_M(x, x') = \phi_M(x)\phi_M(x')$ to approximate $K(x, x')$ in (6) with
\[
\phi_M(x) = \frac{1}{\sqrt{M}} (\psi(x, \omega_1), \cdots, \psi(x, \omega_M)),
\]
where $\omega_1, \cdots, \omega_M$ are sampled independently with respect to $\pi$. 

3
Using random features in (6), the approximate solution of a local estimator $\hat{f}_{D_j,\lambda}$ in (4) is

$$\hat{f}_{D_j,\lambda}^M(x) = \phi_M(x)^T \hat{w}_j, \quad \text{with} \quad \hat{w}_j = (S_M^T S_M + \lambda I)^{-1} S_M^T \hat{y}_n,$$

where $\lambda > 0$. Note that for $j$-th subset $D_j$, $\forall (x, y) \in D_j$, $S_M^T = \frac{1}{\sqrt{n}} (\phi_M(x_1), \cdots, \phi_M(x_n))$ and $\hat{y}_n = \frac{1}{\sqrt{n}} (y_1, \cdots, y_n)$.

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

$$\hat{f}_{D,\lambda}^M = \frac{1}{m} \sum_{j=1}^{m} \hat{f}_{D_j,\lambda}^M.$$  \hfill (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 $\hat{f}_{D,\lambda}^M$, the excess risk is defined as

$$\mathbb{E}[\mathcal{E}(\hat{f}_{D,\lambda}^M)] - \mathcal{E}(f_H).$$  \hfill (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 $\psi$ is continuous and there is a $\kappa \in [1, \infty)$, such that $|\psi(x, \omega)| \leq \kappa$, $\forall x \in \mathcal{X}, \omega \in \Omega$.

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

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

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 $p \in \mathbb{N}$,

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

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 $O(1/\sqrt{N})$ 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 $f_H$, respectively.

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

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

where $L^2(\mathcal{X}, \rho_X) = \{ f : \mathcal{X} \to \mathbb{R} \mid \| f \|_p^2 = \int |f(x)|^2 d\rho_X < \infty \}$, $K$ is the induced kernel and $\rho_X$ is the marginal distribution of $\rho$ on $\mathcal{X}$.
Assumption 5 controls the bias of the estimator and is commonly used in approximation theory [10]. The effective dimension is defined as

\[ N(\lambda) = \text{Tr} \left( (L + \lambda I)^{-1} L \right), \quad \lambda > 0. \]

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

\[ N(\lambda) \leq Q^2 \lambda^{-\gamma}. \]

Assumption 5 (Regularity assumption). Assume there exists \( r \in [1/2, 1] \) and \( g \in L^2(\mathcal{X}, \rho_\mathcal{X}) \) 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 \( \gamma \) 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 7, 2, 3, 4 and 5, for the following condition \( n \geq n_0, \lambda = N^{-\frac{1}{2r+c+1}} \), and the number of partitions \( m \) respectively corresponds to

\[ M \gtrless N^{\frac{2r+c+1}{2-2r-c-1}}, \quad m \lesssim N^{\frac{2r-c-1}{2r+c+1}}, \]

suffice to guarantee with high probability that

\[ \mathbb{E}[\mathcal{E}(\hat{f}^M_{D,\lambda})] - \mathcal{E}(f_\mathcal{H}) = O\left( N^{-\frac{2r}{2r+c+1}} \right). \]

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 \( m \lesssim N^{\frac{2r-c+1}{2-2r-c+1}} \), and KRR-RF [13] under the same restriction on the number of random features that \( M \gtrsim N^{\frac{2r+c+1}{2-2r-c-1}} \). 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 \( \gamma \), 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 \( O(1/N) \) can be achieved by \( O(\sqrt{N}) \) random features and \( O(\sqrt{N}) \) partitions. Note that a smaller RKHS (\( \gamma = 0 \)) provides optimal learning rates \( O(1/N) \) despite the value of \( r \) as shown in the right of Figure 1. Moreover, lower regularity (\( r = 1/2 \)) leads to \( O(1) \) 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 $N^* = N^{1 + \frac{1}{2m+c}}$. 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 3 and 2 shows that $O(\sqrt{N})$ random features and a constant number of local estimators can guarantee $O(1/\sqrt{N})$ learning rate. The number of local estimators in the worst case is $m = O(1)$, independent on sample size $N$, which is very restrictive in large scale settings. In our follow-up work, we employ additional unlabeled samples to relax the restriction $m \lesssim N^{\frac{1}{2m+c}}$, as done in [20, 81].

