Consistency of Multiple Kernel Clustering

Weixuan Liang [1] **Xinwang Liu** [1] **Yong Liu** [2 3] **Chuan Ma** [4] **Yunping Zhao** [1] **Zhe Liu** [4] **En Zhu** [1]

Abstract

Consistency plays an important role in learning theory. However, in multiple kernel clustering (MKC), the consistency of kernel weights has not been sufficiently investigated. In this work, we fill this gap with a non-asymptotic analysis on the consistency of kernel weights of a novel method termed SimpleMKKM. Under the assumptions of the eigenvalue gap, we give an infinity norm bound as $\widetilde{\mathcal{O}}(k/\sqrt{n})$, where k is the number of clusters and n is the number of samples. On this basis, we establish an upper bound for the excess clustering risk. Moreover, we study the difference of the kernel weights learned from n samples and r points sampled without replacement, and derive its upper bound as $\widetilde{\mathcal{O}}(k\cdot\sqrt{1/r-1/n})$. Based on the above results, we propose a novel strategy with Nyström method to enable SimpleMKKM to handle large-scale datasets with a theoretical learning guarantee. Finally, extensive experiments are conducted to verify the theoretical results and the effectiveness of the proposed large-scale strategy.

1. Introduction

Multiple kernel clustering (MKC) (Zhao et al., 2009) is proposed for better performance by searching for an optimal kernel from several base kernels. In recent years, researchers have made great progress in MKC. Huang et al. (2011) propose the multiple kernel k-means algorithm (MKKM), which unifies all base kernels into a consensus one based on a linear combination. Subsequently, several works (Liu et al., 2016; Du et al., 2015; Liu et al., 2021; Wang et al., 2021) enhance MKKM from different perspectives. Among them, Du et al. (2015) improve the robustness of MKKM by using $\ell_{2,1}$-norm. Liu et al. (2016) introduce a matrix-induced regularization to increase the diversity of the optimal kernel. Wang et al. (2021) improve the two-stage strategy into a single step, further reducing redundancy in kernel fusion. A recently proposed algorithm termed SimpleMKKM (Liu, 2022) greatly promotes the clustering performance by a minimization-maximization framework.

Although there are various improvements in MKC, some vital statistical properties of it are not sufficiently studied, especially the consistency of the kernel weights. We usually say that a learning algorithm is consistent, i.e., the parameters learned from the training set will converge to the parameters from the whole sample space when the training sample number $n \to \infty$. Consistency is an important property in statistical learning, as we can estimate whether the learned parameters are effective by studying the consistency of a learning algorithm. In the existing literature, the consistency of clustering centroids of k-means has been studied in (Pollard, 1981a). Von Luxburg et al. (2008) establish several important results about the consistency of spectral clustering. The consistency of kernel weights in MKC is also a key research problem, as it can be used to derive other important statistical properties, such as excess risk bound. In this paper, we attempt to address this issue. We bound the difference of the weights learned from the training set and the sample space with a non-asymptotic analysis. Under some assumptions about the gap of eigenvalues of the kernel matrix, we establish an infinity bound as $\widetilde{\mathcal{O}}(k/\sqrt{n})$[1], where k is the number of clusters and n is the number of samples. Based on the results of consistency, we derive the excess risk bound of SimpleMKKM.

The difference of the kernel weights learned from the training set and its subset is another interesting problem. Utilizing a concentration inequality for sampling without replacement (Bardenet & Maillard, 2015), this difference can be bounded by $\widetilde{\mathcal{O}}(k\cdot\sqrt{1/r-1/n})$, where r is the number of points sampled without placement from the training set with size n. It is illustrated that the kernel weights learned from the selected subset have a fast convergence rate to the whole training set. Based on this, we propose a new algorithm with Nyström method that can enable Sim-

[1]College of Computer, National University of Defense Technology, Changsha, China [2]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China [3]Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing, China [4]Zhejiang Laboratory, Hangzhou, China. Correspondence to: Xinwang Liu <xinwangliu@nudt.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

[1]$\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic terms.

pleMKKM to handle large-scale datasets. Specifically, we perform SimpleMKKM on the selected subset for a group of approximated kernel weights. Then we make weighted combinations of the base kernel similarity matrices consisting of all n samples and the selected subset. Finally, we use the standard Nyström method to obtain the clustering results. Our algorithm can reduce the complexity of SimpleMKKM from $\mathcal{O}(n^3)$ to be linear with n. Thus it can cluster large-scale datasets. In addition, we derive the excess risk bound of the proposed algorithm for a theoretical learning guarantee. Consequently, when the number of the selected samples is $\Theta(\sqrt{n})$, the proposed algorithm will have a favorable statistical and computational trade-off. By the selection, the excess risk bound is the same as single kernel clustering, which is $\mathcal{O}(k/\sqrt{n})$.

To verify the proposed theoretical results, we conduct experiments on commonly used datasets. The numerical experiments substantiate the correctness of the derived bounds. Moreover, we perform our algorithm on large-scale datasets to verify its effectiveness and efficiency.

The contributions of this paper can be summarized as

1. This paper theoretically analyzes the consistency of the kernel weights of an MKC algorithm, and derives its excess risk bound.

2. This paper studies the difference of the kernel weights learned from the whole training set and its subset. Based on this, a method enabling MKC to handle large-scale datasets is proposed. In addition, the generalization ability of the proposed method is studied, and the optimal number of the selected subset is also given by theoretical analysis.

3. Extensive experiments are conducted to verify the correctness of our theoretical results, as well as the effectiveness and efficiency of the proposed large-scale algorithm.

The paper is organized as follows. Section 2 introduces the notations, general assumptions, and the problem of the consistency of multiple kernel clustering. Section 4 states the main results. Section 5 establishes the excess risk bound of SimpleMKKM. Section 6 proposes the large-scale strategy of SimpleMKKM and derives the corresponding excess risk bound. Section 7 reports the experimental results. Section 8 summarizes the paper and discusses future works.

2. Preliminaries

In this section, we first introduce the main notations and general assumptions. Then, we describe multiple kernel clustering and the consistency problem of the kernel weights.

2.1. Notations and General Assumptions

Mathematical notations. We introduce the used mathematical notations across the whole paper for easy reading. We use $\mathcal{X}$ to represent the sample space, and $\rho(x)$ is the corresponding probability measure. We use $\rho_n(x)$ to denote the empirical distribution, i.e., $\rho_n(x) = \frac{1}{n}$ if point x belongs the training set, otherwise $\rho_n(x) = 0$. The definitions of the asymptotic notations $\mathcal{O}, \Theta$, and Ω can be referred to in Chapter 3 of (Cormen et al., 2022). $g(n) = \mathcal{O}(f(n))$ means $g(n) \leq cf(n)$ for some constant c, and we also write it as $g(n) \precsim f(n)$. $g(n) = \Omega(f(n))$ indicates $g(n) \geq cf(n)$ for some constant c. If there exist two constant c_1, c_2 such that $c_1 f(n) \leq g(n) \leq c_2 f(n)$, we denote that $g(n) = \Theta(f(n))$. $\|A\|$ is the operator norm if A is a matrix or a operator, and if A is a vector, $\|A\|$ denotes the 2-norm.

General Assumptions. The general assumptions partially refer to (Von Luxburg et al., 2008). The sample space $\mathcal{X}$ is supposed to be compact. The base kernel functions $\{K_p(\cdot,\cdot)\}_{p=1}^m$ are bounded, positive-definite, and conjugate symmetric. We assume that $K_p(x, y) \leq 1$, for any $x, y \in \mathcal{X}$ and $p \in [m]$. The elements of training set $S_n = \{x_i\}_{i=1}^n$ are drawn i.i.d. from $\mathcal{X}$ with the distribution ρ. The basic notations used in this paper are summarized in Section C.

2.2. Multiple Kernel Clustering

Multiple kernel clustering (MKC) aims to combine several base kernel matrices into a unified one for better clustering performance. Assume that we have m base kernel functions $\{K_p(\cdot,\cdot)\}_{p=1}^m$, and the corresponding feature map of $K_p(\cdot,\cdot)$ is $\phi_p(\cdot)$. For any point x in sample space $\mathcal{X}$, we can obtain its feature map in multiple kernel scenarios as

$$[\phi_1^\top(x), \cdots, \phi_m^\top(x)]^\top.$$

To reflect the different importance of base kernels, we impose kernel weights $\{\alpha_p\}_{p=1}^m$ on them as

$$[\alpha_1\phi_1^\top(x), \cdots, \alpha_m\phi_m^\top(x)]^\top,$$

where $\sum_{p=1}^m \alpha_p = 1$ and $\alpha_p \geq 0$. Suppose that sample set is $S_n = \{x_i\}_{i=1}^n$ and the kernel matrix computed by the p-th base kernel is $\frac{1}{n}\mathbf{K}_p$. Then the combination of base kernel matrices can be represented by

$$\frac{1}{n}\mathbf{K}_{\boldsymbol{\alpha}} = \frac{1}{n}\sum_{p=1}^m \alpha_p^2 \mathbf{K}_p.$$

In existing MKC algorithms, researchers mainly design an objective function $f_n(\boldsymbol{\alpha})$ w.r.t. the kernel weights $\boldsymbol{\alpha}$ and minimize it to obtain a group of desirable kernel weights. There are two mainstream categories:

1. Coordinate descent based multiple kernel clustering:

$$f_n(\boldsymbol{\alpha}) = \min_{\mathbf{H}^\top\mathbf{H}=\mathbf{I}_k} \frac{1}{n}\mathrm{Tr}(\mathbf{K}_{\boldsymbol{\alpha}}(\mathbf{I}_k - \mathbf{H}\mathbf{H}^\top)),$$

2. Kernel alignment based multiple kernel clustering:

$$f_n(\boldsymbol{\alpha}) = \max_{\mathbf{H}^\top \mathbf{H} = \mathbf{I}_k} \frac{1}{n} \mathrm{Tr}(\mathbf{K}_{\boldsymbol{\alpha}} \mathbf{H} \mathbf{H}^\top). \tag{1}$$

The second method outperforms the first one in terms of clustering performance. Thus we focus on studying the consistency of kernel alignment based MKC, which is termed SimpleMKKM (Liu, 2022). Next, we introduce the objective function when the input training set is the whole sample space $\mathcal{X}$. As $n \to \infty$, the empirical kernel matrix will converge to an integral operator $L_K g(x) := \int_{\mathcal{X}} K(x, y) g(y) d\rho(y)$, where $K(\cdot, \cdot)$ is the corresponding kernel function (Rosasco et al., 2010). Assume that the first k non-zero eigenvalues of the integral operator L_K are $\{\lambda_j\}_{j=1}^k$, and the corresponding eigenvectors are $\{h_j\}_{j=1}^k$. Then, we have

$$h_j(x) = \frac{1}{\lambda_j} \int_{\mathcal{X}} K(x, y) h_j(y) d\rho(y).$$

Moreover,

$$\{h_j\}_{j=1}^k = \operatorname*{argmax}_{\{g_j\}_{j=1}^k \in \Gamma} \sum_{j=1}^k \iint_{\mathcal{X}} K(x, y) g_j(x) g_j(y) d\rho(x) d\rho(y),$$

where Γ denotes the orthonormal constraint on $L^2(\mathcal{X}, \rho)$ space.

For any weights $\boldsymbol{\alpha}$, we assume that $K_{\boldsymbol{\alpha}}(x, y) = \sum_{p=1}^m \alpha_p^2 K_p(x, y)$. When the training set is the sample space $\mathcal{X}$, the objective function of kernel alignment based multiple kernel clustering is

$$f(\boldsymbol{\alpha}) = \max_{\{h_j\}_{j=1}^k \in \Gamma} \sum_{j=1}^k \iint_{\mathcal{X}} K_{\boldsymbol{\alpha}}(x, y) h_j(x) h_j(y) d\rho(x) d\rho(y), \tag{2}$$

where $\{h_j\}_{j=1}^k$ are termed clustering indicator functions.

We denote the first k eigenvalues of $\frac{1}{n}\mathbf{K}$ as $\{\hat{\lambda}_j\}_{j=1}^k$, and the corresponding eigenvectors are $\{\mathbf{h}_j\}_{j=1}^k$. By the definition in (Bengio et al., 2004), the empirical eigenfunctions of operator $\frac{1}{n}\mathbf{K}$ are given by

$$\hat{h}_j(x) = \frac{1}{n\hat{\lambda}_j} \sum_{i=1}^n K(x, x_i) \hat{h}_j(x_i),$$

where $\hat{h}_j(x_i) = \sqrt{n} h_{ij}$, and h_{ij} is the i-the element of $\mathbf{h}_j$. Consequently, the objective in Eq. (1) can be rewritten as

$$f_n(\boldsymbol{\alpha}) = \max_{\{\hat{h}_j\}_{j=1}^k} \frac{1}{n^2} \sum_{i=1}^n \sum_{t=1}^n \sum_{j=1}^k K_{\boldsymbol{\alpha}}(x_i, x_t) \hat{h}_j(x_i) \hat{h}_j(x_t), \tag{3}$$

where $\hat{h}_j(x_i) = \sqrt{n} h_{ij}$, h_{ij} is the i-th row and the j-th column of $\mathbf{H}$, and $\mathbf{H}^\top \mathbf{H} = \mathbf{I}_k$. $\{\hat{h}_j\}_{j=1}^k$ are termed approximated clustering indicator functions.

Key problems. In this paper, we focus on the following two key problems:

1. We denote $\hat{\boldsymbol{\alpha}}_n = \mathrm{argmin}_{\boldsymbol{\alpha}} f_n(\boldsymbol{\alpha})$ and $\boldsymbol{\alpha}^* = \mathrm{argmin}_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha})$ in which f_n, f are given by Eq.(3) and Eq.(2), respectively. To study the consistency of kernel weights, we try to establish a non-asymptotic bound of $\|\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^*\|_\infty$.
2. Suppose that $\hat{\boldsymbol{\alpha}}_r$ is the kernel weights learned from r points, which are sampled from S_n without replacement. We also aim to bound $\|\hat{\boldsymbol{\alpha}}_r - \hat{\boldsymbol{\alpha}}_n\|_\infty$.

2.3. The Optimization of SimpleMKKM

Before stating our results, we first introduce how to optimize the objective function f_n in Eq.(3).

In (Liu, 2022), the author first proves the differentiability of $f_n(\boldsymbol{\alpha})$ and utilizes the reduced gradient descent method to minimize it w.r.t. $\boldsymbol{\alpha}$. With some fixed $u \in [m]$, $\forall p \neq u$, its reduced gradient is

$$[\bigtriangledown f_n(\boldsymbol{\alpha})]_p = \frac{\partial f_n(\boldsymbol{\alpha})}{\partial \alpha_p} - \frac{\partial f_n(\boldsymbol{\alpha})}{\partial \alpha_u}.$$

To satisfy the simplex constraint, we know the gradient of the u-th component is

$$[\bigtriangledown f_n(\boldsymbol{\alpha})]_u = \sum_{p \neq u} \left(\frac{\partial f_n(\boldsymbol{\alpha})}{\partial \alpha_u} - \frac{\partial f_n(\boldsymbol{\alpha})}{\partial \alpha_p} \right),$$

where

$$\frac{\partial f_n(\boldsymbol{\alpha})}{\partial \alpha_p} = \frac{2\alpha_p}{n(m-1)} \mathrm{Tr}\left(\mathbf{K}_p \mathbf{H}^{\boldsymbol{\alpha}} \mathbf{H}^{\boldsymbol{\alpha}\top} \right),$$

when the kernel weights are $\boldsymbol{\alpha}$ and the corresponding clustering indicator matrix is $\mathbf{H}^{\boldsymbol{\alpha}}$. Then, the final descent direction is

$$\hat{d} = \begin{cases} 0, & \text{if } \alpha_p = 0 \text{ and } [\bigtriangledown f_n(\boldsymbol{\alpha})]_p \geq 0, \\ -[\bigtriangledown f_n(\boldsymbol{\alpha})]_p, & \text{if } \alpha_p \geq 0 \text{ and } p \neq u, \\ -[\bigtriangledown f_n(\boldsymbol{\alpha})]_u, & \text{if } p = u. \end{cases}$$

Denoting γ as the learning step length, we can update $\boldsymbol{\alpha}$ by $\boldsymbol{\alpha} = \boldsymbol{\alpha} + \gamma \hat{d}$. For ease of analysis, we let $\gamma \leq c$, where c is some positive constant.

