Approximate Kernel Selection via Matrix Approximation

Lizhong Ding, Shizhong Liao, Yong Liu, Li Liu, Fan Zhu, Yazhou Yao, Ling Shao, and Xin Gao

***Abstract*—Kernel selection is of fundamental importance for the generalization of kernel methods. This paper proposes an approximate approach for kernel selection by exploiting the approximability of kernel selection and the computational virtue of kernel matrix approximation. We define approximate consistency to measure the approximability of the kernel selection problem. Based on the analysis of approximate consistency, we solve the theoretical problem of whether, under what conditions and at what speed, the approximate criterion is close to the accurate one, establishing the foundations of approximate kernel selection. We introduce two selection criteria based on error estimation and prove the approximate consistency of multilevel circulant matrix (MCM) approximation and Nyström approximation under these criteria. Under the theoretical guarantees of the approximate consistency, we design approximate algorithms for kernel selection, which exploit the computational advantages of MCM and Nyström approximations to conduct kernel selection in a linear or quasi-linear complexity. We experimentally validate the theoretical results for the approximate consistency and evaluate the effectiveness of the proposed kernel selection algorithms.**

***Index Terms*—Kernel selection, kernel matrix approximation, approximate consistency, approximate algorithms**

I. Introduction

FOR A FINITE sample of data, learning involves finding a function that yields good predictions on unknown data [1], [2]. Learning is an ill-posed problem and it is impossible to obtain a unique solution based only on data itself [3]. Additional assumptions are necessary to make learning possible. We call this set of assumptions the *inductive bias* [4]. The task of model selection involves determining the inductive bias, which is critical to the performance of learning algorithms. Kernel methods play an important role in the machine learning community, which have recently been applied in multitask learning [5], k-means clustering [6], logistic regression [7], hypothesis testing [8], [9], et al. For supervised kernel-based learning algorithms [10]–[13], the aim of model selection is to choose the kernel function and the regularization parameter. In this paper, we focus on kernel selection, which has an essential influence on kernel-based learning [14]–[16].

Kernel selection is closely related to the generalization error of learning algorithms. The kernel with the smallest generalization error is usually regarded as the optimal kernel [17], [18]. However, we cannot directly compute the generalization error because the underlying distribution of the given data is often unknown. Using the theoretical upper bounds of the generalization error is a common strategy in kernel selection. The upper bounds are composed of the empirical error and the hypothesis space complexity [17]. Different measurements of the complexity constitute different kernel selection methods. These include Rademacher complexity [18], local Rademacher complexity [19], [20], maximal discrepancy [17], [18], roughness function [21], and covering number [22]. Although kernel selection is closely linked to the generalization error, kernel selection criteria are not required to be unbiased estimates of the generalization error [23], [24]. The main requirement for kernel selection criteria is that they give an indication of the generalization error. Therefore, it is sufficient to compute approximate kernel selection criteria, which discriminate the best kernel from other candidates. We refer to this property as the *approximability* of kernel selection. On the other hand, the computational complexities of the existing kernel selection criteria are at least $O(l^2)$, where l is the number of samples. This kind of scalability is prohibitive for big data. Such considerations drive the study of this paper.

By exploiting the approximability of the kernel selection problem and the computational virtue of kernel matrix approximation, we propose an approximate approach for kernel selection. We define approximate consistency, which measures the approximability of kernel selection. Then, with the notion of approximate consistency, we answer the theoretical question of approximate kernel selection: under what conditions and at what speed, the approximate criterion is close to the accurate one, if at all. We introduce two criteria defined by error estimation, as two cases for studying the approximate consistency of multilevel circulant matrix (MCM) approximation and Nyström approximation. The results demonstrate the rationality of introducing matrix approximation into kernel selection. Based on theoretical findings on the approximate consistency, we design approximate kernel selection algorithms to conduct kernel selection, which alleviate the computational bottleneck issue faced by the accurate kernel selection procedures. The designed algorithms exploit the computational virtues of MCM approximation and Nyström approximation to conduct kernel selection in $O(l\log(l))$ or $O(l)$ time complexity. We conduct experiments on benchmark and synthetic data to verify the theoretical findings for the approximate consistency and evaluate the effectiveness of the proposed approximate algorithms.

We organize the rest of this paper as follows. Section II introduces related work. In Section III, we discuss the relationship between this paper and our previous works.

L. Ding, L. Liu, F. Zhu, Y. Yao and L. Shao are with Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, UAE.

Y. Liu is with Institute of Information Engineering, CAS, Beijing, China.

S. Liao is with School of Computer Science and Technology, Tianjin University, Tianjin, China.

X. Gao is with King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center (CBRC), Saudi Arabia.

Corresponding author: Xin Gao.

Section IV presents two kernel selection criteria. Section V defines approximate consistency and analyzes the approximate consistency of MCM and Nyström approximation. Section VI elaborates on the designs of the approximate algorithms for kernel selection. In Section VII and VIII, we conduct experiments and conclude the paper.

II. Related Work

Traditional kernel-based learning algorithms suffer from high time and space complexity due to their usage of the kernel matrix. Kernel matrix approximation can be adopted to effectively reduce the computational and storage burdens. Here, we introduce two types of kernel matrix approximation methods, multilevel circulant matrix (MCM) approximation and column sampling based approximation. Using MCM approximation [25], rather than the original kernel matrix, we can approximately solve the eigensystems of the kernel matrix in a time complexity of $O(l\log(l))$ by employing a fast Fourier transform (FFT) [26]–[29]. Approximations based on column sampling have also been extensively studied, with representative methods including the classical Nyström method with different kinds of sampling strategies [30]–[35], modified Nyström method [36], [37], incomplete Cholesky decomposition [38], [39], sparse greedy approximations [40], matrix least squares approximation [41] and CUR matrix decomposition [42], [43]. The time complexity of these methods is $O(p^2l)$, where p is the rank of the approximate matrix, such as the Nyström approximate matrix.

Aside from their computational analyses, the kernel matrix approximation methods have also been theoretically studied [26], [28], [32]–[34], [36], [44]. Most existing theoretical analyses for kernel matrix approximation provide bounds for the discrepancy between the approximate matrix and the original kernel matrix for an appropriate norm (such as the Frobenius norm, spectral norm or trace norm). However, these analyses are independent of the learning problem at hand and cannot reveal the impact of the approximation of kernel matrix on learning algorithms. Recent studies [28], [45]–[47] demonstrate the influence of kernel matrix approximation on the learned hypothesis, but none of these measure the influence of kernel matrix approximation on kernel selection. The introduction of matrix approximation in kernel selection is an effective and promising strategy. However, the existing theoretical analyses on kernel matrix approximation are not sufficient to justify the appositeness of introducing matrix approximation in kernel selection. This paper defines approximate consistency to measure the discrepancy between the approximate criterion computed with the approximate matrix and the accurate criterion computed with the kernel matrix. We also analyze the convergence speed of the discrepancy for different matrix approximation algorithms.

III. Relations to Previous Works

The idea of approximate kernel selection was first proposed in [28]. To theoretically study the rationale behind approximate kernel selection, the notion of approximate consistency was defined [48], [49]. This paper refines the definition in [48] and takes approximate consistency as a basic property of approximate kernel selection algorithms rather than matrix approximation algorithms. Besides refining the definition, we have the following additional contributions. First, we provide detailed explanations and theoretical analyses for MCM to demonstrate its utility in kernel selection. Second, we design novel approximate algorithms for kernel selection by exploiting the computational virtues of MCM and Nyström approximation. Third, we provide theoretical analyses for the approximate kernel selection algorithms. Fourth, we provide extensive empirical evidence to support the theoretical findings for the approximate consistency.

From an algorithmic perspective, the MCM approximate kernel selection algorithm and the Nyström approximate kernel selection algorithm designed in this paper are inspired by the computational skills in [28] and [50], respectively. However, there are two main differences between this paper and [28], [50]. First, no kernel selection criteria were given in [28] and [50] and kernel matrix approximations were adopted to accelerate the training of LSSVM. In this paper, two kernel selection criteria based on error estimation are presented. MCM and Nyström approximate kernel selection algorithms are designed for accelerating the computation of the kernel selection criteria. Second, the theoretical bounds given in [28] and [50] measured the discrepancy between the approximate hypothesis and the accurate one, quantifying the impact of MCM and Nyström approximations on the hypothesis of LSSVM. In this paper, approximate consistency is used to measure the discrepancy between the approximate criterion on the MCM or Nyström approximate matrix and the accurate one on the kernel matrix. This investigates a fundamental theoretical problem of approximate kernel selection.

The previous work [51] and this paper both adopt the ingenious algorithm proposed in [26] to generate MCMs. However, three points distinguish this paper from [51]. First, [51] studied the *kernel combination* problem specifically for regression tasks, whereas this paper studies the *kernel selection* problem for both classification and regression tasks. Second, no kernel selection criteria were given in [51], and the combination weights of base kernels were optimized using kernel ridge regression (KRR). In contrast, two kernel selection criteria are proposed in this paper, and the optimal kernel is selected by minimizing the proposed criteria. Third, the approximation error bound given in [51] was for the discrepancy between the approximate and accurate hypotheses, measuring the impact of the approximation on the hypothesis of KRR. In this paper, the approximate consistency determines the discrepancy between the approximate and accurate criteria, accessing the impact of MCM or Nyström approximation on kernel selection criteria.

IV. Kernel Selection Criteria

We first provide some notations. The set of l labeled data points is denoted as $\mathcal{S} = ((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_l, y_l)) \in (\mathcal{X} \times \mathcal{Y})^l$, where $\mathcal{X} \subseteq \mathbb{R}^d$ is the input space and $\mathcal{Y}$ is the output domain. For the classification case, $\mathcal{Y} = \{-1, 1\}$ and for the regression case, $\mathcal{Y} = \mathbb{R}$. We assume $|y| \leq M$, $y \in \mathcal{Y}$, where M is a given constant. We consider the Mercer kernel κ in this paper, which

is a continuous, symmetric and positive definite function from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$ [2]. The kernel matrix $\mathbf{K} = [\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1}^{l}$, defined on a finite set of inputs $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l\} \subseteq \mathcal{X}$, is symmetric and positive definite (SPD). We denote the reproducing kernel Hilbert space (RKHS) of the kernel κ as $\mathcal{H}_\kappa$ [52], which is defined as $\mathcal{H}_\kappa = \overline{\text{span}}\{\kappa(\boldsymbol{x}, \cdot) : \boldsymbol{x} \in \mathcal{X}\}$. For $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_\kappa}$ of $\mathcal{H}_\kappa$ is $\langle \kappa(\boldsymbol{x}, \cdot), \kappa(\boldsymbol{x}', \cdot) \rangle_{\mathcal{H}_\kappa} = \kappa(\boldsymbol{x}, \boldsymbol{x}')$.

