Journal of Machine Learning Research 24 (2023) 1-39 Submitted 9/21; Revised 12/22; Published 4/23

Optimal Convergence Rates for Distributed Nyström Approximation

Jian Li LIJIAN9026@IIE.AC.CN
Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China

Yong Liu* LIUYONGGSAI@RUC.EDU.CN
Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing, China

Weiping Wang WANGWEIPING@IIE.AC.CN
Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China

Editor: Lorenzo Rosasco

Abstract

The distributed kernel ridge regression (DKRR) has shown great potential in processing complicated tasks. However, DKRR only made use of the local samples that failed to capture the global characteristics. Besides, the existing optimal learning guarantees were provided in expectation and only pertain to the attainable case that the target regression lies exactly in the kernel space. In this paper, we propose distributed learning with globally-shared Nyström centers (`DNyström`), which utilizes global information across the local clients. We also study the statistical properties of `DNyström` in expectation and in probability, respectively, and obtain several state-of-the-art results with the minimax optimal learning rates. Note that, the optimal convergence rates for `DNyström` pertain to the non-attainable case, while the statistical results allow more partitions and require fewer Nyström centers. Finally, we conduct experiments on several real-world datasets to validate the effectiveness of the proposed algorithm, and the empirical results coincide with our theoretical findings.

1. Introduction

Kernel methods are one of the most successful approaches to learning complicated patterns via implicit feature mappings and their statistical properties have been well analyzed using statistical learning theory (Vapnik, 1999). For example, using the integral operator theory, researchers have proven the minimax optimal convergence rates for kernel ridge regression (KRR)(Caponnetto and De Vito, 2007; Smale and Zhou, 2007). Despite their excellent theoretical properties, kernel methods are typically unfeasible in large-scale settings due to high training time and storage requirements. To overcome the scalability issues, researchers have developed a wide range of practical algorithms for kernel methods: distributed learning, low-rank approximation algorithms including random features and Nyström method, and stochastic optimization methods. Distributed learning produces a global model after training disjoint subsets on individual machines with necessary communications (Zhang et al., 2015; Lin et al., 2017). Nyström approximation (Williams and Seeger, 2001; Zhang

*. Corresponding author

et al., 2008; Bach, 2013) and random features (Rahimi and Recht, 2007; Rudi and Rosasco, 2017) alleviate memory bottlenecks via low-rank approximation, while stochastic optimization methods (Raskutti et al., 2014; Lin and Cevher, 2018) improve computational efficiency via iterative solutions. The optimal theoretical guarantees for KRR together with accelerated techniques, such as distributed learning (Zhang et al., 2015; Guo et al., 2017; Lin et al., 2017; Chang et al., 2017; Lin and Cevher, 2020), Nyström approximation (Bach, 2013; Alaoui and Mahoney, 2015; Rudi et al., 2015, 2017), random features (Rudi and Rosasco, 2017; Liu et al., 2021) and stochastic optimization (Lin and Cevher, 2018, 2020), have also been established.

Distributed kernel ridge regression (DKRR) is one of the most popular topics in nonparametric statistical learning (Zhang et al., 2015). DKRR has been incorporated with several techniques that can still achieve the same optimal rates as the exact KRR, including random features (Li et al., 2019; Liu et al., 2021), stochastic gradient methods (Lin and Cevher, 2018, 2020), multi-pass SGD (Lin and Cevher, 2018), Nyström approximation (Yin et al., 2020), random sketching (Lian et al., 2021) and multiple communications (Lin et al., 2020). Even though several algorithms were devised and optimal learning properties for DKRR methods were obtained, some problems remain yet to be settled down: 1) DKRR can only characterize local information from local training samples that are not good enough to capture the global characteristics from the entire training samples. 2) The optimal convergence rates for DKRR were derived in expectation that describe the average error rather than the error of a single trial in practice. 3) The optimal theoretical guarantees only apply to the attainable case, assuming the target regression lies exactly in the kernel space. However, the non-attainable case covers many challenging problems and deserves more attention (Lin and Cevher, 2020; Sun and Wu, 2021). 4) The strict restriction on the number of partitions limits the improvements in computational efficiency (Guo et al., 2017; Lin et al., 2017; Lin and Cevher, 2020). There are natural questions whether we can devise a distributed algorithm with one communication that can characterize the global information from all subsets and how to achieve the optimal generalization rates in a high probability that can be applied to the non-attainable case.

1.1 Contributions

In this paper, we propose a distributed Nyström approximation framework, namely `DNyström`, which can make use of global information via the globally-shared Nyström centers that are sampled from the entire training data rather than the local data. Then, we provide the excess risk bounds with the optimal theoretical guarantees for `DNyström` in expectation and in probability, respectively. Specifically, we relax the strict restriction on the number of partitions such that the optimal rates for `DNyström` pertains to both the attainable case and the non-attainable case. We also conduct experiments to explore the impacts of the number of partitions and the number of random centers, respectively. The experimental results verify the superiority of `DNyström` over the compared algorithms.

1) On the algorithmic front: globally-shared Nyström centers. The existing DC-NY (Yin et al., 2020) sampled Nyström centers from local examples that can only use the local information that lacks the global properties of the task. We proposed the globally-shared Nyström centers that contain local information from all clients to improve

the generalization ability of local clients. As shown in Figure 1, the proposed `DNyström` outperforms DC-NY and DKRR owing to the global characteristics from the globally-shared Nyström centers.

2) On the statistical front: applying to the non-attainable case. The theoretical error bounds for KRR methods (Caponnetto and De Vito, 2007; Rudi et al., 2015; Guo et al., 2017) were derived in expectation and assumed the target regression lies in the induced kernel space. However, the expected error bounds only reflect the average error and the target regression is usually out of the induced kernel space for complicated tasks. In this paper, we prove that the optimal theoretical properties of `DNyström` in expectation and in probability, respectively, which apply to both the attainable and non-attainable cases.

3) On the computational front: higher computational efficiency. The classical DKRR (Zhang et al., 2015) and DC-NY (Yin et al., 2020) still suffered from high computational requirements due to the strict constraints on the number of partitions, i.e. a constant $\mathcal{O}(1)$ number of partitions in the general case. Using a finer-grained estimate of the capacity of Hilbert space and novel proof techniques, we improve the number of partitions and thus improve the computational efficiency.

4) Novel proof techniques. Using explicit intermediate estimators, we introduce novel error decompositions for the excess risk bounds in expectation and in probability, respectively. From the error decompositions, one can specifically quantify the errors caused by different components. We also bound the distributed error in the excess risk bound in probability by estimating the difference between empirical and expected covariance operators via contraction inequality for the self-adjoint operators. Moreover, we estimate the Nyström error term in the non-attainable case for the first time.

1.2 Related Work

The related work includes: distributed learning, Nyström approximation, leverage scores sampling and preconditioned conjugate gradient methods (PCG).

1) Distributed learning. Based on certain eigenfunction assumptions, the optimal learning rates for DKRR were first proven in the seminal work (Zhang et al., 2015), and was extended to features space (Wang, 2019). The conventional integral operator theory was applied to DKRR (Lin et al., 2017; Guo et al., 2017) to derive improved error bounds. Using integral operator theory, optimal learning rates for distributed learning with other tools were established, including DKRR with spectral algorithms (Lin and Cevher, 2020), distributed semi-supervised KRR (Chang et al., 2017), DKRR with stochastic gradient methods (Lin and Cevher, 2018, 2020), DKRR with random features (Li et al., 2019; Liu et al., 2021) and DKRR with Nyström approximation (Yin et al., 2020). However, the existing theoretical findings imposed strict conditions on the number of partitions.

2) Nyström approximation. Nyström approximation is a common tool to approximate kernel matrix with low-rank decomposition (Williams and Seeger, 2001; Drineas et al., 2012). The optimal learning guarantees of the combination of KRR and Nyström approximation (KRR-Nyström) were first established in (Rudi et al., 2015) for both uniform sampling and approximate leverage scores sampling. KRR-Nyström was incorporated with PCG to achieve better computational efficiency (Rudi et al., 2017). The analysis was also extended into coefficient-based regularization (Ma et al., 2019) and manifold regularization

(Sivananthan et al., 2020). Recent work (Kriukova et al., 2017; Lu et al., 2019) also studied the low smoothness of Nyström subsample for the misspecified models.

3) Leverage scores sampling. In the classic Nyström method and its variants (Platt, 2005; Bach, 2013), Nyström landmarks are selected uniformly at random. Uniform sampling is fast to compute, but it fails to capture the low-rank nature of the matrix and thus usually requires a larger sampling number of training examples to achieve the specific approximation accuracy. Therefore, researchers proposed data-dependent sampling strategies (Zhang et al., 2008; Kumar et al., 2012; Alaoui and Mahoney, 2015; Gittens and Mahoney, 2016), such that the sampled Nyström landmarks can more closely approximate the kernel matrix than uniform sampling. Leverage scores sampling has been proven strong guarantees for both kernel approximation (Gittens and Mahoney, 2016) and generalization performance (Alaoui and Mahoney, 2015; Rudi et al., 2015). However, the exact leverage scores are prohibitively expensive to compute, and thus recent studies proposed fast leverage scores sampling methods by approximate leverage scores (Musco and Musco, 2017; Calandriello et al., 2017; Rudi et al., 2018; Chen and Yang, 2021). Lee et al. applied leverage scores sampling to neural networks (Lee et al., 2020). There are also many studies on data-dependent sampling for random features with leverage scores.

4) Preconditioned conjugate gradient (PCG). Since the closed-form solutions for KRR-related methods involving the inverse of kernel matrix term exhibit high computational complexity, some iterative methods are proposed to reduce the complexity, for example, conjugate gradient (CG) methods (Hestenes and Stiefel, 1952; Møller, 1993; Hanke, 2017). PCG introduced a suitable preconditioner to obtain a better condition number and thus reduce the number of iterations for iteratively solving the linear system (Saad, 2003). The optimal theoretical guarantees for PCG-based methods have been recently proven, including KRR integrations with sketching (Avron et al., 2017), Nyström (Rudi et al., 2017, 2018), and both divide-and-conquer and Nyström (Yin et al., 2020).

2. Distributed Nyström Approximation

We consider the supervised learning problem of estimating a predictive function from a fixed but unknown distribution ρ over a probability space $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the output space. The training set $D = (\boldsymbol{X}_N, \boldsymbol{y}_N) = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ is drawn i.i.d from $\mathcal{X} \times \mathcal{Y}$ with respect to ρ. For the regression tasks, we assume the input space is $\mathcal{X} = \mathbb{R}^d$ and the output space is $\mathcal{Y} = \mathbb{R}$. We denote $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) (Steinwart and Christmann, 2008) induced by a Mercer kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that

$$\mathcal{H} = \overline{\operatorname{span}\{K_x | \boldsymbol{x} \in \mathcal{X}\}}, \qquad \text{completed with} \quad \langle K_x, K_{\boldsymbol{x}'} \rangle_K = K(\boldsymbol{x}, \boldsymbol{x}') \quad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}.$$

Here, the inner product in $\mathcal{H}$ is denoted as $\langle \cdot, \cdot \rangle_K$ and the corresponding norm $\| \cdot \|_K$. For any two sample vectors $\boldsymbol{a} = (\boldsymbol{a}_1, \cdots, \boldsymbol{a}_p)^\top \in \mathcal{X}^p$ and $\boldsymbol{b} = (\boldsymbol{b}_1, \cdots, \boldsymbol{b}_q)^\top \in \mathcal{X}^q$, we denote $K(\boldsymbol{a}, \boldsymbol{b})$ as the $p \times q$ kernel matrix whose (i, j)-th component is $K(\boldsymbol{a}_i, \boldsymbol{b}_j)$ for $i \in [p]$ and $j \in [q]$.

2.1 Kernel Ridge Regression (KRR) with Nyström Approximation

KRR is a standard nonparametric regression in supervised learning (Vapnik, 1999), which can be stated as

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

where the square loss is used and λ is the regularization parameter. The representer theorem for kernel methods (Schölkopf et al., 2001) illustrates that KRR admits a closed-form solution

$$\widehat{f}_{D,\lambda}(\boldsymbol{x})=\sum_{i=1}^{N}\alpha_i K(\boldsymbol{x}_i,\boldsymbol{x}),\qquad \text{with}\qquad \boldsymbol{\alpha}=(\mathbf{K}_{NN}+\lambda N\mathbf{I})^{-1}\boldsymbol{y}_N, \tag{2}$$

where $\mathbf{K}_{NN}:=K(\boldsymbol{X}_N,\boldsymbol{X}_N)$ is the kernel matrix and $\boldsymbol{y}_N=(y_1,\cdots,y_N)^\top$ is the label vector. Since KRR requires $\mathcal{O}(N^3)$ time to compute the inverse of $\mathbf{K}_{NN}+\lambda N\mathbf{I}$ and $\mathcal{O}(N^2)$ space to store the kernel matrix, it is unfeasible as n increases in the large-scale settings.

Nyström methods replace the empirical kernel matrix with a smaller matrix obtained by subsampling, which is widely used to reduce the memory/time requirements (Williams and Seeger, 2001; Kumar et al., 2012). Specifically, we sample M Nyström landmarks from the rows of the feature matrix $\widetilde{\boldsymbol{X}}_M:=\left(\tilde{\boldsymbol{x}}_1,\cdots,\tilde{\boldsymbol{x}}_M\right)^\top\subseteq\boldsymbol{X}_N$ where $M\leq N$. The approximation solution with Nyström approach for (1) can be written as

$$\widehat{f}^M_{D,\lambda}(\boldsymbol{x})=\sum_{i=1}^{M}\alpha_i K(\widetilde{\boldsymbol{x}}_i,\boldsymbol{x})\qquad\text{with}\qquad\boldsymbol{\alpha}=(\mathbf{K}_{NM}^\top\mathbf{K}_{NM}+\lambda N\mathbf{K}_{MM})^\dagger\mathbf{K}_{NM}^\top\boldsymbol{y}_N, \tag{3}$$

where $\mathbf{H}\dagger$ denotes the Moore-Penrose inverse of the matrix $\mathbf{H}$, and $\mathbf{K}_{NM}:=K(\boldsymbol{X}_N,\widetilde{\boldsymbol{X}}_M)$, $\mathbf{K}_{MM}=K(\widetilde{\boldsymbol{X}}_M,\widetilde{\boldsymbol{X}}_M)$. Using Nyström centers, we solve the closed-form solution in $\mathcal{O}(NM^2)$ time complexity and $\mathcal{O}(NM)$ space complexity.

The sampling strategy for Nyström landmark points $\{\tilde{\boldsymbol{x}}_1,\cdots,\tilde{\boldsymbol{x}}_M\}$ is crucial to the approximation ability of Nyström methods. We introduce two popular sampling strategies.

- **Nyström landmarks with uniform sampling (Bach, 2013).** Let the Nyström centers $\{\tilde{\boldsymbol{x}}_1,\cdots,\tilde{\boldsymbol{x}}_M\}$ be uniformly sampled from the training set $\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N\}$.
- **Nyström landmarks with leverage scores sampling (Alaoui and Mahoney, 2015)**. Let the random Nyström centers $\{\tilde{\boldsymbol{x}}_1,\cdots,\tilde{\boldsymbol{x}}_M\}$ be selected according to the probability $p_i=\widehat{l}_\lambda(i)/\sum_{i=1}^N\widehat{l}_\lambda(i)$ where the λ-ridge leverage scores of $\boldsymbol{x}_i$ is defined as

$$l_i^\lambda(\mathbf{K}_{NN})=\left(\mathbf{K}_{NN}(\mathbf{K}_{NN}+\lambda N\mathbf{I})^{-1}\right)_{ii},\qquad\forall\ i\in[N]. \tag{4}$$

2.2 Distributed Nyström Approximation (DNyström)

DKRR directly averaged the local KRR solutions on local clients that only used the limited information from the local data and ignore the global characteristics from the entire data. To utilize the global information from other clients, we present a distributed Nyström

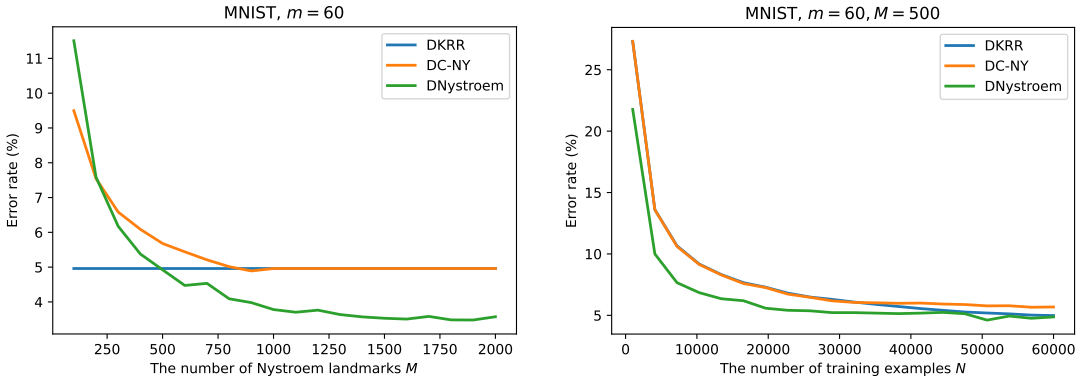

Figure 1: `DNyström` is compared to DKRR (Zhang et al., 2015) and DC-NY (Yin et al., 2020) with respect to the error rate vs. the number of Nyström landmarks (left) and the number of training examples (right) on the MNIST dataset (60000 examples).

approximation (`DNyström`) with globally-shared Nyström centers in Algorithm 1. We first sample M Nyström landmarks from the entire training examples $\widetilde{\boldsymbol{X}}_M := \left(\tilde{\boldsymbol{x}}_1, \cdots, \tilde{\boldsymbol{x}}_M\right)^\top \subseteq \boldsymbol{X}_N$, and we send Nyström landmarks to all clients. Then, we separate the training set into m disjoint training subsets uniformly such that $D = \bigcup_{j=1}^m D_j$ where $D_j = \left(\boldsymbol{X}_j, \boldsymbol{y}_j\right)$ and send them to their corresponding clients.

We consider the divide-and-conquer framework (only once communication), where the global solution is the average of local ones. We define `DNyström` as

$$\bar{f}_{D,\lambda}^M(\boldsymbol{x}) = \sum_{i=1}^M \alpha_i K(\widetilde{\boldsymbol{x}}_i, \boldsymbol{x}), \qquad \text{with} \quad \boldsymbol{\alpha} = \sum_{j=1}^m \frac{|D_j|}{|D|} \boldsymbol{\beta}_j \tag{5}$$

and the local weight on the j-th partition is given by

$$\boldsymbol{\beta}_j = (\mathbf{K}_{jM}^\top \mathbf{K}_{jM} + \lambda |D_j| \mathbf{K}_{MM})^\dagger \mathbf{K}_{jM}^\top \boldsymbol{y}_j, \tag{6}$$

where $\mathbf{K}_{jM} = K(\boldsymbol{X}_j, \widetilde{\boldsymbol{X}}_M) \in \mathbb{R}^{|D_j| \times M}$, $\boldsymbol{y}_j = (y_1, \cdots, y_{|D_j|})^\top$, $\boldsymbol{\beta}_j \in \mathbb{R}^M$ and $\boldsymbol{\alpha} \in \mathbb{R}^M$. For the sake of simplification, we assume the entire training set be partitioned equally, i.e. $|D_j| = N/m$. Since the local linear systems (6) can be solved in parallel, the time and space complexities of `DNyström` are $\mathcal{O}(NM^2/m + M^3)$ and $\mathcal{O}(NM/m)$, respectively. Compared to DKRR and DC-NY, the globally-shared Nyström centers introduce additional communication burden, but it dose not dominate the communication complexity when $M \leq N/m$.