Similarly, because the eigenfunctions corresponding to the first k eigenvalues of the operator $L_{K_{\boldsymbol{\alpha}}}$: $L_{K_{\boldsymbol{\alpha}}} g(x) = \int_{\mathcal{X}} K_{\boldsymbol{\alpha}}(x, y) g(y) d\rho(y)$ are unique, we know that $f(\boldsymbol{\alpha})$ is also differentiable by Theorem 4.1 in (Bonnans & Shapiro, 1998). Fixed some $u \in [m]$, we can compute the reduced gradient of Eq.(2) as follows,

$$[\bigtriangledown f(\boldsymbol{\alpha})]_p = \frac{\partial f(\boldsymbol{\alpha})}{\partial \alpha_p} - \frac{\partial f(\boldsymbol{\alpha})}{\partial \alpha_u}, (\forall p \in [m], p \neq u),$$

and

$$[\bigtriangledown f(\boldsymbol{\alpha})]_u = \sum_{p\neq u}\left(\frac{\partial f(\boldsymbol{\alpha})}{\partial \alpha_u} - \frac{\partial f(\boldsymbol{\alpha})}{\partial \alpha_p}\right),$$

where $\forall p \in [m]$,

$$\frac{\partial f(\boldsymbol{\alpha})}{\partial \alpha_p} = \frac{2\alpha_p}{m-1}\sum_{j=1}^{k}\iint_{\mathcal{X}} K_{\boldsymbol{\alpha}}(x,y)h_j^{\boldsymbol{\alpha}}(x)h_j^{\boldsymbol{\alpha}}(y)d\rho(x)d\rho(y).$$

where $h_j^{\boldsymbol{\alpha}}(\cdot)$ is the j-th clustering indicator function when the kernel weights are $\boldsymbol{\alpha}$. The optimization of $f(\boldsymbol{\alpha})$ is similar to $f_n(\boldsymbol{\alpha})$.

3. Related Work

In the general setting of learning tasks, the training set is drawn from an underlying probability distribution. In such cases, clustering algorithms should satisfy the following fundamental consistency criteria. *When the sample number goes to infinite, the parameters, such as clustering centroids and eigenvectors of graph Laplacian matrices, constructed by the clustering algorithm should converge to the parameters of the whole underlying space.* In the existing literature, several studies are proposed to derive the consistency of classic clustering algorithms.

3.1. Consistency of k-means

Pollard (1981b) shows the consistency of the global minimizer of the objective function for k-means clustering. Specifically, given a set of points $\{x_i\}_{i=1}^n$, the objective function of k-means is

$$W(A,\rho_n) = \int_{\mathcal{X}} \min_{a\in A}\|x-a\|^2 d\rho_n(x),$$

where $A = \{a_1, \cdots, a_k\}$ is a group of clustering centriods. For a fixed A, by the law of large numbers,

$$W(A,\rho_n) \to W(A,\rho) := \int_{\mathcal{X}} \min_{a\in A}\|x-a\|^2 d\rho(x).$$

Denote that $A_n = \operatorname{argmin}_A W(A,\rho_n)$ and $A^* = \operatorname{argmin}_A W(A,\rho)$. The author of (Pollard, 1981b) shows that A_n can converge almost surely A^*. However, the convergence rate is not studied in (Pollard, 1981b). Subsequently, variants of k-means are proven to have consistency guarantees (Sun et al., 2012; Georgogiannis, 2016; Paul et al., 2023).

3.2. Consistency of Spectral Clustering

Spectral clustering is another important algorithm for partitioning non-linear datasets. The consistency of spectral clustering is studied in (Von Luxburg et al., 2008). The authors in (Von Luxburg et al., 2008) show that the normalized graph Laplacian matrix's eigenfunctions converge to a Laplacian operator's eigenfunctions. In particular, assume that $\mathbf{K} \in \mathbb{R}^{n\times n}$ is the data similarity matrix, and $\mathbf{D}$ is a diagonal matrix with elements $d_{ii} := \sum_{j=1}^n K_{ij}$. The normalized graph Laplacian is

$$U_n = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{K}\mathbf{D}^{-1/2}.$$

Then, the corresponding Laplacian operator U is defined by

$$Uf(x) = f(x) - \int_{\mathcal{X}} K(x,y)f(y)/\sqrt{d(x)d(y)}d\rho(y),$$

where $d(x) = \int_{\mathcal{X}} K(x,y)d\rho(y)$. Then the spectra relation between U_n and U is theoretically derived in (Von Luxburg et al., 2008), and the convergence rate is also given for the Gaussian kernel. The consistency properties of other categories of algorithms based on spectral clustering are studied in (Ghoshdastidar & Dukkipati, 2017; Zhixin Zhou;Amini, 2019).

Although the consistency of k-means and spectral clustering is well studied in existing research, multiple kernel clustering (MKC) still lacks consistency guarantees. To fill this gap, we theoretically study the consistency of kernel weights of MKC and derive the corresponding convergence rate.

4. Main Results

In this section, we present our main results. Besides the general assumptions, we need the following assumption.

Assumption 4.1. For any vector $\boldsymbol{\gamma} \in \mathbb{R}^m$, let $\delta_j(\boldsymbol{\gamma})$ be the gap between the j-th eigenvalue and the $(j+1)$-th eigenvalue of $\frac{1}{n}\mathbf{K}_{\boldsymbol{\gamma}}$. For any $j \in [k]$, there exists some constant $c \geq 0$ such that $\delta_j(\boldsymbol{\gamma}) \geq 1/c$ holds with any $\boldsymbol{\gamma} \in \triangle$.

Remark. This assumption commonly appears in matrix perturbation theory (Stewart, 1990; Chen et al., 2016). In the study of the perturbation of eigenvectors and orthogonal projectors, the gaps of eigenvalues are usually regarded as constants. For example, Mitz & Shkolnisky (2022) derive the bounded difference of eigenvectors in Nyström approximation by assuming the eigen gaps are constants. We use a similar technique of (Von Luxburg et al., 2008) to prove our main results. Notice that Assumption 4.1 also implies all the eigenvalues are separated as the assumption made in (Von Luxburg et al., 2008).

The following two theorems are our main results. The first is the difference between the kernel weights learned from the training set S_n and the sample space $\mathcal{X}$.

Theorem 4.2. *Under Assumption 4.1, with the same initialization and learning step length $\gamma \leq c$, after convergence, the solutions of SimpleMKKM on S_n and $\mathcal{X}$ are $\hat{\boldsymbol{\alpha}}_n, \boldsymbol{\alpha}^*$, respectively. Then*

$$\|\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^*\|_\infty \precsim k\sqrt{\frac{\log(k/\delta)}{n}},$$

holds with probability at least $1-\delta$.

Remark. As far as we know, this is the first result about the consistency of the kernel weights in MKC algorithms. This result can be used to obtain some essential properties of MKC algorithms. In the next section, we can obtain the excess risk bound by utilizing this theorem, which is the same as the single kernel clustering proposed in (Biau et al., 2008). The proof can be found in Section B.3 of the appendix.

Theorem 4.3. *Under Assumption 4.1, with the same initialization and learning step length* $\gamma \leq c$*, after convergence, the solutions of SimpleMKKM on* S_n *and its subset* S_r *(sampling without replacement) are* $\hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\alpha}}_r$*, respectively. Then*

$$\|\hat{\boldsymbol{\alpha}}_r - \hat{\boldsymbol{\alpha}}_n\|_\infty \precsim k\sqrt{\left(\frac{1}{r}-\frac{1}{n}\right)\log\left(\frac{k}{\delta}\right)},$$

holds with probability at least $1-\delta$.

Remark. Theorem 4.3 implies that the difference of the kernel weights learned from S_n and S_r has a fast convergence rate as $r \to n$. Thus, we can approximate the kernel weights learned from S_n when r is sufficiently large. Consequently, we propose a large-scale extension with a learning guarantee for SimpleMKKM, which will be described in Section 6.

5. The analysis of Excess Clustering Risk

In (Liu, 2022), the author gives an upper bound of the generalization clustering risk. However, the objective function of (Liu, 2022) being analyzed is not a standard clustering risk function. Moreover, the excess risk bound of SimpleMKKM has not been studied, which is a more general form than the generalization bound.

We first define the loss function for any $x \in \mathcal{X}$. When the kernel weights are $\boldsymbol{\alpha} \in \mathbb{R}^m$, the feature map of x is

$$\phi_{\boldsymbol{\alpha}}(x) = [\alpha_1\phi_1^\top(x), \cdots, \alpha_m\phi_m^\top(x)]^\top.$$

We denote the Hilbert space that $\phi_{\boldsymbol{\alpha}}(x)$ belongs to as $\mathcal{H}_{\boldsymbol{\alpha}}$. For some clustering centroids $\mathbf{C} = \{\mathbf{c}_j\}_{j=1}^k \in \mathcal{H}_{\boldsymbol{\alpha}}^k$, the loss function of x is

$$l(x, \mathbf{C}, \boldsymbol{\alpha}) = \min_{j\in[k]} \|\phi_{\boldsymbol{\alpha}}(x) - \mathbf{c}_j\|^2. \tag{4}$$

Accordingly, the empirical clustering risk can be expressed as

$$\mathcal{W}_n(\mathbf{C}, \boldsymbol{\alpha}, \rho_n) = \frac{1}{n}\sum_{i=1}^n \min_{j\in[k]} \|\phi_{\boldsymbol{\alpha}}(x_i) - \mathbf{c}_j\|^2,$$

and the expected clustering risk is denoted as

$$\mathcal{W}(\mathbf{C}, \boldsymbol{\alpha}, \rho) = \int_{\mathcal{X}} \min_{j\in[k]} \|\phi_{\boldsymbol{\alpha}}(x) - \mathbf{c}_j\|^2 d\rho(x).$$

We denote that the kernel weights learned by performing SimpleMKKM on the training set S_n are $\hat{\boldsymbol{\alpha}}_n$, and the homologous clustering centroids are $\hat{\mathbf{C}} = \{\hat{\mathbf{c}}_j\}_{j=1}^k \in \mathcal{H}_{\hat{\boldsymbol{\alpha}}_n}^k$. When the input is the sample space, the output kernel weights are denoted as $\boldsymbol{\alpha}^*$, and the clustering centroids are $\mathbf{C}^*$ which satisfies $\mathbf{C}^* = \operatorname{argmin}_{\mathbf{C}\in\mathcal{H}^k} \mathcal{W}(\mathbf{C}, \boldsymbol{\alpha}^*, \rho)$. To verify the generalization ability of the learned kernel weights and clustering centroids, we need to upper bound the following formula

$$\mathbb{E}_{S_n}[\mathcal{W}(\hat{\mathbf{C}}, \hat{\boldsymbol{\alpha}}_n, \rho)] - \mathcal{W}(\mathbf{C}^*, \boldsymbol{\alpha}^*, \rho). \tag{5}$$

We must perform the standard k-means to obtain the clustering centroids $\hat{\mathbf{C}}$. It is well known that finding the optimal solution of k-means is an NP-hard problem. This paper does not discuss the relationship between the clustering centroids obtained by standard k-means and the optimal ones. Consequently, we simply assume that $\hat{\mathbf{C}} = \operatorname{argmin}_{\mathbf{C}\in\mathcal{H}_{\hat{\boldsymbol{\alpha}}_n}^k} \mathcal{W}_n(\mathbf{C}, \hat{\boldsymbol{\alpha}}_n, \rho_n)$.

To bound Eq.(5), we process a decomposition as follows

$$\begin{aligned}
&\mathbb{E}_{S_n}[\mathcal{W}(\hat{\mathbf{C}}, \hat{\boldsymbol{\alpha}}_n, \rho)] - \mathcal{W}(\mathbf{C}^*, \boldsymbol{\alpha}^*, \rho) \\
=&\underbrace{\mathbb{E}_{S_n}[\mathcal{W}(\hat{\mathbf{C}}, \hat{\boldsymbol{\alpha}}_n, \rho) - \mathcal{W}_n(\hat{\mathbf{C}}, \hat{\boldsymbol{\alpha}}_n, \rho_n)]}_{\mathcal{A}} \\
+&\underbrace{\mathbb{E}_{S_n}[\mathcal{W}_n(\hat{\mathbf{C}}, \hat{\boldsymbol{\alpha}}_n, \rho_n) - \mathcal{W}_n(\mathbf{C}^*, \boldsymbol{\alpha}^*, \rho_n)]}_{\mathcal{B}} \\
+&\underbrace{\mathbb{E}_{S_n}[\mathcal{W}_n(\mathbf{C}^*, \boldsymbol{\alpha}^*, \rho_n)] - \mathcal{W}(\mathbf{C}^*, \boldsymbol{\alpha}^*, \rho)}_{\mathcal{C}}.
\end{aligned}$$

Term $\mathcal{A}$ and Term $\mathcal{C}$ are the same as the generalization risk of single kernel clustering, and their upper bound is $\mathcal{O}(k/\sqrt{n})$ (Biau et al., 2008).

Term $\mathcal{B}$ can be bounded by $\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_\infty$ multiplying a constant. Combining with Theorem 4.2, we can obtain the excess risk bound of SimpleMKKM by the following theorem.

Theorem 5.1. *The excess clustering risk of SimpleMKKM can be upper bounded by* $\widetilde{\mathcal{O}}(k/\sqrt{n})$.

Remark. Theorem 5.1 gives the upper bound of the excess risk of SimpleMKKM with a standard clustering loss function as is represented in Eq. (4). Meanwhile, the objective function analyzed in (Liu, 2022) is in the form of the inner product, and it is uncommon among the studies of clustering risk bound. Thus, the proposed risk bound is more rational than the one in (Liu, 2022). The detailed proof of Theorem 5.1 is in Section B.5 of the appendix.

6. Large-Scale Extension with Nyström Method

In this section, we use Nyström method to propose a large-scale strategy for SimpleMKKM. Nyström method is usu-

ally used to accelerate kernel k-means such that the time complexity is linear with the sample number n (Calandriello & Rosasco, 2018). However, existing works rarely apply Nyström method to multiple kernel clustering. In a recent paper (Lu et al., 2022), the authors make the first attempt and propose a scalable multiple kernel k-means clustering. However, their method lacks a theoretical learning guarantee, and the number of sampled points will be an additional hyperparameter. Selecting an optimal hyperparameter in the unsupervised scenario is still an open problem. To address the above issues, we propose a novel algorithm, and the detailed process is listed as follows.

1. Sampling r points $\{y_i\}_{i=1}^r$ (without replacement) from n points $\{x_i\}_{i=1}^n$.
2. Perform SimpleMKKM on $\{y_i\}_{i=1}^r$, and the output kernel weights are denoted as $\hat{\boldsymbol{\alpha}}_r$.
3. In the p-th kernel, we denote the kernel similarity matrix consisting of all n samples and r selected samples as $\mathbf{R}_p \in \mathbb{R}^{n\times r}$, and the kernel matrix consisting of r samples is $\mathbf{W}_p$. We then use $\hat{\boldsymbol{\alpha}}_r$ to make weighted combinations of the above matrices which are denoted as $\mathbf{R}_{\hat{\boldsymbol{\alpha}}_r}$ and $\mathbf{W}_{\hat{\boldsymbol{\alpha}}_r}$, respectively.
4. Performing eigen decomposition on $\mathbf{W}_{\hat{\boldsymbol{\alpha}}_r}$, we have $\mathbf{W}_{\hat{\boldsymbol{\alpha}}_r} = \mathbf{U}_r\boldsymbol{\Lambda}_r\mathbf{U}_r^\top$.
5. Perform standard k-means on the rows of $\mathbf{R}_{\hat{\boldsymbol{\alpha}}_r}\mathbf{U}_r\boldsymbol{\Lambda}_r^{-1/2}$ to obtain the final clustering results.

The above algorithm is simple but efficient and effective. We then give an analysis of the complexity and its excess risk bound.

Computational and storage complexity. In Step 2, performing SimpleMKKM on m base kernel matrices of size $r \times r$ costs $\mathcal{O}(Tr^3 + Tmr^2)$ time, and occupies $\mathcal{O}(mr^2)$ space, where T is the iteration number of SimpleMKKM. In Step 3, the computations of $\mathbf{R}_{\hat{\boldsymbol{\alpha}}_r}$ and $\mathbf{W}_{\hat{\boldsymbol{\alpha}}_r}$ cost $\mathcal{O}((m-1)(nr+r^2))$ time, and the space complexity is $\mathcal{O}(mnr + mr^2)$. Step 4 is the eigen decomposition of $\mathbf{W}_{\hat{\boldsymbol{\alpha}}_r}$ whose time consumption is $\mathcal{O}(r^3)$. In Step 5, computing $\mathbf{R}_{\hat{\boldsymbol{\alpha}}_r}\mathbf{U}_r\boldsymbol{\Lambda}_r^{-1/2}$ costs $\mathcal{O}(nr^2)$ time and the subsequent k-means consumes $\mathcal{O}(nrkT_1)$ time, where k is the cluster number and T_1 is the iteration number of k-means. Above all, if $r \ll n$, the computational and storage complexity is linear with n. As a result, the proposed method can handle large-scale datasets.