Now we provide two criteria for kernel selection. The first criterion is based on the regularized square loss $\mathcal{E}(f) = \frac{1}{l}\sum_{i=1}^{l}(f(\boldsymbol{x}_i) - y_i)^2 + \mu \|f\|_{\mathcal{H}_\kappa}^2$, where μ denotes the regularization parameter and $\|\cdot\|_{\mathcal{H}_\kappa}$ is the norm in $\mathcal{H}_\kappa$ induced by $\langle \cdot, \cdot \rangle_{\mathcal{H}_\kappa}$. The optimal function is $f_\kappa = \arg\min_{f \in \mathcal{H}_\kappa} \mathcal{E}(f)$. By the representer theorem [53], we have $f_\kappa = \sum_{i=1}^{l} \alpha_i \kappa(\boldsymbol{x}_i, \cdot)$ with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_l)^{\mathrm{T}} = (\mathbf{K} + \mu l \mathbf{I})^{-1}\boldsymbol{y}$, where $\boldsymbol{y} = (y_1 \ldots, y_l)^{\mathrm{T}}$ and $\mathbf{I}$ denotes the identity matrix. Therefore, $\|f_\kappa\|_{\mathcal{H}_\kappa}^2 = \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}\boldsymbol{\alpha} = \boldsymbol{y}^{\mathrm{T}}\mathbf{K}_\mu^{-1}\mathbf{K}\mathbf{K}_\mu^{-1}\boldsymbol{y}$, where $\mathbf{K}_\mu = \mathbf{K} + \mu l \mathbf{I}$. Denoting $\boldsymbol{f}_\kappa = (f_\kappa(\boldsymbol{x}_1), \ldots, f_\kappa(\boldsymbol{x}_l))^{\mathrm{T}}$, we have $\boldsymbol{f}_\kappa = \mathbf{K}\boldsymbol{\alpha} = \mathbf{K}\mathbf{K}_\mu^{-1}\boldsymbol{y}$, which implies $\boldsymbol{f}_\kappa - \boldsymbol{y} = \mathbf{K}\mathbf{K}_\mu^{-1}\boldsymbol{y} - \mathbf{K}_\mu\mathbf{K}_\mu^{-1}\boldsymbol{y} = -\mu l \mathbf{K}_\mu^{-1}\boldsymbol{y}$. Now,

$$\begin{aligned}\mathcal{E}(f_\kappa) &= \frac{1}{l}(\boldsymbol{f}_\kappa - \boldsymbol{y})^{\mathrm{T}}(\boldsymbol{f}_\kappa - \boldsymbol{y}) + \mu\|f_\kappa\|_{\mathcal{H}_\kappa}^2 \\ &= \mu^2 l \boldsymbol{y}^{\mathrm{T}}\mathbf{K}_\mu^{-1}\mathbf{K}_\mu^{-1}\boldsymbol{y} + \mu \boldsymbol{y}^{\mathrm{T}}\mathbf{K}_\mu^{-1}\mathbf{K}\mathbf{K}_\mu^{-1}\boldsymbol{y} \\ &= \mu \boldsymbol{y}^{\mathrm{T}}\mathbf{K}_\mu^{-1}\boldsymbol{y}.\end{aligned}$$

$\mathcal{E}(f_\kappa)$ is the regularized empirical error of the optimal function f_κ. For a fixed regularization parameter μ, $\mathcal{E}(f_\kappa)$ only depends on the kernel κ. It is known that the kernel function κ has a one-to-one correspondence to RKHS $\mathcal{H}_\kappa$ [52]. Different kernels correspond to different RKHSs. In different RKHSs, different optimal functions are derived. We can select the optimal function f_κ, which makes $\mathcal{E}(f_\kappa)$ the smallest, from all optimal functions. We write

$$\mathcal{C}_{\text{ree}}(\mathbf{K}) = \mathcal{E}(f_\kappa) = \mu \boldsymbol{y}^{\mathrm{T}}\mathbf{K}_\mu^{-1}\boldsymbol{y}. \tag{1}$$

Here "ree" stands for "regularized empirical error". The optimal kernel can be found by $\kappa^* = \arg\min_{\kappa \in \mathcal{K}} \mathcal{C}_{\text{ree}}(\mathbf{K})$, where $\mathcal{K} = \{\kappa_1, \ldots, \kappa_N\}$ is a prescribed set of kernels.

In the following, we present the second kernel selection criterion. We consider the case where the observed output is corrupted by noise. Specifically, we always assume $y_i = \dot{y}_i + \xi_i$, $1 \le i \le l$, where $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_l]^{\mathrm{T}}$ are random variables with finite covariance matrix $\mathbf{C}$ and mean 0. $\dot{\boldsymbol{y}} = [\dot{y}_1, \ldots \dot{y}_l]^{\mathrm{T}}$ is the underlying true output. Now $\boldsymbol{f}_\kappa = \mathbf{K}\boldsymbol{\alpha} = \mathbf{K}\mathbf{K}_\mu^{-1}\boldsymbol{y}$ is referred to as an estimate of $\dot{\boldsymbol{y}}$. We have

$$\begin{aligned}\frac{1}{l}\mathbb{E}_{\boldsymbol{\xi}}\|\boldsymbol{f}_\kappa - \dot{\boldsymbol{y}}\|^2 =& \frac{1}{l}\|\mathbb{E}_{\boldsymbol{\xi}}\boldsymbol{f}_\kappa - \dot{\boldsymbol{y}}\|^2 + \frac{1}{l}\text{trace}(\text{var}_{\boldsymbol{\xi}}(\boldsymbol{f}_\kappa)) \\ =& \frac{1}{l}\|\mathbf{K}\mathbf{K}_\mu^{-1}\dot{\boldsymbol{y}} - \dot{\boldsymbol{y}}\|^2 + \frac{1}{l}\text{trace}(\mathbf{C}\mathbf{K}^2\mathbf{K}_\mu^{-2}) \\ =& \underbrace{\mu^2 l \dot{\boldsymbol{y}}^{\mathrm{T}}\mathbf{K}_\mu^{-2}\dot{\boldsymbol{y}}}_{\text{bias}(\mathbf{K})} + \underbrace{\frac{1}{l}\text{trace}(\mathbf{C}\mathbf{K}^2\mathbf{K}_\mu^{-2})}_{\text{variance}(\mathbf{K})}.\end{aligned}$$

If $\mathbf{C}$ is equal to $\sigma^2\mathbf{I}$, we have

$$\mathcal{C}_{\text{ipe}}(\mathbf{K}) = \underbrace{\mu^2 l \dot{\boldsymbol{y}}^{\mathrm{T}}\mathbf{K}_\mu^{-2}\dot{\boldsymbol{y}}}_{\text{bias}(\mathbf{K})} + \underbrace{\frac{\sigma^2}{l}\text{trace}(\mathbf{K}^2\mathbf{K}_\mu^{-2})}_{\text{variance}(\mathbf{K})}. \tag{2}$$

Here "ipe" stands for "in-sample prediction error". The optimal kernel $\kappa^* = \arg\min_{\kappa \in \mathcal{K}} \mathcal{C}_{\text{ipe}}(\mathbf{K})$.

V. Approximate Consistency

If we have a prescribed kernel set $\mathcal{K}$, a selection criterion $\mathcal{C}(\mathbf{K})$, and a matrix approximation algorithm $\mathcal{A}$, which generates the approximation $\widetilde{\mathbf{K}}$, the approximate kernel selection is designed to select the kernel κ^* as

$$\kappa^* = \arg\min_{\kappa \in \mathcal{K}} \mathcal{C}(\mathcal{A}(\mathcal{S}, \kappa)) = \arg\min_{\kappa \in \mathcal{K}} \mathcal{C}(\tilde{\mathbf{K}}). \tag{3}$$

Let us look at the criteria $\mathcal{C}_{\text{ree}}(\mathbf{K})$ and $\mathcal{C}_{\text{ipe}}(\mathbf{K})$, presented in Section IV. Since the matrix inverse is required, the time complexities of $\mathcal{C}_{\text{ree}}(\mathbf{K})$ and $\mathcal{C}_{\text{ipe}}(\mathbf{K})$ are both $O(l^3)$, which is prohibitive for large-scale data. In Section VI, we will design approximate kernel selection algorithms by employing MCM and Nyström approximation, which exploit the specific structure of $\widetilde{\mathbf{K}}$ to efficiently conduct kernel selection.

However, before designing the algorithms, we solve the theoretical problems faced by approximate kernel selection, i.e., investigate how the approximation on the kernel matrix impacts the criterion. To do so, we analyze the discrepancy between the approximate criterion $\mathcal{C}(\widetilde{\mathbf{K}})$ and the accurate one $\mathcal{C}(\mathbf{K})$. More specifically, for finite samples, we should give the upper bound of the discrepancy between the approximate and accurate criteria; for large samples, we need to show under what conditions and at what speed, the discrepancy between the approximate and accurate criteria converges to 0. We will define and analyze the approximate consistency to solve these problems.

We denote an approximate kernel selection algorithm $\mathcal{AKS}$ as a 2-tuple: $\mathcal{AKS} = (\mathcal{C}(\mathbf{K}), \mathcal{A})$. The following is the definition of approximate consistency.

Definition 1: For an approximate kernel selection algorithm $\mathcal{AKS} = (\mathcal{C}(\mathbf{K}), \mathcal{A})$, we say $\mathcal{AKS}$ is of strong approximate consistency, if $|\mathcal{C}(\mathbf{K}) - \mathcal{C}(\widetilde{\mathbf{K}})| \le \varepsilon(l)$, where $\lim_{l\to\infty} \varepsilon(l) \to 0$. We say $\mathcal{AKS}$ is of p-order approximate consistency if $|\mathcal{C}(\mathbf{K}) - \mathcal{C}(\widetilde{\mathbf{K}})| \le \varepsilon(l)$, where $\lim_{l\to\infty} \varepsilon(l)/l^p \to 0$[1].

A. Approximate Consistency of MCM Approximation

Here we provide the definition of MCM, introduce the MCM approximation and analyze the approximate consistency of MCM approximation under $\mathcal{C}_{\text{ree}}(\mathbf{K})$ and $\mathcal{C}_{\text{ipe}}(\mathbf{K})$.

To facilitate representation, we introduce the following notations. We use $\mathbb{N}$ to denote a set of positive integers. For $m \in \mathbb{N}$, $[m] = \{0, 1, \ldots, m-1\}$. For a positive integer p and $\boldsymbol{m} = (m_0, m_1, \ldots, m_{p-1}) \in \mathbb{N}^p$, we denote

$$\begin{aligned}&\Pi_{\boldsymbol{m}} = m_0 m_1 \ldots m_{p-1}, && \text{(continued product)} \\ &[\boldsymbol{m}] = [m_0] \times [m_1] \times \cdots \times [m_{p-1}]. && \text{(Cartesian product)}\end{aligned}$$

A circulant matrix can be represented in the following form

$$\mathbf{C} = \begin{bmatrix} c_0 & c_{m-1} & \ldots & c_1 \\ c_1 & c_0 & \ldots & c_2 \\ \vdots & \vdots & \ddots & \vdots \\ c_{m-1} & c_{m-2} & \ldots & c_0 \end{bmatrix}.$$

[1]This definition is a refinement of that in [48]. Taking the approximate consistency as a basic property of $\mathcal{AKS}$ instead of $\mathcal{A}$ is more appropriate, since the approximate consistency is closely related to both $\mathcal{C}(\mathbf{K})$ and $\mathcal{A}$.

