Randomized Sketches for Clustering: Fast and Optimal Kernel k-Means

Rong Yin [1,2], Yong Liu [3,4,*], Weiping Wang [1,2], Dan Meng [1,2]
[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3] Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
[4] Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing, China
yinrong@iie.ac.cn, liuyonggsai@ruc.edu.cn, {wangweiping, mengdan}@iie.ac.cn

Abstract

Kernel k-means is arguably one of the most common approaches to clustering. In this paper, we investigate the efficiency of kernel k-means combined with randomized sketches in terms of both statistical analysis and computational requirements. More precisely, we propose a unified randomized sketches framework to kernel k-means and investigate its excess risk bounds, obtaining the state-of-the-art risk bound with only a fraction of computations. Indeed, we prove that it suffices to choose the sketch dimension $\Omega(\sqrt{n})$ to obtain the same accuracy of exact kernel k-means with greatly reducing the computational costs, for sub-Gaussian sketches, the randomized orthogonal system (ROS) sketches, and Nyström kernel k-means, where n is the number of samples. To the best of our knowledge, this is the first result of this kind for unsupervised learning. Finally, the numerical experiments on simulated data and real-world datasets validate our theoretical analysis.

1 Introduction

Kernel learning is an important field of machine learning Yin et al. (2020b,a, 2021, 2022). Kernel k-means is one of the fundamental approaches in unsupervised learning and has been widely used in numerous applications Zhang & Rudnicky (2002); Dhillon et al. (2004); Chitta et al. (2011); Li & Liu (2021), whose basic idea is to classify similar samples into the same cluster, and there is a large difference between samples in different clusters.

The statistical properties of kernel k-means have been studied for decades, but they may appear to be not sufficient. Consistency of the empirical minimizer of the clustering risk was shown in Abaya & Wise (1984); Pollard (1981, 1982). Rates of convergence and nonasymptotic performance bounds have been considered by Antos (2005); Antos et al. (2005); Bartlett et al. (1998); Linder (2000, 2002). The existing excess risk bounds are mostly dependent upon the dimension of the hypothesis space. For example, in Bartlett et al. (1998), the clustering risk upper bound is $\mathcal{O}(\sqrt{kd/n})$, where n is the number of samples, k is the number of clusters, and d is the dimension of the hypothesis space. Note that the hypothesis space of kernel k-means is typically an infinite-dimensional Hilbert space and the upper bound become useless when d is very large. Subsequently, some researchers deduced dimension-independent upper bounds for kernel k-means Koltchinskii (2006); Biau et al. (2008); Maurer & Pontil (2010); Canas et al. (2012); Levrard et al. (2015); Fefferman et al. (2016); Calandriello & Rosasco (2018); Liu (2021). However, the existing excess clustering risk bounds either have a slow convergence rate $\mathcal{O}(k/\sqrt{n})$ Biau et al. (2008); Calandriello & Rosasco (2018) or require pretty strong assumptions on the underlying distribution or large approximate dimensions m

*Corresponding author.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

to get the faster convergence rate. Specifically, in Calandriello & Rosasco (2018), if the approximate dimension reaches $\Omega(\sqrt{n})$, the clustering risk upper bound is $\mathcal{O}(k/\sqrt{n})$, which is proportional to k. Based on its method, Liu Yong Liu (2021) further improves the convergence rate to $\mathcal{O}(\sqrt{k/n})$, but the corresponding approximate dimension is increased to $\Omega(\sqrt{nk})$. Meanwhile, in order to reduce the approximate dimension to $\Omega(\sqrt{n})$ with the convergence rate unchanged, this paper Liu (2021) requires a stronger assumption of algebraically decreasing eigenvalues of the kernel matrix.

From the perspective of computational requirements, kernel k-means requires manipulating and storing an empirical kernel matrix, which is unfeasible for large-scale problems. Exploring approximate kernel k-means algorithms to scale to large-scale application scenarios has become a subject of recent works, see for example Nyström approximations Williams & Seeger (2001); Fowlkes et al. (2004); Pourkamali-Anaraki et al. (2018); Calandriello & Rosasco (2018); Wang et al. (2019); Liu (2021), randomized sketches Biau et al. (2008); Wang et al. (2019), random features Rahimi & Recht (2008); Chitta et al. (2012); Pham & Pagh (2013); Atarashi et al. (2019), incremental clustering Can (1993); Bradley et al. (2000), and reference therein. This paper focuses on the excess risk bound and computational requirements for kernel k-means. Although there are many studies on the approximate kernel k-means, these approximate works pay little attention to the excess risk of clusters with the exception of Biau et al. (2008); Calandriello & Rosasco (2018); Liu (2021). For example, the works in Wang et al. (2019) establish the $1+\varepsilon$ relative-error bound for randomized sketches kernel k-means instead of excess risk bound. Therefore, in this paper, we mainly introduce the most related approximate kernel k-means with excess risk guarantees. In Biau et al. (2008), they employ the randomized sketches method to project the data in Hilbert space so as to approximate kernel k-means. However, the data in Hilbert space are implicit and infinite-dimensional, and its sketch matrix is dense and unstructured. In Calandriello & Rosasco (2018), the excess risk upper bound is $\mathcal{O}(k/\sqrt{n})$ when the approximate dimension reaches $\Omega(\sqrt{n})$. The upper bound of clustering risk in Biau et al. (2008) and Calandriello & Rosasco (2018) does not reach the optimal $\mathcal{O}(\sqrt{k/n})$ Bartlett et al. (1998). In Liu (2021), the approximate Nyström kernel k-means obtains the risk upper bound $\mathcal{O}(\sqrt{k/n})$ with the approximate dimension $\Omega(\sqrt{nk})$. Although this paper Liu (2021) further reduces the approximate dimension to $\Omega(\sqrt{n})$ by introducing a stronger assumption, this is not universal. In addition, the computational requirements in Biau et al. (2008); Calandriello & Rosasco (2018); Liu (2021) are still high.

Motivated by these issues, in this paper, we focus on improving the statistical analysis and computational approximations of kernel k-means. We propose a randomized sketches framework to kernel k-means and construct three novel and specific examples: sub-Gaussian sketches, the randomized orthogonal system (ROS) sketches, and Nyström kernel k-means. Theoretical analysis shows that the proposed three randomized sketches methods obtain the optimal excess clustering risk upper bound $\mathcal{O}(\sqrt{k/n})$ with the sketch dimension (i.e. approximate dimension) of $\Omega(\sqrt{n})$ (see Theorem 2). To the best of our knowledge, this is the first optimal excess risk bound with the least approximate dimension and no strong assumptions for general approximate kernel k-means. From a computational point of view, the proposed methods lead to massive improvements reducing the time complexity from $\mathcal{O}(n^2kt)$ to at least $\mathcal{O}(n\sqrt{n}+n\sqrt{n}kt)$ and the memory complexity from $\mathcal{O}(n^2)$ to $\mathcal{O}(n\sqrt{n})$, where t is the number of iterations. Moreover, we further derive the similarity bound of approximate solutions in the general case, which can be effectively calculated by k-means++ (see Theorem 3). Experimental results verify and illustrate our theoretical analysis.

The rest of the paper is organized as follows. Section 2 is the background of kernel k-means. Section 3 describes the proposed randomized sketches kernel k-means framework and provides three novel examples. In section 4, we mainly show excess risk bounds of the proposed randomized sketches kernel k-means and the further theoretical analysis in the general case of k-means++. Sections 5 and 6 are the experiments and conclusions.

2 Background

2.1 Notation

Given a sampling distribution μ on an arbitrary input space $\mathcal{X}$ and n samples $\mathcal{S}=\{\mathbf{x}_i\}_{i=1}^n \in \mathcal{X}$ drawn i.i.d. from μ, we denote with $\mu_n(\mathcal{S})=\frac{1}{n}\sum_{i=1}^n \mathbb{I}\{\mathbf{x}_i \in \mathcal{X}\}$ the empirical distribution, where $\mathbb{I}(\cdot)$ is the indicator function. In this paper, we use the feature map $\varphi(\cdot) : \mathcal{X} \to \mathcal{H}$ to map $\mathcal{X}$

into a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ Schölkopf et al. (2002); Scholkopf & Smola (2018), and assume that $\mathcal{H}$ is separable, such that for any $\mathbf{x} \in \mathcal{X}$, we have $\Phi_{\mathbf{x}} = \varphi(\mathbf{x})$. Let $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a mercer kernel. We denote the inner product of $\mathcal{H}$ by $\langle \cdot, \cdot \rangle$, the associated norm by $\| \cdot \|$, the Cartesian product of $\mathcal{H}$ by $\mathcal{H}^k = \otimes_{i=1}^k \mathcal{H}$, and with $\mathbf{K}$ the kernel matrix, where $\mathbf{K}_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi_i, \Phi_j \rangle = \Phi_i^T \Phi_j$. This paper assumes that $\|\Phi_{\mathbf{x}}\| \leq 1$ for any $\mathbf{x} \in \mathcal{X}$.

2.2 Kernel k-Means

Let $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_k]$ be a collection of k centroids from $\mathcal{H}$. We divide the given dataset into k disjoint clusters, each characterized by its centroid $\mathbf{c}_j$. The Voronoi cell associated with a centroid $\mathbf{c}_j$ is defined as Calandriello & Rosasco (2018)

$$\mathcal{C}_j := \{i : j = \arg\min_{s=[k]} \|\Phi_i - \mathbf{c}_s\|^2\}, \tag{1}$$

where $[k] = 1, 2, \ldots, k$. That is, the point Φ_i belongs to the j-th cluster if $\mathbf{c}_j$ is its closest centroid. Now we formalize the criterion used to measure the clustering quality. The empirical squared norm criterion is defined as

$$W(\mathbf{C}, \mu_n) := \frac{1}{n} \sum_{i=1}^{n} \min_{j=[k]} \|\Phi_i - \mathbf{c}_j\|^2, \tag{2}$$

and the expected squared norm criterion is defined as

$$W(\mathbf{C}, \mu) := \mathbb{E}_{\Phi \sim \mu}[\min_{j=[k]} \|\Phi - \mathbf{c}_j\|^2]. \tag{3}$$

The empirical risk minimizer (ERM) is defined as

$$\mathbf{C}_n := \arg\min_{\mathbf{C} \in \mathcal{H}^k} W(\mathbf{C}, \mu_n). \tag{4}$$

The sub-script n in $\mathbf{C}_n$ indicates that it minimizes $W(\mathbf{C}, \mu_n)$ for n samples in $\mathcal{S}$.

In this paper, we bound the excess clustering risk $\mathcal{E}(\mathbf{C}_n)$ of the empirical risk minimizer Calandriello & Rosasco (2018):

$$\mathcal{E}(\mathbf{C}_n) := \mathbb{E}_{\mathcal{S} \sim \mu}[W(\mathbf{C}_n, \mu)] - W^*(\mu), \tag{5}$$

where $W^*(\mu) := \inf_{\mathbf{C} \in \mathcal{H}^k} W(\mathbf{C}, \mu)$ is the optimal clustering risk. In the following, we will ignore the subscript $\mathcal{S} \sim \mu$ if the input dataset $\mathcal{S}$ is clear.

From a computational perspective, one cannot compute $\mathbf{C}_n$ directly, since the points Φ_i in $\mathcal{H}$ cannot be explicitly represented. However, due to the properties of the squared norm criterion and the kernel trick, one can reformulate the objective $W(\cdot, \mu_n)$ of kernel k-means.

Proposition 1 (Proposition 2 of Calandriello & Rosasco (2018)). *Let $\mathbf{K}_{nn} \in \mathbb{R}^{n \times n}$ be the empirical kernel matrix, and $\mathbf{k}_i$ its i-th columns. Then*

$$\begin{aligned} \min_{\mathbf{C} \in \mathcal{H}} W(\mathbf{C}, \mu_n) &= \frac{1}{n} \min_{\nu} \sum_{j=1}^{k} \sum_{i \in \mathcal{C}_j} \left\| \Phi_i - \frac{1}{|\mathcal{C}_j|} \sum_{s \in \mathcal{C}_j} \Phi_s \right\|^2 \\ &= \frac{1}{n} \min_{\nu} \sum_{j=1}^{k} \sum_{i \in \mathcal{C}_j} \left\| \mathbf{k}_i - \frac{1}{|\mathcal{C}_j|} \sum_{s \in \mathcal{C}_j} \mathbf{k}_s \right\|^2 . \end{aligned} \tag{6}$$

This approach constructs an n-dimensional embedding $\mathbf{k}_i$ for each point i, namely the i-th columns of the kernel matrix $\mathbf{K}_{nn}$, which can be explicitly computed, and perfectly preserves $W(\cdot, \mu_n)$ and its minimizer $\mathbf{C}_n$. However, it requires $\mathcal{O}(n^2)$ time and space to construct and store the kernel matrix $\mathbf{K}$, which is not scalable to large-scale scenarios.