Excess risk bound. Besides the complexity analysis, to illustrate the effectiveness of the proposed method, we also give an upper bound of it. We assume that the clustering centroids learned by our method are $\widetilde{\mathbf{C}}_{n,r}$ which belong to the space $\underbrace{\mathbb{R}^r \times \cdots \times \mathbb{R}^r}_{k}$[2]. Then, we can compute the corresponding clustering centroids $\hat{\mathbf{C}}_{n,r}$ in the Hilbert space $\mathcal{H}^k_{\hat{\boldsymbol{\alpha}}_r}$ by $\hat{\mathbf{C}}_{n,r} = \boldsymbol{\Phi}_r^{\hat{\boldsymbol{\alpha}}_r}\mathbf{U}_r\boldsymbol{\Lambda}_r^{-1/2}$, where $\boldsymbol{\Phi}_r^{\hat{\boldsymbol{\alpha}}_r}$ is the feature map of r selected points when the kernel weights are $\hat{\boldsymbol{\alpha}}_r$. Specifically, the i-th column of $\boldsymbol{\Phi}_r^{\hat{\boldsymbol{\alpha}}_r}$ is

$$[\hat{\alpha}_r(1)\phi_1^\top(y_i), \cdots, \hat{\alpha}_r(m)\phi_m^\top(y_i)]^\top,$$

where $\hat{\alpha}_r(p)$ denotes the p-th component of $\hat{\boldsymbol{\alpha}}_r$.

We should give the upper bound of the following formula to verify the generalization ability of the learned kernel weights and clustering centroids.

$$\mathbb{E}_S[\mathcal{W}(\hat{\mathbf{C}}_{n,r}, \hat{\boldsymbol{\alpha}}_r, \rho)] - \mathcal{W}(\mathbf{C}^*, \boldsymbol{\alpha}^*, \rho).$$

We can deduce Theorem 6.1 by our theoretical result about the consistency of kernel weights.

Theorem 6.1. *When the number of selected points is r, then*

$$\mathbb{E}_{S_n}[\mathcal{W}(\hat{\mathbf{C}}_{n,r}, \hat{\boldsymbol{\alpha}}_r, \rho)] - \mathcal{W}(\mathbf{C}^*, \boldsymbol{\alpha}^*, \rho)$$

can be upper bounded by

$$\widetilde{\mathcal{O}}\left(\frac{k}{r} + \frac{k}{\sqrt{r}} + \frac{k}{\sqrt{n}}\right).$$

Remark. Usually, we let $r = \Theta(\sqrt{n})$. In this case, Theorem 6.1 gives an upper bound as $\widetilde{\mathcal{O}}(kn^{-1/4})$, which is very loose and unsatisfactory. Fortunately, we can further improve this result.

Tighter risk bound. In (Yin et al., 2022a;b; 2020b), Yin et al. give several randomized sketching methods for clustering, and provide a risk bound for the sketching dimension. Surprisingly, we can use a similar technique of (Yin et al., 2022a;b; 2020b) to establish the following theorem, providing a tighter bound than Theorem 6.1.

Theorem 6.2. *When the number of selected points $r \geq \frac{\log(2/\delta)}{\varepsilon - \log(1+\varepsilon)}$,*

$$\mathbb{E}_{S_n}[\mathcal{W}(\hat{\mathbf{C}}_{n,r}, \hat{\boldsymbol{\alpha}}_r, \rho)] - \mathcal{W}(\mathbf{C}^*, \boldsymbol{\alpha}^*, \rho)$$

can be bounded by

$$\widetilde{\mathcal{O}}\left(\frac{k}{r} + \frac{k}{\sqrt{n}} + k\varepsilon\right).$$

Remark. Larger r can make the above bound tighter but cost more time. To make a favorable statistical and computational trade-off, we should select an appropriate r. In Theorem 6.2, let $\varepsilon = 1/\sqrt{n}$, then $r = \Omega(\sqrt{n})$ and the excess risk bound is $\widetilde{\mathcal{O}}(k/\sqrt{n})$. The proof of Theorem 6.2 is in Section B.6 of the appendix.

[2] $\underbrace{\mathbb{R}^r \times \cdots \times \mathbb{R}^r}_{k}$ is the k-multiple Cartesian products of $\mathbb{R}^r$.

7. Experiments

The experiments compose of two parts. The first part validates the non-asymptotic bound of the difference between the kernel weights learned from the whole dataset and its subset as shown in Theorem 4.3. The second part is the numerical experiment of the proposed large-scale algorithm. All experiments are conducted on a laptop with Intel(R) Core(TM)-i7-10870H CPU.

7.1. The Difference of Kernel Weights

Table 1. Benchmark datasets

Datasets	Samples	Kernels	Clusters
Flo17	1360	7	17
Flo102	8189	4	102
DIGIT	2000	3	10
PFold	694	12	27
CCV	6773	3	20
Reuters	18758	5	6

We conduct experiments on 6 benchmark datasets, including *Flo17*, *Flo102*, *DIGIT*, *PFold*, *CCV* and *Reuters*. We report the detailed information in Table 1, and their URLs can be found in Appendix D. For each dataset, we first perform SimpleMKKM on the whole samples to obtain the kernel weights $\hat{\boldsymbol{\alpha}}_n$. Then, we sample r points without replacement. We run SimpleMKKM on these r samples and record the corresponding kernel weights. To reduce the influence of randomness, we repeat the above process 50 times and compute the average kernel weights, denoted as $\hat{\boldsymbol{\alpha}}_r$. The number of selected points r varies in $[100, 200, \cdots, 3000]$. We let r be smaller than the whole sample number for the small datasets. In addition, for the datasets with large cluster numbers k, we let r be bigger than k. We compute the values of $\|\hat{\boldsymbol{\alpha}}_n - \hat{\boldsymbol{\alpha}}_r\|_\infty$ for different r, and illustrate them in Figure 1.

In Figure 1, the blue points reflect the variation trend of $\|\hat{\boldsymbol{\alpha}}_n - \hat{\boldsymbol{\alpha}}_r\|_\infty$ as r becomes larger. As seen, $\|\hat{\boldsymbol{\alpha}}_n - \hat{\boldsymbol{\alpha}}_r\|_\infty$ tends to be smaller and converges to a small value when r is large enough. As a reference, we plot the image of the function $f(r) = a\sqrt{1/r - 1/n}$ by a red curve, and report the different a's of all the 6 datasets in the right-hand corner of Figure 1. It can be observed that the blue points are bounded by the red curve. This verifies the correctness of the bound in Theorem 4.3. Moreover, the difference of the blue points and the red curve is small when r is small. It shows the tightness of the proposed bound. Meanwhile, this difference becomes larger as r is larger.

Table 2. Large-scale datasets used in the experiments

Dataset	Samples	Views	Clusters
NUSWIDE	30000	5	31
AwA	30475	6	50
CIFAR10	50000	3	10
YtVideo	101499	5	31
Winnipeg	325834	2	7
Covertype	581012	2	10

7.2. Experiments on Large-Scale Datasets

In Section 6, we propose a method with Nyström that can make SimpleMKKM able to deal with large-scale datasets. Furthermore, we analyze the proposed method's complexity and excess risk bound in theory. We conduct experiments on 6 large-scale datasets in this subsection to test the actual clustering performance and running time. The used datasets are NUSWIDE, AwA, CIFAR10, YtVideo, Winnipeg, and Covertype. Their sample numbers, cluster numbers, and view numbers can be found in Table 2. The number of samples ranges from 30000 to 581012. Such large-scale datasets are rarely seen in the research of kernel clustering, especially for multiple kernel clustering. Despite this, the proposed algorithm is still able to handle them. Due to the limited space, the URLs of these datasets are listed in Appendix D.

For each view, we use the Gaussian RBF kernel to construct the kernel similarity matrix between the whole training set S_n and the selected subset S_r, i.e.,

$$K(x, z) = \exp\left(-\frac{\|x - z\|^2}{2\sigma^2}\right),$$

where $x \in S_n$, $z \in S_r$, and σ is the square root of the average interpoint distance between S_n and S_r, i.e.,

$$\sigma = \sqrt{\frac{1}{nr}\sum_{x\in S_n}\sum_{z\in S_r}\|x - z\|^2}.$$

As proven in Section 6, by setting $r = \Theta(\sqrt{n})$, the proposed algorithm can have a favorable statistical and computational trade-off. For a sufficient r, we let $r = 3 \cdot \lceil\sqrt{n}\rceil$. Because the used datasets are extra large, existing multiple kernel clustering methods can rarely be performed on them. Thus, we only perform single kernel clustering with Nyström on each kernel for comparisons. Our experiments use three frequently-used clustering metrics: accuracy (ACC), normalized mutual information (NMI), and purity. We also record the execution time of all experiments. The experimental results are reported in Table 3. Notice that the best results are in bold fonts, and we underline the second-best results for each dataset.

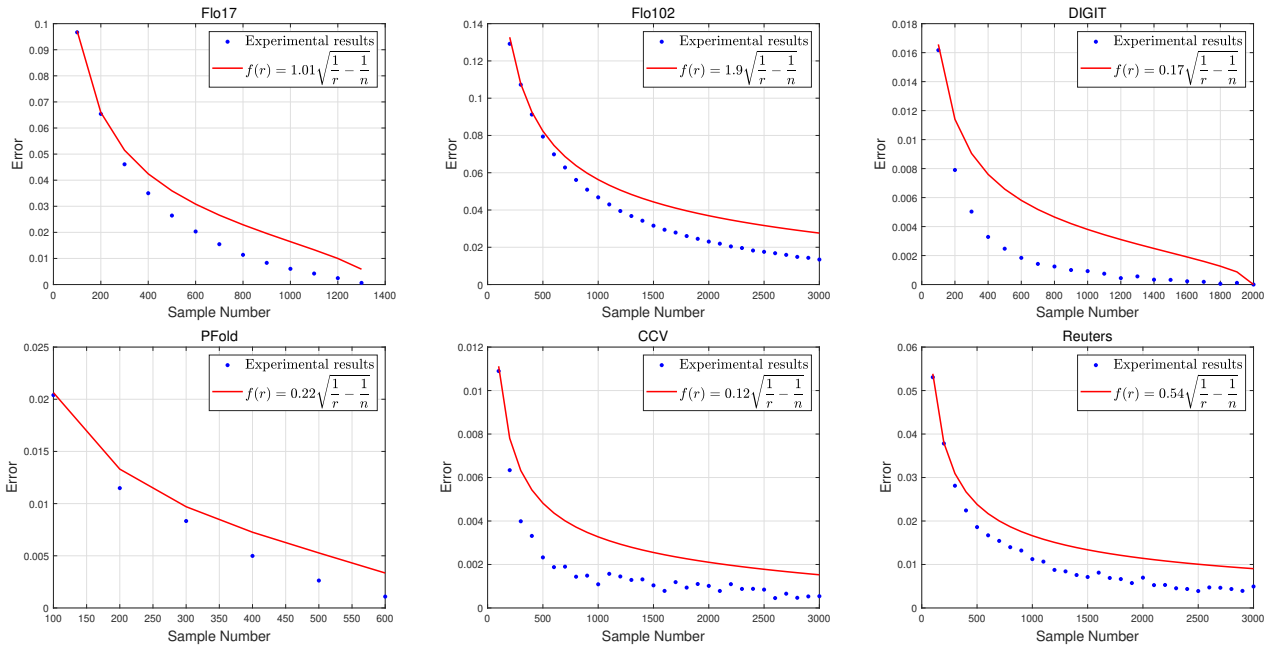

Figure 1. The difference of the kernel weights learned from 6 benchmark datasets and their respective subsets. The blue points record the true difference of the numerical experiments, while the red curves are the image of $f(r) = a\sqrt{1/r - 1/n}$, where the value a differs in all 6 datasets.

As seen from Table 3, the experimental results show that the proposed method achieves the best clustering performance. At the same time, the time consumption is slightly larger than the single kernel k-means with Nyström. Specifically, from this table, we have the following observations:

1. The proposed method outperforms the second-best results by 1.36%, 1.83%, 15.42%, 1.72%, 8.27%, and 3.91% in terms of NMI on all six datasets. On the other two clustering metrics, the proposed method also performs best.

2. Notice that the sample numbers of the last three datasets are enormous. Although such large-scale datasets are rarely seen in the studies of kernel clustering, the proposed method can also handle them within hundreds of seconds.

In summary, the proposed method demonstrates superior clustering performance and has a high execution efficiency on all benchmark datasets.

8. Conclusions and Future Works

In the paper, we study the consistency of multiple kernel clustering. We have obtained two important bounds, i.e., the bounds of $\|\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^*\|_\infty$ and $\|\hat{\boldsymbol{\alpha}}_r - \hat{\boldsymbol{\alpha}}_n\|_\infty$, where $\hat{\boldsymbol{\alpha}}_r, \hat{\boldsymbol{\alpha}}_n, \boldsymbol{\alpha}^*$ are respectively the kernel weights learned from $S_r, S_n, \mathcal{X}$ by SimpleMKKM. We then use the derived bounds to obtain the excess risk bound of SimpleMKKM. Moreover, we make a large-scale extension of SimpleMKKM with a theoretical learning guarantee. Besides the theoretical analysis, we conduct experiments to verify the proposed results and the large-scale algorithm.

In the future, we will focus on the following three aspects to improve our work.

1. The proposed bound of the kernel weights could be further improved for a tighter excess risk bound. In our proofs, we use the excess risk bound of single kernel clustering, which is $\widetilde{\mathcal{O}}(k/\sqrt{n})$ (Biau et al., 2008). However, Liu (2021) improves the above bound as $\widetilde{\mathcal{O}}(\sqrt{k/n})$, and prove its tightness. We want to know whether the proposed bound of the difference of the kernel weights can be improved from $\widetilde{\mathcal{O}}(k/\sqrt{n})$ to $\widetilde{\mathcal{O}}(\sqrt{k/n})$. Accordingly, the excess risk bound of multiple kernel clustering can also be improved.

2. The proposed large-scale algorithm is effective and efficient, but there is room for improvement in multiple kernel clustering. We will try to design more large-scale algorithms for better clustering performance and higher operating efficiency.

3. As is well known, kernel clustering has a significant connection with spectral clustering (Dhillon et al., 2004). Meanwhile, this connection still exists among multiple kernel clustering and multi-view spectral clustering (MVSC). Although the consistency of spectral clustering has been discovered by (Von Luxburg et al., 2008), the consistency of MVSC needs to be further studied. We will explore whether MVSC has consistency with the techniques proposed in this paper.

Table 3. The clustering performance on large-scale datasets.

(a) NUSWIDE

Method	ACC	NMI	Purity	Time(s)
View1	13.58	8.59	20.30	7.84
View2	12.67	9.23	20.97	7.87
View3	13.65	9.48	21.79	8.10
View4	16.16	12.08	24.69	8.26
View5	14.28	10.21	23.61	10.34
Proposed	**16.64**	**13.44**	**25.19**	12.09

(b) AwA

Method	ACC	NMI	Purity	Time(s)
View1	7.56	7.05	8.69	13.18
View2	7.65	7.79	8.79	13.33
View3	7.23	6.16	7.85	14.01
View4	7.82	8.08	9.04	17.47
View5	8.15	9.06	9.27	15.92
View6	8.09	8.21	9.44	13.47
Proposed	**9.42**	**10.89**	**10.78**	22.41

(c) CIFAR10

Method	ACC	NMI	Purity	Time(s)
View1	73.34	65.87	74.22	6.35
View2	71.54	62.36	73.50	6.57
View3	63.63	54.41	65.29	6.55
Proposed	**81.36**	**81.29**	**85.53**	12.35

(d) YtVideo

Method	ACC	NMI	Purity	Time(s)
View1	9.79	5.37	26.87	30.85
View2	17.81	15.32	29.09	29.27
View3	13.74	11.29	27.07	33.93
View4	18.29	16.51	29.76	33.75
View5	18.97	11.67	28.59	30.22
Proposed	**18.97**	**17.04**	**29.36**	60.26

(e) Winnipeg

Method	ACC	NMI	Purity	Time(s)
View1	60.20	47.72	71.84	188.79
View2	56.84	44.60	65.42	121.81
Proposed	**68.93**	**55.99**	**76.33**	164.81

(f) Covertype

Method	ACC	NMI	Purity	Time(s)
View1	36.34	10.64	54.23	391.18
View2	**48.76**	11.32	48.76	402.54
Proposed	45.26	**15.23**	**55.21**	485.52

9. Acknowledgments

This work was supported by the National Key R&D Program of China 2020AAA0107100, Youth Foundation Project of Zhejiang Lab (No. K2023PD0AA01), and the National Natural Science Foundation of China (project no. 62002170, 61773392, 61872377, 61922088, 61976196, and 62006237).