**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 $\{\tilde{D}_j\}_{j=1}^m$ are drawn identically and independently from the conditional distribution $\rho_X$ and are stored in local processors. Consider the merged dataset $D^*$ on the $j$-th processor,

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

with

$$x^*_i = \begin{cases} x_i, & \text{if } (x_i, y_i) \in D_j, \\ \tilde{x}_i, & \text{otherwise,} \end{cases} \quad \text{and} \quad y^*_i = \begin{cases} \frac{\left| D^*_j \right|}{D_j}\gamma_i, & \text{if } (x_i, y_i) \in D_j, \\ 0, & \text{otherwise.} \end{cases}$$

Let $D^* = \bigcup_{j=1}^m D^*_j$, $|D^*_1| = N^*$ 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

$$\tilde{f}_{D^*, \lambda}^M = \frac{1}{m} \sum_{j=1}^m \tilde{f}_{D^*_j, \lambda}^M. \quad (10)$$

According to error decomposition in Lemma 4 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 3, 4, and 5 if $n \geq n_0$, $\lambda = N^{-\frac{1}{2m+c}}$, and the number of random features $M$, the number of partitions in corresponds to

$$M \gtrsim N^{\frac{3r-1}{2r+1} + \frac{1}{2r+1}}, \quad m \leq \min \left\{ N^{\frac{2r+2r+1}{2r+1}}, N^* N^{\frac{1}{2m+c}} \right\}$$

then the following holds with high probability,

$$\mathbb{E}[\mathcal{E}(\tilde{f}_{D^*, \lambda}^M)] - \mathcal{E}(f_U) = O\left(N^{-\frac{2r}{2r+1}}\right).$$
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 \( N^* = N^{1+\frac{\gamma}{2}} \) 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 \( O(1) \) 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 7 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**

\[
M \gtrsim \sqrt{N}, \quad m \lesssim \min \left\{ N, \frac{N^*}{N} \right\}
\]

**is enough to guarantee with high probability,**

\[
E \left[ \mathcal{E}(\hat{f}^M_{D^*}) \right] - \mathcal{E}(f_H) = O \left( \frac{1}{\sqrt{N}} \right).
\]

The learning rate \( O(1/\sqrt{N}) \) 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 \( O(N^\beta) \) partitions and \( O(\sqrt{N}) \) 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**

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

**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** \( n \geq n_0, \lambda = N^{-1/\gamma} \), **and the number of random features \( M \), the number of partitions \( m \) corresponds to**

\[
M \gtrsim N \frac{(2r-1)(\gamma-\alpha+1)+\alpha}{2r+\gamma}, \quad m \lesssim \min \left\{ N^{\frac{2r+2}{2r+\gamma}-1}, N^* N^\frac{2-\gamma}{2r+\gamma} \right\}
\]

**then the following holds with high probability,**

\[
E \left[ \mathcal{E}(\hat{f}^M_{D^*}) \right] - \mathcal{E}(f_H) = O \left( N^{-\frac{2r}{2r+\gamma}} \right).
\]
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 (α = 1) and in data-dependent way (α = γ). 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 $O(1/N)$ learning rate when $r = 1/2$ and $γ = 0$. 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 $\mathbb{E}[\mathcal{E}(\hat{f}_D^M,\lambda)] - \mathbb{E}(f_H)$ 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 $\hat{f}_D^M,\lambda$ in (10) in primal form and denote other useful estimators as follows

$$\hat{f}_D^M,\lambda = \frac{1}{m} \sum_{j=1}^{m} \langle \tilde{w}_j, \phi_M(\cdot) \rangle, \quad \tilde{w}_j = \arg \min_{w \in \mathbb{R}^M} \left\{ \frac{1}{n^*} \sum_{i=1}^{n^*} ((w, \phi_M(x_i^*)) - y_i^*)^2 + \lambda \|w\|^2 \right\}$$

$$\tilde{f}_D^M,\lambda = \frac{1}{m} \sum_{j=1}^{m} \langle \tilde{w}_j, \phi_M(\cdot) \rangle, \quad \tilde{w}_j = \arg \min_{w \in \mathbb{R}^M} \left\{ \frac{1}{n^*} \sum_{i=1}^{n^*} ((w, \phi_M(x_i^*)) - f_H(x_i^*))^2 + \lambda \|w\|^2 \right\}$$

$$f_M^\lambda = \langle \tilde{u}, \phi_M(\cdot) \rangle, \quad \tilde{u} = \arg \min_{u \in \mathbb{R}^M} \int_{\mathcal{X}} ((u, \phi_M(x)) - f_H(x))^2 d\rho_X(x) + \lambda \|u\|^2,$$

$$f_\lambda = \langle \tilde{u}, \phi(\cdot) \rangle, \quad \tilde{u} = \arg \min_{u \in \mathcal{H}_K} \int_{\mathcal{X}} ((u, \phi(x)) - f_H(x))^2 d\rho_X(x) + \lambda \|u\|^2,$$

where $\phi : \mathbb{R}^d \rightarrow \mathcal{H}_K$ is feature map associated to the kernel $K$ by $K(x,x') = \langle \phi(x), \phi(x') \rangle$. The empirical estimator $\hat{f}_D^M,\lambda$ focuses on noise-free data. The last two vectors are both expected estimators defined by random features $\phi_M$ and implicit feature map $\phi$. From [1,33], there holds

$$\mathbb{E}[\mathcal{E}(\hat{f}_D^M,\lambda)] - \mathbb{E}(f_H) = \mathbb{E}\|\hat{f}_D^M,\lambda - f_H\|_\rho^2.$$ (11)

Combining (11) and the identity $\hat{f}_D^M,\lambda - f_H = \tilde{f}_D^M,\lambda - f_M^\lambda + f_M^\lambda - f_\lambda + f_H - f_H$, we obtain the error decomposition in Lemma 1 and its proof is provided in appendix.

