Improved Bounds for the Nyström Method with Application to Kernel Classification

Rong Jin, *Member, IEEE,* Tianbao Yang, Mehrdad Mahdavi,
Yu-Feng Li, Zhi-Hua Zhou, *Fellow, IEEE*

***Abstract*—We develop two approaches for analyzing the approximation error bound for the Nyström method that approximates a positive semidefinite (PSD) matrix by sampling a small set of columns, one based on the concentration inequality of integral operator, and one based on the random matrix theory. We show that the approximation error, measured in the spectral norm, can be improved from $O(N/\sqrt{m})$ to $O(N/m^{1-\rho})$ in the case of large eigengap, where N is the total number of data points, m is the number of sampled data points, and $\rho \in (0, 1/2)$ is a positive constant that characterizes the eigengap. When the eigenvalues of the kernel matrix follow a p-power law, our analysis based on random matrix theory further improves the bound to $O(N/m^{p-1})$ under an incoherence assumption. We present a kernel classification approach based on the Nyström method and derive its generalization performance using the improved bound. We show that when the eigenvalues of kernel matrix follow a p-power law, we can reduce the number of support vectors to $N^{2p/(p^2-1)}$, which is sublinear in N when $p > 1 + \sqrt{2}$, without seriously sacrificing its generalization performance.**

***Index Terms*—Nyström method, approximation error, concentration inequality, kernel methods, random matrix theory**

I. Introduction

The Nyström method has been widely applied in machine learning to efficiently approximate large kernel matrices with low rank matrices in order to speed up kernel algorithms [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11]. In order to evaluate the quality of the Nyström method, we typically bound the norm of the difference between the original kernel matrix and the low rank approximate matrix generated by the Nyström method. Several analysis were developed to bound the approximation error of the Nyström method [2], [4], [9], [12], [10], [13], [14]. Most of them focus on additive error bound, and base on the theoretical results from [2]. When the target matrix is of low rank, significantly better bounds for the approximation error of the Nyström method were given in [10] and [13]. These results are further generalized to kernel matrices of an arbitrary rank by a relative error bound in [14]. In this study, we focus on the additive error bound of the Nyström method for PSD matrices, and will compare our results mainly to the ones stated in [2]. Although a relative error bound is usually tighter than an additive bound [15], we show in this paper that the approximation error resulting from our additive error analysis is better than that indicated by the relative error analysis in [14], when the kernel matrix follows a power law eigen-value distribution [1]. Below, we review the main results in [2] and their limitations.

Let $K \in \mathbb{R}^{N\times N}$ be a PSD kernel matrix to be approximated, and $\lambda_i, i = 1, \ldots, N$ be the eigenvalues of K ranked in the descending order. Let $\widetilde{K}(r)$ be an approximate kernel matrix of rank r generated by the Nyström method. Let m be the number of columns sampled from K used to construct $\widetilde{K}(r)$. Then, under the assumption $K_{i,i} = O(1)$, Drineas and Mahoney [2] showed that for any m uniformly sampled columns [2], with a high probability,

$$\|K - \widetilde{K}(r)\|_2 \leq \lambda_{r+1} + O\left(\frac{N}{\sqrt{m}}\right),$$

where $\|\cdot\|_2$ stands for the spectral norm of a matrix. In particular, by setting $r = m$, the bound becomes

$$\|K - \widetilde{K}(m)\|_2 \leq \lambda_{m+1} + O\left(\frac{N}{\sqrt{m}}\right). \quad (1)$$

The main problem with the bound in (1) is its slow reduction rate in the number of sampled columns (i.e., $O(m^{-1/2})$), implying that a large number of samples is needed in order to achieve a small approximation error.

In this study, we aim to improve the approximation error bound in (1) by considering two special cases of the kernel matrix K. In the first case, we assume there is a large eigengap in the spectrum of K. More specifically, we assume there exists a rank $r \in [N]$ such that $\lambda_r = \Omega(N/m^{\rho})$ and $\lambda_{r+1} = O(N/m^{1-\rho})$, where $\rho < 1/2$. Here, parameter ρ is introduced to characterize the eigengap $\lambda_r - \lambda_{r+1}$: the smaller the ρ, the larger the eigengap will be. We show that the approximation error bound can be improved to $O(N/m^{1-\rho})$ in the case of large eigengap. The second case assumes that the eigenvalues of K follow a p-power law with $p > 1$. Under this assumption, we obtain an improved $O(N/m^{p-1})$ bound on the approximation error, provided that the eigenvector matrix sat-

Rong Jin and Mehrdad Mahrdavi are with the Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, 48824 USA. E-mail: {rongjin, mahdavim}@cse.msu.edu.

Tianbao Yang is with GE Global Research, San Ramon, CA, 94583 USA. Email: tyang@ge.com

Yu-Feng Li and Zhi-Hua Zhou are with National Key Laboratory of Novel Software Technology, Nanjing University, Nanjing, 210023, China. Email: {liyf, zhouzh}@lamda.nju.edu.cn

[1] For completeness, we did include the comparison to the relative error bound in [14] in the later remarks.

[2] Although the main results in [2] use a data dependent sampling scheme, they also hold for the uniform sampling mentioned by the authors of [2] and explicitly established in [16].

isfies an incoherence assumption [3]. This result may shed light on why the Nyström method works well for kernel matrices with power law eigenvalue distributions, a phenomenon that is often observed in practice [10].

The second contribution of this study is a kernel classification algorithm that explicitly explores the improved bounds of the Nyström method developed here. We show that when the eigenvalues of the kernel matrix follow a p-power law with $p > 1$, we can construct a kernel classifier that yields a similar generalization performance as the full version of kernel classifier but with no more than $N^{2p/(p^2-1)}$ support vectors, which is sublinear in N when $p > (1+\sqrt{2})$. Although the generalization error bound of using the Nyström method for classification has been studied in [11], to the best of knowledge, this is the first work that bounds the number of support vectors using the analysis of the Nyström method.

Although the results presented in this paper are theoretical, they can find applications in the real world. The result for a PSD matrix with a large eigen-gap can find applications in document analysis [17] and social network analysis [18]. Several works [19], [20] have observed that a Guassian kernel matrix usually exhibits a skewed eigen-value distribution on real data sets, which make our result for a PSD matrix with a power law eigen-value distribution attractive.

Finally, as we prepare the final version of the manuscript, two related works [21], [22] have been accepted for publication. In [21], the authors compared the Nyström method with different sampling and projection strategies and established improved bounds for these strategies. Their result for the spectral norm error bound with the uniform sampling strategy, as considered in this work, is similar to the one presented in [14]. In [22], the author also addresses the generalization error of kernel learning using the Nyström method. The key result of [22] indicates that the minimum number of sampled columns depends on the degree of freedom. Our result addresses the sample complexity of kernel learning from a different aspect.

II. Notations and Background

Let $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ be a collection of N samples, where $\mathbf{x}_i \in \mathcal{X}$, and $K = [\kappa(\mathbf{x}_i, \mathbf{x}_j)]_{N\times N}$ be the kernel matrix for the samples in $\mathcal{D}$, where $\kappa(\cdot,\cdot)$ is a kernel function. For simplicity, we assume $\kappa(\mathbf{x},\mathbf{x}) \leq 1$ for any $\mathbf{x} \in \mathcal{X}$. We denote by $(\mathbf{v}_i, \lambda_i), i = 1, \ldots, N$ the eigenvectors and eigenvalues of K ranked in the descending order of eigenvalues, and by $V = (\mathbf{v}_1, \cdots, \mathbf{v}_N)$ the orthonormal eigenvector matrix. In order to build the low rank approximation of kernel matrix K, the Nyström method first samples $m < N$ examples randomly from $\mathcal{D}$, denoted by $\widehat{\mathcal{D}} = \{\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_m\}$. Let $\widehat{K} = [\kappa(\widehat{\mathbf{x}}_i, \widehat{\mathbf{x}}_j)]_{m\times m}$ measure the kernel similarity between any two samples in $\widehat{\mathcal{D}}$ and $K_b = [\kappa(\mathbf{x}_i, \widehat{\mathbf{x}}_j)]_{N\times m}$ measure the similarity between the samples in $\mathcal{D}$ and $\widehat{\mathcal{D}}$. Using the samples in $\widehat{\mathcal{D}}$, with rank r set to m (or the rank of $\widehat{K}$ if it is less than m), the Nyström method approximates K by $K_b\widehat{K}^\dagger K_b^\top$, where $\widehat{K}^\dagger$ denote the pseudo inverse of $\widehat{K}$. Our goal is to provide a high probability bound for the approximation error $\left\|K - K_b\widehat{K}^\dagger K_b^\top\right\|_2$. We choose $r = m$ (or the rank of $\widehat{K}$) because according to [2], [4], it yields the best approximation error for a non-singular kernel matrix. In this study, we focus on the spectral norm for measuring the approximation error, which is particularly suitable for kernel classification [11]. We also restrict the analysis to the uniform sampling for the Nyström method. This is because according to [4], [16], for most real-world datasets, uniform sampling is computationally more efficient, and yields performance that is comparable to, if not better than, the non-uniform sampling strategies [2], [8], [23], [24].