Remark 1 (Global information from globally-shared Nyström centers) *To accelerate the local computation in DKRR, DC-NY (Yin et al., 2020) sampled Nyström landmarks from local training examples $\widetilde{\boldsymbol{X}}_M = \left(\tilde{\boldsymbol{x}}_1, \cdots, \tilde{\boldsymbol{x}}_M\right)^\top \subseteq \boldsymbol{X}_j$ on the j-th client. Therefore, the computation of local Nyström approximation in DC-NY can only make use of local information from local training examples. In this paper, `DNyström` generated Nyström centers from the entire training examples $\widetilde{\boldsymbol{X}}_M \subseteq \boldsymbol{X}_N$ that are globally-shared across all devices.*

Algorithm 1 Distributed Nyström approximation (`DNyström`)

Require: Labeled training dataset $D = (\boldsymbol{X}_N, \boldsymbol{y}_N)$, kernel function $K(\cdot, \cdot)$, regularization parameter λ, sampling scale M, the number of local clients m, and the sampling probability $\{p_i\}_{i=1}^N$.

Ensure: Model coefficients $\boldsymbol{\alpha}$

1: Sample Nyström landmarks $\widetilde{\boldsymbol{X}}_M \subseteq \boldsymbol{X}_N$ according to the sampling probability $\{p_i\}_{i=1}^N$ and send $\widetilde{\boldsymbol{X}}_M$ to all local clients.
2: Randomly separate the training set into m disjoint subsets $D = \bigcup_{j=1}^m D_j$ and send $D_j = \big(\boldsymbol{X}_j, \boldsymbol{y}_j\big)$ to the corresponding the j-th client.
3: **In parallel: on the j-th client** $\forall j \in [m]$
4: Compute kernel matrices $\mathbf{K}_{jM} = K(\boldsymbol{X}_j, \widetilde{\boldsymbol{X}}_M)$ and $\mathbf{K}_{MM} = K(\widetilde{\boldsymbol{X}}_M, \widetilde{\boldsymbol{X}}_M)$.
5: Solve the linear system (6) according to *the solver* and obtain local coefficients $\boldsymbol{\beta}_j$.
6: Send $\boldsymbol{\beta}_j$ to the global server.
7: **End parallelism**
8: Average the local model coefficients $\boldsymbol{\alpha} = \sum_{i=1}^m \frac{|D_j|}{|D|}\, \boldsymbol{\beta}_j$.

Compared to DKRR (Zhang et al., 2015) and DC-NY (Yin et al., 2020), `DNyström` *can employ the global information from other clients and be used to tackle statistical heterogeneity in the federated learning scenario. Specifically, DC-NY can only sample $M \leq |D_j|$ Nyström centers on the j-th device while* `DNyström` *allows a larger number of Nyström landmarks. As shown in the left of Figure 1, when we fixed the local sample size $n = N/m = 1000$ and vary M,* `DNyström` *leads to lower error rates than DC-NY if $M \geq \frac{1}{4}n$ and leads to lower error rates than DKRR if $M \geq \frac{1}{2}n$. From the right of Figure 1, as the number of training examples increases, the error rates of all methods decrease but* `DNyström` *leads to lower errors (especially when N is small). Therefore,* `DNyström` *outperforms both DKRR and DC-NY owing to the characterization of global information from the globally-shared Nyström centers.*

The proposed `DNyström` is a flexible framework computed in parallel, which can be incorporated with data-dependent sampling strategies (Alaoui and Mahoney, 2015; Musco and Musco, 2017; Rudi et al., 2018; Chen and Yang, 2021) and iterative methods (Shalev-Shwartz et al., 2011; Rudi et al., 2017; Ma and Belkin, 2017, 2019). These integrations can further improve the computational efficiency, but data-dependent sampling requires additional sample complexity due to the computation of leverage scores (4) and stochastic optimization algorithms introduce optimization error. Since these integrations are orthogonal to `DNyström`, for the sake of simplification, we focus on the closed-form solution of `DNyström` in (6) and discuss the possible integrations.

Remark 2 (Integration with gradient methods) *While the linear system* (6) *can be solved by a direct closed-form solution, the computational requirements are related to the number of examples, making it impractical for large-scale data. Gradient descent methods decouple the kernel model from the scale of the training set by iteratively solving the linear systems. Popular gradient descent kernel methods include SDCA (Hsieh et al., 2008), Pe-*

gasos (Shalev-Shwartz et al., 2011), FALKON (Rudi et al., 2017) and EigenPro (Ma and Belkin, 2017, 2019). These methods improve the computational efficiency with early stopping and enable efficient GPU implementations (Rudi et al., 2018; Ma and Belkin, 2019). For example, we can incorporate `DNyström` *with preconditioned conjugate gradient methods (Rudi et al., 2017) or preconditioned stochastic gradient methods (Ma and Belkin, 2019), which requires* $\mathcal{O}(NMt/m + M^3)$ *time and* $\mathcal{O}(NM/m)$ *space. Specifically, (Rudi et al., 2017) proved that* $t = \Omega(\log(N))$ *iterations guarantee good approximation between the PCG solution and* (6)*.*

Remark 3 (Integration with data-dependent sampling) *Statistical leverage scores that measure the matrix coherence have also proved crucial recently in the development of improved worst-case randomized matrix algorithms (Drineas et al., 2012). The exact leverage scores* (4) *are prohibitively expensive to compute, consuming* $\mathcal{O}(N^3)$ *time. To accelerate the computation of leverage scores, researchers have proposed several approximate leverage scores algorithms, including recursive sampling (Musco and Musco, 2017), SQUEAK (Calandriello et al., 2017), BLESS (Rudi et al., 2018) and spectral analysis (Chen and Yang, 2021). For example, the sampling complexity of BLESS is reduced from* $\mathcal{O}(N^3)$ *to* $\mathcal{O}(M^2/\lambda)$ *where* λ *is the regularization parameter in KRR. When we set* $\lambda = M/N$*, the complexity of BLESS would be* $\mathcal{O}(NM)$*.*

3. Main Results

In this section, we focus on the generalization properties of the closed-form solutions of `DNyström`. We first recover the existing bounds for DC-NY in expectation (Yin et al., 2020), which only pertains to the attainable case. We then present our theoretical results for `DNyström` in expectation and in high probability, respectively. Specifically, we prove the minimax optimal convergence rates for `DNyström` under certain constraints including the allowed number of partitions and the required number of Nyström centers in both expectation and probability. The optimal theoretical guarantees for `DNyström` in expectation and high probability apply to the non-attainable case, respectively.

The ideal learning target of KRR is to find a predictor that minimizes the expected risk $\min_{f\in\mathcal{F}} \mathcal{E}(f), \quad \mathcal{E}(f) = \int_{\mathcal{X}\times\mathcal{Y}}(f(\boldsymbol{x}) - y)^2 d\rho(\boldsymbol{x}, y)$, where $\mathcal{F}$ is the class of all measurable functions from $\mathcal{X}$ to $\mathcal{Y}$ and ρ is the joint probability distribution over $\mathcal{X} \times \mathcal{Y}$. The target regression that minimizes the expected risk over all measurable functions $f : \mathcal{X} \to \mathbb{R}$ is

$$f_\rho(\boldsymbol{x}) = \int_{\mathcal{Y}} y d\rho(y|\boldsymbol{x}), \qquad \forall \boldsymbol{x} \in \mathcal{X}. \tag{7}$$

Here, f_ρ is the true regression without noise labels and belongs to the Hilbert space of square integral functions $L^2_{\rho_X} = \{f : \mathcal{X} \to \mathbb{R} \mid \|f\|_\rho^2 = \int |f(\boldsymbol{x})|^2 d\rho_X < \infty\}$ with respect to the marginal distribution ρ_X of ρ on $\mathcal{X}$, where the $L^2_{\rho_X}$-norm is defined as $\|f\|_\rho^2 = \langle f, f\rangle_\rho = \int_X |f(\boldsymbol{x})|^2 d\rho_X(\boldsymbol{x})$, $\forall f \in L^2_{\rho_X}$. Notably, since the joint probability distribution ρ is unknown, we employ the solution from empirical risk minimization (ERM) (6) to approximate the target function f_ρ, and we investigate the generalization gaps between ERM solution and f_ρ. The generalization ability of a regression estimator $f \in L^2_{\rho_X}$ is measured by the *excess risk*, i.e. $\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|_\rho^2$ (Smale and Zhou, 2007). Throughout this paper, we

assume the outputs are bounded almost surely for some constant $B > 0$ and $\mathcal{X}$ is compact, which implies $\|f_\rho\|_\infty \leq B$ and $|y| \leq B$. We also assume $K(\boldsymbol{x}, \boldsymbol{x}) \leq \kappa^2 < \infty$ for any $\boldsymbol{x} \in \mathcal{X}$.

Definition 4 (Effective dimension) *The integral operator and covariance operator are defined as*

$$L : L^2_{\rho_X} \to L^2_{\rho_X}, \quad (Lf)(\cdot) = \int_X K(\boldsymbol{x}, \cdot) f(\boldsymbol{x}) d\rho_X(\boldsymbol{x}), \qquad \forall\ f \in L^2_{\rho_X}(X, \rho_X)$$
$$C : \mathcal{H} \to \mathcal{H}, \qquad \langle h, Cg \rangle = \int_X h(\boldsymbol{x}) g(\boldsymbol{x}) d\rho_X(\boldsymbol{x}), \qquad \forall\ g, h \in \mathcal{H}.$$

For $\lambda > 0$, we define the random variable $\mathcal{N}_{\boldsymbol{x}}(\lambda) = \langle K_x, (C + \lambda I)^{-1} K_x \rangle$ with $\boldsymbol{x} \in \mathcal{X}$ drawn from ρ_X. Finally we define the quantities $\mathcal{N}(\lambda) = \mathbb{E}\, \mathcal{N}_{\boldsymbol{x}}(\lambda)$, $\mathcal{N}_\infty(\lambda) = \sup_{\boldsymbol{x} \in \mathcal{X}} \mathcal{N}_{\boldsymbol{x}}(\lambda)$.

The effective dimension $\mathcal{N}(\lambda) = \mathrm{Tr}(C(C + \lambda I)^{-1}) = \mathrm{Tr}(L(L + \lambda I)^{-1})$ measures the average capacity of RKHS $\mathcal{H}$, while $\mathcal{N}_\infty(\lambda)$ measures the maximal capacity of RKHS.

Assumption 5 (Regularity assumption) *Assume there exists $R > 0$, $r > 0$, and $g \in L^2_{\rho_X}$, such that*

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

where $\|g\|_\rho \leq R$ 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 regularity assumption is also called source condition, where the value of r measures the regularity of f_ρ. Let $\overline{\mathcal{H}}$ be the closure of $\mathcal{H}$ in $L^2_{\rho_X}$, and then the condition $r = \frac{1}{2}$ means the existence $f_H = L^{\frac{1}{2}} g \in \mathcal{H}$ such that $f_H = f_\rho$ and $\overline{\mathcal{H}} = L^2_{\rho_X}$. Since the fact $L^r(L^2_{\rho_X}) \subseteq L^{r'}(L^2_{\rho_X})$ if $r \geq r'$, the smaller r corresponds to the larger subspace where the target regression lies. More examples refers to (Lin and Cevher, 2020; Sun and Wu, 2021). The case $r \in (0, 1/2)$ is the non-attainable case, where $f_\rho \notin \mathcal{H}$ and the learning tasks are difficult, while the case $r \in [1/2, 1]$ is the attainable case, corresponding to $f_\rho \in \mathcal{H}$. If $r > 1$, the convergence rates of DKRR are same as $r = 1$ duo to the saturation phenomenon in DKRR (Zhang et al., 2015; Lin et al., 2017; Lin and Cevher, 2020; Sun and Wu, 2021) The conventional optimal generalization analysis for KRR focused on the attainable case $r \in [1/2, 1]$ (Caponnetto and De Vito, 2007; Rudi et al., 2015; Guo et al., 2017). In this paper, we extend the source condition from $r \in [1/2, 1]$ to $r \in (0, 1]$ that also covers the special case $r > 1$ due to the saturation effect.

Assumption 6 (Capacity assumption) *There exists $C_0 > 0$ and $\gamma \in (0, 1]$, such that*

$$\mathcal{N}(\lambda) \leq C_0 \lambda^{-\gamma},$$

where C_0 is a constant independent of λ.

Assumption 7 (Compatibility assumption) *Assume there exists $\alpha \in [\gamma, 1]$ and $C_1 > 0$, such that*

$$\mathcal{N}_\infty(\lambda) \leq C_1 \lambda^{-\alpha},$$

where C_1 is a constant independent of λ.

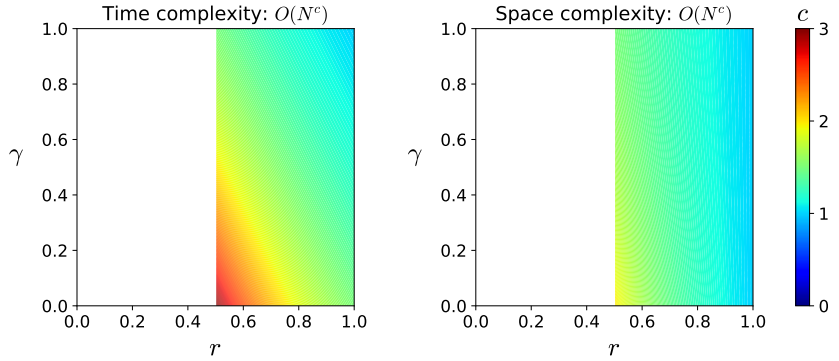

Figure 2: Time complexity and space complexity of Proposition 8. The color closer to red represents higher complexity. Blank areas represent unfeasible situations. The time complexity is $\mathcal{O}\big(N^{\frac{3+\gamma}{2r+\gamma}}\big)$ and space complexity is $\mathcal{O}\big(N^{\frac{2+\gamma}{2r+\gamma}}\big)$.

Note that, the effective dimension $\mathcal{N}(\lambda)$ provides an measure of the average capacity of $\mathcal{H}$ while the quantity $\mathcal{N}_\infty(\lambda)$ considers the worst case. Assumption 6 is ensured if the eigenvalues of the covariance operator C exhibit a polynomial decay $\sigma_i \lesssim i^{-\frac{1}{\gamma}}$ (Caponnetto and De Vito, 2007; Rudi et al., 2015). Since the covariance operator C is a trace class, Assumptions 6-7 are always satisfied with $\gamma = \alpha = 1$. Specifically, if the kernel is bounded $\sup_{\boldsymbol{x}\in\mathcal{X}} K(\boldsymbol{x},\boldsymbol{x}) \leq \kappa^2$, the effective dimensions are upper bounded by $\mathcal{N}(\lambda) \leq \mathcal{N}_\infty(\lambda) = \sup_{\boldsymbol{x}\in\mathcal{X}} \langle K_x, (C+\lambda I)^{-1} K_x\rangle \leq \kappa^2/\lambda$. To obtain a fine-grained estimate for $\mathcal{N}_\infty(\lambda)$, Rudi and Rosasco introduced compatibility assumption $\mathcal{N}_\infty(\lambda) = \mathcal{O}(\lambda^{-\alpha})$ for random features (Rudi and Rosasco, 2017), where $\gamma \leq \alpha \leq 1$. Note that, $\mathcal{N}_\infty(\lambda) \lesssim \lambda^{-\alpha}$ is slightly stronger than the basic condition $\mathcal{N}_\infty(\lambda) \lesssim \lambda^{-1}$ but reasonable. The value γ reflects the size of RKHS $\mathcal{H}$, whereas a larger γ corresponds to a larger RKHS. The case $\gamma = 1$ is capacity-independent case and the effective dimension saturates when $\gamma > 1$ as $i^{-1} < i^{-\frac{1}{\gamma}}$ for any $\gamma > 1$. The capacity assumption is standard for the generalization analysis of KRR algorithms (Caponnetto and De Vito, 2007; Rudi et al., 2015; Guo et al., 2017) while the compatibility assumption was proposed to obtain a fine-grained analysis for random features (Rudi and Rosasco, 2017).

The worst case is $\alpha = 1$ with the uniform sampling and the benign case is $\alpha = \gamma$ when $\mathcal{N}_\infty(\lambda)$ is close to $\mathcal{N}(\lambda)$ with the data-dependent sampling. Following Example 2 of (Rudi and Rosasco, 2017), one can obtain the favorable situation $\alpha = \gamma$ when the Nyström centers are sampled according to the probability $q(\boldsymbol{x}) = \mathcal{N}_x(\lambda)/\mathcal{N}(\lambda)$. Intuitively, the leverage score $l_i^\lambda(\mathbf{K}_{NN})$ is the empirical version of the probability $q(\boldsymbol{x})$ given the training sample $\boldsymbol{X}_N$.

3.1 Existing Results for DC-NY in Expectation

We recall the theoretical results for the combination of DKRR and Nyström approximation (DC-NY) (Yin et al., 2020), where Nyström centers was sampled from local data $\widetilde{\boldsymbol{X}}_M \subseteq \boldsymbol{X}_j$ to approximate local kernel matrix $K(\boldsymbol{X}_j, \boldsymbol{X}_j) \approx K(\boldsymbol{X}_j, \widetilde{\boldsymbol{X}}_M) K(\widetilde{\boldsymbol{X}}_M, \widetilde{\boldsymbol{X}}_M)^\dagger K(\widetilde{\boldsymbol{X}}_M, \boldsymbol{X}_j)$.

Proposition 8 (Theorem 1 in (Yin et al., 2020)) *Assume that* $f_{\mathcal{H}} = \arg\min_{f\in\mathcal{H}} \mathcal{E}(f)$. *Under Assumptions 5-6, if* $\lambda = N^{-\frac{1}{2r+\gamma}}$, *then the following conditions*

$$r \in [1/2, 1], \qquad \gamma \in (0, 1], \qquad M \gtrsim N^{\frac{1}{2r+\gamma}}, \qquad m \lesssim N^{\frac{2r-1}{2r+\gamma}}$$

are sufficient to guarantee the optimal rates in expectation with a high probability, that

$$\mathbb{E}\,\|\hat{f}^m_{D,\lambda} - f_{\mathcal{H}}\|^2_\rho = \mathcal{O}\left(N^{-\frac{2r}{2r+\gamma}}\right).$$

Here, $\hat{f}^m_{D,\lambda}$ *is the estimator of DC-NY and* $f_{\mathcal{H}}$ *minimizes the expected risk in RKHS.*

We use the notations $a_1 = \mathcal{O}(a_2)$ and $a_1 \lesssim a_2$ to represent $a_1 \leq ca_2$ for some positive constant c, while $a_1 \gtrsim a_2$ means $a_1 \geq ca_2$. The learning rate $\mathcal{O}\left(N^{\frac{-2r}{2r+\gamma}}\right)$ is optimal in a minimax sense (Caponnetto and De Vito, 2007), which is the same rate as the exact KRR. Note that, a minimax lower rate of convergence has been proved in Theorem 2 of (Caponnetto and De Vito, 2007). For the sake of comparison, we leave out the PCG term in (Yin et al., 2020), where the condition on iterations is the same as (Rudi et al., 2017).

We depict the computational complexities of Proposition 8 in Figure 2. We find that 1) The optimal rates only apply to the attainable case $r \in [1/2, 1]$ due to the restriction on the number of partitions $m \lesssim N^{\frac{2r-1}{2r+\gamma}}$; 2) The error bounds were derived in expectation that capture the average error but fail to measure the error of one trial; 3) In the general case ($r = 1/2, \gamma = 1$), DC-NY leads to a constant number $\mathcal{O}(1)$ partitions, degrading to KRR with Nyström approach (Rudi et al., 2015), with $\mathcal{O}(N^2)$ time and $\mathcal{O}(N^{1.5})$ space; 4) DC-NY only considered uniform sampling, ignoring more efficient data-dependent sampling strategies.

3.2 Optimal Convergence Rates for `DNyström` in Expectation

We analyze the generalization performance of `DNyström` in expectation. Note that, Proposition 8 requires the existence of $f_{\mathcal{H}} = \arg\min_{f\in\mathcal{H}} \mathcal{E}(f)$ where $f \in \mathcal{H}$, while we remove this strict condition and extend the analysis to the non-attainable where $f \notin \mathcal{H}$.