Each column of $\mathbf{C}$ is a cyclic shift of its left column. It is fully determined by its first column.

Multilevel circulant matrices [54] are defined recursively. For an integer $s \geq 1$, an $(s+1)$-level circulant matrix is a block matrix, where each block is an s-level circulant matrix. For $\boldsymbol{m} \in \mathbb{N}^p$, we use multi-dimensional indices $\boldsymbol{i} = (i_0, \ldots, i_{p-1}), \boldsymbol{j} = (j_0, \ldots, j_{p-1}) \in [\boldsymbol{m}]$ to locate the entries of a p-level circulant matrix $\mathbf{A}_{\boldsymbol{m}}$. According to [55], for $\boldsymbol{m} \in \mathbb{N}^p$, $\mathbf{A}_{\boldsymbol{m}} = [a_{\boldsymbol{i},\boldsymbol{j}} : \boldsymbol{i}, \boldsymbol{j} \in [\boldsymbol{m}]]$ is a p-level circulant matrix if, for any $\boldsymbol{i}, \boldsymbol{j} \in [\boldsymbol{m}]$,

$$a_{\boldsymbol{i},\boldsymbol{j}} = a_{i_0 - j_0 (\mathrm{mod}\ m_0), \ldots, i_{p-1} - j_{p-1} (\mathrm{mod}\ m_{p-1})}.$$

We can fully determine $\mathbf{A}_{\boldsymbol{m}}$ by its first column $a_{\boldsymbol{i},\mathbf{0}}$ with $\mathbf{0} = (0, \ldots, 0) \in \mathbb{R}^p$, so we write $\mathbf{A}_{\boldsymbol{m}} = \mathrm{circ}_{\boldsymbol{m}}[a_{\boldsymbol{i}} : \boldsymbol{i} \in [\boldsymbol{m}]]$, where $a_{\boldsymbol{i}} = a_{\boldsymbol{i},\mathbf{0}}$, for $\boldsymbol{i} \in [\boldsymbol{m}]$. We introduce an example of the MCM to explain the above notations [51] in the Supplement.

In the following, we will show how to approximate the kernel matrix with an MCM. We consider the radial basis function (RBF) kernels for MCM approximation, such as the Gaussian kernel [56]. For $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, we assume $K \in L^1(\mathbb{R})$ on $\mathcal{X}$, such that $\kappa(\boldsymbol{x}, \boldsymbol{x}') = K(\|\boldsymbol{x} - \boldsymbol{x}'\|_2)$. Since $\kappa(\boldsymbol{x}, \boldsymbol{x}') = \kappa(\boldsymbol{x}', \boldsymbol{x})$, K is always an even function. Without loss of generality, for $\boldsymbol{m} \in \mathbb{N}^p$, it is assumed that the number of data points in $\mathcal{S}$ is $\Pi_{\boldsymbol{m}}$, that is, $|\mathcal{S}| = l = \Pi_{\boldsymbol{m}}$. The data points in $\mathcal{S}$ can be relabeled as, $\mathcal{S} = \{(\boldsymbol{x}_{\boldsymbol{i}}, y_{\boldsymbol{i}}) : \boldsymbol{i} \in [\boldsymbol{m}]\}$. Now, $\mathbf{K}$ is rewritten as $\mathbf{K}_{\boldsymbol{m}} = [K(\|\boldsymbol{x}_{\boldsymbol{i}} - \boldsymbol{x}_{\boldsymbol{j}}\|_2) : \boldsymbol{i}, \boldsymbol{j} \in [\boldsymbol{m}]]$.

We introduce Algorithm 1 to demonstrate the construction of an MCM $\mathbf{U}_{\boldsymbol{m}}$ as the approximation of $\mathbf{K}_{\boldsymbol{m}}$ [26]. We give an example [51] to explain Algorithm 1 in the Supplement.

Algorithm 1 The construction of an MCM

Input: p, $\boldsymbol{m} \in \mathbb{N}^p$, K, a sequence of positive numbers $\boldsymbol{h}_{\boldsymbol{m}} = (h_{\boldsymbol{m},0}, \ldots, h_{\boldsymbol{m},p-1}) \in \mathbb{R}^p$;

Output: The MCM $\mathbf{U}_{\boldsymbol{m}}$;

1: For any $\boldsymbol{i} \in [\boldsymbol{m}]$, calculate

$$t_{\boldsymbol{i}} = K(\|[i_s h_{\boldsymbol{m},s} : s \in [p]]\|_2);$$

2: For any $\boldsymbol{i} \in [\boldsymbol{m}]$ and $s \in [p]$, let

$$D_{\boldsymbol{i},s} = \begin{cases} \{0\}, & i_s = 0, \\ \{i_s, m_s - i_s\}, & 1 \leq i_s \leq m_s - 1, \end{cases}$$

and $D_{\boldsymbol{i}} = D_{\boldsymbol{i},0} \times D_{\boldsymbol{i},1} \times \cdots \times D_{\boldsymbol{i},p-1}$;

3: For any $\boldsymbol{i} \in [\boldsymbol{m}]$ calculate $u_{\boldsymbol{i}} = \sum_{\boldsymbol{j} \in D_{\boldsymbol{i}}} t_{\boldsymbol{j}}$;

4: **Return** $\mathbf{U}_{\boldsymbol{m}} = \mathrm{circ}_{\boldsymbol{m}}[u_{\boldsymbol{i}} : \boldsymbol{i} \in [\boldsymbol{m}]]$;

$\mathbf{U}_{\boldsymbol{m}}$, shown in Algorithm 1, is an MCM specifically designed for kernel approximation. First, $\mathbf{U}_{\boldsymbol{m}}$ is symmetric and $\mathbf{U}_{\boldsymbol{m}} + \mu l \mathbf{I}_{\boldsymbol{m}}$ is invertible. The invertibility of $\mathbf{U}_{\boldsymbol{m}} + \mu l \mathbf{I}_{\boldsymbol{m}}$ was proved in [26]. Since $\mu l \mathbf{I}_{\boldsymbol{m}}$ is symmetric, we only need to guarantee the symmetry of $\mathbf{U}_{\boldsymbol{m}}$. The symmetry of $\mathbf{U}_{\boldsymbol{m}}$ is proved in Proposition 1. We provide the proof of Proposition 1 in the Supplement [51]. It is worth noting that Step 2 in Algorithm 2 guarantees the symmetry of the MCM $\mathbf{U}_{\boldsymbol{m}}$. Second, we can see that all the approximate elements $t_{\boldsymbol{i}} = K(\|[i_s h_{\boldsymbol{m},s} : s \in [p]]\|_2)$ of the original kernel matrix have been summed into the first column of the MCM $\mathbf{U}_{\boldsymbol{m}}$. This is carried out in Step 1 and Step 3 of Algorithm 2.

Proposition 1: For p and $\boldsymbol{m} \in \mathbb{N}^p$, the MCM $\mathbf{U}_{\boldsymbol{m}}$ constructed by Algorithm 1 is symmetric.

Theorem 1 shows the strong approximate consistency of MCM approximation under $\mathcal{C}_{\mathrm{ree}}(\mathbf{K}_{\boldsymbol{m}})$. We prove the upper bound of the discrepancy between the approximate and accurate criteria, and then analyze the convergence of the discrepancy. The proof of Theorem 1 is given in the Supplement.

Theorem 1: If the assumptions:

(H1) there are positive constants c_0 and β such that $|K(x) - K(x')| \leq c_0 |x - x'|^\beta$ for $x, x' \in \mathbb{R}$;

(H2) there is a positive constant h such that $h_{\boldsymbol{m},j} \geq h$ for $\boldsymbol{m} \in \mathbb{N}^p$ and $j \in [p]$;

(H3) there are positive constants λ_1 and c_1 such that $|K(x)| \leq c_1 \mathrm{e}^{-\lambda_1 |x|}$ for $x \in \mathbb{R}$;

(H4) there are positive constants λ_2 and c_2 such that

$$\begin{aligned} &|\|x_{\boldsymbol{i}} - x_{\boldsymbol{j}}\|_2 - \|[(i_s - j_s) h_{\boldsymbol{m},s} : s \in [p]]\|_2| \\ &\leq c_2 \sum_{s \in [p]} \left(\mathrm{e}^{-\lambda_2 \delta_{m_s}(i_s)} + \mathrm{e}^{-\lambda_2 \delta_{m_s}(j_s)} \right), \end{aligned}$$

for $\boldsymbol{m} \in \mathbb{N}^p$ and $\boldsymbol{i}, \boldsymbol{j} \in [\boldsymbol{m}]$, where $\delta_m(j) = \frac{m}{2} - |\frac{m}{2} - j|$ for $m \in \mathbb{N}$ and $j \in [m]$;

hold and furthermore, there are positive constants c_3 and r_1 satisfying $|y_{\boldsymbol{i}}| \leq c_3 \mathrm{e}^{-r_1 \nu_{\boldsymbol{m}}(\boldsymbol{i})}$, for any $\boldsymbol{m} \in \mathbb{N}^p$ and $\boldsymbol{i} \in [\boldsymbol{m}]$, where $\nu_{\boldsymbol{m}}(\boldsymbol{i}) = \left\| \frac{\boldsymbol{m}}{2} - \boldsymbol{i} \right\|_2$, then we have

$$\lim_{\boldsymbol{m} \to \infty} |\mathcal{C}_{\mathrm{ree}}(\mathbf{K}_{\boldsymbol{m}}) - \mathcal{C}_{\mathrm{ree}}(\mathbf{U}_{\boldsymbol{m}})| = 0,$$

where $\boldsymbol{m} \to \infty$ indicates that all components of $\boldsymbol{m}$ go to infinity.

We further study the approximate consistency of MCM approximation under $\mathcal{C}_{\mathrm{ipe}}(\mathbf{K}_{\boldsymbol{m}})$. Theorem 2 shows the upper bound on the variance and the bias term of $\mathcal{C}_{\mathrm{ipe}}(\mathbf{K}_{\boldsymbol{m}})$. Based on Theorem 2, we can obtain Theorem 3, which shows the strong approximate consistency of MCM approximation under $\mathcal{C}_{\mathrm{ipe}}(\mathbf{K}_{\boldsymbol{m}})$. The proof of Theorem 2 is in the Supplement.

Theorem 2: If the assumptions in Theorem 1 hold, we have

$$|\mathrm{variance}(\mathbf{K}_{\boldsymbol{m}}) - \mathrm{variance}(\mathbf{U}_{\boldsymbol{m}})| \leq c\sigma^2 (m_{\min})^{-1}$$

for a positive constant c, where $m_{\min} = \min\{m_s : s \in [p]\}$. If for any $\boldsymbol{m} \in \mathbb{N}^p$ and $\boldsymbol{i} \in [\boldsymbol{m}]$, there exist positive constants c_3 and r_1 such that $|\dot{y}_{\boldsymbol{i}}| \leq c_3 \mathrm{e}^{-r_1 \nu_{\boldsymbol{m}}(\boldsymbol{i})}$, where $\nu_{\boldsymbol{m}}(\boldsymbol{i}) = \left\| \frac{\boldsymbol{m}}{2} - \boldsymbol{i} \right\|_2$, then for any $\boldsymbol{m} \in \mathbb{N}^p$,

$$|\mathrm{bias}(\mathbf{K}_{\boldsymbol{m}}) - \mathrm{bias}(\mathbf{U}_{\boldsymbol{m}})| \leq c\mu^2 \Pi_{\boldsymbol{m}}^{3/2} \mathrm{e}^{-r' m_{\min}},$$

for positive constants c and r'.

