Towards Sharp Analysis for Distributed Learning with Random Features

Jian Li[1] and **Yong Liu**[2*]

[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]Gaoling School of Artificial Intelligence, Renmin University of China
lijian9026@iie.ac.cn, liuyonggsai@ruc.edu.cn

Abstract

In recent studies, the generalization properties for distributed learning and random features assumed the existence of the target concept over the hypothesis space. However, this strict condition is not applicable to the more common non-attainable case. In this paper, using refined proof techniques, we first extend the optimal rates for distributed learning with random features to the non-attainable case. Then, we reduce the number of required random features via data-dependent generating strategy, and improve the allowed number of partitions with additional unlabeled data. Theoretical analysis shows these techniques remarkably reduce computational cost while preserving the optimal generalization accuracy under standard assumptions. Finally, we conduct several experiments on both simulated and real-world datasets, and the empirical results validate our theoretical findings.

1 Introduction

A fundamental problem in machine learning is to achieve tradeoffs between statistical properties and computational costs [Bottou and Bousquet, 2008; Li *et al.*, 2018], while this challenge is more severe in kernel methods. Despite the excellent theoretical guarantees, kernel methods do not scale well in large-scale settings because of high time and memory complexities, typically at least quadratic in the number of examples. To break the scalability bottlenecks, researchers developed a wide range of practical algorithms, including distributed learning, which produces a global model after training disjoint subset on individual machines with necessary communications [Zhang *et al.*, 2015; Lin *et al.*, 2017], Nyström approximation [Williams and Seeger, 2001; Rudi *et al.*, 2015; Li *et al.*, 2019a] and random Fourier features [Rahimi and Recht, 2007; Rudi and Rosasco, 2017] to alleviate memory bottleneck, as well as stochastic methods [Lin and Cevher, 2020] to improve the training efficiency.

From the theoretical perspective, many researchers have studied the statistical properties of those large-scale approaches together with kernel ridge regression (KRR) [Rudi *et al.*, 2015; Lin and Rosasco, 2016; Lin *et al.*, 2017]. Using integral operator techniques [Smale and Zhou, 2007] and the effective dimension to control the capability of RKHS [Caponnetto and De Vito, 2007], the generalization bounds have achieved the optimal learning rates. Recent statistical learning studies on KRR together with large-scale approaches demonstrate that these approaches can not only obtain great computational gains but still remain the optimal theoretical properties, such as KRR together with divide-and-conquer [Guo *et al.*, 2017; Mücke and Blanchard, 2018], with random projections including Nyström approximation [Rudi *et al.*, 2015] and random features [Rudi and Rosasco, 2017; Carratino *et al.*, 2018; Li *et al.*, 2020; Li, 2021]. Since the communication cost is high to combine local kernel estimators in RKHS, it's more practical to combine the linear estimator in the feature space, e.g. federated learning [McMahan *et al.*, 2017].

The existing works on DKRR [Guo *et al.*, 2017; Lin *et al.*, 2017; Mücke and Blanchard, 2018] and random features [Rudi and Rosasco, 2017; Li *et al.*, 2019b; Li, 2021] have primarily focused on attainable cases, ignoring the non-attainable cases where the true regression is out of the hypothesis space. Since it's hard to select the suitable kernels to guarantee the attainable cases, the non-attainable cases are more common in practice. Therefore, the statistical guarantees for the non-attainable are of practical and theoretical interest in the context of the statistical learning theory. The optimal rates for DKRR have been extended to a part of the non-attainable case via sharp analysis for the distributed error [Lin and Cevher, 2020] and multiple communications [Lin *et al.*, 2020; Liu *et al.*, 2021], but these techniques are hard to improve the results for random features. Meanwhile, recent studies extended the capacity-independent optimality to the non-attainable cases, including distributed learning [Sun and Wu, 2020], random features [Sun *et al.*, 2018] and Nyström approximation [Kriukova *et al.*, 2017], but these capacity-independent results are suboptimal when the capacity of RKHS is small. *The capacity-optimality for the combination of distributed learning and random features to the non-attainable case is still an open problem.*

In this paper [1], we aim at extending the capacity-dependent optimal guarantees to the non-attainable case and improve the

[*] Corresponding author

[1] Full version: https://arxiv.org/abs/1906.03155

computational efficiency with more partitions and fewer random features. Firstly, using the refined estimation of operators' similarity, we refine the optimal generalization error bound that allows much more partitions and pertains to a part of the non-attainable case. Then, generating random features in a data-dependent manner, we relax the restriction on the dimension of random features, and thus fewer random features are sufficient to reach the optimal rates. By using additional unlabeled data to reduce label-independent error terms, we further enlarge the number of partitions and improve the applicable scope in the non-attainable case. Finally, we validate our theoretical findings with extensive experiments. Note that, we leave proofs and experiments in the full version

1.1 Our Contributions

We highlight our contributions as follows:

1) On algorithmic front: higher computational efficiency
This work presents the currently maximum number of partitions and the minimal dimension of random features, extremely improving the computational efficiency.

More partitions. To achieve the optimal learning rate, the traditional distributed KRR methods [Lin *et al.*, 2017; Guo *et al.*, 2017] impose a strict constraint on the number of partitions $m \lesssim N^{\frac{2r-1}{2r+\gamma}}$, which heavily limits the computational efficiency. In this paper, using a novel estimation of the key quantity, we first relax the restriction to $m \lesssim N^{\frac{2r+\gamma-1}{2r+\gamma}}$. Then, introducing a few additional unlabeled examples, we improve the number of partitions to $m \lesssim N^{\frac{2r+2\gamma-1}{2r+\gamma}}$ for the first time.

Fewer random features. By generating random features in a data-dependent manner rather than in a data-independent manner, we reduce the requirement on the number of random features from $M \gtrsim N^{\frac{(2r-1)\gamma+1}{2r+\gamma}} \quad \forall r \in [1/2, 1]$ to $M \gtrsim N^{\frac{2r+\gamma-1}{2r+\gamma}} \vee N^{\frac{\gamma}{2r+\gamma}} \quad \forall r \in (0,1]$, where M is the number of random features and $\vee$ indicates the bigger one.