Using the representer theorem, there is a reduced RKHS with Nyström approximation:

$$\mathcal{H}_M = \left\{ f \in \mathcal{H} \mid f(\boldsymbol{x}) = \sum_{i=1}^{M} \alpha'_i K(\tilde{\boldsymbol{x}}_i, \boldsymbol{x}),\ \boldsymbol{\alpha}' \in \mathbb{R}^M \right\},$$

where $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^M$ is the subset of inputs in training set.

Definition 9 *On any training set D_j, we define the following estimators*

$$\widehat{f}^M_{D_j,\lambda} = \arg\min_{f\in\mathcal{H}_M}\left\{\frac{1}{|D_j|}\sum_{i=1}^{|D_j|}(\langle f, K_{\boldsymbol{x}_i}\rangle - y_i)^2 + \lambda\|f\|_K^2\right\}, \quad (\boldsymbol{x}_i, y_i)\in D_j,$$

$$\widetilde{f}^M_{D_j,\lambda} = \arg\min_{f\in\mathcal{H}_M}\left\{\frac{1}{|D_j|}\sum_{i=1}^{|D_j|}(\langle f, K_{\boldsymbol{x}_i}\rangle - f_\rho(\boldsymbol{x}_i))^2 + \lambda\|f\|_K^2\right\}, \quad (\boldsymbol{x}_i, y_i)\in D_j,$$

$$\widetilde{f}_{D_j,\lambda} = \arg\min_{f\in\mathcal{H}}\left\{\frac{1}{|D_j|}\sum_{i=1}^{|D_j|}(\langle f, K_{\boldsymbol{x}_i}\rangle - f_\rho(\boldsymbol{x}_i))^2 + \lambda\|f\|_K^2\right\}, \quad (\boldsymbol{x}_i, y_i)\in D_j,$$

$$f_\lambda = \arg\min_{f\in\mathcal{H}}\left\{\int_X(\langle f, K_{\boldsymbol{x}}\rangle - f_\rho(\boldsymbol{x}))^2 d\rho_X(\boldsymbol{x}) + \lambda\|f\|_K^2\right\}.$$

Similarly, we also denote $\widetilde{f}^M_{D,\lambda}$ and $\widetilde{f}_{D,\lambda}$ as the counterparts of $\widetilde{f}^M_{D_j,\lambda}$ and $\widetilde{f}_{D_j,\lambda}$ on the entire dataset D, respectively. The estimator of the proposed `DNyström` is $\bar{f}^M_{D,\lambda} = \sum_{j=1}^m \frac{|D_j|}{|D|}\widehat{f}^M_{D_j,\lambda}$.

Lemma 10 (Error decomposition for `DNyström` in expectation) *Using the estimators defined in Definition 9, if $|D_j| = |D|/m,\ \forall j\in[m]$, we have*

$$\frac{1}{4}\,\mathbb{E}\|\bar{f}^M_{D,\lambda} - f_\rho\|_\rho^2 \le \frac{1}{m}\underbrace{\|\widehat{f}^M_{D_j,\lambda} - \widetilde{f}^M_{D_j,\lambda}\|_\rho^2}_{\textit{Sample variance}} + \underbrace{\|\widetilde{f}^M_{D_j,\lambda} - \widetilde{f}_{D_j,\lambda}\|_\rho^2}_{\textit{Nyström error}} + \underbrace{\|\widetilde{f}_{D_j,\lambda} - f_\lambda\|_\rho^2}_{\textit{Empirical error}} + \underbrace{\|f_\lambda - f_\rho\|_\rho^2}_{\textit{Approximation error}}. \tag{8}$$

The above error decomposition employs intermediate estimators with explicit definitions. From that, one can identify the source of error terms, including local sample variance $\|\widehat{f}^M_{D_j,\lambda} - \widetilde{f}^M_{D_j,\lambda}\|_\rho^2$ from noisy labels, local Nyström error $\|\widetilde{f}^M_{D_j,\lambda} - \widetilde{f}_{D_j,\lambda}\|_\rho^2$ resulted from Nyström approximation, local empirical error $\|\widetilde{f}_{D_j,\lambda} - f_\lambda\|_\rho^2$ from empirical examples drawn w.r.t ρ, and the approximation error (bias) $\|f_\lambda - f_\rho\|_\rho^2$.

Theorem 11 (Excess risk bound for `DNyström` in expectation) *Let $\delta > 0$, $\lambda = N^{\frac{-1}{2r+\gamma}}$ and $|D_1| = \cdots = |D_m| = N/m$. Under Assumptions 5-7, if $\lambda = N^{-\frac{1}{2r+\gamma}}$,*

$$r\in(0,1], \qquad \gamma\in(0,1], \qquad 2r+\gamma\ge\alpha, \qquad m\lesssim N^{\frac{2r+\gamma-\alpha}{2r+\gamma}},$$

$M \gtrsim N^{\frac{\alpha}{2r+\gamma}}$ for the uniform sampling, and $M\gtrsim N^{\frac{\gamma}{2r+\gamma}}$ for the data-dependent sampling, then with probability $1-4\delta$, there exists

$$\mathbb{E}\|\bar{f}^M_{D,\lambda} - f_\rho\|_\rho^2 \lesssim N^{\frac{-2r}{2r+\gamma}}\ \log^2(2/\delta).$$

Here, $\bar{f}^M_{D,\lambda}$ is the estimator of `DNyström` (5) and f_ρ is the true regression.

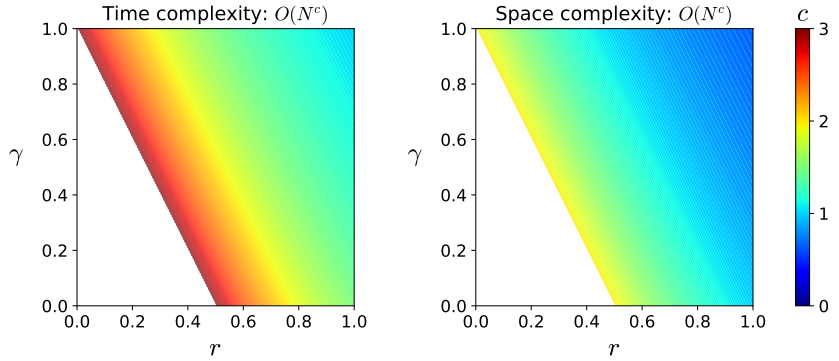

Figure 3: Computational complexities of Corollary 12 with $\alpha = 1$. The average time complexity is $\mathcal{O}(N^{\frac{3}{2r+\gamma}})$ and space complexity is $\mathcal{O}(N^{\frac{2}{2r+\gamma}})$.

Compared with the existing work in DKRR (Guo et al., 2017; Lin et al., 2017; Lin and Cevher, 2020), Nyström approximation (Rudi et al., 2015) and DC-NY (Yin et al., 2020), we relax the strict restriction on the number of partitions from $m \lesssim N^{\frac{2r-1}{2r+\gamma}}$ to $m \lesssim N^{\frac{2r+\gamma-\alpha}{2r+\gamma}}$, leading to two improvements: 1) On the computational front, Theorem 11 guarantees `DNyström` allows more partitions than DC-NY and thus higher computational efficiency; 2) On the theoretical front, beyond the attainable case $r \in [1/2, 1]$ assuming $f_\rho \in \mathcal{H}$, the optimal learning guarantees in Theorem 11 also apply to the non-attainable cases $r \in (0, 1]$ with the restriction $2r+\gamma \geq \alpha$. Meanwhile, the required number of Nyström centers is reduced from $M \gtrsim N^{\frac{1}{2r+\gamma}}$ in DC-NY (Yin et al., 2020) to $M \gtrsim N^{\frac{\alpha}{2r+\gamma}}$ for uniform sampling and $M \gtrsim N^{\frac{\gamma}{2r+\gamma}}$ for data-dependent sampling.

Corollary 12 (The worst case $\alpha = 1$) *Under Assumptions 5-6 and the same settings as Theorem 11, with the uniform sampling, if*

$$r \in (0, 1], \qquad \gamma \in (0, 1], \qquad 2r + \gamma \geq 1, \qquad m \lesssim N^{\frac{2r+\gamma-1}{2r+\gamma}}, \qquad M \gtrsim N^{\frac{1}{2r+\gamma}},$$

then with a high probability, the proposed `DNyström` achieves the optimal rates in expectation.

Without Assumption 7, we consider the worst case that $\alpha = 1$ due to $\mathcal{N}_\infty(\lambda) \leq \kappa^2/\lambda$, which was also used in Nyström methods with uniform sampling (Bach, 2013; Yin et al., 2020). We report the computational complexities and applicable area for the worst case in Figure 3. Compared to the existing results for DC-NY (Yin et al., 2020), the computational efficiency and the applicable area of Corollary 12 are much better. Note that, since $M \approx N/m$ in the worst case, DC-NY method needs to sample all the local examples as Nyström centers that degrades to the exact DKRR, while `DNyström` still works even when $M \geq N/m$.

Corollary 13 (The benign case $\alpha = \gamma$) *Under Assumptions 5-7 and the same settings as Theorem 11, with the uniform sampling, if*

$$r \in (0, 1], \qquad \gamma \in (0, 1], \qquad m \lesssim N^{\frac{2r}{2r+\gamma}}, \qquad M \gtrsim N^{\frac{\gamma}{2r+\gamma}},$$

then with a high probability, the proposed `DNyström` achieves the optimal rates in expectation.

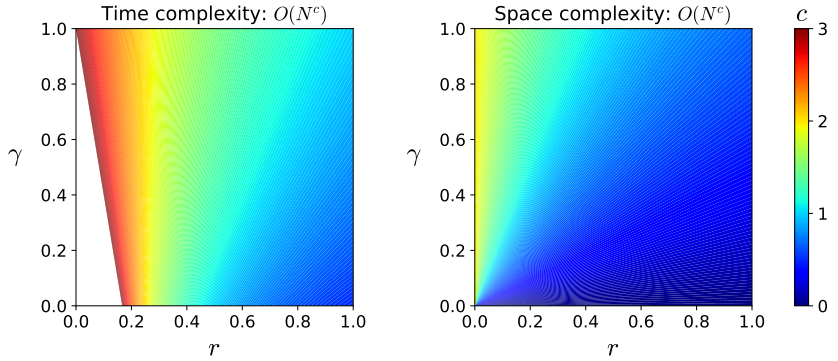

Figure 4: Computational complexities of Corollary 13 with $\alpha = \gamma$. The time complexity is $\mathcal{O}(N^{\frac{1+2\gamma}{2r+\gamma}})$ and space complexity is $\mathcal{O}(N^{\frac{2\gamma}{2r+\gamma}})$. The applicable area for Corollary 13 is $r \in (0, 1]$ and $\gamma \in [0, 1]$, and we clip the time complexities that are bigger than $\mathcal{O}(N^3)$ that is the time complexity of the exact KRR.

In the benign case, $\mathcal{N}_\infty(\lambda)$ is close to $\mathcal{N}(\lambda)$ with the data-dependent sampling. For example, Example 2 of (Rudi and Rosasco, 2017) devised an ideal example to guarantee $\alpha = \gamma$ for random features. When the Nyström centers are sampled according to the probability $q(\boldsymbol{x}) = \mathcal{N}_x(\lambda)/\mathcal{N}(\lambda)$, one can obtain $\alpha = \gamma$ for Nyström approximation. Intuitively, the leverage score $l_i^\lambda(\mathbf{K}_{NN})$ is the empirical version of the probability $q(\boldsymbol{x})$.

Data-dependent sampling introduces additional sampling complexity, for example, BLESS consumes $\mathcal{O}(\widetilde{\mathcal{N}}(\lambda)^2/\lambda) = \mathcal{O}(N^{\frac{1+2\gamma}{2r+\gamma}})$ time to compute approximate leverage scores, which is bigger than the computation of the direct closed-solution of `DNyström` $\mathcal{O}(N^{\frac{3\gamma}{2r+\gamma}})$ and thus dominates the entire time complexity. As shown in Figure 4, the benign case applies to the entire range of the source condition and leads to much higher computational efficiency. The time complexity is smaller than $\mathcal{O}(N^2)$ when $r > 1/4$.

Remark 14 (The combination of `DNyström` and PCG) *We can use PCG to accelerate the solve local closed-form solution* (6) *for* ***`DNyström`****, which consumes* $\mathcal{O}(NMt/m+M^3)$ *time. Based on Theorem 3 of (Rudi et al., 2017), the number of iterations is* $t = \Omega(\log N)$ *for the combination. For uniform sampling, the time complexity of* ***`DNyström`*** *with PCG is* $\mathcal{O}(N^{\frac{2}{2r+\gamma}} \log N + N^{\frac{3}{2r+1}})$*, which has no improvement compared to* ***`DNyström`****. For leverage scores sampling, since the computation of approximate leverage scores dominates the time complexity and PCG is irrelevant to the leverage scores sampling, the time complexity is still* $\mathcal{O}(N^{\frac{1+2\gamma}{2r+\gamma}})$*. Therefore, PCG cannot further reduce the time complexity for* ***`DNyström`*** *with either uniform sampling or leverage scores sampling.*

3.3 Optimal Convergence Rates for `DNyström` in Probability

In Theorem 11, the optimal rates for `DNyström` are in expectation that describe the average error, but fail to quantify the generalization performance of `DNyström` in a single trail. Therefore, we analyze the error decomposition and learning rates for `DNyström` in probability.

Lemma 15 (Error decomposition for `DNyström` in probability) *Let $\widehat{f}^M_{D_j,\lambda}, \widehat{f}^M_{D,\lambda}, \widetilde{f}_{D,\lambda}$ and f_λ be defined in Definition 9. The following error decomposition holds for* ***`DNyström`***

$$\|\bar{f}^M_{D,\lambda} - f_\rho\|_\rho \leq \underbrace{\left\|\bar{f}^M_{D,\lambda} - \widehat{f}^M_{D,\lambda}\right\|_\rho}_{\textit{Distributed error}} + \underbrace{\left\|\widehat{f}^M_{D,\lambda} - f_\lambda\right\|_\rho}_{\textit{Global variance}} + \underbrace{\left\|f_\lambda - f_\rho\right\|_\rho}_{\textit{Approximation error}} . \tag{9}$$

Here, the distributed error can be bounded by

$$\left\|\bar{f}^M_{D,\lambda} - \widehat{f}^M_{D,\lambda}\right\|_\rho \leq 4\left\|C_\lambda^{-1/2}(C - \widehat{C}_{D_j})C_\lambda^{-1/2}\right\| \underbrace{\left\|C_\lambda^{1/2}\ (\widehat{f}^M_{D_j,\lambda} - f_\lambda)\right\|_K}_{\textit{Local variance}}, \tag{10}$$

where $\widehat{C}_{D_j}$ is the empirical covariance operator on D_j and $C_\lambda = C + \lambda I$.

Using the triangle inequality, one can prove the error decomposition (9) easily. The upper bound of the distributed error (10) is proven in Lemma 23, where $\|C_\lambda^{-1/2}(C - \widehat{C}_{D_j})C_\lambda^{-1/2}\|$ measures the gap between expected and empirical covariance operators via concentration inequalities. Distributed error measures the performance gap between the divide-and-conquer strategy and centralized learning. The global variance measures the discrepancy between the expected estimator f_λ and the ERM estimator $\widehat{f}^M_{D,\lambda}$ on the datasets D. The variance consists of sample variance, Nyström error, and empirical error.

Theorem 16 (Excess risk bound of `DNyström` in Probability) *Let $\delta > 0$, $\lambda = N^{\frac{-1}{2r+\gamma}}$ and $|D_1| = \cdots = |D_m| = N/m$. Under Assumptions 5-7, if $\lambda = N^{-\frac{1}{2r+\gamma}}$,*

$$r \in (0,1], \qquad \gamma \in (0,1], \qquad 2r + \gamma \geq \alpha, \qquad m \lesssim N^{\frac{2r+\gamma-\alpha}{4r+2\gamma}},$$

$M \gtrsim N^{\frac{\alpha}{2r+\gamma}}$ for the uniform sampling, and $M \gtrsim N^{\frac{\gamma}{2r+\gamma}}$ for the data-dependent sampling, then with probability $1 - 4\delta$, there exists

$$\left\|\bar{f}^M_{D,\lambda} - f_\rho\right\|_\rho \lesssim N^{\frac{-r}{2r+\gamma}} \log(2/\delta).$$

The above excess risk bound in probability also achieves the optimal rates but allows fewer partitions. The applicable area and the number of Nyström centers are the same as Theorem 11, but the allowed number of partitions is smaller $m \lesssim N^{\frac{2r+\gamma-1}{4r+2\gamma}}$, which is the square root of that in Theorem 11.

Remark 17 *Note that, we usually estimate the key quantity $\|C_\lambda^{-1/2}(C - \widehat{C}_{D_j})C_\lambda^{-1/2}\|$ as a constant with a sufficient number of local examples. However, if we directly estimate $\|C_\lambda^{-1/2}(C - \widehat{C}_{D_j})C_\lambda^{-1/2}\|$ as a constant, the distributed error depends on $\|C_\lambda^{1/2}\ (\widehat{f}^M_{D_j,\lambda} - f_\lambda)\|_K = \mathcal{O}((N/m)^{\frac{-r}{2r+\gamma}})$ that is suboptimal. Therefore, we keep this key quantity in* (10) *and compute the multiplication $\left\|C_\lambda^{-1/2}(C - \widehat{C}_{D_j})C_\lambda^{-1/2}\right\| \left\|C_\lambda^{1/2}\ (\widehat{f}^M_{D_j,\lambda} - f_\lambda)\right\|_K$ to obtain the optimal rates $\mathcal{O}(N^{\frac{-r}{2r+\gamma}})$. More details refer to* (55) *in the proof of Theorem 16.*

3.4 Compared with Related Work

Both distributed learning and Nyström approximation are typical techniques to further reduce the computational burdens for kernel methods. For example, Yin et al. combined divide-and-conquer with Nyström approximation (DC-NY) and proved the optimal rates for DC-NY (Yin et al., 2020), Lian et al. combined divided-and-conquer with random sketching (DC-Sketch) and derived the optimal learning rates (Lian et al., 2021) with $m \lesssim N^{\frac{2r-1}{2r+\gamma}}$ and $M \gtrsim N^{\frac{\gamma}{2r+\gamma}}$ where M is the sketching size. We compare the proposed `DNyström` with DKRR, DC-NY, and DC-Sketch.

1) On the algorithmic front. DKRR averaged the local estimators on each subset, while DC-NY and DC-Sketch are used to accelerate the computation of the local estimators on each subset by using Nyström approximation and random sketching, respectively. They only considered local information from each subset and failed to characterize the global information. However, the proposed `DNyström` sampled Nyström centers from all subsets that can capture the global characteristics of the training data.

2) On the statistical front. The traditional theoretical results for DKRR (Guo et al., 2017; Lin et al., 2017; Chang et al., 2017), DC-NY (Yin et al., 2020) and DC-Sketch (Lian et al., 2021) derived the optimal learning rates in expectation, which only reflect the average errors of the algorithms rather than a single trial in practice. In this paper, we derive the excess risk bounds in expectation and in probability, respectively. Even though DKRR with communications (Lin et al., 2020) provided error bounds in probability, the error decomposition of `DNyström` is different from (Lin et al., 2020) where the Nyström error is derived in this paper. Besides, the existing work usually assumed that the target regression belongs to the RKHS, i.e. $f_\rho \in \mathcal{H}$, but we remove this strict condition for `DNyström` where the optimal convergence rates of excess risk bounds pertain to the non-attainable case $f_\rho \notin \mathcal{H}$.

3) Proof techniques. The error decompositions for DKRR, DC-NY, and DC-Sketch are usually implicit, but we provide intermediate estimators and explicit decompositions such that one can quantify the errors caused by different components. We derive the Nyström error that is resulted from Nyström approximation in the non-attainable case for the first time, while the estimates of error terms for DKRR, DC-NY, and DC-Sketch only applied to the attainable case. The distributed error (10) for `DNyström` in probability is suboptimal if we directly use the traditional proof techniques, i.e. $\|C_\lambda^{-1/2}(C - \widehat{C}_{D_j})C_\lambda^{-1/2}\|$ as a constant. We compute the multiplication and obtain the optimal rates, as discussed in Remark 17.