Theorem 3: If the assumptions in Theorem 1 hold, we have

$$\lim_{\boldsymbol{m} \to \infty} |\mathcal{C}_{\mathrm{ipe}}(\mathbf{K}_{\boldsymbol{m}}) - \mathcal{C}_{\mathrm{ipe}}(\mathbf{U}_{\boldsymbol{m}})| = 0.$$

B. Approximate Consistency of Nyström Approximation

We now briefly review Nyström approximation [30]. We first randomly select c columns of $\mathbf{K}$. We denote $\mathbf{C}$ as an $l \times c$ matrix formed by the selected columns. We use $\mathbf{W}$ to denote the $c \times c$ matrix composed of the intersection between the selected c columns and the corresponding c rows of $\mathbf{K}$. The Nyström approximate matrix is

$$\widetilde{\mathbf{K}} = \mathbf{C} \mathbf{W}_k^+ \mathbf{C}^{\mathrm{T}} \approx \mathbf{K},$$

where $\mathbf{W}_k$ is the optimal rank k approximation to $\mathbf{W}$ and $\mathbf{W}_k^+$ is the generalized inverse of $\mathbf{W}_k$.

Modified Nyström approximation [36] presents a tighter approximation error bound than the classical Nyström approximation but has a higher computational cost. We can write the approximate matrix of modified Nyström approximation as

$$\widetilde{\mathbf{K}} = \mathbf{C}\left(\mathbf{C}^{+}\mathbf{K}\left(\mathbf{C}^{+}\right)^{\mathrm{T}}\right)\mathbf{C}^{\mathrm{T}}.$$

Although there are many different versions of Nyström approximation with different sampling strategies [30]–[34], [36], we concentrate on the approximations with $(1+\epsilon)$ relative-error bounds, where ϵ is independent of l. The bound for Nyström approximation [34] is derived using leverage score based column sampling [42], which states that for $\epsilon \in (0,1]$ and a failure probability $\delta \in (0,1]$,

$$\|\mathbf{K} - \widetilde{\mathbf{K}}\|_* \leq (1+\epsilon)\|\mathbf{K} - \mathbf{K}_k\|_* \tag{4}$$

holds with a probability of at least $0.6-\delta$. For modified Nyström approximation [36] the bound is derived by combining the near-optimal sampling [57] and the error-driven adaptive sampling [58],

$$\mathbb{E}\left(\|\mathbf{K} - \widetilde{\mathbf{K}}\|_{\mathrm{F}}\right) \leq (1+\epsilon)\|\mathbf{K} - \mathbf{K}_k\|_{\mathrm{F}}.$$

Before analyzing the approximate consistency of the classical and modified Nyström approximations, we introduce two assumptions.

Assumption 1. For the rank $k \leq c \ll l$ and $\rho \in (0, 1/2)$, we assume $\lambda_k(\mathbf{K}) = \Omega(l/c^{\rho})$ and $\lambda_{k+1}(\mathbf{K}) = O(l/c^{1-\rho})$, where ρ is used to characterize the gap between the k-th and $(k+1)$-th eigenvalues.

Assumption 2. We assume that the sampling size c is a small ratio r of l and the rank parameter k is a constant.

Assumption 1 is not a strong assumption. As suggested in [38], the eigenvalues of the kernel matrix have polynomial or exponential decay. The eigenvalues of Gaussian kernels have exponential decay [19]. Assumption 1 is always weaker than exponential decay, even when ρ goes to 0. When ρ is close to 1/2, Assumption 1 is weaker than polynomial decay. Assumption 1 was adopted in [46] and [47] and experimentally tested in [46]. Assumption 2 is a common setting of Nyström approximation. The constant rank was adopted in [36].

Theorem 4 shows the $\frac{1}{2}$-order approximate consistency of the Nyström approximation under $\mathcal{C}_{\mathrm{ree}}(\mathbf{K})$. The proof is given in the Supplement.

Theorem 4: For $\mathcal{C}_{\mathrm{ree}}(\mathbf{K})$, if Assumption 1 and Assumption 2 hold, we have that $|\mathcal{C}_{\mathrm{ree}}(\mathbf{K}) - \mathcal{C}_{\mathrm{ree}}(\widetilde{\mathbf{K}})| \leq \varepsilon(l)$ for $\delta \in (0,1]$ and $\epsilon \in (0,1]$, holds with a probability of at least $0.6-\delta$, where $\widetilde{\mathbf{K}}$ is produced by Nyström approximation with leverage score sampling,

$$\varepsilon(l) = \frac{\tau M^2(1+\epsilon)}{\mu r^{1-\rho} l^{1-\rho}}(l-k)$$

for constant τ and $\lim_{l\to\infty} \varepsilon(l)/l^{\frac{1}{2}} \to 0$.

Theorem 5 shows the strong approximate consistency of modified Nyström approximation under $\mathcal{C}_{\mathrm{ree}}(\mathbf{K})$. The proof is also given in the Supplement.

Algorithm 2 MCM Approximate Kernel Selection

Input: $\boldsymbol{y} = \{y_{\boldsymbol{j}} : \boldsymbol{j} \in [\boldsymbol{m}]\}$, $\mathcal{K} = \{\kappa_1, \ldots, \kappa_N\}$, μ;
Output: The optimal kernel κ^*;

```
Initialize: C* = ∞;
for each κ ∈ K do
    Calculate [u_j : j ∈ [m]] according to Algorithm 1;
    Calculate v = Φ[u_j : j ∈ [m]] by mFFT;
    Calculate η = Φy using mFFT;
    Calculate τ = diag(1/(v_j + μl) : j ∈ [m]) η;
    Calculate ζ = (1/Π_m) Φ^H τ using inverse mFFT;
    C_ree(U_m) = μ y^T ζ;
    if C_ree(U_m) ≤ C* then
        C* = C_ree(U_m);
        κ* = κ;
Return κ*;
```

Theorem 5: For $\mathcal{C}_{\mathrm{ree}}(\mathbf{K})$, if Assumption 1 and Assumption 2 hold, we have $\mathbb{E}\left(|\mathcal{C}_{\mathrm{ree}}(\mathbf{K}) - \mathcal{C}_{\mathrm{ree}}(\widetilde{\mathbf{K}})|\right) \leq \varepsilon(l)$, where $\widetilde{\mathbf{K}}$ is produced by modified Nyström approximation,

$$\varepsilon(l) = \frac{\tau M^2(1+\epsilon)}{\mu r^{1-\rho} l^{1-\rho}}\sqrt{l-k}$$

for constant τ and $\lim_{l\to\infty} \varepsilon(l) \to 0$.

VI. Approximate Kernel Selection Algorithms

Under the theoretical guarantee of approximate consistency, we design approximate kernel selection algorithms using MCM and Nyström approximations.

A. Approximate Kernel Selection with MCM Approximation

Here, we adopt the inverse of $\mathbf{U}_{\boldsymbol{m}} + \mu l \mathbf{I}_{\boldsymbol{m}}$ to approximate the inverse of $\mathbf{K}_{\boldsymbol{m}} + \mu l \mathbf{I}_{\boldsymbol{m}}$, where $\mathbf{I}_{\boldsymbol{m}}$ is an identity matrix. The eigenvalues and eigenvectors of an MCM $\mathbf{U}_{\boldsymbol{m}}$ [55][2] can be represented as $\mathbf{U}_{\boldsymbol{m}} = \frac{1}{\Pi_{\boldsymbol{m}}}\mathbf{\Phi}^{\mathrm{H}}\mathrm{diag}(\boldsymbol{v})\mathbf{\Phi}$, where the vector of eigenvalues is $\boldsymbol{v} = \mathbf{\Phi}[u_{\boldsymbol{j}} : \boldsymbol{j} \in [\boldsymbol{m}]]$. $\mathbf{\Phi}$ is the Kronecker product of the Fourier matrices. For any vector $\boldsymbol{x} = [x_{\boldsymbol{i}} : \boldsymbol{i} \in [\boldsymbol{m}]]$, $\mathbf{\Phi}\boldsymbol{x}$ is the multidimensional discrete Fourier transform (mDFT) of $\boldsymbol{x}$. Therefore, we can compute $\mathbf{\Phi}\boldsymbol{x}$ through a multidimensional fast Fourier transform (mFFT). We can compute the eigenvalues of an MCM by applying mFFT to its first column [55]. It follows that

$$(\mathbf{U}_{\boldsymbol{m}} + \mu l \mathbf{I}_{\boldsymbol{m}})^{-1} = \frac{1}{\Pi_{\boldsymbol{m}}}\mathbf{\Phi}^{\mathrm{H}}\mathrm{diag}\left(\frac{1}{v_{\boldsymbol{j}} + \mu l} : \boldsymbol{j} \in [\boldsymbol{m}]\right)\mathbf{\Phi}.$$

Now we design an MCM approximate kernel selection algorithm (Algorithm 2) for $\mathcal{C}_{\mathrm{ree}}$. The time complexity of Algorithm 2 is shown in Theorem 6.

Theorem 6: The time complexity of Algorithm 2 is $O(N(l\max\{\log(l), p\}))$.

Proof: We assume that $l = \Pi_{\boldsymbol{m}}$. The time complexity of step 3 is $O(lp)$ for RBF kernels, where p is the number of levels of the MCM. We know that the time complexity of steps 4, 5 and 7 is $O(l\log(l))$, since mFFT can be applied with the $O(l\log(l))$ [59] complexity. The time complexity of

[2]For complete lemmas, please see Section V of the Supplement.