References

Bardenet, R. and Maillard, O.-A. Concentration inequalities for sampling without replacement. In *Bernoulli*, volume 21, pp. 1361–1385, 2015.

Bengio, Y., Delalleau, O., Roux, N. L., Paiement, J.-F., Vincent, P., and Ouimet, M. Learning eigenfunctions links spectral embedding and kernel pca. *Neural computation*, 16(10):2197–2219, 2004.

Biau, G., Devroye, L., and Lugosi, G. On the performance of clustering in hilbert spaces. In *IEEE Transactions on Information Theory (TIT)*, pp. 781–790, 2008.

Bonnans, J. F. and Shapiro, A. Optimization problems with perturbations: A guided tour. In *SIAM review*, volume 40, pp. 228–264. SIAM, 1998.

Calandriello, D. and Rosasco, L. Statistical and computational trade-offs in kernel k-means. In *Advances in neural information processing systems (NeurIPS)*, volume 31, 2018.

Chen, Y. M., Chen, X. S., and Li, W. On perturbation bounds for orthogonal projections. In *Numerical Algorithms*, pp. 433–444, 2016.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. Introduction to algorithms. MIT press, 2022.

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 (KDD)*, pp. 551–556, 2004.

Du, L., Zhou, P., Shi, L., Wang, H., Fan, M., Wang, W., and Shen, Y.-D. Robust multiple kernel k-means using l21-norm. In *Twenty-fourth international joint conference on artificial intelligence (IJCAI)*, 2015.

Georgogiannis, A. Robust k-means: a theoretical revisit. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016.

Ghoshdastidar, D. and Dukkipati, A. Consistency of spectral hypergraph partitioning under planted partition model. In *The Annals of Statistics*, volume 45, pp. 289 – 315, 2017.

Huang, H.-C., Chuang, Y.-Y., and Chen, C.-S. Multiple kernel fuzzy clustering. In *IEEE Transactions on Fuzzy Systems (TFS)*, pp. 120–134, 2011.

Liu, J., Liu, X., Wang, S., Zhou, S., and Yang, Y. Hierarchical multiple kernel clustering. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, volume 35, pp. 8671–8679, 2021.

Liu, X. Simplemkkm: Simple multiple kernel k-means. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*. IEEE, 2022.

Liu, X., Dou, Y., Yin, J., Wang, L., and Zhu, E. Multiple kernel k-means clustering with matrix-induced regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 1888–1894, 2016.

Liu, Y. Refined learning bounds for kernel and approximate k-means. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Lu, Y., Xin, H., Wang, R., Nie, F., and Li, X. Scalable multiple kernel k-means clustering. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM)*, pp. 4279–4283, 2022.

McDiarmid, C. On the method of bounded differences. In *Surveys in combinatorics*, volume 141, pp. 148–188. Norwich, 1989.

Mitz, R. and Shkolnisky, Y. A perturbation-based kernel approximation framework. In *Journal of Machine Learning Research (JMLR)*, volume 23, pp. 1–26, 2022.

Paul, D., Chakraborty, S., Das, S., and Xu, J. Implicit annealing in kernel spaces: A strongly consistent clustering approach. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, volume 45, pp. 5862–5871, 2023.

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

Pollard, D. Strong consistency of k-means clustering. In *The Annals of Statistics*, volume 9, pp. 135–140, 1981b.

Rosasco, L., Belkin, M., and Vito, E. D. On learning with integral operators. In *Journal of Machine Learning Research (JMLR)*, pp. 905–934, 2010.

Stewart, G. W. Matrix perturbation theory. Citeseer, 1990.

Sun, W., Wang, J., and Fang, Y. Regularized k-means clustering of high-dimensional data and its asymptotic consistency. In *Electronic Journal of Statistics*, volume 6, pp. 148 – 167, 2012.

Von Luxburg, U., Belkin, M., and Bousquet, O. Consistency of spectral clustering. In *The Annals of Statistics*, pp. 555–586, 2008.

Wang, R., Lu, J., Lu, Y., Nie, F., and Li, X. Discrete multiple kernel k-means. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3111–3117, 2021.

Yin, R., Liu, Y., Wang, W., and Meng, D. Extremely sparse johnson-lindenstrauss transform: From theory to algorithm. In *2020 IEEE International Conference on Data Mining (ICDM)*, pp. 1376–1381, 2020a. doi: 10.1109/ICDM50108.2020.00180.

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

Yin, R., Liu, Y., Wang, W., and Meng, D. Randomized sketches for clustering: Fast and optimal kernel k-means. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022a.

Yin, R., Liu, Y., Wang, W., and Meng, D. Scalable kernel k-means with randomized sketching: From theory to algorithm. In *IEEE Transactions on Knowledge and Data Engineering (TKDE)*. IEEE, 2022b.

Yu, Y., Wang, T., and Samworth, R. J. A useful variant of the davis-kahan theorem for statisticians. In *Biometrika*, pp. 315–323, 2014.

Zhao, B., Kwok, J. T., and Zhang, C. Multiple kernel clustering. In *International Conference on Data Mining (ICDM)*, pp. 638–649, 2009.

Zhixin Zhou;Amini, A. A. Analysis of spectral clustering algorithms for community detection: the general bipartite setting. In *Journal of Machine Learning Research (JMLR)*, volume 20, pp. 1–47, 2019.

Zwald, L. and Blanchard, G. On the convergence of eigenspaces in kernel principal component analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 18, 2005.

A. Overview of the Proofs

In this section, we outline the proofs of the main results proposed in Section 4.

A.1. Proof Technique of Theorem 4.2

To study the difference between $\hat{\boldsymbol{\alpha}}_n$ and $\boldsymbol{\alpha}^*$, we should quantify the difference of alignment of the p-th base kernel and the clustering indicator functions, which are respectively defined as

$$\hat{\epsilon}(K_p, \{\hat{h}_j^{\boldsymbol{\alpha}}\}_{j=1}^k) = \frac{1}{n^2}\sum_{j=1}^{k}\sum_{i=1}^{n}\sum_{t=1}^{n} K_p(x_i, x_t)\hat{h}_j^{\boldsymbol{\alpha}}(x_i)\hat{h}_j^{\boldsymbol{\alpha}}(x_t),$$

$$\epsilon(K_p, \{h_j^{\boldsymbol{\alpha}}\}_{j=1}^k) = \sum_{j=1}^{k}\iint_{\mathcal{X}} K_p(x, y)h_j^{\boldsymbol{\alpha}}(x)h_j^{\boldsymbol{\alpha}}(y)d\rho(x)d\rho(y),$$

where $\hat{h}_j^{\boldsymbol{\alpha}}, h_j^{\boldsymbol{\alpha}}$ are the clustering indicator functions of the operators of $\frac{1}{n}\mathbf{K}_{\boldsymbol{\alpha}}$ and $L_{K_{\boldsymbol{\alpha}}}$.

In each iteration, we denote $\boldsymbol{\alpha}$ is the kernel weights obtained from S_n, and the corresponding first k eigenfunctions are $\{\hat{h}_j^{\boldsymbol{\alpha}}(\cdot)\}_{j=1}^k$. Accordingly, we denote $\boldsymbol{\beta}$ as the kernel weights learned from the sample space $\mathcal{X}$ in the same iteration, and $\{h_j^{\boldsymbol{\beta}}(\cdot)\}_{j=1}^k$. We then bound $\left|\hat{\epsilon}(K_p, \{\hat{h}_j^{\boldsymbol{\alpha}}\}_{j=1}^k) - \epsilon(K_p, \{h_j^{\boldsymbol{\beta}}\}_{j=1}^k)\right|$. By decoupling the kernel weights and alignment, we can deduce that

$$\begin{aligned}
&\hat{\epsilon}(K_p, \{\hat{h}_j^{\boldsymbol{\alpha}}\}_{j=1}^k) - \epsilon(K_p, \{h_j^{\boldsymbol{\beta}}\}_{j=1}^k) \\
=&\underbrace{\hat{\epsilon}(K_p, \{\hat{h}_j^{\boldsymbol{\alpha}}\}_{j=1}^k) - \hat{\epsilon}(K_p, \{\hat{h}_j^{\boldsymbol{\beta}}\}_{j=1}^k)}_{\mathcal{A}} + \underbrace{\hat{\epsilon}(K_p, \{\hat{h}_j^{\boldsymbol{\beta}}\}_{j=1}^k) - \epsilon(K_p, \{h_j^{\boldsymbol{\beta}}\}_{j=1}^k)}_{\mathcal{B}}.
\end{aligned} \tag{6}$$

By the assumption of the eigen gap of $\frac{1}{n}\mathbf{K}_{\boldsymbol{\alpha}}$, we can use matrix perturbation theory to bound Term $\mathcal{A}$ as

$$\mathcal{A} \precsim \|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_\infty. \tag{7}$$

The following lemma reflects how the alignment level of arbitrary kernel function $K(\cdot, \cdot)$ and $\{h_j\}_{j=1}^k$ differs from the alignment level of $K(\cdot, \cdot)$ and $\{\hat{h}_j\}_{j=1}^k$.

Lemma A.1. *Under Assumption 4.1,*

$$\left|\hat{\epsilon}(K, \{\hat{h}_j\}_{j=1}^k) - \epsilon(K, \{h_j\}_{j=1}^k)\right| \precsim k\sqrt{\frac{\log(k/\delta)}{n}}$$

holds with probability at least $1 - \delta$.

By Lemma A.1, for any $\boldsymbol{\beta}$, we have

$$\mathcal{B} \leq k\sqrt{\frac{\log(k/\delta)}{n}}. \tag{8}$$

Above all, we have the following theorem.

Theorem A.2. *Under Assumption 4.1, for any* $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \triangle$,

$$\begin{aligned}
&\left|\hat{\epsilon}(K_p, \{\hat{h}_j^{\boldsymbol{\alpha}}\}_{j=1}^k) - \epsilon(K_p, \{h_j^{\boldsymbol{\beta}}\}_{j=1}^k)\right| \\
\precsim&\|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_\infty + k\sqrt{\frac{\log(k/\delta)}{n}},
\end{aligned}$$

holds with probability at least $1 - \delta$.

The sketching proof of Theorem 4.2. In all iterations, denote that the kernel weights learned from the training set are respectively

$$\boldsymbol{\alpha}^{(0)}, \cdots, \boldsymbol{\alpha}^{(T)}.$$

Meanwhile, the kernel weights obtained from the whole sample space are

$$\boldsymbol{\beta}^{(0)}, \cdots, \boldsymbol{\beta}^{(T)}.$$

With the same initialization, we know $\boldsymbol{\alpha}^0 = \boldsymbol{\beta}^0$. Then, it is easy to check that for any integer $t \geq 0$,

$$\left\|\boldsymbol{\alpha}^{(t+1)} - \boldsymbol{\beta}^{(t+1)}\right\|_\infty \precsim \left\|\boldsymbol{\alpha}^{(t)} - \boldsymbol{\beta}^{(t)}\right\|_\infty + \left|\hat{\epsilon}_p(\boldsymbol{\alpha}^{(t)}) - \epsilon_p(\boldsymbol{\beta}^{(t)})\right|,$$

where $\hat{\epsilon}_p(\boldsymbol{\alpha}^{(t)})$ is a brief notation of $\hat{\epsilon}(K_p, \{\hat{h}_j^{\boldsymbol{\alpha}^{(t)}}\}_{j=1}^k)$ as well as for $\epsilon_p(\boldsymbol{\beta}^{(t)})$.

With Theorem A.2, we then have

$$\left\|\boldsymbol{\alpha}^{(t)} - \boldsymbol{\beta}^{(t)}\right\|_\infty \precsim \left\|\boldsymbol{\alpha}^{(t-1)} - \boldsymbol{\beta}^{(t-1)}\right\|_\infty + k\sqrt{\frac{\log(k/\delta)}{n}}.$$

By repeating the above process and the convergence of the reduced gradient descent algorithm, the result of Theorem 4.2 follows. The detailed proof is in Section B.3 of the appendix.

A.2. Proof Technique of Theorem 4.3

The proof process is similar to Theorem 4.2. By the utilization of the similar notations as the above subsection, we define the alignment level of the training set of n points as

$$\hat{\epsilon}(K_p, \{\hat{h}_{n,j}\}_{j=1}^k) = \frac{1}{n^2}\sum_{j=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{n} K_p(x_i, x_j)\hat{h}_{n,j}(x_i)\hat{h}_{n,j}(x_j),$$

where $\hat{h}_{n,j}$ is the approximated clustering indicator function. Similar, the alignment level of r points (sampling from S_n without replacement) can be defined as

$$\hat{\epsilon}(K_p, \{\hat{h}_{r,j}\}_{j=1}^k) = \frac{1}{r^2}\sum_{j=1}^{k}\sum_{i=1}^{r}\sum_{t=1}^{r} K_p(x_i, x_t)\hat{h}_{r,j}(x_i)\hat{h}_{r,j}(x_t).$$

We first quantify the difference between the above two terms by the concentration properties for sampling without replacement, as shown in the following lemma.

Lemma A.3. *Under Assumption 4.1,*

$$\left|\hat{\epsilon}(K_p, \{\hat{h}_{n,j}\}_{j=1}^k) - \hat{\epsilon}(K_p, \{\hat{h}_{r,j}\}_{j=1}^k)\right| \precsim k\sqrt{\left(\frac{1}{r} - \frac{1}{n}\right)\log\left(\frac{k}{\delta}\right)},$$

holds with probability at least $1 - \delta$.

Let $\boldsymbol{\alpha}$ be some kernel weights obtained from the training set S_n and $\boldsymbol{\beta}$ be the kernel weights from S_r. Similar to Theorem A.2, we have the important undermentioned theorem.

Theorem A.4. *Under Assumption 4.1,*

$$\left|\hat{\epsilon}(K_p, \{\hat{h}_{n,j}^{\boldsymbol{\alpha}}\}_{j=1}^k) - \hat{\epsilon}(K_p, \{\hat{h}_{r,j}^{\boldsymbol{\beta}}\}_{j=1}^k)\right| \precsim \|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_\infty + k\sqrt{\left(\frac{1}{r} - \frac{1}{n}\right)\log\left(\frac{k}{\delta}\right)},$$

holds with probability at least $1 - \delta$.

By Theorem A.4, Theorem 4.3 can be obtained similarly as the proof of Theorem 4.2.

B. Detailed Proof

B.1. The Proof of Lemma A.1

The proof is based on a concentration inequality and the bounded difference of two operators. We first introduce the two lemmas which are important to our proof.

The first lemma is the famous McDiarmid's inequality.

Lemma B.1. *(McDiarmid, 1989) If f has c-bounded differences on the sample space $\mathcal{X}$, then for all $\varepsilon > 0$:*

$$\Pr(|f(S_n) - \mathop{\mathbb{E}}_{S_n}[f(S_n)]| \geq \varepsilon) \leq 2\exp\left(-\frac{2\varepsilon^2}{nc^2}\right),$$

where f is some function of S_n.

The second lemma is about the bounded difference of an operator defined in Hilbert space and its empirical version.

Lemma B.2. *(Rosasco et al., 2010) Define two operators as*

$$T_n : \mathcal{H} \to \mathcal{H},\ T_n = \frac{1}{n}\sum_{i=1}^{n}\langle \cdot, K_{x_i}\rangle K_{x_i},$$

and

$$T_{\mathcal{H}} : \mathcal{H} \to \mathcal{H},\ T_{\mathcal{H}} = \int_{\mathcal{X}} \langle \cdot, K_x\rangle K_x d\rho(x).$$

The operators $T_{\mathcal{H}}$ and T_n are Hilbert-Schmidt. Assume that $K(x,x) \leq 1$, then

$$\|T_n - T_{\mathcal{H}}\| \leq \frac{2\sqrt{2}\log(2/\delta)}{\sqrt{n}}$$

holds with probability at least $1-\delta$.