2.3 The Existing Excess Risk Bounds of Kernel k-Means

Here we provide the existing upper bound and lower bound of kernel k-means.

According to Bartlett et al. (1998), we know that there exists a collection of centroids $\mathbf{C}_l \in \mathcal{H}^k$ and $\|\Phi_{\mathbf{x}}\| \leq 1$ for any $\mathbf{x} \in \mathcal{X}$, such that

$$\mathbb{E}[W(\mathbf{C}_l, \mu)] - W^*(\mu) = \Omega\left(\sqrt{\frac{k^{1-4/d}}{n}}\right), \tag{7}$$

where d is the dimension of $\Phi_{\mathbf{x}}$. In general, d is very large or even infinite. Therefore, the lower bound of kernel k-means is $\Omega\left(\sqrt{\frac{k}{n}}\right)$. The following is the upper bound of kernel k-means.

Theorem 1 (**Theorem 1 in Liu (2021)**). *If $\|\Phi_{\mathbf{x}}\| \leq 1$ for any $\mathbf{x} \in \mathcal{X}$, then for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have,*

$$\begin{aligned}\mathcal{E}(\mathbf{C}_n) &= \mathbb{E}[W(\mathbf{C}_n, \mu)] - W^*(\mu) \\ &= \mathcal{O}\left(\sqrt{\frac{k}{n}}\log^2(\sqrt{n})\right) = \tilde{\mathcal{O}}\left(\sqrt{\frac{k}{n}}\right).\end{aligned} \tag{8}$$

Note that, $p = \mathcal{O}(u)$ means that there exists an constant c such that $p \leq cu$. $\tilde{\mathcal{O}}(\cdot)$ means to hide the logarithmic terms. This upper bound matches the theoretical lower bound $\Omega\left(\sqrt{\frac{k}{n}}\right)$, and therefore shows that the ERM $\mathbf{C}_n$ achieve an excess risk (nearly) optimal in n.

3 The Proposed Algorithms

Kernel k-means is one of the most popular clustering methods Yin et al. (2020c). However, it is non-scalable to large scenarios due to computing the exact embedding $\mathbf{k}_i$. To reduce the computational requirements, we propose novel approximate embeddings by using randomized sketches. In this section, we propose a unified randomized sketches kernel k-means. In addition, three specific examples of randomized sketches algorithms and the corresponding complexity analysis are provided.

3.1 Framework of Randomized Sketches Kernel k-Means

We consider an approximation based on reducing the original column $\mathbf{k}_i \in \mathbb{R}^n$ to an m-dimensional subspace of $\mathbb{R}^n$, where $m \ll n$ is the sketch dimension. More precisely, the proposed approximation is defined via a sketch matrix $\mathbf{R} \in \mathbb{R}^{m\times n}$, such that the m-dimensional subspace is generated by the span of $\mathbf{R}$. Therefore, the proposed randomized sketches method can be described as:

$$\tilde{\mathbf{K}} = \mathbf{R}\mathbf{K} = \mathbf{S}\mathbf{Q}\mathbf{K} \in \mathbb{R}^{m\times n}, \tag{9}$$

where $\mathbf{K} = \mathbf{K}_{nn}$ and $\mathbf{Q} \in \mathbb{R}^{m\times n}$ is a sampling matrix. The rows of $\mathbf{Q}$ are composed of m rows sampled uniformly from the $n \times n$ identity matrix without replacement. The matrix $\mathbf{S} \in \mathbb{R}^{m\times m}$ is constructed in three ways, which will be introduced in detail in the following section.

Then the unified randomized sketches kernel k-means can be written as (similarly to Proposition 1):

$$\begin{aligned}\bar{\mathbf{C}}_{n,m} &= \arg\min_{\bar{\mathbf{C}}\in\mathbb{R}^{m\times k}} \frac{1}{n}\sum_{i=1}^{n} \min_{j=1,\ldots,k} \left\|\tilde{\mathbf{k}}_i - \bar{\mathbf{c}}_j\right\|^2 \\ &= \frac{1}{n}\min_{\nu} \sum_{j=1}^{k}\sum_{i\in\mathcal{C}_j} \left\|\tilde{\mathbf{k}}_i - \frac{1}{|\mathcal{C}_j|}\sum_{s\in\mathcal{C}_j}\tilde{\mathbf{k}}_s\right\|^2,\end{aligned} \tag{10}$$

where $\tilde{\mathbf{k}}_i$ is the column of the approximate kernel matrix $\tilde{\mathbf{K}}$ in Eq.(9) and $\bar{\mathbf{C}}_{n,m} = [\bar{\mathbf{c}}_1, \ldots, \bar{\mathbf{c}}_k]$ is the empirical clustering centers associated with the m-dimensional $\tilde{\mathbf{k}}_1, \ldots, \tilde{\mathbf{k}}_n$. Each $\bar{\mathbf{c}}_j$ is the mean of those $\tilde{\mathbf{k}}_i$'s in the Voronoi cell $\tilde{\mathcal{C}}_j$.

Define the clustering centers by

$$\tilde{\mathbf{c}}_j = \frac{\sum_{i=1}^{n}\mathbf{k}_i\mathbb{I}_{\{\tilde{\mathbf{k}}_i\in\tilde{\mathcal{C}}_j\}}}{\sum_{i=1}^{n}\mathbb{I}_{\{\tilde{\mathbf{k}}_i\in\tilde{\mathcal{C}}_j\}}}, \quad j = 1, \ldots, k, \tag{11}$$

where $\tilde{\mathbf{C}}_{n,m} = [\tilde{\mathbf{c}}_1, \ldots, \tilde{\mathbf{c}}_j, \ldots, \tilde{\mathbf{c}}_k]$ and $\mathbb{I}_{\{\cdot\}}$ is the indicator function. $\mathbb{I}_{\left\{\tilde{\mathbf{k}}_i \in \tilde{\mathcal{C}}_j\right\}} = 1$ if $\tilde{\mathbf{k}}_i \in \tilde{\mathcal{C}}_j$ and $\mathbb{I}_{\left\{\tilde{\mathbf{k}}_i \in \tilde{\mathcal{C}}_j\right\}} = 0$ otherwise.

The n-dimensional embeddings $\mathbf{k}_i$ are replaced with the lower m-dimensional enbeddings $\tilde{\mathbf{k}}_i$. We can perform any k-means algorithms over $\{\tilde{\mathbf{k}}_1, \ldots, \tilde{\mathbf{k}}_n\}$ and compute the clustering centers in Eq.(11).

Algorithm 1 Unified Randomized Sketches Kernel k-Means

Input: dataset $\mathcal{S} = \{\mathbf{x}_i\}_{i=1}^n$, number of clusters k, kernel parameter, and sketch dimension m.
Output: centroids $\tilde{\mathbf{C}}_{n,m}$.

1: Sample m data points from $\mathcal{S}$ according to the sampling matrix $\mathbf{Q}$ in Eq.(9).
2: Compute the approximate kernel matrix $\hat{\mathbf{K}} \in \mathbb{R}^{m \times n}$ between the m sampling data points and the all data points in $\mathcal{S}$.
3: Construct the matrix $\mathbf{S} \in \mathbb{R}^{m \times m}$ (See Section 3.2 for specific construction methods).
4: Compute $\mathbf{S}\hat{\mathbf{K}}$, namely $\mathbf{S}\hat{\mathbf{K}} = \mathbf{S}(\mathbf{Q}\mathbf{K}) = \tilde{\mathbf{K}}$ is Eq.(9).
5: Perform k-means algorithm over the columns of $\tilde{\mathbf{K}}$.
6: Compute centroids $\tilde{\mathbf{C}}_{n,m}$ in Eq.(11).

The detail of the proposed randomized sketches kernel k-means is shown in Algorithm 1. The proposed algorithm is mainly divided into two parts. The first part is from step 1 to step 4, which is mainly to construct the sketch matrix $\mathbf{R} = \mathbf{S}\mathbf{Q}$ and obtain the variant kernel matrix $\tilde{\mathbf{K}} = \mathbf{S}\mathbf{Q}\mathbf{K}$. The second part is from step 5 to step 6, mainly performing k-means over the columns of $\tilde{\mathbf{K}}$ and obtaining centroids. In step 1, one samples m data points from $\mathcal{S}$ according to the sampling matrix $\mathbf{Q}$. Then, computing the variant kernel matrix $\hat{\mathbf{K}} \in \mathbb{R}^{m \times n}$ by m sampling data points and all n data points. From a mathematical point of view, this step can be expressed as $\hat{\mathbf{K}} = \mathbf{Q}\mathbf{K}$. In step 3, we construct a matrix $\mathbf{S}$, whose specific expression will be given in Section 3.2. This paper provides three different examples of $\mathbf{S}$, which brings different effects in the approximate kernel k-mean algorithms. In step 5, take the columns $\tilde{\mathbf{k}}_i$ of $\tilde{\mathbf{K}} = \mathbf{S}\hat{\mathbf{K}}$ generated in step 4 as the processing objects and execute k-means algorithm on them. Finally, compute the centroids $\tilde{\mathbf{C}}_{n,m}$ in Eq.(11).

3.2 Examples of Randomized Sketches Kernel k-Means

Here, we introduce three examples of randomized sketches kernel k-means, which are constructed by three different matrices $\mathbf{S}$ in Eq.(9). In addition, the detailed complexity analysis of the corresponding three approximate kernel k-means is provided.

Example 1: Sub-Gaussian Sketches Kernel k-Means The first example of approximate kernel k-means is called sub-Gaussian sketches kernel k-means, whose matrix $\mathbf{S} \in \mathbb{R}^{m \times m}$ in Eq.(9) is described by a hash function. Let σ be a hash function and $\sigma(i) \in \{+1, -1\}$ is 2-wise independent hash function. The entries $\mathbf{S}_{i,j} = \sigma(i)/\sqrt{m}$ with a probability of $\frac{1}{\sqrt{n}}$ and $\mathbf{S}_{i,j} = 0$ with a probability of $1 - \frac{1}{\sqrt{n}}$.

Complexity analysis: In the terms of time, we first sample the data by $\mathbf{Q}$, then generate the variant kernel matrix $\hat{\mathbf{K}}$. Therefore, the time cost of computing $\mathbf{S}\hat{\mathbf{K}}$ should be $\mathcal{O}(nm^2)$. However, due to the sparsity of the sub-Gaussian matrix $\mathbf{S}$, we only need to compute the non-zero elements instead of the total elements, which can further reduce the computational requirements of $\mathbf{S}\hat{\mathbf{K}}$ from $\mathcal{O}(nm^2)$ to $\mathcal{O}(\sqrt{n}m^2)$. In the iteration operation of performing k-means algorithm over the columns $\tilde{\mathbf{k}}_i$ of $\tilde{\mathbf{K}}$, one needs $\mathcal{O}(nmkt)$ time. Combining the above, the total time cost of sub-Gaussian sketches kernel k-means is $\mathcal{O}(\sqrt{n}m^2 + nmkt)$. In terms of space, due to the operation of sampling, the key of the space cost is changed to $\hat{\mathbf{K}}$ and $\tilde{\mathbf{K}}$ instead of $\mathbf{K}$. Therefore, the space complexity of the proposed sub-Gaussian sketches kernel k-means is $\mathcal{O}(nm)$.

