Optimal Convergence Rates for Agnostic Nyström Kernel Learning

Jian Li[1] **Yong Liu**[2 3] **Weiping Wang**[1]

Abstract

Nyström low-rank approximation has shown great potential in processing large-scale kernel matrix and neural networks. However, there lacks a unified analysis for Nyström approximation, and the asymptotical minimax optimality for Nyström methods usually require a strict condition, assuming that the target regression lies exactly in the hypothesis space. In this paper, to tackle these problems, we provide a refined generalization analysis for Nyström approximation in the agnostic setting, where the target regression may be out of the hypothesis space. Specifically, we show Nyström approximation can still achieve the capacity-dependent optimal rates in the agnostic setting. To this end, we first prove the capacity-dependent optimal guarantees of Nyström approximation with the standard uniform sampling, which covers both loss functions and applies to some agnostic settings. Then, using data-dependent sampling, for example, leverage scores sampling, we derive the capacity-dependent optimal rates that apply to the whole range of the agnostic setting. To our best knowledge, the capacity-dependent optimality for the whole range of the agnostic setting is first achieved and novel in Nyström approximation.

1. Introduction

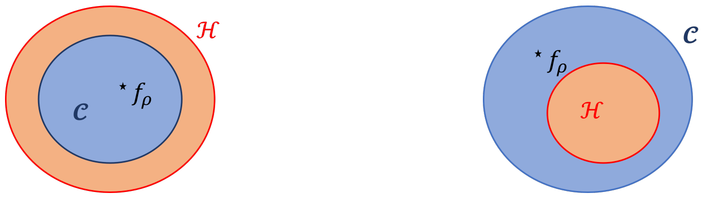
In statistical learning theory, only a limited number of input-output pairs can be observed from a fixed but unknown distribution. As one of the most popular nonparametric statistical approaches, the kernel method offers an elegant paradigm and solid theoretical guarantees (Vapnik, 1999; Shawe-Taylor & Cristianini, 2004). Despite its excellent theoretical properties, the kernel method is unfeasible in large-scale settings because of high computational and storage requirements. To overcome these scalability issues, researchers improved kernel method with accelerated techniques: distributed learning (Zhang et al., 2015; Lin et al., 2017), Nyström approximation (Williams & Seeger, 2001; Rudi et al., 2015) and random features (Rahimi & Recht, 2007; Rudi & Rosasco, 2017) to alleviate memory bottlenecks via low-rank approximation, as well as stochastic and preconditioned extensions (Raskutti et al., 2014; Lin & Rosasco, 2016) to improve computational efficiency via iterative solutions.

Recent theoretical works have extensively studied the statistical properties of kernel methods together with distributed learning (Zhang et al., 2013; Guo et al., 2017a; Lin & Cevher, 2020), Nyström approximation (Bach, 2013; Alaoui & Mahoney, 2015; Rudi et al., 2015; 2017), random features (Rudi & Rosasco, 2017; Sun et al., 2018; Li et al., 2019a; Liu et al., 2021) and stochastic optimization (Carratino et al., 2018; Lin & Cevher, 2018). Specifically, the optimal theoretical guarantees for kernel ridge regression (KRR) (Caponnetto & De Vito, 2007; Smale & Zhou, 2007) and its variants has attracted increasing attentions in statistical learning, including Nyström approximation (Rudi et al., 2015), Nyström approximation with stochastic optimization (Rudi et al., 2017) and Nyström approximation with data-dependent sampling (Rudi et al., 2018), respectively. These optimal guarantees for KRR rely on a strict assumption that the concept (the true regression) f_ρ lies in the hypothesis space $f_\rho \in \mathcal{H}$ associated with the selected kernel. As shown in Figure 1, since the joint distribution is unknown, it's hard to select a perfect kernel to guarantee the realistic setting $f_\rho \in \mathcal{H}$ via prior knowledge or kernel selection. In practice, as stated in PAC theories (Maass, 1994; Auer et al., 1995), the selected kernel is usually imperfect but good enough to approximate the true regression and thus leads to the agnostic learning $f_\rho \notin \mathcal{H}$. Therefore, the statistical guarantees in the agnostic setting are of practical and theoretical interest in the context of kernel methods.

Nyström methods sample landmarks to generate low-rank approximation of large matrix. Remarkably, Nyström method is not only used in kernel matrix approximation but also has shown significant potentials in accelerating complex neural network models, such as Nyströmformer (Xiong et al., 2021) and Nyström attention (Samarakoon

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China [2] Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China [3] Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing, China. Correspondence to: Yong Liu <liuyonggsai@ruc.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).



(a) Realistic setting

(b) Agnostic setting

Figure 1. In in the realistic setting, the concept class is a subset of hypothesis space $\mathcal{C} \subseteq \mathcal{H}$ and thus the concept belongs to the hypothesis space $f_\rho \in \mathcal{H}$. In the agnostic setting, the concept class is larger than the hypothesis space and there exists the situation $f_\rho \notin \mathcal{H}$.

& Leung, 2022). Therefore, in this work, we focus on the optimal statistical guarantees for Nyström approximation in the agnostic learning. From the perspective of generalization analysis, the sharp generalization error bounds in expectation for a fixed design regression setting were derived in (Bach, 2013; Alaoui & Mahoney, 2015), and then (Rudi et al., 2015) extended the former literature with KRR and high probability estimates. However, these optimal statistical guarantees assumed that the target function belongs to the reproducing kernel Hilbert space (RKHS) generated by the selected kernel, which only applied to the realistic setting. Recent works (Kriukova et al., 2017; Lu et al., 2019) derived the generalization guarantees for Nyström approximation in the agnostic setting, but they are either capacity-independent (Kriukova et al., 2017) or suboptimal (Lu et al., 2019). However, the existing capacity-independent optimality works in agnostic learning faced a fatal drawback: Since it doesn't measure the capacity of the RKHS, the capacity-independent learning rates obtained in the above literature are suboptimal when the size of the RKHS is small. *Therefore, the capacity-dependent optimality for Nyström approximation with general loss functions in agnostic settings is rather important but still an open problem. In this paper, we aim to derive general optimal statistical guarantees for Nyström approximation, including the following improvements: 1) The optimal statistical guarantees apply to both realistic settings and agnostic settings; 2) The optimal rates are capacity-dependent, where the capacity-independent results are the special cases.*

1.1. Related Work

The related work includes: Nyström approximation and leverage scores sampling.

1) Nyström approximation. Nyström approximation is a common tool to approximate kernel matrix with low-rank decomposition (Williams & Seeger, 2001; Drineas et al., 2012). The approximation ability of Nyström approaches has been theoretically analyzed by many literature (Drineas & Mahoney, 2005; Drineas et al., 2012; Gittens & Mahoney, 2013; Cohen et al., 2015). From the perspective of generalization ability, the optimal generalization properties in expectation were achieved for Nyström approximation for fixed design regression with uniform sampling in (Bach, 2013) and with data-dependent sampling (Alaoui & Mahoney, 2015), while (Rudi et al., 2015) extended these results to the random design and the high probability estimates. Nyström approximation was incorporated with preconditioned conjugate gradient (PCG) method 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).

2) Leverage scores sampling. Statistical leverage scores that measure the matrix coherence have also proved crucial recently in the development of improved worst-case randomized matrix algorithms (Boutsidis et al., 2009; Drineas et al., 2012). To accelerate the computation of leverage scores, researchers proposed several approximate leverage scores algorithms, including recursive sampling (Musco & Musco, 2017), SQUEAK (Calandriello et al., 2017), BLESS (Rudi et al., 2018) and spectral analysis (Chen & Yang, 2021).

3) Agnostic Kernel Learning. The capacity-independent optimal results for the agnostic kernel learning have been established for KRR (Smale & Zhou, 2007), random features (Sun et al., 2018), Nyström approximation (Kriukova et al., 2017) and distributed kernel ridge regression (DKRR) (Sun & Wu, 2021). In particular, recent works (Kriukova et al., 2017; Lu et al., 2019) studied the generalization properties of Nyström approximation for the agnostic setting, namely low smoothness of target function f_ρ, but the convergence rates of these results are capacity-independent (Kriukova et al., 2017) or suboptimal (Lu et al., 2019). The capacity-dependent optimal results of Nyström approximation for agnostic learning have been scarcely studied.

1.2. Contributions

In this paper, we aim to provide optimal theoretical guarantees to the agnostic setting for KRR-Nyström. For the sake of comparison, we first restate the existing generalization bound for KRR-Nyström (Rudi et al., 2015). With tighter estimates for the similarity between empirical and expected covariance operators and the sample variance, we refine the theoretical guarantees for KRR-Nyström, which improves applicability to the agnostic setting. Specifically, we derive the theoretical guarantees for uniform sampling (the worst case) and data-dependent sampling (the benign case), respectively. We only present the main results in the main paper and leave the proofs in the appendix [1].

1) On the statistical front: capacity-dependent optimality for the agnostic setting. The existing work on capacity-dependent optimal rates for KRR methods (Caponnetto & De Vito, 2007; Rudi et al., 2015; Guo et al., 2017b) focused on the easy problems in the realistic setting, assuming the target regression lies in the induced kernel space. However, the target regression is usually out of the induced kernel space for complicated tasks. The existing works studied the statistical properties for KRR methods, but the rates are either capacity-independent (Smale & Zhou, 2007; Kriukova et al., 2017) or suboptimal (Lu et al., 2019). In this paper, by relaxing the restriction on the regularity assumption, we prove the capacity-dependent optimality for KRR-Nyström that applies to both realistic and agnostic settings.

2) On the computational front: tradeoffs of accelerated techniques. The classical Nyström approximation methods (Rudi et al., 2015; 2017; 2018) focused on the number of Nyström centers. In this paper, we throughout study the computational efficiency with Nyström approximation, data-dependent sampling, and PCG.

3) Novel proof techniques. Using explicit intermediate estimators, we introduce a novel error decomposition for the excess risk to achieve tighter generalization analysis for KRR-Nyström. The similarity of covariance operators and sample variance are two bottleneck factors to relax the restriction on the applicability to the agnostic setting. Instead of second order decomposition of the operator similarity in (Guo et al., 2017b; Lin et al., 2017), we directly estimate it via concentration inequalities for self-adjoint operators. We also consider a tighter estimate for the maximal effective dimension via data-dependent sampling, which guarantees tighter estimate by data-dependent sampling.

2. Backgrounds

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 $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ is drawn i.i.d from $\mathcal{X} \times \mathcal{Y}$ w.r.t. ρ. For the regression tasks, the output space is $\mathcal{Y} = \mathbb{R}$. We denote $\mathcal{H}$ a reproducing kernel Hilbert space (RKHS) (Steinwart & Christmann, 2008) induced by a Mercer kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that $\mathcal{H} = \overline{\text{span}\{K_x | \boldsymbol{x} \in \mathcal{X}\}}$ completed with

$$\langle K_x, K_{\boldsymbol{x}'} \rangle_K = K(\boldsymbol{x}, \boldsymbol{x}') \quad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}.$$

Here, the inner product in $\mathcal{H}$ is denoted as $\langle \cdot, \cdot \rangle_K$ and the corresponding norm $\| \cdot \|_K$.

2.1. Kernel Ridge Regression (KRR)

KRR is a standard nonparametric regression in supervised learning (Vapnik, 1999; Shawe-Taylor & Cristianini, 2000), 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) illustrate that KRR admits a unique closed form solution

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

where $\mathbf{K}_{nn} = [K(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1}^n$ is the $n \times n$ kernel matrix and $\boldsymbol{y}_n = [y_1, \cdots, y_n]^\top$.

Although KRR has been proven with optimal learning bounds (Smale & Zhou, 2007; Caponnetto & De Vito, 2007), it is unfeasible in large-scale settings due to high computational complexity. Precisely, KRR requires $\mathcal{O}(n^3)$ time to solve the inverse of $\mathbf{K}_{nn} + \lambda n \mathbf{I}$ and $\mathcal{O}(n^2)$ space to storage $\mathbf{K}_{nn}$. To reduce the computational costs, we introduce KRR with Nyström approximation (KRR-Nyström).

2.2. KRR with Nyström Method (KRR-Nyström)

To relieve the computational bottlenecks, several approximate approaches were incorporated with KRR that reduce the computational complexity while keeping the optimal generalization properties, including distributed learning (Zhang et al., 2015; Lin et al., 2017), Nyström subsampling (Rudi et al., 2015) and random features (Rudi & Rosasco, 2017). However, the computational efficiency of those work can be further improved by their combinations.

Nyström methods use $\{\tilde{\boldsymbol{x}}_1, \cdots, \tilde{\boldsymbol{x}}_M\}$ the subset of the input points from n the training samples and $M \leq n$ (Williams & Seeger, 2001). The solution of KRR-Nyström is

$$\widehat{f}_{M,\lambda}(\boldsymbol{x}) = \sum_{i=1}^M \alpha_i K(\tilde{\boldsymbol{x}}_i, \boldsymbol{x}) \qquad \text{with} \\ \boldsymbol{\alpha} = (\mathbf{K}_{nM}^\top \mathbf{K}_{nM} + \lambda n \mathbf{K}_{MM})^\dagger \mathbf{K}_{nM}^\top \boldsymbol{y}_n, \tag{3}$$

where $\dagger$ denotes the Moore-Penrose pseudoinverse of a matrix, and $(\mathbf{K}_{nM})_{ij} = K(\boldsymbol{x}_i, \tilde{\boldsymbol{x}}_j)$, $(\mathbf{K}_{MM})_{kj} = K(\boldsymbol{x}_k, \boldsymbol{x}_j)$

[1] Available at https://lijian.ac.cn/files/2023/2023_ICML_Nystroem.pdf

with $i \in \{1, \cdots, n\}$ and $j, k \in \{1, \cdots, M\}$. Nyström methods are different from the sampling strategies to select the input subset, mainly including uniform sampling and sampling with approximate leverage scores (Rudi et al., 2015; 2017). The sharp generalization properties of Nyström approximation with different sampling strategies have been proven, including both uniform sampling (Bach, 2013) and data-dependent sampling (Alaoui & Mahoney, 2015). The closed-form solution of KRR-Nyström requires $\mathcal{O}(nM^2 + M^3)$ time and $\mathcal{O}(nM)$ space.

2.3. Sampling Strategies for Nyström Landmarks

The sampling strategy for Nyström landmarks $\{\tilde{\boldsymbol{x}}_1, \cdots, \tilde{\boldsymbol{x}}_M\}$ is crucial to the approximation ability of Nyström methods. Alaoui et al. proposed the leverage sampling for Nyström approximation (Alaoui & Mahoney, 2015). Let $n \in \mathbb{R}, \lambda > 0$. Let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ be the training points and the exact leverage scores are defined by

$$l_\lambda(i) = \left(\mathbf{K}_{nn}(\mathbf{K}_{nn} + \lambda n \mathbf{I})^{-1}\right)_{ii}, \quad \forall\, i \in [n]. \tag{4}$$

In practice, the compute of leverage scores is overly time consuming. Recent work considered approximate leverage scores $\widehat{l}_\lambda(i)$, for example Recursive-RLS (Musco & Musco, 2017), SQUEAK (Calandriello et al., 2017), BLESS (Rudi et al., 2018), and SA (Chen & Yang, 2021). We recall the Recursive-RLS which defined the approximate leverage scores as $\widehat{l}_\lambda(i) = \frac{2}{3}(B(B^\top \hat{S}\hat{S}^\top B + \lambda I)^{-1}B^\top)_{ii}$, where $\hat{S}$ is the sample matrix given by Recursive-RLS approach and $BB^\top = K$. Nyström centers are selected according to the probability $p_i = \frac{\widehat{l}_\lambda(i)}{\sum_{i=1}^n \widehat{l}_\lambda(i)}$. To measure the sampling complexity for approximate leverage score approaches, we introduce the empirical effective dimension.

