Scalable Kernel k-Means with Randomized Sketching: From Theory to Algorithm

Rong Yin, Yong Liu*, Weiping Wang, and Dan Meng

Abstract—Kernel k-means is a fundamental unsupervised learning in data mining. Its computational requirements are typically at least quadratic in the number of data, which are prohibitive for large-scale scenarios. To address these issues, we propose a novel randomized sketching approach SKK based on the circulant matrix. SKK projects the kernel matrix left and right according to the proposed sketch matrices to obtain a smaller one and accelerates the matrix-matrix product by the fast Fourier transform based on the circulant matrix, which can greatly reduce the computational requirements of the approximate kernel k-means estimator with the same generalization bound as the exact kernel k-means in the statistical setting. In particular, theoretical analysis shows that taking the sketch dimension of $\sqrt{n}$ is sufficient for SKK to achieve the optimal excess risk bound with only a fraction of computations, where n is the number of data. The extensive experiments verify our theoretical analysis, and SKK achieves the state-of-the-art performances on 12 real-world datasets. To the best of our knowledge, in randomized sketching, this is the first time that unsupervised learning makes such a significant breakthrough.

Index Terms—kernel k-means, randomized sketching, statistical and computational trade-offs, excess risk bound.

1 Introduction

K-means clustering, a popular nonparametric approach in the knowledge and data engineering community, divides the datasets into dissimilar groups according to the distance between data points [1], [2], [3]. The kernel version of k-means projects data points into a high-dimensional non-linear manifold, which makes clusters more easily separated [4], [5], [6], [7], [8]. Kernel k-means has been applied in various practical applications and made remarkable achievements [9], [10], [11], [12], [13], [14], [15], [16]. However, with the increasing of the size of datasets, the computational requirements are prohibitive, typically at least quadratic in the number of data.

To overcome these limitations, many researchers have made various explorations. The main popular approaches are as follows. Parallel computing divides clustering tasks into several small computational parts and then distributed processes them so as to reduce time complexity [17], [18], [19]. Random features represent the data points in Hilbert space explicitly and approximately [20], [21], [22], [23], and the dimension of the approximate data points is much smaller than the original data dimension in Hilbert space, which can reduce the computational requirements. Nyström is sampling a subset of training set points (landmarks) to approximate the kernel matrix [24], [25], [26], [27], [28], [29], [30], [31], [32], [33]. Incremental clustering algorithms utilize progressive calculation to accelerate the computing speed [34], [35]. And randomized sketching [25], [36] provides the projection strategy to approximate the large-scale kernel matrix so as to reduce the computational requirements. From a theoretical perspective, the key strategy is to characterize statistical and computational trade-offs, that is if, or under which conditions, computational gains come at the expense of statistical accuracy. Various studies have shown that randomized sketching successfully projects large matrices into smaller matrices, which is more efficient and proved to maintain satisfactory accuracy [25], [36], [37], [38], [39], [40]. However, the existing randomized sketching kernel k-means is still prohibitive for large-scale scenarios and fails to obtain the optimal excess risk bound [25], [36].

In this paper, we quantify the efficiency of approximate kernel k-means from the perspective of theoretical analysis and computational requirements. Arguably, an approximate estimator may incur some accuracy loss. In fact, our theoretical analysis proves that there is a favorable mechanism for keeping optimal statistical accuracy with significantly reducing computational requirements. The phenomenon was also shown in supervised learning [41], [42], [43], [44], [45].

The proposed approach SKK considers a randomized sketching to kernel k-means based on the circulant matrix [46], [47], [48], [49], which significantly accelerates the calculation of kernel k-means, reduces the space complexity, and has the optimal excess risk bound. By constructing the randomized sketching matrices and utilizing the fast Fourier transform (FFT) based on the circulant matrix, SKK can quickly obtain a smaller approximate kernel matrix and greatly reduce the computational cost during the process of iteration in kernel k-means clustering. From a computational point of view, while maintaining the optimal excess risk bound, the proposed approach obtains the time complexity $\mathcal{O}(nkt + n\log\sqrt{n})$ and space complexity $\mathcal{O}(n)$, which are the optimal compared to the existing state-of-the-art approximate kernel k-means estimators, where n, k, and t represent the number of data points, clusters, and iteration in kernel k-means clustering, respectively. The statistical analysis shows that SKK maintains the optimal excess risk bound $\tilde{\mathcal{O}}(\sqrt{k/n})$[1] with only the sketch dimension of $\sqrt{n}$. In this case, the corresponding computational costs are greatly reduced. To the best of our knowledge, in

- R. Yin, W. Wang and D. Meng are with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China.
E-mail: yinrong@iie.ac.cn; wangweiping@iie.ac.cn; mengdan@iie.ac.cn
- Y. Liu is with Renmin University of China, Beijing 100872, China and corresponding author.

[1] $\tilde{\mathcal{O}}$ hides logarithmic term.

unsupervised learning, this is the first optimal generalization bound for randomized sketching kernel k-means estimators. Most importantly, the empirical performances are thoroughly tested on real-world datasets. An extensive experimental analysis indicates that SKK can keep the optimal state compared to the existing state-of-the-art approximate kernel k-means estimators on most problems both in time consumption and prediction accuracy. The main contributions are as follows:

1) A novel randomized sketching kernel k-means approach is proposed, which utilizes the circulant matrix to get a smaller variant kernel matrix, thus bringing the benefits of computational requirements.

2) An optimal excess risk bound $\tilde{\mathcal{O}}(\sqrt{k/n})$ for randomized sketching kernel k-means is achieved with the sketch dimension $\sqrt{n}$. To the best of our knowledge, this is the first time that kernel k-means with randomized sketching achieves such an optimal statistical accuracy (see Theorem 4).

3) Compared to the state-of-the-art approximate estimators in kernel k-means, while keeping the optimal statistical accuracy, the proposed approach achieves the time complexity $\mathcal{O}(nkt + n\log\sqrt{n})$ and space complexity $\mathcal{O}(n)$, which are the optimal (see RELATED WORK).

4) The empirical performances on 12 real-world datasets shows that SKK significantly outperforms the state-of-the-art approximate kernel k-means estimators in terms of time, while maintaining satisfactory accuracy.

In sections 2 and 3, we introduce the related work and the background of kernel k-means. In the following section, the approximate kernel k-means estimator SKK is proposed. In section 5, the statistical guarantees and the related discussion of the proposed approach are presented. Finally, the extensive experiments, proof, and conclusion are presented.

2 Related Work

To deal with the computational requirements bottleneck of kernel k-means, practical approximate kernel k-means estimators are developed [17], [22], [24], [25], [27], [36]. Although there have been many studies on the approximate kernel k-means estimators, the excess risk bound is mainly obtained in the papers [36], [27], and [24]. For example, the paper [25] established the $1+\varepsilon$ relative-error bound for Nyström approximation to kernel k-means instead of excess risk bound, where $\varepsilon \in (0,1)$. The proposed approach in this paper is based on randomized sketching to accelerate the computation and provides the excess risk guarantees for approximate kernel k-means. Therefore, in this part, we mainly introduce the most relevant approximate kernel k-means with excess risk guarantees: [36] based on randomized sketching and [24], [27] based on Nyström.

In the paper [36], a typical randomized sketching approximation approach, called PKK, was proposed in kernel clustering. It constructed an unstructured matrix to act as a sketch matrix, which is used to project data one-time. PKK extracted features from the data in Hilbert space by the sketch matrix, then performed clustering operations on them. As its sketch matrix is unstructured, PKK cannot accelerate kernel k-means by FFT. Meantime, the scale of the projected data matrix is still large so that the computational requirements are still high. PKK assumed that data in Hilbert space are explicit and infinite-dimensional, but did not give any specific expression. For analyzing its complexity, one can use the columns of the kernel matrix, whose dimension is obviously smaller than that of data in Hilbert space [50], as the object of randomized sketching. The space and time consumption of PKK is $\mathcal{O}\left(n^2\right)$ and $\mathcal{O}\left(nmkt + n^2m\right)$ with the excess risk bound $\mathcal{O}(k/\sqrt{n})$, under the sketch dimension[2] $m = \tilde{\Omega}(\log(n)/\varepsilon^2)$. Compared to it, the proposed approach improves the excess risk bound from $\mathcal{O}(k/\sqrt{n})$ of PKK to the optimal $\tilde{\mathcal{O}}(\sqrt{k/n})$, reduces the space complexity by a factor of n, and reduces the time complexity by a factor of $\frac{kt+n}{kt+\log\sqrt{n}} \cdot \frac{\log n}{\varepsilon^2}$.

The state-of-the-art Nyström approach to kernel k-means [27] employed the uniform sampling technique to obtain the Nyström landmarks that can be used to approximate the kernel matrix. Combined with probabilistic results to show that excess risk bound $\tilde{\mathcal{O}}(k/\sqrt{n})$ can be obtained considering $\tilde{\Omega}(\sqrt{n})$ Nyström landmark points. The corresponding space complexity is $\mathcal{O}\left(n\sqrt{n}\right)$ and the time complexity is $\mathcal{O}\left(n\sqrt{n}kt + n^2\right)$. Compared to [27], the proposed approach improves the excess risk bound from $\tilde{\mathcal{O}}(k/\sqrt{n})$ to the optimal $\tilde{\mathcal{O}}(\sqrt{k/n})$. In computational requirements, the proposed approach reduces the time complexity by a factor of $\frac{\sqrt{n}kt+n}{kt+\log\sqrt{n}}$ and space complexity by a factor of $\sqrt{n}$ respectively, while maintaining the better excess risk bound.

Subsequently, based on the Nyström approach in [27], Liu et al. [24] further improved the excess risk bound from $\tilde{\mathcal{O}}(k/\sqrt{n})$ to the optimal $\tilde{\mathcal{O}}(\sqrt{k/n})$, which is linearly dependent on $\sqrt{k}$ instead of k, with the number of Nyström landmark points $\tilde{\Omega}(\sqrt{nk})$. The corresponding space and time complexity is $\mathcal{O}\left(n\sqrt{nk}\right)$ and $\mathcal{O}\left(nk\sqrt{nk}t + n^2k\right)$. Although Liu et. al obtain the optimal excess risk bound for kernel k-means, the corresponding landmarks increased to $\tilde{\Omega}(\sqrt{nk})$ which causes the higher time and memory costs. Compared to [24], the proposed approach reduces the space and time cost by factors of $\sqrt{nk}$ and $\frac{\sqrt{nk}kt+nk}{kt+\log\sqrt{n}}$ at the same optimal statistical accuracy. Therefore, the existing approaches need to be further improved in terms of computational requirements with high theoretical accuracy.

In this paper, a novel approximate kernel k-means estimator is proposed, which is based on the randomized sketching techniques and the structured matrix to approximate kernel k-means with high computation gains and sound statistical guarantees. More precisely, we utilize the sketch matrices to project the kernel matrix left and right to obtain a smaller approximate kernel matrix. In addition, the proposed sketch matrices are based on the circulant matrix, which is a structured matrix, to process data. In terms of time complexity of the circulant matrix, it can realize loglinear computation in the matrix-matrix product by FFT. In terms of space complexity, it can realize linear storage due to the structural characteristics. By carefully constructing the sketch matrix and randomized sketching approach for kernel k-means clustering, SKK only costs space complexity of $\mathcal{O}\left(n\right)$ and time complexity of $\mathcal{O}\left(nkt + n\log\sqrt{n}\right)$ for approximate kernel k-means estimator with the optimal excess risk bound, which significantly outperforms other randomized sketching and the classical Nyström approaches. Compared to the exact kernel k-means, SKK reduces the space requirement from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$, and the time requirement from $\mathcal{O}(n^2kt)$ to $\mathcal{O}(nkt + n\log\sqrt{n})$. From a theoretical perspective, this paper shows a provable guarantee for the proposed approach SKK. Given the sketch dimension of $\sqrt{n}$, SKK obtains the same statistical accuracy of the exact kernel k-means, improving the excess risk bound by a factor of $\sqrt{k}$ compared to [36] and [27]. SKK is the first

[2] $\tilde{\Omega}$ hides logarithmic term.

TABLE 1
Comparison of the classical approximate kernel k-means approaches. The third and fourth columns represent the space and time complexity of each approach. The fifth and sixth columns correspond to the excess risk bounds and the corresponding m. m denotes the sketch dimension in randomized sketching approaches such as PKK and SKK, and Nyström landmark points in Nyström approaches such as NKK- and NKK. n and k represent the number of data points and clusters respectively. t represents the number of iteration in kernel k-means. $\varepsilon \in (0,1)$.

Reference	Approach	Space	Time	Excess Risk Bound	m
Kernel k-Means	KK	$\mathcal{O}\left(n^2\right)$	$\mathcal{O}\left(n^2kt\right)$	$\tilde{\mathcal{O}}\left(\sqrt{\frac{k}{n}}\right)$	/
[36]	PKK	$\mathcal{O}\left(n^2\right)$	$\mathcal{O}\left((nkt+n^2)\cdot\frac{\log n}{\varepsilon^2}\right)$	$\mathcal{O}\left(\frac{k}{\sqrt{n}}\right)$	$\frac{\log n}{\varepsilon^2}$
[27]	NKK-	$\mathcal{O}\left(n\sqrt{n}\right)$	$\mathcal{O}\left(nkt\sqrt{n}+n^2\right)$	$\tilde{\mathcal{O}}\left(\frac{k}{\sqrt{n}}\right)$	$\sqrt{n}$
[24]	NKK	$\mathcal{O}\left(n\sqrt{nk}\right)$	$\mathcal{O}\left(nkt\sqrt{nk}+n^2k\right)$	$\tilde{\mathcal{O}}\left(\sqrt{\frac{k}{n}}\right)$	$\sqrt{nk}$
This Paper	SKK	$\mathcal{O}\left(n\right)$	$\mathcal{O}\left(nkt+n\log\sqrt{n}\right)$	$\tilde{\mathcal{O}}\left(\sqrt{\frac{k}{n}}\right)$	$\sqrt{n}$

to achieve the optimal excess risk bound with only a fraction of computations in randomized sketching kernel k-means. The detailed time complexity, space complexity, excess risk bound, and the corresponding value of m of the classical approximate kernel k-means estimators mentioned above are summarized in TABLE 1. The empirical results on 12 common datasets indicate that SKK outperforms other approximate approaches, which verify our theoretical analysis.