4. Experiments

In this section, we study the generalization performance on real-world datasets. We implement all methods based on Pytorch 1.13 [1] and run experiments on a Linux Server with an Nvidia RTX 2080Ti GPU. We compare the proposed `DNyström` with random features method (RF) (Rahimi and Recht, 2007), Nyström approximation (Bach, 2013), DKRR (Zhang et al., 2015), DC-NY (Yin et al., 2020) and DC-RF (Li et al., 2019). In all experiments, we use uniform sampling to sample Nyström centers and random features.

1. Publicly available at `https://github.com/superlj666/DNystroem`

Dataset	Classes	N_{train}	N_{test}	d	σ	λ
usps	10	7291	2007	256	10	10^{-6}
pendigits	10	7494	3498	16	100	10^{-6}
letter	26	15000	5000	16	1	10^{-7}
MNIST	10	60000	10000	784	10	10^{-6}

Table 1: The statistics and tuned hyperparameters in datasets

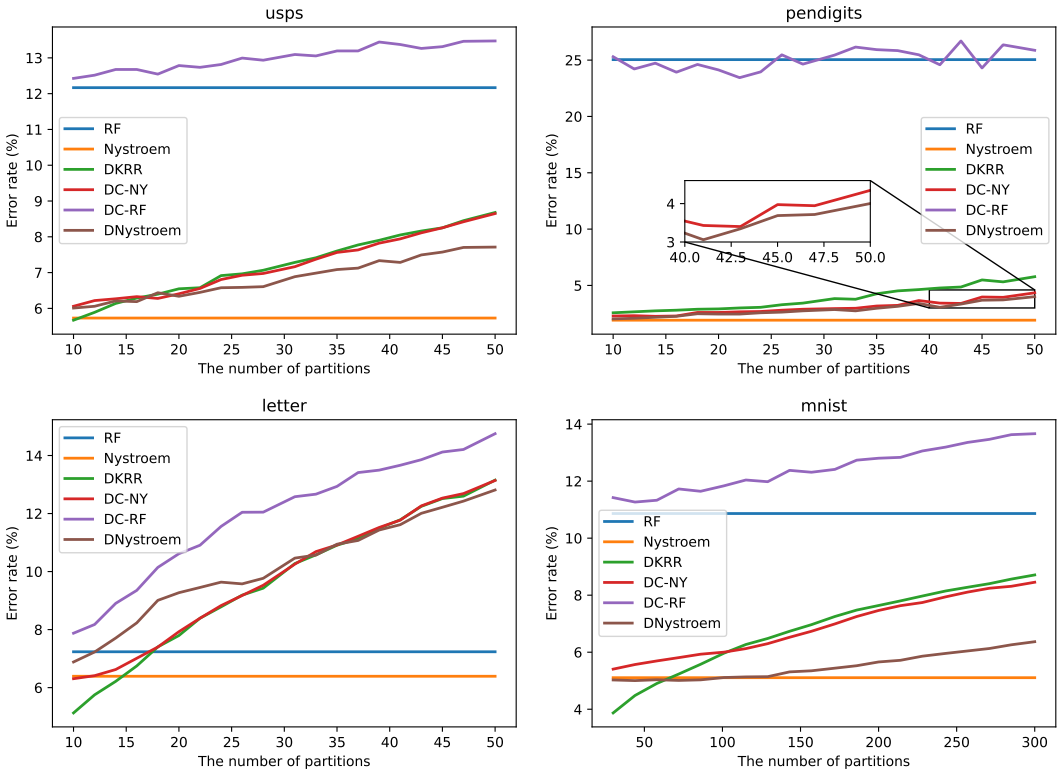

Figure 5: Comparison of the classification error rates vs. the number of partitions m.

We evaluate the compared algorithms on real-world classification datasets, which are publicly available from UCI datasets [2]. Based on Gaussian kernel $K(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\|\boldsymbol{x} - \boldsymbol{x}'\|^2/2\sigma^2)$ we conduct empirical evaluations and repeat the training 10 times to record the average test error rates. Using the toolbox NNI [3], we tune the optimal hyperparameters over the grids $\sigma \in \{10^i, i = -4, -3, \cdots, 4\}$ and $\lambda \in \{10^i, i = -10, \cdots, -1\}$. The statistics information and optimal hyperparameters for datasets are recorded in Table 1.

2. Available at `http://archive.ics.uci.edu/ml/datasets.php`
3. Available at `https://github.com/microsoft/nni`

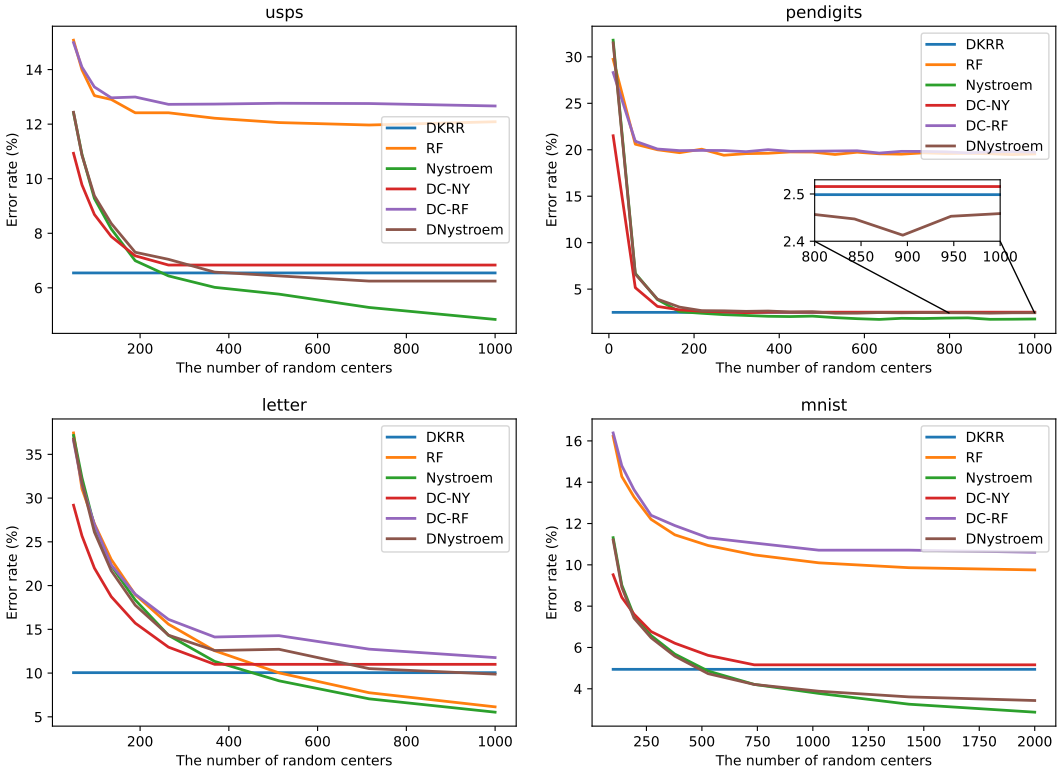

Figure 6: Comparison on the classification error rates vs. the number of Nyström centers M, i.e. the number of Nyström landmarks or the number of random features.

The impact of the number of partitions. We first explore the impact of the number of partitions with a fixed number of Nyström centers or random features $M = 500$. We carry out the compared methods on the classification datasets 10 trials and record the average error rates in Figure 5. From that, we can conclude the following assertions. 1) With the same sample size M, the test accuracies of the random features method are always worse than that of Nyström method, especially on the datasets usps, pendigits, and mnist. This observation verifies the theoretical results in Theorem 11 that the required number of Nyström centers $M \gtrsim N^{\frac{1}{2r+\gamma}}$ is smaller than the required number of random features $M \gtrsim N^{\frac{1+\gamma(2r-1)}{2r+\gamma}}$ (Rudi and Rosasco, 2017) in the case of optimal rates. 2) As the number of partitions increases, the error rates of distributed methods, including DKRR, DC-NY, DC-RF, and `DNyström`, become larger. Specifically, DC-NY is more closed to DKRR since DC-NY approximated local estimators of DKRR via Nyström approximation, while `DNyström` achieves the lower test error rates than them when m is large. 3) When $m \geq 20$ for usps, $m \geq 10$ for pendigits, $m \geq 37$ for letter, and $m \geq 72$ for mnist, `DNyström` achieves better performance than DKRR and DC-NY, and thus `DNyström` is more flexible in the setting of distributed learning.

The impact of the number of random centers. We then fixed the number of partitions and explore the impact of the number of random centers M. We set $m = 20$ for the datasets usps, pendigits and letter, and $m = 60$ for mnist. We carry out the

compared methods on the classification datasets 10 trials and record the average error rates vs. the number of random centers in Figure 6. We find that 1) The predictive accuracies of random features based methods are always worse than that of Nyström based methods. Especially on the datasets usps, pendigits, and mnist, even random features method without partitions perform much worse than Nyström based algorithms. 2) As the number of random centers increases, the error rates of both Nyström and random features based methods decrease. Specifically, the error rates of DC-NY converge to that of DKRR when the number of Nyström centers M is bigger than the local sample size N/m, while the error rates of `DNyström` can still decrease when $M \geq N/m$ owing the Nyström centers are sampled from the entire training set. 3) When M is very small, DC-NY performs better than `DNyström` because these algorithms have not fully characterized the local information. As the increase of Nyström centers, `DNyström` outperforms DC-NY owing to the characterization of global information (from other devices), and finally defeats DKRR when $M \geq 513$ for usps, $M \geq 531$ for pendigits, $M \geq 716$ for letter, and $M \geq 528$ for mnist.

5. Conclusion

In this paper, we propose distributed Nyström approximation approach with the globally-shared Nyström centers, which can capture the global characteristics from all training samples. We then study the generalization properties for `DNyström`, and obtain the optimal convergence rates in both expectation and expectation, respectively. Note that, the derived optimal rates apply to the non-attainable case where the target regression may be out of the hypothesis space. Compared to DKRR and DC-NY, the proposed `DNyström` requires fewer Nyström centers and allows more partitions to achieve the same optimal learning rates. The experimental results also validate the advantage of `DNyström` over the compared methods in both the number of partitions and the number of random centers. In the future, one can use the globally-shared Nyström centers to reduce the effects of data heterogeneity in federated learning and decentralized learning.

Acknowledgments

The work of Jian Li is supported partially by National Natural Science Foundation of China (No. 62106257), Excellent Talents Program of Institute of Information Engineering, CAS, and the Special Research Assistant Project of CAS (No. E0YY231114). The work of Yong Liu is supported partially by National Natural Science Foundation of China (No. 62076234), Beijing Outstanding Young Scientist Program (No. BJJWZYJH012019100020098), the "Intelligent Social Governance Interdisciplinary Platform, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, Renmin University of China", the Huawei-Renmin University joint program on Information Retrieval, the Unicom Innovation Ecological Cooperation Plan, and the CCF-Huawei Populus Grove Fund.

Appendix A. Proofs

In this section, we begin introducing some operators and the the discrepancies between their expected and empirical counterparts. We then provide error decomposition for the excess risk bound of `DNyström` in expectation and in high probability, respectively. Finally, we upper bound the error terms and prove the main results.

A.1 Operators

We define expected operators, and empirical operators based on local training examples and Nyström approximation, respectively. For the sake of simplification, we let the primal training set be equally divided, such that $n = |D_j| = N/m + M,\ \forall j \in [m]$ and $|D| = m|D_j|$.

Definition 18 (Expected operators) *For any $g \in L^2_{\rho_X}$ and $\beta \in \mathcal{H}$, we have*

- $S : \mathcal{H} \to L^2_{\rho_X},\ (S\beta)(\boldsymbol{x}) = \langle \beta, K_x \rangle$.
- $S^* : L^2_{\rho_X} \to \mathcal{H},\ S^* g = \int_X K_x g(\boldsymbol{x})\, d\rho_X(\boldsymbol{x})$.
- $L : L^2_{\rho_X} \to L^2_{\rho_X},\ (Lg)(\boldsymbol{x}) = \int_X K(\boldsymbol{x}, \boldsymbol{z}) g(\boldsymbol{z})\, d\rho_X(\boldsymbol{z})$.
- $C : \mathcal{H} \to \mathcal{H},\ C = \int_X K_x \otimes K_x\, d\rho_X(\boldsymbol{x})$.

It holds that for the integral operator $L = SS^*$ and for the covariance operator $C = S^*S$.

Definition 19 (Empirical operators) *For any $g \in L^2_{\rho_X}$, $\beta \in \mathcal{H}$, $\alpha \in \mathbb{R}^n$ and $\alpha' \in \mathbb{R}^M$, with the training examples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ from the local set D_j, we have*

- $\widehat{S}_{D_j} : \mathcal{H} \to \mathbb{R}^n,\ \widehat{S}_{D_j}\beta = \frac{1}{\sqrt{n}} \left(\langle \beta, K_{x_i} \rangle\right)_{i=1}^n$.
- $\widehat{S}^*_{D_j} : \mathbb{R}^n \to \mathcal{H},\ \widehat{S}^*_{D_j}\alpha = \frac{1}{\sqrt{n}} \sum_{i=1}^n K_{x_i}\alpha_i$.
- $\bar{S}^*_{D_j} : L^2_{\rho_X} \to \mathcal{H},\ \bar{S}^*_{D_j} g = \frac{1}{n} \sum_{i=1}^n K_{x_i} g(\boldsymbol{x}_i)$.
- $\widehat{C}_{D_j} : \mathcal{H} \to \mathcal{H},\ \widehat{C}_{D_j} = \frac{1}{n} \sum_{i=1}^n K_{x_i} \otimes K_{x_i}$.
- $\widehat{L}_D : L^2_{\rho_X} \to L^2_{\rho_X},\ \widehat{L}_D g(\cdot) = \frac{1}{n} \sum_{i=1}^n K(\boldsymbol{x}_i, \cdot) g(\boldsymbol{x}_i)$.
- $\widehat{S}_M : \mathcal{H} \to \mathbb{R}^M,\ \widehat{S}_M\beta = \frac{1}{\sqrt{M}} \left(\langle \beta, K_{x_i} \rangle\right)_{i=1}^M$.
- $\widehat{S}^*_M : \mathbb{R}^M \to \mathcal{H},\ \widehat{S}^*_M\alpha' = \frac{1}{\sqrt{M}} \sum_{i=1}^M K_{x_i}\alpha'_i$.
- $\widehat{C}_M : \mathcal{H} \to \mathcal{H},\ \widehat{C}_M = \frac{1}{M} \sum_{i=1}^M K_{x_i} \otimes K_{x_i}$.

It holds that for the kernel matrices $\mathbf{K}_{D_j} = |D_j| \widehat{S}_{D_j} \widehat{S}^*_{D_j}$, $\mathbf{K}_{MM} = M \widehat{S}_M \widehat{S}^*_M$, $\mathbf{K}_{jM} = \sqrt{|D_j| M} \widehat{S}_{D_j} \widehat{S}^*_M$ and for the covariance operators $\widehat{C}_{D_j} = \widehat{S}^*_{D_j} \widehat{S}_{D_j}$, $\widehat{C}_M = \widehat{S}^*_M \widehat{S}_M$.

We denote with $\|\cdot\|$ the operatorial norm, and specifically the norm $\|\cdot\|$ to represent the $L^2_{\rho_X}$ norm $\|\cdot\|_\rho$ in the estimate of error terms. Let $\mathcal{L}$ be a Hilbert space, we denote with $\langle\cdot,\cdot\rangle_{\mathcal{L}}$ the associated inner product, with $\|\cdot\|_{\mathcal{L}}$ the norm and with $\mathrm{Tr}(\cdot)$ the trace. Moreover, we denote with Q_λ the operator $Q+\lambda I$, where Q is a linear operator, $\lambda\in\mathbb{R}$ and I the identity operator, so for example $C_\lambda := C+\lambda I$, $\widehat{C}_{D,\lambda} := \widehat{C}_D+\lambda I$, $\widehat{C}_{D_j,\lambda} := \widehat{C}_{D_j}+\lambda I$, $L_\lambda := L+\lambda I$, $\widehat{L}_{D,\lambda} := \widehat{L}_D+\lambda I$, and $\widehat{L}_{D_j,\lambda} := \widehat{L}_{D_j}+\lambda I$.

Proposition 20 (Characterizations of estimators) *Let $\sqrt{M}\widehat{S}_M = U\Sigma V^*$ be the SVD of the empirical sampling operator. Using operators in Definitions 18, 19, the estimators can be represented as*

$$\widehat{f}^M_{D_j,\lambda} = V(V^*\widehat{C}_{D_j}V+\lambda I)^{-1}V^*\widehat{S}^*_{D_j}\boldsymbol{y}_{D_j}, \tag{11}$$

$$\widehat{f}^M_{D,\lambda} = V(V^*\widehat{C}_{D}V+\lambda I)^{-1}V^*\widehat{S}^*_{D}\boldsymbol{y}_{D}, \tag{12}$$

$$\widetilde{f}^M_{D,\lambda} = V(V^*\widehat{C}_{D}V+\lambda I)^{-1}V^*\bar{S}^*_{D}f_\rho, \tag{13}$$

$$\widetilde{f}_{D,\lambda} = (\widehat{C}_D+\lambda I)^{-1}\bar{S}^*_D f_\rho, \tag{14}$$

$$f_\lambda = (C+\lambda I)^{-1}S^*f_\rho. \tag{15}$$

Here, $\boldsymbol{y}_{D_j} = \frac{1}{\sqrt{|D_j|}}(y_1,\cdots,y_{|D_j|})^\top$ and $\boldsymbol{y}_D = \frac{1}{\sqrt{|D|}}(y_1,\cdots,y_{|D|})$.

Proof The RKHS solution $\widehat{f}^M_{D_j,\lambda} = \sum_{i=1}^{|D_j|}\alpha'_i K(\widetilde{\boldsymbol{x}}_i,\cdot)$ admits

$$\boldsymbol{\alpha}' = (\mathbf{K}_{jM}^\top\mathbf{K}_{jM}+\lambda n\mathbf{K}_{MM})^\dagger\mathbf{K}_{jM}^\top\boldsymbol{y}_{D_j} = [M(\widehat{S}_M\widehat{S}^*_{D_j})(\widehat{S}_{D_j}\widehat{S}^*_M)+\lambda M(\widehat{S}_M\widehat{S}^*_M)]^\dagger(\sqrt{M}\widehat{S}_M\widehat{S}^*_{D_j})\boldsymbol{y}_{D_j}.$$

Then, there exists

$$\begin{aligned}\widehat{f}^M_{D_j,\lambda} = \sqrt{M}\widehat{S}^*_M\boldsymbol{\alpha}' &= \widehat{S}^*_M[(\widehat{S}_M\widehat{S}^*_{D_j})(\widehat{S}_{D_j}\widehat{S}^*_M)+\lambda(\widehat{S}_M\widehat{S}^*_M)]^\dagger(\widehat{S}_M\widehat{S}^*_{D_j})\boldsymbol{y}_{D_j}\\ &= \widehat{S}^*_M[\widehat{S}_M(\widehat{C}_{D,\lambda})\widehat{S}^*_M]^\dagger(\widehat{S}_M\widehat{S}^*_{D_j})\boldsymbol{y}_{D_j}.\end{aligned}$$

Following the step of proof in Lemma 3 (Rudi et al., 2015), we have

$$[M\widehat{S}_M(\widehat{C}_{D,\lambda})\widehat{S}^*_M]^\dagger = (FGH)^\dagger = H^\dagger(FG)^\dagger = H^\dagger G^{-1}F^\dagger = U\Sigma^{-1}(V^*\widehat{C}_{D_j}V+\lambda I)^{-1}\Sigma^{-1}U^*,$$

where $\sqrt{M}\widehat{S}_M = U\Sigma V^*$, $F=U\Sigma$, $G = V^*\widehat{C}_{D_j}V+\lambda I$, $H=\Sigma U^\top$ and F, GH, G and H are full-rank matrices. Simplifying U and Σ, we prove (11) with

$$\begin{aligned}\widehat{f}^M_{D_j,\lambda} &= \sqrt{M}\widehat{S}^*_M[M\widehat{S}_M(\widehat{C}_{D,\lambda})\widehat{S}^*_M]^\dagger(\sqrt{M}\widehat{S}_M\widehat{S}^*_{D_j})\boldsymbol{y}_{D_j}\\ &= V\Sigma U^*U\Sigma^{-1}(V^*\widehat{C}_{D_j}V+\lambda I)^{-1}\Sigma^{-1}U^*U\Sigma V^*\widehat{S}^*_{D_j}\boldsymbol{y}_{D_j}\\ &= V(V^*\widehat{C}_{D_j}V+\lambda I)^{-1}V^*\widehat{S}^*_{D_j}\boldsymbol{y}_{D_j}.\end{aligned}$$

Similarly, we can prove the empirical estimator $\widehat{f}^M_{D,\lambda}$ on D.