Our analysis for the Nyström method extensively exploits the properties of the integral operator. This is in contrast to most of the previous studies for the Nyström method that rely on matrix analysis. The main advantage of using the integral operator is its convenience in handling the unseen data points (i.e., test data), making it attractive for the analysis of generalization error bounds. In particular, we introduce a linear operator L_N defined over the samples in $\mathcal{D}$. For any function $f(\cdot)$, operator L_N is defined as

$$L_N[f](\cdot) = \frac{1}{N}\sum_{i=1}^{N}\kappa(\mathbf{x}_i,\cdot)f(\mathbf{x}_i).$$

It can be shown that the eigenvalues of the operator L_N are $\lambda_i/N, i = 1, \ldots, N$ [25]. Let $\varphi_1(\cdot), \ldots, \varphi_N(\cdot)$ be the corresponding eigenfunctions of L_N that are normalized by functional norm, i.e., $\langle\varphi_i, \varphi_j\rangle_{\mathcal{H}_\kappa} = \delta(i,j), 1 \leq i \leq j \leq N$, where $\langle\cdot,\cdot\rangle_{\mathcal{H}_\kappa}$ denotes the inner product in $\mathcal{H}_\kappa$ and $\delta(i,j)$ is the Kronecker delta function. According to [25], the eigenfunctions satisfy

$$\sqrt{\lambda_j}\varphi_j(\cdot) = \sum_{i=1}^{N} V_{i,j}\kappa(\mathbf{x}_i,\cdot), j = 1, \cdots, N, \tag{2}$$

where $V_{i,j}$ is the (i,j)th element in V. Similarly, we can write $\kappa(\mathbf{x}_j,\cdot)$ by its eigen-expansion as

$$\kappa(\mathbf{x}_j,\cdot) = \sum_{i=1}^{N}\sqrt{\lambda_i}V_{j,i}\varphi_i(\cdot), j = 1, \ldots, N. \tag{3}$$

Furthermore, let L_m be an operator defined on the samples in $\widehat{\mathcal{D}}$, i.e.,

$$L_m[f](\cdot) = \frac{1}{m}\sum_{i=1}^{m}\kappa(\widehat{\mathbf{x}}_i,\cdot)f(\widehat{\mathbf{x}}_i).$$

Finally we denote by $\langle f, g\rangle_{\mathcal{H}_\kappa}$ and $\|f\|_{\mathcal{H}_\kappa}$ the inner product and function norm in Hilbert space $\mathcal{H}_\kappa$, respectively, and denote by $\|L\|_{HS}$ and $\|L\|_2$ the Hilbert Schmid norm and spectral norm of a linear operator L, respectively, i.e.

$$\|L\|_{HS} = \sqrt{\sum_{i,j}\langle\varphi_i, L\varphi_j\rangle^2_{\mathcal{H}_\kappa}} \text{ and } \|L\|_2 = \max_{\|f\|_{\mathcal{H}_\kappa}\leq 1}\|Lf\|_{\mathcal{H}_\kappa},$$

where $\{\varphi_i, i = 1, \cdots,\}$ is a complete orthogonal basis of $\mathcal{H}_\kappa$. The two norms are the analogs of Frobenius and spectral norm in Euclidean space, respectively. In the following analysis, omitted proofs are presented in the appendix.

[3] A similar assumption was used in the previous analysis of the Nyström method [10], [13], [14].

III. Approximation Error Bound by the Nyström Method

Our first step is to turn $\|K - K_b\widehat{K}^\dagger K_b^\top\|_2$ into a functional approximation problem. To this end, we introduce two sets:

$$\mathcal{H}_a = \text{span}\left(\kappa(\widehat{\mathbf{x}}_1,\cdot),\ldots,\kappa(\widehat{\mathbf{x}}_m,\cdot)\right)$$
$$\mathcal{H}_b = \left\{ f(\cdot) = \sum_{i=1}^N u_i\kappa(\mathbf{x}_i,\cdot) : \sum_{i=1}^N u_i^2 \le 1 \right\},$$

where $\mathcal{H}_a$ is the subspace spanned by kernel functions defined on the samples in $\widehat{\mathcal{D}}$, and $\mathcal{H}_b$ is a subset of a functional space spanned by kernel functions defined on the samples in $\mathcal{D}$ with bounded coefficients. Using the eigen-expansion of $\kappa(\mathbf{x}_j,\cdot)$ in (3), it is straightforward to show that $\mathcal{H}_b$ can be rewritten in the basis of the eigenfunctions $\{\varphi_i\}_{i=1}^N$

$$\mathcal{H}_b = \left\{ f(\cdot) = \sum_{i=1}^N w_i\sqrt{\lambda_i}\varphi_i(\cdot) : \sum_{i=1}^N w_i^2 \le 1 \right\}.$$

Define $\mathcal{E}(g,\mathcal{H}_a)$ as the minimum error in approximating a function $g\in\mathcal{H}_b$ by functions in $\mathcal{H}_a$, i.e.,

$$\begin{aligned}\mathcal{E}(g,\mathcal{H}_a) &= \min_{f\in\mathcal{H}_a} \|f-g\|_{\mathcal{H}_\kappa}^2 \\ &= \|f\|_{\mathcal{H}_\kappa}^2 + \|g\|_{\mathcal{H}_\kappa}^2 - 2\langle f,g\rangle_{\mathcal{H}_\kappa}.\end{aligned}$$

Define $\mathcal{E}(\mathcal{H}_a)$ as the worst error in approximating any function $g\in\mathcal{H}_b$ by functions in $\mathcal{H}_a$, i.e.,

$$\mathcal{E}(\mathcal{H}_a) = \max_{g\in\mathcal{H}_b} \mathcal{E}(g,\mathcal{H}_a). \tag{4}$$

The following proposition connects $\left\|K - K_b\widehat{K}^\dagger K_b^\top\right\|_2$ with $\mathcal{E}(\mathcal{H}_a)$.

Proposition 1: For any random samples $\widehat{\mathbf{x}}_1,\ldots,\widehat{\mathbf{x}}_m$, we have

$$\left\|K - K_b\widehat{K}^\dagger K_b^\top\right\|_2 = \mathcal{E}(\mathcal{H}_a).$$

Proof: Since $g\in\mathcal{H}_b$ and $f\in\mathcal{H}_a$, we can write g and f as

$$g = \sum_{i=1}^N u_i\kappa(\mathbf{x}_i,\cdot) \quad\text{and}\quad f = \sum_{i=1}^m z_i\kappa(\widehat{\mathbf{x}}_i,\cdot),$$

where $\mathbf{u} = (u_1,\ldots,u_N)^\top \in \mathbb{R}^N$ satisfies $\|\mathbf{u}\|_2 \le 1$ and $\mathbf{z} = (z_1,\ldots,z_m)^\top \in \mathbb{R}^m$. We thus can rewrite $\mathcal{E}(g,\mathcal{H}_a)$ as an optimization problem in terms of $\mathbf{z}$, i.e.,

$$\begin{aligned}\mathcal{E}(g,\mathcal{H}_a) &= \min_{\mathbf{z}\in\mathbb{R}^m} \mathbf{z}^\top\widehat{K}\mathbf{z} - 2\mathbf{u}^\top K_b\mathbf{z} + \mathbf{u}^\top K\mathbf{u} \\ &= \mathbf{u}^\top\left(K - K_b\widehat{K}^\dagger K_b^\top\right)\mathbf{u},\end{aligned}$$

and therefore

$$\begin{aligned}\mathcal{E}(\mathcal{H}_a) &= \max_{g\in\mathcal{H}_b}\mathcal{E}(g,\mathcal{H}_a) \\ &= \max_{\|\mathbf{u}\|_2\le 1} \mathbf{u}^\top\left(K - K_b\widehat{K}^\dagger K_b^\top\right)\mathbf{u} \\ &= \left\|K - K_b\widehat{K}^\dagger K_b^\top\right\|_2.\end{aligned}$$

■

Remark 1: We can restrict the space $\mathcal{H}_a$ to its subspace $\mathcal{H}_a^r = \left\{\sum_{i=1}^m z_i\kappa(\widehat{\mathbf{x}}_i,\cdot) : \mathbf{z}\in\text{span}(\widehat{\mathbf{v}}_1,\ldots,\widehat{\mathbf{v}}_r)\right\}$, where $\widehat{\mathbf{v}}_i, i = 1,\ldots,r$ are the first r eigenvectors of $\widehat{K}$, to conduct the analysis for the rank $r<m$ approximation of the Nyström method.

To proceed our analysis, for any $r\in[N]$ we define

$$\begin{aligned}\mathcal{H}_r &= \text{span}(\varphi_1(\cdot),\cdots,\varphi_r(\cdot)),\\ \overline{\mathcal{H}}_r &= \text{span}(\varphi_{r+1}(\cdot),\cdots,\varphi_N(\cdot)),\\ \mathcal{H}_b^r &= \left\{ f(\cdot) = \sum_{i=1}^r w_i\sqrt{\lambda_i}\varphi_i(\cdot) : \sum_{i=1}^r w_i^2 \le 1\right\},\\ \overline{\mathcal{H}}_b^r &= \left\{ f(\cdot) = \sum_{i=1}^{N-r} w_i\sqrt{\lambda_{i+r}}\varphi_{i+r}(\cdot) : \sum_{i=1}^{N-r} w_i^2 \le 1\right\}.\end{aligned}$$

Define $\mathcal{E}(\mathcal{H}_a,r) = \max_{g\in\mathcal{H}_b^r}\mathcal{E}(g,\mathcal{H}_a)$ as the worst error in approximating any function $g\in\mathcal{H}_b^r$ by functions in $\mathcal{H}_a$. The proposition below bounds $\mathcal{E}(\mathcal{H}_a)$ by $\mathcal{E}(\mathcal{H}_a,r)$.

Proposition 2: For any $r\in[N]$, we have

$$\mathcal{E}(\mathcal{H}_a) \le \max\left(\mathcal{E}(\mathcal{H}_a,r),\lambda_{r+1}\right) \le \mathcal{E}(\mathcal{H}_a,r) + \lambda_{r+1}.$$

Proof: We first note that any $f\in\mathcal{H}_a$ can be written as $f = f_1 + f_2$, where $f_1\in\mathcal{H}_a\cap\mathcal{H}_r$, and $f_2\in\mathcal{H}_a\cap\overline{\mathcal{H}}_r$. For any $g\in\mathcal{H}_b$, we can write $g = g_1 + g_2$, where $g_1\in\sqrt{1-\delta}\mathcal{H}_b^r$, $g_2\in\sqrt{\delta}\overline{\mathcal{H}}_b^r$, and $\delta\in[0,1]$. Using these notations, we rewrite $\mathcal{E}(\mathcal{H}_a)$ as

$$\begin{aligned}&\mathcal{E}(\mathcal{H}_a)\\ &= \max_{\substack{\delta\in[0,1]\\ g_1\in\sqrt{1-\delta}\mathcal{H}_b^r\\ g_2\in\sqrt{\delta}\overline{\mathcal{H}}_b^r}} \min_{\substack{f_1\in\mathcal{H}_a\cap\mathcal{H}_r\\ f_2\in\mathcal{H}_a\cap\overline{\mathcal{H}}_r}} \|f_1-g_1\|^2 + \|f_2-g_2\|_{\mathcal{H}_\kappa}^2\\ &\le \max_{\delta\in[0,1]}(1-\delta)\max_{g\in\mathcal{H}_b^r}\min_{f\in\mathcal{H}_a\cap\mathcal{H}_r}\|f-g\|_{\mathcal{H}_\kappa}^2 + \delta\max_{g\in\overline{\mathcal{H}}_b^r}\|g\|_{\mathcal{H}_\kappa}^2\\ &= \max_{\delta\in[0,1]}\left\{(1-\delta)\max_{g\in\mathcal{H}_b^r}\min_{f\in\mathcal{H}_a}\|f-g\|_{\mathcal{H}_\kappa}^2 + \delta\max_{g\in\overline{\mathcal{H}}_b^r}\|g\|_{\mathcal{H}_\kappa}^2\right\}\\ &= \max_{\delta\in[0,1]}(1-\delta)\mathcal{E}(\mathcal{H}_a,r) + \delta\lambda_{r+1} = \max\left(\mathcal{E}(\mathcal{H}_a,r),\lambda_{r+1}\right),\end{aligned}$$

where the second equality follows that for any $g\in\mathcal{H}_b^r$, $\min_{f\in\mathcal{H}_a}\|f-g\|_{\mathcal{H}_\kappa}^2 = \min_{f\in\mathcal{H}_a\cap\mathcal{H}_r}\|f-g\|_{\mathcal{H}_\kappa}^2$, and the last inequality follows the definition of $\mathcal{E}(\mathcal{H}_a,r)$. ■

As indicated by Proposition 2, in order to bound the approximation error $\mathcal{E}(\mathcal{H}_a)$, we can bound $\mathcal{E}(\mathcal{H}_a,r)$, namely the approximation error for functions in the subspace spanned by the top eigenfunctions of L_N. In the next two subsections, we discuss two approaches for bounding $\mathcal{E}(\mathcal{H}_a,r)$: the first approach relies on the concentration inequality of integral operator [25], and the second approach explores the random matrix theory [26]. Before proceeding to upper bound $\mathcal{E}(\mathcal{H}_a)$, we first provide a lower bound for $\mathcal{E}(\mathcal{H}_a)$.