3 Background

3.1 Notation

Given a fixed but unknown probability distribution μ on the input space $\mathcal{X}$ and a feature map $\varphi(\cdot):\mathcal{X}\to\mathcal{H}$, one draws a sample $\mathbf{X}=\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ i.i.d from μ and maps $\mathcal{X}$ into a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ [50], [51] by $\varphi(\cdot)$. Therefore, one has $\phi=\varphi(\mathbf{x})$ for any $\mathbf{x}\in\mathcal{X}$. The n clustering data $\phi_1,\ldots,\phi_n$ are independent in a separable Hilbert space $\mathcal{H}$ with distribution μ. By the kernel trick one has that $\langle\phi_i,\phi_j\rangle=\mathcal{K}(\mathbf{x}_i,\mathbf{x}_j)$, $\mathcal{K}$ is the kernel function associated with $\mathcal{H}$, $\langle .,.\rangle$ denotes the notation of inner product. We denote with $\mathbf{K}_{ij}=\langle\phi_i,\phi_j\rangle=\mathcal{K}(\mathbf{x}_i,\mathbf{x}_j)$ the kernel matrix.

3.2 Kernel k-Means

The key idea of k-means clustering is to obtain k clustering centroid $\mathbf{c}$, then partition the data $\phi_1,\ldots,\phi_n$ into the k clusters according to the similarity with c_j. This can be described by the following mathematical expression:

$$W(\mathbf{c},\mu_n)=\frac{1}{n}\sum_{i=1}^{n}\min_{j\in[k]}\parallel \phi_i-c_j\parallel^2, \tag{1}$$

where $\mathbf{c}=\{c_1,\ldots,c_k\}\in\mathcal{H}^k$ denote the clustering centers, μ_n denotes the empirical distribution of the data. That is to say, if the least empirical squared norm of point ϕ_i corresponding to the clustering center c_j, ϕ_i belongs to the j-th cluster.

The quality of a clustering scheme is usually measured by the clustering risk [27], [36], the form of which is as follows.

Definition 1 (**[27], [36]**). *Define the clustering risk as*

$$W(\mathbf{c},\mu)=\int\min_{j\in[k]}\parallel \phi-c_j\parallel^2\,d\mu(\phi). \tag{2}$$

For introducing the theoretical target, we first show the notions of the empirical risk minimizer and the optimal clustering risk.

Definition 2 (**[36]**). *Define the empirical risk minimizer (ERM) as*

$$\mathbf{c}_n=\arg\min_{\mathbf{c}\in\mathcal{H}^k}W(\mathbf{c},\mu_n). \tag{3}$$

Definition 3 (**[27]**). *Define the optimal clustering risk as*

$$W^*(\mu)=\inf_{\mathbf{c}\in\mathcal{H}^k}W(\mathbf{c},\mu). \tag{4}$$

From a theoretical perspective, this paper aims to bound the excess risk $\mathcal{E}(\mathbf{c}_n)$ of the ERM:

$$\mathcal{E}(\mathbf{c}_n)=\mathbb{E}\left[W(\mathbf{c}_n,\mu)\right]-W^*(\mu).$$

3.3 The Lower and Upper Bounds of Kernel k-Means

The upper bound on excess risk of the ERM to kernel k-means is (nearly) the same as the lower bound, which means that the upper bound is (nearly) the optimal excess risk bound to kernel k-means. In the following, we introduce the detail of the excess risk bound.

Theorem 1 (**Lower Bound [8]**). *There exists $\|\phi\|\leq 1$ and a constant C such that*

$$\mathbb{E}\left[W(\mathbf{c},\mu)\right]-W^*(\mu)\geq C\sqrt{\frac{k^{1-4/d}}{n}}, \tag{5}$$

where d is the dimension of ϕ.

In general, the dimension d of ϕ is very large or even infinite. Therefore, Theorem 1 shows that the lower bound on excess risk to kernel k-means is $\Omega\left(\sqrt{\frac{k}{n}}\right)$.

For a long time, the excess risk of ERM to kernel k-means kept the following upper bound [27], [36]:

$$\mathbb{E}\left[W(\mathbf{c}_n,\mu)\right]-W^*(\mu)\leq C\frac{k}{\sqrt{n}}=\mathcal{O}\left(\frac{k}{\sqrt{n}}\right),$$

where C is a constant, which is linearly related to the number of clusters k.

Recently, Liu et al. [24] further improved the upper bound to kernel k-means, which is shown as below.

Theorem 2 (**Upper bound [24]**). *Given $\delta \in (0,1)$ and $\|\phi\| \le 1$. We have, with the probability at least $1-\delta$,*

$$\begin{aligned}&\mathbb{E}\left[W(\mathbf{c}_n,\mu)\right]-W^*(\mu)\\ \le & C\sqrt{\frac{k}{n}}\log^{\frac{3}{2}+\frac{\delta}{2}}\left(\frac{\sqrt{n}}{3}\right)+C\sqrt{\frac{\log\frac{1}{\delta}}{n}}\\ =&\tilde{\mathcal{O}}\left(\sqrt{\frac{k}{n}}\right),\end{aligned} \tag{6}$$

where C is a constant.

Theorem 2 shows that the upper excess risk bound of kernel k-means can reach $\tilde{\mathcal{O}}\left(\sqrt{\frac{k}{n}}\right)$, that is, it improves the upper bound from k-related to $\sqrt{k}$-related. Combining Theorem 1 and Theorem 2, we know that the lower excess risk bound and upper excess risk bound of kernel k-means are of the same order, which indicates that $\tilde{\mathcal{O}}\left(\sqrt{\frac{k}{n}}\right)$ is the (nearly) optimal excess risk bound to kernel k-means.

3.4 The Computation of Kernel k-Means

From a computational perspective, the clustering center c_j in Eq.(1) cannot be directly computed since that the points ϕ_i cannot be directly represented. To solve the problem, combining the key idea of kernel k-means and the kernel trick, one has

Proposition 1 (**Proposition 2 of [27]**). *Let $\mathbf{K}$ be the kernel matrix. Denote the i-th column of $\mathbf{K}$ by $\mathbf{k}_i$. Then*

$$\min_{\mathbf{c}\in\mathcal{H}} W(\mathbf{c},\mu_n)=\frac{1}{n}\min\sum_{j=1}^{k}\sum_{i\in\mathcal{S}_j}\left\|\mathbf{k}_i-\frac{1}{|\mathcal{S}_j|}\sum_{s\in\mathcal{S}_j}\mathbf{k}_s\right\|^2, \tag{7}$$

where $\mathcal{S}_j$ denotes the Voronoi cell (namely, j-th cluster) and $|\mathcal{S}_j|$ denotes the number of data in $\mathcal{S}_j$.

Proposition 1 shows that the point-oriented $\{\phi_i\}$ kernel clustering problem in Hilbert space can be transformed into the related kernel matrix column-oriented $\{\mathbf{k}_i\}$ problem. The space and time complexity of computing the kernel matrix $\mathbf{K}$ are all $\mathcal{O}(n^2)$ due to the $n\times n$ scale of kernel matrix. When n is very large, the requirements of calculation is prohibitive.

4 Kernel k-Means with Randomized Sketching (SKK)

The proposed approach SKK uses a novel randomized sketching method based on the circulant matrix to approximate kernel k-means clustering, which has the same statistical accuracy as the exact kernel k-means clustering, the optimal time complexity and the optimal space complexity compared to the state-of-the-art estimators. In this section, we mainly introduce the proposed approximate kernel k-means estimator in detail. This paper assumes that $\mathbb{E}\|\phi\|^2<\infty$.

4.1 Randomized Sketching

We use the randomized sketching method to accelerate the computation of kernel k-means, which is mainly to construct a novel sketch matrix $\mathbf{S}$, then project the kernel matrix to a smaller scale. The specific form of the proposed randomized sketching approach is as follows.

$$\hat{\mathbf{K}}=\mathbf{S}\mathbf{K}\mathbf{S}^T\in\mathbb{R}^{m\times m}, \tag{8}$$

with

$$\mathbf{S}=\mathbf{D}\mathbf{A}\mathbf{Q}, \tag{9}$$

where $\mathbf{D}\in\mathbb{R}^{m\times m}$ is a diagonal matrix, $\mathbf{A}\in\mathbb{R}^{m\times m}$ is a circulant matrix, and $\mathbf{Q}\in\mathbb{R}^{m\times n}$ is a sampling matrix.

The diagonal elements of $\mathbf{D}$ are i.i.d, which belong to $\{+1,-1\}$ with the same probability. The first column $\mathbf{a}_1$ of $\mathbf{A}$ obeys the normal distribution with expectation 0 and variance $1/m$, namely $\mathbf{a}_1\sim\mathcal{N}(0,1/m)$. Sample m different rows with the same probability from the identity matrix $\mathbf{I}\in\mathbb{R}^{n\times n}$, then construct the sampling matrix $\mathbf{Q}$ by the m rows of $\mathbf{I}$.

The circulant matrix is the key technique to construct the proposed randomized sketching approach, which generates the complete matrix by cycling the elements of the first column. The form of $\mathbf{A}$ is as below,

$$\mathbf{A}=\begin{bmatrix} a_1 & a_m & a_{m-1} & & \cdots & a_2\\ a_2 & a_1 & a_m & a_{m-1} & & \vdots\\ & a_2 & a_1 & a_m & \ddots & \\ \vdots & \ddots & \ddots & \ddots & \ddots & a_{m-1}\\ & & & \ddots & \ddots & a_m\\ a_m & a_{m-1} & \cdots & & a_2 & a_1 \end{bmatrix}. \tag{10}$$

The complete information of the circulant matrix can be preserved by only storing the first column. Therefore, its space complexity is $\mathcal{O}(m)$, which can save the storage space.

One can transform the circulant matrix into the form of discrete Fourier transform [52], $\mathbf{A}=\frac{1}{m}\mathbf{G}^*diag(\mathbf{G}\mathbf{a})\mathbf{G}$, where $\mathbf{G}=\left[e^{i\frac{2\pi}{m}kt}\right]_{k,t=1}^{m}$ and $\mathbf{a}=[a_1,a_2,\dots,a_m]^T$ is the first column of $\mathbf{A}$. $\mathbf{G}^*$ is conjugate transpose of $\mathbf{G}$. We can use FFT to accelerate a matrix-vector product ($\mathbf{A}\mathbf{v}$, $\mathbf{v}\in\mathbb{R}^m$ is a vector) [52]. The corresponding time complexity is $\mathcal{O}(m\log m)$. Nevertheless, if $\mathbf{A}$ is a traditional unstructured matrix, the time consumption of $\mathbf{A}\mathbf{v}$ is $\mathcal{O}(m^2)$. Therefore, the circulant matrix can greatly save time and space complexity during the process of matrix operation.

4.2 Kernel k-Means with Randomized Sketching

In the following, we introduce the proposed algorithm SKK based on the randomized sketching method mentioned above.

After generating the matrix $\hat{\mathbf{K}}$, we bring the columns $\hat{\mathbf{k}}_i$ of $\hat{\mathbf{K}}$ into k-means algorithm to determine a collection of clustering centers $\bar{\mathbf{c}}_n=(\bar{c}_{n1},\dots,\bar{c}_{nk})$ which minimizes the empirical clustering risk in $\mathbb{R}^m$. And their associated Voronoi cells are expressed as $\hat{\mathcal{S}}_1,\dots,\hat{\mathcal{S}}_k\subset\mathbb{R}^m$. Let the clustering centers be

$$\hat{c}_{nj}=\frac{\sum_{i=1}^{m}\mathbf{k}_i\mathbb{I}_{\hat{\mathbf{k}}_i\in\hat{\mathcal{S}}_j}}{|\hat{\mathcal{S}}_j|}, j=[k], \tag{11}$$

where $\hat{\mathbf{c}}_n$ denotes the collection of these k centers and $\mathbb{I}_{\hat{\mathbf{k}}_i\in\hat{\mathcal{S}}_j}$ is the indicator function. $\mathbb{I}_{\hat{\mathbf{k}}_i\in\hat{\mathcal{S}}_j}=1$ if $\hat{\mathbf{k}}_i\in\hat{\mathcal{S}}_j$ and $\mathbb{I}_{\hat{\mathbf{k}}_i\in\hat{\mathcal{S}}_j}=0$ otherwise.

Here we introduce the theoretical guarantee for the clustering centers and the proposed randomized sketching mentioned above.