The noise-free estimator $\widetilde{f}^M_{D,\lambda}$ depends on the labels $f_\rho(\boldsymbol{x}_i)$ instead of noisy labels y_i, where $(\boldsymbol{x}_i,y_i)\in D$. Since $\bar{f}^M_{D,\lambda} = V(V^*\widehat{C}_DV+\lambda I)^{-1}V^*\left(\sum_{j=1}^m\frac{|D_j|}{|D|}K_{\boldsymbol{x}_i}y_i\right)$, the corresponding noise-free estimator is

$$\widetilde{f}^M_{D,\lambda} = V(V^*\widehat{C}_DV+\lambda I)^{-1}V^*\left(\sum_{j=1}^m\frac{|D_j|}{|D|}K_{\boldsymbol{x}_i}f_\rho(\boldsymbol{x}_i)\right) = V(V^*\widehat{C}_DV+\lambda I)^{-1}V^*\bar{S}^*_Df_\rho.$$

We present the representation for $\widetilde{f}_{D,\lambda}$ without Nyström approximation $\widetilde{f}_{D,\lambda} = \sum_{i=1}^{M} \alpha_i K(\tilde{\boldsymbol{x}}, \cdot)$ with $\boldsymbol{\alpha} = (\mathbf{K}_{D_j} + \lambda |D| I)^{-1} \boldsymbol{y}_j$, and thus

$$\begin{aligned}
\widetilde{f}_{D,\lambda} &= \sqrt{|D|}\widehat{S}_D^*(|D|\widehat{S}_D\widehat{S}_D^* + \lambda|D|I)^{-1}\left[f_\rho(\boldsymbol{x}_1), \cdots, f_\rho(\boldsymbol{x}_{|D|})\right]^\top \\
&= \frac{1}{\sqrt{|D|}}(\widehat{S}_D^*\widehat{S}_D + \lambda I)^{-1}(\widehat{S}_D^*\widehat{S}_D + \lambda I)\widehat{S}_D^*(\widehat{S}_D\widehat{S}_D^* + \lambda I)^{-1}\left[f_\rho(\boldsymbol{x}_1), \cdots, f_\rho(\boldsymbol{x}_{|D|})\right]^\top \\
&= \frac{1}{\sqrt{|D|}}(\widehat{S}_D^*\widehat{S}_D + \lambda I)^{-1}\widehat{S}_D^*(\widehat{S}_D\widehat{S}_D^* + \lambda I)(\widehat{S}_D\widehat{S}_D^* + \lambda I)^{-1}\left[f_\rho(\boldsymbol{x}_1), \cdots, f_\rho(\boldsymbol{x}_{|D|})\right]^\top \\
&= \frac{1}{\sqrt{|D|}}(\widehat{C}_D + \lambda I)^{-1}\widehat{S}_D^*\left[f_\rho(\boldsymbol{x}_1), \cdots, f_\rho(\boldsymbol{x}_{|D|})\right]^\top \\
&= (\widehat{C}_D + \lambda I)^{-1}\bar{S}_D^* f_\rho.
\end{aligned}$$

It is well know the estimator f_λ in $L^2_{\rho_X}$ space is equal to

$$Sf_\lambda = L(L+\lambda I)^{-1} f_\rho = SS^*(SS^* + \lambda I)^{-1} f_\rho = S(S^*S + \lambda I)^{-1} S^* f_\rho = S(C + \lambda I)^{-1} S^* f_\rho.$$

Then, we have $f_\lambda = (C + \lambda I)^{-1} S^* f_\rho$. ■

Lemma 21 *Let $K_{\boldsymbol{x}_1}, \cdots, K_{\boldsymbol{x}_n}$ with $n \geq 1$, be i.i.d random vectors on a separable Hilbert space $\mathcal{H}$ such that $C = \mathbb{E}_{\rho_X}[K_x \otimes K_x]$, $\widehat{C}_{D_j} = \frac{1}{|D_j|}\sum_{i=1}^{|D_j|} K_{\boldsymbol{x}_i} \otimes K_{\boldsymbol{x}_i}$, $(Lg)(\cdot) = \mathbb{E}_{\rho_X}[K(\boldsymbol{x}, \cdot) g(\boldsymbol{x})]$ and $(\widehat{L}_{D_j} g)(\cdot) = \frac{1}{|D_j|}\sum_{i=1}^{|D_j|} K(\boldsymbol{x}_i, \cdot) g(\boldsymbol{x}_i)$ are trace class. Then for any $\delta \in (0,1)$, with the probability at least $1-\delta$, the following holds*

$$\left\|C_\lambda^{-1/2}(C - \widehat{C}_{D_j})C_\lambda^{-1/2}\right\| \leq \left\|C_\lambda^{-1}(C - \widehat{C}_{D_j})\right\| \leq \frac{2\mathcal{N}_\infty(\lambda)\log(2/\delta)}{|D_j|} + \sqrt{\frac{2\mathcal{N}_\infty(\lambda)\log(2/\delta)}{|D_j|}}, \tag{16}$$

and it also holds with the probability at least $1-\delta$ that

$$\left\|L_\lambda^{-1/2}(L - \widehat{L}_{D_j})L_\lambda^{-1/2}\right\| \leq \left\|L_\lambda^{-1}(L - \widehat{L}_{D_j})\right\| \leq \frac{2\mathcal{N}_\infty(\lambda)\log(2/\delta)}{|D_j|} + \sqrt{\frac{2\mathcal{N}_\infty(\lambda)\log(2/\delta)}{|D_j|}}. \tag{17}$$

Proof Using the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\left\|(C+\lambda I)^{-1/2}(C - \widehat{C}_{D_j})(C+\lambda I)^{-1/2}\right\| \\
=&\left\|(C+\lambda I)^{-1/2}(C - \widehat{C}_{D_j})^{1/2}(C - \widehat{C}_{D_j})^{1/2}(C+\lambda I)^{-1/2}\right\| \\
\leq&\left\|(C+\lambda I)^{-1/2}(C - \widehat{C}_{D_j})^{1/2}\right\|^2.
\end{aligned} \tag{18}$$

Recall that the norm on a matrix or operator A can be defined By

$$\|A\| := \sup_x \frac{\|Ax\|_2}{\|x\|_2}.$$

For $K > 1$ and a nonzero vector x, we get

$$\|A^k x\|_2 = \|AA^{k-1}x\|_2 \leq \|A\|\|A^{k-1}x\|_2 \leq \cdots \leq \|A\|^k\|x\|_2.$$

Therefore, it holds $\frac{\|A^k x\|_2}{\|x\|_2} \leq \|A\|^k$ and thus

$$\|A^k\| = \sup_x \frac{\|A^k x\|_2}{\|x\|_2} \leq \|A\|^k. \tag{19}$$

Assuming $A = (C + \lambda I)^{-1/2}$ and substituting (19) to (18), we get

$$\left\|(C + \lambda I)^{-1/2}(C - \widehat{C}_{D_j})(C + \lambda I)^{-1/2}\right\| \leq \left\|(C + \lambda I)^{-1}(C - \widehat{C}_{D_j})\right\|. \tag{20}$$

Let $\xi = (C + \lambda I)^{-1}K_{\boldsymbol{x}} \otimes K_{\boldsymbol{x}}$, thus we have

$$\frac{1}{|D_j|}\sum_{i=1}^{|D_j|}\xi_i = \frac{1}{|D_j|}\sum_{i=1}^{|D_j|}(C + \lambda I)^{-1}[K_{\boldsymbol{x}_i} \otimes K_{\boldsymbol{x}_i}] = (C + \lambda I)^{-1}\widehat{C}_{D_j},$$
$$\mathbb{E}(\xi) = (C + \lambda I)^{-1}\mathbb{E}[K_{\boldsymbol{x}} \otimes K_{\boldsymbol{x}}] = (C + \lambda I)^{-1}C.$$

To bound $\left\|(C + \lambda I)^{-1}(C - \widehat{C}_{D_j})\right\| = \left\|\frac{1}{|D_j|}\sum_{i=1}^{|D_j|}\xi_i - \mathbb{E}(\xi)\right\|$, we estimate the maximal eigenvalue and the moments of random operators $\xi_i - \mathbb{E}(\xi_i)$, such that

$$\lambda_{max}(\xi_i - \mathbb{E}(\xi_i)) \leq \|(C + \lambda I)^{-1/2}K_{\boldsymbol{x}}\|^2 \leq \mathcal{N}_\infty(\lambda).$$
$$\begin{aligned}\mathbb{E}\big(\xi_i - \mathbb{E}(\xi_i)\big)^2 &= \left\|\mathbb{E}\left[\langle(C + \lambda I)^{-1}K_{\boldsymbol{x}}, K_{\boldsymbol{x}}\rangle(C + \lambda I)^{-1}K_{\boldsymbol{x}} \otimes K_{\boldsymbol{x}}\right]\right\| - \left\|C_\lambda^{-2}C^2\right\| \\ &\leq \mathcal{N}_\infty(\lambda)\left\|\mathbb{E}\left[(C + \lambda I)^{-1}K_{\boldsymbol{x}} \otimes K_{\boldsymbol{x}}\right]\right\| \leq \mathcal{N}_\infty(\lambda)\|C_\lambda^{-1}C\| \leq \mathcal{N}_\infty(\lambda).\end{aligned}$$

Then, using Bernstein's inequality for random operators (Proposition 3 of (Rudi and Rosasco, 2017)), with the probability at least $1 - \delta$, we have

$$\left\|(C + \lambda I)^{-1}(C - \widehat{C}_{D_j})\right\| \leq \frac{2\mathcal{N}_\infty(\lambda)\log(2/\delta)}{|D_j|} + \sqrt{\frac{2\mathcal{N}_\infty(\lambda)\log(2/\delta)}{|D_j|}}. \tag{21}$$

Combining (20) and (21), we obtain the result in (16). Then, following the above proof, we can prove (17) by setting $\xi_i = L_\lambda^{-1/2}K(\boldsymbol{x}_i, \cdot)L_\lambda^{-1/2}$. ■

Note that, the above lemma is the key to obtain the sharper estimates for the key quantities $\|(C + \lambda I)^{-1/2}(\widehat{C}_{D_j} + \lambda I)^{1/2}\|$ and $\|(L + \lambda I)^{-1/2}(\widehat{L}_D + \lambda I)^{1/2}\|$, which should be bounded as a constant when estimating the error terms. Traditional DKRR work (Guo et al., 2017; Yin et al., 2020) estimated the key quantities after decomposition, and obtain $\|(C + \lambda I)^{-1/2}(\widehat{C}_{D_j} + \lambda I)^{1/2}\|^2 \leq \|(C + \lambda I)^{-1/2}\|\|(C + \lambda I)^{-1/2}(C - \widehat{C}_{D_j})\| + 1 = \mathcal{O}\big(\frac{1}{\lambda|D_j|} + \sqrt{\frac{\mathcal{N}(\lambda)}{\lambda|D_j|}}\big)$, leading to the restriction $m \lesssim N^{\frac{2r-1}{2r+\gamma}}$. Here, using the concentration inequalities for self-adjoint operators and obtain $\|(C + \lambda I)^{-1/2}(\widehat{C}_{D_j} + \lambda I)^{1/2}\|^2 = \|(C + \lambda I)^{-1/2}(\widehat{C}_{D_j} + \lambda I)(C + \lambda I)^{-1/2}\| = \|I + (C + \lambda I)^{-1/2}(C - \widehat{C}_{D_j})(C + \lambda I)^{-1/2}\| = \mathcal{O}\big(\frac{\mathcal{N}_\infty(\lambda)}{|D_j|} + \sqrt{\frac{\mathcal{N}_\infty(\lambda)}{|D_j|}}\big)$, where the restriction is relaxed to $|D_j| \gtrsim \mathcal{N}_\infty(\lambda)$.

Lemma 22 *When the sample size satisfies* $|D_j| \geq 16\mathcal{N}_\infty(\lambda)\log(2/\delta)$*, then* $\forall\ \delta \in (0,1)$*, there exists with the confidence* $1-\delta$

$$\|C_\lambda^{-1/2}(C-\widehat{C}_{D_j})C_\lambda^{-1/2}\| \leq \frac{1}{2}, \quad \|C_\lambda^{1/2}\widehat{C}_{D_j,\lambda}^{-1/2}\| \leq \sqrt{2}, \quad \|C_\lambda^{-1/2}\widehat{C}_{D_j,\lambda}^{1/2}\| \leq 2,$$
$$\|L_\lambda^{-1/2}(L-\widehat{L}_{D_j})L_\lambda^{-1/2}\| \leq \frac{1}{2}, \quad \|L_\lambda^{1/2}\widehat{L}_{D_j,\lambda}^{-1/2}\| \leq \sqrt{2}, \quad \|L_\lambda^{-1/2}\widehat{L}_{D_j,\lambda}^{1/2}\| \leq 2.$$

Proof From Lemma 21, we set $|D_j| \geq 16\mathcal{N}_\infty(\lambda)\log(2/\delta)$ and obtain that

$$\|C_\lambda^{-1/2}(\widehat{C}_{D_j}-C)C_\lambda^{-1/2}\| \leq \frac{2\mathcal{N}_\infty(\lambda)\log(2/\delta)}{|D_j|} + \sqrt{\frac{2\mathcal{N}_\infty(\lambda)\log(2/\delta)}{|D_j|}} \leq \frac{1}{2}.$$

From Proposition 7 of (Rudi et al., 2015) and the above inequality, there exists

$$\|C_\lambda^{1/2}\widehat{C}_{D,\lambda}^{-1/2}\| \leq \left(1-\frac{1}{2}\right)^{-\frac{1}{2}} = \sqrt{2}.$$

Meanwhile, from Cordes inequality (Fujii et al., 1993) and Lemma 21, when $|D_j| \geq 16\mathcal{N}_\infty(\lambda)\log(2/\delta)$, we have

$$\|C_\lambda^{-1/2}\widehat{C}_{D,\lambda}^{1/2}\| \leq \|(C+\lambda I)^{-1}(\widehat{C}_D+\lambda I)\|^{1/2} = \|I+(C+\lambda I)^{-1}(\widehat{C}_{D,\lambda}-C)\|^{1/2} \leq 2.$$

Similarly, we can prove results for integral operators L and $\widehat{L}_D$. ■

A.2 Estimates for Error Terms

Note that, in Definition 9, we define the estimators in the RKHS where the $\mathcal{H}$-norm can be related to $L^2_{\rho_X}$-norm by the inclusion operator S (Lin and Cevher, 2018) that $\forall f \in \mathcal{H}$,

$$\|f\|_\rho = \|Sf\|_\rho = \|C^{1/2}f\|_K \leq \|(C+\lambda I)^{1/2}f\|_K. \tag{22}$$

A.2.1 Estimate for Distributed error

Lemma 23 (Distributed error) *When* $|D_j| \geq 16\mathcal{N}_\infty(\lambda)\log(2/\delta)$ *and* $|D_1| = \cdots = |D_m| = N/m$*, with the probability at least* $1-\delta$*, we have*

$$\|\bar{f}_{D,\lambda}^M - \widehat{f}_{D,\lambda}^M\|_\rho \leq 4\left\|C_\lambda^{-1/2}(C-\widehat{C}_{D_j})C_\lambda^{-1/2}\right\|\left\|C_\lambda^{1/2}(\widehat{f}_{D_j,\lambda}^M - f_\lambda)\right\|_K. \tag{23}$$

Proof For the sake of simplification, we denote $G_D = V(V^*\widehat{C}_D V+\lambda I)^{-1}V^*$. From $\bar{f}_{D,\lambda}^M = \sum_{j=1}^m \frac{|D_j|}{|D|}\widehat{f}_{D_j,\lambda}^M$ and the definition of $\widehat{f}_{D_j,\lambda}^M$ in (11), using the facts $\widehat{S}_D^*\boldsymbol{y}_D = \sum_{j=1}^m \frac{|D_j|}{|D|}\widehat{S}_{D_j}^*\boldsymbol{y}_{D_j}$

and $A^{-1} - B^{-1} = B^{-1}(B-A)A^{-1}$ for positive operators A and B, we have

$$\begin{aligned}
&\bar{f}^M_{D,\lambda} - \widehat{f}^M_{D,\lambda} \\
=&\sum_{j=1}^m \frac{|D_j|}{|D|} V(V^*\widehat{C}_{D_j}V + \lambda I)^{-1}V^*\widehat{S}^*_{D_j}\boldsymbol{y}_{D_j} - V(V^*\widehat{C}_D V + \lambda I)^{-1}V^*\widehat{S}^*_D\boldsymbol{y}_D \\
=&\sum_{j=1}^m \frac{|D_j|}{|D|} V\left[(V^*\widehat{C}_{D_j}V + \lambda I)^{-1} - (V^*\widehat{C}_D V + \lambda I)^{-1}\right]V^*\widehat{S}^*_{D_j}\boldsymbol{y}_{D_j} \\
=&\sum_{j=1}^m \frac{|D_j|}{|D|} V(V^*\widehat{C}_D V + \lambda I)^{-1}V^*(\widehat{C}_D - \widehat{C}_{D_j})V(V^*\widehat{C}_{D_j}V + \lambda I)^{-1}V^*\widehat{S}^*_{D_j}\boldsymbol{y}_{D_j} \\
=&\sum_{j=1}^m \frac{|D_j|}{|D|} G_D(\widehat{C}_D - \widehat{C}_{D_j})\widehat{f}^M_{D_j,\lambda} \\
=&\sum_{j=1}^m \frac{|D_j|}{|D|} G_D(\widehat{C}_D - C)(\widehat{f}^M_{D_j,\lambda} - f_\lambda) + \sum_{j=1}^m \frac{|D_j|}{|D|} G_D(\widehat{C}_D - C)f_\lambda + \sum_{j=1}^m \frac{|D_j|}{|D|} G_D(C - \widehat{C}_{D_j})\widehat{f}^M_{D_j,\lambda} \\
=&\sum_{j=1}^m \frac{|D_j|}{|D|} G_D(\widehat{C}_D - C)(\widehat{f}^M_{D_j,\lambda} - f_\lambda) + \sum_{j=1}^m \frac{|D_j|}{|D|} G_D(C - \widehat{C}_{D_j})(\widehat{f}^M_{D_j,\lambda} - f_\lambda).
\end{aligned}$$

From the above inequality and (22), we then have

$$\begin{aligned}
&\|\bar{f}^M_{D,\lambda} - \widehat{f}^M_{D,\lambda}\|_\rho \\
\le&\sum_{j=1}^m \frac{|D_j|}{|D|}\left(\left\|C_\lambda^{1/2}G_D(\widehat{C}_D - C)(\widehat{f}^M_{D_j,\lambda} - f_\lambda)\right\|_K + \left\|C_\lambda^{1/2}G_D(C - \widehat{C}_{D_j})(\widehat{f}^M_{D_j,\lambda} - f_\lambda)\right\|_K\right).
\end{aligned} \tag{24}$$