Theorem 1: There exists a kernel matrix $K\in\mathbb{R}^{N\times N}$ with all its diagonal entries being 1 such that for any sampling strategy that selects m columns, the approximation error of the Nyström method is lower bounded by $\Omega(\frac{N}{m})$, i.e.,

$$\left\|K - K_b\widehat{K}^\dagger K_b^\top\right\|_2 \ge \Omega\left(\frac{N}{m}\right),$$

provided $N > 64[\ln 4]^2 m^2$.

Remark 2: Theorem 1 shows that the lower bound for the approximation error of the Nyström method is $\Omega(N/m)$. The analysis developed in this work aims to bridge the gap between the known upper bound (i.e., $O(N/\sqrt{m})$) and the obtained lower bound.

A. Bound for $\mathcal{E}(\mathcal{H}_a, r)$ using Concentration Inequality of Integral Operator

In this section, we bound $\mathcal{E}(\mathcal{H}_a, r)$ using the concentration inequality of integral operator. We show that the approximation error of the Nyström method can be improved to $O(N/m^{1-\rho})$ when there is a large eigengap in the spectrum of kernel matrix K, where $\rho < 1/2$ is introduced to characterize the eigengap. We first state the concentration inequality of a general random variable.

Proposition 3: (Proposition 1 [25]) Let ξ be a random variable on $(\mathcal{X}, P_{\mathcal{X}})$ with values in a Hilbert space $(\mathcal{H}, \|\cdot\|)$. Assume $\|\xi\| \leq M < \infty$ is almost sure, then with a probability at least $1-\delta$, we have

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \xi(\mathbf{x}_i) - \mathrm{E}[\xi] \right\| \leq \frac{4M \ln(2/\delta)}{\sqrt{m}}.$$

The approximation error of the Nyström method using the concentration inequality is given in the following theorem.

Theorem 2: With a probability at least $1-\delta$, for any $r \in [N]$, we have

$$\left\| K - K_b \widehat{K}^{\dagger} K_b^{\top} \right\|_2 \leq \frac{16[\ln(2/\delta)]^2 N^2}{m\lambda_r} + \lambda_{r+1}.$$

We consider the scenario where there is very large eigengap in the spectrum of kernel matrix K. In particular, we assume that there exists a rank r and $\rho \in (0, 1/2)$ such that $\lambda_r = \Omega(N/m^{\rho})$ and $\lambda_{r+1} = O(N/m^{1-\rho})$. Parameter ρ is introduced to characterize the eigengap which is given by

$$\lambda_r - \lambda_{r+1} = \Omega\left(\frac{N}{m^{\rho}} - \frac{N}{m^{1-\rho}}\right) = \Omega\left(\frac{N}{m^{\rho}}\left[1 - \frac{1}{m^{1-2\rho}}\right]\right)$$

Evidently, the smaller the ρ, the larger the eigengap. When $\rho = 1/2$, the eigengap is small. Under the large eigengap assumption, the bound in Theorem 2 is simplified as

$$\left\| K - K_b \widehat{K}^{\dagger} K_b^{\top} \right\|_2 \leq O\left(\frac{N}{m^{1-\rho}}\right). \tag{5}$$

Compared to the bound in (1), the bound in (5) improves the approximation error from $O(N/\sqrt{m})$ to $O(N/m^{1-\rho})$, when $\rho < 1/2$.

To prove Theorem 2, we define two sets of functions

$$\mathcal{H}_c^r = \left\{ h = \sum_{i=1}^{r} w_i \sqrt{\lambda_i} \varphi_i(\cdot) : \frac{1}{N^2} \sum_{i=1}^{r} w_i^2 \lambda_i^2 \leq 1 \right\},$$
$$\mathcal{H}_d^r = \left\{ f \in \mathcal{H}_\kappa : \|f\|_{\mathcal{H}_\kappa}^2 \leq N^2/\lambda_r \right\}.$$

where r corresponds to the rank with a large eigengap. It is evident that $\mathcal{H}_c^r \subseteq \mathcal{H}_d^r$; and for any $g \in \mathcal{H}_b^r$, it can also be written as $g = L_N[h]$, where $h \in \mathcal{H}_c^r$.

Using $\mathcal{H}_c^r$ and $\mathcal{H}_d^r$, we have

$$\begin{aligned}
\mathcal{E}(\mathcal{H}_a, r) &= \max_{g \in \mathcal{H}_b^r} \mathcal{E}(g, \mathcal{H}_a) \\
&= \max_{h \in \mathcal{H}_c^r} \min_{f \in \mathcal{H}_a} \|L_N h - f\|_{\mathcal{H}_\kappa}^2 \\
&\leq \max_{h \in \mathcal{H}_d^r} \min_{f \in \mathcal{H}_a} \|L_N h - f\|_{\mathcal{H}_\kappa}^2.
\end{aligned}$$

By constructing f as $L_m[h]$ we can bound $\mathcal{E}(\mathcal{H}_a, r)$ as

$$\begin{aligned}
\mathcal{E}(\mathcal{H}_a, r) &\leq \max_{h \in \mathcal{H}_d^r} \min_{f \in \mathcal{H}_a} \|L_N(h) - f\|_{\mathcal{H}_\kappa}^2 \\
&\leq \max_{h \in H_d^r} \|(L_N - L_m)h\|_{\mathcal{H}_\kappa}^2 \\
&\leq \|L_N - L_m\|_2^2 \frac{N^2}{\lambda_r} \\
&\leq \|L_N - L_m\|_{HS}^2 \frac{N^2}{\lambda_r}, \qquad (6)
\end{aligned}$$

where the last step follows the fact $\|L_N - L_m\|_2 \leq \|L_N - L_m\|_{HS}$. The corollary below followed immediately from Proposition 3, allows us to bound the difference between L_N and L_m and .

Corollary 1: With a probability $1-\delta$, we have

$$\|L_N - L_m\|_{HS} \leq \frac{4\ln(2/\delta)}{\sqrt{m}}.$$

Finally, Theorem 2 follows directly the inequality in (6) and the result in Corollary 1.

B. Bound for $\mathcal{E}(\mathcal{H}_a, r)$ using Random Matrix Theory

In this subsection, we aim to develop a better error bound for the Nyström method for kernel matrices with eigenvalues that follow a power law distribution. Our analysis explicitly explores some of the key results in random matrix theory that has been the main building blocks in the theory of compressive sensing [26], [27]. To this end, we first introduce the definition of the power law distribution of eigenvalues [28], [29]. The eigenvalues $\sigma_i, i = 1, \ldots$ ranked in the non-increasing order follows a p-power law (distribution) if there exists a constant $c > 0$ such that

$$\sigma_k \leq ck^{-p}.$$

In the sequel, we assume the normalized eigenvalues $\lambda_i/N, i = 1, \ldots, N$ (i.e., the eigenvalues of the operator L_N), follow a p-power law distribution [4]. A well-known example of kernel with a power law eigenvalue distribution [28] is the kernel function that generates Sobolev Spaces $W^{\alpha,2}(\mathbb{T}^d)$ of smoothness $\alpha > d/2$, where $\mathbb{T}^d$ is d-dimensional torus. Its eigenvalues follow a p-power law with $p = 2\alpha > d$. It is also observed that the eigenvalues of a Gaussian kernel by appropriately setting the width parameter follow a power law distribution [19].

In order to exploit the result of random matrices from [26], we introduce the definition of the coherence μ for the eigeven-vector matrix $V = (\mathbf{v}_1, \ldots, \mathbf{v}_N)$ as

$$\mu = \sqrt{N} \max_{1 \leq i,j \leq N} |V_{i,j}|.$$

[4]We assume a power law distribution for the normalize eigenvalues λ_i/N because the eigenvalues λ_i of K scales in N.

Intuitively, the coherence measures the degree to which the eigenvectors in V are correlated with the canonical bases. According to the result from compressive sensing, highly coherent matrices are difficult (even impossible) to be recovered by matrix completion with random sampling.

The coherence measure was first introduced into the error analysis of the Nyström method by Talwalkar and Rostamizadeh [10]. Their analysis shows that a low rank kernel matrix with incoherent eigvenvectors (i.e., with low coherence) can be accurately approximated by the Nyström method using the uniform sampling. This result is generalized to noisy observation in [13] for low rank matrix. The main limitation of these results is that they only apply to low rank matrices. Recently, A. Gittens [14] developed a relative error bound of the Nyström method for kernel matrices with an arbitrary rank using a slightly different coherence measure. Similar to [10], [13], [14], coherence measure also plays an important role in our analysis. However, unlike the previous studies, we focus on the error bound of the Nyström method for kernel matrices with an arbitrary rank and a skewed eigenvalue distribution.

The main result of our analysis is given in the following theorem.