Proof. By the triangle inequality, we have

$$\begin{aligned}
&\left|\hat{\epsilon}(K, \{\hat{h}_j\}_{j=1}^k) - \epsilon(K, \{h_j\}_{j=1}^k)\right| \\
\leq &\sum_{j=1}^{k}\left|\frac{1}{n^2}\sum_{i=1}^{n}\sum_{t=1}^{n} K(x_i,x_t)\hat{h}_j(x_i)\hat{h}_j(x_t) - \iint_{\mathcal{X}} K(x,y)h_j(x)h_j(y)d\rho(x)d\rho(y)\right|
\end{aligned}$$

For the j-th term,

$$\begin{aligned}
&\left|\frac{1}{n^2}\sum_{i=1}^{n}\sum_{t=1}^{n} K(x_i,x_t)\hat{h}_j(x_i)\hat{h}_j(x_t) - \iint_{\mathcal{X}} K(x,y)h_j(x)h_j(y)d\rho(x)d\rho(y)\right| \\
\leq &\underbrace{\left|\frac{1}{n^2}\sum_{i=1}^{n}\sum_{t=1}^{n} K(x_i,x_t)\hat{h}_j(x_i)\hat{h}_j(x_t) - \iint_{\mathcal{X}} K(x,y)\hat{h}_j(x)\hat{h}_j(y)d\rho(x)d\rho(y)\right|}_{\mathcal{C}} \\
+ &\underbrace{\left|\iint_{\mathcal{X}} K(x,y)\hat{h}_j(x)\hat{h}_j(y)d\rho(x)d\rho(y) - \iint_{\mathcal{X}} K(x,y)h_j(x)h_j(y)d\rho(x)d\rho(y)\right|}_{\mathcal{D}}.
\end{aligned}$$

In $\mathcal{C}$, notice that the latter term is the expectation of the ahead one, so we can give an upper bound of it by McDiarmid's inequality. Denote that

$$G(S_n) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{t=1}^{n} K(x_i,x_t)\hat{h}_j(x_i)\hat{h}_j(x_t).$$

We replace the l-th sample x_l of S_n with x'_l and denote the new training set as S'_n. It can be checked that

$$\begin{aligned} & G(S_n) - G(S'_n) \\ \leq & \frac{2}{n^2} \sum_{t=1}^{n} \left| K(x_l, x_t)\hat{h}_j(x_l)\hat{h}_j(x_t) - K(x'_l, x_t)\hat{h}_j(x'_l)\hat{h}_j(x_t) \right| \\ \leq & \frac{4}{n}. \end{aligned} \tag{9}$$

By McDiarmid's inequality (Lemma B.1), we know that there exists a constant $c > 0$ such that

$$\mathcal{C} \leq c\sqrt{\frac{\log(2/\delta)}{n}},$$

holds with probability at least $1 - \delta$.

Then we bound Term $\mathcal{D}$, we have

$$\begin{aligned} \mathcal{D} &= \left| \iint_{\mathcal{X}} K(x, y)(\hat{h}_j(x)\hat{h}_j(y) - h_j(x)h_j(y))d\rho(x)d\rho(y) \right| \\ &\leq \sup_{x,y} |\hat{h}_j(x)\hat{h}_j(y) - h_j(x)h_j(y)|. \end{aligned}$$

We first define the following operator.

$$\hat{L}_K f(x) := \frac{1}{n} \sum_{i=1}^{n} K(x, x_i) f(x_i).$$

$\hat{h}_j$ is the eigenfunction of $\hat{L}_K$, because

$$\hat{L}_K \hat{h}_j(x) = \frac{1}{n} \sum_{i=1}^{n} K(x, x_i)\hat{h}_j(x_i) = \hat{\lambda}_j \hat{h}_j(x).$$

The following proof is similar to the method in (Von Luxburg et al., 2008), which discusses the consistency of spectral clustering. For completeness, we also give detailed proof. By Theorem 7 and Proposition 18 of (Von Luxburg et al., 2008), we know that there exists a sequence $\{a_j\}_j \in \{1, -1\}$ and a constant C such that

$$\begin{aligned} \|a_j\hat{h}_j - h_j\|_\infty \leq & 2\|h_j - P_j(K)h_j\|_\infty \\ \leq & 2C(\|(\hat{L}_K - L_K)h_j\|_\infty + \|(\hat{L}_K - L_K)\hat{L}_K\|) \\ \leq & 2C(\|\hat{L}_K - L_K\|\|h_j\|_\infty + \|\hat{L}_K - L_K\|\|\hat{L}_K\|), \end{aligned}$$

where $P_j(K)$ is the orthogonal projector onto the subspace spanned by j-th eigenvector of the kernel function K.

Because

$$\sup_{\|f\|_\infty=1} \|\hat{L}_K f\|_\infty = \sup_{\|f\|_\infty=1} \left| \frac{1}{n} \sum_{i=1}^{n} K(x, x_i) f(x_i) \right| \leq \sup_{\|f\|_\infty=1} \frac{1}{n} \sum_{i=1}^{n} |K(x, x_i)|\,|f(x_i)| \leq 1,$$

we can obtain $\|\hat{L}_K\| \leq 1$.

By the definition of $T_{\mathcal{H}}$ and T_n in Lemma B.2, we can scale $\|a_j\hat{h}_j - h_j\|_\infty$ as

$$\begin{aligned}
&\|a_j\hat{h}_j - h_j\|_\infty \\
\leq& 4C\|\hat{L}_K - L_K\| \\
=& 4C \sup_{\substack{x\in\mathcal{X}\\ \|f\|_\infty=1}} \left|\frac{1}{n}\sum_{i=1}^{n} K(x,x_i)f(x_i) - \int_{\mathcal{X}} K(x,y)f(y)d\rho(y)\right| \\
\leq& 4C \sup_{\substack{x\in\mathcal{X}\\ \|f\|_\infty=1}} \left|\left\langle K_x, \frac{1}{n}\sum_{i=1}^{n} K_{x_i}f(x_i) - \int_{\mathcal{X}} K_y f(y)d\rho(y)\right\rangle_{\mathcal{H}_K}\right| \\
=& 4C \sup_{\substack{K_x, f\in\mathcal{H}_K\\ f\neq 0}} \frac{|\langle K_x, (T_n - T_{\mathcal{H}})f\rangle_{\mathcal{H}_K}|}{\|f\|_\infty} \\
=& 4C \sup_{\substack{K_x, f\in\mathcal{H}_K\\ f\neq 0}} \frac{|\langle K_x, (T_n - T_{\mathcal{H}})f\rangle_{\mathcal{H}_K}|}{\sup_{x\in X}|\langle K_x, f\rangle_{\mathcal{H}_K}|} \\
\leq& 4C \sup_{\substack{K_x, f\in\mathcal{H}_K\\ f\neq 0}} \left|\frac{\langle K_x, (T_n - T_{\mathcal{H}})f\rangle_{\mathcal{H}_K}}{\langle f, f\rangle_{\mathcal{H}_K}}\right| \\
\leq& 4C \sup_{\substack{K_x, f\in\mathcal{H}_K\\ f\neq 0}} \frac{\|K_x\|_{\mathcal{H}_K}\|(T_n - T_{\mathcal{H}})f\|_{\mathcal{H}_K}}{\|f\|_{\mathcal{H}_K}} \\
\leq& 4C\|T_n - T_{\mathcal{H}}\|.
\end{aligned}$$

Combining Lemma B.2, we know that there exists a constant $c > 0$ such that

$$\|a_j\hat{h}_j - h_j\|_\infty \leq c\sqrt{\frac{\log(2/\delta)}{n}}$$

holds with probability at least $1-\delta$.

On the other hand, we can obtain

$$\begin{aligned}
\mathcal{D} &\leq \sup_{x,y} |a_j\hat{h}_j(x)\cdot a_j\hat{h}_j(y) - a_j\hat{h}_j(x)h_j(y) + a_j\hat{h}_j(x)h_j(y) - h_j(x)h_j(y)| \\
&\leq \sup_{x,y} |a_j\hat{h}_j(x)||a_j\hat{h}_j(y) - h_j(y)| + |h_j(y)||a_j\hat{h}_j(x) - h_j(x)| \\
&\leq 2\|a_j\hat{h}_j - h_j\|_\infty.
\end{aligned}$$

Thus, there exists $c > 0$ such that

$$\mathcal{D} \leq c\sqrt{\frac{\log(2/\delta)}{n}}$$

holds with probability at least $1-\delta$.

Above all, there is also a constant $c \geq 0$ such that

$$\mathcal{C} + \mathcal{D} \leq c\sqrt{\frac{\log(4/\delta)}{n}}$$

holds with probability at least $1-\delta$.

Then, we can obtain the final bound as

$$\left|\hat{\epsilon}(K, \{\hat{h}_j\}_{j=1}^k) - \epsilon(K, \{h_j\}_{j=1}^k)\right| \leq ck\sqrt{\frac{\log(4k/\delta)}{n}}$$

holds with probability at least $1-\delta$. This because $(1-\delta/k)^k \geq 1-\delta$. The proof is complete. □

B.2. The Proof of Theorem A.2

The following lemma is about the perturbation of eigenvectors of Hermitian matrices, which is useful to our proof.

Lemma B.3. *(Yu et al., 2014) Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$ be Hermitian, with eigenvalues $\lambda_1 \geq \cdots \lambda_n$ and $\hat{\lambda}_1 \geq \cdots \hat{\lambda}_n$ respectively. Fixed $1 \leq r \leq s \leq n$ and assume that $\min(\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1}) > 0$, where $\lambda_0 := \infty$ and $\lambda_{n+1} := -\infty$. Let $d := s - r + 1$, let $\mathbf{H} = [\mathbf{h}_r, \mathbf{h}_{r+1}, \cdots, \mathbf{h}_s] \in \mathbb{R}^{n\times d}$ and $\hat{\mathbf{H}} = [\hat{\mathbf{h}}_r, \hat{\mathbf{h}}_{r+1}, \cdots, \hat{\mathbf{h}}_s] \in \mathbb{R}^{n\times d}$ have orthonormal columns satisfying $\mathbf{A}\mathbf{h}_j = \lambda_j \mathbf{h}_j$ and $\mathbf{B}\hat{\mathbf{h}}_j = \hat{\lambda}_j \hat{\mathbf{h}}_j$ for $j = r, r+1, \cdots, s$. Then*

$$\left\|\sin\theta(\mathbf{H}, \hat{\mathbf{H}})\right\|_F \leq \frac{2\min(d^{1/2}\|\mathbf{A} - \mathbf{B}\|_{\text{op}}, \|\mathbf{A} - \mathbf{B}\|_{\text{F}})}{\min(\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1})},$$

where $\theta(\mathbf{H}, \hat{\mathbf{H}}) \in \mathbb{R}^{d\times d}$ is the diagonal matrix whose j-th diagonal entry is the j-th principal angle, i.e., $\arccos(\mathbf{h}_j^\top \hat{\mathbf{h}}_j)$.

Proof. We make a decomposition as

$$\begin{aligned}
&\left|\hat{\epsilon}(K_p, \{\hat{h}_j^{\boldsymbol{\alpha}}\}_{j=1}^k) - \epsilon(K_p, \{h_j^{\boldsymbol{\beta}}\}_{j=1}^k)\right| \\
=&\frac{1}{n}\text{Tr}\left(\mathbf{K}_p\mathbf{H}^{\boldsymbol{\alpha}}\mathbf{H}^{\boldsymbol{\alpha}\top}\right) - \sum_{j=1}^k \iint_{\mathcal{X}} K_p(x,y)h_j^{\boldsymbol{\beta}}(x)h_j^{\boldsymbol{\beta}}(y)d\rho(x)d\rho(y) \\
=&\underbrace{\frac{1}{n}\text{Tr}\left(\mathbf{K}_p\mathbf{H}^{\boldsymbol{\alpha}}\mathbf{H}^{\boldsymbol{\alpha}\top}\right) - \frac{1}{n}\text{Tr}\left(\mathbf{K}_p\mathbf{H}^{\boldsymbol{\beta}}\mathbf{H}^{\boldsymbol{\beta}\top}\right)}_{\mathcal{A}} \\
&+\underbrace{\frac{1}{n^2}\sum_{j=1}^k\sum_{i=1}^n\sum_{t=1}^n K_p(x_i,x_t)\hat{h}_j^{\boldsymbol{\beta}}(x_i)\hat{h}_j^{\boldsymbol{\beta}}(x_t) - \sum_{j=1}^k \iint_{\mathcal{X}} K_p(x,y)h_j^{\boldsymbol{\beta}}(x)h_j^{\boldsymbol{\beta}}(y)d\rho(x)d\rho(y)}_{\mathcal{B}}.
\end{aligned}$$

It is sufficient to bound Term $\mathcal{B}$ by Lemma A.1. We will use matrix perturbation theory (Stewart, 1990) to bound Term $\mathcal{A}$, and we can deduce that

$$\begin{aligned}
\mathcal{A} &= \frac{1}{n}\text{Tr}\left(\mathbf{K}_p\mathbf{H}^{\boldsymbol{\alpha}}\mathbf{H}^{\boldsymbol{\alpha}\top}\right) - \frac{1}{n}\text{Tr}\left(\mathbf{K}_p\mathbf{H}^{\boldsymbol{\beta}}\mathbf{H}^{\boldsymbol{\beta}\top}\right) \\
&\leq \left\|\frac{\mathbf{K}_p}{n}\right\|_{\text{F}} \left\|\mathbf{H}^{\boldsymbol{\alpha}}\mathbf{H}^{\boldsymbol{\alpha}\top} - \mathbf{H}^{\boldsymbol{\beta}}\mathbf{H}^{\boldsymbol{\beta}\top}\right\|_{\text{F}} \\
&\leq \sqrt{\frac{\text{Tr}(\mathbf{K}_p^2)}{n^2}} \left\|\mathbf{H}^{\boldsymbol{\alpha}}\mathbf{H}^{\boldsymbol{\alpha}\top} - \mathbf{H}^{\boldsymbol{\beta}}\mathbf{H}^{\boldsymbol{\beta}\top}\right\|_{\text{F}} \\
&\leq \sqrt{\frac{\text{Tr}^2(\mathbf{K}_p)}{n^2}} \left\|\mathbf{H}^{\boldsymbol{\alpha}}\mathbf{H}^{\boldsymbol{\alpha}\top} - \mathbf{H}^{\boldsymbol{\beta}}\mathbf{H}^{\boldsymbol{\beta}\top}\right\|_{\text{F}} \\
&\leq \left\|\mathbf{H}^{\boldsymbol{\alpha}}\mathbf{H}^{\boldsymbol{\alpha}\top} - \mathbf{H}^{\boldsymbol{\beta}}\mathbf{H}^{\boldsymbol{\beta}\top}\right\|_{\text{F}} \\
&= \sqrt{2\sum_{j=1}^k (1 - (\mathbf{h}_j^{\boldsymbol{\alpha}\top}\mathbf{h}_j^{\boldsymbol{\beta}})^2)} \\
&= \sqrt{2\sum_{j=1}^k (1 - \cos^2\theta(\mathbf{h}_j^{\boldsymbol{\alpha}}, \mathbf{h}_j^{\boldsymbol{\beta}}))} \\
&= \sqrt{2}\left\|\sin\theta(\mathbf{H}^{\boldsymbol{\alpha}}, \mathbf{H}^{\boldsymbol{\beta}})\right\|_{\text{F}}.
\end{aligned}$$

For any vector $\boldsymbol{\alpha} \in \mathbb{R}^m$, denote that $\delta(\boldsymbol{\alpha})$ is the gap of the k-th and $(k+1)$-th eigenvalues of matrix $\frac{1}{n}\mathbf{K}_{\boldsymbol{\alpha}}$. By the assumption that there exists a constant $c \geq 0$ such that for any $\boldsymbol{\alpha} \in \triangle$, $\delta(\boldsymbol{\alpha}) \geq 1/c$. By Theorem B.3, letting $r = 1, s = k$,

we have

$$\begin{aligned}
\mathcal{A} &\leq \sqrt{2}\left\|\sin\theta(\mathbf{H}^{\boldsymbol{\alpha}}, \mathbf{H}^{\boldsymbol{\beta}})\right\|_{\mathrm{F}} \\
&\leq \frac{2\sqrt{2}\left\|\frac{1}{n}\mathbf{K}_{\boldsymbol{\alpha}} - \frac{1}{n}\mathbf{K}_{\boldsymbol{\beta}}\right\|_{\mathrm{F}}}{\delta(\boldsymbol{\alpha})} \\
&\leq 2\sqrt{2}c\sqrt{\sum_{p=1}^{m}\sum_{i=1}^{n}\sum_{t=1}^{n}(\alpha_p^2 - \beta_p^2)^2\frac{K^2(x_i, x_t)}{n^2}} \\
&\leq 2\sqrt{2}c\sqrt{\sum_{p=1}^{m}(\alpha_p^2 - \beta_p^2)^2} \\
&= 2\sqrt{2}c\sqrt{\sum_{p=1}^{m}(\alpha_p - \beta_p)^2(\alpha_p + \beta_p)^2} \\
&\leq 2\sqrt{2}c\max_{p\in[m]}|\alpha_p - \beta_p|\sqrt{\sum_{p=1}^{m}(\alpha_p + \beta_p)^2} \\
&\leq 2\sqrt{2}c\max_{p\in[m]}|\alpha_p - \beta_p|\sqrt{2\sum_{p=1}^{m}(\alpha_p + \beta_p)} \\
&\leq 4\sqrt{2}c\max_{p\in[m]}|\alpha_p - \beta_p| \\
&= 4\sqrt{2}c\|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_{\infty}.
\end{aligned}$$

Above all, we know that

$$\frac{1}{n}\mathrm{Tr}\left(\mathbf{K}_p\mathbf{H}^{\boldsymbol{\alpha}}\mathbf{H}^{\boldsymbol{\alpha}\top}\right) - \sum_{j=1}^{k}\iint_{\mathcal{X}} K_p(x,y)h_j^{\boldsymbol{\beta}}(x)h_j^{\boldsymbol{\beta}}(y)d\rho(x)d\rho(y) \leq c\|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_{\infty} + ck\sqrt{\frac{\log(4k/\delta)}{n}}$$

holds with probability at least $1-\delta$. The proof is complete. □

B.3. The Proof of Theorem 4.2

Proof. For ease of proof, we briefly denote $\epsilon(K_p, \{h_j^{\boldsymbol{\alpha}^{(t)}}\}_{j=1}^k)$ and $\hat{\epsilon}(K_p, \{\hat{h}_j^{\boldsymbol{\alpha}^{(t)}}\}_{j=1}^k)$ as $\epsilon_p(\boldsymbol{\alpha}^{(t)})$ and $\hat{\epsilon}_p(\boldsymbol{\alpha}^{(t)})$ respectively. When the training set is S_n, assume that the kernel weights of each iteration are

$$\boldsymbol{\alpha}^{(0)}, \cdots, \boldsymbol{\alpha}^{(T)},$$

Meanwhile, the kernel weights of each iteration learned from the sample space are denoted as

$$\boldsymbol{\beta}^{(0)}, \cdots, \boldsymbol{\beta}^{(T)}.$$

where $\boldsymbol{\alpha}^0 = \boldsymbol{\beta}^0$ due to the same initialization.