**Lemma 1.** Let $\hat{f}_D^M,\lambda, \tilde{f}_D^M,\lambda, f_M^\lambda$ and $f_\lambda$ be defined as the above, we have

$$\mathbb{E}[\mathcal{E}(\hat{f}_D^M,\lambda)] - \mathbb{E}(f_H)$$

$$\leq \frac{6}{m^2} \sum_{j=1}^{m} \mathbb{E}\|\tilde{f}_D^M,\lambda - \hat{f}_D^M,\lambda\|_\rho^2$$ (Variance) (13)

$$+ \frac{6}{m^2} \sum_{j=1}^{m} \mathbb{E}\|\tilde{f}_D^M,\lambda - f_M^\lambda\|_\rho^2$$ (Empirical error) (14)

$$+ \frac{3}{m} \sum_{j=1}^{m} \mathbb{E}\|\tilde{f}_D^M,\lambda - f_\lambda\|_\rho^2$$ (Distributed Error) (15)

$$+ 6 \|f_M^\lambda - f_\lambda\|_\rho^2$$ (Random Features Error) (16)

$$+ 3 \|f_M^\lambda\|_\rho^2$$ (Approximation Error). (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
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 $2.5 \times 10^5$ data points on binary classification datasets covtype, SUSY and HIGGS where $\sqrt{N} = 500$. We use random Fourier features to approximate Gaussian kernel $K(x, x’) = \exp^{-\|x-x’\|^2/2\sigma^2}$. Random fourier features are in the form $\psi(x, \omega) = \cos(\omega^Tx+b)$, where $\omega$ is drawn from the normal distribution and $b$ is drawn from uniform distribution $[0, 2\pi]$. In the following experiments, we tune parameters $\sigma$ and $\lambda$ 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 lose too much accuracy. Then, we study empirical performance in terms of different numbers of random features. Results 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.

---

*https://archive.ics.uci.edu/ml/datasets/covertype
*https://archive.ics.uci.edu/ml/datasets/susy
*https://archive.ics.uci.edu/ml/datasets/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


Distributed Learning with Random features

Supplementary Materials

We prove 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 affect 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 \( f_{D, \lambda} \) and excess bound of \( f_{D^*, \lambda} \) 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 \( N(\lambda) \) and maximum random feature dimension \( F_\infty \). Combining Assumptions 4 and 6, Theorem 3 is proved. Finally, other theorems are proved as special cases of Theorem 5.

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 \( L \) be a Hilbert space, we denote with \( \langle \cdot, \cdot \rangle_L \) the associated inner product, with \(|\cdot|_L \) the norm and with \( \text{Tr}(\cdot) \) the trace. Let \( Q \) be a bounded self-adjoint linear operator on a separable Hilbert space \( L \), we denote with \( \lambda_{\max}(Q) \) the biggest eigenvalue of \( Q \), that is \( \lambda_{\max}(Q) = \sup \| f \|_{L} \leq 1 \langle f, Qf \rangle_L \).

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

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

For any \( \lambda > 0 \) define the effective dimension \( N_M(\lambda) \) induced by the kernel \( K_M \) as follows,

\[
N_M(\lambda) = \text{Tr}((L_M + \lambda I)^{-1}L_M).
\]

Remark 3. Under Assumption 7, the linear operators \( L \) is trace class and \( L_M, C_M, S_M, \hat{C}_M, \hat{S}_M \) are finite dimensional. Moreover we have that \( L = SS^* \), \( L_M = S_M S_M^* \), \( C_M = S_M^* S_M \), \( \hat{C}_M = \hat{S}_M^* \hat{S}_M \). Finally, \( L_M, C_M, \hat{C}_M \) are self-adjoint and positive operators, with spectrum is \([0, \kappa^2]\). Moreover, we denote with \( Q_\lambda \) the operator \( Q + \lambda I \), where \( Q \) is a linear operator, \( \lambda \in \mathbb{R} \) and \( I \) the identity operator, so for example \( C_{M, \lambda} := \hat{C}_M + \lambda I \).

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

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

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

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 [13]

\[
P f_\rho = S f_H.
\]
Then, Assumption 5 is equivalent to

\[ Pf_\rho = L^T g, \]  

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

Applying operators defined in Definition 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:

\[
\begin{align*}
\tilde{f}_{D_j, \lambda}^M &= S_M \hat{C}_{M, \lambda}^{-1} \hat{g}, \\
\hat{f}_{D_j, \lambda}^M &= S_M \hat{C}_{M, \lambda}^{-1} \hat{S}_M P_{f_\rho}, \\
f_\lambda^M &= L_{1, M}^{-1} L_M P_{f_\rho}, \\
f_\lambda &= L_{1, -1} L_P f_\rho.
\end{align*}
\]