From Lemma 22, if $|D_j| \ge 16\mathcal{N}_\infty(\lambda)\log(2/\delta)$, with the probability at least $1-\delta$, we have

$$\begin{aligned}
&\left\|C_\lambda^{1/2}G_D(\widehat{C}_D - C)(\widehat{f}^M_{D_j,\lambda} - f_\lambda)\right\| \\
=&\left\|C_\lambda^{1/2}\widehat{C}_{D,\lambda}^{-1/2}\widehat{C}_{D,\lambda}^{1/2}G_D\widehat{C}_{D,\lambda}^{1/2}\widehat{C}_{D,\lambda}^{-1/2}C_\lambda^{1/2}C_\lambda^{-1/2}(\widehat{C}_D - C)C_\lambda^{-1/2}C_\lambda^{1/2}(\widehat{f}^M_{D_j,\lambda} - f_\lambda)\right\| \\
\le&\left\|C_\lambda^{1/2}\widehat{C}_{D,\lambda}^{-1/2}\right\|\left\|\widehat{C}_{D,\lambda}^{1/2}G_D\widehat{C}_{D,\lambda}^{1/2}\right\|\left\|\widehat{C}_{D,\lambda}^{-1/2}C_\lambda^{1/2}\right\|\left\|C_\lambda^{-1/2}(\widehat{C}_D - C)C_\lambda^{-1/2}\right\|\left\|C_\lambda^{1/2}(\widehat{f}^M_{D_j,\lambda} - f_\lambda)\right\| \\
\le&2\left\|C_\lambda^{-1/2}(\widehat{C}_D - C)C_\lambda^{-1/2}\right\|\left\|C_\lambda^{1/2}(\widehat{f}^M_{D_j,\lambda} - f_\lambda)\right\|,
\end{aligned} \tag{25}$$

and

$$\begin{aligned}
&\left\|C_\lambda^{1/2}G_D(C - \widehat{C}_{D_j})(\widehat{f}^M_{D_j,\lambda} - f_\lambda)\right\| \\
=&\left\|C_\lambda^{1/2}\widehat{C}_{D,\lambda}^{-1/2}\widehat{C}_{D,\lambda}^{1/2}G_D\widehat{C}_{D,\lambda}^{1/2}\widehat{C}_{D,\lambda}^{-1/2}C_\lambda^{1/2}C_\lambda^{-1/2}(C - \widehat{C}_{D_j})C_\lambda^{-1/2}C_\lambda^{1/2}(\widehat{f}^M_{D_j,\lambda} - f_\lambda)\right\| \\
\le&\left\|C_\lambda^{1/2}\widehat{C}_{D,\lambda}^{-1/2}\right\|\left\|\widehat{C}_{D,\lambda}^{1/2}G_D\widehat{C}_{D,\lambda}^{1/2}\right\|\left\|\widehat{C}_{D,\lambda}^{-1/2}C_\lambda^{1/2}\right\|\left\|C_\lambda^{-1/2}(C - \widehat{C}_{D_j})C_\lambda^{-1/2}\right\|\left\|C_\lambda^{1/2}(\widehat{f}^M_{D_j,\lambda} - f_\lambda)\right\| \\
\le&2\left\|C_\lambda^{-1/2}(C - \widehat{C}_{D_j})C_\lambda^{-1/2}\right\|\left\|C_\lambda^{1/2}(\widehat{f}^M_{D_j,\lambda} - f_\lambda)\right\|.
\end{aligned} \tag{26}$$

Note that, $\|\widehat{C}_{D,\lambda}^{1/2} G_D \widehat{C}_{D,\lambda}^{1/2}\| \leq 1$ from Lemma 8 of (Rudi et al., 2015).

Substituting (25) and (26) to (24), if $|D_j| \geq 16\mathcal{N}_\infty(\lambda)\log(2/\delta)$ and $|D_1| = \cdots = |D_m| = |D|/m$, with the probability at least $1-\delta$ we have

$$\|\bar{f}_{D,\lambda}^M - \widehat{f}_{D,\lambda}^M\|_\rho \leq 4 \left\| C_\lambda^{-1/2}(C - \widehat{C}_{D_j}) C_\lambda^{-1/2} \right\| \left\| C_\lambda^{1/2}(\widehat{f}_{D_j,\lambda}^M - f_\lambda) \right\|.$$

The last step is due to the fact $\left\| C_\lambda^{-1/2}(\widehat{C}_D - C) C_\lambda^{-1/2} \right\| \leq \left\| C_\lambda^{-1/2}(C - \widehat{C}_{D_j}) C_\lambda^{-1/2} \right\|$ where $|D_j| \leq |D|$. ■

A.2.2 Estimate for sample variance

Lemma 24 (Sample variance) *Let $\widehat{f}_{D,\lambda}^M$ and $\widetilde{f}_{D,\lambda}^M$ be defined by (12) and (13). For $\delta \in (0,1)$, if $|D| \geq 16\mathcal{N}_\infty(\lambda)\log(2/\delta)$, with the probability at least $1-\delta$, the local sample variance holds*

$$\|\widehat{f}_{D,\lambda}^M - \widetilde{f}_{D,\lambda}^M\|_\rho \leq \|C_\lambda^{1/2}(\widehat{f}_{D,\lambda}^M - \widetilde{f}_{D,\lambda}^M)\|_K \leq 8B \left(\frac{\sqrt{\mathcal{N}_\infty(\lambda)}}{|D|} + \sqrt{\frac{\mathcal{N}(\lambda)}{|D|}} \right) \log\frac{2}{\delta}. \tag{27}$$

Proof Recall the representations of $\widehat{f}_{D,\lambda}^M$ and $\widetilde{f}_{D,\lambda}^M$ that are

$$\widehat{f}_{D,\lambda}^M = V(V^*\widehat{C}_D V + \lambda I)^{-1} V^* \widehat{S}_D^* \boldsymbol{y}_D, \qquad \widetilde{f}_{D,\lambda}^M = V(V^*\widehat{C}_D V + \lambda I)^{-1} V^* \bar{S}_D^* f_\rho.$$

To simply the representations, we characterize $\widehat{f}_{D,\lambda}^M = G_D \widehat{S}_D^* \boldsymbol{y}_D$ and $\widetilde{f}_{D,\lambda}^M = G_D \bar{S}_D^* f_\rho$ with $G_D = V(V^*\widehat{C}_D V + \lambda I)^{-1} V^*$. Then, from (22), the following inequalities hold

$$\begin{aligned} \|\widehat{f}_{D,\lambda}^M - \widetilde{f}_{D,\lambda}^M\|_\rho \leq & \|C_\lambda^{1/2}(\widehat{f}_{D,\lambda}^M - \widetilde{f}_{D,\lambda}^M)\|_K \leq \|C_\lambda^{1/2} G_D(\widehat{S}_D^* \boldsymbol{y}_D - \bar{S}_D^* f_\rho)\|_K \\ \leq & \|C_\lambda^{1/2}\widehat{C}_{D,\lambda}^{-1/2}\| \|\widehat{C}_{D,\lambda}^{1/2} G_D \widehat{C}_{D,\lambda}^{1/2}\| \|\widehat{C}_{D,\lambda}^{-1/2} C_\lambda^{1/2}\| \|C_\lambda^{-1/2}(\widehat{S}_D^* \boldsymbol{y}_D - \bar{S}_D^* f_\rho)\|. \end{aligned} \tag{28}$$

Then, from Lemma 8 of (Rudi et al., 2015) and Lemma 22, when $|D| \geq 16\mathcal{N}_\infty(\lambda)\log(2/\delta)$, with the probability at least $1-\delta$ we have

$$\|\widehat{f}_{D,\lambda}^M - \widetilde{f}_{D,\lambda}^M\|_\rho \leq 2 \, \|C_\lambda^{-1/2}(\widehat{S}_D^* \boldsymbol{y}_D - \bar{S}_D^* f_\rho)\| \tag{29}$$

Substituting the results in Lemma 25 to (29) with the fact $\|C_\lambda^{-1/2}(\widehat{S}_D^* \boldsymbol{y}_D - \bar{S}_D^* f_\rho)\| \leq \|C_\lambda^{-1/2}(\widehat{S}_D^* \boldsymbol{y}_D - S^* f_\rho)\| + \|C_\lambda^{-1/2}(S^* f_\rho - \bar{S}_D^* f_\rho)\|$, we upper bound the local sample variance with the probability at least $1-\delta$:

$$\|\widehat{f}_{D,\lambda}^M - \widetilde{f}_{D,\lambda}^M\|_\rho \leq \|C_\lambda^{1/2} G_D(\widehat{S}_D^* \boldsymbol{y}_D - \bar{S}_D^* f_\rho)\|_K \leq 8B \left(\frac{\sqrt{\mathcal{N}_\infty(\lambda)}}{|D|} + \sqrt{\frac{\mathcal{N}(\lambda)}{|D|}} \right) \log\frac{2}{\delta}.$$

■

Using Bernstein's inequality and following the proof of Lemma 6 in (Rudi and Rosasco, 2017), we prove the following lemmas to estimate terms in (29).

Lemma 25 *Assume there exists $\kappa \geq 1$ such that $K(\boldsymbol{x}, \boldsymbol{x}) \leq \kappa^2,\ \forall \boldsymbol{x} \in \mathcal{X}$ and $|y| \leq B$. For $\delta \in (0, 1]$, the following holds with the probability at least $1 - \delta$*

$$\|C_\lambda^{-1/2}(\widehat{S}_D^* \boldsymbol{y}_D - S^* f_\rho)\| \leq 2B \left(\frac{\sqrt{\mathcal{N}_\infty(\lambda)}}{|D|} + \sqrt{\frac{\mathcal{N}(\lambda)}{|D|}} \right) \log \frac{2}{\delta},$$

and

$$\|C_\lambda^{-1/2}(S^* f_\rho - \bar{S}_D^* f_\rho)\| \leq 2B \left(\frac{\sqrt{\mathcal{N}_\infty(\lambda)}}{|D|} + \sqrt{\frac{\mathcal{N}(\lambda)}{|D|}} \right) \log \frac{2}{\delta}.$$

Proof Let $\xi_i = C_\lambda^{-1/2} K_{\boldsymbol{x}_i} y_i$ in the Hilbert space $\mathcal{H}_M$. We see that

$$\frac{1}{|D|} \sum_{i=1}^{|D|} \xi_i = \frac{1}{|D|} \sum_{i=1}^{|D|} C_\lambda^{-1/2} K_{\boldsymbol{x}_i} y_i = C_\lambda^{-1/2} \widehat{S}_{D_j}^* \boldsymbol{y}_{D_j},$$
$$\mathbb{E}\, \xi = \int_X C_\lambda^{-1/2} K_x f_\rho(\boldsymbol{x}) d\rho_X(\boldsymbol{x}) = C_\lambda^{-1/2} S^* f_\rho.$$

Thus, the error term to bound can be stated as

$$\|C_\lambda^{-1/2}(\widehat{S}_{D_j}^* \boldsymbol{y}_{D_j} - S^* f_\rho)\| = \left\| \frac{1}{|D|} \sum_{i=1}^{|D|} \xi_i - \mathbb{E}\xi_i \right\|. \tag{30}$$

By Jensen's inequality, we thus have

$$\|\xi_i - \mathbb{E}(\xi_i)\| \leq \|C_\lambda^{-1/2} K_{x_i}\| |y_i| + \mathbb{E}\|C_\lambda^{-1/2} K_{x_i}\| |y_i| \leq 2B\sqrt{\mathcal{N}_\infty(\lambda)}. \tag{31}$$

Note that

$$\begin{aligned} \mathbb{E}\big(\xi_i - \mathbb{E}(\xi_i)\big)^2 &\leq 2 \int_X \|C_\lambda^{-1/2} K_{x_i}\|^2 |y_i|^2 d\rho_X(\boldsymbol{x}) \\ &\leq 2B^2 \int_X \|C_\lambda^{-1/2} K_{x_i}\|^2 d\rho_X(\boldsymbol{x}) \leq 2B^2 \mathcal{N}(\lambda). \end{aligned} \tag{32}$$

Substituting (31) and (32) to (30), by Proposition 3 in (Rudi and Rosasco, 2017), we have

$$\|C_\lambda^{-1/2}(\widehat{S}_{D_j}^* \boldsymbol{y}_{D_j} - S^* f_\rho)\| \leq 2 \left(\frac{B\sqrt{\mathcal{N}_\infty(\lambda)}}{|D|} + \sqrt{\frac{B^2 \mathcal{N}(\lambda)}{|D|}} \right) \log \frac{2}{\delta}.$$

Let $\xi_i = C_\lambda^{-1/2} K_{\boldsymbol{x}_i} f_\rho(\boldsymbol{x}_i)$ on $\mathcal{X}$ in the Hilbert space $\mathcal{H}_M$. We see that

$$\frac{1}{|D|} \sum_{i=1}^{|D|} \xi_i = \frac{1}{|D|} \sum_{i=1}^{|D|} C_\lambda^{-1/2} K_{\boldsymbol{x}_i} f_\rho(\boldsymbol{x}_i) = C_\lambda^{-1/2} \bar{S}_D^* f_\rho,$$
$$\mathbb{E}\xi_i = \int_X C_\lambda^{-1/2} K_x f_\rho(\boldsymbol{x}) d\rho_X(\boldsymbol{x}) = C_\lambda^{-1/2} S^* f_\rho.$$

Thus, the error term to bound can be stated as

$$\|C_\lambda^{-1/2}(S^*f_\rho - \bar{S}_D^* f_\rho)\| = \left\| \frac{1}{|D|}\sum_{i=1}^{|D|}\xi_i - \mathbb{E}\xi_i \right\|. \tag{33}$$

Similarly, using Bernstein's inequality, we have

$$\|C_\lambda^{-1/2}(S^*f_\rho - \bar{S}_D^* f_\rho)\| \le 2\left(\frac{B\sqrt{\mathcal{N}_\infty(\lambda)}}{|D|} + \sqrt{\frac{B^2\mathcal{N}(\lambda)}{|D|}}\right)\log\frac{2}{\delta}.$$

■

A.2.3 Estimate for Nyström error

Lemma 26 (Nyström error) *Let $\widetilde{f}_{D,\lambda}^M$ and $\widetilde{f}_{D,\lambda}$ be defined by (12) and (13). Under Assumption 5 and the condition $|D| \ge 16\mathcal{N}_\infty(\lambda)\log(2/\delta)$, for any $\delta \in (0,1)$, the local Nyström error holds with probability at least $1-\delta$,*

$$\|\widetilde{f}_{D,\lambda}^M - \widetilde{f}_{D,\lambda}\|_\rho \le \|C_\lambda^{1/2}(\widetilde{f}_{D,\lambda}^M - \widetilde{f}_{D,\lambda})\|_K \le \begin{cases} 11R\lambda^{-1/2}\|(I-VV^*)C_\lambda^{1/2}\|^{2r+1}, & \text{when } r \in (0,1/2); \\ 8R\,\|(I-VV^*)C_\lambda^{1/2}\|^{2r}, & \text{when } r \in [1/2,1]. \end{cases}$$

Proof Recall the characterizations of $\widehat{f}_{D_j,\lambda}^M$ and $\widetilde{f}_{D_j,\lambda}$ in Proposition 20, it holds

$$\widetilde{f}_{D,\lambda}^M = V(V^*\widehat{C}_D V + \lambda I)^{-1}V^*\bar{S}_D^* f_\rho, \qquad \widetilde{f}_{D,\lambda} = (\widehat{C}_D + \lambda I)^{-1}\bar{S}_D^* f_\rho.$$

We use $G_D = V(V^*\widehat{C}_D V + \lambda I)^{-1}V^*$ and then $\widetilde{f}_{D,\lambda}^M = G_D\bar{S}_D^* f_\rho$. Using $Z^*f(ZZ^*) = f(Z^*Z)Z^*$, we have

$$\widehat{C}_{D,\lambda}^{-1}\bar{S}_D^* f_\rho = (\bar{S}_D^* S + \lambda I)^{-1}\bar{S}_D^* f_\rho = \bar{S}_D^*(S\bar{S}_D^* + \lambda I)^{-1} f_\rho = \bar{S}_D^*\widehat{L}_{D,\lambda}^{-1} f_\rho.$$

From (22), we estimate the Nyström error as follows with

$$\begin{aligned} &\|\widetilde{f}_{D,\lambda}^M - \widetilde{f}_{D,\lambda}\|_\rho \le \|C_\lambda^{1/2}(\widetilde{f}_{D,\lambda}^M - \widetilde{f}_{D,\lambda})\|_K = \|C_\lambda^{1/2}(G_D - \widehat{C}_{D,\lambda}^{-1})\bar{S}_D^* f_\rho\|_K \\ =&\|C_\lambda^{1/2}(G_D\widehat{C}_{D,\lambda} - I)\widehat{C}_{D,\lambda}^{-1}\bar{S}_D^* f_\rho\|_K = \|C_\lambda^{1/2}(G_D\widehat{C}_{D,\lambda} - I)\bar{S}_D^*\widehat{L}_{D,\lambda}^{-1} f_\rho\|_K. \end{aligned} \tag{34}$$

Then, we bound $\|\widetilde{f}_{D,\lambda}^M - \widetilde{f}_{D,\lambda}\|$ for $r \in (0,1/2)$ and $r \in [1/2,1]$, respectively.

- When $r \in (0,1/2)$, the true regression f_ρ is out of the deduced RKHS $f_\rho \notin \mathcal{H}$.