Theorem 3. *The clustering risks of $\mathbf{K}$ and $\mathbf{S}\mathbf{K}\mathbf{S}^T$ are denoted by $W(\mathbf{c}_n,\mu_n)$ and $W(\hat{\mathbf{c}}_n,\mu_n)$. Given any $\varepsilon,\delta\in(0,1)$, let*

$$m\ge\frac{4\log n-2\log\delta}{\varepsilon-\log(1+\varepsilon)}.$$

We have, with probability at least $1-\delta$,

$$W(\hat{\mathbf{c}}_n,\mu_n)-W(\mathbf{c}_n,\mu_n)\leq\frac{4\varepsilon}{(1-\varepsilon)^2}. \tag{12}$$

Proof. The proof is shown in the section of PROOF. □

Remark 1. *ε is a small value between 0 and 1. Theorem 3 shows that the theoretical loss in clustering risk, caused by the proposed randomized sketching method and the carefully constructed clustering centers, is $\frac{4\varepsilon}{(1-\varepsilon)^2}$, which is a small value and is a part of excess risk bound in SKK. This means that the designed clustering centers based on the randomized sketching method are sound and the proposed randomized sketching method to clustering estimator is effective.*

The detailed process of SKK is summarized in Algorithm 1. Instead of using all the data directly to generate a complete kernel matrix $\mathbf{K}$, we sample the data points to construct a variant kernel matrix $\mathbf{K}^{'}$ for subsequent sketching, as step 4. Note that in the mathematical expression, $\mathbf{K}^{'}$ is equivalent to $\mathbf{QKQ}^T$, but we do not explicitly calculate the kernel matrix $\mathbf{K}$. As the kernel matrix is dense, this way can greatly reduce the computational requirements of the kernel matrix. Meanwhile, due to the proposed sketch matrix is based on the circulant matrix, one can use FFT to compute sketch kernel matrix $\hat{\mathbf{K}}$, as step 5. Then, we use the sketch kernel matrix $\hat{\mathbf{K}}$ to calculate iteratively in k-means algorithm, which can significantly reduce the time complexity, as step 6. Step 7 obtains the clustering centers based on Eq.(11).

Algorithm 1 randomized sketching kernel k-means (SKK)

```
Input: $\{\mathbf{x}_i\}_{i=1}^n$, $k$, kernel parameter and sketch dimension $m$.
Output: centroids $\hat{\mathbf{c}}_n$.
1: Construct the matrices $\mathbf{D}\in\mathbb{R}^{m\times m}$ and $\mathbf{Q}\in\mathbb{R}^{m\times n}$ described in Eq.(9),
2: Construct a vector $\mathbf{a}_1\in\mathbb{R}^m$, the entries of which obey the standard normal distribution,
3: Generate the circulant matrix $\mathbf{A}\in\mathbb{R}^{m\times m}$, where $\mathbf{a}_1$ acts as the first column of $\mathbf{A}$,
4: Sample $m$ data points by the sampling matrix $\mathbf{Q}$ then construct a variant kernel matrix $\mathbf{K}^{'}\in\mathbb{R}^{m\times m}$,
5: Compute $\hat{\mathbf{K}}=\frac{1}{m}(\mathbf{DAK}^{'})\mathbf{A}^T\mathbf{D}^T\in\mathbb{R}^{m\times m}$ with FFT,
6: Perform $k$-means over the columns of $\hat{\mathbf{K}}$,
7: Compute centroids $\hat{\mathbf{c}}_n$ in Eq.(11).
```

4.3 Complexity Analysis

The proposed approach reduces the number of data involved in the k-means iteration and reduces the corresponding time and storage space by utilizing the randomized sketching method to construct a smaller variant kernel matrix. The detailed consumption of SKK in terms of time and space is as follows.

4.3.1 Space Complexity

The determinant of space complexity in kernel k-means is the scale of the kernel matrix. To avoid this bottleneck, we firstly sample the datasets and then generate the kernel matrix based on the sampled data, which can save space cost from $\mathcal{O}(n^2)$ to $\mathcal{O}(m^2)$.

Additionally, the proposed sketch matrix is constructed based on the circulant matrix, whose space complexity is $\mathcal{O}(m)$. Finally, the space complexity of the proposed approach can be abbreviated as $\mathcal{O}(m^2)$, which has an n^2/m^2 improvement over the n^2 space of the exact kernel k-means.

4.3.2 Time Complexity

Because of the introduction of the circulant matrix, FFT can be used to calculate the multiplication of correlation matrices in SKK. In order to minimize the time cost, the expression $\hat{\mathbf{K}}=\mathbf{SKS}^T\in\mathbb{R}^{m\times m}$ can be rewritten as $\mathbf{S}(\mathbf{SK})^T$. Then, we can utilize FFT multiple times during the calculation. Due to early sampling, the time complexity of $\mathbf{S}(\mathbf{SK})^T$ is $O(m^2\log m)$. The time complexity of bringing the processed data into the subsequent t steps of k-means algorithm is $O(m^2kt)$.

Therefore, the time complexity in the proposed approach can be simplified into $O(m^2kt+m^2\log m)$. Once the sketching data $\hat{\mathbf{k}}_i$ are computed they can be manipulated in m^2, with an n^2/m^2 improvement over the n^2 time required by the exact embeddings $\mathbf{k}_i$. The proposed approach can also be implemented by parallel operation, and the corresponding time cost will be reduced by a certain multiple. However, this paper is mainly to validate the effectiveness of the randomized sketching-based approximation for kernel k-means with circulant matrix. Therefore, we do not consider bringing in parallel in this paper.

5 The Excess Risk Bound of SKK

This section mainly describes the generalization properties of SKK showing it achieves the same generalization error as exact kernel k-means, with dramatically reduced computations. This result is given in Theorem 4. In particular, with sketch dimension $\sqrt{n}$, SKK has essentially the same optimal excess risk bond as exact kernel k-means [24].

Theorem 4. *Given any n data and $\varepsilon,\delta\in(0,1)$. Let*

$$m\geq\frac{4\log n-2\log\delta}{\varepsilon-\log(1+\varepsilon)}, \tag{13}$$

and the clustering centers $\hat{\mathbf{c}}_n$ are found by the clustering algorithm based on randomized sketching in Eq.(8). Then, with probability at least $1-\delta$, we have

$$\begin{aligned}&\mathbb{E}\left[W(\hat{\mathbf{c}}_n,\mu)\right]-W^*(\mu)\\=&\tilde{\mathcal{O}}\left(\sqrt{\frac{k}{n}}\right)+\mathcal{O}\left(\frac{\varepsilon}{(1-\varepsilon)^2}\right).\end{aligned} \tag{14}$$

Proof. The proof is shown in the section of PROOF. □

Remark 2. *According to Theorem 2, the upper excess risk bound of the exact kernel k-means is $\mathcal{O}(\sqrt{\frac{k}{n}})$. Note that ε is a small value. Choosing $\varepsilon=\frac{1}{\sqrt{n}}$ (In general, $n>100$), the solution $\hat{\mathbf{c}}_n$ achieves the excess risk bound $\tilde{\mathcal{O}}\left(\sqrt{\frac{k}{n}}\right)$. From a statistical point of view, Theorem 4 shows that, with the suitable sketch dimension, the proposed SKK achieves the same optimal excess risk bound as the exact kernel k-means [24]. Compared to the representative randomized sketching method [36], we improve the excess risk bound from $\mathcal{O}\left(\frac{k}{\sqrt{n}}\right)$ to $\tilde{\mathcal{O}}\left(\sqrt{\frac{k}{n}}\right)$ with the smaller sketch dimension. This paper is the first result of the optimal excess risk bound based on the randomized sketching method. This illuminates that the approximate approach is sound.*

Remark 3. *From a computational point of view, Theorem 4 shows that taking the sketch dimension of $\sqrt{n}$ is sufficient for optimal*

TABLE 2
The main information of datasets used in this paper.

Datasets	Instance	Feature	Class
dna	2000	180	3
segment	2310	19	7
mushrooms	8124	112	2
pendigits	10992	16	10
protein	17766	357	3
a8a	32561	123	2
w7a	49749	300	2
connect-4	67557	126	3
mnist	60000	780	10
SVHN	73257	3072	10
skin-nonskin	245057	3	2
covtype	581012	54	7

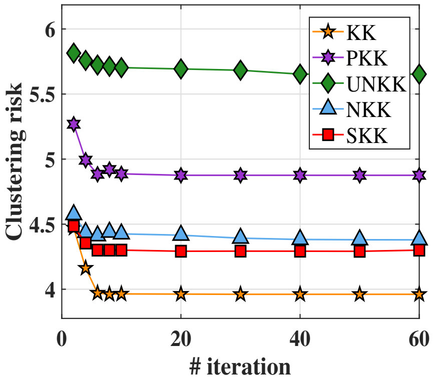


Fig. 1. Clustering risk and different number of iterations of PKK, UNKK, NKK, KK and SKK (ours) approaches on protein datasets with $m = 150$.

statistical accuracy, which greatly reduces the scale of kernel matrix from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$. Each iterations of approximate k-means algorithm only requires $\mathcal{O}(n)$ time instead of $\mathcal{O}(n^2)$. Compared to the exact kernel k-means clustering, the proposed SKK reduces the space and time all by a factor of n, with the same excess risk bounds. The detailed comparison with the related works is shown in TABLE 1 and RELATED WORK.

The result shows that the proposed approach takes a substantial step in provably reducing the computational requirements with the optimal generalization statistical accuracy.

Note that, Theorem 1 and Theorem 2 show the lower and upper bounds on excess risk of the exact kernel k-means. Theorem 3 shows that the proposed clustering centers and the proposed randomized sketching method are sound and effective, which is a part of the excess risk bound of the proposed SKK in Theorem 4. Combining Theorem 1 and Theorem 2, Theorem 4 shows that the proposed SKK achieves the same optimal excess risk bound as the exact kernel k-means with the suitable sketch dimension.

6 Experiments

This section empirically verifies the effectiveness of SKK and compares the performance of SKK with the state-of-the-art approaches of approximate kernel clustering on 12 real-world datasets. The hardware configuration of each experiment is 32 cores (2.40GHz), and RAM is 32 GB.

In this paper, each experimental value is the average of 30 experiments to avoid randomness. The statistical significance of differences among approaches in performance can be estimated by multiple training/prediction partitions. In partition i, let the error of method τ be τ_i and the error of method ρ be ρ_i. Let $\varpi_i = \tau_i - \rho_i, i \in \{1, \ldots, 30\}$. Denote the mean and standard deviation of ϖ_i as $\overline{\varpi}$ and ϱ. Under p-test, if the p-statistic $\frac{\overline{\varpi}}{\varrho/\sqrt{30}} > 1.699$, then with confidence level 95%, ρ is obviously better than τ. In the following, we use the 95% level of significance as the statistical significance. In addition, Friedman test [53], [54] compares the average ranks of algorithms. To measure the algorithms performance more comprehensively, we also use Friedman test to examine. Given r_i^j be the rank of the j-th of u algorithms on i-th of v datasets. Under the null-hypothesis, which states that all the algorithms are equivalent, Friedman statistic $F_F = \frac{(v-1)\chi_F^2}{v(u-1)-\chi_F^2}$ is distributed according to the F-distribution with $u-1$ and $(u-1)(v-1)$ degrees of freedom, where $\chi_F^2 = \frac{12}{vu(u+1)} \sum_{j=1}^{u} (\sum_{i=1}^{v} r_i^j)^2 - 3v(u+1)$.

6.1 Baselines and Parameter Settings

In this part, we mainly introduce parameter settings and the compared approaches, which include the representative randomized sketching kernel k-means approaches and the state-of-the-art Nyström kernel k-means approximation. The details are shown as follows.

1) PKK: PKK approach employed randomized sketching technique for approximate kernel k-means clustering [36].

2) UNKK: It is a representative approach combining randomized sketching and Nyström [25]. Its time complexity and space complexity are $\mathcal{O}\left(nmkt + nc^2\right)$ and $\mathcal{O}\left(nc\right)$ respectively with the uniform sampling, where c is the sketch size and $c > m$. Compared to it, we reduce the space complexity by a factor of nc/m^2 and the time complexity by a factor of $\frac{n(mkt+c^2)}{m(mkt+m\log m)}$ with the same m. Although UNKK did not get the excess risk bound and there is no comparability with it in theoretical accuracy, to fully validate the performance of the proposed approach, we make comparisons between SKK and UNKK in the experiment. According to [25], taking the parameter of sketch size $c = 2m$. Other parameters are the same as those in [25].

3) NKK: NKK is the state-of-the-art approach Nyström approximation to kernel k-means clustering [24].

4) KK: It is the abbreviation of the exact kernel k-means. The code is from websites[3].