2) On theoretical front: covering non-attainable cases
The conventional optimal properties for KRR [Caponnetto and De Vito, 2007; Rudi and Rosasco, 2017; Guo *et al.*, 2017] only pertain to the attainable case $r \in [1, 1/2]$, assuming the true regression belongs to the hypothesis space $f_\rho \in \mathcal{H}$ where the problems can not be too difficult. However, the condition $f_\rho \in \mathcal{H}$ is too ideal and the non-attainable $r \in (0, 1/2)$ assuming $f_\rho \notin \mathcal{H}$ deserve more attention. In this paper, we first restate the classic results in the attainable $r \in [1/2, 1]$. Then, by relaxing the restriction on the number of partitions, we extend the optimal theoretical guarantees to the non-attainable case with the constraints $2r + \gamma \geq 1$ and $2r + 2\gamma \geq 1$. Note that we prove KRR with random features applies to all non-attainable cases $r \in (0, 1/2)$.

3) Extensive experimental validation
To validate our theoretical findings, we conduct extensive experiments on simulated data and real-world data. We first construct simulated experiments under different difficulties to validate the learning rate and training time. Then, we perform comparison on a small real-world dataset to verify the effectiveness of data-dependence random features (with a novel approximate leverage score function) and additional unlabeled examples. Finally, we compare the proposed `DKRR-RF` with related work in terms of the performance on three real-world datasets.

4) Technical challenges
More partitions with additional unlabeled examples. In the error decomposition, only sample variance is label-dependent. At the same time, other terms are label-independent, and thus we employ additional unlabeled examples to reduce the estimation of label-independent error terms. We further improve the applicable scope in the non-attainable case to $m \lesssim N^{\frac{2r+2\gamma-1}{2r+\gamma}}$.

Random features error in all non-attainable cases. Using an appropriate decomposition on the operatorial level for random features error, we prove KRR with random features pertains to both attainable and non-attainable case $r \in (0, 1]$.

Overall, by overcoming several technical hurdles, we present the optimal theoretical guarantees for the combination of DKRR and RF. With more partitions and fewer random features, the theoretical results not only obtain significant computational gains but also preserve the optimal learning properties to both the attainable and non-attainable case $r \in (0, 1]$. Indeed, KRR [Caponnetto and De Vito, 2007], DKRR [Guo *et al.*, 2017], and KRR-RF [Rudi and Rosasco, 2017] are special cases of this paper. Thus, the techniques presented here pave the way for studying the statistical guarantees of other types kernel approaches (even neural networks) that can apply to the non-attainable case.

2 Distributed Learning with Random Feature

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 $\mathcal{Y} = \mathbb{R}$ is the output space. The training set $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ is sampled i.i.d. from $\mathcal{X} \times \mathcal{Y}$ with respect to ρ. The primary objective is to fit the target regression f_ρ on $\mathcal{X} \times \mathcal{Y}$. The Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ induced by a Mercer kernel K is defined as the completion of the linear span of $\{K(\boldsymbol{x}, \cdot), \boldsymbol{x} \in \mathcal{X}\}$ with respect to the inner product $\langle K(\boldsymbol{x}, \cdot), K(\boldsymbol{x}', \cdot)\rangle_{\mathcal{H}} = K(\boldsymbol{x}, \boldsymbol{x}')$. In the view of feature mappings, an underlying nonlinear feature mapping $\phi : \mathcal{X} \to \mathcal{H}$ associated with the kernel K is $\phi(\boldsymbol{x}) := K(\boldsymbol{x}, \cdot)$, so it holds $f(\boldsymbol{x}) = \langle f, \phi(\boldsymbol{x})\rangle_{\mathcal{H}}$.

2.1 Kernel Ridge Regression (KRR)

With an RKHS norm term, kernel ridge regression (KRR) is one of the popular empirical approaches to conducting a nonparametric regression. KRR can be stated as

$$\widehat{f}_\lambda := \operatorname*{arg\,min}_{f \in \mathcal{H}} \left\{ \frac{1}{N} \sum_{i=1}^N (f(\boldsymbol{x}_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}. \quad (1)$$

Using the representation theorem, the nonlinear regression problem (1) admits a closed form solution $\widehat{f}_\lambda(\boldsymbol{x}) = \sum_{i=1}^N \widehat{\alpha}_i K(\boldsymbol{x}_i, \boldsymbol{x})$ with

$$\widehat{\alpha} = (\mathbf{K}_N + \lambda N \mathbf{I})^{-1} \mathbf{y}_N, \quad (2)$$

where $\lambda > 0, \mathbf{y}_N = [y_1, \cdots, y_N]^T$ and $\mathbf{K}_N$ is the $N \times N$ kernel matrix with $\mathbf{K}_N(i,j) = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Although KRR characterizes optimal statistical properties [Smale and Zhou, 2007; Caponnetto and De Vito, 2007], it is unfeasible for large-scale settings because of $\mathcal{O}(N^2)$ memory to store kernel matrix and $\mathcal{O}(N^3)$ time to solve the linear system (2).

2.2 Distributed KRR with Random Features (DKRR-RF)

Assume that the kernel K have an integral representation

$$K(\boldsymbol{x}, \boldsymbol{x}') = \int_\Omega \psi(\boldsymbol{x}, \omega)\psi(\boldsymbol{x}', \omega)p(\omega)d\omega, \ \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}, \quad (3)$$

where (Ω, π) is a probability space and $\psi : \mathcal{X} \times \Omega \to \mathbb{R}$. We define analogous operators for the constructed kernel $K_M(\boldsymbol{x}, \boldsymbol{x}') = \phi_M(\boldsymbol{x})^\top \phi_M(\boldsymbol{x}')$ to approximate the primal kernel $K(\boldsymbol{x}, \boldsymbol{x}')$ in (3) with its corresponding random features via Monte Carlo sampling

$$\phi_M(\boldsymbol{x}) = \frac{1}{\sqrt{M}}\big(\psi(\boldsymbol{x}, \omega_1), \cdots, \psi(\boldsymbol{x}, \omega_M)\big)^\top, \quad (4)$$

where $\omega_1, \cdots, \omega_M \in \Omega$ are sampled w.r.t. $p(\omega)$.