Definition 2.1 (Empirical effective dimension (Alaoui & Mahoney, 2015)**.** The exact leverage scores $l_\lambda(i), \forall i \in [n]$ are defined in (4), and then we define empirical effective dimension as

$$\begin{aligned}
\widetilde{\mathcal{N}}(\lambda) &= \sum_{i=1}^n l_\lambda(i) = \operatorname{Tr}\left(\mathbf{K}_{nn}(\mathbf{K}_{nn} + \lambda n \mathbf{I})^{-1}\right), \\
\widetilde{\mathcal{N}}_\infty(\lambda) &= n \max_{1 \le i \le n} l_\lambda(i) \\
&= n \left\| \operatorname{diag}\left(\mathbf{K}_{nn}(\mathbf{K}_{nn} + \lambda n \mathbf{I})^{-1}\right)\right\|_\infty.
\end{aligned}$$

As illustrated in BLESS (Rudi et al., 2018), the time complexity for approximate leverage scores sampling is $\mathcal{O}(\widetilde{\mathcal{N}}(\lambda)^2/\lambda)$, which often plays an dominant role for data-dependent sampling Nyström methods. The above empirical averaged effective dimension $\widetilde{\mathcal{N}}(\lambda)$ was proposed to analyze the generalization performance of Nyström approximation with uniform sampling (Bach, 2013), while the empirical maximal effective dimension $\widetilde{\mathcal{N}}_\infty(\lambda)$ was used to derive the generalization performance of Nyström approximation with data-dependent sampling (Alaoui & Mahoney, 2015).

3. Main Results

In this section, we first recover the existing bounds for KRR-Nyström in (Rudi et al., 2015). We then present our theoretical results for KRR-Nyström, which applies to the agnostic settings that the true regression may not lie in the hypothesis space. We then consider the worst case (with uniform sampling) of the main result, which applies to one half agnostic settings. We finally derive the benign case (with data-dependent sampling) that characterizes significant computational gains and pertains to the whole range of source condition (all agnostic settings).

The learning target of KRR is to find a predictor that minimizes the expected risk

$$\mathcal{E}(f) = \int_{\mathcal{X}\times\mathcal{Y}} (f(\boldsymbol{x}) - y)^2 d\rho(\boldsymbol{x}, y). \tag{5}$$

The regression function that minimizes the expected risk over all measurable functions $f : \mathcal{X} \to \mathbb{R}$ is given by

$$f_\rho(\boldsymbol{x}) = \int_{\mathcal{Y}} y d\rho(y|\boldsymbol{x}), \qquad \forall \boldsymbol{x} \in \mathcal{X}. \tag{6}$$

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 ρ_X, where the $L^2_{\rho_X}$-norm is defined as $\|f\|_\rho^2 = \langle f, f\rangle_\rho = \int_X |f(\boldsymbol{x})|^2 d\rho_X(\boldsymbol{x}),\ \forall f \in L^2_{\rho_X}$. The generalization ability of a KRR estimator $f \in L^2_{\rho_X}$ is measured by the *excess risk*, i.e. $\mathcal{E}(f) - \mathcal{E}(f_\rho)$. Throughout this paper, we assume the outputs are bounded almost surely for some constant $B > 0$ and $\mathcal{X}$ is compact, which implies $\|f_\rho\|_\infty \le B$. We assume $K(\boldsymbol{x}, \boldsymbol{x}) \le \kappa^2 < \infty, \forall\, \boldsymbol{x} \in \mathcal{X}$.

Definition 3.1 (Integral operator and covariance operator)**.** We define the integral operator $L : L^2_{\rho_X} \to L^2_{\rho_X}$ and the covariance operator $C : \mathcal{H} \to \mathcal{H}$ as

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

The covariance matrix $\widehat{C}_n : \mathcal{H} \to \mathcal{H}$, $\widehat{C}_n = \frac{1}{n}\sum_{i=1}^n K_{x_i} \otimes K_{x_i}$ is the empirical version of covariance operator.

Definition 3.2 (Expected effective dimension)**.** 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 define the following quantities

$$\mathcal{N}(\lambda) = \mathbb{E}\, \mathcal{N}_{\boldsymbol{x}}(\lambda), \quad \mathcal{N}_\infty(\lambda) = \sup_{\boldsymbol{x} \in \mathcal{X}} \mathcal{N}_{\boldsymbol{x}}(\lambda)$$

The quantity $\mathcal{N}(\lambda) = \mathrm{Tr}(C(C+\lambda I)^{-1}) = \mathrm{Tr}(L(L+\lambda I)^{-1})$ is also called degree of freedom, which has been widely used to measure the capacity of the RKHS $\mathcal{H}$ in expectation (Zhang, 2002; Caponnetto & De Vito, 2007; Bach, 2013; Della Vecchia et al., 2021). The maximal effective dimension $\mathcal{N}_\infty(\lambda)$ is used to provide a uniform bound on the leverage scores and it holds $\mathcal{N}(\lambda) \leq \mathcal{N}_\infty(\lambda) \leq \kappa^2\lambda^{-1}$ (Rudi et al., 2015).

Assumption 3.3 (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 smoothness of f_ρ (Lu et al., 2019). If the target function is with high smoothness $r \in [1/2, 1]$, corresponding to the realistic setting $f_\rho \in \mathcal{H}$ where the problems are easier. The low smoothness case $r \in (0, 1/2)$ corresponds to the agnostic setting $f_\rho \notin \mathcal{H}$ that are more general in real tasks. Only with the source condition, KRR-type methods achieve the optimal capacity-independent rate $\mathcal{O}(n^{\frac{-2r}{2r+1}})$ (Smale & Zhou, 2007; Kriukova et al., 2017; Sun & Wu, 2021).

Remark 3.4. From Assumption 3.3, the concept class can be stated as $\mathcal{C} = L^r(L^2_{\rho_X})$ and the hypothesis space $\mathcal{H} = L^{1/2}(L^2_{\rho_X})$ (Steinwart & Christmann, 2008). We denote the concept class as $\mathcal{C} = L^r(L^2_{\rho_X})$ where the concept class contains the target regression $f_\rho \in \mathcal{C}$. We also define $\mathcal{H}_\rho = \{f : \mathcal{X} \to \mathbb{R} | f(x) = \langle w, K_x\rangle, \ \forall w \in \mathcal{H}\}$ as the projection of the RKHS $\mathcal{H}$ on $L^2_{\rho_X}$. Specifically, because $L^a(L^2_{\rho_X}) \subseteq L^b(L^2_{\rho_X})$ when $a \geq b$ (Lin & Rosasco, 2016; 2017), the bigger r is, the stricter the assumption is. We discuss Assumption 3.3 in four cases:

- If $r = 0$, we make no assumption due to $\|f_\rho\|_\rho \leq R$.
- If $r = 1/2$, since $\mathcal{H}_\rho = L^{1/2}(L^2_{\rho_X})$ (Rosasco & Villa, 2015), we obtain $f_\rho \in \mathcal{H}_\rho$.
- If $r \in (1/2, 1]$, using the fact $L^a(L^2_{\rho_X}) \subseteq L^b(L^2_{\rho_X})$ when $a \geq b$, we have $\mathcal{C} = L^r(L^2_{\rho_X}) \subseteq L^{1/2}(L^2_{\rho_X}) = \mathcal{H}_\rho$ and thus $f_\rho \in \mathcal{H}_\rho$, as shown in Figure 1 (a).
- If $r \in (0, 1/2)$, using $L^a(L^2_{\rho_X}) \subseteq L^b(L^2_{\rho_X})$ when $a \geq b$, we have $\mathcal{H}_\rho = L^{1/2}(L^2_{\rho_X}) \subseteq L^r(L^2_{\rho_X}) = \mathcal{C}$ which exists $f_\rho \notin \mathcal{H}_\rho$, as shown in Figure 1 (b).

Note that, many integral operator theory work makes the assumption on the existence of the minimizer $\mathcal{E}(f_\mathcal{H}) = \min_{h\in\mathcal{H}} \mathcal{E}(f)$ and the excess risk is stated as $\mathcal{E}(f) - \mathcal{E}(f_\mathcal{H})$ rather than $\mathcal{E}(f) - \mathcal{E}(f_\rho)$. Under the assumption $r \in [1/2, 1]$, f_ρ belongs to $\mathcal{H}$, such that $f_\mathcal{H} = f_\rho$ (Steinwart & Christmann, 2008). The optimal learning guarantees for Nyström (Rudi et al., 2015) only pertain to $f_\rho \in \mathcal{H}$ in the realistic setting $r \in [1/2, 1]$, assuming the problem cannot be too difficult. In this paper, we employ the target regression f_ρ instead of $f_\mathcal{H}$ and study the optimal statistical guarantees on both the realistic and agnostic setting $r \in (0, 1]$.

Assumption 3.5 (Capacity assumption)**.** Assume $Q > 0$ and $\gamma \in [0, 1]$, such that

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

The above capacity assumption is satisfied when the eigenvalues σ_i of covariance operator C decays polynominally $\sigma_i \leq i^{-1/\gamma}$. More examples for the capacity assumption are referred to (Alaoui & Mahoney, 2015; Rudi et al., 2015). Note that, Assumption 3.5 always holds for the capacity-independent case $\gamma = 1$ as the covariance operator C is trace class. Under Assumptions 3.3 and 3.5, KRR and its accelerated variants reach the minimax optimal capacity-dependent rate $O(n^{\frac{-2r}{2r+\gamma}})$ (Caponnetto & De Vito, 2007; Rudi et al., 2015; Rudi & Rosasco, 2017; Guo et al., 2017b). However, the conventional optimal generalization analysis for kernel methods focused on the realistic setting (Caponnetto & De Vito, 2007; Rudi et al., 2015; Guo et al., 2017b).

3.1. Existing Theoretical Results for KRR-Nyström

Proposition 3.6 (Nyström approximation with the squared loss, Theorem 1 in (Rudi et al., 2015))**.** *Assume there exists* $\mathcal{E}(f_\mathcal{H}) = \min_{f\in\mathcal{H}} \mathcal{E}(f)$, $K(\boldsymbol{x}, \boldsymbol{x}) \leq \kappa^2$ *with* $\kappa \in [1, +\infty)$ *and the outputs are bounded. Under Assumptions 3.3 and 3.5, if*

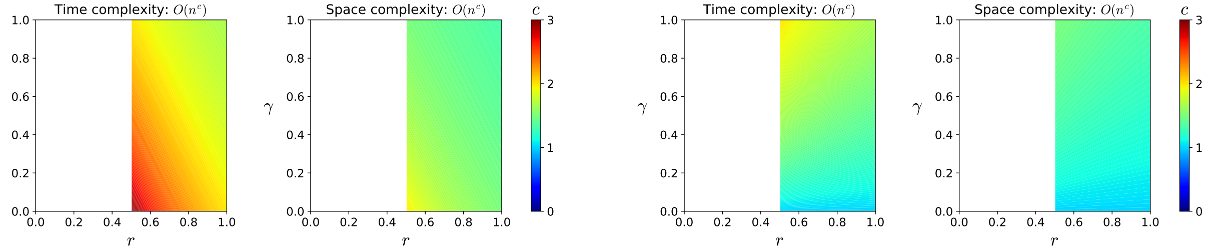
$$r \in [1/2, 1], \quad \gamma \in [0, 1],$$

and $\lambda = n^{-\frac{1}{2r+\gamma}}$*, then with a high probability the following conditions* $M \gtrsim \mathcal{N}_\infty(\lambda)$ *for uniform sampling and* $M \gtrsim \mathcal{N}(\lambda)$ *for data-dependent sampling are sufficient to guarantee the optimal learning rate, respectively*

$$\mathcal{E}(\widehat{f}_{M,\lambda}) - \mathcal{E}(f_\mathcal{H}) = \mathcal{O}\left(n^{-\frac{2r}{2r+\gamma}}\right).$$

Although the existing generalization bound for KRR-Nyström above achieved the optimal learning rates $\mathcal{O}\big(n^{-\frac{2r}{2r+\gamma}}\big)$ (Caponnetto & De Vito, 2007; Wainwright, 2019), it only pertained to the realistic setting $r \in [1/2, 1]$.

To qualify the computational efficiency of Nyström approximation, we compute computational complexities for KRR as an example. Since the fact $M \leq n$, the computation of KRR-Nyström (3) requires $\mathcal{O}(nM^2)$ time to solve the linear system and $\mathcal{O}(nM)$ space to store $\mathbf{K}_{nM}$. Therefore,



(a) Uniform sampling

(b) Leverage scores sampling

Figure 2. Computational complexities and applicable area of KRR-Nyström in Proposition 3.6 with the uniform sampling (left) and with the leverage scores sampling (right).

with $M \gtrsim \mathcal{N}_\infty(\lambda)$ and the fact $\mathcal{N}_\infty(\lambda) \leq \kappa^2/\lambda$, the computational complexities for KRR-Nyström with uniform sampling (Bach, 2013) are:

$$\text{Time: } \mathcal{O}\Big(n^{\frac{2r+\gamma+2}{2r+\gamma}}\Big), \qquad \text{Space: } \mathcal{O}\Big(n^{\frac{2r+\gamma+1}{2r+\gamma}}\Big). \tag{7}$$

Data-dependent sampling introduces additional computations, for example, BLESS requires to compute leverage scores of time $\mathcal{O}(\widetilde{\mathcal{N}}(\lambda)^2/\lambda)$ (Rudi et al., 2018).

Since Proposition 3.6 assumes that $r \geq 1/2$ and $\gamma \in [0,1]$, it holds the fact $2r + 3\gamma \geq 1 + 2\gamma$ and thus the computation of the closed-form KRR-Nyström $\mathcal{O}(n^{\frac{2r+3\gamma}{2r+\gamma}})$ dominates the computational complexity. We depict the computational complexities of Proposition 3.6 in Figure 2.

Remark 3.7. The existing bounds for KRR-Nyström (Rudi et al., 2015) has one fatal drawback: the above theoretical result is only applicable to the realistic setting $r \in [1/2, 1]$ assuming $f_\rho \in \mathcal{H}$, and fails to apply to the agnostic setting $r \in (0, 1/2)$ where the concept f_ρ may not belongs to the hypothesis space $\mathcal{H}$.

3.2. Refined Results for KRR-Nyström

To relax the restriction $r \geq 1/2$, we introduce the compatibility assumption for tighter estimate of maximal effective dimension $\mathcal{N}_\infty(\lambda)$. We introduce high probability excess risk bounds in following theorems for KRR-Nyström.

Assumption 3.8 (Compatibility assumption)**.** Assume there exists $\alpha \in [\gamma, 1]$ and $F > 0$, such that

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

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. Since the covariance operator C is a trace class, Assumption 3.8 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 & 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 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 & 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 $\{\boldsymbol{x}_i\}_{i=1}^n$.