Example 2: ROS Sketches Kernel k-Means The second example of the random sketches kernel k-means is based on the randomized orthogonal system (ROS) sketches. The corresponding matrix

$\mathbf{S} \in \mathbb{R}^{m\times m}$ in Eq.(9) can be defined as below:

$$\mathbf{S} = \mathbf{DA}, \tag{12}$$

where $\mathbf{D} \in \mathbb{R}^{m\times m}$ is a random diagonal matrix whose entries are i.i.d. Rademacher variables. $\mathbf{A} \in \mathbb{R}^{m\times m}$ is an orthogonal matrix with uniformly bounded entries, for example the Hadamard matrix Wallis (1976) and the discrete Fourier transform matrix. We use the Hadamard matrix in this paper. The Hadamard matrix is defined recursively as: $\mathbf{A}_m = \begin{bmatrix} \mathbf{A}_{m/2} & \mathbf{A}_{m/2} \\ \mathbf{A}_{m/2} & -\mathbf{A}_{m/2} \end{bmatrix}$ with $\mathbf{A}_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$, and $\mathbf{A} = \frac{1}{\sqrt{m}}\mathbf{A}_m$.

Due to the constructed property of the Hadamard matrix $\mathbf{A}$, we can use FFT (Fast Fourier Transform algorithm) to compute the matrix-vector product, such as $\mathbf{Au}$ for any $\mathbf{u} \in \mathbb{R}^m$, whose time complexity is $\mathcal{O}(m \log m)$ instead of $\mathcal{O}(m^2)$. Therefore, in step 4 of Algorithm 1, the computation of $\mathbf{S}\hat{\mathbf{K}}$ can be realized by the fast FFT, which is another way to further reduce the time cost, in addition to the sparsity mentioned above.

Complexity analysis: In terms of time cost, due to the use of the Hadamard matrix, we can compute $\mathbf{S}\hat{\mathbf{K}}$ by FFT, whose time cost is $\mathcal{O}(nm \log m)$. In the iteration operation of k-means algorithm over the columns $\tilde{\mathbf{k}}_i$, the time cost is $\mathcal{O}(nmkt)$. Therefore, the total time cost of ROS sketches kernel k-means is $\mathcal{O}(nm \log m + nmkt)$. In terms of space, the key to the space cost is to store the matrices $\hat{\mathbf{K}}$ and $\tilde{\mathbf{K}}$, whose space requirements is $\mathcal{O}(nm)$. Therefore, the space complexity of the proposed ROS sketches kernel k-means is $\mathcal{O}(nm)$.

Example 3: Nyström Kernel k-Means The third example of the approximate kernel k-means is Nyström kernel k-means, whose matrix $\mathbf{S} \in \mathbb{R}^{m\times m}$ in Eq.(9) can be defined as: $\mathbf{S} = \mathbf{I}$, where $\mathbf{I}$ is an identify matrix. That is, we only use the sample matrix $\mathbf{Q}$ in Eq.(9). Therefore, the proposed Nyström kernel k-means can be converted into:

$$\begin{aligned} \tilde{\mathbf{C}}_{n,m} &= \arg\min_{\tilde{\mathbf{C}}\in\mathbb{R}^{m\times k}} \frac{1}{n}\sum_{i=1}^{n} \min_{j=1,\ldots,k} \left\| \tilde{\mathbf{k}}_i - \tilde{\mathbf{c}}_j \right\|^2 \\ &= \arg\min_{\tilde{\mathbf{C}}\in\mathbb{R}^{m\times k}} \frac{1}{n}\sum_{i=1}^{n} \min_{j=1,\ldots,k} \left\| \tilde{\Phi}_i - \tilde{\mathbf{c}}_j \right\|^2, \end{aligned} \tag{13}$$

where $\tilde{\Phi}_i = \mathbf{\Phi}_m^T \Phi_i$, $\tilde{\mathbf{c}}_j = \mathbf{\Phi}_m^T \mathbf{c}_j$, $\mathbf{\Phi}_m = [\Phi_{\pi(1)}, \ldots, \Phi_{\pi(m)}]$, $\pi(i) \in [1, n]$, and the dictionary (i.e., subset) $\{\Phi_{\pi(i)}\}_{i=1}^m$ is m points Φ_j sampled from $\{\Phi_j\}_{j=1}^n$ through the sampling matrix $\mathbf{Q}$.

Note that, the proposed Nyström kernel k-means can also be understood as a variant ROS, based on the identity matrix as an orthonormal matrix and not using the Rademacher randomization.

Complexity analysis: In terms of time cost, the matrix $\mathbf{S}$ is a scaled identity matrix so that the computation of step 3 and step 4 in Algorithm 1 is not needed. Therefore, the time complexity of Nyström kernel k-means is decided by the iteration operation of k-means algorithm over the columns $\tilde{\mathbf{k}}_i$, which is $\mathcal{O}(nmkt)$. In terms of space, the key is the matrix $\tilde{\mathbf{K}}$. Therefore, the space complexity of the proposed Nyström kernel k-means is $\mathcal{O}(nm)$.

In algorithm, the function of $\mathbf{Q}$ is to reduce the scale of data. The function of $\mathbf{S}$ is to fuse data features. In complexity, the proposed randomized sketches can reduce the time and space complexity. We sample data points according to $\mathbf{Q}$, then generate the variant kernel matrix, instead of generating and processing the kernel matrix directly, which can greatly reduce the time and space complexity. In addition, our matrices $\mathbf{S}$ are structured (in ROS) or sparse (in sub-Gaussian and Nyström), which can speed up kernel k-means by FFT or sparsity. In theoretical analysis, we obtain the optimal excess risk bound with a small sketch dimension based on the proposed randomized sketches, which can further reduce the time and space complexity. Overall, the proposed randomized sketches can greatly reduce the time and space complexity with the optimal excess risk bound.

4 Theoretical Analysis

In this section, we exploit the excess risk bound of the proposed randomized sketches kernel k-means. Theoretical analysis shows that we can improve the computational requirements of kernel k-means using sub-Gaussian, ROS, and Nyström, while maintaining optimal generalization guarantees.

Theorem 2. *If $\|\Phi_{\mathbf{x}}\| \leq 1$ for any $\mathbf{x} \in \mathcal{X}$, $\varepsilon \in (0,1)$, $\delta \in (0,1)$, and, in either one of the three cases of sub-Gaussian, ROS, and Nyström, the sketch dimension is $m = \Omega\left(\frac{4\log n - 2\log\delta}{\varepsilon - \log(1+\varepsilon)}\right)$, then, with probability at least $1-\delta$, we have*

$$\mathbb{E}[W(\tilde{\mathbf{C}}_{n,m}, \mu)] - W^*(\mu) = \tilde{\mathcal{O}}\left(\sqrt{\frac{k}{n}}\right) + \mathcal{O}\left(\frac{\varepsilon}{1-\varepsilon}\right). \tag{14}$$

Remark 1. *From a statistical point of view, let $\varepsilon = 1/\sqrt{n}$, Theorem 2 shows that when the sketch dimension is $m = \Omega(\sqrt{n})$, the proposed randomized sketches (sub-Gaussian, ROS, and Nyström) kernel k-means achieve the same excess risk bound $\tilde{\mathcal{O}}\left(\sqrt{k/n}\right)$ as the exact kernel k-means.*

Remark 2. *From a computational point of view, we can construct the $\sqrt{n}$-dimension randomized sketches simply, which can greatly reduces the total required space from $\mathcal{O}(n^2)$ to $\mathcal{O}(n\sqrt{n})$ and the total required time from $\mathcal{O}(n^2kt)$ to $\mathcal{O}(n\sqrt{n} + n\sqrt{n}kt)$ at least, with the optimal excess risk bound.*

Remark 3. *In Calandriello & Rosasco (2018), when the approximate dimension m is $\Omega(\sqrt{n})$, the excess risk bound can reach $\tilde{\mathcal{O}}\left(k/\sqrt{n}\right)$, which is linearly dependent on k and fail to reach the optimal bound. The corresponding space complexity and time complexity are $\mathcal{O}\left(n\sqrt{n}\right)$ and $\mathcal{O}\left(nkt\sqrt{n} + n^2\right)$, respectively. Compared to it, our proposed methods obtain the better excess risk bound and reduce the time complexity from $\mathcal{O}\left(nkt\sqrt{n} + n^2\right)$ to $\mathcal{O}\left(nkt\sqrt{n} + n\sqrt{n}\right)$ at least. Subsequently, Liu Yong Liu (2021) further improves the excess risk bound of the method in Calandriello & Rosasco (2018) to $\tilde{\mathcal{O}}\left(\sqrt{k/n}\right)$, but the corresponding approximate dimension m is increased to $\Omega(\sqrt{nk})$. Meanwhile, its space complexity and time complexity increase to $\mathcal{O}\left(n\sqrt{nk}\right)$ and $\mathcal{O}\left(nkt\sqrt{nk} + n^2k\right)$. Compared to it, the proposed methods reduce the time complexity from $\mathcal{O}\left(nkt\sqrt{nk} + n^2k\right)$ to $\mathcal{O}\left(nkt\sqrt{n} + n\sqrt{n}\right)$ at least and reduce the space complexity from $\mathcal{O}\left(n\sqrt{nk}\right)$ to $\mathcal{O}\left(n\sqrt{n}\right)$ while maintaining the optimal excess risk bound $\tilde{\mathcal{O}}\left(\sqrt{k/n}\right)$ and the smaller $m = \Omega(\sqrt{n})$. To the best of our knowledge, the proposed methods are the first time that they are always possible to maintain the optimal excess risk bound $\tilde{\mathcal{O}}\left(\sqrt{k/n}\right)$ in unsupervised non-parametric problem with smaller $m = \Omega\left(\sqrt{n}\right)$, while greatly reducing the time and space requirements. In Table 1, we show the detail space complexity, time complexity, excess risk bounds, and m of the approximate kernel k-means.*

4.1 Further Results: k-Means++

We adopt the improved kernel k-means++ sampling Lattanzi & Sohler (2019), which has a local search strategy, for the proposed randomized sketches kernel k-mean. Here is its theoretical analysis.

Lemma 1 (Lattanzi & Sohler (2019)). *If C_n^+ is obtained by the improved k-means++ algorithm with a local search strategy Lattanzi & Sohler (2019), then $\mathbb{E}_{\mathcal{J}}\left[W(\mathbf{C}_n^+, \mu_n)\right] \leq \varpi \cdot W(\mathbf{C}_n, \mu_n)$, where ϖ is a constant and $\mathcal{J}$ is the randomness derived from the k-means++ initialization.*

Note that, this is a multiplicative error bound on the empirical risk.

Theorem 3. *Let $C_{n,m}^+$ be obtained by the improved k-means++ algorithm with a local search strategy Lattanzi & Sohler (2019). If $\|\Phi_{\mathbf{x}}\| \leq 1$ for any $\mathbf{x} \in \mathcal{X}$, $\varepsilon \in (0,1)$, $\delta \in (0,1)$, and, in either one of the three cases of sub-Gaussian, ROS, and Nyström, the sketch dimension is $m = \Omega\left(\frac{4\log n - 2\log\delta}{\varepsilon - \log(1+\varepsilon)}\right)$, then, with probability at least $1-\delta$, we have*

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\mathcal{J}}\left[W(\mathbf{C}_{n,m}^+, \mu)\right]\right] = \tilde{\mathcal{O}}\left(\sqrt{\frac{k}{n}} + W^*(\mu)\right) + \mathcal{O}\left(\frac{\varepsilon}{1-\varepsilon}\right), \tag{15}$$

where $\mathcal{J}$ is the randomness derived from the k-means++ initialization.

Table 1: Comparison of the approximate kernel k-means. The second and third columns represent the space and time complexity. The fourth and fifth columns represent the excess risk bounds and m.