Let the training set D be randomly partitioned into m disjoint subsets $\{D_j\}_{j=1}^m$ with $|D_1| = \cdots = |D_m| = n$. The local estimator $\widehat{\boldsymbol{w}}_j$ on the subset D_j is defined as

$$\widehat{\boldsymbol{w}}_j = \underset{\boldsymbol{w} \in \mathbb{R}^M}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^n (f(\boldsymbol{x}_i) - y_i)^2 + \lambda \|f\|^2 \right\}, \quad (5)$$

where the estimator is $f(\boldsymbol{x}) = \langle \boldsymbol{w}, \phi_M(\boldsymbol{x}) \rangle$. It admits a closed-form solution

$$\widehat{\boldsymbol{w}}_j = \left[\Phi_M^\top \Phi_M + \lambda I\right]^{-1} \Phi_M^\top \widehat{y}_n, \quad (6)$$

where $\lambda > 0$. Note that for j-th subset D_j, it holds $\forall(\boldsymbol{x}, y) \in D_j, \Phi_M = \frac{1}{\sqrt{n}}[\phi_M(\boldsymbol{x}_1), \cdots, \phi_M(\boldsymbol{x}_n)]^\top \in \mathbb{R}^{n \times M}$ and $\widehat{y}_n = \frac{1}{\sqrt{n}}(y_1, \cdots, y_n)^\top$. The average of local estimators (6) yields a global estimator

$$\widehat{f}_{D,\lambda}^M(\boldsymbol{x}) = \frac{1}{m} \sum_{j=1}^m \widehat{f}_{D_j,\lambda}^M(\boldsymbol{x}). \quad (7)$$

3 Theoretical Assessment

In this section, we present the theoretical analysis on the generalization performance of kernel ridge regression with divide-and-conquer and random features.

The generalization ability of a regression predictor $f : \mathcal{X} \to \mathbb{R}$ is measured in terms of the *expected risk*

$$\mathcal{E}(f) = \int_{\mathcal{X} \times \mathcal{Y}} (f(\boldsymbol{x}) - y)^2 d\rho(\boldsymbol{x}, y). \quad (8)$$

In this case, the target regression $f_\rho = \int_\mathcal{Y} y d\rho(y|\boldsymbol{x}), \ \forall \boldsymbol{x} \in \mathcal{X}$ minimizes the *expected risk* over all measurable functions $f : \mathcal{X} \to \mathbb{R}$. The generalization ability of a KRR estimator $f \in L^2_{\rho_X}$ is measured by the *excess risk*, i.e. $\mathcal{E}(f) - \mathcal{E}(f_\rho)$, where $L^2_{\rho_X} = \{f : \mathcal{X} \to \mathbb{R} \mid \|f\|_\rho^2 = \int_X |f(\boldsymbol{x})|^2 d\rho_X < \infty\}$ is the square integral Hilbert space with respect to the marginal distribution ρ_X on the input space $\mathcal{X}$.

3.1 Assumptions

We first introduce two standard assumptions, which are also used in statistical learning theory [Smale and Zhou, 2007; Caponnetto and De Vito, 2007; Rudi and Rosasco, 2017].

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

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

$$\int_\mathbb{R} |y|^p d\rho(y|\boldsymbol{x}) \leq \frac{1}{2} p! B^{p-2} \sigma^2. \quad (9)$$

According to Assumption 1, the kernel K is bounded by $K(\boldsymbol{x}, \boldsymbol{x}) \leq \kappa^2$. The moment assumption on the output y holds when y is bounded, sub-gaussian or sub-exponential. Assumptions 1 and 2 are standard in the generalization analysis of KRR, always leading to the learning rate $\mathcal{O}(1/\sqrt{N})$ [Smale and Zhou, 2007] in general cases.

Definition 1 (Integral operators). $\forall\ g \in L^2_{\rho_X}(X, \rho_X)$*, the integral operators L, L_M are defined by the kernel K and the random features ϕ_M, respectively*

$$(Lg)(\cdot) = \int_X K(\cdot, \boldsymbol{x}) g(\boldsymbol{x}) d\rho_X(\boldsymbol{x}),$$
$$(L_M g)(\cdot) = \int_X \langle \phi_M(\cdot), \phi_M(\boldsymbol{x}) \rangle g(\boldsymbol{x}) d\rho_X(\boldsymbol{x}).$$

Definition 2 (Effective dimension)**.** *The effective dimension of the RKHS $\mathcal{H}$ induced by the kernel K is defined as*

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

The effective dimension $\mathcal{N}(\lambda)$ is used to measure the complexity of RKHS $\mathcal{H}$, and its empirical counterpart is also called degree of freedom [Bach, 2013]. Similarly, we define the effective dimension $\mathcal{N}_M(\lambda)$ for the random features mapping ϕ_M to measure the size of the approximate RKHS $\mathcal{H}_M$, which is induced by finite dimensional random features $\phi_M : \mathcal{X} \to \mathbb{R}^M$.

Assumption 3 (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 4 (Regularity assumption)**.** *Assume there exists $R > 0$, $r > 0$, and $g \in L^2_{\rho_X}$, such that*

$$f_\rho = L^r g,$$

where f_ρ is the target regression, $\|g\|_\rho \leq R$ and the operator L^r denotes the r-th power of the integral operator $L : L^2_{\rho_X} \to L^2_{\rho_X}$, thus it is also a positive trace class operator.

The above two conditions are commonly used to prove the optimal statistical properties of the exact KRR [Caponnetto and De Vito, 2007; Smale and Zhou, 2007], and its large-scale extensions including divide-and-conquer [Lin *et al.*, 2017; Guo *et al.*, 2017] and random features [Rudi and Rosasco,

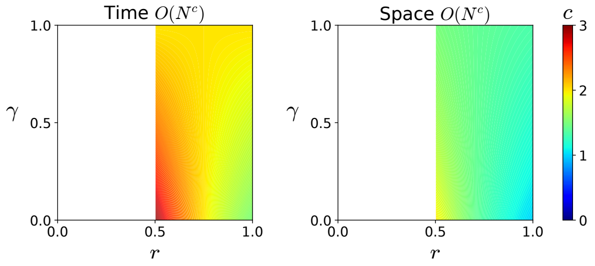

Figure 1: Time complexity and space complexity of Theorem 1 in different settings. The color closer to red represents higher complexity. Blank areas represent unfeasible situations.