Theorem 3: Assume the eigenvalues $\lambda_i/N, i = 1, \ldots, N$ follow a p-power law with $p > 1$. Given a sufficiently large number of samples, i.e.,

$$m > \mu^2 \max\left(16\left(\frac{\ln N}{\gamma}\right)^2, 2C_{ab}\ln(3N^3), 4C_{ab}^2\ln^2(3N^3)\right)$$

we have, with a probability $1 - 2N^{-3}$,

$$\left\|K - K_b\widehat{K}^{\dagger}K_b^{\top}\right\|_2 \leq \widetilde{O}\left(\frac{N}{m^{p-1}}\right),$$

where $\widetilde{O}(\cdot)$ suppresses the polynomial factor that depends on $\ln N$, and C_{ab} is a numerical constant as revealed in our later analysis.

Remark 3: Compared to the approximation error in (1), Theorem 3 improves the bound from $O(N/\sqrt{m})$ to $O(N/m^{p-1})$ provided the eigenvalues of kernel matrix follow a power law. It is worth noting that similar to [10], [13], [14], the bound in Theorem 3 is meaningful only when the coherence μ of the eigenvector matrix is small (i.e., the eigenvector matrix satisfies the incoherence assumption). We also would like to compare the absolute error bound in Theorem 3 with the relative error bound in [14], which is dominated by $O(N^2/m^{p+1})$ when eigenvalues follow a p-power law. This bound is worse than the $O(N/m^{p-1})$ bound obtained in Theorem 3 when $m \leq \sqrt{N}$, a favorable setting when N is very large and m is small. Although we notice that the relative error bound in [14] holds more generally than the error bound in Theorem 3, the comparison may shed light on how a power law eigen-value distribution could yield an improved approximation error bound of the Nyström method, and could possibly inspire new error analysis of other random matrix approximation methods. In Remark 4, we make a comparison between a more general result in Theorem 7 and [14].

We emphasize that the result in Theorem 3 does not contradict the lower bound given in Theorem 1 because Theorem 3 holds only for the cases when eigenvalues of the kernel matrix follow a power law. In fact, an updated lower bound for kernel matrix with a power law eigenvalue distribution is given in the following theorem.

Theorem 4: There exists a kernel matrix $K \in \mathbb{R}^{N\times N}$ with all its diagonal entries being 1 and its eigenvalues following a p-power law such that for any sampling strategy that selects m columns, the approximation error of the Nyström method is lower bounded by $\Omega(\frac{N}{m^p})$, i.e.,

$$\left\|K - K_b\widehat{K}^{\dagger}K_b^{\top}\right\|_2 \geq \Omega\left(\frac{N}{m^p}\right)$$

provided $N > 64[\ln 4]^2 m^2$.

We skip the proof of this theorem as it is almost identical to that of Theorem 1. The gap between the upper bound and the lower bound given in Theorems 3 and 4 indicates that there is potentially a room for further improvement.

Next, we present several theorems and corollaries to pave the path for the proof of Theorem 3. We borrow the following two theorems used by the compressive sensing theory [26].

Theorem 5: (Theorem 1.2 from [26]) Let V be an $N \times N$ orthogonal matrix ($V^{\top}V = I$) with coherence μ. Fix a subset T of the signal domain. Choose a subset S of the measurement domain of size $|S| = m$ uniformly at random. Suppose that the number of measurements m obeys $m \geq |T|\mu^2 \max(C_a \ln|T|, C_b \ln(3/\delta))$ for some positive constants C_a and C_b. Then

$$\Pr\left(\left\|\frac{N}{m}V_{S,T}^{\top}V_{S,T} - I\right\|_2 \geq 1/2\right) \leq \delta.$$

Theorem 6: (Lemma 3.3 from [26]) Let V, S, and T be the same as defined in Theorem 5. Let $\mathbf{u}_k^{\top}$ be the k-th row of $V_{S,*}^{\top}V_{S,T}$. Define $\sigma^2 = \mu^2 m \max(1, \mu|T|/\sqrt{m})$. Fix $a > 0$ obeying $a \leq (m/\mu^2)^{1/4}$ if $\mu|T|/\sqrt{m} > 1$ and $a \leq (m/[\mu^2|T|])^{1/2}$ otherwise. Let $\mathbf{z}_k = (V_{S,T}^{\top}V_{S,T})^{-1}\mathbf{u}_k$. Then, we have

$$\Pr\left(\sup_{k\in T^c}\|\mathbf{z}_k\|_2 \geq 2\mu\sqrt{|T|/m} + 2a\sigma/m\right) \leq N\exp(-\gamma a^2) + \Pr\left(\|V_{S,T}^{\top}V_{S,T}\|_2 \leq \frac{m}{2N}\right)$$

for some positive constant γ, where T^c stands for the complementary set to T.

We note that Theorem 5 has been explored in [10] for showing that the Nyström approximation is exact when the target PSD matrix is of low rank. However, our analysis does not restrict to a low rank PSD matrix. Instead, by combining the results in Theorem 5 and Theorem 6, we have the following high probability bound for $\sup_{k\in T^c}\|\mathbf{z}_k\|_2$, which serves the key to prove a general result stated in Theorem 7 for a PSD matrix with an arbitrary rank.

Corollary 2: If $|T| \geq \max\left(C_{ab}\ln(3N^3), 4\frac{\ln N}{\gamma}\right)$, and

$$\mu^2 \max\left(|T|C_{ab}\ln(3N^3), 16\left(\frac{\ln N}{\gamma}\right)^2\right) \leq m < \mu^2|T|^2,$$

where $C_{ab} = \max(C_a, C_b)$, then with a probability $1-2N^{-3}$, we have

$$\sup_{k\in T^c}\|\mathbf{z}_k\|_2 \leq 4\mu\sqrt{\frac{|T|}{m}}.$$

Using Corollary 2, we have the following bound for $\mathcal{E}(\mathcal{H}_a, r)$.

Theorem 7: If $r > \max(C_{ab}\ln(3N^3), 4\ln N/\gamma)$ and

$$\mu^2 \max\left(rC_{ab}\ln(3N^3), 16\left(\frac{\ln N}{\gamma}\right)^2\right) \le m < \mu^2 r^2,$$

then, with a probability $1-2N^{-3}$, we have

$$\mathcal{E}(\mathcal{H}_a, r) \le \frac{16\mu^2 r}{m}\sum_{i=r+1}^{N}\lambda_i.$$

Remark 4: Before presenting the proof, it is worthwhile to compare the approximation error bound by combing the results in Theorem 7 and Proposition 2 to Gittens' relative error bound [14], i.e., with a probability at least $1-\delta$ it holds that $\left\|K - K_b\widehat{K}^{\dagger}K_b^{\top}\right\|_2 \le \lambda_{r+1}\left(1+\frac{2N}{m}\right)$ provided $m \ge 8\tau^2 r\log(r/\delta)$, where τ is a coherence measure of the top-r eigen-space of K defined by $\tau^2 = \frac{N}{r}\max_{1\le i\le N}\sum_{j=1}^{r}V_{ij}^2$. In the case when the eigenvalues decay fast (e.g., eigenvalues follow a power law), we have $\sum_{i=r+1}^{N}\lambda_i \ll N\lambda_{r+1}$, and therefore our bound could be significantly better than Gittens' relative error bound. On the other hand, when eigenvalues follow a flat distribution (e.g., $\lambda_i \approx \lambda_{r+1}$ for all $i \in [r+2, N]$), we have $\sum_{i=r+1}^{N}\lambda_i \approx N\lambda_{r+1}$, and therefore our bound is worse than Gittens' relative bound by only a factor of $\mu^2 r$.

Proof of Theorem 7: For the sake of simplicity, we assume that the first m examples are sampled, i.e., $\widehat{\mathcal{D}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$. For any $g \in \mathcal{H}_b^r$, we have $g(\cdot) = \sum_{i=1}^{r} w_i\lambda_i^{1/2}\varphi_i(\cdot)$, with $\sum_{i=1}^{r}w_i^2 \le 1$. Below, we will make specific construction of f based on g that ensures a small approximation error. Let f be

$$\begin{aligned} f(\cdot) &= \sum_{j=1}^{m} a_j\kappa(\mathbf{x}_j, \cdot) \\ &= \sum_{i=1}^{N}\varphi_i(\cdot)\lambda_i^{1/2}\left(\sum_{j=1}^{m}a_jV_{j,i}\right) \\ &= \sum_{i=1}^{N} b_i\lambda_i^{1/2}\varphi_i(\cdot), \end{aligned}$$

where $b_i = \sum_{j=1}^{m}a_jV_{j,i}, i = 1, \ldots, N$, and the value of $\mathbf{a} = (a_1, \ldots, a_m)^{\top}$ will be given later. Define $T = \{1, \ldots, r\}$ and $S = \{1, \ldots, m\}$. Under the condition that

$$\begin{aligned} m &\ge r\mu^2\max\left(C_a, C_b\right))\ln(3N^3) \\ &\ge r\mu^2\max\left(C_a\ln r, C_b\ln(3N^3)\right), \end{aligned}$$

Theorem 5 holds, and therefore with a probability at least $1-N^{-3}$,

$$\frac{m}{2N} \le \lambda_{\min}\left(V_{S,T}^{\top}V_{S,T}\right) \le \lambda_{\max}\left(V_{S,T}^{\top}V_{S,T}\right) \le \frac{3m}{2N}. \quad (7)$$

We construct $\mathbf{a}$ as $\mathbf{a} = V_{S,T}\left[V_{S,T}^{\top}V_{S,T}\right]^{-1}\mathbf{w}$, where $\mathbf{w} = (w_1, \ldots, w_r)^{\top}$. Since

$$\mathbf{b} = V_{S,*}^{\top}\mathbf{a} = V_{S,*}^{\top}V_{S,T}\left(V_{S,T}^{\top}V_{S,T}\right)^{-1}\mathbf{w},$$

where $\mathbf{b} = (b_1, \cdots, b_N)^{\top}$, it is straightforward to see that $b_j = w_j$ for $j \in T$. Using the result from Corollary 2, we have, with a probability at least $1-2N^{-3}$,

$$\max_{j\in T^c}|b_j| \le \max_{j\in T^c}\|\mathbf{z}_j\|_2\|\mathbf{w}\|_2 \le 4\mu\sqrt{\frac{r}{m}},$$

where $\mathbf{z}_j^{\top}$ is the j-th row of matrix $V_{S,*}^{\top}V_{S,T}\left(V_{S,T}^{\top}V_{S,T}\right)^{-1}$. We thus obtain

$$\|f-g\|_{\mathcal{H}_\kappa}^2 = \left\|\sum_{i\in T^c}\lambda_i^{1/2}b_i\varphi_i(\cdot)\right\|_{\mathcal{H}_\kappa}^2 \le \frac{16\mu^2 r}{m}\sum_{i=r+1}^{N}\lambda_i.$$

Hence,

$$\mathcal{E}(\mathcal{H}_a, r) = \max_{g\in\mathcal{H}_b^r}\min_{f\in\mathcal{H}_a}\|f-g\|_{\mathcal{H}_\kappa}^2 \le \frac{16\mu^2 r}{m}\sum_{i=r+1}^{N}\lambda_i.$$

■

Finally, we show the proof of Theorem 3 using Theorem 7.