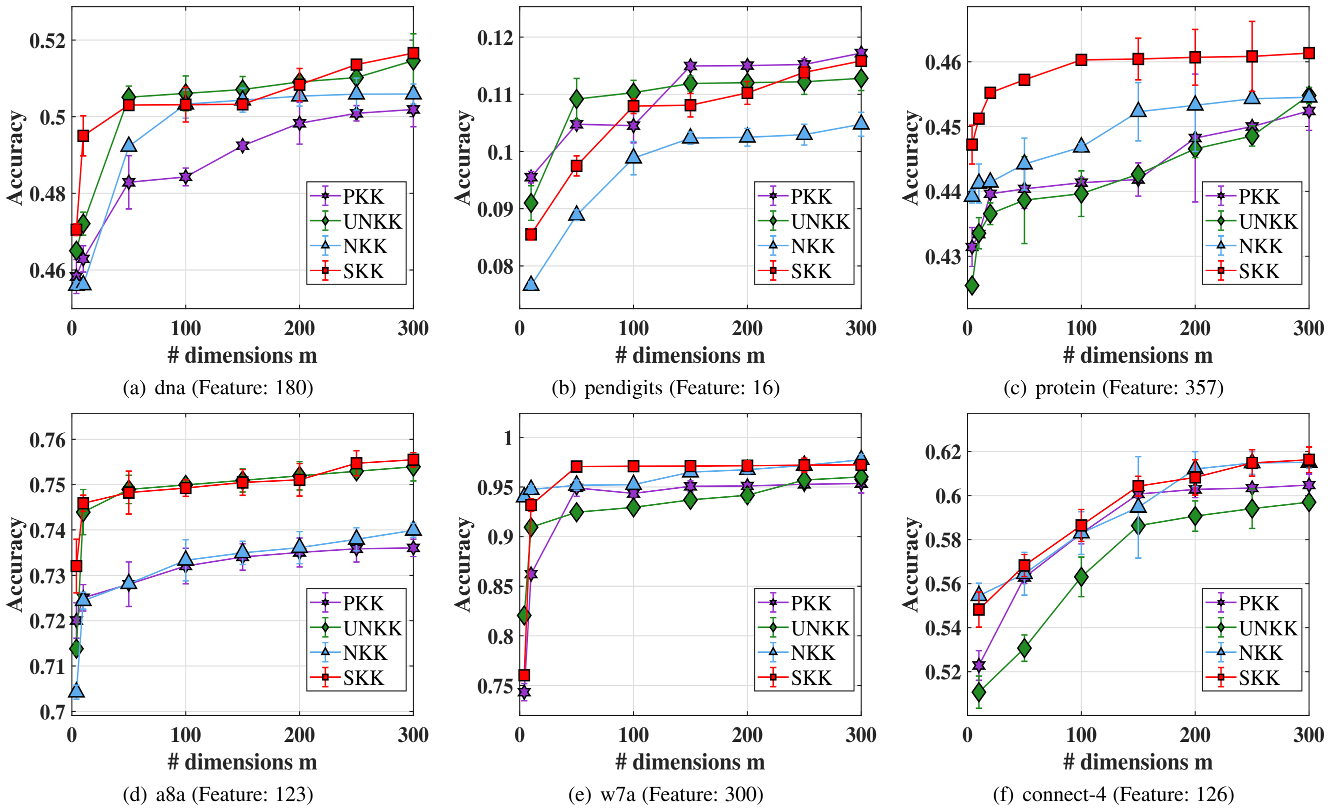
5) SKK: SKK represents the proposed approach which uses the randomized sketching based on the circulant matrix to approximate kernel k-means.

In the experiments, we utilize the Gaussian kernel for the approximate kernel k-means estimators on 12 real-world datasets. The kernel bandwidth σ of the Gaussian kernel is expressed in the following form, which is related to the average interpolation distance between data points, $\sigma = \sqrt{\frac{\sum_{ij}\|\mathbf{x}_i - \mathbf{x}_j\|^2}{n}}$.

6.2 Datasets

The experiments compare SKK with the state-of-the-art approximate kernel k-means approaches on 12 conventional and widely used datasets: dna, segment, mushrooms, pendigits, protein, a8a, w7a, connect-4, mnist, SVHN, skin-nonskin, and covtype datasets,

[3] www.cad.zju.edu.cn/home/dengcai/Data/data.html



(a) dna (Feature: 180)

(b) pendigits (Feature: 16)

(c) protein (Feature: 357)

(d) a8a (Feature: 123)

(e) w7a (Feature: 300)

(f) connect-4 (Feature: 126)

Fig. 2. Clustering accuracy and different dimensions m of PKK, UNKK, NKK, and SKK (ours) approaches on dna, pendigits, protein, a8a, w7a, and connect-4 datasets.

which are from LIBSVM website[4]. TABLE 2 shows the main information of the datasets. All datasets are normalized. Each datasets is divided into two parts. The first part is used for training experiments, accounting for 70 percent of the instances. The second part is used for prediction experiments. Each dataset is randomly divided into training set and prediction set according to the same rules in every approach. And the training/prediction partitions are the same given a specific approach.

6.3 Evolution Methodologies and Results

We use the clustering risk and the number of iterations to verify the convergence of kernel k-means estimators.

Fig.1 shows the relationship between the clustering risk and the number of iterations of PKK, UNKK, NKK, KK and SKK (ours) approaches on protein datasets with $m = 150$. When the number of iterations is very small, SKK has converged. The clustering risk of SKK is better than that of PKK, UNKK, and NKK.

We use the clustering accuracy and running time to evaluate the effectiveness of kernel k-means estimators.

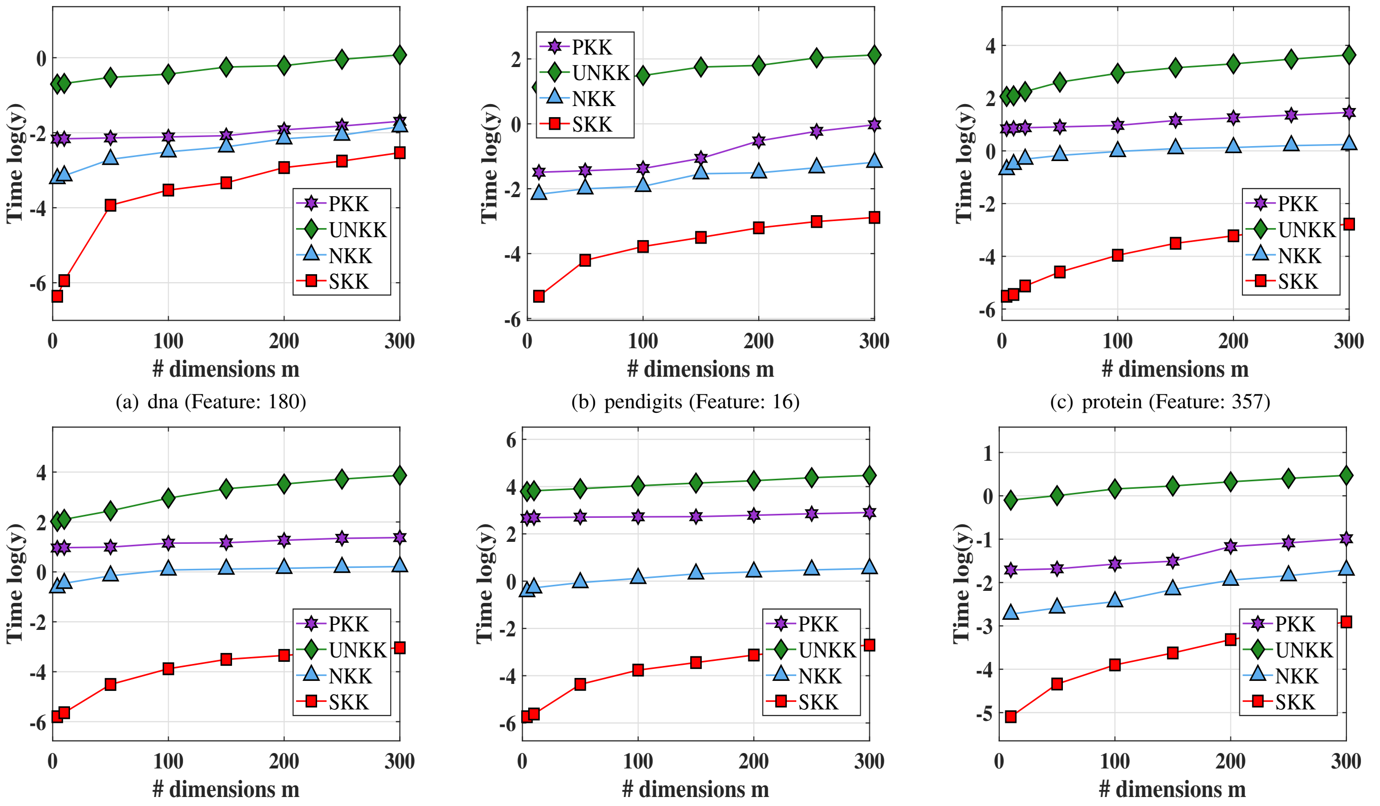
Define the clustering accuracy as $Acc = \frac{\sum_{i=1}^{\ddot{n}} v(\hat{y}, map(y))}{\ddot{n}}$, where $\ddot{n}$ is the number of data in prediction experiments, $\hat{y}$ and y are the real label and the derived label of the ith data. If $p = q$, function $v(p, q) = 1$, otherwise $v(p, q) = 0$. The mapping function $map(\cdot)$ represents the best mapping to match $\hat{y}$ and y. The formula of accuracy is the same as that in [33]. The higher the accuracy, the better the approach. For the sake of fairness, in Fig.2 and Fig.3, we use the default iteration number 100 and the same cluster centers initialization method on each approximation algorithm.

Fig.2, Fig.3, and TABLE 3 show the detail numerical results. The Y-axes represent clustering accuracy and logarithmic to running time (in seconds) of the training process. m denotes the sketch dimension in randomized sketching approaches such as PKK and SKK, and Nyström landmark points in Nyström approaches such as NKK and UNKK. The X-axes are the value of m. For the convenience of expression, we call m as "dimensions" in the figures. TABLE 3 shows the detailed accuracy and time numerical results of approximate kernel k-means approaches when $m = 150$ on every dataset. The missing experimental data in TABLE 3 are due to the too long running time (more than 90 seconds) of the algorithms or the inability of server memory to support them. See 3) below for specific explanation.

Based on the experimental results, we have the following analysis:

1) From Fig. 2 we know that the proposed approach always keeps the best or approximate best accuracy between approximate kernel k-means approaches. On protein datasets, the clustering accuracy of Nyström approach (NKK) is poor within the interval of m (the value of $m < \sqrt{nk}$) in this Figure. This is consistent with its theoretical analysis that when $m > \sqrt{nk}$, NKK obtains the optimal excess risk bound. Under the same conditions, the proposed approach has an advantage over the state-of-the-art Nyström approach NKK in most cases. Especially, when m is a very small value, our algorithm has reached a high accuracy on datasets such as protein, a8a, and w7a. Namely, for a small m, SKK has obtained the satisfactory accuracy, and there is no need to further increase the number of m for accuracy, which will cause

[4] http://www.csie.ntu.edu.tw/∼cjlin/libsvm.



(a) dna (Feature: 180) (b) pendigits (Feature: 16) (c) protein (Feature: 357)

(d) a8a (Feature: 123) (e) w7a (Feature: 300) (f) connect-4 (Feature: 126)

Fig. 3. Running time and different dimensions m of PKK, UNKK, NKK, and SKK (ours) approaches on dna, pendigits, protein, a8a, w7a, and connect-4 datasets.

TABLE 3
Clustering accuracy and Time in solving kernel k-means between PKK, UNKK, NKK and SKK (ours) approaches on 12 datasets with $m = 150$. The missing experimental values are due to the too long running time of the approaches or the inability of server memory to support them. If the running time of the approach is more than 90 seconds, the experiment will be stopped. "TO" is short for the timeout. "OM" is short for out of memory. The bold values represent the best experimental results. The underlined values indicate that the results of this approach in accuracy are not significantly worse than those of the best approach.

Dataset	PKK		UNKK		NKK		SKK (Ours)	
	Time	Accuracy	Time	Accuracy	Time	Accuracy	Time	Accuracy
dna	0.12	0.49±0.0009	0.78	**0.51± 0.0034**	0.09	0.50±0.0031	**0.04**	0.50±0.0101
segment	0.09	**0.45±0.0043**	0.31	0.39±0.0032	0.05	0.43±0.0090	**0.02**	0.37±0.0110
mushrooms	0.32	0.63±0.0063	2.54	0.53±0.0019	0.11	0.61±0.0027	**0.03**	**0.64±0.0062**
pendigits	0.34	**0.11 ±0.0005**	5.76	**0.11±0.0013**	0.21	0.10± 0.0011	**0.03**	**0.11±0.0021**
protein	3.16	0.44±0.0026	23.4	0.44±0.0005	1.09	0.45±0.0045	**0.03**	**0.46±0.0032**
a8a	3.21	0.73±0.0029	27.9	**0.75±0.0026**	1.12	0.73±0.0026	**0.03**	**0.75±0.0028**
w7a	15.3	0.95±0.0055	63.3	0.94±0.0041	1.36	0.96± 0.0000	**0.03**	**0.97±0.0043**
connect-4	0.22	**0.60±0.0017**	1.25	0.59±0.0010	0.11	0.59±0.0230	**0.03**	**0.60±0.0045**
mnist	TO	TO	TO	TO	0.95	0.17±0.0096	**0.06**	**0.22±0.0048**
SVHN	OM	OM	TO	TO	6.67	0.11±0.0014	**0.12**	**0.14±0.0021**
skin-nonskin	OM	OM	TO	TO	TO	TO	**0.07**	**0.63±0.0031**
covtype	OM	OM	TO	TO	TO	TO	**0.27**	**0.32±0.0035**

the increase of the computational cost. They are in line with the theoretical guarantees in this paper and [24]. Due to data noise or maybe other reasons, the accuracy of SKK is not significantly better than other algorithms on dna and pendigits datasets, but it is still at the same level as that of the state-of-the-art approximate kernel k-means estimators.

2) The empirical results in Fig. 3 show that SKK is significantly faster than other approximate approaches in running time with the same m, which can obviously accelerate the kernel k-means. The larger m is, the longer the algorithms run. The proposed approach is even more than 3,000 times faster than UNKK approach on a8a and w7a datasets. This verifies the theoretical analysis of computational requirements. The running time of UNKK is high, which is caused by the high-consuming operation of multiple matrix decompositions. The running time of the proposed approach is always smaller than other approximate

approaches on the small and large scale datasets. This means that SKK is scalable not only to large scale datasets but also to small scale datasets for kernel k-means. Combining the results in Fig. 2 and Fig. 3, we know that at the same m, SKK obtains satisfactory accuracy with the obviously time advantage over the state-of-the-art approximate clustering estimators. Therefore, we obtain that SKK can obviously speed up the kernel k-means while maintaining sound clustering performance. This verifies the theoretical analysis.