Algorithm 3 Nyström Approximate Kernel Selection

Input: $\mathcal{S} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{l}$, $\mathcal{K} = \{\kappa_1, \ldots, \kappa_N\}$, c, k, μ;
Output: The optimal kernel κ^*;
1: **Initialize**: $\mathcal{C}^* = \infty$;
2: **for** each $\kappa \in \mathcal{K}$ **do**
3: Sample c indices from $\{1, \ldots, l\}$ to form the index set $\mathcal{I}$;
4: Generate $\mathbf{C}$ and $\mathbf{W}$ using $\mathcal{S}$ and κ according to $\mathcal{I}$;
5: Calculate the SVD of $\mathbf{W}$ as $\mathbf{W} = \mathbf{U}_{\mathbf{W}}\mathbf{\Sigma}_{\mathbf{W}}\mathbf{U}_{\mathbf{W}}^{\mathrm{T}}$;
6: Let $\mathbf{V} = \mathbf{C}\mathbf{U}_{\mathbf{W},k}\sqrt{\mathbf{\Sigma}_{\mathbf{W},k}^{+}}$;
7: Solve $\left(\mu l\mathbf{I}_k + \mathbf{V}^{\mathrm{T}}\mathbf{V}\right)\boldsymbol{t} = \mathbf{V}^{\mathrm{T}}\boldsymbol{y}$ to obtain $\boldsymbol{t}$;
8: $\boldsymbol{u} = \frac{1}{\mu l}(\boldsymbol{y} - \mathbf{V}\boldsymbol{t})$;
9: $\mathcal{C}_{\mathrm{ree}}(\widetilde{\mathbf{K}}) = \mu\boldsymbol{y}^{\mathrm{T}}\boldsymbol{u}$;
10: **if** $\mathcal{C}_{\mathrm{ree}}(\widetilde{\mathbf{K}}) \leq \mathcal{C}^*$ **then**
11: $\mathcal{C}^* = \mathcal{C}_{\mathrm{ree}}(\widetilde{\mathbf{K}})$;
12: $\kappa^* = \kappa$;
13: **Return** κ^*;

step 6 is $O(l)$. Because the number of candidate kernels is N, the total time complexity is $O(N(l\max\{\log(l), p\}))$. ∎

The time complexity of computing $\mathcal{C}_{\mathrm{ree}}(\mathbf{U}_{\boldsymbol{m}})$ for each candidate kernel κ is $O(l\max\{\log(l), p\})$, which is quasi-linear in the number of samples l and much lower than $O(l^3)$. The space complexity of $\mathcal{C}_{\mathrm{ree}}(\mathbf{U}_{\boldsymbol{m}})$ is $O(l)$, since we only need to store the first column of $\mathbf{U}_{\boldsymbol{m}}$, which is lower than the $O(l^2)$ space complexity required to store the kernel matrix.

We further discuss the approximate computation of the kernel selection criterion $\mathcal{C}_{\mathrm{ipe}}$, defined in equation (2). We can approximately compute the bias term $\mu^2 l\dot{\boldsymbol{y}}^{\mathrm{T}}(\mathbf{K}_{\boldsymbol{m}}+\mu l\mathbf{I}_{\boldsymbol{m}})^{-2}\dot{\boldsymbol{y}}$ of $\mathcal{C}_{\mathrm{ipe}}$, using steps 5-7 of Algorithm 2, twice. For the variance term $\frac{\sigma^2}{l}\mathrm{trace}(\mathbf{K}_{\boldsymbol{m}}^2(\mathbf{K}_{\boldsymbol{m}} + \mu l\mathbf{I}_{\boldsymbol{m}})^{-2})$ of $\mathcal{C}_{\mathrm{ipe}}$, we compute the eigenvalues of $\mathbf{U}_{\boldsymbol{m}}$ as the approximations of the eigenvalues of the kernel matrix $\mathbf{K}_{\boldsymbol{m}}$. Therefore,

$$\mathrm{trace}(\mathbf{K}_{\boldsymbol{m}}^2(\mathbf{K}_{\boldsymbol{m}} + \mu l\mathbf{I}_{\boldsymbol{m}})^{-2}) \approx \sum_{\boldsymbol{i}\in[\boldsymbol{m}]}\left(\frac{v_{\boldsymbol{i}}}{v_{\boldsymbol{i}} + \mu l}\right)^2.$$

B. Approximate Kernel Selection with Nyström Approximation

The Nyström approximate matrix can be represented as $\widetilde{\mathbf{K}} = \mathbf{C}\mathbf{W}_k^{+}\mathbf{C}^{\mathrm{T}} \approx \mathbf{K}$. The SVD of $\mathbf{W}$ is $\mathbf{W} = \mathbf{U}_W\mathbf{\Sigma}_W\mathbf{U}_W^{\mathrm{T}}$, so $\mathbf{W}_k^{+} = \mathbf{U}_{\mathbf{W},k}\mathbf{\Sigma}_{\mathbf{W},k}^{+}\mathbf{U}_{\mathbf{W},k}^{\mathrm{T}}$. Usually, we consider the case where $k < \mathrm{rank}(\mathbf{W})$, since when $k \geq \mathrm{rank}(\mathbf{W})$, the Nyström approximation is exact [60]. Therefore, all elements of $\mathbf{\Sigma}_{\mathbf{W},k}$ are positive. Now we have

$$\widetilde{\mathbf{K}} = \underbrace{\mathbf{C}\mathbf{U}_{\mathbf{W},k}\sqrt{\mathbf{\Sigma}_{\mathbf{W},k}^{+}}}_{\mathbf{V}}\underbrace{\left(\mathbf{C}\mathbf{U}_{\mathbf{W},k}\sqrt{\mathbf{\Sigma}_{\mathbf{W},k}^{+}}\right)^{\mathrm{T}}}_{\mathbf{V}^{\mathrm{T}}}.$$

We adopt $\widetilde{\mathbf{K}} + \mu l\mathbf{I}$ as an approximation of $\mathbf{K} + \mu l\mathbf{I}$. Since $\widetilde{\mathbf{K}}$ is positive semi-definite, the invertibility of $\widetilde{\mathbf{K}} + \mu l\mathbf{I}$ is guaranteed. Using the Woodbury formula, we can obtain

$$\left(\mu l\mathbf{I} + \widetilde{\mathbf{K}}\right)^{-1} = \frac{1}{\mu l}\left(\mathbf{I} - \mathbf{V}\left(\mu l\mathbf{I}_k + \mathbf{V}^{\mathrm{T}}\mathbf{V}\right)^{-1}\mathbf{V}^{\mathrm{T}}\right), \quad (5)$$

where $\mathbf{I}_k$ is a $k \times k$ identity matrix. We let $\boldsymbol{u} = (\mu l\mathbf{I} + \widetilde{\mathbf{K}})^{-1}\boldsymbol{y}$. According to equation (5), we have $\boldsymbol{u} = \frac{1}{\mu l}\left(\boldsymbol{y} - \mathbf{V}\left(\mu l\mathbf{I}_k + \mathbf{V}^{\mathrm{T}}\mathbf{V}\right)^{-1}\mathbf{V}^{\mathrm{T}}\boldsymbol{y}\right)$. To avoid direct matrix multiplication, we introduce a temporary variable $\boldsymbol{t}$: $\left(\mu l\mathbf{I}_k + \mathbf{V}^{\mathrm{T}}\mathbf{V}\right)\boldsymbol{t} = \mathbf{V}^{\mathrm{T}}\boldsymbol{y}$, $\boldsymbol{u} = \frac{1}{\mu l}(\boldsymbol{y} - \mathbf{V}\boldsymbol{t})$. We present a Nyström approximate kernel selection algorithm (Algorithm 3) for $\mathcal{C}_{\mathrm{ree}}$.

Theorem 7: The computational complexity of Algorithm 3 is $O(N(c^3 + lc\max\{d, k\}))$.

Proof: The complexity of step 4 is $O(lcd)$, where d is the dimension of the input data. The complexity of step 5 is $O(c^3)$, since this step conducts SVD. Step 6 has a complexity of $O(lck)$. In step 7, we solve the inverse of $\left(\mu l\mathbf{I}_k + \mathbf{V}^{\mathrm{T}}\mathbf{V}\right)$ with the complexity $O(k^3)$. Computing the matrix of the linear system takes $O(lk^2)$ multiplications. Therefore, the total complexity of step 7 is $O(lk^2)$. The complexity of step 8 is $O(lk)$. Since $k < c \ll l$ and the number of candidate kernels is N, the total time complexity of Algorithm 3 is $O(N(c^3 + lc\max\{d, k\}))$. ∎

We further discuss the approximate computation of the kernel selection criterion $\mathcal{C}_{\mathrm{ipe}}$. The computation of the bias term of $\mathcal{C}_{\mathrm{ipe}}$ is similar to that of $\mathcal{C}_{\mathrm{ree}}$. For the variance term of $\mathcal{C}_{\mathrm{ipe}}$, we need the sum of the eigenvalues of $\mathbf{K}^2\mathbf{K}_{\mu}^{-2}$. Actually, the Nyström approximate matrix $\widetilde{\mathbf{K}}$ corresponds to an approximate eigen-decomposition of $\mathbf{K}$ [30],

$$\widetilde{\mathbf{K}} = \underbrace{\sqrt{\frac{c}{l}}\mathbf{C}\mathbf{U}_{\mathbf{W},k}\mathbf{\Sigma}_{\mathbf{W},k}^{+}}_{\widetilde{\mathbf{U}}}\underbrace{\frac{l}{c}\mathbf{\Sigma}_{\mathbf{W},k}}_{\widetilde{\mathbf{\Sigma}}}\underbrace{\left(\sqrt{\frac{c}{l}}\mathbf{C}\mathbf{U}_{\mathbf{W},k}\mathbf{\Sigma}_{\mathbf{W},k}^{+}\right)^{\mathrm{T}}}_{\widetilde{\mathbf{U}}^{\mathrm{T}}}.$$

We use $\widetilde{\mathbf{\Sigma}} = \frac{l}{c}\mathbf{\Sigma}_{\mathbf{W},k} = \frac{l}{c}\mathrm{diag}(\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_k)$ as the eigenvalues of $\widetilde{\mathbf{K}}$ to approximate the eigenvalues of $\mathbf{K}$, where $\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_k$ are the top-k eigenvalues of $\mathbf{W}$. Therefore,

$$\mathrm{trace}(\mathbf{K}^2\mathbf{K}_{\mu}^{-2}) \approx \sum_{i=1}^{k}\left(\frac{(l/c)\tilde{\lambda}_i}{(l/c)\tilde{\lambda}_i + \mu l}\right)^2 = \sum_{i=1}^{k}\left(\frac{\tilde{\lambda}_i}{\tilde{\lambda}_i + \mu c}\right)^2.$$

According to the above analysis, we can obtain the approximate kernel selection algorithm for the criterion $\mathcal{C}_{\mathrm{ipe}}$, which is similar to Algorithm 3 and omitted here. Its computational complexity is $O(N(c^3 + lc\max\{d, k\}))$.

VII. Experiments

In this section, we empirically verified the theoretical findings about the approximate consistency and evaluated the effectiveness of the proposed approximate algorithms.

We conducted experiments on benchmark datasets which are publicly available from UCI Repository, StatLib Datasets and Weka Datasets, for both regression and classification problems. We randomly split each dataset into training and test sets (50% of all samples for training and the other 50% for testing). We adopt $\kappa(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\gamma\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2\right)$ with a variable width γ as the kernel set $\mathcal{K}$. Since the focus of this work is not on tuning the regularization parameter μ, we just set it as a fixed value 0.005. We set the parameter σ in $\mathcal{C}_{\mathrm{ipe}}$ to be 1% of the standard deviation of the response vector $\boldsymbol{y}$. All the implementations were in R language.