2017]. Those two assumptions reflect the capacity of the RKHS $\mathcal{H}$ and the regularity of f_ρ, respectively. We provide some intuitive interpretations of the above assumptions, and more details can be found in [Caponnetto and De Vito, 2007]. Assumption 3 holds when the eigenvalues of the integral operator have a polynomial decay $i^{-1/\gamma}$, $\forall i > 1$ [Rudi and Rosasco, 2017; Li *et al.*, 2019b]. Thus, faster convergence rates are derived when the eigenvalues decay faster, a.k.a. γ approaches 0, while $\gamma = 1$ corresponds to the capacity-independent case. Assumption 4 (source condition) controls the regularity of the target function f_ρ. The bigger the r is, the stronger regularity of the regression is, and the easier the learning problem is. Both these two assumptions are widely used in the optimal theory for KRR [Caponnetto and De Vito, 2007; Rudi and Rosasco, 2017; Guo *et al.*, 2017].

3.2 General Results with Fast Rates

One can prove the optimal generalization guarantees for DKRR-RF by combining the theories in KRR-DC [Lin *et al.*, 2017] and KRR-RF [Rudi and Rosasco, 2017]. The attainable case $r \in [1/2, 1]$ requires the existence of $f_{\mathcal{H}} = \min_{f\in\mathcal{H}} \mathcal{E}(f)$, such that $f_\rho = f_{\mathcal{H}}$ almost surely [Steinwart and Christmann, 2008], which is widely used in KRR and its variants including distributed KRR and random features based KRR [Caponnetto and De Vito, 2007; Rudi and Rosasco, 2017; Guo *et al.*, 2017].

Theorem 1. *Under Assumptions 1, 2, 3 and 4, if* $r \in [1/2, 1], \gamma \in [0, 1]$, *and* $\lambda = N^{-\frac{1}{2r+\gamma}}$, *then*

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

are enough to guarantee, with a high probability, that

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

The optimal learning rate $\mathcal{O}\left(N^{-\frac{2r}{2r+\gamma}}\right)$ stated in Theorem 1 in the above bound is optimal in a minimax sense for KRR approaches [Caponnetto and De Vito, 2007]. Distributed KKR methods have obtained the same optimal error bounds with a stronger condition on the number of partitions, such as KRR-DC [Lin *et al.*, 2017; Mücke and Blanchard, 2018] with $m \lesssim N^{\frac{2r-1}{2r+\gamma}}$. In particular, for the general case $r = 1/2$, the number of local processors $m = \mathcal{O}(1)$ becomes a constant number that is independent of the sample size N. The time complexity of DKRR-RF is $\mathcal{O}(NM^2/m)$ and the space complexity $\mathcal{O}(NM/m)$, thus we report the computational complexities of Theorem 1 in Figure 1.

Remark 1. *The general results in Theorem 1 have three fatal drawbacks: 1) the above bound is only suitable for the attainable case* $r \in [1/2, 1]$ *and fail to apply to the non-attainable case* $r \in (0, 1/2)$ *induced by more complicated problems; 2) random features generated via Monte Carlo are data-independent, which requires much more features than the data-dependent generating features; 3) the constraint on the number of partitions* $m \lesssim N^{\frac{2r-1}{2r+\gamma}}$ *is too strict, leading to a constant number of partitions when* r *is close to* $1/2$.

3.3 Refined Results in the Non-attainable Case

Theorem 2. *Under Assumptions 1, 2, 3 and 4, if* $r \in (0, 1]$, $\gamma \in [0, 1]$, $2r + \gamma \geq 1$ *and* $\lambda = N^{-\frac{1}{2r+\gamma}}$, *then the number of partitions corresponding to*

$$1 \lesssim m \lesssim N^{\frac{2r+\gamma-1}{2r+\gamma}}$$

and the number of random features M *satisfying*

$$M \gtrsim N^{\frac{1}{2r+\gamma}} \quad \text{when } 0 < r < 1/2 \qquad \text{and}$$
$$M \gtrsim N^{\frac{(2r-1)\gamma+1}{2r+\gamma}} \quad \text{when } 1/2 \leq r \leq 1,$$

are enough to guarantee, with a high probability, that

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

Compared to Theorem 1, Theorem 2 allows more partitions and extends the optimal learning guarantees to the non-attainable case $r \in (0, 1/2)$ where the true regression does not lie in RKHS $\mathcal{H}$. Thus, it achieves significant improvements in both computational efficiency and statistical guarantees. With the same optimal learning rates, Theorem 2 relaxes the restriction on m from $m \lesssim N^{\frac{2r-1}{2r+\gamma}}$ to $m \lesssim N^{\frac{2r+\gamma-1}{2r+\gamma}}$, which allows more partitions and relaxes the constraints from $r \geq 1/2$ to $2r + \gamma \geq 1$. More importantly, for the general cases when $r = 1/2$, the number of partitions becomes $m \lesssim N^{\frac{\gamma}{2r+\gamma}}$, which increases as the sample size N becomes larger and avoids the constant number of partitions $m = \mathcal{O}(1)$ in the conventional KRR-DC methods [Guo *et al.*, 2017; Lin *et al.*, 2017]. When $r \in (0, 1/2)$, the number of random features $M \gtrsim N^{\frac{1}{2r+\gamma}}$ increases as the r approaches zero, because f_ρ becomes far away from $\mathcal{H}$ when r is near zero. When $r \in [1/2, 1]$, we obtain the same level of the number of random features $M \gtrsim N^{\frac{(2r-1)\gamma+1}{2r+\gamma}}$ as KRR-RF [Rudi and Rosasco, 2017], which is continuous to $M \gtrsim N^{\frac{1}{2r+\gamma}}$ at the critical points $r = 1/2$. Compared to Figure 1, Figure 2 illustrates Theorem 2 not only enlarge the applicable case but also improve the computational efficiency.