3) TABLE 3 shows the concrete value of clustering accuracy and running time of each approaches on 12 datasets with $m = 150$. The bold values in this table represent the best experimental results. The underlined values indicate that the results of this approach in accuracy are not significantly worse than those of the best approach. The missing experimental values are due to the too long running time of the approaches or the inability of server memory to support them. If the running time of the approach is more than 90 seconds, the experiment will be stopped. "TO" is short for the timeout. "OM" is short for out of memory in TABLE 3. With the increase of data points in datasets, due to the large space and/or time complexity and the limitation of the hardware machine, some approximate approaches cannot achieve the experimental results. Firstly, PKK and UNKK cannot obtain the experimental values on mnist and SVHN datasets due to their high time cost and high space cost of PKK, subsequently, NKK cannot obtain them on the larger datasets skin-nonskin and covtype due to their high time cost. This is in line with the theoretical results that the computational requirements of PKK and UNKK are bigger than NKK, and all of them are bigger than SKK. In particular, processing more than 200000 data points in skin-nonskin dataset, SKK only needs 0.07 seconds, while the state-of-the-art kernel k-means estimates cannot even get the results because of the high computational cost (more than 90 seconds). This means that SKK accelerates the speed of approximate kernel k-means by at least 1200 times, which is scalable. This verifies that the proposed approach is more efficient than other approximate approaches in time cost. In accuracy, this table shows that there is no difference between the proposed approach and the best at the 95 percent level of significance except on segment dataset. With four algorithms and the first 8 datasets in TABLE 3, F_F is distributed with $4-1=3$ and $(4-1)\times(8-1)=21$ degrees of freedom. The critical value of $F(3,21)$ for $\alpha=0.05$ is 3.07. In accuracy, F_F is smaller than 3.07 so we accept the null-hypothesis. In time cost, F_F is bigger than 3.07 so we reject the null-hypothesis. That is, the proposed approach is more efficient in time cost and has no difference in accuracy. This is in line with the theoretical analysis. Those verify that SKK is effective in approximate kernel k-means.

Using less time to obtain sound accuracy highlights the high-cost performance of the algorithms. Combining with the above analysis, we know the proposed algorithm can achieve satisfactory accuracy, prominent advantages in speed and storage space on datasets. This is consistent with the theoretical results.

7 Proof

7.1 Preparations

Before proving the theorems mentioned above, we give some definitions about the circulant matrix and ϕ, and provide some propositions.

Definition 4. *Define a k-valued function*

$$g_{\mathbf{c}} = (g_{c_1}, \dots, g_{c_k}),$$

which is about the collection $\mathbf{c} = \{c_1, \dots, c_k\} \in \mathcal{H}^k$ with $g_{c_j}(\mathbf{x}) = \|\phi_{\mathbf{x}} - c_j\|^2$. Define $\mathcal{G}_{\mathbf{c}}$ be a family of $g_{\mathbf{c}}$:

$$\mathcal{G}_{\mathbf{c}} := \{g_{\mathbf{c}} = (g_{c_1}, \dots, g_{c_k}) : \mathbf{c} \in \mathcal{H}^k\}.$$

Definition 5 (Definition 1 in [52]). *If the entries in the first column of a circulant matrix are i.i.d, the circulant matrix is called a circulant random matrix.*

Proposition 2 (Theorem 1 in [55]). *Let $l : \mathbb{R}^k \to \mathbb{R}$ satisfy $\|l(\boldsymbol{\gamma}) - l(\boldsymbol{\gamma}^{'})\|_\infty \le L \cdot \|\boldsymbol{\gamma} - \boldsymbol{\gamma}^{'}\|_\infty, \forall \boldsymbol{\gamma}, \boldsymbol{\gamma}^{'} \in \mathbb{R}^k$. That is, l is L-Lipschitz with respect to the L_∞ norm. Let $\mathcal{G} \subseteq \{g : \mathcal{X} \to \mathbb{R}^k\}$. If $\max\{|l(g(\mathbf{x}))|, \|g(\mathbf{x})\|_\infty\} \le \rho$, then there exists a constant $C > 0$ such that for any $b > 0$, the following inequality holds:*

$$\mathcal{A}_n(l \circ \mathcal{G}) \le C \cdot L\sqrt{k} \max_i \tilde{\mathcal{A}}_n(\mathcal{G}_i) \log^{\frac{3}{2}+b}\left(\frac{\rho n}{\max_i \tilde{\mathcal{A}}_n(\mathcal{G}_i)}\right),$$

where $\mathcal{A}_n(l \circ \mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} |\sum_{i=1}^n \sigma_i l(g(\mathbf{x}_i))|\right]$, $\tilde{\mathcal{A}}_n(\mathcal{G}_i) = \sup_{\mathbf{X}\in\mathcal{X}^n} \mathcal{A}_n(\mathcal{G}_i)$.

Proposition 3 (Lemma 24(a) in [56]). *Given $\varsigma_1, \dots, \varsigma_n \in \mathcal{H}$, one has*

$$\mathbb{E}_{\boldsymbol{\sigma}}\left\|\sum_{i=1}^n \sigma_i \varsigma_i\right\|^2 \le \sum_{i=1}^n \|\varsigma_i\|^2, \tag{15}$$

and

$$\mathbb{E}_{\boldsymbol{\sigma}}\left\|\sum_{i=1}^n \sigma_i \varsigma_i\right\| \ge \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^n \|\varsigma_i\|^2}, \tag{16}$$

where $\sigma_1, \dots, \sigma_n$ be a sequence of independent Rademacher variables.

7.2 Proof About Randomized Sketching

Here, we theoretically analyze the projection effectiveness of the proposed randomized sketching method.

Theorem 5. *Given any set $\mathcal{D}$ of n points in $\mathcal{H}$, and $\mathbf{K} \in \mathbb{R}^{n\times n}$ its kernel matrix. $\mathbf{k}_i$ denotes i-th column of $\mathbf{K}$. For any $\varepsilon, \delta \in (0,1)$, let*

$$m \ge \frac{4\log n - 2\log\delta}{\varepsilon - \log(1+\varepsilon)}.$$

There exists a map f: $\mathbb{R}^n \to \mathbb{R}^m$ described in Eq.(8). We have, with probability at least $1-\delta$,

$$(1-\varepsilon)\|\mathbf{k}_i - \mathbf{k}_j\|^2 \le \|f(\mathbf{k}_i) - f(\mathbf{k}_j)\|^2 \le (1+\varepsilon)\|\mathbf{k}_i - \mathbf{k}_j\|^2,$$

for all $\mathbf{k}_i, \mathbf{k}_j \in \mathbf{K}$.

Proof. The circulant matrix $\mathbf{A}$ is a circulant random matrix according to Definition 5. The random diagonal matrix $\mathbf{D}$ guarantees the independence among the columns of $\mathbf{A}$. The elements of $\mathbf{DA}$, $(\mathbf{DA})_{ij} = \mathbf{D}_{ii}\mathbf{A}_{ij}$, retain Gaussianity. Therefore $(\mathbf{DA})_{.j}$ is also Gaussian vector, where, for $i, j \in 1, \dots, m$, $\mathbb{E}[(\mathbf{DA})_{ij}] = \mathbb{E}[\mathbf{D}_{ii}\mathbf{A}_{ij}] = 0$, and

$$\begin{aligned}\text{Var}[(\mathbf{DA})_{ij}] &= \text{Var}[\mathbf{D}_{ii}\mathbf{A}_{ij}]\\ =&\mathbb{E}[\mathbf{D}_{ii}^2\mathbf{A}_{ij}^2] - \mathbb{E}[\mathbf{D}_{ii}\mathbf{A}_{ij}]^2 = \frac{1}{m}.\end{aligned}$$

Combining the properties of diagonal matrix and circulant matrix, one can obtain $\mathbf{S}_{ij} \sim \mathcal{N}(0, 1/m)$.

In the following, $\mathbf{SKS}^T$ is divided into two parts for analysis, namely $\mathbf{KS}^T$ and $\mathbf{S}(\mathbf{KS}^T)$.

Firstly, analyzing the random projection of $\mathbf{KS}^T$, which can be expressed as a linear map $h : \mathbb{R}^n \to \mathbb{R}^m$.

In this part, the target is to prove the following inequalities: For any $\varepsilon_2 > 0$,

$$\mathbb{P}\Big[\frac{\|h(\mathbf{k}_{i.})\|^2}{\|\mathbf{k}_{i.}\|^2} > 1+\varepsilon_2\Big] < \Big(\frac{1}{1+\varepsilon_2}\Big)^{-m/2}\exp\Big(\frac{-m\varepsilon_2}{2}\Big). \tag{17}$$

$$\mathbb{P}\Big[\frac{\|h(\mathbf{k}_{i.})\|^2}{\|\mathbf{k}_{i.}\|^2} < 1-\varepsilon_2\Big] <\Big(\frac{1}{1+\varepsilon_2}\Big)^{-m/2}\exp\Big(\frac{-m\varepsilon_2}{2}\Big). \tag{18}$$

We start with the upper tail Eq.(17).

For $j = 1, \dots, m$, let $\mathbf{R}_j(\mathbf{k}_{i.}) = \mathbf{k}_{i.} \cdot \mathbf{S}_j^T$. Note that

$$\mathbb{E}(\mathbf{S}_{ij}^4) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}\exp(-\lambda^2/2)\left(\frac{\lambda^4}{m^2}\right)d\lambda = \frac{3}{m^2}, \tag{19}$$

and

$$\mathbb{E}(\mathbf{S}_{ij}^4) \geq \mathbb{E}(\mathbf{R}_1(\mathbf{k}_{i.})^4), \tag{20}$$

therefore, we have

$$\mathbb{E}(\mathbf{R}_1(\mathbf{k}_{i.})^4) \leq \frac{3}{m^2}. \tag{21}$$

For any $p \in [0, m/2)$, the integral in Eq.(22) is convergent, so that we have

$$\begin{aligned}\mathbb{E}(\exp(p\mathbf{S}_{ij}^2)) &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}\exp(-\lambda^2/2)\exp(p\frac{\lambda^2}{m})d\lambda \\ &= \frac{1}{\sqrt{1-2p/m}}.\end{aligned} \tag{22}$$

For any p and any random variable U, $\mathbb{E}(\exp(pU^2))$ is bounded. According to the Monotone Convergence Theorem (MCT), we know that

$$\mathbb{E}(\exp(pU^2)) = \mathbb{E}\left(\sum_{k=0}^{\infty}\frac{(pU^2)^k}{k!}\right) = \sum_{k=0}^{\infty}\frac{p^k}{k!}\mathbb{E}(U^{2k}).$$

Replacing U with $\mathbf{S}_{ij}$, we obtain

$$\begin{aligned}\mathbb{E}(\exp(p\mathbf{S}_{ij}^2)) &= \sum_{k=0}^{\infty}\frac{p^k}{k!}\mathbb{E}(\mathbf{S}_{ij}^{2k}) \\ \geq \sum_{k=0}^{\infty}\frac{p^k}{k!}&\mathbb{E}(\mathbf{R}_1(\mathbf{k}_{i.})^{2k}) = \mathbb{E}(\exp(p\mathbf{R}_1(\mathbf{k}_{i.})^2)).\end{aligned} \tag{23}$$

Here, combining Eq.(22) and Eq.(23), we obtain

$$\mathbb{E}(\exp(p\mathbf{R}_1(\mathbf{k}_{i.})^2)) \leq \frac{1}{\sqrt{1-2p/m}}. \tag{24}$$

Let $\|h(\mathbf{k}_{i.})\|^2 = \sum_{j=1}^m(\mathbf{k}_{i.} \cdot \mathbf{S}_j^T)^2 = \sum_{j=1}^m \mathbf{R}_j^2(\mathbf{k}_{i.})$. For arbitrary $p > 0$, we can write

$$\begin{aligned}&\mathbb{P}\Big[\frac{\|h(\mathbf{k}_{i.})\|^2}{\|\mathbf{k}_{i.}\|^2} > 1+\varepsilon_2\Big] \\ =&\mathbb{P}\Big[\exp(p\frac{\|h(\mathbf{k}_{i.})\|^2}{\|\mathbf{k}_{i.}\|^2}) > \exp(p(1+\varepsilon_2))\Big] \\ <&\mathbb{E}\Big(\exp(p\frac{\|h(\mathbf{k}_{i.})\|^2}{\|\mathbf{k}_{i.}\|^2})\Big)\exp\Big(-p(1+\varepsilon_2)\Big).\end{aligned} \tag{25}$$

Since $\{\mathbf{R}_j\}_{j=1}^m$ are i.i.d, let $\|\mathbf{k}_{1.}\|^2 = 1$, we have:

$$\begin{aligned}\mathbb{E}\Big(\exp(p\frac{\|h(\mathbf{k}_{i.})\|^2}{\|\mathbf{k}_{i.}\|^2})\Big) &= \mathbb{E}\Big(\prod_{j=1}^m\exp(p\frac{\mathbf{R}_j^2}{\|\mathbf{k}_{i.}\|^2})\Big) \\ = \left(\mathbb{E}\left(\exp(p\frac{\mathbf{R}_1^2}{\|\mathbf{k}_{1.}\|^2})\right)\right)^m &= \left(\mathbb{E}\left(\exp(p\mathbf{R}_1^2)\right)\right)^m.\end{aligned} \tag{26}$$

To optimize the bound, this gives $p = \frac{m}{2}\cdot\frac{\varepsilon_2}{(1+\varepsilon_2)} < \frac{m}{2}$. Taking Eq.(24) to Eq.(27). Thus, for any $\varepsilon_2 > 0$, one can see that

$$\begin{aligned}&\mathbb{P}\Big[\frac{\|h(\mathbf{k}_{i.})\|^2}{\|\mathbf{k}_{i.}\|^2} > 1+\varepsilon_2\Big] \\ <& \left(\mathbb{E}\left(\exp(p\mathbf{R}_1^2)\right)\right)^m\exp\Big(-p(1+\varepsilon_2)\Big) \\ \leq&\Big(\frac{1}{\sqrt{1-2p/m}}\Big)^m\exp\Big(-p(1+\varepsilon_2)\Big) \\ =&\Big(\frac{1}{1+\varepsilon_2}\Big)^{-m/2}\exp\Big(\frac{-m\varepsilon_2}{2}\Big).\end{aligned} \tag{27}$$

The proof of lower bound in Eq.(18) is similar to Eq.(17).