### B Error decomposition

Applying the identity \( \tilde{f}_{D, \lambda}^M - f_H = \hat{f}_{D_j, \lambda}^M - f_\lambda^M + f_\lambda^M - f_\lambda + f_\lambda - f_H \), to the excess risk (11) and \( (a + b + c)^2 \leq 3a^2 + 3b^2 + 3c^2 \), we have

\[ \mathbb{E}[\mathcal{E}(\hat{f}_{D_j, \lambda}^M) - \mathcal{E}(f_H)] \leq 3 \mathbb{E}\|\tilde{f}_{D_j, \lambda}^M - f_\lambda^M\|_\rho^2 + 3 \mathbb{E}\|f_\lambda^M - f_\lambda\|_\rho^2 + 3 \mathbb{E}\|f_\lambda - f_H\|_\rho^2. \]  

(23)

Note that the norm of \( \tilde{f}_{D_j, \lambda}^M - f_\lambda^M \) contains variance, sample error and distributed error, which coincides to decompose it into three terms in the following Lemma 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 \( \tilde{f}_{D_j, \lambda}^M \) be defined in Section 4 we have

\[
\mathbb{E}\|\tilde{f}_{D_j, \lambda}^M - f_\lambda^M\|_\rho^2 \leq \frac{1}{m^2} \sum_{j=1}^{m} \mathbb{E}\|\tilde{f}_{D_j, \lambda}^M - f_\lambda^M\|_\rho^2 + \frac{1}{m} \sum_{j=1}^{m} \mathbb{E}\|\hat{f}_{D_j, \lambda}^M - f_\lambda^M\|_\rho^2.
\]

For further decomposition on sample error, there exists

\[
\mathbb{E}\|\tilde{f}_{D_j, \lambda}^M - f_\lambda^M\|_\rho^2 \leq \frac{2}{m^2} \sum_{j=1}^{m} \mathbb{E}\|\tilde{f}_{D_j, \lambda}^M - \hat{f}_{D_j, \lambda}^M\|_\rho^2 + \frac{2}{m} \sum_{j=1}^{m} \mathbb{E}\|\hat{f}_{D_j, \lambda}^M - f_\lambda^M\|_\rho^2 + \frac{1}{m} \sum_{j=1}^{m} \mathbb{E}\|\hat{f}_{D_j, \lambda}^M - f_\lambda^M\|_\rho^2.
\]

**Proof.** Since

\[
\|\tilde{f}_{D_j, \lambda}^M - f_\lambda^M\|_\rho^2 = \left\| \frac{1}{m} \sum_{j=1}^{m} (\tilde{f}_{D_j, \lambda}^M - f_\lambda^M) \right\|_\rho^2
\]

\[
= \frac{1}{m^2} \sum_{j=1}^{m} \| (\tilde{f}_{D_j, \lambda}^M - f_\lambda^M) \|_\rho^2 + \frac{1}{m} \sum_{j=1}^{m} \left( \tilde{f}_{D_j, \lambda}^M - f_\lambda^M, \frac{1}{m} \sum_{k \neq j} (\tilde{f}_{D_k, \lambda}^M - f_\lambda^M) \right)_\rho
\]

\[
= \frac{1}{m^2} \sum_{j=1}^{m} \| (\tilde{f}_{D_j, \lambda}^M - f_\lambda^M) \|_\rho^2 + \frac{1}{m} \sum_{j=1}^{m} \left( \tilde{f}_{D_j, \lambda}^M - f_\lambda^M, \frac{1}{m} (\tilde{f}_{D_j, \lambda}^M - f_\lambda^M) \right)_\rho,
\]

14
the expectation of $\| \hat{f}_D^M - f_\lambda^M \|^2_\rho$ becomes

$$E[\| \hat{f}_D^M - f_\lambda^M \|^2_\rho] = \frac{1}{m^2} \sum_{j=1}^{m} E[\| \hat{f}_D^M - f_\lambda^M \|^2_\rho] + \frac{1}{m} \sum_{j=1}^{m} \left( E[\hat{f}_D^M] - f_\lambda^M, E[\hat{f}_D^M] - f_\lambda^M - \frac{1}{m} (E[\hat{f}_D^M] - f_\lambda^M) \right)_\rho.$$

The second part equals

$$\frac{1}{m} \sum_{j=1}^{m} \left( E[\hat{f}_D^M] - f_\lambda^M, E[\hat{f}_D^M] - f_\lambda^M \right)_\rho = \left( E[\hat{f}_D^M] - f_\lambda^M \right)_\rho = \frac{1}{m^2} \sum_{j=1}^{m} \|E[\hat{f}_D^M] - f_\lambda^M\|_\rho$$

$$= \left( \frac{1}{m} \sum_{j=1}^{m} (E[\hat{f}_D^M] - f_\lambda^M) \right)_\rho^2 - \frac{1}{m^2} \sum_{j=1}^{m} \|E[\hat{f}_D^M] - f_\lambda^M\|_\rho^2.$$}

Due to Cauchy–Schwarz inequality, it holds

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

According to Jensen’s inequality, we have

$$\frac{1}{m} \sum_{j=1}^{m} \|E[\hat{f}_D^M] - f_\lambda^M\|_\rho^2 \leq \frac{1}{m} \sum_{j=1}^{m} \|E[\hat{f}_D^M] - f_\lambda^M\|_\rho^2.$$}

