Scalable Multiple Kernel Clustering: Learning Clustering Structure from Expectation

Weixuan Liang [1] **En Zhu** [1] **Shengju Yu** [1] **Huiying Xu** [2] **Xinzhong Zhu** [2] **Xinwang Liu** [1]

Abstract

In this paper, we derive an upper bound of the difference between a kernel matrix and its expectation under a mild assumption. Specifically, we assume that the true distribution of the training data is an unknown isotropic Gaussian distribution. When the kernel function is a Gaussian kernel, and the mean of each cluster is sufficiently separated, we find that the expectation of a kernel matrix can be close to a rank-k matrix, where k is the cluster number. Moreover, we prove that the normalized kernel matrix of the training set deviates (w.r.t. Frobenius norm) from its expectation in the order of $\widetilde{\mathcal{O}}(1/\sqrt{d})$, where d is the dimension of samples. Based on the above theoretical results, we propose a novel multiple kernel clustering framework which attempts to learn the information of the expectation kernel matrices. First, we aim to minimize the distance between each base kernel and a rank-k matrix, which is a proxy of the expectation kernel. Then, we fuse these rank-k matrices into a consensus rank-k matrix to find the clustering structure. Using an anchor-based method, the proposed framework is flexible with the sizes of input kernel matrices and able to handle large-scale datasets. We also provide the approximation guarantee by deriving two non-asymptotic bounds for the consensus kernel and clustering indicator matrices. Finally, we conduct extensive experiments to verify the clustering performance of the proposed method and the correctness of the proposed theoretical results.

[1]College of Computer, National University of Defense Technology, Changsha, China [2]School of Computer Science and Technology, Zhejiang Normal University, Jinhua, China. Correspondence to: Huiying Xu, Xinzhong Zhu, Xinwang Liu <xhy@zjnu.edu.cn,zxz@zjnu.edu.cn,xinwangliu@nudt.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

A fundamental principle of learned models is to study the parameters of the true distribution of sample space from a set of training samples. In the learning task, we should explore more general information about the distribution of samples with its empirical distribution. In kernel clustering (Dhillon et al., 2004), the empirical kernel matrix is essential to improve clustering performance. It is necessary to know how the empirical kernel matrix deviates from its expectation. In this paper, for a multi-view training set, we assume that samples of each view are generated from an unknown isotropic Gaussian distribution. If we use the Gaussian kernel function to compute the kernel matrix, then we can derive an upper bound of the difference between the empirical kernel matrix and its expectation. Assuming that the dimension of the samples is d, the above upper bound is basically $\widetilde{\mathcal{O}}(1/\sqrt{d})$[1]. When the means of the clusters are well separated, the expectation of the kernel matrix can be close to a rank-k matrix.

Based on the above observations, we attempt to devise a novel multiple kernel clustering (MKC) algorithm. MKC (Zhao et al., 2009) is proposed to improve the clustering performance of the single kernel clustering (Dhillon et al., 2004). It constructs several base kernel matrices $\{\mathbf{K}_v\}_{v=1}^V$ with different kernel functions and fuses $\mathbf{K}_v$'s into a consensus one. Then, the standard kernel k-means (Dhillon et al., 2004) is performed on the consensus kernel matrix for the final clustering result. MKC has been extensively studied in recent years (Liang et al., 2023; 2022; Liu, 2023; 2022; Ren & Sun, 2020; Liu et al., 2017; 2016; Li et al., 2016). Among them, Liu (2022) proposes a new fusion style by a min-max optimization objective and improves clustering performance without any hyper-parameter tuning. Liu (2023) further improves the work of (Liu, 2022). Subsequently, Liang et al. (2022) study the stability and generalization of MKC and derive the excess risk bound of MKC for the first time. Liang et al. (2023) establish the strong consistency of MKC by proving the empirical kernel weights can converge to the corresponding expected version.