For arbitrary $p > 0$ and any $\varepsilon_2 > 0$, we obtain that,

$$\begin{aligned}&\mathbb{P}\Big[\frac{\|h(\mathbf{k}_{i.})\|^2}{\|\mathbf{k}_{i.}\|^2} < 1-\varepsilon_2\Big] \\ =&\mathbb{P}\Big[\exp(p\frac{\|h(\mathbf{k}_{i.})\|^2}{\|\mathbf{k}_{i.}\|^2}) < \exp(p(1-\varepsilon_2))\Big] \\ <&\mathbb{E}\Big(\exp(-p\frac{\|h(\mathbf{k}_{i.})\|^2}{\|\mathbf{k}_{i.}\|^2})\Big)\exp\Big(p(1-\varepsilon_2)\Big). \\ =& \left(\mathbb{E}\left(\exp(-p\mathbf{R}_1^2)\right)\right)^m\exp\Big(p(1-\varepsilon_2)\Big).\end{aligned} \tag{28}$$

We know that $\mathbb{E}(\mathbf{S}_{ij}) = 0$, $\mathbb{E}(\mathbf{S}_{ij}^2) = \frac{1}{m}$, and $\mathbb{E}(\mathbf{R}_j) = \mathbb{E}(\mathbf{k}_{i.} \cdot \mathbf{S}_{.j}^T) = \mathbf{k}_{i.}\mathbb{E}(\mathbf{S}_{.j}^T) = 0$. So, we obtain that

$$\begin{aligned}\mathbb{E}(\mathbf{R}_j^2) &= \mathbb{E}\left(\left(\mathbf{k}_{i.}\cdot\mathbf{S}_{.j}^T\right)^2\right) = \mathbb{E}\left(\left(\sum_{t=1}^n \mathbf{k}_{it}\mathbf{S}_{tj}^T\right)^2\right) \\ =&\left[\sum_{t=1}^n \mathbf{k}_{it}^2\mathbb{E}((\mathbf{S}_{tj}^T)^2) + \sum_{l=1}^n\sum_{m=1}^n 2\mathbf{k}_{il}\mathbf{k}_{im}\mathbb{E}(\mathbf{S}_{lj}^T)\mathbb{E}(\mathbf{S}_{mj}^T)\right] \\ =&\frac{\|\mathbf{k}_{i.}\|^2}{m}.\end{aligned} \tag{29}$$

Let us expand $\exp(-p\mathbf{R}_1^2)$ to get

$$\begin{aligned}&\mathbb{P}\left[\frac{\|h(\mathbf{k}_{i.})\|^2}{\|\mathbf{k}_{i.}\|^2} < 1-\varepsilon_2\right] \\ <&\Big(\mathbb{E}\Big(1-p\mathbf{R}_1^2+\frac{(-p\mathbf{R}_1^2)^2}{2!}\Big)\Big)^m\exp\Big(p(1-\varepsilon_2)\Big) \\ =&\left(1-p\mathbb{E}(\mathbf{R}_1^2)+\frac{p^2}{2}\mathbb{E}(\mathbf{R}_1^4)\right)^m\exp\Big(p(1-\varepsilon_2)\Big).\end{aligned} \tag{30}$$

Substituting Eq.(29) for Eq.(30), $\|\mathbf{k}_{i.}\|^2 \leq 1$, we get Eq.(32). Taking $p = \frac{m}{2}\cdot\frac{\varepsilon_2}{(1+\varepsilon_2)} < \frac{m}{2}$, and a series of expansion, we get Eq.(33):

$$\mathbb{P}\Big[\frac{\|h(\mathbf{k}_{i.})\|^2}{\|\mathbf{k}_{i.}\|^2} < 1-\varepsilon_2\Big] \tag{31}$$

$$<\Big(1-\frac{p}{m}+\frac{p^2}{2m^2}\Big)^m\exp\Big(p(1-\varepsilon_2)\Big) \tag{32}$$

$$<\Big(\frac{1}{1+\varepsilon_2}\Big)^{-m/2}\exp\Big(\frac{-m\varepsilon_2}{2}\Big). \tag{33}$$

Here, we complete the proof of upper and lower bounds.

Let

$$2 \times \left(\frac{1}{1+\varepsilon_2}\right)^{-m/2} \exp\left(\frac{-m\varepsilon_2}{2}\right) \le 2\delta_2/n^2,$$

we obtain $m \ge \frac{4\log n - 2\log\delta_2}{\varepsilon_2 - \log(1+\varepsilon_2)}$. Combining Eq.(17) and Eq.(18), for each of the $\binom{n}{2}$ pairs $\mathbf{u},\mathbf{v} \in \{\mathbf{k}_{i.}\}$, the squared norm of the vector $\mathbf{u}-\mathbf{v}$, is maintained within a factor of $1 \pm \varepsilon_2$, with the probability of $1-\binom{n}{2}\times 2\delta_2/n^2 > 1-\delta_2$. Because kernel matrix $\mathbf{K}$ is a symmetric matrix, we have $\mathbf{k}_{i.} = \mathbf{k}_i^T$. Therefore we have, with the probability at least $1-\delta_2$, for all $\mathbf{k}_i,\mathbf{k}_j \in \mathbf{K}$,

$$(1-\varepsilon_2)\|\mathbf{k}_i-\mathbf{k}_j\|^2 \le \|h^{'}(\mathbf{k}_i)-h^{'}(\mathbf{k}_j)\|^2 \le (1+\varepsilon_2)\|\mathbf{k}_i-\mathbf{k}_j\|^2. \tag{34}$$

The relation between h and $h^{'}$ is $h^{'}(\mathbf{k}_i) = h(\mathbf{k}_{i.})$.

Secondly, analyzing the second part $\mathbf{S}(\mathbf{K}\mathbf{S}^T)$.

Let $\mathbf{S}(\mathbf{K}\mathbf{S}^T) = \mathbf{S}\tilde{\mathbf{K}}$, which can be expressed as a linear map $g : \mathbb{R}^n \to \mathbb{R}^m$. The proof is similar to the above. So one can get the following result. Let

$$m \ge \frac{4\log m - 2\log\delta_1}{\varepsilon_1 - \log(1+\varepsilon_1)}. \tag{35}$$

We have, with the probability at least $1-\delta_1$, for all $h^{'}(\mathbf{k}_i), h^{'}(\mathbf{k}_j) \in \tilde{\mathbf{K}}$,

$$\begin{aligned}&(1-\varepsilon_1)\|h^{'}(\mathbf{k}_i)-h^{'}(\mathbf{k}_j)\|^2 \le \left\|g\left(h^{'}(\mathbf{k}_i)\right) - g\left(h^{'}(\mathbf{k}_j)\right)\right\|^2 \\ \le&(1+\varepsilon_1)\|h^{'}(\mathbf{k}_i)-h^{'}(\mathbf{k}_j)\|^2.\end{aligned}$$

Combining the above conclusions, one can obtain

$$\begin{aligned}&(1-\varepsilon_1)(1-\varepsilon_2)\|\mathbf{k}_i-\mathbf{k}_j\|^2 \le \|f(\mathbf{k}_i)-f(\mathbf{k}_j)\|^2 \\ \le&(1+\varepsilon_1)(1+\varepsilon_2)\|\mathbf{k}_i-\mathbf{k}_j\|^2.\end{aligned}$$

Let $\varepsilon_1 = \varepsilon_2, \varepsilon = \varepsilon_1^2 + 2\varepsilon_1 \in (0,1), \delta_1 = \delta_2, \delta = 2\delta_1 - \delta_1^2 \in (0,1)$. Here, the conclusion in this Theorem is obtained. □

7.3 Proof of Theorem 3

Proof. The clustering risks of $\mathbf{K}\mathbf{S}^T$ are denoted by $W(\hat{\mathbf{c}}_{ks},\mu_n)$. Denote by $\tilde{\mathbf{c}}_n = (\tilde{c}_{n1},\ldots,\tilde{c}_{nk})$ the clustering centers of $\tilde{\mathbf{k}}_1,\ldots,\tilde{\mathbf{k}}_n$. Each clustering centers $\tilde{c}_{nj}$ is the mean of data in this cluster $\tilde{S}_{nj}$. The expression is as follows.

$$\tilde{c}_{nj} = \frac{\sum_{i=1}^n \tilde{\mathbf{k}}_i \mathbb{I}_{\left\{\tilde{\mathbf{k}}_i\in\tilde{S}_{nj}\right\}}}{\sum_{i=1}^n \mathbb{I}_{\left\{\tilde{\mathbf{k}}_i\in\tilde{S}_{nj}\right\}}}, \quad j=1,\ldots,k.$$

Define $\tilde{\alpha}_j = \sum_{i=1}^n \mathbb{I}_{\left\{\tilde{\mathbf{k}}_i\in\tilde{S}_{nj}\right\}}$. Then, we obtain

$$\begin{aligned}W\left(\tilde{\mathbf{c}}_n,\mu_n\right) &= \frac{1}{n}\sum_{i=1}^n \min_{j=1,\ldots,k}\left\|\tilde{\mathbf{k}}_i - \tilde{c}_{nj}\right\|^2 \\ &=\frac{1}{n}\sum_{j=1}^k\sum_{i=1}^n\left\|\tilde{\mathbf{k}}_i - \tilde{c}_{nj}\right\|^2 \mathbb{I}_{\left\{\tilde{\mathbf{k}}_i\in\tilde{S}_{nj}\right\}} \\ &=\sum_{j=1}^k \frac{1}{2n\tilde{\alpha}_j}\sum_{i_1,i_2=1}^n \left\|\tilde{\mathbf{k}}_{i_1} - \tilde{\mathbf{k}}_{i_2}\right\|^2 \mathbb{I}_{\left\{\left(\tilde{\mathbf{k}}_{i_1},\tilde{\mathbf{k}}_{i_2}\right)\in\tilde{S}_{nj}^2\right\}}.\end{aligned}$$

Using the optimality of k-means (see Lemma 1 in Linder [57]), we know

$$W\left(\tilde{\mathbf{c}}_n,\mu_n\right) \le \sum_{j=1}^k \frac{1}{2n\beta_j}\sum_{i_1,i_2=1}^n \left\|\tilde{\mathbf{k}}_{i_1} - \tilde{\mathbf{k}}_{i_2}\right\|^2 \mathbb{I}_{\left\{\left(\mathbf{k}_{i_1},\mathbf{k}_{i_2}\right)\in S_{nj}^2\right\}},$$

where the S_{nj} 's are the Voronoi cells associated with $\mathbf{c}_n = (c_{n1},\ldots,c_{nk})$, and $\beta_j = \sum_{i=1}^n \mathbb{I}_{\{\mathbf{k}_i\in S_{nj}\}}$. Consequently, we have, with probability at least $1-\delta$, by Eq.(34),

$$\begin{aligned}&W\left(\tilde{\mathbf{c}}_n,\mu_n\right) \\ \le&(1+\varepsilon)\sum_{j=1}^k \frac{1}{2n\beta_j}\sum_{i_1,i_2=1}^n \|\mathbf{k}_{i_1}-\mathbf{k}_{i_2}\|^2 \mathbb{I}_{\left\{\left(\mathbf{k}_{i_1},\mathbf{k}_{i_2}\right)\in S_{nj}^2\right\}} \\ =&(1+\varepsilon)W\left(\mathbf{c}_n,\mu_n\right).\end{aligned}$$

Similarly $(1-\varepsilon)W\left(\hat{\mathbf{c}}_{ks},\mu_n\right) \le W\left(\tilde{\mathbf{c}}_n,\mu_n\right)$ as desired. Therefore, one can obtain

$$W(\hat{\mathbf{c}}_{ks},\mu_n) \le \frac{1+\varepsilon}{1-\varepsilon}W(\mathbf{c}_n,\mu_n). \tag{36}$$

Combining Theorem 5 and Eq.(36), one has

$$W(\hat{\mathbf{c}}_n,\mu_n) \le \frac{1+\varepsilon}{1-\varepsilon}W(\hat{\mathbf{c}}_{ks},\mu_n) \le \left(\frac{1+\varepsilon}{1-\varepsilon}\right)^2 W(\mathbf{c}_n,\mu_n). \tag{37}$$

Changing the form, we have:

$$W(\hat{\mathbf{c}}_n,\mu_n) - W(\mathbf{c}_n,\mu_n) \le \frac{4\varepsilon}{(1-\varepsilon)^2}W(\mathbf{c}_n,\mu_n) \le \frac{4\varepsilon}{(1-\varepsilon)^2},$$

where ε is a small value, $W(\mathbf{c}_n,\mu_n) \le 1$. □

7.4 Proof of Theorem 4

For proving Theorem 4, we first introduce some lemmas.