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

$$E[\| \hat{f}_D^M - f_\lambda^M \|^2_\rho] \leq \frac{1}{m^2} \sum_{j=1}^{m} E[\| \hat{f}_D^M - f_\lambda^M \|^2_\rho] + \frac{1}{m} \sum_{j=1}^{m} E[\| \hat{f}_D^M - f_\lambda^M \|^2_\rho].$$

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

$$\frac{1}{m^2} \sum_{j=1}^{m} E[\| \hat{f}_D^M - f_\lambda^M \|^2_\rho] \leq \frac{2}{m^2} \sum_{j=1}^{m} E[\| \hat{f}_D^M - f_\lambda^M \|^2_\rho] + \frac{2}{m^2} \sum_{j=1}^{m} E[\| \hat{f}_D^M - f_\lambda^M \|^2_\rho].$$

According to the above Lemma, the error decomposition in Lemma[1] can be easily proved. Compared with the sample error of a local estimator $E[\| \hat{f}_D^M - f_\lambda^M \|]$, the sample error $E[\| \hat{f}_D^M - f_\lambda^M \|$ 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 $E[\| \hat{f}_D^M - f_\lambda^M \|^2_\rho]$ in (15) focuses on noise-free data, therefore it is smaller than $E[\| \hat{f}_D^M - f_\lambda^M \|^2_\rho].$ Then, the distributed error is possible to bounded in $O(N^{-2r/(2r+\gamma)})$ 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 \cite{38,11,10} with a recently developed tool second order decomposition of operator inverses \cite{7,39} to propose an analytic result. There are four terms to bound $E\|\hat{f}^M_{D_j,\lambda} - \hat{f}^M_{D_j,\lambda}\|^2$ for a local subset $D_j$ of the $j$-th local subset.

Lemma 3. Let $\delta \in (0, 1/2)$, $N, M \in \mathbb{N}$ and $\lambda > 0$. Under Assumption 7 on the $j$-th local subset $D_j$ the following holds with probability at least $1 - 2\delta$

\[
E\|\hat{f}^M_{D_j,\lambda} - \hat{f}^M_{D_j,\lambda}\|^2 \leq c_0 \left( \frac{A_{D_j,\lambda}}{\lambda} + 1 \right)^2 B^2_{D_j,\lambda}
\]

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

\[
A_{D_j,\lambda} = \frac{m}{N \sqrt{\lambda}} + \frac{mN_M(\lambda)}{N},
\]

\[
B_{D_j,\lambda} = \frac{mB\kappa}{N \sqrt{\lambda}} + \frac{m\sigma^2N_M(\lambda)}{N}.
\]

Proof. Let $\hat{f}^M_{D_j,\lambda}$ and $\hat{f}^M_{D_j,\lambda}$ be defined as \cite{19} and \cite{20}, we have

\[
\|\hat{f}^M_{D_j,\lambda} - \hat{f}^M_{D_j,\lambda}\| = \|S_M \hat{C}^{-1\,2}_{M,\lambda}(\hat{S}^T_M \hat{y} - S^*_M Pf_p)\| \\
\leq \|S_M \hat{C}^{-1\,2}_{M,\lambda}\| \|C^{-1\,2}_{M,\lambda}(\hat{S}^T_M \hat{y} - S^*_M Pf_p)\| \\
\leq \|S_M \hat{C}^{-1\,2}_{M,\lambda}\| \|C^{-1\,2}_{M,\lambda}(\hat{S}^T_M \hat{y} - S^*_M Pf_p)\| \\
\leq 2\text{log}\left( \frac{b}{m\sqrt{\lambda}} + \sqrt{\frac{\sigma^2N_M(\lambda)}{n}} \right) \log \frac{2}{\delta}.
\]

Combining the above results and \cite{26,27} and \cite{28} to \cite{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 $M \geq 32\left( \frac{\kappa^2 + \kappa}{\|f\| + \kappa} \right) \log \frac{2}{\delta}$. Under Assumptions 7, 8 and 9 on the $j$-th local subset $D_j$ the following holds with probability at least $1 - 2\delta$

\[
E\|f^M_{D_j,\lambda} - f^M_{\lambda}\|^2 \leq c_1 \left( \frac{A^2_{D_j,\lambda}}{\lambda} + 1 \right)^2 A^2_{D_j,\lambda} \lambda^{2r-1}
\]

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

\[
A_{D_j,\lambda} = \frac{m}{N \sqrt{\lambda}} + \sqrt{\frac{mN_M(\lambda)}{N}},
\]
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

\[
\|\tilde{f}_M^{M} - f_X^M\| = \|S_M^{M}\tilde{C}_M^{-1}S_M^{M}Pf\rho - L_{M,\lambda}L_M Pf\rho\| \\
= \|\tilde{C}_M^{-1}(\tilde{C}_M - L_M)Pf\rho + (\tilde{C}_M^{-1} - L_{M,\lambda})L_M Pf\rho\| \\
= \|\tilde{C}_M^{-1}(\tilde{C}_M - L_M)Pf\rho + \tilde{C}_M^{-1}(L_M - \tilde{C}_M)L_{M,\lambda}L_M Pf\rho\| \\
= \|\tilde{C}_M^{-1}(\tilde{C}_M - L_M)(Pf\rho - f_M^M)\| \\
= \|\tilde{C}_M^{-1}(\tilde{C}_M - L_M)\|\|C_{M,\lambda}^{-1/2}(\tilde{C}_M - L_M)(Pf\rho - f_M^M)\| \\
(29)
\]