Approach	Space	Time	Bound	m
Kernel k-Means	$\mathcal{O}\left(n^2\right)$	$\mathcal{O}\left(n^2kt\right)$	$\tilde{\mathcal{O}}\left(\sqrt{\frac{k}{n}}\right)$	/
NytrömCalandriello & Rosasco (2018)	$\mathcal{O}\left(n\sqrt{n}\right)$	$\mathcal{O}\left(nkt\sqrt{n}+n^2\right)$	$\tilde{\mathcal{O}}\left(\frac{k}{\sqrt{n}}\right)$	$\sqrt{n}$
NytrömLiu (2021)	$\mathcal{O}\left(n\sqrt{nk}\right)$	$\mathcal{O}\left(nkt\sqrt{nk}+n^2k\right)$	$\tilde{\mathcal{O}}\left(\sqrt{\frac{k}{n}}\right)$	$\sqrt{nk}$
Sub-Gaussian Sketches (**This Paper**)	$\mathcal{O}\left(n\sqrt{n}\right)$	$\mathcal{O}\left(nkt\sqrt{n}+n\sqrt{n}\right)$	$\tilde{\mathcal{O}}\left(\sqrt{\frac{k}{n}}\right)$	$\sqrt{n}$
ROS Sketches (**This Paper**)	$\mathcal{O}(n\sqrt{n})$	$\mathcal{O}(nkt\sqrt{n}+n\sqrt{n})$	$\tilde{\mathcal{O}}\left(\sqrt{\frac{k}{n}}\right)$	$\sqrt{n}$
Nyström (**This Paper**)	$\mathcal{O}\left(n\sqrt{n}\right)$	$\mathcal{O}\left(nkt\sqrt{n}\right)$	$\tilde{\mathcal{O}}\left(\sqrt{\frac{k}{n}}\right)$	$\sqrt{n}$

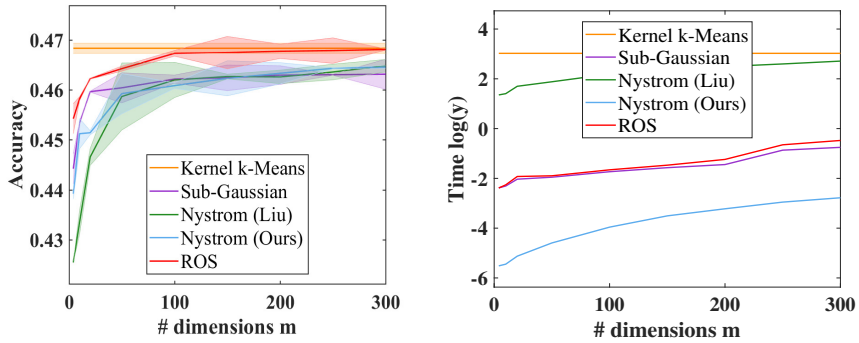

Figure 1: Test accuracy and training time (in seconds) with different dimensions m of kernel k-means, Sub-Gaussian, ROS, Nystrom (Ours), and Nystrom (Liu) on simulated data.

Theorem 3 shows that, if the optimal clustering risk $W^*(\mu)$ is smaller than $\tilde{\mathcal{O}}(\sqrt{k/n})$, the risk of $W(\mathbf{C}^+_{n,m},\mu)$ can reach $\tilde{\mathcal{O}}(\sqrt{k/n})$. Note that ε is small, i.e. $\varepsilon = 1/\sqrt{n}$.

5 Experiments

In this section, we evaluate experimentally our theoretical analysis on both simulated data and real-world data for the proposed methods. The server is 32 cores (2.40GHz) and 32 GB of RAM. The compared methods are the exact kernel k-means, Gaussian Biau et al. (2008), Nyström (Liu) Liu (2021), ROS, Sub-Gaussian, and Nyström (Ours). For the sake of distinguishing, Nyström (Liu) is Nyström Liu (2021) in this paper. Each experiment is repeated 5 times.

5.1 Numerical Experiments on Simulated Data

We conduct the experiments to validate our theoretical analysis of the proposed randomized sketches kernel k-means on simulated data. Now we generate the simulated data. Let $\mathbf{c}_i^* \in \mathbb{R}^8$, $i = [1,k]$, be the clustering centers, where the values of the dimensions are 1 or -1 with the probability of 1/2. The data in ith clustering follows the normal distribution with mean $\mathbf{c}_i^*$ and variance 1. The number of data in each clustering is the same. We use the Gaussian kernel $\kappa(\mathbf{x},\mathbf{x}') = \exp\left(-\|\mathbf{x}-\mathbf{x}'\|^2/8\right)$.

Generating 10,000 samples for training and 10,000 samples for testing. The number of training samples in each clustering is $10000/k$. The accuracy of kernel k-means He & Zhang (2018) on the test set can be written as $\frac{\sum_{i=1}^{\ddot{n}} v(\hat{y},map(y))}{\ddot{n}}$, where y is the solution returned by the (approximate)

Table 2: The datasets used in this paper. Test accuracy and training time (in seconds) of kernel k-means, Gaussian, Nyström (Liu), Sub-Gaussian sketches, ROS sketches, and Nyström (Ours) on real datasets.

Dataset	Instance	Class	Kernel k-Means		Gaussian		Nyström (Liu)	
			Time	Accuracy	Time	Accuracy	Time	Accuracy
dna	2000	3	0.16	0.50±0.01	0.12	0.49± 0.02	0.09	0.50±0.02
segment	2310	7	0.13	0.50±0.02	0.09	0.45±0.03	0.05	0.43±0.01
mushrooms	8124	2	0.56	0.64±0.01	0.32	0.63±0.02	0.11	0.61±0.01
pendigits	10992	10	0.61	0.11 ±0.01	0.34	0.11±0.01	0.21	0.10± 0.02
protein	17766	3	5.07	0.46±0.01	3.16	0.44±0.03	1.09	0.45±0.02
a8a	32561	2	6.47	0.75±0.01	3.21	0.73±0.03	1.12	0.73±0.02
w7a	49749	2	29.7	0.97±0.02	15.3	0.95±0.02	1.36	0.96± 0.01
connect-4	67557	3	0.28	0.61±0.01	0.22	0.60±0.03	0.11	0.59±0.02
covtype	581012	7	/	/	/	/	/	/
Dataset	Instance	Class	Sub-Gaussian (Ours)		ROS (Ours)		Nyström (Ours)	
			Time	Accuracy	Time	Accuracy	Time	Accuracy
dna	2000	3	0.06	0.49± 0.01	0.07	0.50±0.01	0.04	0.50±0.01
segment	2310	7	0.03	0.47±0.03	0.03	0.49±0.01	0.02	0.42±0.01
mushrooms	8124	2	0.04	0.63±0.01	0.04	0.62±0.02	0.03	0.60±0.01
pendigits	10992	10	0.14	0.11±0.01	0.16	0.11± 0.01	0.03	0.11±0.02
protein	17766	3	0.16	0.45±0.01	0.21	0.46±0.01	0.03	0.44±0.02
a8a	32561	2	0.11	0.74±0.01	0.12	0.74±0.02	0.03	0.73±0.02
w7a	49749	2	0.30	0.94±0.02	0.36	0.95± 0.01	0.03	0.97±0.01
connect-4	67557	3	0.05	0.59±0.01	0.06	0.60±0.02	0.03	0.58±0.02
covtype	581012	7	1.02	0.32±0.02	1.36	0.33±0.04	0.66	0.32±0.03

kernel k-means using Lloyd's algorithm Lloyd (1982), $\hat{y}$ is the real label, and $\ddot{n}$ is the number of data in the test set. If $p = q$, $\upsilon(p, q) = 1$, otherwise $\upsilon(p, q) = 0$. $map(\cdot)$ represents the best mapping to match $\hat{y}$ and y. The higher the accuracy, the better the method. The test accuracy and training time of the approximate kernel k-means with different m are given in Figure 1 , which can be summarized as follows: (1) There exists a lower bound of the approximate dimensions $m = \sqrt{n} = 100$. When this lower bound is reached, the accuracy of the proposed methods tends to be stable. This is consistent with our theoretical analysis in Theorem 2. (2) The accuracy of the proposed methods keeps the similar accuracy to the exact kernel k-means. (3) We take the logarithm of the running time (in seconds) in Figure 1. Our methods (ROS, Sub-Gaussian, Nyström) have obvious advantages over other methods in running time. This verifies our complexity analysis.

5.2 Numerical Experiments on Real-World Scenarios

In this subsection, we perform the experiments on the 9 real datasets: dna, segment, mushrooms, pendigits, protein, a8a, w7a, connect-4, and covtype, which are from LIBSVM website [2]. 70 percent of the data in each dataset is used for training experiments, and the rest is used for testing. $m = 150$. The Gaussian kernel is $\exp\left(-\|\mathbf{x} - \mathbf{x}'\|^2/\sigma^2\right)$, where $\sigma = \sqrt{\frac{\sum_{ij}\|x_i - x_j\|^2}{n}}$. The detail of the datasets and experimental results are shown in Table 2. From the above results, we can find that these methods give a similar accuracy as the exact kernel k-means. The proposed methods outperform Nyström (Liu) and Gaussian in time cost, which matches our theoretical analysis. If the training time exceeds 90 seconds or the memory is insufficient, the experiment will be stopped. In the large covtype dataset, kernel k-means, Gaussian, and Nyström (Liu) cannot achieve the experimental results, but our proposed methods can obtain small training time and good accuracy. Those verify the smaller computational requirements of the proposed methods.

6 Conclusions

We propose a unified randomized sketches framework to kernel k-means and provide three specific examples of sub-Gaussian sketches, the randomized orthogonal system (ROS) sketches, and Nyström

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

kernel k-means. Theoretical analysis show that the proposed methods obtain the state-of-the-art risk bound and greatly reduce the computational requirements with sketch dimension $\Omega(\sqrt{n})$. To the best of our knowledge, this is the first optimal excess risk bound with the least approximate dimension and no strong assumptions for general approximate kernel k-means. Moreover, we further derive the similarity optimal bound of approximate solutions in the general case, which can be effectively calculated by k-means++. The extensive experiments illustrate our theoretical analysis.

Acknowledgments and Disclosure of Funding

We appreciate all the anonymous reviewers, ACs, and PCs for their invaluable and constructive comments. This work is supported in part by the Special Research Assistant project of CAS (No.E0YY221-2020000702), the National Natural Science Foundation of China (No.62106259, No.62076234), Beijing Outstanding Young Scientist Program (NO.BJJWZYJH012019100020098), and Beijing Natural Science Foundation (No. 4222029). Thank Intelligent Social Governance Platform and Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" initiative.

References

Abaya, E. F. and Wise, G. L. Convergence of vector quantizers with applications to optimal quantization. *SIAM Journal on Applied Mathematics*, 44(1):183–189, 1984.

Antos, A. Improved minimax bounds on the test and training distortion of empirically designed vector quantizers. *IEEE Transactions on Information Theory*, 51(11):4022–4032, 2005.

Antos, A., Gyorfi, L., and Gyorgy, A. Individual convergence rates in empirical vector quantizer design. *IEEE Transactions on Information Theory*, 51(11):4013–4022, 2005.

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

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Bartlett, P. L., Linder, T., and Lugosi, G. The minimax distortion redundancy in empirical quantizer design. *IEEE Transactions on Information theory*, 44(5):1802–1813, 1998.

Biau, G., Devroye, L., and Lugos, G. On the performance of clustering in hilbert spaces. *IEEE Transactions on Information Theory*, 54(2):781–790, 2008.

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

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

Can, F. Incremental clustering for dynamic information processing. *Acm Transactions on Information Systems*, 11(2):143–164, 1993.

Canas, G. D., Poggio, T., and Rosasco, L. A. Learning manifolds with k-means and k-flats. In *Proceedings of the 25th International Conference on Neural Information Processing Systems-Volume 2*, pp. 2465–2473, 2012.

Chitta, R., Jin, R., Havens, T. C., and Jain, A. K. Approximate kernel k-means: Solution to karge scale kernel clustering. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 895–903. ACM, 2011.

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

Dhillon, I. S., Guan, Y., and Kulis, B. Kernel k-means: Spectral clustering and normalized cuts. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 551–556. ACM, 2004.

Fefferman, C., Mitter, S., and Narayanan, H. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

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

Fowlkes, C., Belongie, S., Chung, F., and Malik, J. Spectral grouping using the nystrom method. *IEEE transactions on pattern analysis and machine intelligence*, 26(2):214–225, 2004.

He, L. and Zhang, H. Kernel k-means sampling for nyström approximation. *IEEE Transactions on Image Processing*, 27(5):2108–2120, 2018.

Koltchinskii, V. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

Lattanzi, S. and Sohler, C. A better k-means++ algorithm via local search. In *International Conference on Machine Learning*, pp. 3662–3671. PMLR, 2019.

Lei, Y., Dogan, Ü., Zhou, D.-X., and Kloft, M. Data-dependent generalization bounds for multi-class classification. *IEEE Transactions on Information Theory*, 65(5):2995–3021, 2019.