Lemma 1. *Let* $b_i := \sup_{\mathbf{x}\in\mathcal{X}} \sup_{g_c\in\mathcal{G}_{\mathbf{c}_i}} |g_c(\mathbf{x})|$ *. For all* $\mathbf{x} \in \mathcal{X}^n$ *and* $\mathbf{c} \in \mathcal{H}^k$*, we have*

$$\frac{\sqrt{n}\sqrt{\max\{b_i, i=1,\ldots,k\}}}{\sqrt{2}} \le \max_i \tilde{\mathcal{A}}_n\left(\mathcal{G}_{\mathbf{c}_i}\right) \le 3\sqrt{n}. \tag{38}$$

Proof. For all j, we have,

$$\begin{aligned}\tilde{\mathcal{A}}_n\left(\mathcal{G}_{\mathbf{c}_j}\right) &= \sup_{\mathbf{X}\in\mathcal{X}^n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g_c\in\mathcal{G}_{\mathbf{c}_j}}\left|\sum_{i=1}^n \sigma_i g_c(\mathbf{x}_i)\right|\right] \\ &\ge \sup_{\mathbf{x}\in\mathcal{X}}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g_c\in\mathcal{G}_{\mathbf{c}_j}}\left|\sum_{i=1}^n \sigma_i g_c(\mathbf{x})\right|\right] \\ &\ge \sup_{\mathbf{x}\in\mathcal{X}, g_c\in\mathcal{G}_{\mathbf{c}_j}}\mathbb{E}_{\boldsymbol{\sigma}}\left|\sum_{i=1}^n \sigma_i g_c(\mathbf{x})\right| &(39) \\ &\ge \frac{\sqrt{n}}{\sqrt{2}}\sup_{\mathbf{x}\in\mathcal{X}, g_c\in\mathcal{G}_{\mathbf{c}_j}}\sqrt{|g_c(\mathbf{x})|} &(40) \\ &= \frac{\sqrt{n}\sqrt{b_j}}{\sqrt{2}}. &(41)\end{aligned}$$

Note that, Eq.(39) is obtained by Jensen's inequality, Eq.(40) is by Eq.(16) of Proposition 3. One can know that

$$\max_i \tilde{\mathcal{A}}_n\left(\mathcal{G}_{\mathbf{c}_i}\right) \ge \frac{\sqrt{n}\sqrt{\max\{b_i, i=1,\ldots,k\}}}{\sqrt{2}}. \tag{42}$$

In the following, we prove the right inequality of Eq.(38).

Note that $\|\phi_{\mathbf{x}}\| \le 1$ for all $\mathbf{x} \in \mathcal{X}$. For $i \in \{1, \dots, k\}$, we have

$$\begin{aligned}\mathcal{A}_n\left(\mathcal{G}_{\mathbf{c}_i}\right) &= \mathbb{E}_{\boldsymbol{\sigma}} \sup_{g_c \in \mathcal{G}_{\mathbf{c}_i}} \left| \sum_{j=1}^{n} \sigma_j g_c(\mathbf{x}_j) \right| \\ =&\mathbb{E}_{\boldsymbol{\sigma}} \sup_{\mathrm{c} \in \mathcal{H}} \left| \sum_{j=1}^{n} \sigma_j \left\|\phi_j - \mathrm{c}\right\|^2 \right| \\ =&\mathbb{E}_{\boldsymbol{\sigma}} \sup_{\mathrm{c} \in \mathcal{H}} \left| \sum_{j=1}^{n} \sigma_j \left[-2\left\langle \phi_j, \mathrm{c}\right\rangle + \|\mathrm{c}\|^2 + \|\phi_j\|^2\right] \right| \\ =&\mathbb{E}_{\boldsymbol{\sigma}} \sup_{\mathrm{c} \in \mathcal{H}} \left| \sum_{j=1}^{n} \sigma_j \left[-2\left\langle \phi_j, \mathrm{c}\right\rangle + \|\mathrm{c}\|^2\right] \right| \\ \le& 2\mathbb{E}_{\boldsymbol{\sigma}} \sup_{\mathrm{c} \in \mathcal{H}} \left| \sum_{j=1}^{n} \sigma_j \left\langle \phi_j, \mathrm{c}\right\rangle \right| + \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\mathrm{c} \in \mathcal{H}} \left| \sum_{j=1}^{n} \sigma_j \|\mathrm{c}\|^2 \right|\end{aligned} \tag{43}$$

According to Eq.(15) of Proposition 3 and $\|c\| \le 1$, we have

$$\mathbb{E}_{\boldsymbol{\sigma}} \sup\nolimits_{c \in \mathcal{H}} \left| \sum_{j=1}^{n} \sigma_j \|c\|^2 \right| \le \mathbb{E}_{\boldsymbol{\sigma}} \left| \sum_{j=1}^{n} \sigma_j \right| \le \sqrt{\mathbb{E}_{\boldsymbol{\sigma}} \left| \sum_{j=1}^{n} \sigma_j \right|^2} \le \sqrt{n}, \tag{44}$$

and

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\sigma}} \sup\nolimits_{c \in \mathcal{H}} \left| \sum_{j=1}^{n} \sigma_j \left\langle \phi_j, c\right\rangle \right| = \mathbb{E}_{\boldsymbol{\sigma}} \sup_{c \in \mathcal{H}} \left| \left\langle \sum_{j=1}^{n} \sigma_j \phi_j, c \right\rangle \right| \\ \le& \mathbb{E}_{\boldsymbol{\sigma}} \left\| \sum_{j=1}^{n} \sigma_j \phi_j \right\| \le \sqrt{\mathbb{E}_{\sigma} \left\| \sum_{j=1}^{n} \sigma_j \phi_j \right\|^2} \le \sqrt{\sum_{i=1}^{n} \|\phi_i\|^2} \\ \le& \sqrt{n}.\end{aligned} \tag{45}$$

Combining Eq.(43), Eq.(44) and Eq.(45), we prove the right inequality of Eq.(38): $\max_i \tilde{\mathcal{A}}_n\left(\mathcal{G}_{\mathbf{c}_i}\right) \le 3\sqrt{n}$. □

Lemma 2. *For every* $\mathbf{c}$ *in* $\mathcal{H}$, *any* $\delta \in (0,1)$, *with probability* $1-\delta$, *we have*

$$\mathbb{E} \sup_{\mathbf{c} \in \mathcal{H}^k} \left| W(\mathbf{c}, \mu_n) - W(\mathbf{c}, \mu) \right| \le \frac{C\sqrt{k}\log^2(\sqrt{n}) + \sqrt{8\log\frac{1}{\delta}}}{\sqrt{n}}, \tag{46}$$

where C *is a constant.*

Proof. Assumed that $\mathbf{x}_1^{'}, \dots, \mathbf{x}_n^{'}$ be an independent copy of $\mathbf{x}_1, \dots, \mathbf{x}_n$, independent of the σ_i's. Therefore, by a standard symmetrization argument, one has the following inequality.

$$\begin{aligned}&\mathbb{E} \sup_{\mathbf{c} \in \mathcal{H}^k} \left| W(\mathbf{c}, \mu_n) - W(\mathbf{c}, \mu) \right| \\ \le& \mathbb{E} \sup_{g_{\mathbf{c}} \in \mathcal{G}_{\mathbf{c}}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left[g_{\mathbf{c}}(\mathbf{x}) - g_{\mathbf{c}}\left(\mathbf{x}^{'}\right)\right] \right| \\ \le& 2\mathbb{E} \sup_{g\mathbf{c} \in \mathcal{G}_{\mathbf{c}}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i g_{\mathbf{c}}(\mathbf{x}) \right| = \frac{2}{n} \mathbb{E} \sup_{g\mathbf{c} \in \mathcal{G}_{\mathbf{c}}} \left| \sum_{i=1}^{n} \sigma_i g_{\mathbf{c}}(\mathbf{x}) \right|.\end{aligned} \tag{47}$$

According to [58], we have, with probability $1-\delta$,

$$\mathbb{E} \sup_{g\mathbf{c} \in \mathcal{G}_{\mathbf{c}}} \left| \sum_{i=1}^{n} \sigma_i g_{\mathbf{c}}(\mathbf{x}) \right| \le \sup_{g\mathbf{c} \in \mathcal{G}_{\mathbf{c}}} \left| \sum_{i=1}^{n} \sigma_i g_{\mathbf{c}}(\mathbf{x}) \right| + \sqrt{2n\log\frac{1}{\delta}}. \tag{48}$$

In the following, we solve the first term of Eq.(48).

Define a minimum function $l: \mathbb{R}^k \to \mathbb{R}$: $l(\boldsymbol{\gamma}) = \min_{i=[k]} \gamma_i$, for all $\boldsymbol{\gamma} \in \mathbb{R}^k$. Assuming that $l(\boldsymbol{\gamma}) \ge l(\boldsymbol{\gamma}^{'})$ and $l(\boldsymbol{\gamma}^{'}) = \gamma_j^{'}$, we have

$$|l(\boldsymbol{\gamma}) - l(\boldsymbol{\gamma}^{'})| = l(\boldsymbol{\gamma}) - \gamma_j^{'} \le \gamma_j - \gamma_j^{'} \le \|\boldsymbol{\gamma} - \boldsymbol{\gamma}^{'}\|_\infty.$$

Therefore, according to Proposition 2, the function $l(\boldsymbol{\gamma})$ is 1-Lipschitz continuous with respect to the L_∞-norm.

Note that $\|\phi_{\mathbf{x}}\| \le 1$. According to Definition 4, one can obtain that $\|c_j\| \le 1$ and

$$g_{c_j}(\mathbf{x}) \le 2\|\phi_{\mathbf{x}}\| + 2\|c_j\| \le 4, \tag{49}$$

for all $\mathbf{x} \in \mathcal{X}$. Therefore, we have $\|g_{\mathbf{c}}(\mathbf{x})\|_\infty = \max_j |g_{c_j}(\mathbf{x})| \le 4$, and $|l(g_{\mathbf{c}}(\mathbf{x}))| = |\min_{j=[k]} g_{c_j}(\mathbf{x})| \le 4$.

According to Proposition 2, let $L = 1, \rho = 4$, and $b = 0.5$, we obtain that

$$\mathcal{A}_n(\mathcal{G}_{\mathbf{c}}) \le C_0 \cdot \sqrt{k} \max_i \tilde{\mathcal{A}}_n\left(\mathcal{G}_{\mathbf{c}_i}\right) \log^2\left(\frac{4n}{\max_i \tilde{\mathcal{A}}_n\left(\mathcal{G}_{\mathbf{c}_i}\right)}\right), \tag{50}$$

where C_0 is a constant, $\mathcal{A}_n(\mathcal{G}_{\mathbf{c}}) = \sup_{g\mathbf{c} \in \mathcal{G}_{\mathbf{c}}} |\sum_{i=1}^n \sigma_i g_{\mathbf{c}}(\mathbf{x})|$.

According to Eq.(49), we obtain that

$$\max\{b_i, i = 1, \dots, k\} \le 4, \tag{51}$$

which is a constant.

Combining Eq.(38), Eq.(50), and Eq.(51), we have

$$\sup_{g\mathbf{c} \in \mathcal{G}_{\mathbf{c}}} \left| \sum_{i=1}^{n} \sigma_i g_{\mathbf{c}}(\mathbf{x}) \right| \le 3C_1\sqrt{kn}\log^2(\sqrt{n}), \tag{52}$$

where C_1 is a constant. Substituting Eq.(48) and Eq.(52) for Eq.(47), we complete this proof. □

Now we begin to formally prove Theorem 4.

Proof. Note that

$$\begin{aligned}&\mathbb{E}\left[W(\hat{\mathbf{c}}_n, \mu)\right] - W^*(\mu) \\ \le& \mathbb{E}\left[W(\hat{\mathbf{c}}_n, \mu) - W(\hat{\mathbf{c}}_n, \mu_n)\right] + \mathbb{E}\left[W(\hat{\mathbf{c}}_n, \mu_n) - W(\mathbf{c}_n, \mu_n)\right] \\ +& \mathbb{E}\left[W(\mathbf{c}_n, \mu_n) - W(\mathbf{c}_n, \mu)\right] + \mathbb{E}\left[W(\mathbf{c}_n, \mu)\right] - W^*(\mu).\end{aligned} \tag{53}$$

According to the standard application of the bounded differences concentration inequality [59] and Eq.(46) in Lemma 2, the first term of Eq.(53) can be written as

$$\begin{aligned}&\mathbb{E}\left[W(\hat{\mathbf{c}}_n, \mu) - W(\hat{\mathbf{c}}_n, \mu_n)\right] \\ \le& \mathbb{E}\left[\sup_{\mathbf{c} \in \mathcal{H}^k} \left(W\left(\mathbf{c}, \mu\right) - W\left(\mathbf{c}, \mu_n\right)\right)\right] \\ \le& \frac{C_1\sqrt{k}\log^2(\sqrt{n}) + \sqrt{8\log\frac{1}{\delta}}}{\sqrt{n}},\end{aligned} \tag{54}$$

where C_1 is a constant.

The second term of Eq.(53) is proved in Eq.(12) of Theorem 3. That is, let $m \ge \frac{4\log n - 2\log\delta}{\varepsilon - \log(1+\varepsilon)}$, with probability at least $1-\delta$, we have $\mathbb{E}\left[W(\hat{\mathbf{c}}_n, \mu_n) - W(\mathbf{c}_n, \mu_n)\right] \le \frac{4\varepsilon}{(1-\varepsilon)^2}$.

The third term of Eq.(53) uses the same principle as the first term. According to Theorem 2, the last term of Eq.(53) is

$$\mathbb{E}\left[W(\mathbf{c}_n, \mu)\right] - W^*(\mu) \le C_2\sqrt{\frac{k}{n}} \log^{\frac{3}{2}+\frac{\delta}{2}}\left(\frac{\sqrt{n}}{3}\right) + C_2\sqrt{\frac{\log\frac{1}{\delta}}{n}},$$

where C_2 is a constant.