Because SimpleMKKM can obtain the globally optimal solution, after T iterations, we have

$$\boldsymbol{\alpha}^{(T)} = \hat{\boldsymbol{\alpha}}_n,\ \boldsymbol{\beta}^{(T)} = \boldsymbol{\alpha}^*.$$

For any integers $t \geq 1$ and $p \in [m]$, if $p = u$, we have

$$\begin{aligned}
&\left|\alpha_u^{(t)} - \beta_u^{(t)}\right| - \left|\alpha_u^{(t-1)} - \beta_u^{(t-1)}\right| \\
\precsim &\frac{1}{m-1}\left|\sum_{p\neq u}\left(\hat{\alpha}_p\hat{\epsilon}_p(\boldsymbol{\alpha}^{(t-1)}) - \hat{\alpha}_u\hat{\epsilon}_u(\boldsymbol{\alpha}^{(t-1)})\right) - \sum_{p\neq u}\left(\beta_p\epsilon_p(\boldsymbol{\beta}^{(t-1)}) - \beta_u\epsilon_u(\boldsymbol{\beta}^{(t-1)})\right)\right| \\
= &\frac{1}{m-1}\left|\sum_{p\neq u}\left(\hat{\alpha}_p\hat{\epsilon}_p(\boldsymbol{\alpha}^{(t-1)}) - \beta_p\epsilon_p(\boldsymbol{\beta}^{(t-1)})\right) - (m-1)\left(\hat{\alpha}_u\hat{\epsilon}_u(\boldsymbol{\alpha}^{(t-1)}) - \beta_u\epsilon_u(\boldsymbol{\beta}^{(t-1)})\right)\right| \\
\precsim &\frac{1}{m-1}\sum_{p\neq u}\left|\left(\hat{\alpha}_p\hat{\epsilon}_p(\boldsymbol{\alpha}^{(t-1)}) - \beta_p\epsilon_p(\boldsymbol{\beta}^{(t-1)})\right)\right| + \left|\left(\hat{\alpha}_u\hat{\epsilon}_u(\boldsymbol{\alpha}^{(t-1)}) - \beta_u\epsilon_u(\boldsymbol{\beta}^{(t-1)})\right)\right| \\
\precsim &\max_{p\in[m]}\left|\hat{\alpha}_p\hat{\epsilon}_p(\boldsymbol{\alpha}^{(t-1)}) - \beta_p\epsilon_p(\boldsymbol{\beta}^{(t-1)})\right| \\
= &\max_{p\in[m]}\left|\hat{\alpha}_p\hat{\epsilon}_p(\boldsymbol{\alpha}^{(t-1)}) - \beta_p\hat{\epsilon}_p(\boldsymbol{\alpha}^{(t-1)}) + \beta_p\hat{\epsilon}_p(\boldsymbol{\alpha}^{(t-1)}) - \beta_p\epsilon_p(\boldsymbol{\beta}^{(t-1)})\right| \\
\precsim &\max_{p\in[m]}\left|(\hat{\alpha}_p - \beta_p)\hat{\epsilon}_p(\boldsymbol{\alpha}^{(t-1)})\right| + \beta_p\left|\hat{\epsilon}_p(\boldsymbol{\alpha}^{(t-1)}) - \epsilon_p(\boldsymbol{\beta}^{(t-1)})\right| \\
\precsim &\left\|\boldsymbol{\alpha}^{(t-1)} - \boldsymbol{\beta}^{(t-1)}\right\|_\infty + \left|\hat{\epsilon}_p(\boldsymbol{\alpha}^{(t-1)}) - \epsilon_p(\boldsymbol{\beta}^{(t-1)})\right|.
\end{aligned}$$

By Theorem A.2, and

$$\left|\alpha_u^{(t-1)} - \beta_u^{(t-1)}\right| \leq \left\|\boldsymbol{\alpha}^{(t-1)} - \boldsymbol{\beta}^{(t-1)}\right\|_\infty,$$

we have

$$\left|\alpha_u^{(t)} - \beta_u^{(t)}\right| \precsim \left\|\boldsymbol{\alpha}^{(t-1)} - \boldsymbol{\beta}^{(t-1)}\right\|_\infty + k\sqrt{\frac{\log(k/\delta)}{n}}.$$

Similarly, for $p \neq u$, we have

$$\left\|\boldsymbol{\alpha}^{(t)} - \boldsymbol{\beta}^{(t)}\right\|_\infty \precsim \left\|\boldsymbol{\alpha}^{(t-1)} - \boldsymbol{\beta}^{(t-1)}\right\|_\infty + k\sqrt{\frac{\log(k/\delta)}{n}}.$$

Finally, due to the convergence of the reduced gradient descent algorithm,

$$\left\|\boldsymbol{\alpha}^{(T)} - \boldsymbol{\beta}^{(T)}\right\|_\infty \precsim \cdots \precsim \left\|\boldsymbol{\alpha}^{(0)} - \boldsymbol{\beta}^{(0)}\right\|_\infty + k\sqrt{\frac{\log(k/\delta)}{n}}.$$

It means that

$$\|\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^*\|_\infty \precsim k\sqrt{\frac{\log(k/\delta)}{n}},$$

holds with probability at least $1 - \delta$.

□

B.4. The Proof of Lemma A.3

Theorem B.4. *(Zwald & Blanchard, 2005) Let A be a symmetric positive Hilbert-Schmidt operator of the Hilbert space with simple positive eigenvalues $\lambda_1 > \lambda_2 > \cdots$. For an integer j such that $\lambda_j > 0$, let $\tilde{\sigma}_j = \sigma_j \wedge \sigma_{j+1}$ where $\sigma_j = \frac{1}{2}(\lambda_j - \lambda_{j+1})$. Let $B \in \mathrm{HS}(\mathcal{H})$ be another symmetric operator such that $\|B\| < \tilde{\sigma}_j/2$ and $A + B$ is still a positive operator with simple nonzero eigenvalues. Let $P_j(A)$ be the orthogonal projector onto the subspace spanned by j-th eigenvector of A. Then, these projectors satisfy the following:*

$$\|P_j(A) - P_j(A+B)\| \leq \frac{2\|B\|}{\tilde{\sigma}_j}.$$

Theorem B.5. *(Bardenet & Maillard, 2015) Let $S_n = \{x_i\}_{i=1}^n$ be a finite sequence of real numbers, and $S_r = \{x_i\}_{i=1}^r$ are r points uniformly selected from it without replacement. Then, for any $t > 0$, the following probability inequality holds*

$$\Pr\left(\left|\frac{1}{r}\sum_{i=1}^{r} x_i - \frac{1}{n}\sum_{i=1}^{n} x_i\right| \geq t\right) \leq 2\exp\left(-\frac{2rt^2}{(1-r/n)(1+1/r)(b-a)^2}\right),$$

where $a = \min_{i\in[n]} x_i$ and $b = \max_{i\in[n]} x_i$.

Proof. Without loss of generality, we assume that the selected points are $\{x_i\}_{i=1}^r$. Then for any kernel function $K(\cdot,\cdot)$, the following inequality holds

$$\begin{aligned}
&\left|\hat{\epsilon}(K, \{\hat{h}_{n,j}\}_{j=1}^k) - \hat{\epsilon}(K, \{\hat{h}_{r,j}\}_{j=1}^k)\right| \\
\leq &\sum_{j=1}^{k}\left|\sum_{i=1}^{n}\sum_{t=1}^{n} K(x_i,x_t)\hat{h}_{n,j}(x_i)\hat{h}_{n,j}(x_t) - \sum_{i=1}^{r}\sum_{t=1}^{r} K(x_i,x_t)\hat{h}_{r,j}(x_i)\hat{h}_{r,j}(x_t)\right|.
\end{aligned}$$

The difference of j-th term in the above formula can be bounded by

$$\begin{aligned}
&\left|\frac{1}{n^2}\sum_{i=1}^{n}\sum_{t=1}^{n} K(x_i,x_t)\hat{h}_{n,j}(x_i)\hat{h}_{n,j}(x_t) - \frac{1}{r^2}\sum_{i=1}^{r}\sum_{t=1}^{r} K(x_i,x_t)\hat{h}_{r,j}(x_i)\hat{h}_{r,j}(x_t)\right| \\
\leq &\underbrace{\left|\frac{1}{n^2}\sum_{i=1}^{n}\sum_{t=1}^{n} K(x_i,x_t)\hat{h}_{n,j}(x_i)\hat{h}_{n,j}(x_t) - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{t=1}^{n} K(x_i,x_t)\hat{h}_{r,j}(x_i)\hat{h}_{r,j}(x_t)\right|}_{\mathcal{C}} \\
+ &\underbrace{\left|\frac{1}{n^2}\sum_{i=1}^{n}\sum_{t=1}^{n} K(x_i,x_t)\hat{h}_{r,j}(x_i)\hat{h}_{r,j}(x_t) - \frac{1}{r^2}\sum_{i=1}^{r}\sum_{t=1}^{r} K(x_i,x_t)\hat{h}_{r,j}(x_i)\hat{h}_{r,j}(x_t)\right|}_{\mathcal{D}}.
\end{aligned}$$

We first bound Term $\mathcal{C}$ as

$$\begin{aligned}
\mathcal{C} = &\left|\frac{1}{n^2}\sum_{i=1}^{n}\sum_{t=1}^{n} K(x_i,x_t)(\hat{h}_{n,j}(x_i)\hat{h}_{n,j}(x_t) - \hat{h}_{r,j}(x_i)\hat{h}_{r,j}(x_t))\right| \\
\leq &\frac{1}{n^2}\sum_{i=1}^{n}\sum_{t=1}^{n}\left|K(x_i,x_t)(\hat{h}_{n,j}(x_i)\hat{h}_{n,j}(x_t) - \hat{h}_{r,j}(x_i)\hat{h}_{r,j}(x_t))\right| \\
\leq &\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sup_{x,y}\left|K(x,y)(\hat{h}_{n,j}(x)\hat{h}_{n,j}(y) - \hat{h}_{r,j}(x)\hat{h}_{r,j}(y))\right| \\
\leq &\sup_{x,y}\left|\hat{h}_{n,j}(x)\hat{h}_{n,j}(y) - \hat{h}_{r,j}(x)\hat{h}_{r,j}(y)\right| \\
= &\sup_{x,y}\left|\hat{h}_{n,j}(x)\hat{h}_{n,j}(y) - \hat{h}_{n,j}(x)\hat{h}_{r,j}(y) + \hat{h}_{n,j}(x)\hat{h}_{r,j}(y) - \hat{h}_{r,j}(x)\hat{h}_{r,j}(y)\right| \\
\leq &\sup_{x,y}|\hat{h}_{n,j}(x)|\left|\hat{h}_{n,j}(y) - \hat{h}_{r,j}(y)\right| + \sup_{x,y}|\hat{h}_{r,j}(y)|\left|\hat{h}_{n,j}(x) - \hat{h}_{r,j}(x)\right| \\
\leq &2\sup_{x}\left|\hat{h}_{n,j}(x) - \hat{h}_{r,j}(x)\right|.
\end{aligned}$$

Similar to the definition of Lemma B.2, we define the following two operators $T_n, T_r : \mathcal{H} \to \mathcal{H}$:

$$T_n = \frac{1}{n}\sum_{i=1}^{n}\langle\cdot, K_{x_i}\rangle K_{x_i},\ T_r = \frac{1}{r}\sum_{i=1}^{r}\langle\cdot, K_{x_i}\rangle K_{x_i}.$$

According to (Rosasco et al., 2010), the above two operators are positive definite, and it can be checked that

$$T_n\hat{h}_{n,j}(x) = \frac{1}{n}\sum_{i=1}^{n} K(x,x_i)\hat{h}_{n,j}(x_i) = \hat{\lambda}_{n,j}.$$

Thus $\hat{h}_{n,j} \in \mathcal{H}$ is the eigenfunction of T_n, while $\hat{h}_{r,j}$ is the eigenfunction of T_r, and their corresponding non-zero eigenvalues are the same as the eigenvalues of $\frac{1}{n}\mathbf{K}_n$ and $\frac{1}{r}\mathbf{K}_r$, respectively. We assume that the minimal gap of the first $k+1$ eigenvalues of $\frac{1}{n}\mathbf{K}_n$ is $\tilde{\sigma}$.

By the reproducing property of the kernel function and Theorem B.4, we have

$$\begin{aligned}
\mathcal{C} =& 2\sup_x \left|\hat{h}_{n,k}(x) - \hat{h}_{r,k}(x)\right| \\
\leq& 2\sup_x \left|\left\langle K_x, \hat{h}_{n,k} - \hat{h}_{r,k}\right\rangle_{\mathcal{H}}\right| \\
\leq& 2\sup_x \|K_x\| \cdot \left\|\hat{h}_{n,k} - \hat{h}_{r,k}\right\| \\
\leq& 2\left\|\hat{h}_{n,k} - \hat{h}_{r,k}\right\| \\
\leq& \frac{2\|T_r - T_n\|}{\tilde{\sigma}}.
\end{aligned} \tag{10}$$

We will give the conditions of the last inequality later. Before this, we aim to bound $\|T_r - T_n\|$.

$$\begin{aligned}
&\|T_r - T_n\| \\
=& \sup_{f\in\mathcal{H},\|f\|=1} \|T_r f - T_n f\| \\
\leq& \sup_{\substack{f,v\in\mathcal{H},\\ \|f\|=\|v\|=1}} \langle T_r f - T_n f, v\rangle_{\mathcal{H}} \\
=& \sup_{\substack{f,v\in\mathcal{H},\\ \|f\|=\|v\|=1}} \left|\frac{1}{r}\sum_{i=1}^{r} f(x_i)\langle K_{x_i}, v\rangle_{\mathcal{H}} - \frac{1}{n}\sum_{i=1}^{n} f(x_i)\langle K_{x_i}, v\rangle_{\mathcal{H}}\right|.
\end{aligned}$$

By Theorem B.5, we have

$$\Pr\left(\left|\frac{1}{r}\sum_{i=1}^{r} f(x_i)\langle K_{x_i}, v\rangle_{\mathcal{H}} - \frac{1}{n}\sum_{i=1}^{n} f(x_i)\langle K_{x_i}, v\rangle_{\mathcal{H}}\right| \geq t\right) \leq 2\exp\left(-\frac{rt^2}{2(1-r/n)(1+1/r)}\right).$$

Then the following inequality holds with probability $1-\delta$:

$$\|T_r - T_n\| \precsim \sqrt{\left(\frac{1}{r} - \frac{1}{n}\right)\log\left(\frac{2}{\delta}\right)}.$$

Now we find the establishment condition of the last inequality in Eq. (10). By the conditions of Theorem B.4, we know that r should satisfy

$$r = \Omega\left(\frac{n}{1+n\tilde{\sigma}^2}\right).$$

We let $n = \Omega(1/\tilde{\sigma}^2)$, and r will be $\Omega(1/\tilde{\sigma}^2)$.