Theorem 3.9. *Assume $\kappa > 1$ such that $K(\boldsymbol{x}, \boldsymbol{x}) \leq \kappa^2$ and the outputs are bounded. Under Assumption 3.3, 3.5 and 3.8, if $\gamma \in [0,1]$, $r \in (0,1], 2r + \gamma \geq \alpha$ and $\lambda = n^{-\frac{1}{2r+\gamma}}$, then the following condition*

$$M \gtrsim n^{\frac{\alpha}{2r+\gamma}},$$

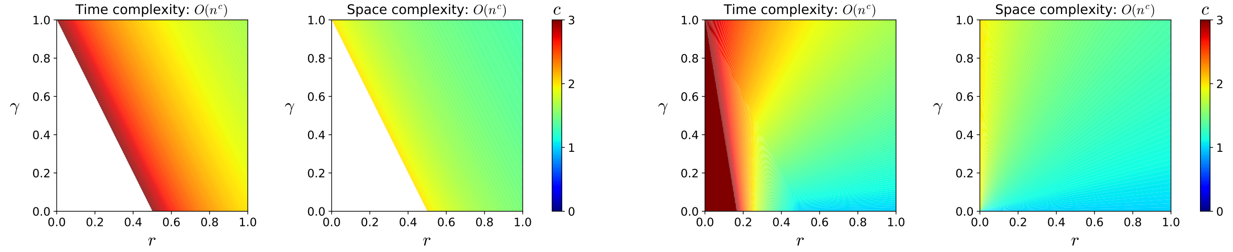
are sufficient to guarantee, with a high probability, that

$$\mathcal{E}(\widehat{f}_{M,\lambda}) - \mathcal{E}(f_\rho) = \mathcal{O}\left(n^{-\frac{2r}{2r+\gamma}}\right).$$

Here $\widehat{f}_{M,\lambda}$ is the estimator of KRR-Nyström (3) *and f_ρ is the target regression.*

Compared to Theorem 1 in (Rudi et al., 2015), the optimal learning guarantees for KRR-Nyström in Theorem 3.9 pertain to the agnostic setting $r \in (0, 1]$ with the condition $2r + \gamma \geq \alpha$, beyond the realistic setting $r \in [1/2, 1]$ for the first time. Due to the fact $\gamma \leq \alpha \leq 1$, both the worst case (uniform sampling) and the benign case (data-dependent sampling) are special cases of Theorem 3.9.

Remark 3.10. Compared with the existing work in KRR (Guo et al., 2017a; Müecke, 2019), KRR-Nyström (Rudi et al., 2015) and KRR-Nyström (Rudi et al., 2015), we relax the strict restriction from $r \geq 1/2$ to $2r + \gamma \geq \alpha$, applying



(a) Uniform sampling

(b) Leverage scores sampling

Figure 3. Computational complexities and applicable area of KRR-Nyström with the uniform sampling (Corollary 3.12, left) and with the leverage scores sampling (Corollary 3.15, right).

to the agnostic setting that the target regression f_ρ may be out of the kernel space $\mathcal{H}$. The improvements come from novel proof techniques: 1) We introduce tighter estimate of the key quantity $\|(C+\lambda I)^{-1/2}(\widehat{C}_n+\lambda I)^{1/2}\|$ via the concentration inequality for self-adjoint operators, leading to the restriction $2r+\gamma \geq \alpha$; 2) Using Assumption 3.8 and Bennett's inequality, we estimate sample variance tightly and relax the constraint from $2r+2\gamma \geq 1$ to $2r+2\gamma \geq \alpha$. Combining the restrictions $2r+\gamma \geq \alpha$ and $2r+2\gamma \geq \alpha$, we obtain that $2r+\gamma \geq \alpha$.

Using $\Omega(n^{\frac{\alpha}{2r+\gamma}})$ Nyström centers, the computational complexities of Theorem 3.9 are:

$$\text{Time} : \mathcal{O}\big(n^{\frac{2r+\gamma+2\alpha}{2r+\gamma}}\big), \quad \text{Space} : \mathcal{O}\big(n^{\frac{2r+\gamma+\alpha}{2r+\gamma}}\big). \qquad (8)$$

Remark 3.11 (Discussion about the integration with PCG). Nyström approximation was often integrated with preconditioned conjugate gradient (PCG) methods to reduce the time complexity, for example FALKON (Rudi et al., 2017), and FALKON-BLESS (Rudi et al., 2018). The use of FALKON still remains the optimal guarantees. The time complexity of KRR-Nyström with PCG is $\mathcal{O}(nMt+M^3)=\mathcal{O}(n^{\frac{2r+\gamma+\alpha}{2r+\gamma}}+n^{\frac{3\alpha}{2r+\gamma}})$, where we omit the log term $t=\log(n)$. Therefore, PCG can improve the computational efficiency for the closed-form solution of KRR-Nyström. However, to achieve smaller effective dimension $\mathcal{N}_\infty(\lambda)$ with $\alpha=\gamma$, we make use of leverage scores sampling and it consumes $\mathcal{O}(\widetilde{\mathcal{N}}(\lambda)^2/\lambda)=\mathcal{O}(n^{\frac{1+2\gamma}{2r+\gamma}})$ time to sample Nyström centers. From Theorem 3.9, the sampling complexity dominates the computational costs for the data-dependent sampling for the agnostic setting $r \in (0,1)$.

Corollary 3.12 (Nyström approximation with uniform sampling)**.** *Assume $K(\boldsymbol{x},\boldsymbol{x}) \leq \kappa^2$, $\forall \kappa > 1$ and the outputs are bounded. Under Assumptions 3.3 and 3.5, if*

$$r \in (0,1], \quad \gamma \in [0,1], \quad 2r+\gamma \geq 1$$

and $\lambda = n^{-\frac{1}{2r+\gamma}}$, then the condition $M \gtrsim n^{\frac{1}{2r+\gamma}}$, is sufficient to guarantee, with a high probability, that

$$\mathcal{E}(\widehat{f}_{M,\lambda}) - \mathcal{E}(f_\rho) = \mathcal{O}\left(n^{-\frac{2r}{2r+\gamma}}\right).$$

Here $\widehat{f}_{M,\lambda}$ is the estimator of KRR-Nyström (3) *with uniform sampling and f_ρ is the target regression.*

Instead of $r \in [1/2,1]$, the optimal guarantees apply to the agnostic setting with the constraint $2r+\gamma \geq 1$. We report the computational complexities and the applicability for Corollary 3.12 in the left of Figure 3.

Remark 3.13 (Nyström approximation in the agnostic setting). Recent work also studied the low smoothness of Nyström subsample (Kriukova et al., 2017; Lu et al., 2019) for misspecified models (the agnostic setting in this work). However, the learning rates in their works are either capacity-independent (Kriukova et al., 2017) or suboptimal (Lu et al., 2019). Those two low smoothness studies on Nyström approximation are special cases of Corollary 3.12.

Remark 3.14 (Beyond square loss). The statistical-computational tradeoffs of low-rank approximation for kernel methods have ben recently explored for the Lipschitz loss, including Nyström approximation (Della Vecchia et al., 2021) and random features (Li et al., 2019b; 2021; Yashima et al., 2021; Li, 2022). However, these works focused on the realistic setting where the target funciton belongs to the RKHS and requires more assumptions, i.e. Bernstein condition and fast eigendecay. Specifically, Nyström approximation with convex Lipschitz loss functions (Della Vecchia et al., 2021) where it only considered a special source condition $r=1/2$. The proof techniques presented here can also be used to prove the fast rates for Lipschitz loss functions. For example, motivated by (Sun et al., 2018; Li et al., 2021; Li, 2022), one can bridge the excess risk for Lipschitz losses with the squared error by introducing an intermedia estimator. We leave the fast rates for agnostic Nyström kernel learning with the Lipschitz loss functions in future work.

Corollary 3.15 (Nyström approximation with data-dependent sampling)**.** *Assume $K(\boldsymbol{x},\boldsymbol{x}) \leq \kappa^2$, $\forall \kappa > 1$ and the outputs are bounded. Under Assumptions 3.3 and 3.5, if*

$$r \in (0,1], \qquad \gamma \in [0,1]$$

and $\lambda = n^{-\frac{1}{2r+\gamma}}$, then the following condition $M \gtrsim n^{\frac{\gamma}{2r+\gamma}}$

Table 1. Summary of statistical and computational properties for related work.

Approaches	Regularity condition	Capacity condition	# Random centers M	Learning rate
KRR (Caponnetto & De Vito, 2007)	$r \in [1/2, 1]$	$\gamma \in [0, 1]$	$\times$	$n^{\frac{-2r}{2r+\gamma}}$
KRR (Smale & Zhou, 2007)	$r \in (0, 1]$	$\gamma = 1$	$\times$	$n^{\frac{-2r}{(2r \vee 1)+1}}$
RF-Uniform (Rudi & Rosasco, 2017)	$r \in [1/2, 1]$	$\gamma \in [0, 1]$	$n^{\frac{(2r-1)\gamma+1}{2r+\gamma}}$	$n^{\frac{-2r}{2r+\gamma}}$
RF-Leverage (Rudi & Rosasco, 2017)	$r \in [1/2, 1]$	$\gamma \in [0, 1]$	$n^{\frac{2r+\gamma-1}{2r+\gamma}}$	$n^{\frac{-2r}{2r+\gamma}}$
Nyström-Uniform (Rudi et al., 2015)	$r \in [1/2, 1]$	$\gamma \in [0, 1]$	$n^{\frac{1}{2r+\gamma}}$	$n^{\frac{-2r}{2r+\gamma}}$
Nyström-Leverage (Rudi et al., 2015)	$r \in [1/2, 1]$	$\gamma \in [0, 1]$	$n^{\frac{\gamma}{2r+\gamma}}$	$n^{\frac{-2r}{2r+\gamma}}$
FALKON-Uniform (Rudi et al., 2017)	$r \in [1/2, 1]$	$\gamma \in [0, 1]$	$n^{\frac{1}{2r+\gamma}}$	$n^{\frac{-2r}{2r+\gamma}}$
FALKON-Leverage (Rudi et al., 2018)	$r \in [1/2, 1]$	$\gamma \in [0, 1]$	$n^{\frac{\gamma}{2r+\gamma}}$	$n^{\frac{-2r}{2r+\gamma}}$
Nyström (Kriukova et al., 2017)	$r \in (0, 1]$	$\gamma = 1$	$\sqrt{n}$	$n^{\frac{-2r}{2r+1}}$
Nyström (Lu et al., 2019)	$r \in (0, 1]$	$\gamma \in [0, 1]$	n	$n^{\frac{\beta-1}{2}}$
DKRR-CM (Lin et al., 2020)	$r \in (0, 1]$, $2r + \gamma \geq 1$	$\gamma \in [0, 1]$	$\times$	$n^{\frac{-2r}{2r+\gamma}}$
Nyström (Theorem 3.9)	$r \in (0, 1]$, $2r + \gamma \geq \alpha$	$\gamma \in [0, 1]$	$n^{\frac{\alpha}{2r+\gamma}}$	$n^{\frac{-2r}{2r+\gamma}}$
Nyström-Uniform (Corollary 3.12)	$r \in (0, 1]$, $2r + \gamma \geq 1$	$\gamma \in [0, 1]$	$n^{\frac{1}{2r+\gamma}}$	$n^{\frac{-2r}{2r+\gamma}}$
Nyström-Leverage (Corollary 3.15)	$r \in (0, 1]$	$\gamma \in [0, 1]$	$n^{\frac{\gamma}{2r+\gamma}}$	$n^{\frac{-2r}{2r+\gamma}}$

Here, $\alpha \in [\gamma, 1]$, "RF" represents the random features methods, "Uniform" denotes the uniform sampling, "Leverage" represents the data-dependent sampling and "DKRR-CM" represents distributed kernel ridge regression (DKRR) with multiple communications.

is sufficient to guarantee, with a high probability, that

$$\mathcal{E}(\widehat{f}_{M,\lambda}) - \mathcal{E}(f_\rho) = \mathcal{O}\left(n^{-\frac{2r}{2r+\gamma}}\right).$$

Here $\widehat{f}_{M,\lambda}$ is the estimator of KRR-Nyström (3) with leverage scores sampling (4) and f_ρ is the target regression.

As shown in Corollary 3.15, we remove the restriction on the range of source condition and extend the optimal theoretical guarantees to the entire agnostic setting $r \in (0, 1/2]$. The time complexity for data-dependent sampling is $\mathcal{O}\Big(n^{\frac{2r+3\gamma}{2r+\gamma}} + n^{\frac{1+2\gamma}{2r+\gamma}}\Big)$. As shown in the right of Figure 3, Corollary 3.15 provides significant computational gains and apply to all agnostic settings via data-dependent sampling.

4. Comparison and Discussion

In this section, we compare this work with the existing theoretical work for KRR and discuss the technical contributions. Compared with related work in Table 1, the theoretical findings remove the strict condition $r \geq 1/2$ and enlarge the applicability area of Nyström approaches:

- **Capacity-dependent optimality.** Compared to recent Nyström approximation in the agnostic setting, this work achieves the capacity-dependent optimality in all agnostic setting, while the learning rate of (Kriukova et al., 2017) is capacity-independent and that of (Lu et al., 2019) is suboptimal.
- **Uniform sampling.** The existing low-rank approximation literature (Rudi et al., 2015; 2017; Rudi & Rosasco, 2017) with uniform sampling only applies to the realistic setting $r \in [1/2, 1]$. We extend the optimal rates for the uniform sampling from the realistic setting $r \in [1/2, 1]$ to a part of agnostic setting $2r + \gamma \geq 1$, where the sample complexity is related to the maximal effective dimension $\mathcal{N}_\infty(\lambda)$ (Bach, 2013).
- **Data-dependent sampling.** Using data-dependent sampling for Nyström centers, one can prove $\mathcal{N}_\infty(\lambda) = \mathcal{N}(\lambda) \leq \mathcal{O}(\lambda^{-\gamma})$ (Alaoui & Mahoney, 2015; Rudi & Rosasco, 2017; Rudi et al., 2017; 2018). We enlarge the applicable area of the optimal rates from only the realistic setting $r \in [1/2, 1]$ to the whole range of the source condition $r \in (0, 1]$.
- **Compared to DKRR-CM (Lin et al., 2020).** Our paper studied the optimality of the Nyström method, while (Lin et al., 2020) focused on the optimality of DKRR that cannot be directly applied to the Nyström method. Our study has a broader applicable area and relaxed constraints from $2r+\gamma \leq 1$ (Lin et al., 2020) to $2r+\gamma \leq \alpha$, where $2r+\gamma \leq 1$ is the worst case of Theorem 3.9. We employ data-dependent sampling to cover all agnostic cases $r \in (0, 1]$ in Corollary 3.15. The

error analysis differs due to the inclusion of Nyström error in our paper.

The theoretical gains of this work comes from two aspects:

- **Explicit intermediate estimators and tight error decomposition.** In the existing KRR work (Caponnetto & De Vito, 2007; Rudi et al., 2015; Guo et al., 2017b), various intermediate estimators were also introduced in Definition A.11, thus we illustrate the source of errors and the approximation relationships between several estimators (more details can be found in Proposition A.12). Then, in Lemma A.13, we tightly decompose the excess risk into: sample variance, Nyström error, empirical error and approximation error.
- **Tight estimate for sample variance $\|\widehat{f}_{M,\lambda} - \widetilde{f}_{M,\lambda}\|$.** The existing KRR relevant work (Caponnetto & De Vito, 2007; Rudi et al., 2018; 2017; Guo et al., 2017b) applied a relatively loose estimate for the sample variance

$$\|\widehat{f}_{M,\lambda} - \widetilde{f}_{M,\lambda}\| \lesssim \frac{1}{n\sqrt{\lambda}} + \sqrt{\frac{\mathcal{N}(\lambda)}{n}}. \tag{9}$$

  In this paper, using Bennett's inequality, we provide a novel estimate for sample variance for the first time

$$\|\widehat{f}_{M,\lambda} - \widetilde{f}_{M,\lambda}\| \lesssim \frac{\sqrt{\mathcal{N}_\infty(\lambda)}}{n} + \sqrt{\frac{\mathcal{N}(\lambda)}{n}}. \tag{10}$$

5. Conclusion

Based on the integral operator techniques, the minimax convergence rates for KRR and KRR variants have been proven in the realistic setting. The existing studies required a strict restriction on the target regression, assuming the concept lies exactly in the hypothesis space. However, according to the PAC theories, this assumption is relatively unreasonable since the joint distribution is unknown and the hypothesis space is usually biased. Therefore, this work explores the optimal statistical guarantees for Nyström-KRR in the agnostic setting, where the concept may be out of the hypothesis space. Overall, the techniques presented in this paper pave the way for studying other types of KRR relevant for the theoretical understanding of agnostic learning.