Remark 2. *The common optimal generalization learning of KRR [Smale and Zhou, 2007] with random features techniques [Rudi and Rosasco, 2017] and divide-and-conquer [Zhang* et al.*, 2015; Lin and Rosasco, 2016] focus on the generalization properties on the standard setting* $f_\rho \in \mathcal{H}$

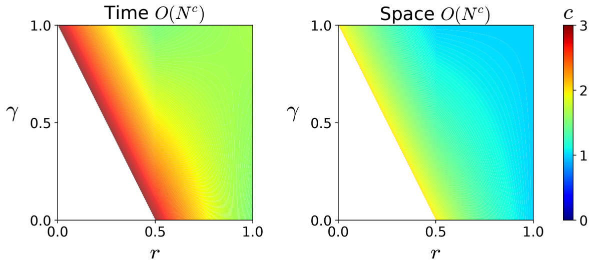

Figure 2: Time complexity and space complexity of Theorem 2 in different settings. The color closer to red represents higher complexity. Blank areas represent unfeasible situations.

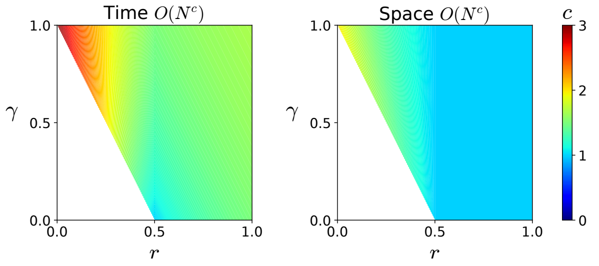

Figure 3: Time complexity and space complexity of Theorem 3 in different settings. The color closer to red represents higher complexity. Blank areas represent unfeasible situations.

under the condition $r \in [1/2, 1]$. Note that, in this paper, we extend the range of regularity assumption to $r \in (0, 1/2)$ for considering the situation $f_\rho \notin \mathcal{H}$. Meanwhile, when $r > 1$, the divide-and-conquer approach can only reduce the sample error and computational variance, but not bias. The situation is called the saturation phenomenon and observed in KRR-DC approaches [Zhang et al.*, 2015; Lin and Cevher, 2020]. In the future, it's worthy to reducing the bias with multi-communications using stochastic methods rather than the divide-and-conquer strategy.*

Remark 3. *Theorem 2 extends the optimal generalization theories from only attainable case $r \in [1/2, 1]$ to the non-attainable case $2r + \gamma \geq 1$, which include a part of difficult problems $r \in (0, 1/2)$. However, there are also many cases satisfying $2r + \gamma < 1$ in the non-attainable case $r \in (0, 1/2)$, where the optimal learning guarantees in Theorem 2 are no longer valid. Inspired the literature [Chang* et al.*, 2017], we employ additional unlabeled samples to relax the restriction $2r + \gamma \geq 1$ in Section 3.5.*

3.4 Fewer Features with Data-dependent Sampling

Assumption 5 (Compatibility assumption [Rudi and Rosasco, 2017])**.** *Define the maximum effective dimension as*

$$\mathcal{N}_\infty(\lambda) = \sup_{\omega \in \Omega} \|(L + \lambda I)^{-1/2} \psi(\cdot, \omega)\|_{\rho_X}^2, \lambda > 0.$$

Assume there exists $\alpha \in [0, 1]$ and $F > 0$, such that

$$\mathcal{N}_\infty(\lambda) \leq F \lambda^{-\alpha}.$$

Using the definition of $\mathcal{N}(\lambda)$, we characterize the lower bounds for $\mathcal{N}_\infty(\lambda)$:

$$\begin{aligned}\mathcal{N}(\lambda) &= \mathbb{E}_\omega \|(L + \lambda I)^{-1/2} \psi(\cdot, \omega)\|_{\rho_X}^2 \\ &\leq \sup_{\omega \in \Omega} \|(L + \lambda I)^{-1/2} \psi(\cdot, \omega)\|_{\rho_X}^2 = \mathcal{N}_\infty(\lambda).\end{aligned}$$

Compared to the (average) effective dimension used in Assumption 3, the maximum effective dimension offers a finer-grained estimate for the capacity of RKHS [Alaoui and Mahoney, 2015; Rudi and Rosasco, 2017; Rudi *et al.*, 2018], which often leads to shaper estimate for the related quantities. Using the compatibility assumption, we relax the constraints on the dimension of random features and the number of partitions by generating features in a data-dependent manner, as shown in [Rudi *et al.*, 2018; Bach, 2017; Li *et al.*, 2019b].

Theorem 3. *Under the same assumptions of Theorem 2 and Assumption 5, if $r \in (0, 1]$, $\gamma \in [0, 1]$, $2r + \gamma \geq 1$ and $\lambda = N^{-\frac{1}{2r+\gamma}}$, then the number of partitions m satisfying*

$$1 \lesssim m \lesssim N^{\frac{2r+\gamma-1}{2r+\gamma}}$$

and the number of random features M satisfying

$$\begin{aligned}M &\gtrsim N^{\frac{\alpha}{2r+\gamma}} && \text{when } \ 0 < r < 1/2 \qquad \text{and} \\ M &\gtrsim N^{\frac{(2r-1)(1+\gamma-\alpha)+\alpha}{2r+\gamma}} && \text{when } \ 1/2 \leq r \leq 1,\end{aligned}$$

is sufficient to guarantee, with a high probability, that

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

The learning rates of the above theorem are optimal, same as Theorems 2. Achieving the same optimal learning rates, Theorem 3 reduce the computational costs with fewer random features. The number of required random features is reduced from $\mathcal{O}\big(N^{\frac{1}{2r+\gamma}}\big)$ to $\mathcal{O}\big(N^{\frac{\alpha}{2r+\gamma}}\big)$ when $r \in (0, 1/2)$ and $\mathcal{O}\big(N^{\frac{(2r-1)\gamma+1}{2r+\gamma}}\big)$ to $\mathcal{O}\big(N^{\frac{(2r-1)\gamma+1+2(r-1)(1-\alpha)}{2r+\gamma}}\big)$ when $r \in [1/2, 1]$, where the term $2(r-1)(1-\alpha) \leq 0$. We report the applicable area and computational complexities of Theorem 3 in Figure 3. It shows the use of data-dependent sampling significantly reduce both the time and space complexities. The situations near the boarderline $2r + \gamma = 1$ are away from the same computational complexities as the exact KRR.