Levrard, C. et al. Nonasymptotic bounds for vector quantization in hilbert spaces. *The Annals of Statistics*, 43(2):592–619, 2015.

Li, S. and Liu, Y. Sharper generalization bounds for clustering. In *International Conference on Machine Learning*, pp. 6392–6402. PMLR, 2021.

Linder, T. On the training distortion of vector quantizers. *IEEE Transactions on Information Theory*, 46(4):1617–1623, 2000.

Linder, T. Learning-theoretic methods in vector quantization. In *Principles of nonparametric learning*, pp. 163–210. Springer, 2002.

Liu, Y. Refined learning bounds for kernel and approximate k-means. *Advances in Neural Information Processing Systems*, 34:6142–6154, 2021.

Lloyd, S. P. Least squares quantization in pcm. *IEEE Transactions on Information theory*, 28(2): 129–137, 1982.

Maurer, A. and Pontil, M. k-dimensional coding schemes in hilbert spaces. *IEEE Transactions on Information Theory*, 56(11):5839–5846, 2010.

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

Pollard, D. Strong consistency of k-means clustering. *The Annals of Statistics*, pp. 135–140, 1981.

Pollard, D. Quantization and the method of k-means. *IEEE Transactions on Information theory*, 28 (2):199–205, 1982.

Pourkamali-Anaraki, F., Becker, S., and Wakin, M. B. Randomized clustered nystrom for large-scale kernel machines. In *Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 3960–3967, 2018.

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

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

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

Wallis, J. S. On the existence of hadamard matrices. *Journal of Combinatorial Theory, Series A*, 21 (2):188–195, 1976.

Wang, S., Gittens, A., and Mahoney, M. W. Scalable kernel k-means clustering with nyström approximation: relative-error bounds. *The Journal of Machine Learning Research*, 20(1):431–479, 2019.

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

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

Yin, R., Liu, Y., Wang, W., and Meng, D. Sketch kernel ridge regression using circulant matrix: Algorithm and theory. *IEEE transactions on neural networks and learning systems*, 31(9):3512–3524, 2020b.

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

Yin, R., Liu, Y., Wang, W., and Meng, D. Distributed nyström kernel learning with communications. In *International Conference on Machine Learning*, pp. 12019–12028, 2021.

Yin, R., Liu, Y., and Meng, D. Distributed randomized sketching kernel learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 8883–8891, 2022.

Zhang, R. and Rudnicky, A. I. A large scale clustering scheme for kernel k-means. In *Object recognition supported by user interaction for service robots*, volume 4, pp. 289–292. IEEE, 2002.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We provide Pseudo code, data, and instructions. Lucky to be accepted, the code will be provided.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes]
(b) Did you mention the license of the assets? [N/A]
(c) Did you include any new assets either in the supplemental material or as a URL? [No]
(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Notations and Preliminaries

Let

$$\mathcal{Q}_{\mathbf{C}} := \{q_{\mathbf{C}} = (q_{\mathbf{c}_1}, \ldots, q_{\mathbf{c}_k}) : \mathbf{C} \in \mathcal{H}^k\} \tag{16}$$

be a *k-valued* function with $q_{\mathbf{c}_j}(\mathbf{x}) = \|\phi_{\mathbf{x}} - \mathbf{c}_j\|^2$.

Proposition 2 (L_∞ **Contraction Inequality, Theorem 1 in Foster & Rakhlin (2019)**). *Let $\mathcal{Q} \subseteq \{q : \mathcal{X} \to \mathbb{R}^k\}$ and $l : \mathbb{R}^k \to \mathbb{R}$ be L-Lipschitz with respect to the L_∞ norm, that is $\|l(\boldsymbol{v}) - l(\boldsymbol{v}')\|_\infty \leq L \cdot \|\boldsymbol{v} - \boldsymbol{v}'\|_\infty, \forall \boldsymbol{v}, \boldsymbol{v}' \in \mathbb{R}^k$. For any $b > 0$, there exists a constant $C > 0$ such that if $\max\{|l(q(\mathbf{x}))|, \|q(\mathbf{x})\|_\infty\} \leq \rho$, then*

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

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

Proposition 3 (**Lemma 24(a) in Lei et al. (2019)**). *Let $\eta_1, \ldots, \eta_n \in \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space with $\|\cdot\|$ being the associated norm. Let $\sigma_1, \ldots, \sigma_n$ be a sequence of independent Rademacher variables. Then, we have*

$$\mathbb{E}_{\boldsymbol{\sigma}} \left\| \sum_{i=1}^n \sigma_i \eta_i \right\|^2 \leq \sum_{i=1}^n \|\eta_i\|^2, \tag{17}$$

and

$$\mathbb{E}_{\boldsymbol{\sigma}} \left\| \sum_{i=1}^n \sigma_i \eta_i \right\| \geq \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^n \|\eta_i\|^2}. \tag{18}$$

B Main Lemmas

To prove the main theorems in this paper, we firstly introduce some lemmas.

Lemma 2 (**Lemma 3 in Yin et al. (2020c)**). *Let r_1, r_2 be any two numbers in $\{+1, -1, 0\}$. For any $a, b \in \mathbb{R}$, let $c = \sqrt{(a^2 + b^2)/2}$. Then $\forall M \in \mathbb{R}$ and $s \in \mathbb{N}_+^0$,*

$$\mathbb{E}\left((M + ar_1 + br_2)^{2s}\right) \leq \mathbb{E}\left((M + cr_1 + cr_2)^{2s}\right). \tag{19}$$

Lemma 3. *Let $T \sim \mathcal{N}(0, 1)$, $\|\mathbf{k}_i\|^2 \leq 1$. $\tilde{\mathbf{k}}_{ij}$ is the element in i-th row and j-th column of $\tilde{\mathbf{K}}$. For all $s \in \mathbb{N}_+^0$, we have*

$$\mathbb{E}(\tilde{\mathbf{k}}_{ij}^{2s}) \leq \mathbb{E}(T^{2s}). \tag{20}$$

Proof. Let a "worst-case" unit vector $\mathbf{w} = \frac{1}{\sqrt{n}}(1, \ldots, 1)^T$. For any vector $\mathbf{k}_j$, $\tilde{\mathbf{k}}_{ij} = \mathbf{R}_{\cdot i}\mathbf{k}_j$, where $\mathbf{R}_{\cdot i}$ is the i-th row of $\mathbf{R}$.

If $\mathbf{k}_j = (\mathbf{k}_{1j}, \ldots, \mathbf{k}_{nj})^T$ is such that $\mathbf{k}_{ij}^2 = \mathbf{k}_{tj}^2$ for all i, t, then by symmetry, $\mathbf{R}_{\cdot i}\mathbf{k}_j$ and $\mathbf{R}_{\cdot i}\mathbf{w}$ are identically distributed and this lemma holds trivially.

Otherwise, we can assume without loss of generality, that $\mathbf{k}_{1j}^2 \neq \mathbf{k}_{2j}^2$ and consider the "more balanced" unit vector $\boldsymbol{\theta} = (c, c, \mathbf{k}_{3j}, \ldots, \mathbf{k}_{nj})^T$, where $c = \sqrt{(\mathbf{k}_{1j}^2 + \mathbf{k}_{2j}^2)/2}$.

We first express $\mathbb{E}(\tilde{\mathbf{k}}_{ij}^{2s})$ as a sum of averages over r_{i1}, r_{i2} and apply Eq.(19) in Lemma 2 to get that each term (average) in the sum, where r_{i1} is the element of i-th row and 1-th column of $\mathbf{R}$.

More precisely, in sub-Gaussian case and ROS case,

$$\begin{aligned}
&\mathbb{E}(\tilde{\mathbf{k}}_{ij}^{2s}) = \mathbb{E}((\mathbf{R}_{\cdot i}\mathbf{k}_j)^{2s}) \\
=&m^{-s}\sum_{M}\mathbb{E}\left((M + \mathbf{k}_{1j}r_{i1} + \mathbf{k}_{2j}r_{i2})^{2s}\right)\cdot\mathbb{P}\left[\sum_{t=3}^{n} r_{it}\mathbf{k}_{tj} = \frac{M}{\sqrt{m}}\right] \\
\leq&m^{-s}\sum_{M}\mathbb{E}\left((M + cr_{i1} + cr_{i2})^{2s}\right)\cdot\mathbb{P}\left[\sum_{t=3}^{n} r_{it}\mathbf{k}_{tj} = \frac{M}{\sqrt{m}}\right] \\
=&\mathbb{E}((\mathbf{R}_{\cdot i}\boldsymbol{\theta})^{2s}).
\end{aligned}$$

In Nyström case,

$$\begin{aligned}
&\mathbb{E}(\tilde{\mathbf{k}}_{ij}^{2s}) = \mathbb{E}((\mathbf{R}_{\cdot i}\mathbf{k}_j)^{2s}) \\
=&\sum_{M}\mathbb{E}\left((M + \mathbf{k}_{1j}r_{i1} + \mathbf{k}_{2j}r_{i2})^{2s}\right)\cdot\mathbb{P}\left[\sum_{t=3}^{n} r_{it}\mathbf{k}_{tj} = M\right] \\
\leq&\sum_{M}\mathbb{E}\left((M + cr_{i1} + cr_{i2})^{2s}\right)\cdot\mathbb{P}\left[\sum_{t=3}^{n} r_{it}\mathbf{k}_{tj} = M\right] \\
=&\mathbb{E}((\mathbf{R}_{\cdot i}\boldsymbol{\theta})^{2s}).
\end{aligned}$$

Applying this argument repeatedly yields the lemma, as $\boldsymbol{\theta}$ eventually becomes $\mathbf{w}$, we obtain

$$\mathbb{E}(\tilde{\mathbf{k}}_{ij}^{2s}) = \mathbb{E}((\mathbf{R}_{\cdot i}\mathbf{k}_j)^{2s}) \leq \mathbb{E}((\mathbf{R}_{\cdot i}\mathbf{w})^{2s}). \tag{21}$$

In the following, we prove $\mathbb{E}((\mathbf{R}_{\cdot i}\mathbf{w})^{2s}) \leq \mathbb{E}(T^{2s})$.

To simplify notation, we write $r_{it} = Y_t$. Thus, in sub-Gaussian case and ROS case, $\mathbf{R}_{\cdot i}\mathbf{w} = \frac{1}{\sqrt{nm}}\sum_{t=1}^{n} Y_t$. In Nyström case, $\mathbf{R}_{\cdot i}\mathbf{w} = \frac{1}{\sqrt{n}}\sum_{t=1}^{n} Y_t$.

Let $\{T_i\}_{i=1}^n$ be a family of i.i.d. standard Normal random variables. Then $\sum_{i=1}^n T_i$ is a Normal random variable with variance n. Therefore, $T = \frac{1}{\sqrt{n}}\sum_{i=1}^n T_i$ and $T \sim \mathcal{N}(0, 1)$.