Acknowledgements

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.

References

Alaoui, A. and Mahoney, M. W. Fast randomized kernel ridge regression with statistical guarantees. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pp. 775–783, 2015.

Auer, P., Holte, R. C., and Maass, W. Theory and applications of agnostic pac-learning with small decision trees. In *Proceedings of the 12th International Conference on Machine Learning (ICML)*, pp. 21–29. Elsevier, 1995.

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

Blanchard, G. and Krämer, N. Optimal learning rates for kernel conjugate gradient regression. In *Advances in Neural Information Processing Systems 23 (NIPS)*, pp. 226–234, 2010.

Boutsidis, C., Mahoney, M. W., and Drineas, P. An improved approximation algorithm for the column subset selection problem. In *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*, pp. 968–977. SIAM, 2009.

Calandriello, D., Lazaric, A., and Valko, M. Distributed adaptive sampling for kernel matrix approximation. In *Artificial Intelligence and Statistics*, pp. 1421–1429. PMLR, 2017.

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

Carratino, L., Rudi, A., and Rosasco, L. Learning with sgd and random features. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pp. 10192–10203, 2018.

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

Cohen, M. B., Lee, Y. T., Musco, C., Musco, C., Peng, R., and Sidford, A. Uniform sampling for matrix approximation. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, pp. 181–190, 2015.

Della Vecchia, A., Mourtada, J., De Vito, E., and Rosasco, L. Regularized erm on random subspaces. In *International Conference on Artificial Intelligence and Statistics*, pp. 4006–4014. PMLR, 2021.

Drineas, P. and Mahoney, M. W. Approximating a gram matrix for improved kernel-based learning. In *International Conference on Computational Learning Theory*, pp. 323–337. Springer, 2005.

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

Fujii, J., Fujii, M., Furuta, T., and Nakamoto, R. Norm inequalities equivalent to heinz inequality. *Proceedings of the American Mathematical Society*, 118(3):827–830, 1993.

Gittens, A. and Mahoney, M. Revisiting the nystrom method for improved large-scale machine learning. In *International Conference on Machine Learning*, pp. 567–575. PMLR, 2013.

Guo, Z.-C., Lin, S.-B., and Shi, L. Distributed learning with multi-penalty regularization. *Applied and Computational Harmonic Analysis*, 2017a.

Guo, Z.-C., Lin, S.-B., and Zhou, D.-X. Learning theory of distributed spectral algorithms. *Inverse Problems*, 33(7): 074009, 2017b.

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

Li, J., Liu, Y., and Wang, W. Distributed learning with random features. *arXiv preprint arXiv:1906.03155*, 2019a.

Li, Z. Sharp analysis of random fourier features in classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7444–7452, 2022.

Li, Z., Ton, J.-F., Oglic, D., and Sejdinovic, D. Towards a unified analysis of random fourier features. In *International Conference on Machine Learning*, pp. 3905–3914. PMLR, 2019b.

Li, Z., Ton, J.-F., Oglic, D., and Sejdinovic, D. Towards a unified analysis of random fourier features. *Journal of Machine Learning Research*, 22(108), 2021.

Lin, J. and Cevher, V. Optimal distributed learning with multi-pass stochastic gradient methods. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 3098–3107, 2018.

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

Lin, J. and Rosasco, L. Optimal learning for multi-pass stochastic gradient methods. In *Advances in Neural Information Processing Systems 29 (NIPS)*, pp. 4556–4564, 2016.

Lin, J. and Rosasco, L. Optimal rates for multi-pass stochastic gradient methods. *The Journal of Machine Learning Research*, 18(1):3375–3421, 2017.

Lin, S.-B., Guo, X., and Zhou, D.-X. Distributed learning with regularized least squares. *The Journal of Machine Learning Research*, 18(1):3202–3232, 2017.

Lin, S.-B., Wang, D., and Zhou, D.-X. Distributed kernel ridge regression with communications. *Journal of Machine Learning Research*, 21(93):1–38, 2020.

Liu, Y., Liu, J., and Wang, S. Effective distributed learning with random features: Improved bounds and algorithms. In *International Conference on Learning Representations*, 2021.

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

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

Maass, W. Efficient agnostic pac-learning with simple hypothesis. In *Proceedings of the seventh annual conference on Computational learning theory*, pp. 67–75, 1994.

Müecke, N. Reducing training time by efficient localized kernel regression. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2603–2610. PMLR, 2019.

Musco, C. and Musco, C. Recursive sampling for the nystrom method. *Advances in neural information processing systems*, 30, 2017.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 21 (NIPS)*, pp. 1177–1184, 2007.

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

Rosasco, L. and Villa, S. Learning with incremental iterative regularization. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pp. 1630–1638, 2015.

Rudi, A. and Rosasco, L. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pp. 3215–3225, 2017.

Rudi, A., Camoriano, R., and Rosasco, L. Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pp. 1657–1665, 2015.

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

Rudi, A., Calandriello, D., Carratino, L., and Rosasco, L. On fast leverage score sampling and optimal learning. In *Advances in Neural Information Processing Systems*, pp. 5672–5682, 2018.

Samarakoon, L. and Leung, T.-Y. Conformer-based speech recognition with linear nyström attention and rotary position embedding. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8012–8016. IEEE, 2022.

Schölkopf, B., Herbrich, R., and Smola, A. J. A generalized representer theorem. In *International conference on computational learning theory*, pp. 416–426. Springer, 2001.

Shawe-Taylor, J. and Cristianini, N. *An introduction to support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000. ISBN 0521780195.

Shawe-Taylor, J. and Cristianini, N. *Kernel methods for pattern analysis*. Cambridge University Press, 2004.

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

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

Steinwart, I. and Christmann, A. *Support vector machines*. Springer Verlag, 2008.

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

Sun, Y., Gilbert, A., and Tewari, A. But how does it work in theory? linear svm with random features. In *Advances in Neural Information Processing Systems*, pp. 3379–3388, 2018.

Tropp, J. A. User-friendly tools for random matrices: An introduction. Technical report, 2012.

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

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Williams, C. K. and Seeger, M. Using the nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 14 (NIPS)*, pp. 682–688, 2001.

Xiong, Y., Zeng, Z., Chakraborty, R., Tan, M., Fung, G., Li, Y., and Singh, V. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 14138–14148, 2021.

Yashima, S., Nitanda, A., and Suzuki, T. Exponential convergence rates of classification errors on learning with sgd and random features. In *International Conference on Artificial Intelligence and Statistics*, pp. 1954–1962. PMLR, 2021.

Yin, R., Liu, Y., Lu, L., Wang, W., and Meng, D. Divide-and-conquer learning with nyström: Optimal rate and algorithm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 6696–6703, 2020.

Zhang, T. Effective dimension and generalization of kernel learning. *Advances in Neural Information Processing Systems*, 15, 2002.

Zhang, Y., Duchi, J., and Wainwright, M. Divide and conquer kernel ridge regression. In *Conference on Learning Theory*, pp. 592–617, 2013.

Zhang, Y., Duchi, J., and Wainwright, M. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research*, 16(1):3299–3340, 2015.

A. Proofs

We define operators in expectation, operators in probability with n training samples and M Nyström landmarks. We then estimate key operators similarities.

Definition A.1 (Operators in expectation)**.** 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 A.2 (Operators in probability)**.** For any $g \in L^2_{\rho_X}$, $\beta \in \mathcal{H}$, $\alpha \in \mathbb{R}^n$ and $\alpha' \in \mathbb{R}^M$, we have

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

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

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 $C_\lambda := C + \lambda I$, $\widehat{C}_{n\lambda} := \widehat{C}_n + \lambda I$, $L_\lambda := L + \lambda I$, and $\widehat{L}_{n\lambda} := \widehat{L}_n + \lambda I$.

The operators similarity quantity $\|(C + \lambda I)^{-1/2} (\widehat{C}_n + \lambda I)^{1/2}\|^2$ is the key to analyze the excess risk bound, and this quantity should be bounded as a constant. Traditional KRR related work (Guo et al., 2017b; Yin et al., 2020) estimated the key quantities after decomposition, and obtain $\|(C + \lambda I)^{-1/2} (\widehat{C}_n + \lambda I)^{1/2}\|^2 \le \|(C + \lambda I)^{-1/2}\| \|(C + \lambda I)^{-1/2} (C - \widehat{C}_n)\| + 1 = \mathcal{O}\big(\frac{1}{\lambda n} + \sqrt{\frac{\mathcal{N}(\lambda)}{\lambda n}}\big)$. To bound the quantity as a constant $\mathcal{O}\big(\frac{1}{\lambda n} + \sqrt{\frac{\mathcal{N}(\lambda)}{\lambda n}}\big) = \mathcal{O}(1)$, it leads to a restriction $n \ge \frac{\mathcal{N}(\lambda)}{\lambda}$. By Assumption 3.5 and setting $\lambda = n^{\frac{-1}{2r+1}}$, we obtain the constraint $n \ge n^{\frac{1+\gamma}{2r+\gamma}}$ and thus $r \ge 1/2$.

A.1. Operators Inequalities

We introduce useful operator inequalities and concentration inequalities to derive tight estimate for operator similarities.

Proposition A.3 (Cordes Inequality (Fujii et al., 1993))**.** *Let A, B two positive semi-definite bounded linear operators on a separable Hilbert space. Then*

$$\|A^s B^s\| \le \|AB\|^s, \qquad \textit{when} \quad 0 \le s \le 1.$$

Proposition A.4 (Bennett's inequality for random variables)**.** *Let $\mathcal{H}$ be a separable Hilbert space and $\{\xi_1, \cdots, \xi_n\}$ be a sequence of i.i.d random variables in $\mathcal{H}$. Assume the bound be $\|\xi - \mathbb{E}(\xi)\| \leq \widetilde{M} \leq \infty$ and the variance be $\tilde{\sigma}^2 = \mathbb{E}(\|\xi - \mathbb{E}(\xi)\|^2)$ for any $i \in [n]$. For any $\delta \in (0, 1)$, with confidence $1 - \delta$,*

$$\left\| \frac{1}{n} \sum_{i=1}^n \xi_i - \mathbb{E}(\xi_i) \right\| \leq \frac{2\widetilde{M} \log(2/\delta)}{3\,n} + \sqrt{\frac{2\tilde{\sigma}^2 \log(2/\delta)}{n}}. \tag{11}$$

The Bennett's inequality is the key to analysis the relationship between the empirical random vector and its expected counterpart, which is used to prove Lemma A.7 and Lemma A.16. The above Bennett's inequality for random vectors was provided in (Smale & Zhou, 2007; Rudi & Rosasco, 2017) and later was extended to the random operator cases in Theorem 7.3.1 in (Tropp, 2012) and Lemma 24 in (Lin & Cevher, 2020).

Proposition A.5 (Proposition 9 in (Rudi & Rosasco, 2017))**.** *Let $\mathcal{H}, \mathcal{K}$ be two separable Hilbert spaces and X, A be bounded linear operators, with $X : \mathcal{H} \to \mathcal{K}$ and $A : \mathcal{H} \to \mathcal{H}$ be positive semi-definite. The following holds*

$$\|XA^\varsigma\| = \|X\|^{1-\varsigma} \|XA\|^\varsigma, \qquad \forall \varsigma \in [0, 1].$$

Proposition A.6 (Lemma E.2 of (Blanchard & Krämer, 2010))**.** *For any self-adjoint and positive semi-definite operators A and B, if there exists $0 < \eta < 1$ such that the following inequality holds*

$$\|(A + \lambda I)^{-1/2}(B - A)(A + \lambda I)^{-1/2}\| \leq 1 - \eta,$$

then

$$\|(A + \lambda I)^{1/2}(B + \lambda I)^{-1/2}\| \leq \frac{1}{\sqrt{\eta}}.$$

The above inequality (Blanchard & Krämer, 2010) was used to establish the connection between $\|(A+\lambda I)^{-1/2}(B-A)(A+\lambda I)^{-1/2}\|$ and $\|(A + \lambda I)^{1/2}(B + \lambda I)^{-1/2}\|$. In this paper, those two terms $\|(A + \lambda I)^{-1/2}(B - A)(A + \lambda I)^{-1/2}\|$ and $\|(A + \lambda I)^{1/2}(B + \lambda I)^{-1/2}\|$ often exist on the left parts of the estimates of error terms, where we make use of Proposition A.6 to guarantee both of two terms of lhs as constants.

Lemma A.7. *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 $C = \mathbb{E}_{\rho_X}[K_x \otimes K_x]$ and $\widehat{C}_n = \frac{1}{n}\sum_{i=1}^n K_{\boldsymbol{x}_i} \otimes K_{\boldsymbol{x}_i}$ are trace class. Then for any $\delta \in (0, 1/2)$ with the probability at least $1 - 2\delta$, the following holds*

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

Proof. Let $C_\lambda^{-1/2} = (C + \lambda I)^{-1/2}$ and

$$\xi_i = C_\lambda^{-1/2} K_{x_i} \otimes C_\lambda^{-1/2} K_{x_i},$$

thus we have

$$\mathbb{E}(\xi_i) = C_\lambda^{-1/2} \mathbb{E}[K_{x_i} \otimes K_{x_i}] C_\lambda^{-1/2} = C_\lambda^{-1/2} C C_\lambda^{-1/2},$$
$$\frac{1}{n}\sum_{i=1}^n \xi_i = \frac{1}{n}\sum_{i=1}^n C_\lambda^{-1/2}[K_{\boldsymbol{x}_i} \otimes K_{\boldsymbol{x}_i}] C_\lambda^{-1/2} = C_\lambda^{-1/2} \widehat{C}_n C_\lambda^{-1/2}.$$

The left of the desired inequality becomes

$$\left\|C_\lambda^{-1/2}(C - \widehat{C}_n)C_\lambda^{-1/2}\right\| = \left\|\mathbb{E}(\xi_i) - \frac{1}{n}\sum_{i=1}^n \xi_i\right\|.$$

Note that

$$\|\xi_i\| = \|C_\lambda^{-1/2} K_{x_i} \otimes C_\lambda^{-1/2} K_{x_i}\| \leq \|C_\lambda^{-1/2} K_x\|^2 \leq \sup_{\boldsymbol{x} \in \mathcal{X}} \|C_\lambda^{-1/2} K_x\|^2 = \mathcal{N}_\infty(\lambda).$$

To make use of Bennett's inequality (Proposition A.4), we bound $\|\xi_i - \mathbb{E}(\xi_i)\|$ and $\mathbb{E}\|\xi_i - \mathbb{E}(\xi_i)\|^2$ as follows

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

Substituting the above two identities to Bennett's inequality (11), we prove the result. □

Lemma A.8. *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 the integral operator is defined as $(Lg)(\cdot) = \mathbb{E}_{\rho_X}[K(\boldsymbol{x}, \cdot)g(\boldsymbol{x})]$ and $(\widehat{L}_n g)(\cdot) = \frac{1}{n}\sum_{i=1}^n K(\boldsymbol{x}_i, \cdot)g(\boldsymbol{x}_i)$ are trace class. Then for any $\delta \in (0,1)$ with the probability at least $1-2\delta$, the following holds*

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

Proof. Let $L_\lambda^{-1/2} = (L+\lambda I)^{-1/2}$ and

$$\xi_i = L_\lambda^{-1/2}K(\boldsymbol{x}_i, \cdot)L_\lambda^{-1/2},$$

thus we have

$$\begin{aligned}
\mathbb{E}(\xi_i) &= L_\lambda^{-1/2}\mathbb{E}[K(\boldsymbol{x}_i, \cdot)]L_\lambda^{-1/2} = L_\lambda^{-1/2}LL_\lambda^{-1/2}, \\
\frac{1}{n}\sum_{i=1}^n \xi_i &= \frac{1}{n}\sum_{i=1}^n L_\lambda^{-1/2}[K(\boldsymbol{x}_i, \cdot)]L_\lambda^{-1/2} = L_\lambda^{-1/2}\widehat{L}_n L_\lambda^{-1/2}.
\end{aligned}$$