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

\[
\|\tilde{C}_{M,\lambda}^{-1/2}C_{M,\lambda}^{-1/2}\| \leq \|C_{M,\lambda}^{-1/2}\|\|C_{M,\lambda}^{-1/2}\| \leq \|C_M^{-1/2}\|\|C_{M,\lambda}^{-1/2}\|^2 \leq \frac{\sqrt{3}}{2}\kappa\|C_{M,\lambda}^{-1/2}\|^2.
\]

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

\[
\|\tilde{C}_{M,\lambda}^{-1/2}C_{M,\lambda}^{-1/2}\| \leq \sqrt{\frac{3}{2}}\kappa\|C_{M,\lambda}^{-1/2}\|^2 \leq \frac{17\sqrt{3}}{4}\kappa^2 \kappa^2 \log\frac{2}{\delta} \left(\frac{A_{D,\lambda}^2}{\kappa^2 + \kappa} + 1\right).
\]

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

\[
\|C_{M,\lambda}^{-1/2}(\tilde{C}_M - L_M)\| \leq 2(2\kappa^2 + \kappa)\log\frac{2}{\delta} \lambda A_{D,\lambda} \leq 4(\kappa^2 + \kappa)\log\frac{2}{\delta} \lambda A_{D,\lambda}.
\]

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

\[
\|Pf\rho - f_M^M\| \leq \lambda^r ||g||
\]

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

\[
\|Pf\rho - f_M^M\| \leq \sqrt{\frac{3}{2}}\kappa\lambda^{-1/2} R.
\]

(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 [7] and [2] for $\delta \in (0, 1/2)$ and $\lambda > 0$, when

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

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

\[
\|f_M^M - f_X^M\|^2 \leq 16R^2\lambda^{2r},
\]

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

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

\[
\|f_M^M - f_X^M\| \leq 4\kappa(\lambda, M),
\]

(33)
where $\mathcal{C}(\lambda, M) = R\kappa^{2r-1}\left(\frac{\sqrt{\lambda F_\infty(\lambda)\log \frac{\lambda}{M}}}{M} + \frac{\lambda^2(\lambda)^{2r-1}F_\infty(\lambda)^{2-2r}\log \frac{\lambda}{M}}{M} \right) t^{1-r}$ and $t := \log \frac{11\kappa^2}{\lambda}$.

Proof details in Theorem 6 of [13] shows that under the condition
\[ M \geq 4\kappa^2\lambda^{-r} \nu(\lambda)^{2r-1}F_\infty(\lambda)^{2-2r}, \] (34)
we have
\[ \mathcal{C}(\lambda, M) \leq R\lambda^r. \] (35)

Then, we complete the proof by applying the following result.

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

**Lemma 6.** Under Assumption 7, the following holds for any $\lambda > 0$,
\[ \|f_\lambda - f_H\|^2_\rho \leq R^2\lambda^2. \]

**Proof.** Using Remark 3, we have $Pf_\rho = L'g$. 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)
\[ \|f_\lambda - f_H\| = \|LL_\lambda^{-1}Pf_\rho - Pf_\rho\| = \|L_\lambda^{-1}Pf_\rho = \|\lambda L_\lambda^{-1}L'g\|
= \|\lambda^{1-r}L_\lambda^{-1}(1-r)(L_\lambda^{-1}L')g\|
\leq \lambda^{r}\|\lambda^{1-r}L_\lambda^{-1}(1-r)\|\|L_\lambda^{-1}L'\|\|g\|
\]
Note that $\|\lambda^{1-r}L_\lambda^{-1}\| \leq 1$ and $\|L_\lambda^{-1}L'\| \leq 1$, while $R := \|g\|_{\rho, \lambda}$ 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 $f_H^{M_{D}, \lambda}$ be defined by (10). Under Assumptions 2, 3 and 4 when $0 < \lambda \leq \frac{3}{4}||L||$ and
\[ M \geq 4\kappa^2\left(\frac{\nu(\lambda)}{\lambda}\right)^{2r-1}F_\infty(\lambda)^{2-2r} \vee (4 + 18F_\infty(\lambda))\log \frac{48\kappa^2}{\lambda^6} \]
then the following holds with probability at least $1 - \delta$,
\[ \mathbb{E}[\mathcal{E}(f_H^{M_{D}, \lambda})] - \mathcal{E}(f_H) \leq C_2 \left[ \left( \frac{A_{D, \lambda}^{m, \lambda}}{\lambda} + 1 \right) \left( \frac{1}{m}B_{D, \lambda}^2 + A_{D, \lambda}^2 \lambda^{2r-1} \right) + \lambda^{2r} \right], \]
where $C_2$ is a constant independent of $m, N, N^*$ that
\[ C_2 = 6(\kappa^2 + \kappa)^{3/2} \left[ 289 + 677\kappa^4(\kappa^2 + \kappa)^2R^2 \right] \log^6 \frac{12}{\delta}, \]
and
\[ A_{D, \lambda} = \frac{m}{N^*} + \sqrt{\frac{m\nu(\lambda)}{N^*}}, \quad B_{D, \lambda} = \frac{mB}{N\sqrt{\lambda}} + \sqrt{\frac{m\nu(\lambda)}{N}}, \]