Remark 4. *From Theorem 1 in [Li* et al.*, 2019b], we find that the requirement on the data-dependent random features is bounded as $M \gtrsim d_{\tilde{l}} := \sup_{\boldsymbol{w} \in \Omega} l_\lambda(\boldsymbol{w})/q(\boldsymbol{w})$, where $d_{\tilde{l}} \propto \mathcal{N}_\infty(\lambda) \leq F N^{\frac{\alpha}{2r+\gamma}}$. The condition is the same as Theorem 3 in the non-attainable $r \in (0, 1/2)$ and milder than Theorem 3 in the attainable case $r \in [1/2, 1]$. However, the theoretical analysis provided in [Li* et al.*, 2019b] only pertains to the general case $(r = 1/2, \gamma = 1)$ and obtains error bounds with the convergence rate $\mathcal{O}(1/\sqrt{N})$.*

Remark 5. *According to the definition of $\mathcal{N}_\infty(\lambda)$, the sampling probability of random features $\pi(\omega)$ is independent of data, which leads to a pessimistic estimate of α. However, generating random features in a data-dependent manner relaxes the estimate of α closer to γ. A theoretical example of data-dependent random features was given in Example 2 [Rudi and Rosasco, 2017], which guarantees $\mathcal{N}_\infty(\lambda) = \mathcal{N}(\lambda)$ (such that $\alpha = \gamma$) by constructing random features*

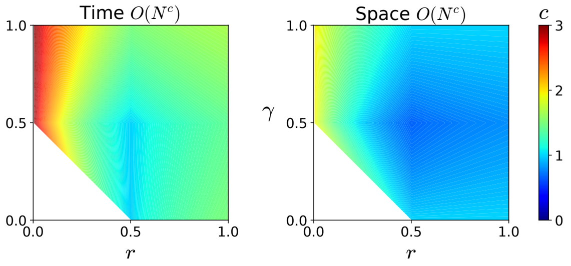

Figure 4: Time complexity and space complexity of Theorem 4 versus different values of r and γ. The color which is closer to red represents higher complexity.

generated in a data-dependent way. In practice, leverage sampling algorithms were proposed to obtain data-dependent random features [Li et al.*, 2019b], where α is close to γ. To intuitively illustrate the improvement of data-dependent random features, we boldly assume $\alpha = \gamma$ by generating data-dependent random features.*

3.5 More Partitions with Unlabeled Data

Theorem 2 illustrates the optimal learning rates for `DKRR-RF` needs the number of local partitions satisfying

$$m \lesssim N^{\frac{2r+\gamma-1}{2r+\gamma}}.$$

The number of partitions avoids a constant number of partitions when $r \in [1/2, 1]$, but it is still unfeasible for many challenging problems in the non-attainable case $r \in (0, 1/2)$. In this part, we introduce the additional unlabeled samples $\widetilde{D}_j$ to relax this restriction further. We consider the merged dataset D^* on the j-th processor, $D_j^* = D_j \cup \widetilde{D}_j$ with

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

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

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

Theorem 4. *Under the same assumptions of Theorem 3, if $r \in (0, 1], \gamma \in [0, 1], 2r + 2\gamma \geq 1$ and $\lambda = N^{-\frac{1}{2r+\gamma}}$, then the total number of samples corresponding to*

$$N^* \gtrsim N N^{\frac{\gamma+\alpha-1}{2r+\gamma}} \vee N,$$

the number of local processors satisfying

$$1 \lesssim m \lesssim N^{\frac{2r+2\gamma-1}{2r+\gamma}}$$

and the number of random features M satisfying

$$M \gtrsim N^{\frac{\alpha}{2r+\gamma}} \quad \textit{when} \;\; 0 < r < 1/2 \qquad \textit{and}$$

$$M \gtrsim N^{\frac{(2r-1)(1+\gamma-\alpha)+\alpha}{2r+\gamma}} \quad \textit{when} \;\; 1/2 \leq r \leq 1,$$

are sufficient to guarantee, with a high probability, that

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

To our best knowledge, for the first time, we prove that the number of partitions can achieve $m \lesssim N^{\frac{2r+2\gamma-1}{2r+\gamma}}$, while the existing constraints on m of the existing work [Lin and Cevher, 2020; Liu *et al.*, 2021] are $m \lesssim N^{\frac{2r+\gamma-1}{2r+\gamma}}$. Such that, much more partitions are allowed in distributed KRR methods. The relaxation of condition on the partition number m can not only lead to better computational efficiency but also covers more difficult problems, where the suitable problems are enlarged from the situation $2r + \gamma \geq 1$ to the situation $2r + 2\gamma \geq 1$. Figure 4 reveals the advantages of `DKRR-RF` with unlabeled data. Theorem 4 provides the largest applicable area $2r + 2\gamma \geq 1$ but also the highest computational efficiency owing to more partitions.

Remark 6. *From the error decomposition, there are two error terms related to the number of partitions m: sample variance and empirical error. Sample variance depends on the number of labeled samples n, while empirical error is input-dependent but output-independent; thus, it is related to the number of total samples n^*. Meanwhile, the similarity between empirical and expected covariance operators $\|\widehat{C}_{M,\lambda}^{-1/2} C_{M,\lambda}^{1/2}\|$ is also label-free, and thus it is related to the total sample size n^* rather than n. To achieve the optimal learning rates, we consider the constraints on both the required labeled samples n and the total samples n^*. Considering both conditions for supervised learning $m = N/n$ and semi-supervised learning $m = N^*/n^*$, we then obtain two constraints on the number of partitions m and consolidate them together.*

4 Compared with Related Work

The existing optimal learning guarantees of KRR [Caponnetto and De Vito, 2007], KRR-DC [Guo *et al.*, 2017; Mücke and Blanchard, 2018] and KRR-RF [Rudi and Rosasco, 2017; Liu *et al.*, 2021] only apply to the attainable case $r \in [1/2, 1]$. In this paper, we apply the optimal generalization error bounds to the non-attainable case $r \in (0, 1/2)$ with some restrictions, including $2r + \gamma \geq 1$ in Theorem 2 and $2r + 2\gamma \geq 1$ in Theorem 4. Using refined estimation, we extend the random features error to the non-attainable case.