For every $s = 0, 1, \ldots,$

$$\mathbb{E}(T^{2s}) = \frac{1}{(\sqrt{n})^{2s}}\sum_{i_1=1}^{n}\cdots\sum_{i_{2s}=1}^{n}\mathbb{E}(T_{i_1}\cdots T_{i_{2s}}), \tag{22}$$

and in sub-Gaussian and ROS cases

$$\mathbb{E}((\mathbf{R}_{\cdot i}\mathbf{w})^{2s}) = \frac{1}{(\sqrt{nm})^{2s}}\sum_{i_1=1}^{n}\cdots\sum_{i_{2s}=1}^{n}\mathbb{E}(Y_{i_1}\cdots Y_{i_{2s}}),$$

in Nyström case

$$\mathbb{E}((\mathbf{R}_{\cdot i}\mathbf{w})^{2s}) = \frac{1}{(\sqrt{n})^{2s}}\sum_{i_1=1}^{n}\cdots\sum_{i_{2s}=1}^{n}\mathbb{E}(Y_{i_1}\cdots Y_{i_{2s}}).$$

To prove this lemma, in the following, we will prove that for every value assignment to the indices $i_1, \ldots, i_{2s}$,

$$\mathbb{E}(Y_{i_1} \cdots Y_{i_{2s}}) \leq \mathbb{E}(T_{i_1} \cdots T_{i_{2s}}). \tag{23}$$

In sub-Gaussian case:

Let $V = \langle v_1, v_2, \ldots, v_{2s} \rangle$ be the value assignment considered. For $i \in \{1, \ldots, n\}$, let $c_V(i)$ be the number of times that i appears in V. Observe that if for some i, $c_V(i)$ is odd then the expectations appearing in Eq.(22) are 0, since $\{Y_i\}_{i=1}^n$ and $\{T_i\}_{i=1}^n$ are independent families and $\mathbb{E}(Y_i) = \mathbb{E}(T_i) = 0$ for all i. Thus, we can assume that there exists a set $\{j_1, j_2, \ldots, j_p\}$ of indices and corresponding values $\{l_1, l_2, \ldots, l_p\}$ such that

$$\mathbb{E}(T_{i_1} \cdots T_{i_{2s}}) = \mathbb{E}(T_{j_1}^{2l_1} T_{j_2}^{2l_2} \cdots T_{j_p}^{2l_p})$$

and

$$\mathbb{E}(Y_{i_1} \cdots Y_{i_{2s}}) = \mathbb{E}(Y_{j_1}^{2l_1} Y_{j_2}^{2l_2} \cdots Y_{j_p}^{2l_p}).$$

Note that since the indices $j_1, j_2, \cdots, j_p$ are distinct, $\{T_{j_t}\}_{t=1}^p$ and $\{Y_{j_t}\}_{t=1}^p$ are families of i.i.d. Therefore,

$$\mathbb{E}(T_{i_1} \cdots T_{i_{2s}}) = \mathbb{E}(T_{j_1}^{2l_1}) \times \cdots \times \mathbb{E}(T_{j_p}^{2l_p}) \tag{24}$$

and

$$\mathbb{E}(Y_{i_1} \cdots Y_{i_{2s}}) = \mathbb{E}(Y_{j_1}^{2l_1}) \times \cdots \times \mathbb{E}(Y_{j_p}^{2l_p}).$$

So, in order to prove Eq.(23) it suffices to prove that for every $l = 0, 1, \ldots$

$$\mathbb{E}(Y_1^{2l}) \leq \mathbb{E}(T_1^{2l}).$$

We know that $(2l)$-th moment of $\mathcal{N}(0,1)$ is

$$(2l-1)!! = (2l)!/(l!2^l) \geq 1. \tag{25}$$

For all $l \geq 0$, we have $\mathbb{E}(Y_1^{2l}) \leq 1$. Therefore, we have $\mathbb{E}(Y_{i_1} \cdots Y_{i_{2s}}) \leq \mathbb{E}(T_{i_1} \cdots T_{i_{2s}})$.

In ROS and Nyström cases:

$\{Y_i\}_{i=1}^n$ is a family of i.i.d. One knows that

$$\mathbb{E}(Y_{i_1} \cdots Y_{i_{2s}}) = \mathbb{E}(Y_{i_1}) \times \cdots \times \mathbb{E}(Y_{i_{2s}}).$$

Combining $-1 \leq \mathbb{E}(Y_{i_1}) \leq 1$, Eq.(24), and Eq.(25), we know that $\mathbb{E}(Y_{i_1} \cdots Y_{i_{2s}}) \leq \mathbb{E}(T_{i_1} \cdots T_{i_{2s}})$.

Here, we complete the proof of $\mathbb{E}(Y_{i_1} \cdots Y_{i_{2s}}) \leq \mathbb{E}(T_{i_1} \cdots T_{i_{2s}})$ and $\mathbb{E}((\mathbf{R}_{\cdot i}\mathbf{w})^{2s}) \leq \mathbb{E}(T^{2s})$.

Combining $\mathbb{E}((\mathbf{R}_{\cdot i}\mathbf{w})^{2s}) \leq \mathbb{E}(T^{2s})$ and Eq.(21), we obtain $\mathbb{E}(\tilde{\mathbf{k}}_{ij}^{2s}) \leq \mathbb{E}(T^{2s})$. □

Lemma 4. *For all $h \in [0, m/2)$, and $\|\mathbf{k}_i\|^2 \leq 1$, we have*

$$\mathbb{E}(\exp(h\tilde{\mathbf{k}}_{ij}^2)) \leq \frac{1}{\sqrt{1-2h/m}}, \tag{26}$$

and

$$\mathbb{E}(\tilde{\mathbf{k}}_{ij}^4) \leq 3/m^2. \tag{27}$$

Proof. According to Lemma 3, we know

$$\mathbb{E}(\tilde{\mathbf{k}}_{ij}^4) \leq \mathbb{E}(T^4), \tag{28}$$

while

$$\mathbb{E}(T^4) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp(-\lambda^2/2) \left(\lambda^4/m^2\right) m\lambda = 3/m^2.$$

The following will prove Eq.(26).

For any real-valued random variable U and for all h such that $\mathbb{E}(\exp(hU^2))$ is bounded. According to the Monotone Convergence Theorem (MCT), we get the formula

$$\mathbb{E}(\exp(hU^2)) = \mathbb{E}\left(\sum_{t=0}^{\infty}\frac{(hU^2)^t}{t!}\right) = \sum_{t=0}^{\infty}\frac{h^t}{t!}\mathbb{E}(U^{2t}).$$

Here we obtain

$$\begin{aligned}\mathbb{E}(\exp(hT^2)) =& \int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\exp(-\lambda^2/2)\exp(h\lambda^2/m)m\lambda \\ =& \frac{1}{\sqrt{1-2h/m}} = \sum_{t=0}^{\infty}\frac{h^t}{t!}\mathbb{E}(T^{2t}) \\ \geq& \sum_{t=0}^{\infty}\frac{h^t}{t!}\mathbb{E}(\tilde{\mathbf{k}}_{ij}^{2t}) = \mathbb{E}(\exp(h\tilde{\mathbf{k}}_{ij}^2)).\end{aligned} \tag{29}$$

For converge, we take $h \in [0, m/2)$ and apply the MCT in Eq.(38). Therefore, we have $\mathbb{E}(\exp(h\tilde{\mathbf{k}}_{ij}^2)) \leq \frac{1}{\sqrt{1-2h/m}}$, for $h \in [0, m/2)$. This proof logic is similar to Yin et al. (2020c). □

Lemma 5. *Let $\mathcal{S}$ be an arbitrary set of n samples in $\mathcal{X}$ and $\mathbf{K} \in \mathbb{R}^{n\times n}$ be its kernel matrix. The i-th column of $\mathbf{K}$ is represented by $\mathbf{k}_i$. Given $\varepsilon, \delta \in (0,1)$, let*

$$m = \Omega\left(\frac{4\log n - 2\log\delta}{\varepsilon - \log(1+\varepsilon)}\right), \tag{30}$$
$$\tilde{\mathbf{K}} = \mathbf{R}\mathbf{K} \in \mathbb{R}^{m\times n},$$

and $\mathbf{R}$ be a $m \times n$ random matrix in one of the three cases of sub-Gaussian, ROS, and Nyström. And let $f : \mathbb{R}^n \to \mathbb{R}^m$ map the i-th column of $\mathbf{K}$ to the i-th column of $\tilde{\mathbf{K}}$.

For all $\mathbf{k}_i, \mathbf{k}_j \in \mathbf{K}$, with probability at least $1-\delta$, we have,

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

Proof. For arbitrary $h > 0$, according to Markov's inequality we get

$$\begin{aligned}\mathbb{P}\Big[\frac{\sum_{i=1}^m(\mathbf{R}_{\cdot i}\cdot\mathbf{k}_j^T)^2}{\|\mathbf{k}_j\|^2} > 1+\varepsilon\Big] =& \mathbb{P}\Big[\exp(h\frac{\sum_{i=1}^m(\mathbf{R}_{\cdot i}\cdot\mathbf{k}_j^T)^2}{\|\mathbf{k}_j\|^2}) > \exp(h(1+\varepsilon))\Big] \\ <& \mathbb{E}\Big(\exp(h\frac{\sum_{i=1}^m(\mathbf{R}_{\cdot i}\cdot\mathbf{k}_j^T)^2}{\|\mathbf{k}_j\|^2})\Big)\exp\Big(-h(1+\varepsilon)\Big).\end{aligned} \tag{31}$$

Let $\|\mathbf{k}_1\|^2 = 1$, we have:

$$\begin{aligned}\mathbb{E}\Big(\exp(h\frac{\sum_{i=1}^m(\mathbf{R}_{\cdot i}\cdot\mathbf{k}_j^T)^2}{\|\mathbf{k}_j\|^2})\Big) =& \mathbb{E}\Big(\prod_{i=1}^m\exp(h\frac{(\mathbf{R}_{\cdot i}\cdot\mathbf{k}_j^T)^2}{\|\mathbf{k}_j\|^2})\Big) \\ =& \left(\mathbb{E}\left(\exp(h\frac{(\mathbf{R}_{\cdot i}\cdot\mathbf{k}_1^T)^2}{\|\mathbf{k}_1\|^2})\right)\right)^m \\ =& \left(\mathbb{E}\left(\exp(h(\mathbf{R}_{\cdot i}\cdot\mathbf{k}_1^T)^2)\right)\right)^m.\end{aligned} \tag{32}$$

According to Eq.(26) of Lemma 4, we have

$$\mathbb{E}\left(\exp(h(\mathbf{R}_{\cdot i}\cdot\mathbf{k}_1^T)^2)\right) \leq \frac{1}{\sqrt{1-2h/m}}. \tag{33}$$

Let $h = \frac{m\varepsilon}{2(1+\varepsilon)} < \frac{m}{2}$. Taking Eq.(31), Eq.(32), and Eq.(33) to Eq.(34), for any $0 < \varepsilon < 1$, we obtain that

$$\begin{aligned}\mathbb{P}\Big[\frac{\sum_{i=1}^m(\mathbf{R}_{\cdot i}\cdot\mathbf{k}_j^T)^2}{\|\mathbf{k}_j\|^2} > 1+\varepsilon\Big] <& \left(\mathbb{E}\left(\exp(h(\mathbf{R}_{\cdot i}\cdot\mathbf{k}_1^T)^2)\right)\right)^m\exp\Big(-h(1+\varepsilon)\Big) \\ \leq& \Big(\frac{1}{\sqrt{1-2h/m}}\Big)^m\exp\Big(-h(1+\varepsilon)\Big) \\ =& \Big(\frac{1}{1+\varepsilon}\Big)^{-m/2}\exp\Big(\frac{-m\varepsilon}{2}\Big).\end{aligned} \tag{34}$$

Similarly, for arbitrary $h > 0$ and $0 < \varepsilon < 1$, we have

$$\begin{aligned}\mathbb{P}\Big[\frac{\sum_{i=1}^{m}(\mathbf{R}_{\cdot i}\cdot\mathbf{k}_j^T)^2}{\|\mathbf{k}_j\|^2} < 1-\varepsilon\Big] =& \mathbb{P}\Big[\exp(h\frac{\sum_{i=1}^{m}(\mathbf{R}_{\cdot i}\cdot\mathbf{k}_j^T)^2}{\|\mathbf{k}_j\|^2}) < \exp(h(1-\varepsilon))\Big] \\ <& \mathbb{E}\Big(\exp(-h\frac{\sum_{i=1}^{m}(\mathbf{R}_{\cdot i}\cdot\mathbf{k}_j^T)^2}{\|\mathbf{k}_j\|^2})\Big)\exp\Big(h(1-\varepsilon)\Big) \\ =& \big(\mathbb{E}\left(\exp(-h(\mathbf{R}_{\cdot i}\cdot\mathbf{k}_1^T)^2)\right)\big)^m \exp\Big(h(1-\varepsilon)\Big).\end{aligned} \tag{35}$$

By expanding $\exp(-h(\mathbf{R}_{\cdot i}\cdot\mathbf{k}_1^T)^2)$, we get that

$$\begin{aligned}&\mathbb{P}\Big[\frac{\sum_{i=1}^{m}(\mathbf{R}_{\cdot i}\cdot\mathbf{k}_j^T)^2}{\|\mathbf{k}_j\|^2} < 1-\varepsilon\Big] \\ <&\Big(\mathbb{E}\Big(1-h(\mathbf{R}_{\cdot i}\cdot\mathbf{k}_1^T)^2+\frac{(-h(\mathbf{R}_{\cdot i}\cdot\mathbf{k}_1^T)^2)^2}{2!}\Big)\Big)^m\exp\Big(h(1-\varepsilon)\Big) \\ =&\left(1-h\mathbb{E}((\mathbf{R}_{\cdot i}\cdot\mathbf{k}_1^T)^2)+\frac{h^2}{2}\mathbb{E}((\mathbf{R}_{\cdot i}\cdot\mathbf{k}_1^T)^4)\right)^m\exp\Big(h(1-\varepsilon)\Big).\end{aligned} \tag{36}$$