So, combining them, we can get that:

$$\begin{aligned}
&\mathbb{E}\left[W(\hat{\mathbf{c}}_n,\mu)\right]-W^*(\mu)\\
\leq&\frac{2C_1\sqrt{k}\log^2(\sqrt{n})+2\sqrt{8\log\frac{1}{\delta}}}{\sqrt{n}}\\
&+C_2\sqrt{\frac{k}{n}}\log^{\frac{3}{2}+\frac{\delta}{2}}\left(\frac{\sqrt{n}}{3}\right)+C_2\sqrt{\frac{\log\frac{1}{\delta}}{n}}+\frac{4\varepsilon}{(1-\varepsilon)^2} \qquad (55)\\
\leq&\frac{C_3\sqrt{k}\log^2(\sqrt{n})+C_4\sqrt{\log\frac{1}{\delta}}}{\sqrt{n}}+\frac{4\varepsilon}{(1-\varepsilon)^2}\\
=&\tilde{\mathcal{O}}\left(\sqrt{\frac{k}{n}}\right)+\mathcal{O}\left(\frac{\varepsilon}{(1-\varepsilon)^2}\right),
\end{aligned}$$

where C_3 and C_4 are constants. Here, we complete this proof.

□

8 Conclusion

Due to the high computational requirements, kernel k-means is not scalable. To get out of the trouble, this paper carefully constructs a novel randomized sketching kernel k-means estimator SKK based on the circulant matrix. To the best of our knowledge, SKK has the same statistical accuracy as exact kernel k-means, the optimal time complexity $\mathcal{O}(nkt+n\log\sqrt{n})$ and the optimal space complexity $\mathcal{O}(n)$. More precisely, taking the sketch dimension of $\sqrt{n}$ is sufficient for optimal statistical accuracy in our approach. Compared to the state-of-the-art approximate kernel k-means estimates, SKK reduces the space and running time at least by factor of $\sqrt{n}$ with the optimal statistical accuracy $\tilde{\mathcal{O}}(\sqrt{\frac{k}{n}})$. Extensive experiments on 12 real datasets show that SKK has significant advantages in running time and memory with satisfactory accuracy. In the future, based on this approach, we can expand to other approximate algorithms with higher accuracy.

Acknowledgments

We appreciate all the anonymous reviewers, ACs, and PCs for their invaluable and constructive comments. This work was supported in part by the Special Research Assistant project of CAS (No.E0YY221), the National Natural Science Foundation of China (No.62106259, No.62076234), the Beijing Municipal Science and Technology Project (No. Z191100007119002), Beijing Outstanding Young Scientist Program (NO.BJJWZYJH012019100020098), Beijing Natural Science Foundation (No. 4222029), the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (2021030199). Thank Intelligent Social Governance Platform and Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" initiative.

References

[1] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl, "Constrained k-means clustering with background knowledge," in *Proceedings of the Eighteenth International Conference on Machine Learning*, vol. 1, 2001, pp. 577–584.

[2] G. Hamerly and C. Elkan, "Learning the k in k-means," in *Advances in neural information processing systems*, 2004, pp. 281–288.

[3] J. Xu and K. Lange, "Power k-means clustering," in *International Conference on Machine Learning*, 2019, pp. 6921–6931.

[4] L. Zhang, W. D. Zhou, and L. C. Jiao, "Kernel clustering algorithm," *Chinese Journal of Computers*, vol. 25, no. 6, pp. 587–590, 2002.

[5] C. Levrard *et al.*, "Nonasymptotic bounds for vector quantization in hilbert spaces," *The Annals of Statistics*, vol. 43, no. 2, pp. 592–619, 2015.

[6] A. Antos, L. Gyorfi, and A. Gyorgy, "Individual convergence rates in empirical vector quantizer design," *IEEE Transactions on Information Theory*, vol. 51, no. 11, pp. 4013–4022, 2005.

[7] A. Maurer and M. Pontil, "k-dimensional coding schemes in hilbert spaces," *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5839–5846, 2010.

[8] P. L. Bartlett, T. Linder, and G. Lugosi, "The minimax distortion redundancy in empirical quantizer design," *IEEE Transactions on Information theory*, vol. 44, no. 5, pp. 1802–1813, 1998.

[9] J. Liu, X. Liu, J. Xiong, Q. Liao, S. Zhou, S. Wang, and Y. Yang, "Optimal neighborhood multiple kernel clustering with adaptive local kernels," *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[10] X. Zhang, B. Chen, H. Sun, Z. Liu, Z. Ren, and Y. Li, "Robust low-rank kernel subspace clustering based on the schatten p-norm and correntropy," *IEEE Transactions on Knowledge and Data Engineering*, 2019.

[11] M. Girolami, "Mercer kernel-based clustering in feature space," *IEEE Transactions on Neural Networks*, vol. 13, no. 3, pp. 780–784, 2002.

[12] Y. Han, K. Yang, Y. Yang, and Y. Ma, "Localized multiple kernel learning with dynamical clustering and matrix regularization," *IEEE Transactions on Neural Networks & Learning Systems*, vol. 29, no. 2, pp. 486–499, 2018.

[13] S. Mehrkanoon, C. Alzate, R. Mall, R. Langone, and J. A. K. Suykens, "Multiclass semisupervised learning based upon kernel spectral clustering," *IEEE Transactions on Neural Networks & Learning Systems*, vol. 26, no. 4, pp. 720–733, 2015.

[14] Z. Wang, B. Du, W. Tu, L. Zhang, and D. Tao, "Incorporating distribution matching into uncertainty for multiple kernel active learning," *IEEE Transactions on Knowledge and Data Engineering*, 2019.

[15] G. Acs, L. Melis, C. Castelluccia, and E. De Cristofaro, "Differentially private mixture of generative neural networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 6, pp. 1109–1121, 2018.

[16] T. Van Laarhoven and E. Marchiori, "Local network community detection with continuous optimization of conductance and weighted kernel k-means," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 5148–5175, 2016.

[17] J. Yuan and Y. Tian, "Practical privacy-preserving mapreduce based k-means clustering over large-scale dataset," *IEEE Transactions on Cloud Computing*, vol. PP, no. 99, pp. 1–1, 2017.

[18] M. Khalilian, N. Mustapha, and N. Sulaiman, "Data stream clustering by divide and conquer approach based on vector model," *Journal of Big Data*, vol. 3, no. 1, p. 1, 2016.

[19] P. Saigal and V. Khanna, "Divide and conquer approach for semi-supervised multi-category classification through localized kernel spectral clustering," *Neurocomputing*, vol. 238, pp. 296–306, 2017.

[20] K. Atarashi, S. Maji, and S. Oyama, "Random feature maps for the itemset kernel," in *The Thirty-Third AAAI Conference on Artificial Intelligence*, 2019, pp. 3199–3206.

[21] N. Pham and R. Pagh, "Fast and scalable polynomial kernels via explicit feature maps," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013, pp. 239–247.

[22] R. Chitta, R. Jin, and A. K. Jain, "Efficient kernel clustering using random fourier features," in *2012 IEEE 12th International Conference on Data Mining*. IEEE, 2012, pp. 161–170.

[23] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in neural information processing systems*, 2008, pp. 1177–1184.

[24] Y. Liu, L. Ding, H. Zhang, W. Ren, X. Zhang, S. Jiang, X. Liu, and W. Wang, "Nearly optimal risk bounds for kernel k-means," *arXiv preprint arXiv:2003.03888*, 2020.

[25] S. Wang, A. Gittens, and M. W. Mahoney, "Scalable kernel k-means clustering with nyström approximation: relative-error bounds," *The Journal of Machine Learning Research*, vol. 20, no. 1, pp. 431–479, 2019.

[26] F. Pourkamali-Anaraki and S. Becker, "Randomized clustered nystrom for large-scale kernel machines," *arXiv preprint arXiv:1612.06470*, 2016.

[27] D. Calandriello and L. Rosasco, "Statistical and computational trade-offs in kernel k-means," in *Advances in Neural Information Processing Systems*, 2018, pp. 9379–9389.

[28] H. Wolfgang, "Integral equations; theory and numerical treatment," 1995.

[29] C. K. Williams and M. Seeger, "Using the nyström method to speed up kernel machines," in *Advances in neural information processing systems*, 2001, pp. 682–688.

[30] 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, no. 2, pp. 214–225, 2004.

[31] L.-L. Liu, X.-B. Wen, and X.-X. Gao, “Segmentation for sar image based on a new spectral clustering algorithm,” in *Life System Modeling and Intelligent Computing*. Springer, 2010, pp. 635–643.

[32] 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, no. Dec, pp. 2153–2175, 2005.

[33] L. He and H. Zhang, “Kernel k-means sampling for nyström approximation,” *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2108–2120, 2018.

[34] F. Can, “Incremental clustering for dynamic information processing,” *Acm Transactions on Information Systems*, vol. 11, no. 2, pp. 143–164, 1993.

[35] P. S. Bradley, U. Fayyad, and C. Reina, “Clustering very large databases using em mixture models,” in *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, vol. 2. IEEE, 2000, pp. 76–80.

[36] G. Biau, L. Devroye, and G. Lugos, “On the performance of clustering in hilbert spaces,” *IEEE Transactions on Information Theory*, vol. 54, no. 2, pp. 781–790, 2008.

[37] R. Yin, Y. Liu, W. Wang, and D. Meng, “Distributed nyström kernel learning with communications,” in *International Conference on Machine Learning*, 2021, pp. 12 019–12 028.

[38] M. W. Mahoney, “Randomized algorithms for matrices and data,” *Advances in Machine Learning & Data Mining for Astronomy*, vol. 3, no. 2, pp. 647–672, 2011.

[39] D. P. Woodruff, “Sketching as a tool for numerical linear algebra,” *Foundations & Trends® in Theoretical Computer Science*, vol. 10, no. 1, pp. 1–157, 2013.

[40] R. Yin, Y. Liu, W. Wang, and D. Meng, “Extremely sparse johnson-lindenstrauss transform: From theory to algorithm,” in *20th IEEE International Conference on Data Mining*. IEEE, 2020, pp. 1376–1381.

[41] R. Yin, Y. Liu, L. Lu, W. Wang, and D. Meng, “Divide-and-conquer learning with nyström: Optimal rate and algorithm.” in *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020, pp. 6696–6703.

[42] 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*, vol. 31, no. 9, pp. 3512–3524, 2020.

[43] A. Rudi, L. Carratino, and L. Rosasco, “Falkon: An optimal large scale kernel method,” in *Advances in Neural Information Processing Systems*, 2017, pp. 3888–3898.

[44] Z. C. Guo, S. B. Lin, and L. Shi, “Distributed learning with multi-penalty regularization,” *Applied & Computational Harmonic Analysis*, p. S1063520317300532, 2017.

[45] A. Rudi, R. Camoriano, and L. Rosasco, “Generalization properties of learning with random features,” in *Advances in Neural Information Processing Systems*, 2016, pp. 3215—3225.

[46] P. J. Davis, *Circulant matrices*. American Mathematical Soc., 2012.

[47] F. Yu, S. Kumar, Y. Gong, and S.-F. Chang, “Circulant binary embedding,” in *International conference on machine learning*, 2014, pp. 946–954.

[48] A. Wilson and H. Nickisch, “Kernel interpolation for scalable structured gaussian processes (kiss-gp),” in *International Conference on Machine Learning*, 2015, pp. 1775–1784.

[49] W. Yin, S. Morgan, J. Yang, and Y. Zhang, “Practical compressive sensing with toeplitz and circulant matrices,” in *Visual Communications and Image Processing*, 2010.

[50] B. Scholkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Adaptive Computation and Machine Learning series, 2018.

[51] B. Schölkopf, A. J. Smola, F. Bach *et al.*, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[52] C. Feng, Q. Hu, and S. Liao, “Random feature mapping with signed circulant matrix projection,” in *International Conference on Artificial Intelligence*, 2015, pp. 3490–3496.

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

[54] M. Friedman, “The use of ranks to avoid the assumption of normality implicit in the analysis of variance,” *Journal of the american statistical association*, vol. 32, no. 200, pp. 675–701, 1937.

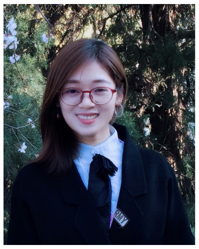
[55] D. J. Foster and A. Rakhlin, “ℓ_∞ vector contraction for rademacher complexity,” *arXiv preprint arXiv:1911.06468*, 2019.

[56] Y. Lei, Ü. Dogan, D.-X. Zhou, and M. Kloft, “Data-dependent generalization bounds for multi-class classification,” *IEEE Transactions on Information Theory*, vol. 65, no. 5, pp. 2995–3021, 2019.

[57] T. Linder, “Learning-theoretic methods in vector quantization,” in *Principles of nonparametric learning*. Springer, 2002, pp. 163–210.

[58] P. L. Bartlett and S. Mendelson, “Rademacher and gaussian complexities: Risk bounds and structural results,” *Journal of Machine Learning Research*, vol. 3, no. Nov, pp. 463–482, 2002.

[59] C. McDiarmid, “On the method of bounded differences,” *Surveys in combinatorics*, vol. 141, no. 1, pp. 148–188, 1989.



Rong Yin received the Ph.D. degree from the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China, and the School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China, in 2020. She is currently an Associate Professor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. Her current research interests include machine learning, data mining, statistical theory, distributed learning, and optimization algorithm.

Yong Liu was born in 1986. He received the Ph.D. degree in computer science from Tianjin University, Tianjin, China, in 2016. He is currently an Associate Professor with the Beijing Key Laboratory of Big Data Management and Analysis Methods, Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China. His current research interests include large-scale kernel methods, large-scale model selection, and machine learning.

Weiping Wang received the Ph.D. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 2006. He is currently a Professor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. His research interests include big data, data security, database, and storage systems. He has more than 100 publications in major journals and international conferences.

Dan Meng received the Ph.D. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1995. He is currently a Professor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. He has more than 100 publications in major journals and international conferences. His current research interests include machine learning, data security, database, and storage systems.