Proof of Theorem 3: Let $r = \left\lfloor \frac{m}{\mu^2 C_{ab}\ln(3N^3)} \right\rfloor$, then

$$\mu^2 rC_{ab}\ln(3N^3) \le m < \mu^2 r^2,$$

where the right inequality follows that $r \ge \frac{m}{2\mu^2 C_{ab}\ln(3N^3)}$, and $m > 4\mu^2 C_{ab}^2\ln^2(3N^3)$. Then the conditions in Theorem 7 hold and we have

$$\begin{aligned} \left\|K - K_b\widehat{K}^{\dagger}K_b^{\top}\right\|_2 &\le \max\left(\mathcal{E}(\mathcal{H}_a, r), \lambda_{r+1}\right) \\ &\le \max\left(\frac{16\mu^2 r}{m}, 1\right)\sum_{i=r+1}^{N}\lambda_i. \end{aligned}$$

Since $\max(16\mu^2 r/m, 1) \le O(1)$ due to the specific value we choose for r, and $\sum_{i=r+1}^{N}\lambda_i \le O(N/r^{p-1})$ due to the power law distribution, then

$$\left\|K - K_b\widehat{K}^{\dagger}K_b^{\top}\right\|_2 \le O\left(\frac{N}{r^{p-1}}\right) \le \widetilde{O}\left(\frac{N}{m^{p-1}}\right).$$

■

IV. Application of the Nyström Method to Kernel Classification

Although the Nyström method was proposed in [1] to speed up kernel machine, few studies examine the application of the Nyström method to kernel classification. In fact, to the best of our knowledge, [1] and [11] are the only two pieces of work that explicitly explore the Nyström method for kernel classification. The key idea of both works is to apply the Nyström method to approximate the kernel matrix with a low rank matrix in order to reduce the computational cost. More specifically, we consider the following optimization problem for kernel classification

$$\min_{f\in\mathcal{H}_\kappa}\mathcal{L}_N(f) = \frac{\lambda}{2}\|f\|_{\mathcal{H}_\kappa}^2 + \frac{1}{N}\sum_{i=1}^{N}\ell(y_i f(\mathbf{x}_i)), \quad (8)$$

where $y_i \in \{-1, +1\}$ is the class label assigned to instance $\mathbf{x}_i$, and $\ell(z)$ is a convex loss function. To facilitate our analysis, we assume (i) $\ell(z)$ is strongly convex with modulus σ, i.e. $|\ell''(z)| \ge \sigma$ [5], and (ii) $\ell(z)$ is Lipschitz continuous, i.e.

[5]Loss functions such as square loss used for regression and logit function used for logistic regression are strongly convex.

$|\ell'(z)| \leq C$ for any z within the domain. Using the convex conjugate of the loss function $\ell(z)$, denoted by $\ell_*(\alpha), \alpha \in \Omega$, where Ω is the domain for dual variable α, we can cast the problem in (8) into the following optimization problem over α

$$\max_{\{\alpha_i \in \Omega\}_{i=1}^N} -\frac{1}{N}\sum_{i=1}^N \ell_*(\alpha_i) - \frac{1}{2\lambda N^2}(\alpha \circ \mathbf{y})^\top K(\alpha \circ \mathbf{y}), \quad (9)$$

with the solution f given by $f = -\frac{1}{N\lambda}\sum_{i=1}^N \alpha_i y_i \kappa(\mathbf{x}_i, \cdot)$. By the Fenchel conjugate theory, we have $\max_{\alpha \in \Omega} |\alpha|^2 \leq C^2$. because $|\ell'(z)| \leq C$.

To reduce the computational cost, [1] and [11] suggest to replace the kernel matrix K with its low rank approximation $\widetilde{K} = K_b \widehat{K}^\dagger K_b^\top$, leading to the following optimization problem for α

$$\max_{\{\alpha_i \in \Omega\}_{i=1}^N} -\frac{1}{N}\sum_{i=1}^N \ell_*(\alpha_i) - \frac{1}{2\lambda N^2}(\alpha \circ \mathbf{y})^\top \widetilde{K}(\alpha \circ \mathbf{y}). \quad (10)$$

One main problem with this approach is that although it simplifies the computation of kernel matrix, it does not simplify the classifier f, because the number of support vectors, after the application of the Nyström method, is not guaranteed to be small [30], [31], leading to a high computational cost in performing prediction on testing examples. We address this difficulty by presenting a new approach to explore the Nyström method for kernel classification. We show that, by an analysis of excessive risk, the number of support vectors for achieving a similar performance as the kernel classifier learned with the full kernel matrix is bounded by $N^{2p/(p^2-1)}$, where p characterizes the power law of the eigen-value distribution of the kernel matrix.

Similar to the previous analysis, we randomly select a subset of training examples, denoted by $\widehat{\mathcal{D}} = (\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_m)$, and restrict the solution of $f(\cdot)$ to the subspace $\mathcal{H}_a = \text{span}(\kappa(\widehat{\mathbf{x}}_1, \cdot), \ldots, \kappa(\widehat{\mathbf{x}}_m, \cdot))$, leading to the following optimization problem

$$\min_{f \in \mathcal{H}_a} \mathcal{L}_N(f) = \frac{\lambda}{2}\|f\|_{\mathcal{H}_\kappa}^2 + \frac{1}{N}\sum_{i=1}^N \ell(y_i f(\mathbf{x}_i)). \quad (11)$$

The following proposition shows that the optimal solution to (11) is closely related to the optimal solution to (10).

Proposition 4: The solution f to (11) is given by

$$f = -\frac{1}{N\lambda}\sum_{i=1}^m z_i y_i \kappa(\widehat{\mathbf{x}}_i, \cdot),$$

where $\mathbf{z} = \widehat{K}^\dagger K_b^\top \alpha$ and α is the optimal solution to (10).

It is important to note that the classifier obtained from (11) is only supported by the sampled training examples in $\widehat{\mathcal{D}}$, which significantly reduces the complexity of the kernel classifier compared to the approach suggested in [1], [11]. We also note that the proposed approach is equivalent to learning a linear classifier by representing each instance $\mathbf{x}$ with the vector

$$\phi(\mathbf{x}) = \widehat{D}^{-1/2}\widehat{V}^\top \left(\kappa(\widehat{\mathbf{x}}_1, \mathbf{x}), \ldots, \kappa(\widehat{\mathbf{x}}_m, \mathbf{x})\right)^\top,$$

where $\widehat{D}$ is a diagonal matrix with non-zero eigenvalues of $\widehat{K}$, and $\widehat{V}$ is the corresponding eigenvector matrix. Although this idea has already been adopted by practitioners, we are unable to find any reference on its empirical study. The remaining of this work is to show that this approach could have a good generalization performance provided that the eigenvalues of kernel matrix follow a skewed distribution. Below, we develop the generalization error bound for the classifier learned from (11).

Let f_N and f_N^a be the optimal solutions to (8) and (11), respectively. Let f^* be the optimal classifier that minimizes the expected loss function, i.e.,

$$f^* = \arg\min_{f \in \mathcal{H}_\kappa} P(\ell \circ f) \triangleq \mathrm{E}_{(\mathbf{x},y)}\left[\ell(yf(\mathbf{x}))\right].$$

Let $\|f\|_{L_2}^2 = \mathrm{E}_{\mathbf{x}}[|f(\mathbf{x})|^2]$ denote the ℓ_2 norm square of f. In order to create a tight bound, we exploit the technique of local Rademacher complexity [32], [33]. Define $\psi(\cdot)$ as

$$\psi(\delta) = \left(\frac{2}{N}\sum_{i=1}^N \min(\delta^2, \lambda_i)\right)^{1/2}.$$

Let $\widetilde{\varepsilon}$ be the solution to $\widetilde{\varepsilon}^2 = \psi(\widetilde{\varepsilon})$ where the existence and uniqueness of $\widetilde{\varepsilon}$ is determined by the sub-root property of $\psi(\delta)$ [32]. Finally we define

$$\epsilon = \max\left(\widetilde{\varepsilon}, \sqrt{\frac{6\ln N}{N}}\right). \quad (12)$$

Theorem 8: Assume with a probability $1 - 2N^{-3}$, $\mathcal{E}(\mathcal{H}_a) \leq \Gamma(N, m)$, where $\Gamma(N, m)$ is some function depending on N and m. Assume that N is sufficiently large such that

$$\max\left(\|f_N^a\|_{\mathcal{H}_\kappa}, \|f^*\|_{\mathcal{H}_\kappa}\right) \leq \frac{e^N N}{12\ln N},$$
$$\max\left(\|f_N^a\|_{L_2}, \|f^*\|_{L_2}\right) \leq \frac{e^N}{2}\sqrt{\frac{N}{6\ln N}}.$$

Then, with a probability at least $1 - 4N^{-3}$, we have

$$P(\ell \circ f_N^a) \leq P(\ell \circ f^*) + 2\lambda\|f^*\|_{\mathcal{H}_\kappa}^2 + \frac{C^2\Gamma(N,m)}{\lambda N} + \frac{2C_1^2C^2\epsilon^4}{\lambda} + \frac{2C_1^2C^2\epsilon^2}{\sigma} + C_1 C e^{-N}$$

where ϵ is given in (12) and C_1 is a constant independent from m and N. By choosing λ that minimizes the above bound, we have

$$P(\ell \circ f_N^a) \leq P(\ell \circ f^*) + 4\|f^*\|_{\mathcal{H}_\kappa}\epsilon^2 C\sqrt{C_1^2 + \frac{\Gamma(N,m)}{2N\epsilon^4}} + \frac{2C_1^2C^2}{\sigma}\epsilon^2 + C_1 C e^{-N}.$$