The left of the desired inequality becomes

$$\left\|L_\lambda^{-1/2}(L - \widehat{L}_n)L_\lambda^{-1/2}\right\| = \left\|\mathbb{E}(\xi_i) - \frac{1}{n}\sum_{i=1}^n \xi_i\right\|.$$

Note that

$$\begin{aligned}
\sup_{\boldsymbol{x}\in\mathcal{X}} \|C_\lambda^{-1/2}K_x\|^2 &= \mathcal{N}_\infty(\lambda). \\
\|C_\lambda^{-1/2}K_{(\cdot)}\|^2 \leq \sup_{\boldsymbol{x}\in\mathcal{X}} \|C_\lambda^{-1/2}K_x\|^2 &= \mathcal{N}_\infty(\lambda).
\end{aligned}$$

To use Bennett's inequality (Proposition A.4), we need to bound $\|\xi_i - \mathbb{E}(\xi_i)\|$ and $\mathbb{E}\|\xi_i - \mathbb{E}(\xi_i)\|^2$ as follows

$$\begin{aligned}
\|\xi_i - \mathbb{E}(\xi_i)\| &= \|C_\lambda^{-1/2}\langle K_{x_i}, K_{(\cdot)}\rangle C_\lambda^{-1/2} - L_\lambda^{-1/2}LL_\lambda^{-1/2}\| \\
&\leq \|C_\lambda^{-1/2}K_{x_i}\|\|C_\lambda^{-1/2}K_{(\cdot)}\| + \|L_\lambda^{-1}L_\lambda\| \leq \mathcal{N}_\infty(\lambda) + 1. \\
\mathbb{E}\|\xi_i - \mathbb{E}(\xi_i)\|^2 &= \left\|\mathbb{E}\left[\langle C_\lambda^{-1/2}K_{x_i}, C_\lambda^{-1/2}K_{(\cdot)}\rangle L_\lambda^{-1/2}K(\boldsymbol{x}_i, \cdot)L_\lambda^{-1/2}\right] - L_\lambda^{-2}L^2\right\| \\
&\leq \mathcal{N}_\infty(\lambda)\left\|\mathbb{E}\left[L_\lambda^{-1/2}K(\boldsymbol{x}_i, \cdot)L_\lambda^{-1/2}\right]\right\| + \|L_\lambda^{-2}L^2\| \\
&\leq \mathcal{N}_\infty(\lambda)\|L_\lambda^{-1}L\| + 1 \leq \mathcal{N}_\infty(\lambda) + 1.
\end{aligned}$$

Substituting the above two identities to Bennett's inequality (11), we prove the result. □

Lemma A.9. *When the number of the training samples $n \geq 16(\mathcal{N}_\infty(\lambda)+1)\log(2/\delta)$, then $\forall\ \delta \in (0,1)$, there exists with the confidence $1-\delta$*

$$\|C_\lambda^{-1/2}(C - \widehat{C}_n)C_\lambda^{-1/2}\| \leq \frac{1}{2} \quad \textit{and} \quad \|C_\lambda^{1/2}\widehat{C}_{n\lambda}^{-1/2}\| \leq \sqrt{2}, \quad \|C_\lambda^{-1/2}\widehat{C}_{n\lambda}^{1/2}\| \leq \sqrt{2}.$$

Proof. From Lemma A.7, we set $n \geq 16(\mathcal{N}_\infty(\lambda)+1)\log(2/\delta)$ and obtain that

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

From Proposition A.6 and the above inequality, there exists

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

□

Lemma A.10. *When the number of the training samples $n \geq 16(\mathcal{N}_\infty(\lambda)+1)\log(2/\delta)$, then $\forall\, \delta \in (0,1)$, there exists with the confidence $1-\delta$*

$$\|L_\lambda^{-1/2}(L - \widehat{L}_n)L_\lambda^{-1/2}\| \leq \frac{1}{2} \quad \textit{and} \quad \|L_\lambda^{1/2}\widehat{L}_{n\lambda}^{-1/2}\| \leq \sqrt{2}, \quad \|L_\lambda^{-1/2}\widehat{L}_{n\lambda}^{1/2}\| \leq \sqrt{2}.$$

Proof. From Lemma A.8, we set $n \geq 16(\mathcal{N}_\infty(\lambda)+1)\log(2/\delta)$ and obtain that

$$\|L_\lambda^{-1/2}(\widehat{L}_n - L)L_\lambda^{-1/2}\| \leq \frac{2(\mathcal{N}_\infty(\lambda)+1)\log(2/\delta)}{n} + \sqrt{\frac{2(\mathcal{N}_\infty(\lambda)+1)\log(2/\delta)}{n}} \leq \frac{1}{2}.$$

From Proposition A.6 and the above inequality, there exists

$$\|L_\lambda^{1/2}\widehat{L}_{n\lambda}^{-1/2}\| \leq \left(1-\frac{1}{2}\right)^{-\frac{1}{2}} = \sqrt{2}.$$
$$\|L_\lambda^{-1/2}\widehat{L}_{n\lambda}^{1/2}\| \leq \left(1-\frac{1}{2}\right)^{-\frac{1}{2}} = \sqrt{2}.$$

□

A.2. Tight Error Decomposition

In this section, using linear operators, we first prove the closed-form solutions of estimators. We then establish the relationship between intermediate estimators. Finally, we provide the tight error decomposition for KRR-Nyström.

A.2.1. Intermediate estimators

We introduce intermediate estimators to bridge the solution of KRR-Nyström $\widehat{f}_{M,\lambda}$ and the concept f_ρ. We measure the the generalization ability of $f \in L^2_{\rho_X}$ in terms of *excess risk* $\mathcal{E}(f) - \mathcal{E}(f_\rho)$ rather than $\mathcal{E}(f) - \mathcal{E}(f_\mathcal{H})$. It was proven that (Smale & Zhou, 2007)

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|_\rho^2, \qquad \forall\, f \in L^2_{\rho_X}. \tag{12}$$

Definition A.11 (Intermedia estimators)**.** Using the representer theorem, there are two reduced RKHS without and with Nyström approximation :

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

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

$$\widehat{f}_{M,\lambda}(\boldsymbol{x}) = \langle \boldsymbol{w}, K_{\boldsymbol{x}} \rangle, \qquad \text{with} \quad \boldsymbol{w} = \operatorname*{arg\,min}_{\boldsymbol{w} \in \mathcal{H}_M} \left\{ \frac{1}{n} \sum_{i=1}^n (\langle \boldsymbol{w}, K_{\boldsymbol{x}_i} \rangle - y_i)^2 + \lambda \|f\|_K^2 \right\}.$$

$$\widetilde{f}_{M,\lambda}(\boldsymbol{x}) = \langle \boldsymbol{u}, K_{\boldsymbol{x}} \rangle, \qquad \text{with} \quad \boldsymbol{u} = \operatorname*{arg\,min}_{\boldsymbol{u} \in \mathcal{H}_M} \left\{ \frac{1}{n} \sum_{i=1}^n (\langle \boldsymbol{u}, K_{\boldsymbol{x}_i} \rangle - f_\rho(\boldsymbol{x}_i))^2 + \lambda \|f\|_K^2 \right\}.$$

$$\widetilde{f}_{\lambda}(\boldsymbol{x}) = \langle \boldsymbol{v}, K_{\boldsymbol{x}} \rangle, \qquad \text{with} \quad \boldsymbol{v} = \operatorname*{arg\,min}_{\boldsymbol{v} \in \mathcal{H}_n} \left\{ \frac{1}{n} \sum_{i=1}^n (\langle \boldsymbol{v}, K_{\boldsymbol{x}_i} \rangle - f_\rho(\boldsymbol{x}_i))^2 + \lambda \|f\|_K^2 \right\}.$$

$$f_{\lambda}(\boldsymbol{x}) = \langle \boldsymbol{n}, K_{\boldsymbol{x}} \rangle, \qquad \text{with} \quad \boldsymbol{n} = \operatorname*{arg\,min}_{\boldsymbol{n} \in \mathcal{H}} \left\{ \int_X (\langle \boldsymbol{n}, K_{\boldsymbol{x}} \rangle - f_\rho(\boldsymbol{x}))^2 \, d\rho_X(\boldsymbol{x}) + \lambda \|f\|_K^2 \right\}.$$

We define the weights $\{\boldsymbol{w}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{n}\}$ in the RKHS, while the estimators $\{\widehat{f}_{M,\lambda}, \widetilde{f}_{M,\lambda}, \widetilde{f}_\lambda, f_\lambda\} \in L^2_{\rho_X}$. Let $Z_n = \sqrt{M}\widehat{S}_M = (\langle \beta, K(\boldsymbol{x}_i, \cdot) \rangle)_{i=1}^M$, such that $Z_n^* = \sqrt{M}\widehat{S}_M^* = \sum_{i=1}^M \alpha_i' K(\boldsymbol{x}_i, \cdot)$ is exactly $\mathcal{H}_M$. Let

$$Z_n = U\Sigma V^*$$

be the SVD of Z_n where $U : \mathbb{R}^t \to \mathbb{R}^M$, $\Sigma : \mathbb{R}^t \to \mathbb{R}^t$, $V : \mathbb{R}^t \to \mathcal{H}$, $t \leq M$ and $\Sigma = \operatorname{diag}(\sigma_1, \cdots, \sigma_t)$ in non-increasing order. It holds $U^*U = I_t$, $V^*V = I_t$ and $VV^* = P_n$ where P_n is the orthogonal projection operator and the range of P_n is exactly $\mathcal{H}_M$.

Proposition A.12. *Using operators in Definitions A.1, A.2, the estimators can be represented as*

$$\widehat{f}_{M,\lambda} = SV(V^*\widehat{C}_nV + \lambda I)^{-1}V^*\widehat{S}_n^*\widehat{y}_n, \tag{13}$$

$$\widetilde{f}_{M,\lambda} = SV(V^*\widehat{C}_nV + \lambda I)^{-1}V^*\bar{S}_n^*f_\rho, \tag{14}$$

$$\widetilde{f}_{\lambda} = S(\widehat{C}_n + \lambda I)^{-1}\bar{S}_n^*f_\rho, \tag{15}$$

$$f_\lambda = S(C + \lambda I)^{-1}S^*f_\rho. \tag{16}$$

Proof of Proposition A.12. The RKHS solution of $\widehat{f}_{M,\lambda}(\boldsymbol{x}) = \langle \boldsymbol{w}, K_x \rangle$ can be stated as

$$\boldsymbol{w} = \sum_{i=1}^M \alpha_i' K(\widetilde{\boldsymbol{x}}_i, \cdot) \qquad \text{with} \quad \boldsymbol{\alpha}' = (\mathbf{K}_{nM}^\top \mathbf{K}_{nM} + \lambda n \mathbf{K}_{MM})^\dagger \mathbf{K}_{nM}^\top \boldsymbol{y}_n,$$

According to the definitions of operators in Definition A.2, we have

$$\begin{aligned} \boldsymbol{\alpha}' &= (\mathbf{K}_{nM}^\top \mathbf{K}_{nM} + \lambda n \mathbf{K}_{MM}) \dagger \mathbf{K}_{nM}^\top \boldsymbol{y}_n \\ &= [M(\widehat{S}_M\widehat{S}_n^*)(\widehat{S}_n\widehat{S}_M^*) + \lambda M(\widehat{S}_M\widehat{S}_M^*)]^\dagger(\sqrt{M}\widehat{S}_M\widehat{S}_n^*)\widehat{y}_n \end{aligned}$$

Then, there exists

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

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

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

where $\sqrt{M}\widehat{S}_M = U\Sigma V^*$, $F = U\Sigma$, $G = V^*\widehat{C}_nV + \lambda I$, $H = \Sigma U^\top$ and F, GH, G and H are full-rank matrices. Simplifying U and Σ, we prove (13) with

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

The difference between $\widehat{f}_{M,\lambda}$ and $\widetilde{f}_{M,\lambda}$ is the labels with noises y_i or the labels without noises $f_\rho(\boldsymbol{x}_i)$. Thus, there exists

$$\begin{aligned}
\widehat{f}_{M,\lambda} &= SV(V^*\widehat{C}_nV+\lambda I)^{-1}V^*\left(\frac{1}{n}\sum_{i=1}^n K_{x_i}y_i\right)\\
\widetilde{f}_{M,\lambda} &= SV(V^*\widehat{C}_nV+\lambda I)^{-1}V^*\left(\frac{1}{n}\sum_{i=1}^n K_{x_i}f_\rho(\boldsymbol{x}_i)\right) = SV(V^*\widehat{C}_nV+\lambda I)^{-1}V^*\bar{S}_n^*f_\rho.
\end{aligned}$$

The estimator $\widetilde{f}_\lambda$ is the solution of KRR with noise-free labels, and it holds

$$\widetilde{f}_\lambda = S\sqrt{n}\widehat{S}_n^*\boldsymbol{\alpha} \qquad \text{with} \qquad \boldsymbol{\alpha} = (\mathbf{K}_{nn}+\lambda nI)^{-1}[f_\rho(\boldsymbol{x}_1),\cdots,f_\rho(\boldsymbol{x}_n)]^\top$$

Then, we have

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

It is well know the estimator f_λ is equal to

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

where the third step is due to $Z^*f(ZZ^*) = f(Z^*Z)Z^*$ for any continuous spectral function and any compact operator Z. □

According the definitions of estimators with operators (Proposition A.12), it is natural to estimate errors w.r.t the difference among the estimators. There is an approximation chain from the KRR-Nyström $\widehat{f}_{M,\lambda}$ to the expected estimator f_λ in terms of expectation and the number of Nyström centers: $\widehat{f}_{M,\lambda} \overset{\rho(y|\boldsymbol{x})}{\longrightarrow} \widetilde{f}_{M,\lambda} \overset{M\to n}{\longrightarrow} \widetilde{f}_\lambda \overset{\rho_X}{\longrightarrow} f_\lambda$.

Integrating the *excess risk* (12) together the with the intermediate estimators defined in Definition A.11, we then decompose the *excess risk* into four parts in terms of the $L^2_{\rho_X}$ norms: $\|\widehat{f}_{M,\lambda}-\widetilde{f}_{M,\lambda}\|_\rho^2$ is the sample variance introduced by noised labels; $\|\widetilde{f}_{M,\lambda}-\widetilde{f}_\lambda\|_\rho^2$ is the error brought by Nyström approximation; $\|\widetilde{f}_\lambda-f_\lambda\|_\rho^2$ is the computational error from empirical samples; and $\|f_\lambda-f_\rho\|_\rho^2$ is the approximation error.