In the first experiment, we evaluated the theoretical findings on the approximate consistency. We compared six kernel matrix approximation algorithms under $\mathcal{C}_{\mathrm{ree}}$ and $\mathcal{C}_{\mathrm{ipe}}$, including the optimal-rank k approximation derived with SVD (OptApp), Nyström approximation with uniform sampling (Uniform)

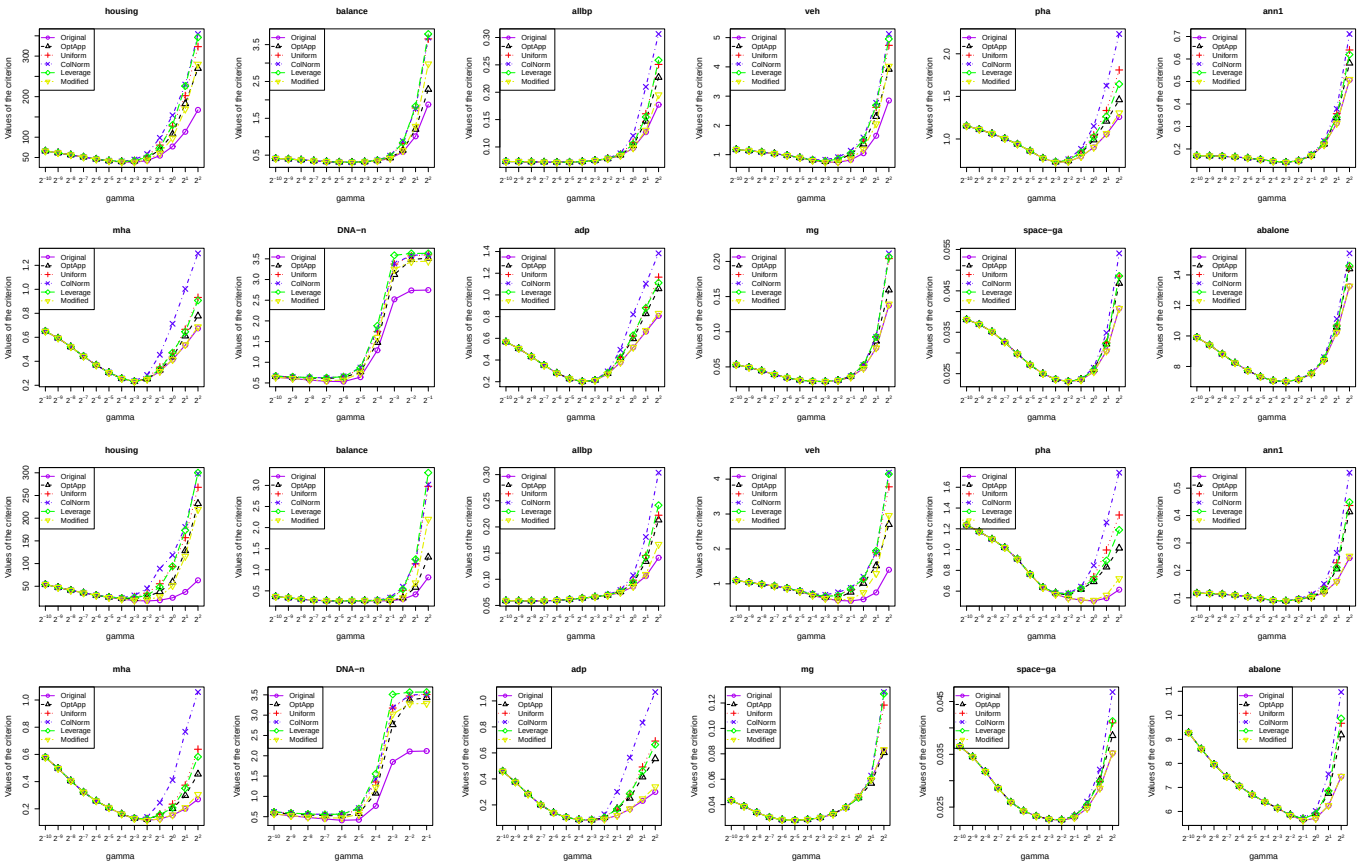


Fig. 1. Approximate consistency of different kernel matrix approximation algorithms under $\mathcal{C}_{\mathrm{ree}}(\mathbf{K})$ and $\mathcal{C}_{\mathrm{ipe}}(\mathbf{K})$ for regression data.

[32], Nyström approximation with column-norm sampling (ColNorm) [31], Nyström approximation with leverage-score sampling (Leverage) [34][3], modified Nyström approximation (Modified)[4] [36] and MCM approximation (MCM) [28]. We set $k = 20$ and $c = 0.2l$. To avoid randomness, all experiments for the compared Nyström methods were repeated 20 times. For MCM approximation, a 3-level circulant matrix was adopted. Based on (H4) of Theorem 1, we can tune h to minimize the Frobenius norm of the difference between $\mathbf{X}_{\boldsymbol{m}} = [\|\boldsymbol{x_i} - \boldsymbol{x_j}\| : \boldsymbol{i}, \boldsymbol{j} \in [\boldsymbol{m}]]$ and $\mathbf{H}_{\boldsymbol{m}} = [h\|\boldsymbol{i} - \boldsymbol{j}\| : \boldsymbol{i}, \boldsymbol{j} \in [\boldsymbol{m}]]$. However, in experiments we fixed $\boldsymbol{h_m} = (5, 5, 5)$, since it is enough to demonstrate the effectiveness of approximate kernel selection algorithms. Similar setting has been adopted in [51].

We generated synthetic data following the settings in [27][5]. The target function [27] is $f(\boldsymbol{x}) = \mathrm{e}^{-8(3-\|\boldsymbol{x}\|_2)^2} - \mathrm{e}^{-8(1.5-\|\boldsymbol{x}\|_2)^2} - \mathrm{e}^{-8(2-\|\boldsymbol{x}\|_2)^2}$. We used $\{(\boldsymbol{x_j}, y_j), \boldsymbol{j} \in [\boldsymbol{m}]\} \in \mathbb{R}^2 \times \mathbb{R}$ for $\boldsymbol{m} = (10, 10), (20, 20), (30, 30), (40, 40)$ as data points. Each dimension of the sampled inputs $\boldsymbol{x_j}$ is centered at 0. For two successive points, there is a fixed difference of 0.1 between them, and $y_j = f(\boldsymbol{x_j}) + \xi$, where ξ is a Gaussian random variable with mean 0 and standard deviation 0.01.

For each kernel parameter γ, we observed the values of the original criterion $\mathcal{C}(\mathbf{K})$ and the approximate criterion $\mathcal{C}(\widetilde{\mathbf{K}})$. The results of $\mathcal{C}_{\mathrm{ree}}$ and $\mathcal{C}_{\mathrm{ipe}}$ for regression were shown in Fig. 1. The results for classification were shown in Fig. 2. We found that the curves of the original criterion and the approximate criteria were close for most datasets. The curves of OptApp and Modified were closer to the curve of the original criterion than the Nyström approximations, which was in accordance with our theoretical findings on the approximate consistency. The results on the synthetic data were also provided (Fig. 3). We observed that the more samples, the closer the curves of the original criteria and approximate ones were.

In the second experiment, we evaluated the performance of the optimal kernels selected by the proposed approximate algorithms in Section VI. We determined the effectiveness to assess the performance. Effectiveness includes efficiency and generalization, where the former is measured by averaging the computational time for kernel selection and the latter is measured by the mean testing error of the trained model with the kernel selected by the kernel selection algorithm. For the regression and classification problems, we used kernel ridge regression (KRR) and least squares support vector machine (LSSVM) as the base models, respectively.

In the first step of the experiment, we selected the optimal kernel from the candidate kernel set $\mathcal{K}$ by minimizing the accurate or approximate kernel selection criterion on the

[3]Although in the theoretical analysis, we only considered the Nyström approximation with $(1 + \epsilon)$ relative-error bound, in experiments, for the purpose of comparison, we also considered the Nyström approximation with other sampling distributions.

[4]For modified Nyström approximation, we used the uniform+adpative^2 sampling [37].

[5]To strictly satisfy (H4) in Theorem 1, experiments for MCM approximation were only conducted on synthetic data in the first experiment.

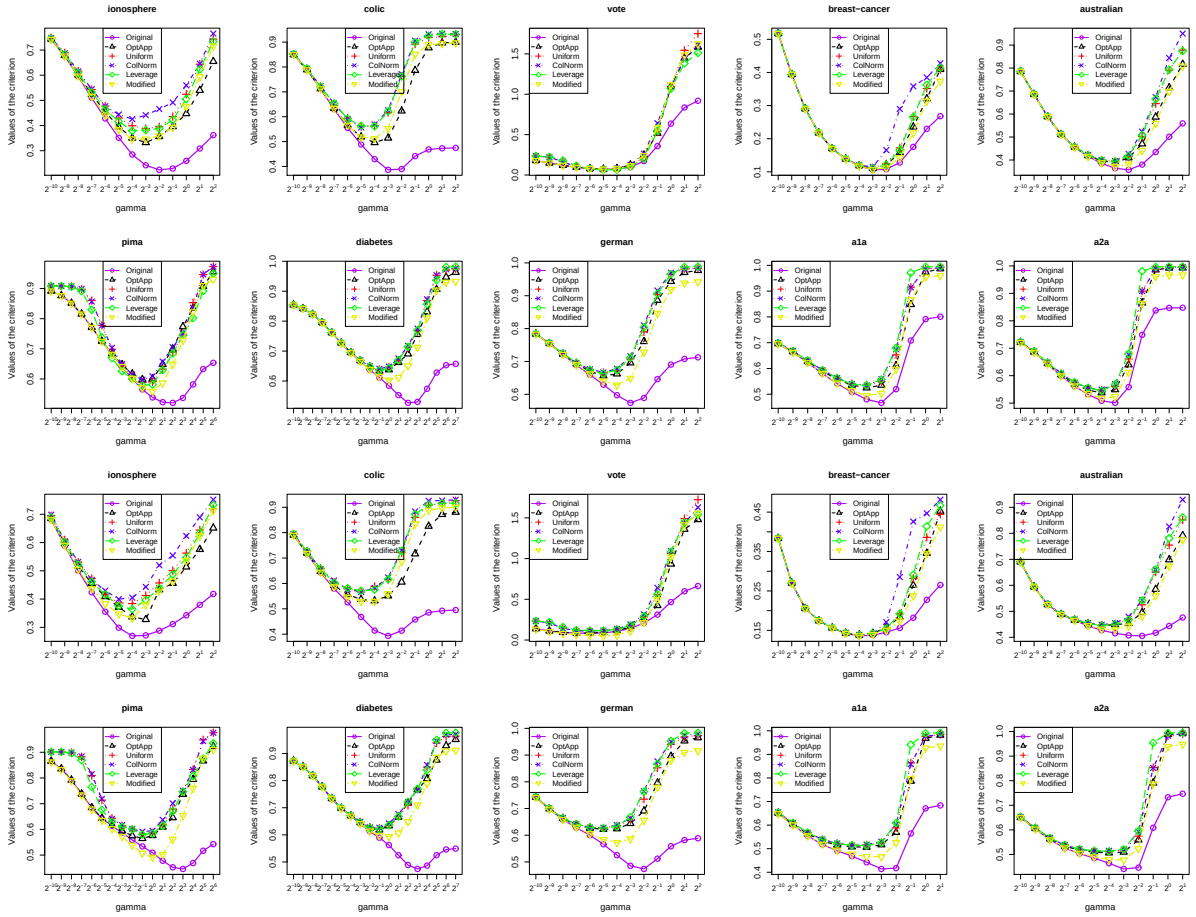
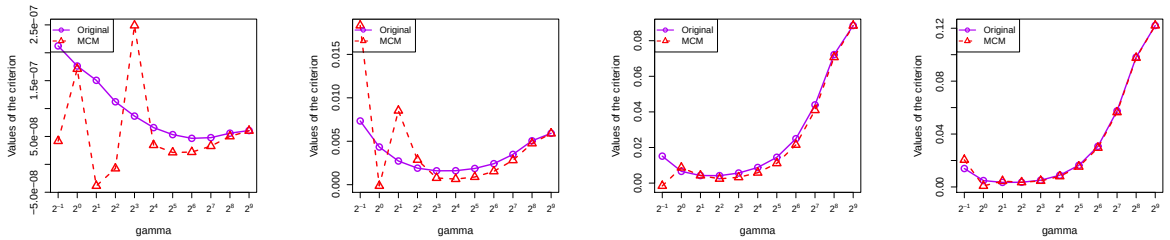


Fig. 2. Approximate consistency of different kernel matrix approximation algorithms under $\mathcal{C}_{\mathrm{ree}}(\mathbf{K})$ and $\mathcal{C}_{\mathrm{ipe}}(\mathbf{K})$ for classification data.