Remark 5: In the case when the eigenvalues of the kernel matrix follow a p-power law with $p > 1$, we have $\epsilon^2 = O(N^{-p/(p+1)})$ according to [28], and $\Gamma(N, m) = O(N/m^{p-1})$ according to Theorem 3. Applying these results to Theorem 8, the generalization performance of f_N^a becomes

$$P(\ell \circ f_N^a) \leq P(\ell \circ f^*) + 2\lambda\|f^*\|_{\mathcal{H}_\kappa}^2 + \frac{C_2C^2}{\lambda m^{p-1}} + C_1 C e^{-N} + \frac{2C_3C^2N^{-2p/(p+1)}}{\lambda} + \frac{2C_4C^2N^{-p/(p+1)}}{\sigma} \quad (13)$$

where C_2, C_3, and C_4 are constants independent from N and m. By choosing λ that minimizes the bound in (13), we have

$$P(\ell \circ f_N^a) \leq P(\ell \circ f^*) + \frac{4\|f^*\|_{\mathcal{H}_\kappa}}{N^{p/(p+1)}} C \sqrt{C_3 + C_2 \frac{N^{2p/(p+1)}}{2m^{p-1}}} + \frac{2C_4 C^2}{\sigma N^{p/(p+1)}} + C_1 C e^{-N} = P(\ell \circ f^*) + O\left(N^{-p/(p+1)} + m^{-(p-1)/2}\right).$$

As indicated by above inequality, when the eigenvalues of the kernel matrix follow a p-power law, by setting $m = N^{2p/(p^2-1)}$, we are able to achieve a similar performance as the full version of kernel classifier (i.e., $O(N^{-p/(p+1)})$). In other words, we can construct a kernel classifier without sacrificing its generalization performance with no more than $N^{2p/(p^2-1)}$ support vectors, which could be significantly smaller than N when $p > (1+\sqrt{2})$. For the example of kernel that generates Sobolev Spaces $W^{\alpha,2}(\mathbb{T}^d)$ of smoothness $\alpha > d/2$, where $\mathbb{T}^d$ is d-dimensional torus, its eigenvalues follow a p-power law with $p = 2\alpha > d$, which is larger than $(1+\sqrt{2})$ when $d \geq 3$.

V. Conclusion

We develop new methods for analyzing the approximation bound for the Nyström method. We show that the approximation error can be improved to $O(N/m^{1-\rho})$ in the case when there is a large eigengap in the spectrum of a kernel matrix, where $\rho \in (0, 1/2)$ is introduced to characterize the eigengap. When the eigenvalues of a kernel matrix follow a p-power law, the approximation error is further reduced to $O(N/m^{p-1})$ under an incoherence assumption. We develop a kernel classification approach based on the Nyström method and show that when the eigenvalues of a kernel matrix follow a p-power law ($p > 1$), we can reduce the number of support vectors to $N^{2p/(p^2-1)}$, which could be significantly less than N if p is large, without seriously sacrificing its generalization performance.

Acknowledgment

Rong Jin and Mehrdad Mahdavi are supported in part by Office of Navy Research (ONR award N00014-09-1-0663 and N00014-12-10431). Yu-Feng Li and Zhi-Hua Zhou are partially supported by the National Fundamental Research Program of China (2010CB327903), the National Science Foundation of China (61073097, 61021062).

References

[1] C. Williams and M. Seeger, "Using the nystrom method to speed up kernel machines," in *Advances in Neural Information Processing Systems 13*. MIT Press, 2001, pp. 682–688.

[2] P. Drineas and M. W. Mahoney, "On the nystrom method for approximating a gram matrix for improved kernel-based learning," *Journal of Machine Learning Research*, vol. 6, p. 2005, 2005.

[3] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using the nystrom method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, p. 2004, 2004.

[4] S. Kumar, M. Mohri, and A. Talwalkar, "Sampling techniques for the nystrom method," in *Proceedings of Conference on Artificial Intelligence and Statistics*, 2009, pp. 304 – 311.

[5] V. D. Silva and J. B. Tenenbaum, "Global versus local methods in nonlinear dimensionality reduction," in *Advances in Neural Information Processing Systems 15*, 2003, pp. 705–712.

[6] J. C. Platt, "Fast embedding of sparse music similarity graphs," in *Advances in Neural Information Processing Systems 16*. MIT Press, 2004, p. 2004.

[7] A. Talwalkar, S. Kumar, and H. A. Rowley, "Large-scale manifold learning," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2008.

[8] K. Zhang, I. W. Tsang, and J. T. Kwok, "Improved nystrom low-rank approximation and error analysis," in *Proceedings of International Conference on Machine Learning*, 2008.

[9] M.-A. Belabbas and P. J. Wolfe, "Spectral methods in machine learning and new strategies for very large data sets," *Proceedings of the National Academy of Sciences of the USA*, vol. 106, pp. 369–374, 2009.

[10] A. Talwalkar and A. Rostamizadeh, "Matrix coherence and the nystrom method," in *Proceedings of Conference on Uncertainty in Artificial Intelligence*, 2010.

[11] C. Cortes, M. Mohri, and A. Talwalkar, "On the impact of kernel approximation on learning accuracy," *Journal of Machine Learning Research - Proceedings Track*, vol. 9, pp. 113–120, 2010.

[12] M. Li, J. T. Kwok, and B.-L. Lu, "Making large-scale nyström approximation possible," in *Proceedings of the 27th international conference on Machine learning*, 2010, pp. 631–638.

[13] L. W. Mackey, A. S. Talwalkar, and M. I. Jordan, "Divide-and-conquer matrix factorization," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Eds., 2011, pp. 1134–1142.

[14] A. Gittens, "The spectral norm error of the naive nystrom extension," *CoRR*, 2011.

[15] M. W. Mahoney, "Randomized algorithms for matrices and data," *Foundations and Trends in Machine Learning*, vol. 3, no. 2, pp. 123–224, 2011.

[16] S. Kumar, M. Mohri, and A. Talwalkar, "Sampling methods for the nystrom method," *J. Mach. Learn. Res.*, vol. 98888, pp. 981–1006, 2012.

[17] J. Chen and Y. Saad, "Lanczos vectors versus singular vectors for effective dimension reduction," *IEEE Trans. on Knowl. and Data Eng.*, vol. 21, pp. 1091–1103, 2009.

[18] L. Wu, X. Ying, X. Wu, and Z.-H. Zhou, "Line orthogonality in adjacency eigenspace with application to community partition," in *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence*, 2011, pp. 2349–2354.

[19] M. Ji, T. Yang, B. Lin, R. Jin, and J. Han, "A simple algorithm for semi-supervised learning with improved generalization error bound," in *Proceedings of the 29th international conference on Machine learning*, 2012, pp. –.

[20] T. Yang, Y.-F. Li, M. Mahdavi, R. Jin, and Z.-H. Zhou, "Nystrom method vs random fourier features: A theoretical and empirical comparison," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 485–493.

[21] A. Gittens and M. W. Mahoney, "Revisiting the nystrom method for improved large-scale machine learning," in *Proceedings of the 30th international conference on Machine learning*, 2013, pp. –.

[22] F. Bach, "Sharp analysis of low-rank kernel matrix approximations," in *Proceedings of the 26th Conference on Learning Theory*, 2013, pp. –.

[23] J. B. Hough, M. Krishnapur, Y. Peres, and B. Virag, "Determinantal processes and independence," *Probability Surveys*, vol. 3, pp. 206–229, 2006.

[24] P. Drineas, M. W. Mahoney, and S. Muthukrishnan, "Relative-error cur matrix decompositions," *SIAM J. Matrix Anal. Appl.*, vol. 30, pp. 844–881, 2008.

[25] S. Smale and D.-X. Zhou, "Geometry on probability spaces," *Constr Approx*, vol. 30, pp. 311–323, 2009.

[26] E. Candés and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inverse Problems*, vol. 23, no. 3, pp. 969–985, 2007.

[27] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[28] V. Koltchinskii and M. Yuan, "Sparsity in multiple kernel learning," *Annuals of Statistics*, vol. 38, pp. 3660–3694, 2010.

[29] M. Kloft and G. Blanchard, "The local rademacher complexity of lp-norm multiple kernel learning," in *Advances in Neural Information Processing Systems 23*, 2011, pp. 2438–2446.

[30] O. Dekel and Y. Singer, "Support vector machines on a budget," in *NIPS*, 2006, pp. 345–352.

[31] T. Joachims and C.-N. J. Yu, "Sparse kernel svms via cutting-plane training," *Mach. Learn.*, vol. 76, pp. 179–193, 2009.

[32] P. L. Bartlett, O. Bousquet, and S. Mendelson, "Local rademacher complexities," *Annals of Statistics*, pp. 44–58, 2002.
[33] V. Koltchinskii, *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer, 2011.

APPENDIX A
PROOF OF THEOREM 1

We argue that there exists a kernel matrix K such that (i) all its diagonal entries equal to 1, and (ii) the first $m+1$ eigenvalues of K are in the order of $\Omega(N/m)$. To see the existence of such a matrix, we sample $m+1$ vectors $\mathbf{u}_1, \cdots, \mathbf{u}_{m+1}$, where $\mathbf{u}_i \in \mathbb{R}^N$, from a Bernoulli distribution, with $\Pr(u_{i,j} = +1) = \Pr(u_{i,j} = -1) = 1/2$. We then construct K as

$$K = \sum_{i=1}^{m+1} \mathbf{u}_i \mathbf{u}_i^\top \frac{1}{m+1} = \frac{1}{m+1} UU^\top, \tag{14}$$

where $U = (\mathbf{u}_1, \cdots, \mathbf{u}_{m+1})$.

First, since $u_{i,j} = \pm 1$, we have $diag(\mathbf{u}_i \mathbf{u}_i^\top) = \mathbf{1}$, where $\mathbf{1}$ is a vector of all ones, and therefore $K_{i,i} = 1$ for $i \in [N]$. Second, we show that with some probability $1-\delta$, all non-zero eigenvalues of $\frac{1}{N} U^\top U$ are bounded between $1/2$ and $3/2$, i.e.,

$$\frac{1}{2} \leq \lambda_{\min}\left(\frac{1}{N} U^\top U\right) \leq \lambda_{\max}\left(\frac{1}{N} U^\top U\right) \leq \frac{3}{2}. \tag{15}$$

To prove (15), we use the concentration inequality in Proposition 3. We define $\xi_i = \mathbf{z}_i \mathbf{z}_i^\top, i = 1, \ldots, N$, where $\mathbf{z}_i \in \mathbb{R}^m$ is the ith row of the matrix U, and $\|\cdot\|$ in the above proposition as the spectral norm of a matrix. Since every element in $\mathbf{z}_i$ is sampled from a Bernoulli distribution with equal probabilities of being ± 1, we have $\mathrm{E}[\mathbf{z}_i \mathbf{z}_i^\top] = I_m$ and $\|\mathbf{z}_i \mathbf{z}_i^\top\| = m$. Thus, with a probability $1-\delta$, we have

$$\left\| \frac{1}{N} U^\top U - I \right\| = \left\| \frac{1}{N} \sum_{i=1}^N \xi_i - \mathrm{E}[\xi] \right\| \leq \frac{4m \ln(2/\delta)}{\sqrt{N}}.$$

When $N > 64m^2[\ln 4]^2$, for any sampled U, with 50% chance, we have

$$\left\| \frac{1}{N} U^\top U - I \right\| \leq \frac{1}{2},$$

which implies (15).