  Note that, there exists $\|g\| \le R$, $\|L_\lambda^{-1}L\| \le 1$, $\|\widehat{L}_{D,\lambda}^{-1/2}\lambda^{1/2}\| \le 1$, $\|\bar{S}_D^*\widehat{L}_{D,\lambda}^{-1/2}\| \le \|\widehat{L}_{D,\lambda}^{-1/2}\widehat{L}_D\widehat{L}_{D,\lambda}^{-1/2}\|^{1/2} \le 1$ and $\|\widehat{C}_{D,\lambda}^{-1/2}\bar{S}_D^*\| = \|\widehat{C}_{D,\lambda}^{-1/2}\widehat{C}_D\widehat{C}_{D,\lambda}^{-1/2}\|^{1/2} \le 1$. From Lemma 22 and (34), when $|D| \ge 16\mathcal{N}_\infty(\lambda)\log(2/\delta)$, with the probability at least $1-\delta$, we

have

$$\begin{aligned}\|\widetilde{f}^M_{D,\lambda}-\widetilde{f}_{D,\lambda}\|_\rho \leq & \|C_\lambda^{1/2}(\widetilde{f}^M_{D,\lambda}-\widetilde{f}_{D,\lambda})\|_K = \|C_\lambda^{1/2}(G_D\widehat{C}_{D,\lambda}-I)\bar{S}_D^*\widehat{L}_{D,\lambda}^{-1}f_\rho\|_K\\ =&\|C_\lambda^{1/2}(G_D\widehat{C}_{D,\lambda}-I)\bar{S}_D^*\widehat{L}_{D,\lambda}^{r-1}(\widehat{L}_{D,\lambda}^{-1/2}L_\lambda^{1/2})^{2r}(L_\lambda^{-1}L)^r g\|\\ \leq&R\|C_\lambda^{1/2}(G_D\widehat{C}_{D,\lambda}-I)\bar{S}_D^*\widehat{L}_{D,\lambda}^{r-1}(\widehat{L}_{D,\lambda}^{-1/2}L_\lambda^{1/2})^{2r}\|\\ \leq&R\|C_\lambda^{1/2}(G_D\widehat{C}_{D,\lambda}-I)C_\lambda^r(C_\lambda^{-1/2}\widehat{C}_{D,\lambda}^{1/2})^{2r}(\widehat{C}_{D,\lambda}^{-1/2}\bar{S}_D^*)^{2r}\\ &(\bar{S}_D^*\widehat{L}_{D,\lambda}^{-1/2})^{1-2r}(\widehat{L}_{D,\lambda}^{-1/2}\lambda^{1/2})\lambda^{-1/2}(\widehat{L}_{D,\lambda}^{-1/2}L_\lambda^{1/2})^{2r}\|\\ \leq&R\lambda^{-1/2}\|C_\lambda^{-1/2}\widehat{C}_{D,\lambda}^{1/2}\|^{2r}\|\widehat{L}_{D,\lambda}^{-1/2}L_\lambda^{1/2}\|^{2r}\|C_\lambda^{1/2}(G_D\widehat{C}_{D,\lambda}-I)C_\lambda^r\|\\ \leq&2\sqrt{2}R\lambda^{-1/2}\|C_\lambda^{1/2}(G_D\widehat{C}_{D,\lambda}-I)C_\lambda^r\|.\end{aligned} \tag{35}$$

Noting that $G_D\widehat{C}_{D,\lambda}VV* = VV*$, we have

$$\begin{aligned}G_D\widehat{C}_{D,\lambda}-I=&G_D\widehat{C}_{D,\lambda}(I-VV^*)+G_D\widehat{C}_{D,\lambda}VV^*-I\\ =&G_D\widehat{C}_{D,\lambda}(I-VV^*)-(I-VV^*).\end{aligned} \tag{36}$$

Using above identity, we have

$$\begin{aligned}&\|C_\lambda^{1/2}(G_D\widehat{C}_{D,\lambda}-I)C_\lambda^r\|\\ \leq&\|C_\lambda^{1/2}G_D\widehat{C}_{D,\lambda}(I-VV^*)C_\lambda^r\|+\|C_\lambda^{1/2}(I-VV^*)C_\lambda^r\|\\ \leq&\|C_\lambda^{1/2}\widehat{C}_{D,\lambda}^{-1/2}\widehat{C}_{D,\lambda}^{1/2}G_D\widehat{C}_{D,\lambda}^{1/2}\widehat{C}_{D,\lambda}^{1/2}C_\lambda^{-1/2}C_\lambda^{1/2}(I-VV^*)C_\lambda^r\|+\|C_\lambda^{1/2}(I-VV^*)C_\lambda^r\|\\ \leq&\|C_\lambda^{1/2}(I-VV^*)C_\lambda^r\|(\|C_\lambda^{1/2}\widehat{C}_{D,\lambda}^{-1/2}\|\|\widehat{C}_{D,\lambda}^{1/2}G_D\widehat{C}_{D,\lambda}^{1/2}\|\|\widehat{C}_{D,\lambda}^{1/2}C_\lambda^{-1/2}\|+1)\\ \leq&\|C_\lambda^{1/2}(I-VV^*)C_\lambda^r\|(\|C_\lambda^{1/2}\widehat{C}_{D,\lambda}^{-1/2}\|\|\widehat{C}_{D,\lambda}^{1/2}C_\lambda^{-1/2}\|+1).\end{aligned} \tag{37}$$

The last step is due to $\|\widehat{C}_{D,\lambda}^{1/2}G_D\widehat{C}_{D,\lambda}^{1/2}\|\leq 1$ in Lemma 8 of (Rudi et al., 2015).

Next, we estimate $\|C_\lambda^{1/2}(I-VV^*)C_\lambda^r\|$. Since VV^* is a projection operator, it holds for any $s>0$ that $(I-VV^*)=(I-VV^*)^s$, therefore

$$\|C_\lambda^{1/2}(I-VV^*)C_\lambda^r\|\leq\|C_\lambda^{1/2}(I-VV^*)\|\|(I-VV^*)C_\lambda^r\|.$$

Using Cordes inequality (Fujii et al., 1993) to $\|(I-VV^*)C_\lambda^r\|$, we have

$$\|(I-VV^*)C_\lambda^r\|=\|(I-VV^*)^{2r}C_\lambda^{\frac{1}{2}2r}\|=\|(I-VV^*)C_\lambda^{1/2}\|^{2r}.$$

Thus, it holds

$$\|C_\lambda^{1/2}(I-VV^*)C_\lambda^r\|\leq\|(I-VV^*)C_\lambda^{1/2}\|^{2r+1}. \tag{38}$$

Substituting (37) and (38) into (35), under the condition $|D|\geq 16\mathcal{N}_\infty(\lambda)\log(2/\delta)$, for $r\in(0,1/2)$, we have with the probability $1-\delta$

$$\|\widetilde{f}^M_{D,\lambda}-\widetilde{f}_{D,\lambda}\|_\rho\|C_\lambda^{1/2}(\widetilde{f}^M_{D,\lambda}-\widetilde{f}_{D,\lambda})\|_K\leq 11R\lambda^{-1/2}\|(I-VV^*)C_\lambda^{1/2}\|^{2r+1}. \tag{39}$$

- When $r \in [1/2, 1]$, the regression function belongs to the hypothesis space $f_\rho \in \mathcal{H}$.

  Note that, there exists $\|g\| \le R$, $\|L_\lambda^{-1} L\| \le 1$, $\|\bar{S}_D^* \widehat{L}_{D,\lambda}^{-1/2}\| \le \|\widehat{L}_{D,\lambda}^{-1/2} \widehat{L}_D \widehat{L}_{D,\lambda}^{-1/2}\|^{1/2} \le 1$, and $\|\widehat{C}_{D,\lambda}^{-1/2} \bar{S}_D^*\| = \|\widehat{C}_{D,\lambda}^{-1/2} \widehat{C}_D \widehat{C}_{D,\lambda}^{-1/2}\|^{1/2} \le 1$. From Lemma 22 and (34), when $|D| \ge 16\mathcal{N}_\infty(\lambda)\log(2/\delta)$, with the probability at least $1-\delta$, we have

$$
\begin{aligned}
&\|\widetilde{f}_{D,\lambda}^M - \widetilde{f}_{D,\lambda}\|_\rho \le \|C_\lambda^{1/2}(\widetilde{f}_{D,\lambda}^M - \widetilde{f}_{D,\lambda})\|_K \\
\le &\|C_\lambda^{1/2}(G_D \widehat{C}_{D,\lambda} - I)\bar{S}_D^* \widehat{L}_{D,\lambda}^{-1} f_\rho\|_K \\
= &\|C_\lambda^{1/2}(G_D \widehat{C}_{D,\lambda} - I)\bar{S}_D^* \widehat{L}_{D,\lambda}^{r-1}(\widehat{L}_{D,\lambda}^{-1/2} L_\lambda^{1/2})^{2r}(L_\lambda^{-1} L)^r g\| \\
\le &R\|C_\lambda^{1/2}(G_D \widehat{C}_{D,\lambda} - I)\widehat{C}_{D,\lambda}^{r-1/2}(\widehat{C}_{D,\lambda}^{-1/2}\bar{S}_D^*)^{2r-1}(\bar{S}_D^* \widehat{L}_{D,\lambda}^{-1/2})^{2-2r}(\widehat{L}_{D,\lambda}^{-1/2} L_\lambda^{1/2})^{2r}\| \\
\le &2R\|C_\lambda^{1/2}(G_D \widehat{C}_{D,\lambda} - I)C_\lambda^{r-1/2}\|.
\end{aligned} \tag{40}
$$

  Using the identity (36), when $|D| \ge 16\mathcal{N}_\infty(\lambda)\log(2/\delta)$, with the probability at least $1-\delta$, we have

$$
\begin{aligned}
&\|C_\lambda^{1/2}(G_D \widehat{C}_{D,\lambda} - I)C_\lambda^{r-1/2}\| \le \|C_\lambda^{1/2}(I - VV^*)C_\lambda^{r-1/2}\| \\
&\quad + \|C_\lambda^{1/2}\widehat{C}_{D,\lambda}^{-1/2}\widehat{C}_{D,\lambda}^{1/2} G_D \widehat{C}_{D,\lambda}^{1/2}\widehat{C}_{D,\lambda}^{1/2} C_\lambda^{-1/2} C_\lambda^{1/2}(I - VV^*)C_\lambda^{r-1/2}\| \\
&\le \|C_\lambda^{1/2}(I - VV^*)C_\lambda^{r-1/2}\|(1 + \|C_\lambda^{1/2}\widehat{C}_{D,\lambda}^{-1/2}\|\|\widehat{C}_{D,\lambda}^{1/2} G_D \widehat{C}_{D,\lambda}^{1/2}\|\|\widehat{C}_{D,\lambda}^{1/2} C_\lambda^{-1/2}\|) \\
&\le 4\|C_\lambda^{1/2}(I - VV^*)C_\lambda^{r-1/2}\|.
\end{aligned} \tag{41}
$$

  Next, we estimate $\|C_\lambda^{1/2}(I - VV^*)C_\lambda^{r-1/2}\|$. Since VV^* is a projection operator, it holds for any $s > 0$ that $(I - VV^*) = (I - VV^*)^s$, therefore

$$
\|C_\lambda^{1/2}(I - VV^*)C_\lambda^{r-1/2}\| \le \|C_\lambda^{1/2}(I - VV^*)\|\|(I - VV^*)C_\lambda^{r-1/2}\|.
$$

  Using Cordes inequality (Fujii et al., 1993) to $\|(I - VV^*)C_\lambda^{r-1/2}\|$, we have

$$
\|(I - VV^*)C_\lambda^{r-1/2}\| = \|(I - VV^*)^{2r-1} C_\lambda^{\frac{1}{2}2r-1}\| = \|(I - VV^*)C_\lambda^{1/2}\|^{2r-1}.
$$

  Thus, it holds

$$
\|C_\lambda^{1/2}(I - VV^*)C_\lambda^{r-1/2}\| \le \|(I - VV^*)C_\lambda^{1/2}\|^{2r}. \tag{42}
$$

  Substituting (41) and (42) into (40), with the condition $|D| \ge 16\mathcal{N}_\infty(\lambda)\log(2/\delta)$, there exists for $r \in [1/2, 1]$ with the probability $1-\delta$

$$
\|\widetilde{f}_{D,\lambda}^M - \widetilde{f}_{D,\lambda}\|_\rho \le \|C_\lambda^{1/2}(\widetilde{f}_{D,\lambda}^M - \widetilde{f}_{D,\lambda})\|_K \le 8R\,\|(I - VV^*)C_\lambda^{1/2}\|^{2r}. \tag{43}
$$

Then, combining (39) and (43), we prove the desired result. ■

A.2.4 Estimate for empirical error

Lemma 27 (Empirical error) *Let $\widetilde{f}_{D,\lambda}$ and f_λ be defined by (14) and (15). Under the condition $|D| \geq 16\mathcal{N}_\infty(\lambda)\log(2/\delta)$, for any $\delta \in (0,1)$, the local empirical error holds with probability at least $1-\delta$,*

$$\|\widetilde{f}_{D,\lambda} - f_\lambda\|_\rho \leq \|C_\lambda^{1/2}(\widetilde{f}_{D,\lambda} - f_\lambda)\|_K \leq \left(2+\sqrt{2}\right)\|f_\lambda - f_\rho\|.$$

Proof Recall the definitions of $\widetilde{f}_{D,\lambda}$ and f_λ with operators in Proposition 20, it holds

$$\widetilde{f}_{D,\lambda} = \widehat{C}_{D,\lambda}^{-1}\bar{S}_D^* f_\rho, \qquad f_\lambda = C_\lambda^{-1}S^* f_\rho.$$

Using the identity $A^{-1} - B^{-1} = A^{-1}(B-A)B^{-1}$ for positive operators A, B, we have

$$\begin{aligned}
&\|\widetilde{f}_{D,\lambda} - f_\lambda\|_\rho \leq \|C_\lambda^{1/2}(\widetilde{f}_{D,\lambda} - f_\lambda)\|_K \\
\leq&\|C_\lambda^{1/2}\widehat{C}_{D,\lambda}^{-1}\bar{S}_D^* f_\rho - C_\lambda^{1/2}C_\lambda^{-1}S^* f_\rho\| \\
=&\|C_\lambda^{1/2}\widehat{C}_{D,\lambda}^{-1}(\bar{S}_D^* - S^*)f_\rho + C_\lambda^{1/2}(\widehat{C}_{D,\lambda}^{-1} - C_\lambda^{-1})S^* f_\rho\|_K \\
=&\|C_\lambda^{1/2}\widehat{C}_{D,\lambda}^{-1}(\bar{S}_D^* - S^*)f_\rho + C_\lambda^{1/2}\widehat{C}_{D,\lambda}^{-1}(C - \widehat{C}_{D_j})C_\lambda^{-1}S^* f_\rho\| \\
=&\|C_\lambda^{1/2}\widehat{C}_{D,\lambda}^{-1}(\bar{S}_D^* - S^*)f_\rho + C_\lambda^{1/2}\widehat{C}_{D,\lambda}^{-1}(S^*C_\lambda^{1/2} - \bar{S}_D^* S)C_\lambda^{-1}S^* f_\rho\| \\
=&\|C_\lambda^{1/2}\widehat{C}_{D,\lambda}^{-1}(\bar{S}_D^* - S^*)f_\rho + C_\lambda^{1/2}\widehat{C}_{D,\lambda}^{-1}(S^* - \bar{S}_D^*)f_\lambda\| \\
=&\|C_\lambda^{1/2}\widehat{C}_{D,\lambda}^{-1}\bar{S}_D^*(f_\rho - f_\lambda) + C_\lambda^{1/2}\widehat{C}_{D,\lambda}^{-1}S^*(f_\lambda - f_\rho)\| \\
=&\|C_\lambda^{1/2}C_\lambda^{-1/2}C_\lambda^{1/2}\widehat{C}_{D,\lambda}^{-1/2}\widehat{C}_{D,\lambda}^{-1/2}\bar{S}_D^*(f_\rho - f_\lambda) + C_\lambda^{1/2}C_\lambda^{-1/2}C_\lambda^{1/2}\widehat{C}_{D,\lambda}^{-1/2}\widehat{C}_{D,\lambda}^{-1/2}C_\lambda^{1/2}C_\lambda^{-1/2}S^*(f_\lambda - f_\rho)\|.
\end{aligned}$$

Note that $\|SC_\lambda^{-1/2}\| = \|C_\lambda^{-1/2}CC_\lambda^{-1/2}\|^{1/2} \leq 1$, $\|\widehat{C}_{D,\lambda}^{-1/2}\bar{S}_D^*\| = \|\widehat{C}_{D,\lambda}^{-1/2}\widehat{C}_{D_j}\widehat{C}_{D,\lambda}^{-1/2}\|^{1/2} \leq 1$, and $\|C_\lambda^{-1/2}S^*\| = \|C_\lambda^{-1/2}CC_\lambda^{-1/2}\|^{1/2} \leq 1$. Thus, we obtain

$$\|\widetilde{f}_{D,\lambda} - f_\lambda\|_\rho \leq \|C_\lambda^{1/2}(\widetilde{f}_{D,\lambda} - f_\lambda)\|_K \leq \left[\|C_\lambda^{1/2}\widehat{C}_{D,\lambda}^{-1/2}\| + \|C_\lambda^{1/2}\widehat{C}_{D,\lambda}^{-1/2}\|^2\right]\|f_\lambda - f_\rho\|.$$

From Lemma 22, if $|D| \geq 16\mathcal{N}_\infty(\lambda)\log(2/\delta)$, we have $\|C_\lambda^{1/2}\widehat{C}_{D,\lambda}^{-1/2}\| + \|C_\lambda^{1/2}\widehat{C}_{D,\lambda}^{-1/2}\|^2 \leq (2+\sqrt{2})$ with the probability at least $1-\delta$. ■

The empirical error is also related to f_ρ that can be estimated by $f_\rho = L^r g$ with $\|g\| \leq R$. Thus, we estimate the empirical error in terms of $r \in (0, 1/2)$ and $r \in [1/2, 1]$, respectively. To bound the empirical error, the restrictions on n influence the number of partitions m.

A.2.5 Estimate for approximation error

The last term we need to estimate is approximation error $\|f_\lambda - f_\rho\|$, whose proof is standard (Smale and Zhou, 2007; Caponnetto and De Vito, 2007; Rudi and Rosasco, 2017).

Lemma 28 (Approximation error) *Let f_λ and f_ρ be defined by (15) and (7). Under Assumption 5, the approximation error holds for any $\lambda > 0$ and $r > 0$,*

$$\|f_\lambda - f_\rho\| \leq R\lambda^r. \tag{44}$$

Proof Under Assumption 5, there exists $g \in L^2_{\rho_X}$ such that $f_\rho = L^r g$ with $\|g\| \leq R$. The identity $A(A+\lambda I)^{-1} = I - \lambda(A+\lambda I)^{-1}$ is valid for $\lambda > 0$ and A the bounded self-adjoint positive operator and by the definition of f_λ (Proposition 20), we have

$$\begin{aligned}\|f_\lambda - f_\rho\| =& \|LL_\lambda^{-1} f_\rho - f_\rho\| = \|(LL_\lambda^{-1} - I) f_\rho\| = \|\lambda L_\lambda^{-1} f_\rho\| \\ =& \|\lambda^r (\lambda^{1-r} L_\lambda^{-(1-r)})(L_\lambda^{-r} L^r) g\| \\ \leq& \|\lambda^r\| \|\lambda^{1-r} L_\lambda^{-(1-r)}\| \|L_\lambda^{-r} L^r\| \|g\|.\end{aligned}$$

Note that $\|\lambda^{1-r} L_\lambda^{-(1-r)}\| \leq 1$ and $\|L_\lambda^{-r} L^r\| \leq 1$, while $R := \|g\|_{L^2_{\rho_X}}$ according to Assumption 5. The proof is completed. ■

The estimate of approximation error is standard and holds for any $r > 0$. When r approaches zero, the approximation error gradually becomes the distance between two unrelated estimators f_λ and f_ρ.

A.3 Proofs of Main Results in Expectation

Proof of Lemma 10. Using the triangle inequalities, we have

$$\begin{aligned}\|\bar{f}^M_{D,\lambda} - f_\rho\|^2_\rho \leq& 4 \left\| \sum_{j=1}^m \frac{|D_j|}{|D|} \left(\widehat{f}^M_{D_j,\lambda} - \widetilde{f}^M_{D_j,\lambda} \right) \right\|^2_\rho + 4 \left\| \sum_{j=1}^m \frac{|D_j|}{|D|} \left(\widetilde{f}^M_{D_j,\lambda} - \widetilde{f}_{D_j,\lambda} \right) \right\|^2_\rho \\ &+ 4 \left\| \sum_{j=1}^m \frac{|D_j|}{|D|} \widetilde{f}_{D_j,\lambda} - f_\lambda \right\|^2_\rho + 4 \|f_\lambda - f_\rho\|^2_\rho .\end{aligned} \tag{45}$$

The local sample sets $\{(\boldsymbol{X}_1, \boldsymbol{y}_1), \cdots, (\boldsymbol{X}_m, \boldsymbol{y}_m)\}$ are independently sampled from $\rho_{X\times Y}$. Note that from Proposition 5 in (Chang et al., 2017), there exists

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

By taking the expectation with respect to ρ_X and $|D_1| = \cdots = |D_m| = |D|/m$, we have

$$\mathbb{E} \left\| \sum_{j=1}^m \frac{|D_j|}{|D|} \left(\widehat{f}^M_{D_j,\lambda} - \widetilde{f}^M_{D_j,\lambda} \right) \right\|^2_\rho = \sum_{j=1}^m \frac{|D_j|^2}{N^2} \mathbb{E} \|\widehat{f}^M_{D_j,\lambda} - \widetilde{f}^M_{D_j,\lambda}\|^2_\rho = \frac{1}{m} \mathbb{E} \|\widehat{f}^M_{D_j,\lambda} - \widetilde{f}^M_{D_j,\lambda}\|^2_\rho . \tag{46}$$

Following the proof of Lemma 1 in (Lin and Cevher, 2018), by Hölder's inequality, we know that

$$\begin{aligned}\mathbb{E}\left\|\sum_{j=1}^{m}\frac{|D_j|}{|D|}\left(\widetilde{f}^M_{D_j,\lambda}-\widetilde{f}_{D_j,\lambda}\right)\right\|_\rho^2 &= \frac{1}{m^2}\mathbb{E}\left\|\sum_{i=1}^{m}(\widetilde{f}^M_{D_j,\lambda}-\widetilde{f}_{D_j,\lambda})\right\|_\rho^2 \\ &\le \frac{1}{m}\mathbb{E}\sum_{i=1}^{m}\left\|\widetilde{f}^M_{D_j,\lambda}-\widetilde{f}_{D_j,\lambda}\right\|_\rho^2 = \mathbb{E}\left\|\widetilde{f}^M_{D_j,\lambda}-\widetilde{f}_{D_j,\lambda}\right\|_\rho^2.\end{aligned} \tag{47}$$

Similarly, we derive relationship between global and local empirical errors

$$\begin{aligned}\mathbb{E}\left\|\sum_{j=1}^{m}\frac{|D_j|}{|D|}\widetilde{f}_{D_j,\lambda}-f_\lambda\right\|_\rho^2 &= \frac{1}{m^2}\mathbb{E}\left\|\sum_{i=1}^{m}(\widetilde{f}_{D_j,\lambda}-f_\lambda)\right\|_\rho^2 \\ &\le \frac{1}{m}\mathbb{E}\sum_{i=1}^{m}\left\|\widetilde{f}_{D_j,\lambda}-f_\lambda\right\|_\rho^2 = \mathbb{E}\left\|\widetilde{f}_{D_j,\lambda}-f_\lambda\right\|_\rho^2.\end{aligned} \tag{48}$$

Substituting (46), (47) and (48) to (45), we obtain the result in (8). ■

Proof of Theorem 11. From the error decomposition (8), we estimate the local error terms $\|\bar{f}^M_{D,\lambda}-\widetilde{f}^M_{D,\lambda}\|_\rho^2$, $\|\widetilde{f}^M_{D,\lambda}-\widetilde{f}_{D,\lambda}\|_\rho^2$, $\|\widetilde{f}_{D,\lambda}-f_\lambda\|_\rho^2$, and $\|f_\lambda-f_\rho\|_\rho^2$, respectively.