(a) 100 (b) 400 (c) 900 (d) 1600

Fig. 3. Evolution of the approximate consistency of the MCM approximation as the number of samples $\Pi_{\boldsymbol{m}}$ increases.. $\Pi_{\boldsymbol{m}}$ increases from 100 to 1600. We can find that the more samples there are, the closer the curves of "MCM" and "Original" are.

training set. Then, we trained the model using the selected optimal kernel, again on the training set. Finally, we evaluated the test performance of the trained model on the test set. The test performance was evaluated in terms of the mean square error (MSE) for the regression problems and the mean classification error rate (MCER) for the classification tasks. The experimental results for the criteria $\mathcal{C}_{\mathrm{ree}}$ and $\mathcal{C}_{\mathrm{ipe}}$ were shown in Table I and Table II, respectively. "Accurate" refers to the optimal kernel selected by minimizing the accurate kernel selection criterion using the original kernel matrix. "MCM" and "Nyström" denote the MCM kernel selection algorithm and the Nyström kernel selection algorithm.

We used the Wilcoxon signed rank test, which is recommended for comparisons over multiple datasets [61], to estimate the statistical significance of differences in performance. According to this test, "Accurate" was statistically superior to neither "MCM" nor "Nyström" on $\mathcal{C}_{\mathrm{ree}}$ and $\mathcal{C}_{\mathrm{ipe}}$ at the 95% level of significance. Meanwhile, Table I and Table II also showed that approximate algorithms were much faster than the accurate ones, especially the "MCM" approximate kernel selection. In summary, the proposed approximate algorithms can significantly reduce the computational cost of kernel selection, while at the same time exhibiting a competitive performance.

TABLE I
COMPARISON OF THE MEAN SQUARE ERRORS (MSE) OR THE MEAN CLASSIFICATION ERROR RATES (MCER), THE TIME (SECONDS) AND THE SPEEDUPS (ACCURATETIME/MCMTIME OR ACCURATETIME/NYSTRÖMTIME) BETWEEN THE ACCURATE AND THE PROPOSED APPROXIMATE KERNEL SELECTION ALGORITHMS FOR $\mathcal{C}_{\mathrm{ree}}$.

regression (♯ samples)	Accurate		MCM		Nyström	
	MSE	Time (sec)	MSE	Speedup	MSE	Speedup
housing (506)	2.79±0.44(1e1)	4.69(1e-2)	2.83±0.35(1e1)	25.35	2.81±0.44(1e1)	1.73
balance (625)	2.42±0.15(1e-1)	5.31(1e-2)	2.42±0.15(1e-1)	41.81	2.41±0.14(1e-1)	2.01
allbp (840)	5.50±0.64(1e-2)	1.59(1e-1)	5.50±0.64(1e-2)	57.40	5.60±0.68(1e-2)	2.45
veh (846)	5.89±0.27(1e-1)	1.65(1e-1)	6.15±0.31(1e-1)	89.18	5.98±0.32(1e-1)	2.48
pha (1070)	5.93±0.35(1e-1)	3.63(1e-1)	5.95±0.36(1e-1)	131.04	5.92±0.29(1e-1)	3.33
ann1 (1131)	8.84±0.79(1e-2)	4.41(1e-1)	8.84±0.79(1e-2)	153.12	8.94±0.86(1e-2)	3.29
mha (1269)	1.16±0.12(1e-1)	4.58(1e-1)	1.16±0.12(1e-1)	165.34	1.11±0.13(1e-1)	3.31
DNA-n (1275)	5.04±0.20(1e-1)	5.62(1e-1)	5.06±0.17(1e-1)	197.19	5.04±0.20(1e-1)	3.30
adp (1351)	8.50±0.74(1e-2)	4.48(1e-1)	8.50±0.74(1e-2)	197.35	7.86±0.86(1e-2)	3.50
mg (1385)	2.06±0.08(1e-2)	3.58(1e-1)	2.07±0.10(1e-2)	186.45	1.93±0.10(1e-2)	3.25
space-ga (3107)	2.04±0.21(1e-2)	4.87(1e0)	2.04±0.21(1e-2)	1305.63	1.98±0.22(1e-2)	5.05
abalone (4177)	5.97±0.22(1e0)	1.02(1e1)	5.97±0.22(1e0)	2434.36	5.59±0.26(1e0)	5.45
classification (♯ samples)	MCER	Time (sec)	MCER	Speedup	MCER	Speedup
ionosphere (351)	5.80%	1.31(1e-2)	6.90%	28.47	6.06%	1.29
colic (368)	23.70%	1.08(1e-2)	18.80%	21.60	19.40%	1.08
vote (435)	4.89%	4.23(1e-2)	9.68%	36.78	4.89%	2.10
breast-cancer (683)	2.96%	4.00(1e-2)	3.06%	34.78	3.06%	1.81
australian (690)	14.30%	1.47(1e-1)	14.00%	72.05	14.00%	2.39
pima (768)	25.90%	1.19(1e-1)	23.70%	98.34	23.70%	2.19
diabetes (768)	26.50%	1.73(1e-1)	23.70%	83.17	23.80%	2.46
german (1000)	27.50%	1.70(1e-1)	25.60%	74.88	25.60%	2.58
a1a (1605)	17.60%	8.17(1e-1)	17.40%	504.32	17.40%	3.68
a2a (2265)	18.60%	2.00(1e0)	18.30%	826.44	18.30%	4.20

VIII. CONCLUSION

In this paper, we proposed a novel approach for kernel selection based on kernel matrix approximation. We theoretically justified the introduction of kernel matrix approximation into kernel selection by defining and analyzing the approximate consistency, which provides the foundations for approximate kernel selection. Under the theoretical guarantee of the approximate consistency, we designed approximate algorithms for kernel selection by exploiting the computational virtues of MCM and Nyström approximation, whose computational complexities are quasi-linear or linear in the number of samples and significantly lower than the accurate approaches. The approximate consistency of different kernel matrix approximation algorithms under two error-minimization criteria was empirically verified and the results showed that even for a not-so-large number of samples, the phenomenon of the consistency between the approximate and accurate criteria was present. Furthermore, the effectiveness experiments demonstrated that the approximate algorithms can significantly improve the kernel selection efficiency as compared to the accurate algorithms, without sacrificing predictive performance.

ACKNOWLEDGMENTS

This work was supported in part by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. URF/1/3007-01-01 and BAS/1/1624-01-01, National Natural Science Foundation of China (No. 61703396), National Natural Science Foundation of China (No. 61673293), the CCF-Tencent Open Fund and Shenzhen Government (GJHZ20180419190732022).

REFERENCES

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler, "Scale-sensitive dimensions, uniform convergence, and learnability," *Journal of the ACM*, vol. 44, no. 4, pp. 615–631, 1997.

[2] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bulletin of the American Mathematical Society*, vol. 39, no. 1, pp. 1–49, 2002.

[3] E. Alpaydin, *Introduction to Machine Learning*. Cambridge, MA: MIT Press, 2004.

[4] T. M. Mitchell, *Machine Learning*. New York, NY: McGraw Hill, 1997.

[5] X. Tian, Y. Li, T. Liu, X. Wang, and D. Tao, "Eigenfunction-based multitask learning in a reproducing kernel Hilbert space," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 6, pp. 1818–1830, 2019.

[6] B. Nguyen and B. De Baets, "Kernel-based distance metric learning for supervised k-means clustering," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 10, pp. 3084–3095, 2019.

[7] F. Liu, X. Huang, C. Gong, J. Yang, and J. A. K. Suykens, "Indefinite kernel logistic regression with concave-inexact-convex procedure," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 3, pp. 765–776, 2019.

[8] L. Ding, Z. Liu, Y. Li, S. Liao, Y. Liu, P. Yang, G. Yu, L. Shao, and X. Gao, "Linear kernel tests via empirical likelihood for high-dimensional data," in *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI)*, 2019, pp. 3454–3461.

[9] L. Ding, M. Yu, L. Liu, F. Zhu, Y. Liu, Y. Li, and L. Shao, "Two generator game: Learning to sample via linear goodness-of-fit test," in *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.

[10] B. Schölkopf and A. Smola, *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. Cambridge, MA, USA: MIT Press, 2002.

[11] S. Zhou, "Sparse LSSVM in primal using Cholesky factorization for large-scale problems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 4, pp. 783–795, 2016.

[12] Y. Xu, Z. Yang, and X. Pan, "A novel twin support-vector machine with pinball loss," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 2, pp. 359–370, 2017.

[13] Y. Liu, S. Jiang, and S. Liao, "Efficient approximation of cross-validation for kernel methods using Bouligand influence function," in *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 2014, pp. 324–332.