4.1 Applicable Area from $r \in [1/2, 1]$ to $2r + \gamma \geq 1$

The key to obtaining the optimal learning rates with integral-operator approach is to bound the identity $\|(\widehat{C}_M + \lambda I)^{-1/2}(C_M + \lambda I)^{1/2}\|$ as a constant, where C_M and $\widehat{C}_M$ are the expected and empirical covariance operators defined in Definition 4. In conventional distributed KRR [Lin *et al.*, 2017; Chang *et al.*, 2017], they estimated the operator difference after first order (or second order) decomposition

$$\begin{aligned} &\|(C_M + \lambda I)^{-1/2}(\widehat{C}_M + \lambda I)^{1/2}\|^2 \\ \leq\ &\|(C_M + \lambda I)^{-1/2}\| \|(C_M + \lambda I)^{-1/2}(C_M - \widehat{C}_M)\| + 1 \\ =\ &\mathcal{O}\left(\frac{m}{\lambda N} + \sqrt{\frac{\mathcal{N}(\lambda) m}{\lambda N}}\right). \quad \text{Section 4 [Guo } \textit{et al.}\text{, 2017].} \end{aligned}$$

To bound the identity as a constant, the local sample size should be larger enough $n \geq \frac{\mathcal{N}(\lambda)}{\lambda}$. It holds $m \lesssim N^{\frac{2r-1}{2r+\gamma}}$

for KRR-DC and only applies to $r \geq 1/2$. However, this paper directly estimates the identity in total (rather than in parts after decomposition) based on concentration inequalities for self-adjoint operators and obtain

$$\begin{aligned} & \|(C+\lambda I)^{-1/2}(\widehat{C}_M+\lambda I)^{1/2}\| \\ \leq \; & \left(1-\left\|(C+\lambda I)^{-1/2}(C-\widehat{C}_M)(\widehat{C}_M+\lambda I)^{-1/2}\right\|\right)^{-1/2} \\ = \; & \mathcal{O}\left(\frac{m}{\lambda N}+\sqrt{\frac{m}{\lambda N}}\right). \qquad \text{Theorem 2} \end{aligned}$$

To bound the identity as a constant, the local sample size only needs $n \geq \frac{1}{\lambda}$, which is smaller than [Guo *et al.*, 2017] with $\mathcal{N}(\lambda)$. Therefore, our estimation of $\|(C_M+\lambda I)^{-1/2}(\widehat{C}_M+\lambda I)^{1/2}\|$ in Theorem 2 is $\sqrt{\mathcal{N}(\lambda)}$ tighter than that in [Guo *et al.*, 2017]. To bound identity as a constant, we then have $m \lesssim N^{\frac{2r+\gamma-1}{2r+\gamma}}$, which is the key to obtain more partitions and extends the optimal learning guarantees to the non-attainable case $2r+\gamma \geq 1$.

4.2 Applicable Area from $2r+\gamma \geq 1$ to $2r+2\gamma \geq 1$

Only sample variance is dependent on the labeled samples, while other error terms involving the estimate of $\|(C+\lambda I)^{-1/2}(\widehat{C}_M+\lambda I)^{1/2}\|$ are label-free. Thus, there are two restrictions on the number of partitions m: sample variance (label-dependent) and the estimate of $\|(C+\lambda I)^{-1/2}(\widehat{C}_M+\lambda I)^{1/2}\|$ (label-free).

As shown in the proof of Theorem 3, the global sample variance (label-dependent) can be estimated

$$\frac{1}{m}\mathbb{E}\|\widehat{f}^M_{D_j,\lambda}-\widetilde{f}^M_{D_j,\lambda}\|^2_\rho \;\leq\; \mathcal{O}\left(mN^{\frac{1-4r-2\gamma}{2r+\gamma}}+N^{\frac{-2r}{2r+\gamma}}\right)$$

To achieve the optimal learning rates $\mathcal{O}(N^{\frac{-2r}{2r+\gamma}})$, the number of partitions should satisfy $m \lesssim N^{\frac{2r+2\gamma-1}{2r+\gamma}}$. Then, we utilize additional unlabeled samples to relax the condition on the estimate of $\|(C+\lambda I)^{-1/2}(\widehat{C}_M+\lambda I)^{1/2}\|$. Using Assumption 5, one can further relax the condition of m due to

$$\begin{aligned} & \|(C+\lambda I)^{-1/2}(\widehat{C}_M+\lambda I)^{1/2}\| \\ \leq \; & \mathcal{O}\left(\frac{m\mathcal{N}_\infty(\lambda)}{N^*}+\sqrt{\frac{m\mathcal{N}_\infty(\lambda)}{N^*}}\right) \\ = \; & \mathcal{O}\left(\frac{m}{\lambda^\alpha N^*}+\sqrt{\frac{m}{\lambda^\alpha N^*}}\right). \qquad \text{Theorem 4} \end{aligned}$$

To guarantee the key quantity $\|(C_M+\lambda I)^{-1/2}(\widehat{C}_M+\lambda I)^{1/2}\|$ be a constant, we have $m \lesssim \lambda^\alpha N^* = \mathcal{O}(N^* N^{\frac{-\alpha}{2r+\gamma}})$. We then consider the dominant constraints:

- The case $\alpha < 1-\gamma$. It holds $2r+2\gamma-1 < 2r+\gamma-\alpha$, thus the number of partition is $m \lesssim N^{\frac{2r+2\gamma-1}{2r+\gamma}}$.
- The case $\alpha \geq 1-\gamma$. It holds $\gamma+\alpha-1 \geq 0$ and we make use of additional unlabeled examples $N^* \gtrsim NN^{\frac{\gamma+\alpha-1}{2r+\gamma}}$ to guarantee $m \lesssim N^{\frac{2r+\gamma-\alpha}{2r+\gamma}} \leq N^{\frac{2r+2\gamma-1}{2r+\gamma}}$.

4.3 Random Features Error in the Non-attainable Case

Using appropriate decomposition on operatorial level, we derive the random features error for both attainable and non-attainable case, where the dimension of random features should satisfy $M \gtrsim N^{\frac{\gamma}{2r+\gamma}}$ for the non-attainable case $r \in (0, 1/2)$. The extension from the attainable case to the non-attainable case is non-trivial, where the non-attainable case requires refined estimations for operators similarity.