Lemma A.13. *Let $\widehat{f}_{M,\lambda}, \widetilde{f}_{M,\lambda}, \widetilde{f}_\lambda$ and f_λ be defined in Definition A.11. The following error decomposition holds for KRR-Nyström*

$$\mathcal{E}(\widehat{f}_{M,\lambda})-\mathcal{E}(f_\rho) \leq \underbrace{4\|\widehat{f}_{M,\lambda}-\widetilde{f}_{M,\lambda}\|_\rho^2}_{\textit{Sample Variance}} + \underbrace{4\|\widetilde{f}_{M,\lambda}-\widetilde{f}_\lambda\|_\rho^2}_{\textit{Nyström Error}} + \underbrace{4\|\widetilde{f}_\lambda-f_\lambda\|_\rho^2}_{\textit{Empirical Error}} + \underbrace{4\|f_\lambda-\ f_\rho\|_\rho^2}_{\textit{Approximation Error}}. \tag{17}$$

Proof. The excess risk is related to the difference between estimators in (12), and thus we have

$$\mathcal{E}(\widehat{f}_{M,\lambda})-\mathcal{E}(f_\rho) = \|\widehat{f}_{M,\lambda}-f_\rho\|_\rho^2, \qquad \forall f\in L^2_{\rho_X}. \tag{18}$$

Introducing the intermediate estimators $\widetilde{f}_{M,\lambda}, \widetilde{f}_\lambda, f_\lambda$, we have

$$\|\widehat{f}_{M,\lambda}-f_\rho\|_\rho^2 = \|\widehat{f}_{M,\lambda}-\widetilde{f}_{M,\lambda}+\widetilde{f}_{M,\lambda}-\widetilde{f}_\lambda+\widetilde{f}_\lambda-f_\lambda+f_\lambda-f_\rho\|_\rho^2. \tag{19}$$

By the fact $(a+b+c+d)^2 \leq 4a^2+4b^2+4c^2, \forall a,b,c,d>0$, we have

$$\|\widetilde{f}_{M,\lambda}-f_\rho\|_\rho^2 \leq 4\|\widehat{f}_{M,\lambda}-\widetilde{f}_{M,\lambda}\|_\rho^2+4\|\widetilde{f}_{M,\lambda}-\widetilde{f}_\lambda\|_\rho^2+4\|\widetilde{f}_\lambda-f_\lambda\|_\rho^2+4\|f_\lambda-f_\rho\|_\rho^2. \tag{20}$$

Substituting (19) and (20) into (18), one can obtain the desired result. □

Here, the first three error terms can be regarded as variance and the approximation error as bias.

A.3. Estimates for Error Terms

In this part, we provide the rough estimates for those four error terms in Lemma A.13 : the sample variance $\|\widehat{f}_{M,\lambda} - \widetilde{f}_{M,\lambda}\|$, the Nyström error $\|\widetilde{f}_{M,\lambda} - \widetilde{f}_{\lambda}\|$, the empirical error $\|\widetilde{f}_{\lambda} - f_{\lambda}\|$ and the approximation error $\|f_{\lambda} - f_{\rho}\|$. Most of the integral-operator theory related work involves the estimation of the quantity $\|(\widehat{C}_n + \lambda I)^{-1/2}(C + \lambda I)^{1/2}\|$. Besides, this work also involves the quantity $\|(\widehat{L}_n + \lambda I)^{-1/2}(L + \lambda I)^{1/2}\|$. Using Bennett's inequality, we upper bound those two quantities by a constant $\sqrt{2}$ under the condition

$$n \geq 16(\mathcal{N}_{\infty}(\lambda) + 1)\log(2/\delta).$$

It holds the restriction $2r + \gamma \geq 1$ due to the fact $\mathcal{N}_{\infty}(\lambda) \lesssim \lambda^{-1}$ for the uniform sampling, while the restriction is relaxed to $r \in (0, 1]$ with the data-dependent sampling. To achieve the optimal convergence rates in sample variance, we also obtain the restriction $2r + 2\gamma \geq 1$ for the uniform sampling and $r \in (0, 1]$ for the data-dependent sampling. This work relaxed the regularity condition for the capacity-dependent optimality of Nyström approximation from $r \in [1/2, 1]$ to $2r + \gamma \geq 1$ with the uniform sampling and $r \in (0, 1]$ with the data-dependent sampling.

A.3.1. Estimate for sample variance

Lemma A.14. *Let $\delta \in (0, 1]$, $\widehat{f}_{M,\lambda}$ and $\widetilde{f}_{M,\lambda}$ be defined by (13) and (14). Then, the sample variance holds with the probability at least $1 - \delta$*

$$\|\widehat{f}_{M,\lambda} - \widetilde{f}_{M,\lambda}\| \leq 4\|C_{\lambda}^{-1/2}\widehat{C}_{n\lambda}^{1/2}\|^2 \left(\frac{B\sqrt{\mathcal{N}_{\infty}(\lambda)}}{n} + \sqrt{\frac{B^2\mathcal{N}(\lambda)}{n}}\right)\log\frac{2}{\delta}.$$

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

$$\begin{aligned}\widehat{f}_{M,\lambda} &= SV(V^*\widehat{C}_nV + \lambda I)^{-1}V^*\widehat{S}_n^*\widehat{y}_n,\\ \widetilde{f}_{M,\lambda} &= SV(V^*\widehat{C}_nV + \lambda I)^{-1}V^*\bar{S}_n^*f_{\rho}.\end{aligned}$$

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

$$\begin{aligned}\|\widehat{f}_{M,\lambda} - \widetilde{f}_{M,\lambda}\| =&\|SG_n(\widehat{S}_n^*\widehat{y}_n - \bar{S}_n^*f_{\rho})\|\\ \leq&\|\big(SG_n\widehat{C}_{n\lambda}^{1/2}\big)\big(\widehat{C}_{n\lambda}^{-1/2}(\widehat{S}_n^*\widehat{y}_n - \bar{S}_n^*f_{\rho})\big)\|\\ \leq&\underbrace{\|SG_n\widehat{C}_{n\lambda}^{1/2}\|}_{\mathcal{A}}\|\widehat{C}_{n\lambda}^{-1/2}C_{\lambda}^{1/2}\|\underbrace{\|C_{\lambda}^{-1/2}(\widehat{S}_n^*\widehat{y}_n - \bar{S}_n^*f_{\rho})\|}_{\mathcal{C}}.\end{aligned} \tag{21}$$

where the last step is due to Cauchy–Schwarz inequality. Note that

$$\mathcal{A} = \|SC_{\lambda}^{-1/2}C_{\lambda}^{1/2}\widehat{C}_{n\lambda}^{-1/2}\widehat{C}_{n\lambda}^{1/2}G_n\widehat{C}_{n\lambda}^{1/2}\| \leq \|SC_{\lambda}^{-1/2}\|\|C_{\lambda}^{1/2}\widehat{C}_{n\lambda}^{-1/2}\|\|\widehat{C}_{n\lambda}^{1/2}G_n\widehat{C}_{n\lambda}^{1/2}\|,$$

where $\|SC_{\lambda}^{-1/2}\| \leq \|C_{\lambda}^{-1/2}S^*SC_{\lambda}^{-1/2}\|^{1/2} \leq 1$. Thus, it holds that

$$\mathcal{A} \leq \|C_{\lambda}^{1/2}\widehat{C}_{n\lambda}^{-1/2}\|\|\widehat{C}_{n\lambda}^{1/2}G_n\widehat{C}_{n\lambda}^{1/2}\|. \tag{22}$$

The part $\mathcal{C}$ can be bounded as follows

$$\begin{aligned}\mathcal{C} =&\|C_{\lambda}^{1/2}(\widehat{S}_n\widehat{y}_n - S^*f_{\rho} + S^*f_{\rho} - \bar{S}_n^*f_{\rho})\|\\ \leq&\|C_{\lambda}^{1/2}(\widehat{S}_n\widehat{y}_n - S^*f_{\rho})\| + \|C_{\lambda}^{1/2}(S^*f_{\rho} - \bar{S}_n^*f_{\rho})\|.\end{aligned} \tag{23}$$

Substituting (22), (23) into (21), we have

$$\|\widehat{f}_{M,\lambda} - \widetilde{f}_{M,\lambda}\| \leq \|C_{\lambda}^{-1/2}\widehat{C}_{n\lambda}^{1/2}\|^2\|\widehat{C}_{n\lambda}^{1/2}G_n\widehat{C}_{n\lambda}^{1/2}\|\Big[\|C_{\lambda}^{1/2}(\widehat{S}_n^*\widehat{y}_n - S^*f_{\rho})\| + \|C_{\lambda}^{1/2}(S^*f_{\rho} - \bar{S}_n^*f_{\rho})\|\Big].$$

Using Lemma A.16 and Lemma A.17, the sample variance holds with the probability at least $1-\delta$

$$\|\widehat{f}_{M,\lambda} - \widetilde{f}_{M,\lambda}\| \leq 4\|C_\lambda^{-1/2}\widehat{C}_{n\lambda}^{1/2}\|^2 \left(\frac{B\sqrt{\mathcal{N}_\infty(\lambda)}}{n} + \sqrt{\frac{B^2\mathcal{N}(\lambda)}{n}}\right) \log\frac{2}{\delta}.$$

□

The sample variance is resulted from the label noises, thus $\widetilde{f}_{M,\lambda}$ use $f_\rho(\boldsymbol{x})$ instead of y. To upper bound the rhs in the above inequality via Bennett's inequality, we introduce an expected term $S^* f_\rho$ to bridge two empirical terms.

Lemma A.15 (Lemma 8 of (Rudi et al., 2015))**.** *For any $\lambda > 0$, let V be such that $V^*V = I$ and $\widehat{C}_n$ be a positive self-adjoint operator. Then, the following holds*

$$\|\widehat{C}_{n\lambda}^{1/2} G_n \widehat{C}_{n\lambda}^{1/2}\| \leq 1.$$

Proof. Let $\widehat{C}_{n\lambda} = \widehat{C}_n + \lambda I$ and $G_n = V(V^*\widehat{C}_n V + \lambda)^{-1}V^*$, then

$$\begin{aligned}
\|\widehat{C}_{n\lambda}^{1/2} G_n \widehat{C}_{n\lambda}^{1/2}\|^2 =& \|\widehat{C}_{n\lambda}^{1/2} G_n \widehat{C}_{n\lambda} G_n \widehat{C}_{n\lambda}^{1/2}\|^2 \\
=& \|\widehat{C}_{n\lambda}^{1/2} V(V^*\widehat{C}_{n\lambda}V)^{-1}(V^*\widehat{C}_{n\lambda}V)(V^*\widehat{C}_{n\lambda}V)^{-1}V^*\widehat{C}_{n\lambda}^{1/2}\| \\
=& \|\widehat{C}_{n\lambda}^{1/2} G_n \widehat{C}_{n\lambda}^{1/2}\|,
\end{aligned}$$

and thus $\|\widehat{C}_{n\lambda}^{1/2} G_n \widehat{C}_{n\lambda}^{1/2}\|$ is 0 or 1. □

Using Bennett's inequality (Proposition A.4) and following the proof of Lemma 6 in (Rudi & Rosasco, 2017), we prove the following lemmas.

Lemma A.16. *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 $1-\delta$*

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

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

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \xi_i = \frac{1}{n}\sum_{i=1}^n C_\lambda^{-1/2} K_{\boldsymbol{x}_i} y_i = C_\lambda^{-1/2}\widehat{S}_n^* \widehat{y}_n, \\
\mathbb{E}\,\xi = \int_X C_\lambda^{-1/2} K_x f_\rho(\boldsymbol{x}) d\rho_X(\boldsymbol{x}) = C_\lambda^{-1/2} S^* f_\rho
\end{aligned}$$

Thus, the error term to bound can be stated as

$$\|C_\lambda^{-1/2}(\widehat{S}_n^* \widehat{y}_n - S^* f_\rho)\| = \left\|\frac{1}{n}\sum_{i=1}^n \xi_i - \mathbb{E}\xi_i\right\|. \tag{24}$$

The rhs of the above identity can be bounded by Bennett's inequality (Proposition A.4), thus we need to estimate $\|\xi_i - \mathbb{E}(\xi_i)\|$ and $\mathbb{E}\,\|\xi_i - \mathbb{E}(\xi_i)\|^2$ first.

We first recall the definitions of $\mathcal{N}(\lambda)$ and $\mathcal{N}_\infty(\lambda)$.

$$\begin{aligned}
\mathcal{N}(\lambda) &= \mathbb{E}\,\langle K_x, (C+\lambda I)^{-1}K_x\rangle_K = \int_X \|(C+\lambda I)^{-1}K_x\|_K^2\, d\rho_X(\boldsymbol{x}), \\
\mathcal{N}_\infty(\lambda) &= \sup_{\boldsymbol{x}\in\mathcal{X}}\ \langle K_x, (C+\lambda I)^{-1}K_x\rangle_K = \sup_{\boldsymbol{x}\in\mathcal{X}}\ \|(C+\lambda I)^{-1}K_x\|_K^2.
\end{aligned}$$

By Jensen's inequality, we thus have

$$\|\xi_i - \mathbb{E}(\xi_i)\| \leq \|C_\lambda^{-1/2} K_{x_i}\||y_i| + \mathbb{E}\|C_\lambda^{-1/2} K_{x_i}\||y_i| \leq 2B\sqrt{\mathcal{N}_\infty(\lambda)}. \tag{25}$$

Note that

$$\begin{aligned}\mathbb{E}\|\xi_i - \mathbb{E}(\xi_i)\|^2 &\leq 2\int_X \|C_\lambda^{-1/2} K_{x_i}\|^2 |y_i|^2 d\rho_X(\boldsymbol{x}) \\ &\leq 2B^2 \int_X \|C_\lambda^{-1/2} K_{x_i}\|^2 d\rho_X(\boldsymbol{x}) \leq 2B^2 \mathcal{N}(\lambda).\end{aligned} \tag{26}$$

Substituting (25) and (26) to (24), by Bennett's inequality (Proposition A.4), we have

$$\begin{aligned}\|C_\lambda^{-1/2}(S_M^* f_\rho - \bar{S}_M^* f_\rho)\| &\leq \frac{2B\sqrt{\mathcal{N}_\infty(\lambda)}\log(2/\delta)}{n\sqrt{\lambda}} + 2\sqrt{\frac{B^2\mathcal{N}(\lambda)\log(2/\delta)}{n}} \\ &\leq 2\left(\frac{B\sqrt{\mathcal{N}_\infty(\lambda)}}{n} + \sqrt{\frac{B^2\mathcal{N}(\lambda)}{n}}\right)\log\frac{2}{\delta}.\end{aligned}$$

□

Lemma A.17. *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 $\delta \in (0,1]$, with the probability at least $1-\delta$, we have*

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

Proof. 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

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \xi_i &= \frac{1}{n}\sum_{i=1}^n C_\lambda^{-1/2} K_{\boldsymbol{x}_i} f_\rho(\boldsymbol{x}_i) = C_\lambda^{-1/2}\bar{S}_n^* f_\rho, \\ \mathbb{E}\xi_i &= \int_X C_\lambda^{-1/2} K_x f_\rho(\boldsymbol{x}) d\rho_X(\boldsymbol{x}) = C_\lambda^{-1/2} S^* f_\rho\end{aligned}$$

Thus, the error term to bound can be stated as

$$\|C_\lambda^{-1/2}(S^* f_\rho - \bar{S}_n^* f_\rho)\| = \left\|\frac{1}{n}\sum_{i=1}^n \xi_i - \mathbb{E}\xi_i\right\|. \tag{27}$$

To apply Bennett's inequality (Proposition A.4), we also need to estimate $\|\xi - \mathbb{E}(\xi)\|$ and $\mathbb{E}(\|\xi - \mathbb{E}(\xi)\|^2)$. Note that $|y| \leq B$ almost surely for some constant $B > 0$ and $\mathcal{X}$ is compact, that indicates $|f_\rho(x)| \leq B$ almost surely.