**Proof.** For SKRR-DC-RF (10), a similar error decomposition holds
\[ \mathbb{E}[\mathcal{E}(f_H^{M_{D}, \lambda})] - \mathcal{E}(f_H) \]
\[ \leq \frac{6}{m^2} \sum_{j=1}^{m} \mathbb{E}\|f^{M_{D}, \lambda} - f^{M_{D}, \lambda}\|^2_\rho \quad \text{(Variance)} \]
\[ + \frac{6}{m^2} \sum_{j=1}^{m} \mathbb{E}\|f^{M_{D}, \lambda} - f^{M_{D}, \lambda}\|^2_\rho \quad \text{(Empirical error)} \]
\[ + \frac{3}{m} \sum_{j=1}^{m} \mathbb{E}\|f^{M_{D}, \lambda} - J^{M}_{\lambda}\|^2_\rho \quad \text{(Distributed Error)} \]
\[ + 3\|f^{M}_{\lambda} - f^{\lambda}_{\rho}\|^2_\rho \quad \text{(Random Feature Error)} \]
\[ + 3\|f_{\lambda} - f_{H}\|^2_\rho \quad \text{(Approximation Error)} \]

(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 $\delta$ with $\tau$, such that both Lemma 5 and Lemma 6 hold with probability at least $1 - 2\tau$ with

$$c_0 = 289(\kappa^2 + \kappa)^4 \log^6 \frac{2}{\tau}, \quad c_1 = 676k^4(\kappa^2 + \kappa)6R^2 \log^6 \frac{2}{\tau}$$

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

Specifically, for estimates of $\mathbb{E}\| \hat{f}_D^M - f_D^M \|$ and $\mathbb{E}\| \hat{f}_D^M - f_D^M \|^2$, additional unlabeled samples have no influence on $A_{D_j,\lambda}$ but $B_{D_j,\lambda}$ is dependent on labels of dataset, which need to be replaced by

$$A_{D_j,\lambda} = \frac{m}{N^* \sqrt{\lambda}} + \sqrt{\frac{mN_M(\lambda)}{N^*}},$$

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

$$\mathbb{E}[\mathcal{E}(\hat{f}_D^M,\lambda)] - \mathcal{E}(f_H) \leq \frac{6}{m} \left[ A_{D_j,\lambda} \right]^2 \mathcal{B}_{D_j,\lambda}^2 + \left( \frac{6}{m} + 3 \right) c_1 \left( A_{D_j,\lambda}^2 \right)^2 \mathcal{A}_{D_j,\lambda}^2 \lambda^{2r-1} + 48R^2 \lambda^{3r} + 3R^2 \lambda^{2r}

\leq 6 \left( A_{D_j,\lambda} \right)^2 \left( \frac{c_0}{m} \mathcal{B}_{D_j,\lambda} + c_1 A_{D_j,\lambda} \lambda^{2r-1} \right) + 51R^2 \lambda^{2r}

\leq 6 \left( c_0 + c_1 + 9R^2 \right) \left[ \left( \frac{A_{D_j,\lambda}}{\lambda} \right)^2 + 1 \right] \left( \frac{1}{m} \mathcal{B}_{D_j,\lambda} + A_{D_j,\lambda} \lambda^{2r-1} \right) + \lambda^{2r}.$$ 

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

$$c_2 = 6(c_0 + c_1 + 9R^2) = 6 \left[ 289(\kappa^2 + \kappa)^4 + 676k^4(\kappa^2 + \kappa)6R^2 + 9R^2 \right] \log^6 \frac{12}{\delta}

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

and then complete the proof. \hfill \Box

**Theorem 5** (General excess risk bound). Let $\delta \in (0,1]$. Under Assumptions 2, 3, 4, 5 and 6, we have

$$n \geq \left( \frac{4}{3\|L\|} \right)^{2r+\gamma}, \quad \lambda = N^{-\frac{4}{\pi \gamma + \psi}},$$

$$M \geq c_3 N \left( \frac{2r-1}{\gamma-1} \right)^{2r+\gamma} \log \frac{56\kappa^2}{\lambda^\delta},$$

$$m \leq \min \left\{ N, N^* \kappa \left( \frac{2r-1}{\gamma-1} \right)^{2r+\gamma} \right\},$$

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

$$\mathbb{E}[\mathcal{E}(\hat{f}_D^M,\lambda)] - \mathcal{E}(f_H) \leq c_4 N^{-\frac{2r}{\pi \gamma + \psi}},$$