Above all, if $\tilde{\sigma} \geq 1/c$ and r, n are $\Omega(1/\tilde{\sigma}^2)$, then we have

$$\mathcal{C} \precsim \sqrt{\left(\frac{1}{r} - \frac{1}{n}\right)\log\left(\frac{2}{\delta}\right)}.$$

In the next, we proceed to bound Term $\mathcal{D}$. Notice that

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{t=1}^{n}K(x_i,x_t)\hat{h}_{r,j}(x_i)\hat{h}_{r,j}(x_t)=\left\|\frac{1}{n}\sum_{i=1}^{n}\hat{h}_{r,j}(x_i)\phi(x_i)\right\|^2.$$

Thus, we have

$$\begin{aligned}\mathcal{D}=&\left|\left\langle\left(\frac{1}{n}\sum_{i=1}^{n}\hat{h}_{r,j}(x_i)\phi(x_i)-\frac{1}{r}\sum_{i=1}^{r}\hat{h}_{r,j}(x_i)\phi(x_i)\right),\left(\frac{1}{n}\sum_{i=1}^{n}\hat{h}_{r,j}(x_i)\phi(x_i)+\frac{1}{r}\sum_{i=1}^{r}\hat{h}_{r,j}(x_i)\phi(x_i)\right)\right\rangle_{\mathcal{H}}\right|\\ \leq&\left\|\frac{1}{n}\sum_{i=1}^{n}\hat{h}_{r,j}(x_i)\phi(x_i)-\frac{1}{r}\sum_{i=1}^{r}\hat{h}_{r,j}(x_i)\phi(x_i)\right\|\cdot\left\|\frac{1}{n}\sum_{i=1}^{n}\hat{h}_{r,j}(x_i)\phi(x_i)+\frac{1}{r}\sum_{i=1}^{r}\hat{h}_{r,j}(x_i)\phi(x_i)\right\|\\ \leq&2\left\|\frac{1}{n}\sum_{i=1}^{n}\hat{h}_{r,j}(x_i)\phi(x_i)-\frac{1}{r}\sum_{i=1}^{r}\hat{h}_{r,j}(x_i)\phi(x_i)\right\|\\ \leq&2\sup_{\|v\|=1,v\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n}\hat{h}_{r,j}(x_i)\langle v,\phi(x_i)\rangle_{\mathcal{H}}-\frac{1}{r}\sum_{i=1}^{r}\hat{h}_{r,j}(x_i)\langle v,\phi(x_i)\rangle_{\mathcal{H}}\right|.\end{aligned}$$

For any $v_0\in\mathcal{H}$, $\|v_0\|=1$, by Theorem B.5, we have

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}\hat{h}_{r,j}(x_i)\langle v_0,\phi(x_i)\rangle_{\mathcal{H}}-\frac{1}{r}\sum_{i=1}^{r}\hat{h}_{r,j}(x_i)\langle v_0,\phi(x_i)\rangle_{\mathcal{H}}\right|\geq t\right)\leq 2\exp\left(-\frac{2rt^2}{4(1-r/n)(1+1/r)}\right).$$

Then, we know that

$$\mathcal{D}\precsim\sqrt{\left(\frac{1}{r}-\frac{1}{n}\right)\log\left(\frac{2}{\delta}\right)}$$

holds with probability at least $1-\delta$.

Combining all things together, the following inequality holds with probability $1-\delta$:

$$\left|\hat{\epsilon}(K_p,\{\hat{h}_{n,j}\}_{j=1}^{k})-\hat{\epsilon}(K_p,\{\hat{h}_{r,j}\}_{j=1}^{k})\right|\precsim k\sqrt{\left(\frac{1}{r}-\frac{1}{n}\right)\log\left(\frac{k}{\delta}\right)}.$$

□

B.5. The Proof of Theorem 5.1

Proof. We make a decomposition as follows:

$$\begin{aligned}&\mathbb{E}_{S_n}[\mathcal{W}(\hat{\mathbf{C}},\hat{\boldsymbol{\alpha}},\rho)]-\mathcal{W}(\mathbf{C}^*,\boldsymbol{\alpha}^*,\rho)\\ =&\underbrace{\mathbb{E}_{S_n}[\mathcal{W}(\hat{\mathbf{C}},\hat{\boldsymbol{\alpha}},\rho)-\mathcal{W}_n(\hat{\mathbf{C}},\hat{\boldsymbol{\alpha}},\rho_n)]}_{\mathcal{A}}+\underbrace{\mathbb{E}_{S_n}[\mathcal{W}_n(\hat{\mathbf{C}},\hat{\boldsymbol{\alpha}},\rho_n)-\mathcal{W}_n(\mathbf{C}^*,\boldsymbol{\alpha}^*,\rho_n)]}_{\mathcal{B}}\\ &+\underbrace{\mathbb{E}_{S_n}[\mathcal{W}_n(\mathbf{C}^*,\boldsymbol{\alpha}^*,\rho_n)]-\mathcal{W}(\mathbf{C}^*,\boldsymbol{\alpha}^*,\rho)}_{\mathcal{C}}.\end{aligned}$$

$\mathcal{A}$ and $\mathcal{C}$ can be bounded by the generalization risk of single kernel clustering (Biau et al., 2008), and their upper bounds are all $\mathcal{O}(k/\sqrt{n})$.

Then we bound Term $\mathcal{B}$. Assume that the clustering indicator matrix corresponding to clustering centroids $\hat{\mathbf{C}}$ is $\hat{\mathbf{H}}_{\hat{\boldsymbol{\alpha}}}\in\mathbb{R}^{n\times k}$, whose element is $h_{ij}=1/\sqrt{|\mathcal{C}_j|}$. When x_i belongs to the j-th cluster $h_{ij}=1/\sqrt{|\mathcal{C}_j|}$, otherwise $h_{ij}^{\hat{\boldsymbol{\alpha}}}=0$. Similarly, we denote the clustering indicator matrix corresponding to $\mathbf{C}^*$ is $\hat{\mathbf{H}}_{\boldsymbol{\alpha}^*}\in\mathbb{R}^{n\times k}$.

Because $\hat{\mathbf{C}} = \operatorname{argmin}_{\mathbf{C}\in\mathcal{H}^k_{\hat{\boldsymbol{\alpha}}}} \mathcal{W}_n(\mathbf{C}, \hat{\boldsymbol{\alpha}}, \rho_n)$, we have

$$
\begin{aligned}
\mathcal{B} =& \mathbb{E}_{S_n}[\mathcal{W}_n(\hat{\mathbf{C}}, \hat{\boldsymbol{\alpha}}, \rho_n) - \mathcal{W}_n(\mathbf{C}^*, \boldsymbol{\alpha}^*, \rho_n)] \\
=& \mathbb{E}_{S_n}\left[\frac{1}{n}\mathrm{Tr}\left(\mathbf{K}_{\hat{\boldsymbol{\alpha}}}(\mathbf{I}_n - \hat{\mathbf{H}}_{\hat{\boldsymbol{\alpha}}}\hat{\mathbf{H}}_{\hat{\boldsymbol{\alpha}}}^\top)\right) - \frac{1}{n}\mathrm{Tr}\left(\mathbf{K}_{\boldsymbol{\alpha}^*}(\mathbf{I}_n - \hat{\mathbf{H}}_{\boldsymbol{\alpha}^*}\hat{\mathbf{H}}_{\boldsymbol{\alpha}^*}^\top)\right)\right] \\
=& \mathbb{E}_{S_n}\left[\frac{1}{n}\mathrm{Tr}\left(\mathbf{K}_{\hat{\boldsymbol{\alpha}}}(\mathbf{I}_n - \hat{\mathbf{H}}_{\hat{\boldsymbol{\alpha}}}\hat{\mathbf{H}}_{\hat{\boldsymbol{\alpha}}}^\top)\right) - \frac{1}{n}\mathrm{Tr}\left(\mathbf{K}_{\hat{\boldsymbol{\alpha}}}(\mathbf{I}_n - \hat{\mathbf{H}}_{\boldsymbol{\alpha}^*}\hat{\mathbf{H}}_{\boldsymbol{\alpha}^*}^\top)\right)\right] \\
&+ \mathbb{E}_{S_n}\left[\frac{1}{n}\mathrm{Tr}\left(\mathbf{K}_{\hat{\boldsymbol{\alpha}}}(\mathbf{I}_n - \hat{\mathbf{H}}_{\boldsymbol{\alpha}^*}\hat{\mathbf{H}}_{\boldsymbol{\alpha}^*}^\top)\right) - \frac{1}{n}\mathrm{Tr}\left(\mathbf{K}_{\boldsymbol{\alpha}^*}(\mathbf{I}_n - \hat{\mathbf{H}}_{\boldsymbol{\alpha}^*}\hat{\mathbf{H}}_{\boldsymbol{\alpha}^*}^\top)\right)\right] \\
\leq& \mathbb{E}_{S_n}\left[\frac{1}{n}\mathrm{Tr}\left((\mathbf{K}_{\hat{\boldsymbol{\alpha}}} - \mathbf{K}_{\boldsymbol{\alpha}^*})(\mathbf{I}_n - \hat{\mathbf{H}}_{\boldsymbol{\alpha}^*}\hat{\mathbf{H}}_{\boldsymbol{\alpha}^*}^\top)\right)\right] \\
=& \mathbb{E}_{S_n}\left[\sum_{p=1}^m (\hat{\alpha}_p^2 - (\alpha_p^*)^2)\cdot\frac{1}{n}\mathrm{Tr}\left(\mathbf{K}_p(\mathbf{I}_n - \hat{\mathbf{H}}_{\boldsymbol{\alpha}^*}\hat{\mathbf{H}}_{\boldsymbol{\alpha}^*}^\top)\right)\right] \\
\leq& \sum_{p=1}^m (\hat{\alpha}_p + \alpha_p^*)|\hat{\alpha}_p - \alpha_p^*| \\
\leq& \sum_{p=1}^m (\hat{\alpha}_p + \alpha_p^*)\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_\infty \\
=& 2\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_\infty \\
\lesssim& k\sqrt{\frac{\log(k/\delta)}{n}}.
\end{aligned}
$$

The proof is complete. □

B.6. The Proof of Theorem 6.1

Proof. When the kernel weights are $\hat{\boldsymbol{\alpha}}_r$, assume that the corresponding clustering centroids are $\hat{\mathbf{C}}_n \in \mathcal{H}^k_{\hat{\boldsymbol{\alpha}}_r}$. Then, we have

$$
\begin{aligned}
&\mathbb{E}_{S_n}[\mathcal{W}(\hat{\mathbf{C}}_{n,r}, \hat{\boldsymbol{\alpha}}_r, \rho)] - \mathcal{W}(\mathbf{C}^*, \boldsymbol{\alpha}^*, \rho) \\
=& \underbrace{\mathbb{E}_{S_n}[\mathcal{W}(\hat{\mathbf{C}}_{n,r}, \hat{\boldsymbol{\alpha}}_r, \rho) - \mathcal{W}_n(\hat{\mathbf{C}}_{n,r}, \hat{\boldsymbol{\alpha}}_r, \rho_n)]}_{\mathcal{A}} + \underbrace{\mathbb{E}_{S_n}[\mathcal{W}_n(\hat{\mathbf{C}}_{n,r}, \hat{\boldsymbol{\alpha}}_r, \rho_n) - \mathcal{W}_n(\hat{\mathbf{C}}_n, \hat{\boldsymbol{\alpha}}_r, \rho_n)]}_{\mathcal{B}} \\
&+ \underbrace{\mathbb{E}_{S_n}[\mathcal{W}_n(\hat{\mathbf{C}}_n, \hat{\boldsymbol{\alpha}}_r, \rho_n) - \mathcal{W}_n(\hat{\mathbf{C}}, \hat{\boldsymbol{\alpha}}, \rho_n)]}_{\mathcal{C}} + \underbrace{\mathbb{E}_{S_n}[\mathcal{W}_n(\hat{\mathbf{C}}, \hat{\boldsymbol{\alpha}}, \rho_n)] - \mathcal{W}(\mathbf{C}^*, \boldsymbol{\alpha}^*, \rho)}_{\mathcal{D}}.
\end{aligned}
$$

According to (Biau et al., 2008), $\mathcal{A}$ can be bounded by $\widetilde{\mathcal{O}}(k/\sqrt{n})$. By Theorem 5.1, $\mathcal{D}$ has an upper bound as $\widetilde{\mathcal{O}}(k/\sqrt{n})$.

Moreover, by Theorem 1 in (Calandriello & Rosasco, 2018), when $r = \Omega(n/\gamma)$, $\mathcal{B}$ can be bounded by $\widetilde{\mathcal{O}}(k\gamma/n)$, where γ is a positive parameter.

Similar to the proof of Theorem 5.1, combining Theorem 4.3, we can bound $\mathcal{C}$ as

$$
\begin{aligned}
\mathcal{C} &= \mathbb{E}_{S_n}[\mathcal{W}_n(\hat{\mathbf{C}}_{\hat{\boldsymbol{\alpha}}_r}, \hat{\boldsymbol{\alpha}}_r, \rho_n) - \mathcal{W}_n(\hat{\mathbf{C}}, \hat{\boldsymbol{\alpha}}, \rho_n)] \\
&\leq 2\|\hat{\boldsymbol{\alpha}}_r - \hat{\boldsymbol{\alpha}}\|_\infty \\
&\lesssim k\sqrt{\left(\frac{1}{r} - \frac{1}{n}\right)\log(k/\delta)}
\end{aligned}
$$

Above all, we can conclude that $\mathbb{E}_{S_n}[\mathcal{W}(\hat{\mathbf{C}}_{n,r}, \hat{\boldsymbol{\alpha}}_r, \rho)] - \mathcal{W}(\mathbf{C}^*, \boldsymbol{\alpha}^*, \rho)$ can be bounded by

$$
\widetilde{\mathcal{O}}\left(\frac{K\gamma}{n} + \frac{K}{\sqrt{r}} + \frac{K}{\sqrt{n}}\right).
$$

Letting $\gamma = \Theta(n/r)$, the desirable result follows.

□

B.7. The Proof of Theorem 6.2

We first bound the difference of T_n and T_r by the following theorem.

Theorem B.6. *When $r \geq \frac{\log(2/\delta)}{\varepsilon - \log(1+\varepsilon)}$, the following holds with probability $1-\delta$:*

$$\|T_n - T_r\| \leq \varepsilon.$$

We introduce the following lemma and its corollary to prove Theorem B.6.

Lemma B.7. *(Yin et al., 2020a) Suppose that $G \sim \mathcal{N}(0, n)$, and a random matrix $\mathbf{T} = [\mathbf{t}_1, \cdots, \mathbf{t}_m] \in \mathbb{R}^{n\times m}$. The j-th column $\mathbf{t}_j$ of $\mathbf{T}$ has only a non-zero element with probability*

$$\Pr\left(t_{ij} = \delta(i)n\right) = \frac{1}{n},\ \Pr(t_{ij} = 0) = 1 - \frac{1}{n},$$

where $\delta(i) = 1$ or -1 with equal probability. Then for any vector $\mathbf{b} \in \mathbb{R}^n$ and non-negative integer S_n, the following inequality holds

$$\mathbb{E}\left[(\mathbf{t}_j^\top \mathbf{b})^{2s}\right] \leq \mathbb{E}\left[G^{2s}\right].$$

Corollary B.8. *Let $T \sim \mathcal{N}(0, 1/r)$, and a random matrix $\mathbf{S} = [\mathbf{s}_1, \cdots, \mathbf{s}_r] \in \mathbb{R}^{n\times r}$. Each column of $\mathbf{S}$ has only a non-zero element with probability*

$$\Pr\left(s_{ij} = \delta(i)\sqrt{\frac{n}{r}}\right) = \frac{1}{n},\ \Pr(s_{ij} = 0) = 1 - \frac{1}{n},$$

where $\delta(i) = 1$ or -1 with equal probability. Then for any vector $\mathbf{b} \in \mathbb{R}^n$ and non-negative integer S_n, the following inequality holds

$$\mathbb{E}\left[(\mathbf{s}_j^\top \mathbf{b})^{2s}\right] \leq \mathbb{E}\left[T^{2s}\right].$$

The proof of Corollary B.8 is as follows.