Although the research mentioned enriches the fields of

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

MKC, the design of MKC algorithms is almost entirely based on heuristic observations. This weakens the explanation of MKC. In this paper, we propose a scalable multiple kernel clustering (SMKC) based on the concentration results of kernel matrices for stronger rationality. By the low-rank property of the expectation kernel, we use a rank-k matrix as its proxy. The proposed objective function aims to minimize the distance between each base kernel and a rank-k matrix. Meanwhile, these rank-k matrices are then fused into a rank-k consensus matrix. The proposed SMKC has two noticeable advantages: free of tuning hyper-parameters and scalable to large-scale datasets. Existing MKC algorithms that have hyper-parameters need true labels for hyper-parameter tuning. In clustering tasks, the lack of true labels makes it impossible for these methods to handle real-world datasets. Our method doesn't need to tune hyper-parameters and can be easily implemented. In addition, the proposed SMKC is flexible to the size of input base kernel matrices. Inspired by some anchor-based large-scale multi-view clustering methods (Yu et al., 2023; 2024a), we can sample s ($s \ll n$) columns from each base kernel matrix of size $n \times n$. When the input kernels are several $n \times s$ matrices $\mathbf{G}_v$'s, the complexity of the corresponding optimization algorithm is basically $\mathcal{O}(ns^2)$. Since the complexity is linear with n, SMKC can handle large-scale datasets. Assume that the consensus kernel learned from $\mathbf{G}_v$'s is $\mathbf{G}^*$. When $s = n$, we denote the output consensus kernel as $\mathbf{K}^*$. For an approximation guarantee, we establish a non-asymptotic bound to depict the difference between $\mathbf{G}^*$ and $\mathbf{K}^*$. We also derive an upper bound of the distance between the clustering indicator matrices obtained from the above two consensus kernel matrices.

Subsequently, we conduct comprehensive experiments to test the clustering performance in 11 benchmark datasets including 6 middle-scale and 5 large-scale datasets. The experimental results show the superiority of the proposed SMKC in both clustering performance and execution efficiency. Moreover, we also perform an experiment for the verification of the theoretical approximation guarantee. The contributions of this paper can be summarized as follows.

1. By assuming the samples are generated from isotropic Gaussian distributions, we obtain the expectation of a kernel matrix computed by the Gaussian kernel function. We then theoretically study how far the empirical kernel matrix is from its expectation.

2. We find the expectation of a kernel matrix has an apparent clustering structure. We use this phenomenon to devise a novel MKC algorithm with theoretical approximation guarantees. Our method is free of tuning hyper-parameters and able to handle large-scale datasets.

3. Extensive experiments on real-world datasets demonstrate the advantages of the proposed SMKC and the correctness of our theoretical results.

The paper is organized as follows. Section 2 gives an introduction to the related work of our work. Section 3 elaborates the notations and assumptions throughout this paper. Section 4 gives two theoretical results about the expectation of an empirical kernel matrix and explains the motivation of the proposed SMKC. The objective function of SMKC and its optimization method is also placed in Section 4. We give the theoretical analysis in Section 5. Section 6 records the results of numerical experiments. Section 7 concludes this paper.

2. Related Work

In this section, we introduce two related works, i.e., the assumption of Gaussian distribution in clustering and multiple kernel clustering.

2.1. Gaussian Distributed Datasets in Clustering

In the existing literature on clustering, a common assumption is that the training datasets obey Gaussian distribution (Shi et al., 2009; Löffler et al., 2021; Ding & Ma, 2023; Srivastava et al., 2023). Specifically, for each point $\mathbf{x}$, a data-generative model assumes that

$$\mathbf{x} = \mathbf{y} + \boldsymbol{\epsilon},$$

where $\mathbf{y}$ is the signal of $\mathbf{x}$, and $\boldsymbol{\epsilon}$ is the noise of $\mathbf{x}$. Usually, it is typically assumed that $\boldsymbol{\epsilon}$ obeys Gaussian distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the covariance matrix. Based on this assumption, the researchers design various clustering algorithms and derive the corresponding theoretical results (Yang et al., 2012; Abbe et al., 2022; Han et al., 2023). However, in the multi-view setting, this assumption is never discussed. If we let the data-generative model of each view be Gaussian distributed, can we design a multi-view clustering algorithm to explore the clustering structure sufficiently from all base views? This paper tries to answer the above question by proposing a novel clustering algorithm.