With the bound in (15) and using the fact that the eigenvalues of $UU^\top$ equal to the eigenvalues of $U^\top U$, it is straightforward to see that the first $m+1$ eigenvalues of K are in the order of $\Omega(N/m)$. Up to this point, we proved the existence of such a kernel matrix. Next, we prove the lower bound for the constructed kernel matrix.

Let $V_{1:(m+1)} = (\mathbf{v}_1, \cdots, \mathbf{v}_{m+1})$ the first $m+1$ eigenvectors of K. We construct $\widehat{g}$ as follows: Let $\mathbf{u} = V_{1:(m+1)}\mathbf{a}$ be a vector in the subspace $\mathrm{span}(\mathbf{v}_1, \cdots, \mathbf{v}_{m+1})$ that satisfies the condition $K_b^\top \mathbf{u} = 0$. The existence of such a vector is guaranteed because $\mathrm{rank}(K_b^\top V_{1:(m+1)}) \leq m$. We normalize $\mathbf{a}$ such that $\|\mathbf{a}\|_2 = 1$. Then we let $\widehat{g} = \sum_{i=1}^N u_i \kappa(\mathbf{x}_i, \cdot) = \sum_{i=1}^{m+1} w_i \sqrt{\lambda_i} \varphi_i(\cdot)$, where $\mathbf{w} = V_{1:(m+1)}^\top \mathbf{u}$. It is easy to verify that (i) $\widehat{g} \in \mathcal{H}_b$ since $\|\mathbf{u}\|_2 = \|V_{1:(m+1)}\mathbf{a}\|_2 = 1$, and (ii) $\widehat{g} \perp \mathcal{H}_a$ since $\mathbf{u}^\top K_b = 0$. Using $\widehat{g}$, we have

$$\begin{aligned}\mathcal{E}(\mathcal{H}_a) &= \max_{g \in \mathcal{H}_b} \min_{f \in \mathcal{H}_a} \|f - g\|_{\mathcal{H}_\kappa}^2 \geq \|\widehat{g}\|_{\mathcal{H}_\kappa}^2 = \sum_{i=1}^{m+1} w_i^2 \lambda_i \\ &= \Omega\left(\frac{N}{m+1}\right) \|\mathbf{w}\|_2^2 \geq \Omega\left(\frac{N}{m}\right),\end{aligned}$$

where we use $\|\mathbf{w}\|_2 = \|V_{1:(m+1)}^\top V_{1:(m+1)}\mathbf{a}\|_2 = \|\mathbf{a}\|_2 = 1$. We complete the proof by using the fact $\mathcal{E}(\mathcal{H}_a) = \left\|K - K_b \widehat{K}^\dagger K_b^\top\right\|_2$.

APPENDIX B
PROOF OF COROLLARY 1

Define $\xi(\widehat{\mathbf{x}}_i)$ to be a rank one linear operator, i.e.,

$$\xi(\widehat{\mathbf{x}}_i)[f](\cdot) = \kappa(\widehat{\mathbf{x}}_i, \cdot) f(\widehat{\mathbf{x}}_i).$$

Apparently, $L_m = \frac{1}{m}\sum_{i=1}^m \xi(\widehat{\mathbf{x}}_i)$ and $\mathrm{E}[\xi(\widehat{\mathbf{x}}_i)] = L_N$. We complete the proof by using the result from Proposition 3 and the fact

$$\begin{aligned}\|\xi(\widehat{\mathbf{x}}_k)\|_{HS} &= \sqrt{\sum_{i,j=1}^N \langle \varphi_i, \kappa(\widehat{\mathbf{x}}_k, \cdot)\varphi_j(\widehat{\mathbf{x}}_k)\rangle^2} \\ &= \sqrt{\sum_{i,j=1}^N \varphi_i(\widehat{\mathbf{x}}_k)^2 \varphi_j(\widehat{\mathbf{x}}_k)^2} = \kappa(\widehat{\mathbf{x}}_k, \widehat{\mathbf{x}}_k) \leq 1,\end{aligned}$$

where the last equality follows equation (3).

APPENDIX C
PROOF OF COROLLARY 2

We choose $a = 2\sqrt{\ln N/\gamma}$ in Theorem 6. Since $m \geq 16\mu^2 \left(\frac{\ln N}{\gamma}\right)^2$, then we have $a \leq \left(\frac{m}{\mu^2}\right)^{1/4}$. Additionally, by having $\mu|T|/\sqrt{m} > 1$, the conditions in Theorem 6 hold, and by setting $\delta = N^{-3}$ in Theorem 5, the condition in Theorem 5 holds, which together implies

$$\begin{aligned}&\Pr\left(\sup_{k \in T^c} \|\mathbf{z}_k\|_2 \geq 2\mu\sqrt{|T|/m} + 2a\sigma/m\right) \\ &\leq N\exp(-\gamma a^2) + \Pr\left(\|V_{S,T}^\top V_{S,T}\|_2 \leq \frac{m}{2N}\right) \\ &\leq N^{-3} + \Pr\left(\left\|\frac{N}{m} V_{S,T}^\top V_{S,T} - I\right\|_2 \geq \frac{1}{2}\right) \\ &\leq 2N^{-3}.\end{aligned}$$

From this we have, with a probability $1 - 2N^{-3}$,

$$\begin{aligned}\sup_{k \in T^c} \|\mathbf{z}_k\|_2 &\leq 2\mu\sqrt{\frac{|T|}{m}} + 2\left(\frac{m}{\mu^2}\right)^{1/4} \frac{\sqrt{\mu^3|T|m^{1/2}}}{m} \\ &= 4\mu\sqrt{\frac{|T|}{m}}.\end{aligned}$$

APPENDIX D
PROOF OF PROPOSITION 4

Since

$$\ell(y_i f(\mathbf{x}_i)) = \max_{\alpha_i \in \Omega} \alpha_i y_i f(\mathbf{x}_i) - \ell_*(\alpha_i),$$

we rewrite the optimization problem in (11) into a convex-concave optimization problem

$$\min_{f\in\mathcal{H}_a} \max_{\{\alpha_i\in\Omega\}_{i=1}^m} \frac{\lambda}{2}\|f\|_{\mathcal{H}_\kappa}^2 + \frac{1}{N}\sum_{i=1}^N \left(\alpha_i y_i f(\mathbf{x}_i) - \ell_*(\alpha_i)\right).$$

Since $f \in \mathcal{H}_a$, we write $f = \sum_{i=1}^m z_i \kappa(\widehat{\mathbf{x}}_i, \cdot)$, resulting in the following optimization problem

$$\min_{\mathbf{z}\in\mathbb{R}^m} \max_{\{\alpha_i\in\Omega\}_{i=1}^m} \frac{\lambda}{2}\mathbf{z}^\top \widehat{K}\mathbf{z} + \frac{1}{N}(\alpha\circ\mathbf{y})^\top K_b \mathbf{z} - \frac{1}{N}\sum_{i=1}^N \ell_*(\alpha_i).$$

Since the above problem in linear (convex) in $\mathbf{z}$ and concave in α, we can switch minimization with maximization. We complete the proof by taking the minimization over $\mathbf{z}$.

APPENDIX E
PROOF OF THEOREM 8

To simply our presentation, we introduce notations

$$P_N(\ell\circ f) = \frac{1}{N}\sum_{i=1}^N \ell(y_i f(\mathbf{x}_i)),$$
$$\Lambda(f) = P(\ell\circ f) - P(\ell\circ f^*).$$

Using $P_N(\ell\circ f)$, we can write $\mathcal{L}_N(f) = P_N(\ell\circ f) + \frac{\lambda}{2}\|f\|_{\mathcal{H}_\kappa}^2$. We first prove that

$$\mathcal{L}_N(f_N) \le \mathcal{L}_N(f_N^a) + \frac{C^2}{2\lambda N}\mathcal{E}(\mathcal{H}_a),$$

where $\max_{z\in\Omega}|z|^2 \le C^2$. Note that

$$\begin{aligned}
&\mathcal{L}_N(f_N) \\
&= \max_{\{\alpha_i\in\Omega\}_{i=1}^N} -\frac{1}{N}\sum_{i=1}^N \ell_*(\alpha_i) - \frac{1}{2\lambda N^2}(\alpha\circ\mathbf{y})^\top K(\alpha\circ\mathbf{y}) \\
&\mathcal{L}_N(f_N^a) \\
&= \max_{\{\alpha_i\in\Omega\}_{i=1}^N} -\frac{1}{N}\sum_{i=1}^N \ell_*(\alpha_i) - \frac{1}{2\lambda N^2}(\alpha\circ\mathbf{y})^\top \widetilde{K}(\alpha\circ\mathbf{y}).
\end{aligned}$$

Then

$$\begin{aligned}
&\mathcal{L}_N(f_N) \\
&= \max_{\{\alpha_i\in\Omega\}_{i=1}^N} -\frac{1}{N}\sum_{i=1}^N \ell_*(\alpha_i) - \frac{1}{2\lambda N^2}(\alpha\circ\mathbf{y})^\top \widetilde{K}(\alpha\circ\mathbf{y}) \\
&\quad + \frac{1}{2\lambda N^2}(\alpha\circ\mathbf{y})^\top(\widetilde{K}-K)(\alpha\circ\mathbf{y}) \\
&\le \max_{\{\alpha_i\in\Omega\}_{i=1}^N} -\frac{1}{N}\sum_{i=1}^N \ell_*(\alpha_i) - \frac{1}{2\lambda N^2}(\alpha\circ\mathbf{y})^\top \widetilde{K}(\alpha\circ\mathbf{y}) \\
&\quad + \max_{\{\alpha_i\in\Omega\}_{i=1}^N} \frac{1}{2\lambda N^2}(\alpha\circ\mathbf{y})^\top(\widetilde{K}-K)(\alpha\circ\mathbf{y}) \\
&\le \mathcal{L}_N(f_N^a) + \frac{1}{2\lambda N^2}\|\alpha\|_2^2\|K-\widetilde{K}\|_2 \\
&\le \mathcal{L}_N(f_N^a) + \frac{C^2}{2\lambda N}\mathcal{E}(\mathcal{H}_a).
\end{aligned}$$