By Jensen's inequality, we thus have

$$\|\xi_i - \mathbb{E}(\xi_i)\| \leq \|C_\lambda^{-1/2} K_{x_i}\||f_\rho(\boldsymbol{x})| + \mathbb{E}\|C_\lambda^{-1/2} K_{x_i}\||f_\rho(\boldsymbol{x})| \leq 2B\sqrt{\mathcal{N}_\infty(\lambda)}. \tag{28}$$

Note that

$$\begin{aligned}\mathbb{E}\,\|\xi_i - \mathbb{E}(\xi_i)\|^2 &\leq 2\int_X \|C_\lambda^{-1/2} K_{x_i}\|^2 |f_\rho(\boldsymbol{x})|^2 d\rho_X(\boldsymbol{x}) \leq 2\int_X \|C_\lambda^{-1/2} K_{x_i}\|^2 d\rho_X(\boldsymbol{x}) \\ &\leq 2B^2 \int_X \|C_\lambda^{-1/2} K_{x_i}\|^2 d\rho_X(\boldsymbol{x}) \leq 2B^2 \mathcal{N}(\lambda).\end{aligned} \tag{29}$$

Substituting (28) and (29) to (27), by Bennett's inequality (Proposition A.4), we have

$$\begin{aligned}\|C_\lambda^{-1/2}(S_M^* f_\rho - \bar{S}_M^* f_\rho)\| &\leq \frac{2B\sqrt{\mathcal{N}_\infty(\lambda)}\log(2/\delta)}{n\sqrt{\lambda}} + 2\sqrt{\frac{B^2\mathcal{N}(\lambda)\log(2/\delta)}{n}} \\ &\leq 2\left(\frac{B\sqrt{\mathcal{N}_\infty(\lambda)}}{n} + \sqrt{\frac{B^2\mathcal{N}(\lambda)}{n}}\right)\log\frac{2}{\delta}.\end{aligned}$$

□

A.3.2. Estimate for Nyström error

The following two lemmas are used to estimate the key term $\|(I - VV^*)C_\lambda^{1/2}\|$ in Nyström error (Lemma A.20) in terms of different subsampling strategies. The first one measures the error term for Nyström method with uniform subsampling, given in Lemma 6 of (Rudi et al., 2015). The second one measures the error term for Nyström method with approximate leverage scores (ALS) subsampling, which was provided in Lemma 7 of (Rudi et al., 2015).

Lemma A.18 (Uniform sampling, Lemma 6 of (Rudi et al., 2015))**.** *Let $\lambda > 0$ for any $\delta > 0$ and the Nyström centers are sampled uniformly from the training examples, such that $M \geq 67 \log \frac{4\kappa^2}{\lambda\delta} \vee 5\mathcal{N}_\infty(\lambda) \log \frac{4\kappa^2}{\lambda\delta}$, the following holds with probability $1 - \delta$*

$$\|(I - VV^*)C_\lambda^{1/2}\|^2 \leq 3\lambda.$$

Lemma A.19 (Data-dependent sampling, Lemma 7 of (Rudi et al., 2015))**.** *Under Assumption 3.5, let $\lambda > 0$ for any $\delta > 0$ and the Nyström centers are sampled according to the leverage scores $p_i = \frac{\widehat{l}_\lambda(i)}{\sum_{i=1}^n \widehat{l}_\lambda(i)}$ in (4), then for any $\delta > 0$ the following holds with probability $1 - 2\delta$*

$$\|(I - VV^*)C_\lambda^{1/2}\|^2 \leq 3\lambda,$$

when the following conditions are satisfied:

- *there exists a $p \geq 1$ and a λ_0 such that $(\widehat{l}_\lambda(i))_{i=1}^n$ are p-approximate leverage scores are used to select random Nyström centers.*
- $n \geq 1665\kappa^2 + 223\kappa^2 \log \frac{2\kappa^2}{\delta}$,
- $\lambda_0 \vee \frac{19\kappa^2}{n} \log \frac{2N}{\delta} \leq \lambda \leq \|C\|$,
- $M \geq 334 \log \frac{8N}{\delta} \vee 78p^2\mathcal{N}(\lambda) \log \frac{8N}{\delta}$.

Lemma A.20. *Let $\delta \in (0, 1]$, $\widetilde{f}_{M,\lambda}$ and $\widetilde{f}_\lambda$ be defined by (14) and (15). When the number of Nyström centers satisfies $M \geq 67 \log \frac{4\kappa^2}{\lambda\delta} \vee 5\mathcal{N}_\infty(\lambda) \log \frac{4\kappa^2}{\lambda\delta}$ for uniform sampling and $M \geq 334 \log \frac{8N}{\delta} \vee 78p^2\mathcal{N}(\lambda) \log \frac{8N}{\delta}$ for data-dependent sampling, the Nyström error holds*

$$\|\widetilde{f}_{M,\lambda} - \widetilde{f}_\lambda\| \leq 6(\|C_\lambda^{-1/2}\widehat{C}_{n\lambda}^{1/2}\|^2 + 1)R\lambda^r, \qquad \textit{when } r \in (0, 1/2).$$
$$\|\widetilde{f}_{M,\lambda} - \widetilde{f}_\lambda\| \leq 3\|C_\lambda^{-1/2}\widehat{C}_{n\lambda}^{1/2}\|\|\widehat{L}_{n\lambda}^{-1/2}L_\lambda^{1/2}\|^2 R\lambda^r, \qquad \textit{when } r \in [1/2, 1].$$

Proof. Recall the definitions of $\widetilde{f}_{M,\lambda}$ and $\widetilde{f}_\lambda$ with operators, it holds

$$\begin{aligned}\widetilde{f}_{M,\lambda} &= SV(V^*\widehat{C}_nV + \lambda I)^{-1}V^*\bar{S}_n^*f_\rho,\\ \widetilde{f}_\lambda &= S(\widehat{C}_n + \lambda I)^{-1}\bar{S}_n^*f_\rho.\end{aligned}$$

We use $G_n = V(V^*\widehat{C}_nV + \lambda I)^{-1}V^*$ and then $\widetilde{f}_{M,\lambda} = SG_n\bar{S}_n^*f_\rho$.

Using $Z^*f(ZZ^*) = f(Z^*Z)Z^*$, we have

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

We estimate the Nyström error as follows with

$$\begin{aligned}\|\widetilde{f}_{M,\lambda} - \widetilde{f}_\lambda\| &= \|S(G_n - \widehat{C}_{n\lambda}^{-1})\bar{S}_n^*f_\rho\|\\ &= \|S(G_n\widehat{C}_{n\lambda} - I)\widehat{C}_{n\lambda}^{-1}\bar{S}_n^*f_\rho\|\\ &= \|S(G_n\widehat{C}_{n\lambda} - I)\bar{S}_n^*\widehat{L}_{n\lambda}^{-1}f_\rho\|\\ &= \|S(G_n\widehat{C}_{n\lambda} - I)\bar{S}_n^*\widehat{L}_{n\lambda}^{-1}L^rg\|.\end{aligned}$$

Then, we bound $\|\widetilde{f}_{M,\lambda} - \widetilde{f}_\lambda\|$ 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 $\|g\| \leq R$, $\|L_\lambda^{-1} L\| \leq 1$, $\|\widehat{L}_{n\lambda}^{-1/2} \lambda^{1/2}\| \leq 1$, $\|S C_\lambda^{-1/2}\| = \|C_\lambda^{-1/2} C_\lambda C_\lambda^{-1/2}\|^{1/2} \leq 1$, $\|\bar{S}_n^* \widehat{L}_{n\lambda}^{-1/2}\| \leq \|\widehat{L}_{n\lambda}^{-1/2} \widehat{L}_n \widehat{L}_{n\lambda}^{-1/2}\|^{1/2} \leq 1$ and $\|\widehat{C}_{n\lambda}^{-1/2} \bar{S}_n^*\| = \|\widehat{C}_{n\lambda}^{-1/2} \widehat{C}_n \widehat{C}_{n\lambda}^{-1/2}\|^{1/2} \leq 1$, and we then have

$$\begin{aligned}
&\|\widetilde{f}_{M,\lambda} - \widetilde{f}_\lambda\| \\
=&\|S(G_n \widehat{C}_{n\lambda} - I)\bar{S}_n^* \widehat{L}_{n\lambda}^{r-1} (\widehat{L}_{n\lambda}^{-1/2} L_\lambda^{1/2})^{2r} (L_\lambda^{-1} L)^r g\| \\
=&\|(S C_\lambda^{-1/2}) C_\lambda^{1/2} (G_n \widehat{C}_{n\lambda} - I) \bar{S}_n^* \widehat{L}_{n\lambda}^{r-1} (\widehat{L}_{n\lambda}^{-1/2} L_\lambda^{1/2})^{2r} (L_\lambda^{-1} L)^r g\| \\
\leq& R\|C_\lambda^{1/2} (G_n \widehat{C}_{n\lambda} - I) C_\lambda^r (C_\lambda^{-1/2} \widehat{C}_{n\lambda}^{1/2})^{2r} (\widehat{C}_{n\lambda}^{-1/2} \bar{S}_n^*)^{2r} \\
&\quad (\bar{S}_n^* \widehat{L}_{n\lambda}^{-1/2})^{1-2r} (\widehat{L}_{n\lambda}^{-1/2} \lambda^{1/2}) \lambda^{-1/2} (\widehat{L}_{n\lambda}^{-1/2} L_\lambda^{1/2})^{2r}\| \\
\leq& R\lambda^{-1/2} \|C_\lambda^{1/2} (G_n \widehat{C}_{n\lambda} - I) C_\lambda^r\| \|C_\lambda^{-1/2} \widehat{C}_{n\lambda}^{1/2}\|^{2r} \|\widehat{L}_{n\lambda}^{-1/2} L_\lambda^{1/2}\|^{2r}.
\end{aligned}$$

  Using Lemma A.9 and Lemma A.10, with the constraint $n \geq 16(\mathcal{N}_\infty(\lambda) + 1) \log(2/\delta)$, we have

$$\begin{aligned}
\|\widetilde{f}_{M,\lambda} - \widetilde{f}_\lambda\| &\leq R\lambda^{-1/2} \|C_\lambda^{1/2} (G_n \widehat{C}_{n\lambda} - I) C_\lambda^r\| 2^{2r} \\
&\leq 2R\lambda^{-1/2} \|C_\lambda^{1/2} (G_n \widehat{C}_{n\lambda} - I) C_\lambda^r\|.
\end{aligned} \tag{30}$$

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

$$\begin{aligned}
G_n \widehat{C}_{n\lambda} - I &= G_n \widehat{C}_{n\lambda}(I - VV^*) + G_n \widehat{C}_{n\lambda} VV^* - I \\
&= G_n \widehat{C}_{n\lambda}(I - VV^*) - (I - VV^*).
\end{aligned}$$

  Multiplying and dividing by $\widehat{C}_{n\lambda}^{1/2}$ and $C_\lambda^{1/2}$ and using above identity, we have

$$\begin{aligned}
&\|C_\lambda^{1/2} (G_n \widehat{C}_{n\lambda} - I) C_\lambda^r\| \\
\leq& \|C_\lambda^{1/2} \widehat{C}_{n\lambda}^{-1/2} \widehat{C}_{n\lambda}^{1/2} G_n \widehat{C}_{n\lambda}^{1/2} \widehat{C}_{n\lambda}^{1/2} C_\lambda^{-1/2} C_\lambda^{1/2} (I - VV^*) C_\lambda^r\| + \|C_\lambda^{1/2} (I - VV^*) C_\lambda^r\| \\
\leq& \|C_\lambda^{1/2} (I - VV^*) C_\lambda^r\| (\|C_\lambda^{1/2} \widehat{C}_{n\lambda}^{-1/2}\| \|\widehat{C}_{n\lambda}^{1/2} G_n \widehat{C}_{n\lambda}^{1/2}\| \|\widehat{C}_{n\lambda}^{1/2} C_\lambda^{-1/2}\| + 1) \\
\leq& \|C_\lambda^{1/2} (I - VV^*) C_\lambda^r\| (\|C_\lambda^{1/2} \widehat{C}_{n\lambda}^{-1/2}\| \|\widehat{C}_{n\lambda}^{1/2} C_\lambda^{-1/2}\| + 1) \\
\leq& (\|C_\lambda^{-1/2} \widehat{C}_{n\lambda}^{1/2}\|^2 + 1) \|C_\lambda^{1/2} (I - VV^*) C_\lambda^r\|.
\end{aligned} \tag{31}$$

  The third step is due to $\|\widehat{C}_{n\lambda}^{1/2} G_n \widehat{C}_{n\lambda}^{1/2}\| \leq 1$ in Lemma A.15.

  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 (Proposition A.3) to $\|(I - VV^*) C_\lambda^r\|$, we have

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

  Thus, it holds

$$\|C_\lambda^{1/2} (I - VV^*) C_\lambda^r\| \leq \|(I - VV^*) C_\lambda^{1/2}\|^{2r+1}. \tag{32}$$

  Substituting (31) and (32) into (30), with the condition $n \geq 16(\mathcal{N}_\infty(\lambda) + 1) \log(2/\delta)$, there exists for $r \in (0, 1/2)$ that

$$\|\widetilde{f}_{M,\lambda} - \widetilde{f}_\lambda\| \leq 2(\|C_\lambda^{-1/2} \widehat{C}_{n\lambda}^{1/2}\|^2 + 1) R\lambda^{-1/2} \|(I - VV^*) C_\lambda^{1/2}\|^{2r+1}. \tag{33}$$

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

  Note that, there exists $\|g\| \leq R$, $\|L_\lambda^{-1}L\| \leq 1$, $\|SC_\lambda^{-1/2}\| = \|C_\lambda^{-1/2}C_\lambda C_\lambda^{-1/2}\|^{1/2} \leq 1$, $\|\bar{S}_n^*\widehat{L}_{n\lambda}^{-1/2}\| \leq \|\widehat{L}_{n\lambda}^{-1/2}\widehat{L}_n\widehat{L}_{n\lambda}^{-1/2}\|^{1/2} \leq 1$ and $\|\widehat{C}_{n\lambda}^{-1/2}\bar{S}_n^*\| = \|\widehat{C}_{n\lambda}^{-1/2}\widehat{C}_n\widehat{C}_{n\lambda}^{-1/2}\|^{1/2} \leq 1$, and we then have

$$\begin{aligned}
&\|\widetilde{f}_{M,\lambda} - \widetilde{f}_\lambda\| \\
=&\|S(G_n\widehat{C}_{n\lambda} - I)\bar{S}_n^*\widehat{L}_{n\lambda}^{r-1}(\widehat{L}_{n\lambda}^{-1/2}L_\lambda^{1/2})^{2r}(L_\lambda^{-1}L)^r g\| \\
=&\|(SC_\lambda^{-1/2})C_\lambda^{1/2}(G_n\widehat{C}_{n\lambda} - I)\bar{S}_n^*\widehat{L}_{n\lambda}^{r-1}(\widehat{L}_{n\lambda}^{-1/2}L_\lambda^{1/2})^{2r}(L_\lambda^{-1}L)^r g\| \\
\leq&R\|C_\lambda^{1/2}(G_n\widehat{C}_{n\lambda} - I)\widehat{C}_{n\lambda}^{r-1/2}(\widehat{C}_{n\lambda}^{-1/2}\bar{S}_n^*)^{2r-1}(\bar{S}_n^*\widehat{L}_{n\lambda}^{-1/2})^{2-2r}(\widehat{L}_{n\lambda}^{-1/2}L_\lambda^{1/2})^{2r}\| \\
\leq&R\|C_\lambda^{1/2}(G_n\widehat{C}_{n\lambda} - I)C_\lambda^{r-1/2}\|\|C_\lambda^{-1/2}\widehat{C}_{n\lambda}^{1/2}\|^{2r-1}\|\widehat{L}_{n\lambda}^{-1/2}L_\lambda^{1/2}\|^{2r}.
\end{aligned} \tag{34}$$

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

$$\begin{aligned}
G_n\widehat{C}_{n\lambda} - I &= G_n\widehat{C}_{n\lambda}(I - VV^*) + G_n\widehat{C}_{n\lambda}VV^* - I \\
&= G_n\widehat{C}_{n\lambda}(I - VV^*) - (I - VV^*).
\end{aligned}$$