According to Eq.(27) in Lemma 4, we know

$$\mathbb{E}((\mathbf{R}_{\cdot i}\cdot\mathbf{k}_1^T)^4) \le 3/m^2. \tag{37}$$

According to Eq.(38), we have

$$\frac{1}{\sqrt{1-2h/m}} \ge \sum_{t=0}^{\infty}\frac{h^t}{t!}\mathbb{E}(\tilde{\mathbf{k}}_{ij}^{2t}) \ge \sum_{t=0}^{1}\frac{h^t}{t!}\mathbb{E}(\tilde{\mathbf{k}}_{ij}^{2t}) = 1+h\mathbb{E}(\tilde{\mathbf{k}}_{ij}^2) = 1+h\mathbb{E}((\mathbf{R}_{\cdot i}\cdot\mathbf{k}_1^T)^2). \tag{38}$$

So, one can obtain

$$\mathbb{E}((\mathbf{R}_{\cdot i}\cdot\mathbf{k}_1^T)^2) \le \frac{1}{h}(\frac{1}{\sqrt{1-2h/m}}-1). \tag{39}$$

Let $h = \frac{m}{2}\cdot\frac{\varepsilon}{(1+\varepsilon)} < \frac{m}{2}$. Taking Eq.(37) and Eq.(39) into Eq.(36), we obtain that

$$\begin{aligned}&\mathbb{P}\Big[\frac{\sum_{i=1}^{m}(\mathbf{R}_{\cdot i}\cdot\mathbf{k}_j^T)^2}{\|\mathbf{k}_j\|^2} < 1-\varepsilon\Big] \\ <&\Big(1-(\frac{1}{\sqrt{1-2h/m}}-1)+\frac{3h^2}{2m^2}\Big)^m\exp\Big(h(1-\varepsilon)\Big) \\ <&\Big(\frac{1}{1+\varepsilon}\Big)^{-m/2}\exp\Big(\frac{-m\varepsilon}{2}\Big).\end{aligned} \tag{40}$$

Let

$$2\times\Big(\frac{1}{1+\varepsilon}\Big)^{-m/2}\exp\Big(\frac{-m\varepsilon}{2}\Big) \le 2\delta/n^2,$$

we obtain

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

According to Eq.(34) and Eq.(40), we know that, for each of the $\binom{n}{2}$ pairs $\mathbf{k}_i, \mathbf{k}_j$, with the probability of $1-\binom{n}{2}\times 2\delta/n^2 > 1-\delta$, the squared norm of the vector $\mathbf{k}_i - \mathbf{k}_j$ is maintained within a factor of $1\pm\varepsilon$. That is, with the probability at least $1-\delta$, for all $\mathbf{k}_i, \mathbf{k}_j \in \mathbf{K}$,

$$(1-\varepsilon)\|\mathbf{k}_i-\mathbf{k}_j\|^2 \le \|\tilde{\mathbf{k}}_i-\tilde{\mathbf{k}}_j\|^2 \le (1+\varepsilon)\|\mathbf{k}_i-\mathbf{k}_j\|^2. \tag{41}$$

□

Lemma 6. *If constructing the random matrix* $\mathbf{R}$ *in one of the three cases of sub-Gaussian, ROS, and Nyström, by*

$$m = \Omega\left(\frac{4\log n - 2\log\delta}{\varepsilon - \log(1+\varepsilon)}\right),$$

given any $\varepsilon, \delta \in (0,1)$*, then we have, with probability at least* $1-\delta$*,*

$$W(\tilde{\mathbf{C}}_{n,m}, \mu_n) - W(\mathbf{C}_n, \mu_n) \leq \frac{2\varepsilon}{1-\varepsilon}. \tag{42}$$

Proof. We denote by $\bar{\mathbf{C}}_{n,m} = [\bar{\mathbf{c}}_1, \ldots, \bar{\mathbf{c}}_k]$ the empirical clustering centers associated with the m-dimensional embeddings $\tilde{\mathbf{k}}_1, \ldots, \tilde{\mathbf{k}}_n$. Each $\bar{\mathbf{c}}_j$ is the mean of those $\tilde{\mathbf{k}}_i$'s in the Voronoi cell $\tilde{\mathcal{C}}_j$, that is

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

Let $\tilde{\alpha}_j = \sum_{i=1}^n \mathbb{I}_{\left\{\tilde{\mathbf{k}}_i \in \tilde{\mathcal{C}}_j\right\}}$ and $\beta_j = \sum_{i=1}^n \mathbb{I}_{\{\mathbf{k}_i \in \mathcal{C}_j\}}$. We have

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

Combining the optimality of the k-means procedure (Lemma 1 in Linder (2002)), we get

$$W\left(\bar{\mathbf{C}}_{n,m}, \mu_n\right) \leq \sum_{j=1}^k \frac{1}{2n\beta_j} \sum_{i_1, i_2 = 1}^n \left\|\tilde{\mathbf{k}}_{i_1} - \tilde{\mathbf{k}}_{i_2}\right\|^2 \mathbb{I}_{\left\{\left(\mathbf{k}_{i_1}, \mathbf{k}_{i_2}\right) \in \mathcal{C}_j^2\right\}}.$$

Therefore, combining Lemma 5, with probability at least $1-\delta$, we have

$$\begin{aligned}
W\left(\bar{\mathbf{C}}_{n,m}, \mu_n\right) \leq& (1+\varepsilon) \sum_{j=1}^k \frac{1}{2n\beta_j} \sum_{i_1, i_2 = 1}^n \|\mathbf{k}_{i_1} - \mathbf{k}_{i_2}\|^2 \mathbb{I}_{\left\{\left(\mathbf{k}_{i_1}, \mathbf{k}_{i_2}\right) \in \mathcal{C}_j^2\right\}} \\
=& (1+\varepsilon) W\left(\mathbf{C}_n, \mu_n\right).
\end{aligned}$$

Using the similar proof methods, we can obtain

$$(1-\varepsilon) W(\tilde{\mathbf{C}}_{n,m}, \mu_n) \leq W\left(\bar{\mathbf{C}}_{n,m}, \mu_n\right).$$

Note that $W(\mathbf{c}_n, \mu_n) \leq 1$ and $\varepsilon \in (0,1)$. So, we have

$$W(\tilde{\mathbf{C}}_{n,m}, \mu_n) - W(\mathbf{C}_n, \mu_n) \leq \frac{2\varepsilon}{1-\varepsilon} W(\mathbf{C}_n, \mu_n) \leq \frac{2\varepsilon}{1-\varepsilon}.$$

□

Lemma 7. *For* $\delta \in (0,1)$*, with probability* $1-\delta$*, we have*

$$\sup_{q_{\mathbf{C}} \in \mathcal{Q}_{\mathbf{C}}} \left|\sum_{i=1}^n \sigma_i q_{\mathbf{C}}(\mathbf{x})\right| \leq \mathcal{O}\left(\sqrt{kn}\log^2(\sqrt{n})\right). \tag{43}$$

Proof. Note that $\|\phi_{\mathbf{x}}\| \leq 1$ and $\mathcal{Q}_{\mathbf{C}} := \{q_{\mathbf{C}} = (q_{\mathbf{c}_1}, \ldots, q_{\mathbf{c}_k}) : \mathbf{C} \in \mathcal{H}^k\}$is a k-*valued* function with $q_{\mathbf{c}_j}(\mathbf{x}) = \|\phi_{\mathbf{x}} - \mathbf{c}_j\|^2$. Therefore, we have $\|\mathbf{c}_j\| \leq 1$, $q_{\mathbf{c}_j}(\mathbf{x}) \leq 2\|\phi_{\mathbf{x}}\| + 2\|\mathbf{c}_j\| \leq 4$, $\|q_{\mathbf{C}}(\mathbf{x})\|_\infty = \max_j |q_{\mathbf{c}_j}(\mathbf{x})| \leq 4$, and $|l(q_{\mathbf{C}}(\mathbf{x}))| = |\min_{j=[k]} q_{\mathbf{c}_j}(\mathbf{x})| \leq 4$, for all $\mathbf{x} \in \mathcal{X}$.

Due to $q_{\mathbf{c}_j}(\mathbf{x}) \leq 2\|\phi_{\mathbf{x}}\| + 2\|\mathbf{c}_j\| \leq 4$, we get

$$\max\left\{\sup_{\mathbf{x}\in\mathcal{X}}\sup_{q_{\mathbf{C}}\in\mathcal{Q}_{\mathbf{C}_i}}|q_{\mathbf{C}}(\mathbf{x})|, i = 1, \ldots, k\right\} \leq 4. \tag{44}$$

According to L_∞ contraction inequality in Proposition 2, with $L = 1, \rho = 4$, and $b = 1/2$, one can get

$$\sup_{q_{\mathbf{C}}\in\mathcal{Q}_{\mathbf{C}}}\left|\sum_{i=1}^n \sigma_i q_{\mathbf{C}}(\mathbf{x})\right| \leq C \cdot \sqrt{k}\max_i \tilde{\mathcal{B}}_n\left(\mathcal{Q}_{\mathbf{C}_i}\right)\log^2\left(\frac{4n}{\max_i \tilde{\mathcal{B}}_n\left(\mathcal{Q}_{\mathbf{C}_i}\right)}\right), \tag{45}$$

where $\tilde{\mathcal{B}}_n\left(\mathcal{Q}_{\mathbf{C}_i}\right) = \sup_{\mathbf{X}\in\mathcal{X}^n}\mathcal{B}_n\left(\mathcal{Q}_{\mathbf{C}_i}\right)$, $\mathcal{B}_n(\mathcal{Q}_{\mathbf{C}}) = \sup_{q_{\mathbf{C}}\in\mathcal{Q}_{\mathbf{C}}}|\sum_{i=1}^n \sigma_i q_{\mathbf{C}}(\mathbf{x})|$, and C is a constant.

For all j, we get,

$$\begin{aligned}
\tilde{\mathcal{B}}_n\left(\mathcal{Q}_{\mathbf{C}_j}\right) &= \sup_{\mathbf{X}\in\mathcal{X}^n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{q_{\mathbf{C}}\in\mathcal{Q}_{\mathbf{C}_j}}\left|\sum_{i=1}^n \sigma_i q_{\mathbf{C}}(\mathbf{x}_i)\right|\right] \\
&\geq \sup_{\mathbf{x}\in\mathcal{X}}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{q_{\mathbf{C}}\in\mathcal{Q}_{\mathbf{C}_j}}\left|\sum_{i=1}^n \sigma_i q_{\mathbf{C}}(\mathbf{x})\right|\right] \\
&\geq \sup_{\mathbf{x}\in\mathcal{X}, q_{\mathbf{C}}\in\mathcal{Q}_{\mathbf{C}_j}}\mathbb{E}_{\boldsymbol{\sigma}}\left|\sum_{i=1}^n \sigma_i q_{\mathbf{C}}(\mathbf{x})\right| \qquad (46) \\
&\geq \frac{\sqrt{n}}{\sqrt{2}}\sup_{\mathbf{x}\in\mathcal{X}, q_{\mathbf{C}}\in\mathcal{Q}_{\mathbf{C}_j}}\sqrt{|q_{\mathbf{C}}(\mathbf{x})|}. \qquad (47)
\end{aligned}$$

According to Jensen's inequality, we obtain Eq.(46). Eq.(47) is obtained by Eq.(18) of Proposition 3. So, we have

$$\max_i \tilde{\mathcal{B}}_n\left(\mathcal{Q}_{\mathbf{C}_i}\right) \geq \frac{\sqrt{n}\sqrt{\max\left\{\sup_{\mathbf{x}\in\mathcal{X}}\sup_{q_{\mathbf{C}}\in\mathcal{Q}_{\mathbf{C}_i}}|q_{\mathbf{C}}(\mathbf{x})|, i = 1, \ldots, k\right\}}}{\sqrt{2}}. \tag{48}$$