The operatorial definitions of intermediate estimators $\widetilde{f}^M_{D_j,\lambda}$, f^M_λ and f_λ in Lemma 1 involve the true regression f_ρ, where $f_\rho = L^r g$ (under Assumption 4) is related to the range of r. Such that, we estimate the last three error items (empirical error $\|\widetilde{f}^M_{D_j,\lambda} - f^M_\lambda\|$, random features error $\|f^M_\lambda - f_\lambda\|$ and approximation error $\|f_\lambda - f_\rho\|$) that involve $\widetilde{f}^M_{D_j,\lambda}$, f^M_λ and f_λ for the non-attainable case. Meanwhile, because the empirical error satisfies $\|\widetilde{f}^M_{D_j,\lambda} - f^M_\lambda\| \leq (\sqrt{2}+2)\left(\|f^M_\lambda - f_\lambda\| + \|f_\lambda - f_\rho\|\right)$ and the approximation error $\|f_\lambda - f_\rho\|$ naturally applies to the non-attainable case, only random features error $\|f^M_\lambda - f_\lambda\|$ is needed to specifically estimated for the non-attainable case.

5 Conclusion

This paper explores the generalization performance of kernel ridge regression with two commonly used efficient large-scale techniques: divide-and-conquer and random features. We first present a general result with the optimal learning rates under standard assumptions. We then refine the theoretical results with more partitions and applicability in the non-attainable case. Further, we reduce the number of random features by generating features in a data-dependent manner. Finally, we present the theoretical results that substantially relax the constraint on the number of partitions with extra unlabeled data, which apply to both the attainable case and non-attainable case. The proposed optimal theoretical guarantees are state-of-the-art in the theoretical analysis for KRR approaches. We validate our theoretical findings with extensive experimental results.

This paper can be extended in several ways: (a) the combination with gradient algorithms such as multi-pass SGD [Lin and Cevher, 2018; Lin and Cevher, 2020] and preconditioned conjugate gradient [Avron *et al.*, 2017] to further reduce the time complexity. (b) using asynchronous distributed methods or a few of communications [Lin *et al.*, 2020; Liu *et al.*, 2021] instead of one-shot approach to alleviate the saturation phenomenon when $r \geq 1$.

Acknowledgments

This work was supported in part by the Excellent Talents Program of Institute of Information Engineering, CAS, the Special Research Assistant Project of CAS, the Beijing Outstanding Young Scientist Program (No. BJJWZYJH012019100020098), Beijing Natural Science Foundation (No. 4222029), and National Natural Science Foundation of China (No. 62076234, No. 62106257).

References

[Alaoui and Mahoney, 2015] Ahmed Alaoui and Michael W Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pages 775–783, 2015.

[Avron *et al.*, 2017] 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.

[Bach, 2013] Francis Bach. Sharp analysis of low-rank kernel matrix approximations. In *Conference on Learning Theory*, pages 185–209, 2013.

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

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

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

[Carratino *et al.*, 2018] 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.

[Chang *et al.*, 2017] 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.

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

[Kriukova *et al.*, 2017] Galyna Kriukova, Sergiy Pereverzyev, and Pavlo Tkachenko. Nyström type subsampling analyzed as a regularized projection. *Inverse Problems*, 33(7):074001, 2017.

[Li *et al.*, 2018] Jian Li, Yong Liu, Rong Yin, Hua Zhang, Lizhong Ding, and Weiping Wang. Multi-class learning: From theory to algorithm. In *Advances in Neural Information Processing Systems 31*, pages 1591–1600, 2018.

[Li *et al.*, 2019a] Jian Li, Yong Liu, Rong Yin, and Weiping Wang. Approximate manifold regularization: Scalable algorithm and generalization analysis. In *IJCAI*, pages 2887–2893, 2019.

[Li *et al.*, 2019b] Zhu Li, Jean-Francois Ton, Dino Oglic, and Dino Sejdinovic. Towards a unified analysis of random fourier features. In *International Conference on Machine Learning*, pages 3905–3914. PMLR, 2019.

[Li *et al.*, 2020] Jian Li, Yong Liu, and Weiping Wang. Automated spectral kernel learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4618–4625, 2020.

[Li, 2021] Zhu Li. Sharp analysis of random fourier features in classification. *arXiv preprint arXiv:2109.10623*, 2021.

[Lin and Cevher, 2018] 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.

[Lin and Cevher, 2020] Junhong Lin and Volkan Cevher. Optimal convergence for distributed learning with stochastic gradient methods and spectral algorithms. *Journal of Machine Learning Research*, 21(147):1–63, 2020.

[Lin and Rosasco, 2016] 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.

[Lin *et al.*, 2017] 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.

[Lin *et al.*, 2020] Shao-Bo Lin, Di Wang, and Ding-Xuan Zhou. Distributed kernel ridge regression with communications. *Journal of Machine Learning Research*, 21(93):1–38, 2020.

[Liu *et al.*, 2021] Yong Liu, Jiankun Liu, and Shuqiang Wang. Effective distributed learning with random features: Improved bounds and algorithms. In *International Conference on Learning Representations*, 2021.

[McMahan *et al.*, 2017] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[Mücke and Blanchard, 2018] Nicole Mücke and Gilles Blanchard. Parallelizing spectrally regularized kernel algorithms. *The Journal of Machine Learning Research*, 19(1):1069–1097, 2018.

[Rahimi and Recht, 2007] 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.

[Rudi and Rosasco, 2017] 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.

[Rudi *et al.*, 2015] 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.

[Rudi *et al.*, 2018] 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.

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

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

[Sun and Wu, 2020] Hongwei Sun and Qiang Wu. Optimal rates of distributed regression with imperfect kernels. *arXiv preprint arXiv:2006.16744*, 2020.

[Sun *et al.*, 2018] Yitong Sun, Anna Gilbert, and Ambuj Tewari. But how does it work in theory? linear svm with random features. In *Advances in Neural Information Processing Systems*, pages 3379–3388, 2018.

[Williams and Seeger, 2001] 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.

[Zhang *et al.*, 2015] 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.