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

19
Combing (37), (38), (39) with Theorem 4, we have

\[
\mathbb{E} [ \mathcal{E}(f_{D^*, \lambda}) ] - \mathcal{E}(f_{\mathcal{H}}) \leq c_2 \left[ \left( \frac{A_{D_{j, \lambda}}^2}{\lambda} + 1 \right)^2 \left( \frac{1}{m} B_{D_{j, \lambda}}^2 + A_{D_{j, \lambda}}^2 \lambda^{2r-1} \right) + \lambda^{2r} \right],
\]

\[
\leq c_5 \left[ \left( \frac{A_{D_{j, \lambda}}^2}{\lambda} + 1 \right)^2 \left( \frac{1}{m} A_{D_{j, \lambda}}^2 + A_{D_{j, \lambda}}^2 \lambda^{2r-1} \right) + \lambda^{2r} \right],
\]

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

\[
A_{D_{j, \lambda}} = \frac{m}{N \sqrt{\lambda}} + \sqrt{\frac{m N(\lambda)}{N}}, \quad A_{D_{j, \lambda}} = \frac{m}{N^* \sqrt{\lambda}} + \sqrt{\frac{m N(\lambda)}{N^*}}.
\]

\[
M \geq 4\kappa^2 \left( \frac{N(\lambda)}{\lambda} \right)^{2r-1} \left( \mathcal{F}_\infty \log \frac{11\kappa^2}{\lambda} \right)^{2-2r} \sqrt{(4 + 18 \mathcal{F}_\infty(\lambda))} \log \frac{56\kappa^2}{\lambda^5}.
\]

Let \( \lambda = N^{-\frac{1}{2r+\gamma}} \), \( |D^*_1| = \cdots = |D^*_n| \) and \( |D_1| = \cdots = |D_m| = N/m \), under Assumption 4 with the fact \( r + s \geq r \geq 1/2 \) we have

\[
A_{D_{j, \lambda}} \leq mN^{-\frac{2r+2\gamma+1}{2r+2\gamma}} + Q \sqrt{mN^{-\frac{2r+2\gamma}{2r+2\gamma}}},
\]

\[
A_{D_{j, \lambda}} \leq mN^{-\frac{1}{2r+\gamma}} N^* \sqrt{\lambda} + Q \sqrt{mN^{-\frac{1}{2r+\gamma}} N^*}. \tag{37}
\]

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

\[
\lambda^{-1/2} A_{D_{j, \lambda}} \leq Q + 1. \tag{39}
\]

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

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

Note that \( n \) need to satisfy the associated constraint with respect to \( \lambda \) that \( \lambda \in (0, \frac{4}{3} \| L \|) \), 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

\[
M \geq 4\kappa^2 \left( \frac{N(\lambda)}{\lambda} \right)^{2r-1} \left( \mathcal{F}_\infty \log \left\{ \frac{11\kappa^2}{\lambda} \right\} \right)^{2-2r} \sqrt{(4 + 18 \mathcal{F}_\infty(\lambda))} \log \frac{56\kappa^2}{\lambda^5},
\]

we get

\[
M \geq c_3 N^{\left( \frac{2r-1}{2r+\gamma} \right) \lambda^{-\frac{1}{2r+\gamma}} + 1} \log \frac{56\kappa^2}{\lambda^5},
\]

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

\[
\square
\]

**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

\[
n \geq \left( \frac{4}{3\| L \|} \right)^{2r+\gamma}, \quad \lambda = N^{-\frac{1}{2r+\gamma}},
\]

\[
M \geq c_6 N^{\left( \frac{2r-1}{2r+\gamma} \right) + \frac{1}{\gamma+1}} \log \frac{56\kappa^2}{\lambda^5},
\]

\[
m \leq \min \left\{ N^{\frac{2r+2\gamma-1}{2r+2\gamma}}, N^* N^{-\frac{1}{2r+\gamma}} \right\}.
\]

20
with \(c_6 = 4(\kappa^2 Q^{4r-2} \kappa^{4-4r} + 4 + 18\kappa^2)\), then the following holds with probability at least \(1 - \delta\),

\[
\mathbb{E}[\mathcal{E}(\hat{f}_{M, \lambda}^M)] - \mathcal{E}(f_H) \leq c_4 N^{\frac{2r}{2r+1}},
\]

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

**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

\[
n \geq \left(\frac{4}{3\|L\|}\right)^{2r+\gamma}, \quad \lambda = \frac{1}{\sqrt{N}},
\]

\[
M \geq c_6 N \log^2 \frac{56\kappa^2}{\lambda \delta},
\]

\[
m \leq \min\left\{N, \frac{N^*}{N}\right\},
\]

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

\[
\mathbb{E}[\mathcal{E}(\hat{f}_{M, \lambda}^M)] - \mathcal{E}(f_H) \leq c_4 \frac{1}{\sqrt{N}},
\]

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

**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

\[
n \geq \left(\frac{4}{3\|L\|}\right)^{2r+\gamma}, \quad \lambda = N^{-\frac{1}{2r+1}},
\]

\[
M \geq c_6 N \frac{(2r-1)\gamma+1}{2r+1} \log^2 \frac{56\kappa^2}{\lambda \delta},
\]

\[
m \leq N^{\frac{2r-1}{2r+1}},
\]

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

\[
\mathbb{E}[\mathcal{E}(\hat{f}_{M, \lambda}^M)] - \mathcal{E}(f_H) \leq c_4 N^{-\frac{2r}{2r+1}},
\]

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

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