[14] C. A. Micchelli and M. Pontil, "Learning the kernel function via

TABLE II
COMPARISON OF THE MEAN SQUARE ERRORS (MSE) OR THE MEAN CLASSIFICATION ERROR RATES (MCER), THE TIME (SECONDS) AND THE SPEEDUPS (ACCURATETIME/MCMTIME OR ACCURATETIME/NYSTRÖMTIME) BETWEEN THE ACCURATE AND THE PROPOSED APPROXIMATE KERNEL SELECTION ALGORITHMS FOR $\mathcal{C}_{\mathrm{ipe}}$.

regression (♯ samples)	Accurate		MCM		Nyström	
	MSE	Time (sec)	MSE	Speedup	MSE	Speedup
housing (506)	2.80±0.62(1e1)	3.15(1e-1)	2.63±0.43(1e1)	157.50	2.59±0.44(1e1)	15.29
balance (625)	2.90±0.54(1e-1)	3.13(1e-1)	2.46±0.17(1e-1)	118.11	2.47±0.15(1e-1)	14.22
allbp (840)	6.03±0.57(1e-2)	1.26(1e0)	5.33±0.47(1e-2)	450.00	5.23±0.46(1e-2)	24.46
veh (846)	6.34±0.84(1e-1)	1.39(1e0)	6.03±0.41(1e-1)	327.83	5.86±0.46(1e-1)	24.51
pha (1070)	6.07±0.55(1e-1)	5.70(1e0)	6.06±0.37(1e-1)	937.50	6.29±0.43(1e-1)	40.14
ann1 (1131)	9.40±0.92(1e-2)	5.49(1e0)	9.40±0.92(1e-2)	1003.46	9.22±1.03(1e-2)	36.84
mha (1269)	1.07±0.14(1e-1)	5.50(1e0)	1.04±0.07(1e-1)	940.17	1.04±0.07(1e-1)	39.85
DNA-n (1275)	5.68±0.20(1e-1)	7.05(1e0)	5.01±0.14(1e-1)	1159.53	5.00±0.14(1e-1)	44.90
adp (1351)	8.52±0.98(1e-2)	5.47(1e0)	7.72±0.45(1e-2)	1226.45	7.72±0.45(1e-2)	39.63
mg (1385)	1.91±0.06(1e-2)	5.41(1e0)	1.94±0.05(1e-2)	1134.17	1.91±0.06(1e-2)	35.35
space-ga (3107)	2.03±0.20(1e-2)	1.30(1e2)	2.03±0.19(1e-2)	7647.05	2.03±0.20(1e-2)	94.89
abalone (4177)	5.50±0.18(1e0)	3.49(1e2)	5.50±0.18(1e0)	23266.66	5.50±0.18(1e0)	121.18
classification (♯ samples)	MCER	Time (sec)	MCER	Speedup	MCER	Speedup
ionosphere (351)	7.04%	6.23(1e-2)	7.61%	31.78	6.73%	6.42
colic (368)	25.10%	6.46(1e-2)	19.00%	39.87	20.60%	7.24
vote (435)	7.03%	3.09(1e-1)	5.98%	109.96	5.75%	14.23
breast-cancer (683)	3.81%	3.14(1e-1)	3.47%	103.28	3.47%	16.16
australian (690)	15.20%	1.29(1e0)	13.30%	479.55	13.30%	25.09
pima (768)	31.40%	1.65(1e0)	23.70%	455.80	24.40%	23.17
diabetes (768)	27.20%	1.44(1e0)	23.30%	360.00	23.50%	24.65
german (1000)	28.10%	1.46(1e0)	25.40%	403.31	25.00%	24.49
a1a (1605)	19.50%	1.43(1e1)	17.20%	3264.84	17.20%	60.59
a2a (2265)	20.00%	4.50(1e1)	18.30%	8181.81	18.30%	87.04

regularization," *Journal of Machine Learning Research*, vol. 6, pp. 1099–1125, 2005.

[15] Y. Liu, H. Lin, L. Ding, W. Wang, and S. Liao, "Fast cross-validation," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, 2018, pp. 2497–2503.

[16] Y. Liu, S. Liao, S. Jiang, L. Ding, H. Lin, and W. Wang, "Fast cross-validation for kernel-based algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[17] P. L. Bartlett, S. Boucheron, and G. Lugosi, "Model selection and error estimation," *Machine Learning*, vol. 48, no. 1–3, pp. 85–113, 2002.

[18] D. Anguita, A. Ghio, L. Oneto, and S. Ridella, "In-sample and out-of-sample model selection and error estimation for support vector machines," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 9, pp. 1390–1406, 2012.

[19] C. Cortes, M. Kloft, and M. Mohri, "Learning kernels using local Rademacher complexity," in *Advances in Neural Information Processing Systems 26*, 2013, pp. 2760–2768.

[20] J. Li, Y. Liu, R. Yin, H. Zhang, L. Ding, and W. Wang, "Multi-class learning: from theory to algorithm," in *Advances in Neural Information Processing Systems 31*, 2018, pp. 1591–1600.

[21] D. You, C. F. Benitez-Quiroz, and A. M. Martinez, "Multiobjective optimization for model selection in kernel methods in regression," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 10, pp. 1879–1893, 2014.

[22] L. Ding and S. Liao, "Model selection with the covering number of the ball of RKHS," in *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management (CIKM)*, 2014, pp. 1159–1168.

[23] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Machine Learning*, vol. 46, no. 1–3, pp. 131–159, 2002.

[24] G. Cawley and N. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research*, vol. 11, pp. 2079–2107, 2010.

[25] L. Ding, S. Liao, Y. Liu, P. Yang, and X. Gao, "Randomized kernel selection with spectra of multilevel circulant matrices," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, 2018, pp. 2910–2917.

[26] G. H. Song and Y. S. Xu, "Approximation of high-dimensional kernel matrices by multilevel circulant matrices," *Journal of Complexity*, vol. 26, no. 4, pp. 375–405, 2010.

[27] G. H. Song, "Approximation of kernel matrices in machine learning," Ph.D. dissertation, Syracuse University, Syracuse, NY, USA, 2010.

[28] L. Ding and S. Liao, "Approximate model selection for large scale LSSVM," *Journal of Machine Learning Research - Proceedings Track*, vol. 20, pp. 165–180, 2011.

[29] R. Yin, Y. Liu, W. Wang, and D. Meng, "Sketch kernel ridge regression using circulant matrix: Algorithm and theory," *IEEE Transactions on Neural Networks and Learning Systems*, 2019.

[30] C. K. I. Williams and M. Seeger, "Using the Nyström method to speed up kernel machines," *Advances in Neural Information Processing Systems*, vol. 13, pp. 682–688, 2001.

[31] P. Drineas and M. W. Mahoney, "On the Nyström method for approximating a Gram matrix for improved kernel-based learning," *Journal of Machine Learning Research*, vol. 6, pp. 2153–2175, 2005.

[32] S. Kumar, M. Mohri, and A. Talwalkar, "Sampling methods for the Nyström method," *Journal of Machine Learning Research*, vol. 13, pp. 981–1006, 2012.

[33] K. Zhang and J. T. Kwok, "Clustered Nyström method for large scale manifold learning and dimension reduction," *IEEE Transactions on Neural Networks*, vol. 21, no. 10, pp. 1576–1587, 2010.

[34] A. Gittens and M. W. Mahoney, "Revisiting the Nyström method for improved large-scale machine learning," in *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013, pp. 567–575.

[35] L. Lan, Z. Wang, S. Zhe, W. Cheng, J. Wang, and K. Zhang, "Scaling up kernel SVM on limited resources: A low-rank linearization approach," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 2, pp. 369–378, 2019.

[36] S. Wang and Z. Zhang, "Improving CUR matrix decomposition and the Nyström approximation via adaptive sampling," *Journal of Machine Learning Research*, vol. 14, pp. 2729–2769, 2013.

[37] ——, "Efficient algorithms and error analysis for the modified Nyström method," in *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2014, pp. 996–1004.

[38] F. Bach, "Sharp analysis of low-rank kernel matrix approximations," in *Proceedings of the 26th Annual Conference on Learning Theory (COLT)*, 2013, pp. 185–209.

[39] S. Fine and K. Scheinberg, "Efficient SVM training using low-rank kernel representations," *Journal of Machine Learning Research*, vol. 2, pp. 243–264, 2002.

[40] A. J. Smola and B. Schölkopf, "Sparse greedy matrix approximation for machine learning," in *Proceedings of the 17th International Conference on Machine Learning (ICML)*, 2000, pp. 911–918.

[41] X. Chang, Y. Zhong, Y. Wang, and S. Lin, "Unified low-rank matrix estimate via penalized matrix least squares approximation," *IEEE Trans-*

actions on Neural Networks and Learning Systems, vol. 30, no. 2, pp. 474–485, 2019.

[42] P. Drineas, M. W. Mahoney, and S. Muthukrishnan, "Relative-error CUR matrix decompositions," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 2, pp. 844–881, 2008.

[43] S. Wang and Z. Zhang, "A scalable CUR matrix decomposition algorithm: Lower time complexity and tighter bound," *Advances in Neural Information Processing Systems*, vol. 24, pp. 647–655, 2012.

[44] M. W. Mahoney and P. Drineas, "CUR matrix decompositions for improved data analysis," *Proceedings of the National Academy of Sciences*, vol. 106, no. 3, pp. 697–702, 2009.

[45] C. Cortes, M. Mohri, and A. Talwalkar, "On the impact of kernel approximation on learning accuracy," in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010, pp. 113–120.

[46] T. B. Yang, Y. F. Li, M. Mahdavi, R. Jin, and Z. H. Zhou, "Nyström method vs random Fourier features: A theoretical and empirical comparison," *Advances in Neural Information Processing Systems*, vol. 24, pp. 1060–1068, 2012.

[47] R. Jin, T. B. Yang, M. Mahdavi, Y. F. Li, and Z. H. Zhou, "Improved bounds for the Nyström method with application to kernel classification," *IEEE Transactions on Information Theory*, vol. 5, no. 10, pp. 6939–6949, 2013.

[48] L. Ding and S. Liao, "Approximate consistency: Towards foundations of approximate kernel selection," in *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Database (ECML PKDD)*, 2014, pp. 354–369.

[49] L. Ding, Y. Liu, S. Liao, Y. Li, P. Yang, Y. Pan, C. Huang, L. Shao, and X. Gao, "Approximate kernel selection with strong approximate consistency," in *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI)*, 2019, pp. 3462–3469.

[50] L. Ding and S. Liao, "Nyström approximate model selection for LSSVM," in *Advances in Knowledge Discovery and Data Mining - 16th Pacific-Asia Conference (PAKDD)*, 2012, pp. 282–293.

[51] ——, "An approximate approach to automatic kernel selection," *IEEE Transactions on Cybernetics*, vol. 47, no. 3, pp. 554–565, 2017.

[52] Y. Xu and H. Zhang, "Refinement of reproducing kernels," *Journal of Machine Learning Research*, vol. 10, pp. 107–140, 2009.

[53] G. S. Kimeldorf and G. Wahba, "A correspondence between Bayesian estimation on stochastic processes and smoothing by splines," *Annals of Mathematical Statistics*, vol. 41, no. 2, pp. 495–502, 1970.

[54] P. J. Davis, *Circulant Matrices*. New York, NY, USA: John Wiley & Sons, 1979.

[55] E. E. Tyrtyshnikov, "A unifying approach to some old and new theorems on distribution and clustering," *Linear Algebra and its Applications*, vol. 232, pp. 1–43, 1996.

[56] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization networks and support vector machines," *Advances in Computational Mathematics*, vol. 13, no. 1, pp. 1–50, 2000.

[57] C. Boutsidis, P. Drineas, and M. Magdon-Ismail, "Near optimal column-based matrix reconstruction," in *Proceedings of the IEEE 52nd Annual Symposium on Foundations of Computer Science (FOCS)*, 2011, pp. 305–314.

[58] A. Deshpande, L. Rademacher, S. Vempala, and G. Wang, "Matrix approximation and projective clustering via volume sampling," in *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithm (SODA)*, 2006, pp. 1117–1126.

[59] R. Singleton, "An algorithm for computing the mixed radix fast fourier transform," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 2, pp. 93–103, 1969.

[60] S. Kumar, M. Mohri, and A. Talwalkar, "On sampling-based approximate spectral decomposition," in *Proceedings of the 26th International Conference on Machine Learning (ICML)*, 2009, pp. 553–560.

[61] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.