  Multiplying and dividing by $\widehat{C}_{n\lambda}^{1/2}$ and $C_\lambda^{1/2}$ and using above identity, we have

$$\begin{aligned}
&\|C_\lambda^{1/2}(G_n\widehat{C}_{n\lambda} - I)C_\lambda^{r-1/2}\| \leq \|C_\lambda^{1/2}(I - VV^*)C_\lambda^{r-1/2}\| \\
&+ \|C_\lambda^{1/2}\widehat{C}_{n\lambda}^{-1/2}\widehat{C}_{n\lambda}^{1/2}G_n\widehat{C}_{n\lambda}^{1/2}\widehat{C}_{n\lambda}^{1/2}C_\lambda^{-1/2}C_\lambda^{1/2}(I - VV^*)C_\lambda^{r-1/2}\| \\
&\leq \|C_\lambda^{1/2}(I - VV^*)C_\lambda^{r-1/2}\|(1 + \|C_\lambda^{1/2}\widehat{C}_{n\lambda}^{-1/2}\|\|\widehat{C}_{n\lambda}^{1/2}G_n\widehat{C}_{n\lambda}^{1/2}\|\|\widehat{C}_{n\lambda}^{1/2}C_\lambda^{-1/2}\|).
\end{aligned} \tag{35}$$

  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

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

  Using Cordes inequality (Proposition A.3) 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

$$\|C_\lambda^{1/2}(I - VV^*)C_\lambda^{r-1/2}\| \leq \|(I - VV^*)C_\lambda^{1/2}\|^{2r}. \tag{36}$$

  Substituting (35) and (36) into (34), there exists for $r \in [1/2, 1]$ that

$$\begin{aligned}
\|\widetilde{f}_{M,\lambda} - \widetilde{f}_\lambda\| &\leq R\|C_\lambda^{-1/2}\widehat{C}_{n\lambda}^{1/2}\|^{2r-1}\|\widehat{L}_{n\lambda}^{-1/2}L_\lambda^{1/2}\|^{2r}\|(I - VV^*)C_\lambda^{1/2}\|^{2r} \\
&= \|C_\lambda^{-1/2}\widehat{C}_{n\lambda}^{1/2}\|\|\widehat{L}_{n\lambda}^{-1/2}L_\lambda^{1/2}\|^2 R\,\|(I - VV^*)C_\lambda^{1/2}\|^{2r}.
\end{aligned}$$

Therefore, the Nyström error holds

$$\begin{aligned}
&\|\widetilde{f}_{M,\lambda} - \widetilde{f}_\lambda\| \leq 2(\|C_\lambda^{-1/2}\widehat{C}_{n\lambda}^{1/2}\|^2 + 1)R\lambda^{-1/2}\|(I - VV^*)C_\lambda^{1/2}\|^{2r+1}, \qquad \text{when } r \in (0, 1/2). \\
&\|\widetilde{f}_{M,\lambda} - \widetilde{f}_\lambda\| \leq \|C_\lambda^{-1/2}\widehat{C}_{n\lambda}^{1/2}\|\|\widehat{L}_{n\lambda}^{-1/2}L_\lambda^{1/2}\|^2 R\,\|(I - VV^*)C_\lambda^{1/2}\|^{2r}, \qquad \text{when } r \in [1/2, 1].
\end{aligned}$$

Together with Lemma A.18 and Lemma A.19, when the number of Nyström centers satisfies $M \gtrsim \mathcal{N}_\infty(\lambda)$ for uniform sampling and $M \gtrsim \mathcal{N}(\lambda)$ for data-dependent sampling, the Nyström error holds

$$\begin{aligned}
&\|\widetilde{f}_{M,\lambda} - \widetilde{f}_\lambda\| \leq 6(\|C_\lambda^{-1/2}\widehat{C}_{n\lambda}^{1/2}\|^2 + 1)R\lambda^r, \qquad \text{when } r \in (0, 1/2). \\
&\|\widetilde{f}_{M,\lambda} - \widetilde{f}_\lambda\| \leq 3\|C_\lambda^{-1/2}\widehat{C}_{n\lambda}^{1/2}\|\|\widehat{L}_{n\lambda}^{-1/2}L_\lambda^{1/2}\|^2 R\lambda^r, \qquad \text{when } r \in [1/2, 1].
\end{aligned}$$

□

In the proof of Lemma A.20, Assumption 3.3 is used such that the estimates of Nyström error is related to the quantity r. We bound the Nyström error for two cases $r \in (0, 1/2)$ ($f_\rho \notin \mathcal{H}$, induced by an imperfect kernel) and $r \in [1/2, 1]$ ($f_\rho \in \mathcal{H}$, induced by a perfect kernel). The key quantity $\|(I - VV^*)C_\lambda^{1/2}\|$ reflects the degree of approximation, and its upper bound depends on the sampling strategy. To guarantee $\|(I - VV^*)C_\lambda^{1/2}\|^2 \leq 3\lambda$, it needs $M = \Omega(n^{\frac{1}{2r+\gamma}})$ Nyström centers for uniform sampling, and $M = \Omega(n^{\frac{\gamma}{2r+\gamma}})$ Nyström centers for approximate leverage scores sampling.

A.3.3. Estimate for empirical error

Lemma A.21. *Let $\widetilde{f}_\lambda$ and f_λ be defined by* (15) *and* (16)*. The empirical error holds*

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

Proof. Recall the definitions of $\widetilde{f}_\lambda$ and f_λ with operators in Proposition A.12, it holds

$$\widetilde{f}_\lambda = S(\widehat{C}_n + \lambda I)^{-1}\bar{S}_n^* f_\rho = S\widehat{C}_{n\lambda}^{-1}\bar{S}_n^* f_\rho,$$
$$f_\lambda = S(C + \lambda I)^{-1}S^* f_\rho = SC_\lambda^{-1}S^* f_\rho.$$

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

$$\begin{aligned}
&\|\widetilde{f}_\lambda - f_\lambda\| \\
=&\|S\widehat{C}_{n\lambda}^{-1}\bar{S}_n^* f_\rho - SC_\lambda^{-1}S^* f_\rho\| \\
=&\|S\widehat{C}_{n\lambda}^{-1}(\bar{S}_n^* - S^*)f_\rho + S(\widehat{C}_{n\lambda}^{-1} - C_\lambda^{-1})S^* f_\rho\| \\
=&\|S\widehat{C}_{n\lambda}^{-1}(\bar{S}_n^* - S^*)f_\rho + S\widehat{C}_{n\lambda}^{-1}(C - \widehat{C}_n)C_\lambda^{-1}S^* f_\rho\| \\
=&\|S\widehat{C}_{n\lambda}^{-1}(\bar{S}_n^* - S^*)f_\rho + S\widehat{C}_{n\lambda}^{-1}(S^*S - \bar{S}_n^*S)C_\lambda^{-1}S^* f_\rho\| \\
=&\|S\widehat{C}_{n\lambda}^{-1}(\bar{S}_n^* - S^*)f_\rho + S\widehat{C}_{n\lambda}^{-1}(S^* - \bar{S}_n^*)f_\lambda\| \\
=&\|S\widehat{C}_{n\lambda}^{-1}\bar{S}_n^*(f_\rho - f_\lambda) + S\widehat{C}_{n\lambda}^{-1}S^*(f_\lambda - f_\rho)\| \\
=&\|SC_\lambda^{-1/2}C_\lambda^{1/2}\widehat{C}_{n\lambda}^{-1/2}\widehat{C}_{n\lambda}^{-1/2}\bar{S}_n^*(f_\rho - f_\lambda) + SC_\lambda^{-1/2}C_\lambda^{1/2}\widehat{C}_{n\lambda}^{-1/2}\widehat{C}_{n\lambda}^{-1/2}C_\lambda^{1/2}C_\lambda^{-1/2}S^*(f_\lambda - f_\rho)\|.
\end{aligned}$$

Note that, the following inequalities holds $\|SC_\lambda^{-1/2}\| = \|C_\lambda^{-1/2}CC_\lambda^{-1/2}\|^{1/2} \leq 1$, $\|\widehat{C}_{n\lambda}^{-1/2}\bar{S}_n^*\| = \|\widehat{C}_{n\lambda}^{-1/2}\widehat{C}_n\widehat{C}_{n\lambda}^{-1/2}\|^{1/2} \leq 1$, and $\|C_\lambda^{-1/2}S^*\| = \|C_\lambda^{-1/2}CC_\lambda^{-1/2}\|^{1/2} \leq 1$. Thus, we obtain

$$\|\widetilde{f}_\lambda - f_\lambda\| \leq [\|C_\lambda^{1/2}\widehat{C}_{n\lambda}^{-1/2}\| + \|C_\lambda^{1/2}\widehat{C}_{n\lambda}^{-1/2}\|^2]\|f_\lambda - f_\rho\|.$$

□

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.

A.3.4. Estimate for approximation error

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

Lemma A.22 (Approximation error)**.** *Let f_λ and f_ρ be defined by* (16) *and* (6)*. Under Assumption 3.3, the approximation error holds for any $\lambda > 0$ and $r > 0$,*

$$\|f_\lambda - f_\rho\| \leq R\lambda^r.$$

Proof. Under Assumption 3.3, 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

A.12), 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 3.3. 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.4. Proof of Main Results

Proof of Theorem 3.9. Firstly, we recall the error decomposition of $\mathcal{E}(\widehat{f}_{M,\lambda}) - \mathcal{E}(f_\rho)$ in Lemma A.13 that is

$$\mathcal{E}(\widehat{f}_{M,\lambda}) - \mathcal{E}(f_\rho) \leq\ 4\|\widehat{f}_{M,\lambda} - \widetilde{f}_{M,\lambda}\|_\rho^2 +\ 4\|\widetilde{f}_{M,\lambda} - \widetilde{f}_\lambda\|_\rho^2 +\ 4\|\widetilde{f}_\lambda - f_\lambda\|_\rho^2 +\ 4\|f_\lambda - f_\rho\|_\rho^2. \tag{37}$$

We need to combine analytical results for those four errors in Lemmas A.14, A.20, A.21 and A.22. We combine the constraints about n used in the first three error terms as $n \geq 16(\mathcal{N}_\infty(\lambda) + 1)\log(2/\delta)$. Let $\lambda = n^{-\frac{1}{2r+\gamma}}$ and $\mathcal{N}_\infty \leq F\lambda^{-\alpha}$ under Assumption 3.8, and then the restrict on n becomes almost surely $n \leq n^{\frac{\alpha}{2r+\gamma}}$. It holds

$$2r + \gamma \geq \alpha. \tag{38}$$

Estimate sample variance. According to Lemma A.14 and Lemma A.9, when $n \geq 16(\mathcal{N}_\infty(\lambda) + 1)\log(2/\delta)$, it holds with the probability at least $1 - \delta$

$$\|\widehat{f}_{M,\lambda} - \widetilde{f}_{M,\lambda}\| \leq 8\left(\frac{B\sqrt{\mathcal{N}_\infty(\lambda)}}{n} + \sqrt{\frac{B^2\mathcal{N}(\lambda)}{n}}\right)\log\frac{2}{\delta}.$$

The sample variance can be bounded by

$$\begin{aligned}\|\widehat{f}_{M,\lambda} - \widetilde{f}_{M,\lambda}\|_\rho^2 \ \leq&\ 64\left(\frac{B\sqrt{\mathcal{N}_\infty(\lambda)}}{n} + \sqrt{\frac{B^2\mathcal{N}(\lambda)}{n}}\right)^2 \log^2\frac{2}{\delta} \\ \leq&\ 128\left(\frac{B^2\mathcal{N}_\infty(\lambda)}{n^2} + \frac{B^2\mathcal{N}(\lambda)}{n}\right)\log^2\frac{2}{\delta} \\ \leq&\ 128\left(\kappa^2 B^2 n^{\frac{\alpha-4r-2\gamma}{2r+\gamma}} + \kappa^2 B^2 n^{\frac{-2r}{2r+\gamma}}\right)\log^2\frac{2}{\delta}.\end{aligned} \tag{39}$$

To ensure the convergence rate for the sample variance in (39) is optimal $\mathcal{O}(n^{\frac{-2r}{2r+\gamma}})$, the above inequality yields $n^{\frac{\alpha-4r-2\gamma}{2r+\gamma}} \leq n^{\frac{-2r}{2r+\gamma}}$, which leads to the following restriction

$$2r + 2\gamma \geq \alpha. \tag{40}$$

Combining the restrictions for covariance operator difference $2r + \gamma \geq \alpha$ (38) and sample variance $2r + 2\gamma \geq \alpha$ (40), we provide the restriction $2r + \gamma \geq \alpha$. We then bound the sample variance as follows:

$$\|\widehat{f}_{M,\lambda} - \widetilde{f}_{M,\lambda}\|_\rho^2 \leq 128\left(\kappa^2 B^2 n^{\frac{\alpha-4r-2\gamma}{2r+\gamma}} + \kappa^2 B^2 n^{\frac{-2r}{2r+\gamma}}\right)\log^2\frac{2}{\delta} \leq c_1 n^{\frac{-2r}{2r+\gamma}}, \tag{41}$$

where $c_1 = 256\kappa^2 B^2$.

Estimate Nyström error. According to Lemma A.20, Lemma A.9, and Lemma A.10, when $M \geq 67 \log \frac{4\kappa^2}{\lambda\delta} \vee 5\mathcal{N}_\infty(\lambda) \log \frac{4\kappa^2}{\lambda\delta}$ and $n \geq 16(\mathcal{N}_\infty(\lambda) + 1)\log(2/\delta)$, Nyström error holds

$$\|\widetilde{f}_{M,\lambda} - \widetilde{f}_\lambda\| \leq 18R\lambda^r, \qquad \text{when } r \in (0, 1/2).$$
$$\|\widetilde{f}_{M,\lambda} - \widetilde{f}_\lambda\| \leq 9R\lambda^r, \qquad \text{when } r \in [1/2, 1].$$

Then, the Nyström error can be stated as

$$\|\widetilde{f}_{M,\lambda} - \widetilde{f}_\lambda\|_\rho^2 \leq 324R^2 n^{\frac{-2r}{2r+\gamma}}. \tag{42}$$

Estimate empirical error. According Lemma A.21 and Lemma A.9, when $n \geq 16(\mathcal{N}_\infty(\lambda) + 1)\log(2/\delta)$, there holds

$$\|\widetilde{f}_\lambda - f_\lambda\|_\rho^2 \leq 4R^2 n^{\frac{-2r}{2r+\gamma}}. \tag{43}$$

Estimate Approximation error. According Lemma A.22, for $\lambda > 0$ and $r > 0$ there exists

$$\|f_\lambda - f_\rho\|_\rho^2 \leq R^2 n^{\frac{-2r}{2r+\gamma}}. \tag{44}$$

Substituting (39), (42), (43) and (44) to (37), we prove the final result. With probability $1 - \delta$, the conditions $n \geq 16(\mathcal{N}_\infty(\lambda) + 1)\log(2/\delta)$ and

$$M \geq 67 \log \frac{4\kappa^2}{\lambda\delta} \vee 5\mathcal{N}_\infty(\lambda) \log \frac{4\kappa^2}{\lambda\delta}.$$

can guarantee the optimal error bound for Nyström approximation with uniform sampling

$$\mathcal{E}(\widehat{f}_{M,\lambda}) - \mathcal{E}(f_\rho) \leq c_2 n^{\frac{-2r}{2r+\gamma}}.$$

where $c_2 = 256\kappa^2 B^2 + 987R^2$. □

Proof of Corollary 3.12. Under Assumption 3.8, using Theorem 3.9 and Nyström approximation with uniform subsampling, we can obtain the desired results when considering the worst case $\mathcal{N}_\infty(\lambda) \leq \frac{\kappa^2}{\lambda}$ with $\alpha = 1$. □

Proof of Corollary 3.15. Let $\lambda = n^{-\frac{1}{2r+\gamma}}$, and $\mathcal{N}_\infty(\lambda) \simeq \mathcal{N}(\lambda)$ for the data-dependent sampling strategy (Rudi & Rosasco, 2017) (Example 2), and then the restrict on n becomes almost surely $n \geq n^{\frac{\gamma}{2r+\gamma}}$ by Assumption 3.5. It holds for the entire range of source condition $r \in (0, 1], \gamma \in [0, 1]$. □