Proof. By construction, we know $t_{ij} = s_{ij}\sqrt{mn}$. According to Lemma B.7, we have

$$\mathbb{E}\left[(\mathbf{s}_j^\top \mathbf{b})^{2s}\right] \leq \mathbb{E}\left[\left(\frac{G}{\sqrt{mn}}\right)^{2s}\right].$$

It is obvious that $\frac{G}{\sqrt{mn}} \sim \mathcal{N}(0, 1/m)$, thus the corollary holds.

□

Now, we prove Theorem B.6. The technique is similar to (Yin et al., 2022a), but we give the detailed process of proof for completeness.

Proof. For ease of proof, we rewrite T_n and T_r as matrix forms. Let $\mathbf{\Phi}_n = [\phi_1, \cdots, \phi_n] \in \mathbb{R}^{d\times n}$, where d denotes the dimension of the Hilbert space corresponding to the kernel function. Then, $T_n = \frac{1}{n}\mathbf{\Phi}_n\mathbf{\Phi}_n^\top$. Similarly, we can rewrite T_r as $T_r = \frac{1}{r}\mathbf{\Phi}_r\mathbf{\Phi}_r^\top$. By the notations of Corollary B.8, we have

$$\begin{aligned}
\|T_n - T_r\| &= \left\|\frac{1}{n}\mathbf{\Phi}_n\mathbf{\Phi}_n^\top - \frac{1}{r}\mathbf{\Phi}_r\mathbf{\Phi}_r^\top\right\| \\
&= \left\|\frac{1}{n}\mathbf{\Phi}_n\mathbf{\Phi}_n^\top - \frac{1}{n}\mathbf{\Phi}_n\mathbf{S}\mathbf{S}^\top\mathbf{\Phi}_n^\top\right\| \\
&= \max_{\|\mathbf{x}\|=1}\left|\frac{1}{n}\mathbf{x}^\top\mathbf{\Phi}_n\mathbf{\Phi}_n^\top\mathbf{x} - \frac{1}{n}\mathbf{x}^\top\mathbf{\Phi}_n\mathbf{S}\mathbf{S}^\top\mathbf{\Phi}_n^\top\mathbf{x}\right| \\
&= \max_{\|\mathbf{x}\|=1}\left|\left\|\frac{1}{\sqrt{n}}\mathbf{\Phi}_n^\top\mathbf{x}\right\|^2 - \left\|\frac{1}{\sqrt{n}}\mathbf{S}^\top\mathbf{\Phi}_n^\top\mathbf{x}\right\|^2\right|
\end{aligned}$$

To bound the above formula, for any $\|\mathbf{x}\| = 1$, we first give the bound of

$$\Pr\left(\frac{\left\|\frac{1}{\sqrt{n}}\mathbf{S}^\top\mathbf{\Phi}_n^\top\mathbf{x}\right\|^2 - \left\|\frac{1}{\sqrt{n}}\mathbf{\Phi}_n^\top\mathbf{x}\right\|^2}{\left\|\frac{1}{\sqrt{n}}\mathbf{\Phi}_n^\top\mathbf{x}\right\|^2} \geq \varepsilon\right) = \Pr\left(\frac{\left\|\frac{1}{\sqrt{n}}\mathbf{S}^\top\mathbf{\Phi}_n^\top\mathbf{x}\right\|^2}{\left\|\frac{1}{\sqrt{n}}\mathbf{\Phi}_n^\top\mathbf{x}\right\|^2} \geq 1+\varepsilon\right).$$

Setting $\mathbf{b} = \frac{1}{\sqrt{n}}\mathbf{\Phi}_n^\top\mathbf{x}/\left\|\frac{1}{\sqrt{n}}\mathbf{\Phi}_n^\top\mathbf{x}\right\|$, for any $\lambda < m/2$, the following inequality holds according to Markov's inequality.

$$\begin{aligned}
\Pr(\|\mathbf{S}^\top\mathbf{b}\| \geq 1+\varepsilon) &= \Pr(\exp(\lambda\|\mathbf{S}^\top\mathbf{b}\|^2) \geq \exp(\lambda(1+\varepsilon))) \\
&\leq \mathbb{E}\left[\exp(\lambda\|\mathbf{S}^\top\mathbf{b}\|^2)\right]\exp(-\lambda(1+\varepsilon)) \\
&= \mathbb{E}\left[\exp(\lambda\|\mathbf{S}^\top\mathbf{b}\|^2)\right]\exp(-\lambda(1+\varepsilon)) \\
&= \mathbb{E}\left[\exp(\lambda\sum_{j=1}^{m}(\mathbf{s}_j^\top\mathbf{b})^2)\right]\exp(-\lambda(1+\varepsilon)) \\
&= \exp(-\lambda(1+\varepsilon))\prod_{j=1}^{m}\mathbb{E}\left[\exp(\lambda(\mathbf{s}_j^\top\mathbf{b})^2)\right].
\end{aligned} \tag{11}$$

Furthermore, by Taylor's formula and Corollary B.8, we have

$$\begin{aligned}
\mathbb{E}\left[\exp(\lambda T^2)\right] &= \sum_{s=0}^{+\infty}\frac{\lambda^s}{s!}\mathbb{E}[T^{2s}] \\
&\geq \sum_{s=0}^{\infty}\frac{\lambda^s}{s!}\mathbb{E}[(\mathbf{s}_j^\top\mathbf{b})^{2s}] \\
&= \mathbb{E}\left[\exp(\lambda(\mathbf{s}_j^\top\mathbf{b})^2)\right].
\end{aligned}$$

In addition, we have

$$\begin{aligned}
\mathbb{E}\left[\exp(\lambda T^2)\right] &= \int_{-\infty}^{+\infty}\sqrt{\frac{m}{2\pi}}\exp(-mx^2/2)\exp(\lambda x^2)dx \\
&= \int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\exp(-t^2/2)\exp\left(\lambda t^2/m\right)dt \\
&= \frac{1}{\sqrt{1-2\lambda/m}}.
\end{aligned}$$

Thus, $\mathbb{E}\left[\exp(\lambda(\mathbf{s}_j^\top\mathbf{b})^2)\right] \leq \frac{1}{\sqrt{1-2\lambda/m}}$. Combining Eq. (11), we have

$$\Pr(\|\mathbf{S}^\top\mathbf{b}\| \geq 1+\varepsilon) \leq \left(\frac{1}{\sqrt{1-2\lambda/m}}\right)^m\exp(-\lambda(1+\varepsilon)).$$

Setting $\lambda = \frac{\varepsilon}{1+\varepsilon}\cdot\frac{m}{2}$, we can obtain

$$\Pr(\|\mathbf{S}^\top\mathbf{b}\| \geq 1+\varepsilon) \leq (1+\varepsilon)^{m/2}\exp\left(\frac{-m\varepsilon}{2}\right). \tag{12}$$

On the other hand, for any $\|\mathbf{x}\| = 1$, we turn to bound

$$\Pr\left(\frac{\left\|\frac{1}{\sqrt{n}}\mathbf{S}^\top\mathbf{\Phi}_n^\top\mathbf{x}\right\|^2 - \left\|\frac{1}{\sqrt{n}}\mathbf{\Phi}_n^\top\mathbf{x}\right\|^2}{\left\|\frac{1}{\sqrt{n}}\mathbf{\Phi}_n^\top\mathbf{x}\right\|^2} \leq -\varepsilon\right) = \Pr\left(\frac{\left\|\frac{1}{\sqrt{n}}\mathbf{S}^\top\mathbf{\Phi}_n^\top\mathbf{x}\right\|^2}{\left\|\frac{1}{\sqrt{n}}\mathbf{\Phi}_n^\top\mathbf{x}\right\|^2} \leq 1-\varepsilon\right).$$

Similar to the above procedures, for any $\lambda < m/2$, we have

$$\begin{aligned}
\Pr(\|\mathbf{S}^\top \mathbf{b}\| \leq 1-\varepsilon) &= \Pr(\exp(\lambda(1-\varepsilon)) \geq \exp(\lambda\|\mathbf{S}^\top \mathbf{b}\|^2)) \\
&\leq \exp(\lambda(1-\varepsilon))\, \mathbb{E}\left[\exp(-\lambda\|\mathbf{S}^\top \mathbf{b}\|^2)\right] \\
&= \exp(\lambda(1-\varepsilon))\, \mathbb{E}\left[\exp(-\lambda\|\mathbf{S}^\top \mathbf{b}\|^2)\right] \\
&= \exp(\lambda(1-\varepsilon))\, \mathbb{E}\left[\exp(-\lambda\sum_{j=1}^{m}(\mathbf{s}_j^\top \mathbf{b})^2)\right] \\
&= \exp(\lambda(1-\varepsilon)) \prod_{j=1}^{m} \mathbb{E}\left[\exp(-\lambda(\mathbf{s}_j^\top \mathbf{b})^2)\right] \\
&\leq \exp(\lambda(1-\varepsilon)) \prod_{j=1}^{m} \mathbb{E}\left[1-\lambda(\mathbf{s}_j^\top \mathbf{b})^2 + \frac{\lambda^2(\mathbf{s}_j^\top \mathbf{b})^4}{2}\right].
\end{aligned} \tag{13}$$

Moreover, it can be checked that

$$\begin{aligned}
\mathbb{E}\left[(\mathbf{s}_j^\top \mathbf{b})^2\right] &= \sum_{i=1}^{n}\sum_{t=1}^{n} \mathbb{E}\left[(s_{ij}b_i)(s_{tj}b_t)\right] \\
&= \sum_{i=1}^{n} \mathbb{E}\left[s_{ij}^2 b_i^2\right] \\
&= \sum_{i=1}^{n} \left(\frac{1}{n}\left(\sqrt{\frac{n}{m}}\right)^2 b_i^2\right) \\
&= \frac{1}{m},
\end{aligned} \tag{14}$$

and

$$\mathbb{E}\left[(\mathbf{s}_j^\top \mathbf{b})^4\right] \leq \mathbb{E}\left[T^4\right] = \frac{3}{m^2}. \tag{15}$$

Substituting Eq. (14) and Eq. (15) into Eq. (13), we have

$$\Pr(\|\mathbf{S}^\top \mathbf{b}\| \leq 1-\varepsilon) \leq \exp(\lambda(1-\varepsilon)) \left(1 - \frac{\lambda}{m} + \frac{3\lambda^2}{2m^2}\right)^m. \tag{16}$$

Setting $\lambda = \frac{\varepsilon}{1+\varepsilon} \cdot \frac{m}{2}$ and $w = \frac{\varepsilon}{1+\varepsilon}$, we have

$$\begin{aligned}
\Pr(\|\mathbf{S}^\top \mathbf{b}\| \leq 1-\varepsilon) &\leq \exp\left(\frac{\varepsilon(1-\varepsilon)}{1+\varepsilon} \cdot \frac{m}{2}\right) \left(1 - \frac{\varepsilon}{2(1+\varepsilon)} + \frac{3\varepsilon^2}{8(1+\varepsilon)^2}\right)^m \\
&\leq \exp\left(-\frac{m\varepsilon}{2}\right)(1+\varepsilon)^{m/2}.
\end{aligned} \tag{17}$$

Combining Eq. (12) and Eq. (17), we can obtain

$$\Pr\left(\left|\frac{\left\|\frac{1}{\sqrt{n}}\mathbf{S}^\top \mathbf{\Phi}_n^\top \mathbf{x}\right\|^2 - \left\|\frac{1}{\sqrt{n}}\mathbf{\Phi}_n^\top \mathbf{x}\right\|^2}{\left\|\frac{1}{\sqrt{n}}\mathbf{\Phi}_n^\top \mathbf{x}\right\|^2}\right| \geq \varepsilon\right) \leq 2\exp\left(\frac{m\log(1+\epsilon) - m\varepsilon}{2}\right). \tag{18}$$

Above all, it can be deduced that if $m \geq \frac{\log(2/\delta)}{\varepsilon - \log(1+\varepsilon)}$,

$$\left|\left\|\frac{1}{\sqrt{n}}\mathbf{S}^\top \mathbf{\Phi}_n^\top \mathbf{x}\right\|^2 - \left\|\frac{1}{\sqrt{n}}\mathbf{\Phi}_n^\top \mathbf{x}\right\|^2\right| \leq \varepsilon \left\|\frac{1}{\sqrt{n}}\mathbf{\Phi}_n^\top \mathbf{x}\right\|^2 \leq \varepsilon$$

holds with probability at least $1-\delta$.

Because of the arbitrariness of $\mathbf{x}$, we have

$$\|T_n - T_r\| \leq \varepsilon.$$

□

Finally, we complete the proof of Theorem 6.2.

Proof. We can use Theorem B.6 to improve the result of Lemma A.3. In the proof of Lemma A.3, Term $\mathcal{C}$ can be bounded by $\mathcal{O}(\|T_n - T_r\|)$, and Term $\mathcal{D}$ in Lemma A.3 also has the same upper bound. This is because

$$\begin{aligned}
\mathcal{D} \leq & 2\left\|\frac{1}{n}\sum_{i=1}^{n}\hat{h}_{r,j}(x_i)\phi(x_i) - \frac{1}{r}\sum_{i=1}^{r}\hat{h}_{r,j}(x_i)\phi(x_i)\right\| \\
= & 2\left\|(T_n - T_r)\hat{h}_{r,j}\right\| \\
\leq & 2\left\|\hat{h}_{r,j}\right\| \cdot \|T_n - T_r\| \\
\precsim & \|T_n - T_r\|.
\end{aligned}$$

By Theorem B.6, when $r \geq \frac{\log(2/\delta)}{\varepsilon - \log(1+\varepsilon)}$,

$$\left|\hat{\epsilon}(K_p, \{\hat{h}_{n,j}\}_{j=1}^{k}) - \hat{\epsilon}(K_p, \{\hat{h}_{r,j}\}_{j=1}^{k})\right| \precsim k\|T_n - T_r\| \leq k\varepsilon.$$

By the same decomposition of the proof of Theorem 6.1, we can improve the bound of $\mathcal{C}$ in Theorem 6.1 as

$$\mathcal{C} \precsim \|\hat{\boldsymbol{\alpha}}_r - \hat{\boldsymbol{\alpha}}\|_\infty \precsim k\|T_n - T_r\| \leq k\varepsilon.$$

Combining all things together, we can conclude the conclusion.

□

C. Notation Table

Table 4. Basic notations.

Notations	Meaning
k	Clustering number
n	Sample number
$\mathcal{X}$	Sample space
$K_p(\cdot,\cdot)$	The p-th base kernel function
$\mathbf{K}_p$	The p-th base kernel matrix
ρ	Real distribution
ρ_n	Empirical distribution
$\mathbf{H}$	Clustering indicator matrix
$\{h_j(\cdot)\}_{j=1}^k$	Clustering indicator functions
$\{\hat{h}_j(\cdot)\}_{j=1}^k$	Approximated clustering indicator functions
L_K	The integral operator associated with kernel function K.

D. URLs of the Used Datasets

In this section, we give a detailed introduction to the datasets used in our experiments. The URLs of the datasets in Table 1 are as follows:

1. Flo17: http://www.robots.ox.ac.uk/œvgg/data/flowers/17/,
2. Flo102: http://www.robots.ox.ac.uk/œvgg/data/flowers/102/,
3. DIGIT: http://ss.sysu.edu.cn/py/,
4. PFold: http://mkl.ucsd.edu/dataset/protein-fold-prediction,
5. CCV: http://www.ee.columbia.edu/ln/dvmm/CCV/,
6. Reuters: http://kdd.ics.uci.edu/databases/reuters21578/.

All the kernel matrices are pre-computed and available on websites. The large-scale datasets in Table 2 can be downloaded from

1. NUSWIDE: http://lms.comp.nus.edu.sg/wp-content/uploads/2019/research/nuswide/NUS-WIDE.html,
2. AwA: http://cvml.ist.ac.at/AwA/,
3. CIFAR10: http://www.cs.toronto.edu/~kriz/cifar.html,
4. YtVideo: http://archive.ics.uci.edu/ml/datasets/YouTube+Multiview+Video+Games+Dataset,
5. Winnipeg: https://archive.ics.uci.edu/ml/datasets/Crop+mapping+using+fused+optical-radar+data+set,
6. Covertype: http://archive.ics.uci.edu/ml/datasets/Covertype.