Estimate the local sample variance. According to Lemma 24, under Assumptions 6 and 7, if $2r+2\gamma\ge\alpha$ and $\lambda=N^{\frac{-1}{2r+\gamma}}$, when the local sample size is large enough $|D_j|\ge 16\mathcal{N}_\infty(\lambda)\log(2/\delta)$, it holds with the probability at least $1-\delta$

$$\begin{aligned}\frac{1}{m}\mathbb{E}\|\widehat{f}^M_{D_j,\lambda}-\widetilde{f}^M_{D_j,\lambda}\|_\rho^2 \le& \frac{1}{m}\mathbb{E}\|C_\lambda^{1/2}(\widehat{f}^M_{D_j,\lambda}-\widetilde{f}^M_{D_j,\lambda})\|_K^2 \\ \le& \frac{64B^2}{m}\left(\frac{\sqrt{\mathcal{N}_\infty(\lambda)}}{|D_j|}+\sqrt{\frac{\mathcal{N}(\lambda)}{|D_j|}}\right)^2\log^2\frac{2}{\delta} \\ \le& 128B^2\left(\frac{\mathcal{N}_\infty(\lambda)}{|D||D_j|}+\frac{\mathcal{N}(\lambda)}{|D|}\right)\log^2\frac{2}{\delta} \\ \le& 128B^2\left(\frac{C_1}{16\log(2/\delta)}+C_0\right)N^{\frac{-2r}{2r+\gamma}}\log^2\frac{2}{\delta}.\end{aligned} \tag{49}$$

Estimate the local Nyström error. Combing the results in Lemma 26, Lemma 6 of (Rudi et al., 2015) and Lemma 7 of (Rudi et al., 2015), when $|D_j|\ge 16\mathcal{N}_\infty(\lambda)\log(2/\delta)$, $M\ge 67\log\frac{4\kappa^2}{\lambda\delta}\vee 5\mathcal{N}_\infty(\lambda)\log\frac{4\kappa^2}{\lambda\delta}$ for the uniform sampling and $M\ge 334\log\frac{8|D_j|}{\delta}\vee 78\mathcal{N}(\lambda)\log\frac{8|D_j|}{\delta}$ for the data-dependent sampling, we obtain local Nyström error with the probability at least $1-2\delta$ that

$$\mathbb{E}\|\widetilde{f}^M_{D,\lambda}-\widetilde{f}_{D,\lambda}\|_\rho^2\le\|C_\lambda^{1/2}(\widetilde{f}^M_{D,\lambda}-\widetilde{f}_{D,\lambda})\|_K^2\le 1089R^2\lambda^{2r}=1089R^2N^{\frac{-2r}{2r+\gamma}}. \tag{50}$$

Estimate the empirical error. According Lemmas 27 and 28, when the sample size satisfies $|D|\ge 16\mathcal{N}_\infty(\lambda)\log(2/\delta)$, there holds with the probability at least $1-\delta$

$$\|\widetilde{f}_{D,\lambda}-f_\lambda\|_\rho^2\le 16R^2\lambda^{2r}=16R^2N^{\frac{-2r}{2r+\gamma}}. \tag{51}$$

Substituting (44), (49), (50) and (51) to (8), we prove the following result. If $M \geq 67 \log \frac{4\kappa^2}{\lambda\delta} \vee 5\mathcal{N}_\infty(\lambda) \log \frac{4\kappa^2}{\lambda\delta}$ for uniform sampling, $M \geq 334 \log \frac{8|D_j|}{m\delta} \vee 78\mathcal{N}(\lambda) \log \frac{8|D_j|}{m\delta}$ for leverage scores sampling, and $m \lesssim N^{\frac{2r+\gamma-\alpha}{2r+\gamma}}$, then, with probability $1-4\delta$, there exists

$$\mathbb{E}\|\bar{f}_{D,\lambda}^M - f_\rho\|_\rho^2 \leq C_2 N^{\frac{-2r}{2r+\gamma}} \log^2(2/\delta),$$

where $C_2 = 512B^2 \left(\frac{C_1}{16\log(2/\delta)} + C_0\right) + 4424R^2$. ■

A.4 Proofs of Main Results in Probability

Proof of Theorem 16. To derive the excess risk of `DNyström` in probability, we recall its error decomposition in probability in Lemma 15. We first estimate the global error terms on the entire set D and then the distributed error. Let $\lambda = N^{\frac{-1}{2r+\gamma}}$, $\delta > 0$ and $|D_j| \geq 16\mathcal{N}_\infty(\lambda)\log(2/\delta)$.

Estimate the sample variance. According to Lemma 24, under Assumptions 6 and 7, it holds with the probability at least $1-\delta$

$$\begin{aligned}\|\widehat{f}_{D,\lambda}^M - \widetilde{f}_{D,\lambda}^M\|_\rho \leq \|C_\lambda^{1/2}(\widehat{f}_{D,\lambda}^M - \widetilde{f}_{D,\lambda}^M)\|_K \leq & 8B\left(\frac{\sqrt{\mathcal{N}_\infty(\lambda)}}{|D|} + \sqrt{\frac{\mathcal{N}(\lambda)}{|D|}}\right)\log\frac{2}{\delta} \\ \leq & 8B(\sqrt{C_0}+\sqrt{C_1})N^{\frac{-r}{2r+\gamma}}\log\frac{2}{\delta}.\end{aligned} \tag{52}$$

Estimate the Nyström error. Combing the results in Lemma 26, Lemma 6 of (Rudi et al., 2015) and Lemma 7 of (Rudi et al., 2015), when $M \geq 67\log\frac{4\kappa^2}{\lambda\delta} \vee 5\mathcal{N}_\infty(\lambda)\log\frac{4\kappa^2}{\lambda\delta}$ for the uniform sampling and $M \geq 334\log\frac{8|D|}{\delta} \vee 78\mathcal{N}(\lambda)\log\frac{8|D|}{\delta}$ for the data-dependent sampling, we obtain the Nyström error with the probability at least $1-2\delta$ that

$$\|\widetilde{f}_{D,\lambda}^M - \widetilde{f}_{D,\lambda}\|_\rho \leq \|C_\lambda^{1/2}(\widetilde{f}_{D,\lambda}^M - \widetilde{f}_{D,\lambda})\|_K \leq 33R\lambda^r = 33RN^{\frac{-r}{2r+\gamma}}. \tag{53}$$

Estimate the empirical error. According to Lemmas 27 and 28, there holds with the probability at least $1-\delta$

$$\|\widetilde{f}_{D,\lambda} - f_\lambda\| \leq 4R\lambda^r = 4RN^{\frac{-r}{2r+\gamma}}. \tag{54}$$

Estimate the distributed error. From the result in Lemma 23, we find that the distributed error is related to $\|C_\lambda^{1/2}(\widehat{f}_{D_j,\lambda}^M - f_\lambda)\|$. From (23), (52), (53) and (54), when $M \geq 67\log\frac{4\kappa^2}{\lambda\delta} \vee 5\mathcal{N}_\infty(\lambda)\log\frac{4\kappa^2}{\lambda\delta}$ for the uniform sampling and $M \geq 334\log\frac{8|D_j|}{\delta} \vee 78\mathcal{N}(\lambda)\log\frac{8|D_j|}{\delta}$

for the data-dependent sampling, with the probability at least $1-4\delta$, we have

$$\begin{aligned}
&\|\bar{f}^M_{D,\lambda} - \widehat{f}^M_{D,\lambda}\| \\
\leq& 4\left\|C_\lambda^{-1/2}(C-\widehat{C}_{D_j})C_\lambda^{-1/2}\right\| \|C_\lambda^{1/2}(\widehat{f}^M_{D_j,\lambda} - f_\lambda)\| \\
\leq& 4\left\|C_\lambda^{-1/2}(C-\widehat{C}_{D_j})C_\lambda^{-1/2}\right\| \Big(\|C_\lambda^{1/2}(\widehat{f}^M_{D_j,\lambda} - \widetilde{f}^M_{D_j,\lambda})\| \\
&\qquad + \|C_\lambda^{1/2}(\widetilde{f}^M_{D_j,\lambda} - \widetilde{f}_{D_j,\lambda})\| + \|C_\lambda^{1/2}(\widetilde{f}_{D_j,\lambda} - f_\lambda)\|\Big) \\
\leq& 4\left\|C_\lambda^{-1/2}(C-\widehat{C}_{D_j})C_\lambda^{-1/2}\right\| \left(8B\left(\frac{\sqrt{\mathcal{N}_\infty(\lambda)}}{|D_j|} + \sqrt{\frac{\mathcal{N}(\lambda)}{|D_j|}}\right)\log\frac{2}{\delta} + 37R\lambda^r\right) \qquad (55) \\
\leq& 32B\left\|C_\lambda^{-1/2}(C-\widehat{C}_{D_j})C_\lambda^{-1/2}\right\| \left(\frac{\sqrt{\mathcal{N}_\infty(\lambda)}}{|D_j|} + \sqrt{\frac{\mathcal{N}(\lambda)}{|D_j|}}\right) + 74R\lambda^r \\
\leq& 64B\left(\frac{\mathcal{N}_\infty(\lambda)}{|D_j|} + \sqrt{\frac{\mathcal{N}_\infty(\lambda)}{|D_j|}}\right)\left(\frac{\sqrt{\mathcal{N}_\infty(\lambda)}}{|D_j|} + \sqrt{\frac{\mathcal{N}(\lambda)}{|D_j|}}\right)\log(2/\delta) + 74R\lambda^r.
\end{aligned}$$

The last two steps are due to the results in Lemma 22 and Lemma 21, respectively. Using the condition $|D_j| \geq 16\mathcal{N}_\infty(\lambda)\log(2/\delta)$, we have $\frac{\mathcal{N}_\infty(\lambda)}{|D_j|} \leq \sqrt{\frac{\mathcal{N}_\infty(\lambda)}{|D_j|}}$. Then, under Assumptions 6 and 7, if $\lambda = N^{\frac{-1}{2r+\gamma}}$ and $|D_1| = |D_2| = \cdots = |D_m| = |D|/m$, we have

$$\begin{aligned}
\frac{\mathcal{N}_\infty(\lambda)}{|D_j|} + \sqrt{\frac{\mathcal{N}_\infty(\lambda)}{|D_j|}} &\leq C_1 m N^{\frac{\alpha-2r-\gamma}{2r+\gamma}} + \sqrt{C_1 m}N^{\frac{\alpha-2r-\gamma}{4r+2\gamma}}, \\
\frac{\sqrt{\mathcal{N}_\infty(\lambda)}}{|D_j|} + \sqrt{\frac{\mathcal{N}(\lambda)}{|D_j|}} &\leq \sqrt{C_1} m N^{\frac{\alpha-4r-2\gamma}{4r+2\gamma}} + \sqrt{C_0 m}N^{\frac{-r}{2r+\gamma}}.
\end{aligned}$$

Then, when $m \lesssim N^{\frac{2r+\gamma-\alpha}{4r+2\gamma}}$, we have

$$\left(\frac{\mathcal{N}_\infty(\lambda)}{|D_j|} + \sqrt{\frac{\mathcal{N}_\infty(\lambda)}{|D_j|}}\right)\left(\frac{\sqrt{\mathcal{N}_\infty(\lambda)}}{|D_j|} + \sqrt{\frac{\mathcal{N}(\lambda)}{|D_j|}}\right) \leq \sqrt{C_0 C_1}N^{\frac{-r}{2r+\gamma}}. \qquad (56)$$

Combing (55) and (56), under the same conditions, with the probability at least $1-4\delta$, we have

$$\|\bar{f}^M_{D,\lambda} - \widehat{f}^M_{D,\lambda}\| \leq (256BC_1\sqrt{C_0C_1} + 74R)N^{\frac{-r}{2r+\gamma}}\log(2/\delta). \qquad (57)$$

Since $m \lesssim N^{\frac{2r+\gamma-\alpha}{4r+2\gamma}}$ is more strict than the condition $|D_j| \leq 16\mathcal{N}_\infty(\lambda)\log(2/\delta)$, we impose the condition on the number of partitions as $m \lesssim N^{\frac{2r+\gamma-\alpha}{4r+2\gamma}}$. Substituting (44), (52), (53), (54) and (57) to (9), we prove the result.

Assume there exists $\kappa > 1$ such that $K(\boldsymbol{x},\boldsymbol{x}) \leq \kappa^2$ for any $\boldsymbol{x} \in \mathcal{X}$ and $|y| \leq B$. Let $\delta > 0$, $\lambda = N^{\frac{-1}{2r+\gamma}}$ and $|D_1| = \cdots = |D|_m = |D|/m$. Under Assumptions 5 – 7 with $r \in (0,1]$ and $\gamma \in [0,1]$, if $2r+\gamma \geq \alpha$, $M \geq 67\log\frac{4\kappa^2}{\lambda\delta} \vee 5\mathcal{N}_\infty(\lambda)\log\frac{4\kappa^2}{\lambda\delta}$ for uniform sampling, $M \geq 334\log\frac{8|D_j|}{m\delta} \vee 78\mathcal{N}(\lambda)\log\frac{8|D_j|}{m\delta}$ for leverage scores sampling, and $m \lesssim N^{\frac{2r+\gamma-\alpha}{4r+2\gamma}}$, then

with probability $1-4\delta$, there exists

$$\|\bar{f}_{D,\lambda}^{M}-f_{\rho}\|_{\rho}\leq C_{3}N^{\frac{-r}{2r+\gamma}}\log(2/\delta),$$

where $C_3=8B(\sqrt{C_0}+\sqrt{C_1})+256BC_1\sqrt{C_0C_1}+112R$. ■

References

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.

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.

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

Daniele Calandriello, Alessandro Lazaric, and Michal Valko. Distributed adaptive sampling for kernel matrix approximation. In *Artificial Intelligence and Statistics*, pages 1421–1429. PMLR, 2017.

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

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.

Yifan Chen and Yun Yang. Fast statistical leverage score approximation in kernel ridge regression. In *International Conference on Artificial Intelligence and Statistics*, pages 2935–2943. PMLR, 2021.

Petros Drineas, Malik Magdon-Ismail, Michael W Mahoney, and David P Woodruff. Fast approximation of matrix coherence and statistical leverage. *The Journal of Machine Learning Research*, 13(1):3475–3506, 2012.

Junichi Fujii, Masatoshi Fujii, Takayuki Furuta, and Ritsuo Nakamoto. Norm inequalities equivalent to heinz inequality. *Proceedings of the American Mathematical Society*, 118 (3):827–830, 1993.

Alex Gittens and Michael W Mahoney. Revisiting the nyström method for improved large-scale machine learning. *Journal of Machine Learning Research*, 17(1):3977–4041, 2016.

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

Martin Hanke. *Conjugate gradient type methods for ill-posed problems*. Routledge, 2017.

Magnus Rudolph Hestenes and Eduard Stiefel. *Methods of conjugate gradients for solving linear systems*, volume 49. NBS Washington, DC, 1952.

Cho-Jui Hsieh, Kai-Wei Chang, Chih-Jen Lin, S Sathiya Keerthi, and Sellamanickam Sundararajan. A dual coordinate descent method for large-scale linear svm. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pages 408–415. ACM, 2008.

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

Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Sampling methods for the nyström method. *Journal of Machine Learning Research*, 13(Apr):981–1006, 2012.

Jason D Lee, Ruoqi Shen, Zhao Song, Mengdi Wang, et al. Generalized leverage score sampling for neural networks. *Advances in Neural Information Processing Systems*, 33: 10775–10787, 2020.

Jian Li, Yong Liu, and Weiping Wang. Distributed learning with random features. *arXiv preprint arXiv:1906.03155*, 2019.

Heng Lian, Jiamin Liu, and Zengyan Fan. Distributed learning for sketched kernel regression. *Neural Networks*, 143:368–376, 2021.

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.

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.

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.

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.

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

Shuai Lu, Peter Mathé, and Sergiy Pereverzyev Jr. Analysis of regularized nyström subsampling for regression functions of low smoothness. *Analysis and Applications*, 17(06): 931–946, 2019.

Longda Ma, Lei Shi, and Zongmin Wu. Nyström subsampling method for coefficient-based regularized regression. *Inverse Problems*, 35(7):075002, 2019.

Siyuan Ma and Mikhail Belkin. Diving into the shallows: a computational perspective on large-scale shallow learning. *Advances in neural information processing systems*, 30, 2017.

Siyuan Ma and Mikhail Belkin. Kernel machines that adapt to gpus for effective large batch training. *Proceedings of Machine Learning and Systems*, 1:360–373, 2019.

Martin Fodslette Møller. A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks*, 6(4):525–533, 1993.

Cameron Musco and Christopher Musco. Recursive sampling for the nyström method. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 3836–3848, 2017.

John Platt. Fastmap, metricmap, and landmark mds are all nyström algorithms. In *International Workshop on Artificial Intelligence and Statistics*, pages 261–268. PMLR, 2005.

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.

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

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.

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.

Alessandro Rudi, Luigi Carratino, and Lorenzo Rosasco. Falkon: An optimal large scale kernel method. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 3888–3898, 2017.

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

Yousef Saad. *Iterative methods for sparse linear systems*, volume 82. siam, 2003.

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

Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming*, 127(1):3–30, 2011.

S Sivananthan et al. Manifold regularization based on nyström type subsampling. *Applied and Computational Harmonic Analysis*, 49(1):152–179, 2020.

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

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Verlag, 2008.

Hongwei Sun and Qiang Wu. Optimal rates of distributed regression with imperfect kernels. *Journal of Machine Learning Research*, 22:171–1, 2021.

Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer Science & Business Media, 1999.

Shusen Wang. A sharper generalization bound for divide-and-conquer ridge regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5305–5312, 2019.

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.

Rong Yin, Yong Liu, Lijing Lu, Weiping Wang, and Dan Meng. Divide-and-conquer learning with nyström: Optimal rate and algorithm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6696–6703, 2020.

Kai Zhang, Ivor W Tsang, and James T Kwok. Improved nyström low-rank approximation and error analysis. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pages 1232–1239. ACM, 2008.

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.