Then we proceed the proof as follows

$$\begin{aligned}
&\frac{\lambda}{2}\|f_N^a\|_{\mathcal{H}_\kappa}^2 + P(\ell\circ f_N^a) \\
&\le P_N(\ell\circ f_N^a) + \frac{\lambda}{2}\|f_N^a\|_{\mathcal{H}_\kappa}^2 + (P-P_N)(\ell\circ f_N^a) \\
&\le P_N(\ell\circ f_N) + \frac{\lambda}{2}\|f_N\|_{\mathcal{H}_\kappa}^2 + \frac{C^2}{2\lambda N}\mathcal{E}(\mathcal{H}_a) \\
&\quad + (P-P_N)(\ell\circ f_N^a) \\
&\le P_N(\ell\circ f^*) + \frac{\lambda}{2}\|f^*\|_{\mathcal{H}_\kappa}^2 + \frac{C^2}{2\lambda N}\mathcal{E}(\mathcal{H}_a) \\
&\quad + (P-P_N)(\ell\circ f_N^a),
\end{aligned}$$

where the third inequality follows from the fact that f_N is the minimizer of $P_N(\ell\circ f) + \frac{\lambda}{2}\|f\|_{\mathcal{H}_\kappa}^2$. Hence,

$$\begin{aligned}
\Lambda(f_N^a) \le &\frac{\lambda}{2}\|f^*\|_{\mathcal{H}_\kappa}^2 - \frac{\lambda}{2}\|f_N^a\|_{\mathcal{H}_\kappa}^2 + \frac{C^2}{2\lambda N}\mathcal{E}(\mathcal{H}_a) \\
&+ (P-P_N)(\ell\circ f_N^a - \ell\circ f^*).
\end{aligned}$$

Let $r = \|f^* - f_N^a\|_{L_2}$ and $R = \|f^* - f_N^a\|_{\mathcal{H}_\kappa}$. Define

$$\mathcal{G}(r,R) = \left\{f\in\mathcal{H}_\kappa : \|f-f^*\|_{L_2} \le r, \|f^*-f\|_{\mathcal{H}_\kappa} \le R\right\}.$$

Using the domain $\mathcal{G}$, we rewrite the bound for $\Lambda(f_N^a)$ by

$$\begin{aligned}
\Lambda(f_N^a) \le &\frac{\lambda}{2}\|f^*\|_{\mathcal{H}_\kappa}^2 - \frac{\lambda}{2}\|f_N^a\|_{\mathcal{H}_\kappa}^2 + \frac{C^2}{2\lambda N}\mathcal{E}(\mathcal{H}_a) \\
&+ \sup_{f\in\mathcal{G}(r,R)} (P-P_N)(\ell\circ f - \ell\circ f^*).
\end{aligned}$$

Since $\epsilon r \le e^N$ and $\epsilon^2 R \le e^N$, using Lemma 9 from [28], we have, with a probability $1-2N^{-3}$, for any

$$\sup_{f\in\mathcal{G}(r,R)} (P-P_N)(\ell\circ f - \ell\circ f^*)) \le C_1 C(r\epsilon + R\epsilon^2 + e^{-N}),$$

where C_1 is a constant independent from N. Thus, with a probability at least $1-4N^{-3}$, we have

$$\begin{aligned}
&\Lambda(f_N^a) - C_1 C e^{-N} \\
&\le \frac{\lambda}{2}\|f^*\|_{\mathcal{H}_\kappa}^2 - \frac{\lambda}{2}\|f_N^a\|_{\mathcal{H}_\kappa}^2 + \frac{C^2\Gamma(N,m)}{2\lambda N} \\
&\quad + C_1 C\epsilon\|f_N^a - f^*\|_{L_2} + C_1 C\epsilon^2\|f^* - f_N^a\|_{\mathcal{H}_\kappa} \\
&\le \frac{\lambda}{2}\|f^*\|_{\mathcal{H}_\kappa}^2 - \frac{\lambda}{2}\|f_N^a\|_{\mathcal{H}_\kappa}^2 + \frac{C^2\Gamma(N,m)}{2\lambda N} \\
&\quad + \frac{C_1^2C^2\epsilon^2}{\sigma} + \frac{\sigma}{4}\|f_N^a - f^*\|_{L_2}^2 + \frac{C_1^2C^2\epsilon^4}{\lambda} + \frac{\lambda}{4}\|f^* - f_N^a\|_{\mathcal{H}_\kappa}^2 \\
&\le \frac{\lambda}{2}\|f^*\|_{\mathcal{H}_\kappa}^2 - \frac{\lambda}{2}\|f_N^a\|_{\mathcal{H}_\kappa}^2 + \frac{C^2\Gamma(N,m)}{2\lambda N} + \frac{\lambda}{2}\|f^*\|_{\mathcal{H}_\kappa}^2 \\
&\quad + \frac{C_1^2C^2\epsilon^2}{\sigma} + \frac{\sigma}{4}\|f_N^a - f^*\|_{L_2}^2 + \frac{C_1^2L^2\epsilon^4}{\lambda} + \frac{\lambda}{2}\|f_N^a\|_{\mathcal{H}_\kappa}^2 \\
&\le \lambda\|f^*\|_{\mathcal{H}_\kappa}^2 + \frac{C^2\Gamma(N,m)}{2\lambda N} + \frac{C_1^2C^2\epsilon^2}{\sigma} + \frac{C_1^2C^2\epsilon^4}{\lambda} \\
&\quad + \frac{\sigma}{4}\|f_N^a - f^*\|_{L_2}^2 \\
&\le \lambda\|f^*\|_{\mathcal{H}_\kappa}^2 + \frac{C^2\Gamma(N,m)}{2\lambda N} + \frac{C_1^2C^2\epsilon^2}{\sigma} + \frac{C_1^2C^2\epsilon^4}{\lambda} + \frac{1}{2}\Lambda(f_N^a),
\end{aligned}$$

where in the second inequality we apply Young's inequality $ab \le \frac{a^2}{2\epsilon} + \frac{\epsilon b^2}{2}$ twice, the last inequality follows from the strong convexity of $\ell(\mathbf{z})$ and f^* is the minimizer of $P(\ell\circ f) =$

$\mathrm{E}_{(\mathbf{x},y)}[\ell(yf(\mathbf{x}))]$. Thus, with a probability at least $1-4N^{-3}$, we have

$$P(\ell \circ f_N^a) \leq P(\ell \circ f^*) + 2\lambda \|f^*\|_{\mathcal{H}_\kappa}^2 + \frac{C^2\Gamma(N,m)}{\lambda N} + \frac{2C_1^2C^2\epsilon^2}{\sigma} + \frac{2C_1^2C^2\epsilon^4}{\lambda} + C_1Ce^{-N}.$$

We complete the proof by minimizing over λ in the R.H.S. of the above inequality.

Rong Jin is a professor of the Computer and Science Engineering Dept. at Michigan State University. He has been working in the areas of statistical machine learning and its application to information retrieval. He has extensive research experience in a variety of machine learning algorithms such as conditional exponential models, support vector machine, boosting and optimization for different applications including information retrieval. Dr. Jin is an associative editor of ACM Transactions on Knowledge Discovery from Data, and received NSF Career Award in 2006. Dr. Jin obtained his Ph.D. degree from Carnegie Mellon University in 2003, and received best paper award from Conference of Learning Theory (COLT) in 2012.

Tianbao Yang received the Ph.D. degree in Computer Science from Michigan State University in 2012. He joined GE Global Research in 2012 and works as a machine learning researcher. Dr. Yang has board interests in machine learning and he has focused on several research topics, including social network analysis and large scale optimization in machine learning. He has published over 20 papers in prestigious machine learning conferences and journals. He has won the Mark Fulk Best student paper award at 25th Conference on Learning Theory (COLT) in 2012. Tianbao Yang also served as program committee for several conferences and journals, including AAAI'12, CIKM'12,'13, IJCAI'13, ACML'12, TKDD, TKDE, and etc..

Mehrdad Mahdavi received the M.Sc. degree from Sharif University of Technology, Tehran, Iran. Since 2009, he has been pursuing a Ph.D. in Computer Science at the Michigan State University and before that he spent two years as a P.h.D. candidate at Sharif University of Technology. His current research interests include Machine Learning focused on Online (Stochastic) Convex Optimization, Learning Theory, and Learning in Games. He has won the Mark Fulk Best Student Paper Award at the Conference on Learning Theory (COLT) in 2012.

Yu-Feng Li received the BSc and PhD degrees in computer science from Nanjing University, China, in 2006 and 2013, respectively. His main research interests include machine learning and data mining. He has won the Microsoft Fellowship Award (2009). He has been Program Committee member of several conferences including IJCAI'11, ICDM'11, AAAI'12 and IJCNN'13.

Zhi-Hua Zhou (S'00-M'01-SM'06) received the BSc, MSc and PhD degrees in computer science from Nanjing University, China, in 1996, 1998 and 2000, respectively, all with the highest honors. He joined the Department of Computer Science & Technology at Nanjing University as an assistant professor in 2001, and is currently professor and Director of the LAMDA group. His research interests are mainly in artificial intelligence, machine learning, data mining, pattern recognition and multimedia information retrieval. In these areas he has published more than 100 papers in leading international journals or conference proceedings, and holds 12 patents. He has won various awards/honors including the IEEE CIS Outstanding Early Career Award, the National Science & Technology Award for Young Scholars of China, the Fok Ying Tung Young Professorship Award, the Microsoft Young Professorship Award and nine international journals/conferences paper or competition awards. He is an Associate Editor-in-Chief of the *Chinese Science Bulletin*, Associate Editor of the *ACM Transactions on Intelligent Systems and Technology* and on the editorial boards of various other journals. He is the founder and Steering Committee Chair of ACML, and Steering Committee member of PAKDD and PRICAI. He serves/ed as General Chair/Co-chair of ACML'12, ADMA'12 and PCM'13, Program Chair/Co-Chair for PAKDD'07, PRICAI'08, ACML'09, SDM'13, etc., Workshop Chair of KDD'12, Program Vice Chair or Area Chair of various conferences, and chaired many domestic conferences in China. He is the Chair of the Machine Learning Technical Committee of the Chinese Association of Artificial Intelligence, Chair of the Artificial Intelligence & Pattern Recognition Technical Committee of the China Computer Federation, Vice Chair of the Data Mining Technical Committee of IEEE Computational Intelligence Society and the Chair of the IEEE Computer Society Nanjing Chapter. He is a fellow of the IAPR, the IEEE, and the IET/IEE.