For $i \in \{1, \ldots, k\}$, we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\sigma}}\sup_{q_{\mathbf{C}}\in\mathcal{Q}_{\mathbf{C}_i}}\left|\sum_{j=1}^n \sigma_j q_{\mathbf{C}}(\mathbf{x}_j)\right| &= \mathbb{E}_{\boldsymbol{\sigma}}\sup_{\mathbf{c}\in\mathcal{H}}\left|\sum_{j=1}^n \sigma_j\|\phi_j - \mathbf{c}\|^2\right| \\
&\leq 2\mathbb{E}_{\boldsymbol{\sigma}}\sup_{\mathbf{c}\in\mathcal{H}}\left|\sum_{j=1}^n \sigma_j\langle\phi_j, \mathbf{c}\rangle\right| + \mathbb{E}_{\boldsymbol{\sigma}}\sup_{\mathbf{c}\in\mathcal{H}}\left|\sum_{j=1}^n \sigma_j\|\mathbf{c}\|^2\right|
\end{aligned} \tag{49}$$

According to Eq.(17) of Proposition 3 and $\|\mathbf{c}\| \leq 1$, we have

$$\mathbb{E}_{\boldsymbol{\sigma}\sup_{\mathbf{c}\in\mathcal{H}}}\left|\sum_{j=1}^n \sigma_j\|\mathbf{c}\|^2\right| \leq \mathbb{E}_{\boldsymbol{\sigma}}\left|\sum_{j=1}^n \sigma_j\right| \leq \sqrt{\mathbb{E}_{\boldsymbol{\sigma}}\left|\sum_{j=1}^n \sigma_j\right|^2} \leq \sqrt{n}, \tag{50}$$

and

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\sigma}\sup_{\mathbf{c}\in\mathcal{H}}}\left|\sum_{j=1}^n \sigma_j\langle\phi_j, \mathbf{c}\rangle\right| &= \mathbb{E}_{\boldsymbol{\sigma}}\sup_{\mathbf{c}\in\mathcal{H}}\left|\left\langle\sum_{j=1}^n \sigma_j\phi_j, \mathbf{c}\right\rangle\right| \leq \sqrt{\mathbb{E}_\sigma\left\|\sum_{j=1}^n \sigma_j\phi_j\right\|^2} \\
&\leq \sqrt{\sum_{i=1}^n\|\phi_i\|^2} \leq \sqrt{n}.
\end{aligned} \tag{51}$$

Combining Eq.(49), Eq.(50) and Eq.(51), we obtain

$$\max_i \tilde{\mathcal{B}}_n\left(\mathcal{Q}_{\mathbf{C}_i}\right) \leq 3\sqrt{n}. \tag{52}$$

Combining Eq.(44), Eq.(45), Eq.(48), and Eq.(52), we have

$$\sup_{q_{\mathbf{C}}\in\mathcal{Q}_{\mathbf{C}}}\left|\sum_{i=1}^{n}\sigma_i q_{\mathbf{C}}(\mathbf{x})\right| \leq 3C_1\sqrt{kn}\log^2(\sqrt{n}), \tag{53}$$

where C_1 is a constant. Here we complete this proof. □

Lemma 8. *For $\delta \in (0,1)$, with probability $1-\delta$, we have*

$$\mathbb{E}\left[W(\tilde{\mathbf{C}}_{n,m},\mu) - W(\tilde{\mathbf{C}}_{n,m},\mu_n)\right] \leq \mathcal{O}\left(\frac{\sqrt{k}\log^2(\sqrt{n}) + \sqrt{\log\frac{1}{\delta}}}{\sqrt{n}}\right). \tag{54}$$

Proof. Let $\mathbf{x}_1^{'},\ldots,\mathbf{x}_n^{'}$ be a copy of $\mathbf{x}_1,\ldots,\mathbf{x}_n$, independent of the σ_i's. According to a standard symmetrization argument Bartlett & Mendelson (2002), one can obtain that

$$\begin{aligned}
&\mathbb{E}\sup_{\mathbf{C}\in\mathcal{H}^k}|W(\mathbf{C},\mu) - W(\mathbf{C},\mu_n)| \\
\leq&\mathbb{E}\sup_{q_{\mathbf{C}}\in\mathcal{Q}_{\mathbf{C}}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left[q_{\mathbf{C}}(\mathbf{x}) - q_{\mathbf{C}}\left(\mathbf{x}^{'}\right)\right]\right| \\
\leq&2\mathbb{E}\sup_{q_{\mathbf{C}}\in\mathcal{Q}_{\mathbf{C}}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i q_{\mathbf{C}}(\mathbf{x})\right| = \frac{2}{n}\mathbb{E}\sup_{q_{\mathbf{C}}\in\mathcal{Q}_{\mathbf{C}}}\left|\sum_{i=1}^{n}\sigma_i q_{\mathbf{C}}(\mathbf{x})\right|.
\end{aligned} \tag{55}$$

According to Bartlett & Mendelson (2002), we have, with probability $1-\delta$,

$$\mathbb{E}\sup_{q_{\mathbf{C}}\in\mathcal{Q}_{\mathbf{C}}}\left|\sum_{i=1}^{n}\sigma_i q_{\mathbf{C}}(\mathbf{x})\right| \leq \sup_{q_{\mathbf{C}}\in\mathcal{Q}_{\mathbf{C}}}\left|\sum_{i=1}^{n}\sigma_i q_{\mathbf{C}}(\mathbf{x})\right| + \sqrt{2n\log\frac{1}{\delta}}. \tag{56}$$

According to Lemma 7, we know, with probability $1-\delta$, $\sup_{q_{\mathbf{C}}\in\mathcal{Q}_{\mathbf{C}}}|\sum_{i=1}^{n}\sigma_i q_{\mathbf{C}}(\mathbf{x})| \leq \mathcal{O}\left(\sqrt{kn}\log^2(\sqrt{n})\right)$. Therefore, combining Eq.(55), Eq.(56), and Lemma 7, one can obtain

$$\mathbb{E}\sup_{\mathbf{C}\in\mathcal{H}^k}|W(\mathbf{C},\mu_n) - W(\mathbf{C},\mu)| \leq \mathcal{O}\left(\frac{\sqrt{k}\log^2(\sqrt{n}) + \sqrt{\log\frac{1}{\delta}}}{\sqrt{n}}\right). \tag{57}$$

Note that $\mathbb{E}\left[W(\tilde{\mathbf{C}}_{n,m},\mu) - W(\tilde{\mathbf{C}}_{n,m},\mu_n)\right] \leq \mathbb{E}\sup_{\mathbf{C}\in\mathcal{H}^k}|W(\mathbf{C},\mu) - W(\mathbf{C},\mu_n)|$. So we obtain the result in this lemma.

□

C Proof of Theorem 2

Proof. Note that

$$\begin{aligned}
&\mathbb{E}\left[W(\tilde{\mathbf{C}}_{n,m},\mu)\right] - W^*(\mu) \\
\leq&\underbrace{\mathbb{E}\left[W(\tilde{\mathbf{C}}_{n,m},\mu) - W(\tilde{\mathbf{C}}_{n,m},\mu_n)\right]}_{\textbf{Term-A}} + \underbrace{\mathbb{E}\left[W(\tilde{\mathbf{C}}_{n,m},\mu_n) - W(\mathbf{C}_n,\mu_n)\right]}_{\textbf{Term-B}} \\
&+\underbrace{\mathbb{E}\left[W(\mathbf{C}_n,\mu_n) - W(\mathbf{C}_n,\mu)\right]}_{\textbf{Term-C}} + \underbrace{\mathbb{E}\left[W(\mathbf{C}_n,\mu)\right] - W^*(\mu)}_{\textbf{Term-D}}.
\end{aligned} \tag{58}$$

According to Lemma 8, with probability $1-\delta$, we have **Term-A** $\leq \mathcal{O}\left(\frac{\sqrt{k}\log^2(\sqrt{n})+\sqrt{\log\frac{1}{\delta}}}{\sqrt{n}}\right)$.

According to Lemma 6, for $m = \Omega\left(\frac{4\log n - 2\log\delta}{\varepsilon - \log(1+\varepsilon)}\right)$, we have, with probability at least $1-\delta$, **Term-B** $\leq \frac{2\varepsilon}{1-\varepsilon}$.

Note that **Term-C** $= \mathbb{E}\left[W(\mathbf{C}_n, \mu_n) - W(\mathbf{C}_n, \mu)\right] \leq \mathbb{E}\sup_{\mathbf{C}\in\mathcal{H}^k}|W(\mathbf{C}, \mu_n) - W(\mathbf{C}, \mu)|$. Therefore, according to Eq.(57), we can obtain **Term-C** $\leq \mathcal{O}\left(\frac{\sqrt{k}\log^2(\sqrt{n})+\sqrt{\log\frac{1}{\delta}}}{\sqrt{n}}\right)$.

According to Theorem 1, with probability at least $1-\delta$, we have, **Term-D** $\leq \mathcal{O}\left(\sqrt{\frac{k}{n}}\log^2(\sqrt{n})\right)$.

Here, we complete this proof.

□

D Proof of Theorem 3

Proof. We have

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\mathcal{J}}\left[W(\mathbf{C}^+_{n,m}, \mu)\right]\right] = \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\mathcal{J}}\left[W(\mathbf{C}^+_{n,m}, \mu)\right] - \mathbb{E}_{\mathcal{J}}\left[W(\mathbf{C}^+_{n,m}, \mu_n)\right]\right] + \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\mathcal{J}}\left[W(\mathbf{C}^+_{n,m}, \mu_n)\right]\right]. \tag{59}$$

According to Lemma 1, one can obtain that

$$\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\mathcal{J}}\left[W(\mathbf{C}^+_{n,m}, \mu_n)\right]\right] \leq \varpi\cdot\mathbb{E}[W(\tilde{\mathbf{C}}_{n,m}, \mu_n)] = \varpi\cdot\mathbb{E}[W(\tilde{\mathbf{C}}_{n,m}, \mu_n) - W(\tilde{\mathbf{C}}_{n,m}, \mu)] + \varpi\cdot\mathbb{E}[W(\tilde{\mathbf{C}}_{n,m}, \mu)],$$

Therefore, Eq.(59) can be transferred into

$$\begin{aligned}\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\mathcal{J}}\left[W(\mathbf{C}^+_{n,m}, \mu)\right]\right] \leq & \underbrace{\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\mathcal{J}}\left[W(\mathbf{C}^+_{n,m}, \mu)\right] - \mathbb{E}_{\mathcal{J}}\left[W(\mathbf{C}^+_{n,m}, \mu_n)\right]\right]}_{\textbf{Term-A}} \\ & + \varpi\cdot\underbrace{\mathbb{E}[W(\tilde{\mathbf{C}}_{n,m}, \mu_n) - W(\tilde{\mathbf{C}}_{n,m}, \mu)]}_{\textbf{Term-B}} \\ & + \varpi\cdot\underbrace{\mathbb{E}[W(\tilde{\mathbf{C}}_{n,m}, \mu)]}_{\textbf{Term-C}},\end{aligned}$$

Note that **Term-A** $\leq \mathbb{E}\sup_{\mathbf{C}\in\mathcal{H}^k}|W(\mathbf{C}, \mu_n) - W(\mathbf{C}, \mu)|$. Therefore, according to Eq.(57), we can obtain **Term-A** $\leq \mathcal{O}\left(\frac{\sqrt{k}\log^2(\sqrt{n})+\sqrt{\log\frac{1}{\delta}}}{\sqrt{n}}\right)$.

According to Lemma 8, with probability $1-\delta$, we have **Term-B** $\leq \mathcal{O}\left(\frac{\sqrt{k}\log^2(\sqrt{n})+\sqrt{\log\frac{1}{\delta}}}{\sqrt{n}}\right)$.

According to Theorem 2, we have **Term-C** $= \mathbb{E}\left[W(\tilde{\mathbf{C}}_{n,m}, \mu)\right] \leq W^*(\mu) + \tilde{\mathcal{O}}\left(\sqrt{\frac{k}{n}}\right) + \mathcal{O}\left(\frac{\varepsilon}{1-\varepsilon}\right)$.

We complete this proof.

□