2.2. Multiple Kernel Clustering

Multiple kernel clustering (MKC) can be used to handle multi-view clustering tasks (Yu et al., 2023; 2024b). MKC algorithms construct a kernel matrix for each view and then fuse them on different principles. Existing MKC methods give a weight for each kernel matrix and try to optimize the weight by a unified objective. We will briefly introduce two MKC methods as follows. Assuming that the $\boldsymbol{\gamma} \in \mathbb{R}^V$ are the weights and $\{\mathbf{K}_v\}_{v=1}^V$, multiple kernel k-means (Huang et al., 2012) (MKKM) optimizes the following objective

function:

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

where $\mathbf{K}_{\boldsymbol{\gamma}}=\sum_{v=1}^{V}\gamma_v^2\mathbf{K}_v$, $\mathbf{H}\in\mathbb{R}^{n\times k}$, and Δ is the constraint of simplex. Liu (2022) modifies the above method in a min-max manner:

$$\min_{\boldsymbol{\gamma}\in\Delta}\max_{\mathbf{H}^\top\mathbf{H}=\mathbf{I}_k} \frac{1}{n}\mathrm{Tr}(\mathbf{K}_{\boldsymbol{\gamma}}\mathbf{H}\mathbf{H}^\top).$$

The work in (Liu, 2022) improves MKKM and becomes one of the most popular MKC algorithms. Despite the success of these MKC methods, they lack the exploration of the statistical properties of kernel matrices. We will fill this gap by learning the expectation of kernel matrices.

3. Notations and Assumptions

In this section, we introduce the main notations and general assumptions.

Mathematical notations. This section presents the mathematical notations utilized throughout the paper to enhance readability. For the asymptotic notations $\mathcal{O},\Theta$, we refer to Chapter 3 of (Cormen et al., 2022). The notation $u(n)=\mathcal{O}(v(n))$ implies $u(n)\leq cv(n)$ for some constant c, also denoted as $u(n)\precsim v(n)$. $\tilde{\mathcal{O}}(\cdot)$ is similar to $\mathcal{O}$ but hides logarithmic terms. $u(n)=\Theta(v(n))$ means $u(n)\precsim v(n)\precsim u(n)$. The operator norm of a matrix or operator $\mathbf{A}$ is defined as $\|\mathbf{A}\|:=\max_{\|\mathbf{x}\|=1}\|\mathbf{A}\mathbf{x}\|$.

In the multi-view setting, there are V views $\{\mathbf{X}^{(v)}\}_{v=1}^{V}$ in a training set. The v-th view is denoted by $\mathbf{X}^{(v)}=\{\mathbf{x}_i^{(v)}\}_{i=1}^{n}$, where $\mathbf{x}_i^{(v)}\in\mathbb{R}^{d^{(v)}}$ and $d^{(v)}$ denotes the dimension of the v-th view. As the following assumption, we suppose all the samples are composed of signal and noise.

Assumption 3.1. For i-th point of the v-th view, if it belongs to the p-th cluster, we assume that

$$\mathbf{x}_i^{(v)}=\boldsymbol{\mu}_p^{(v)}+\boldsymbol{\epsilon}_i^{(v)},$$

where $\boldsymbol{\mu}_p^{(v)}$ is the mean of the p-th cluster in the v-th view, and $\boldsymbol{\epsilon}_i^{(v)}\sim\mathcal{N}(\mathbf{0}_{d^{(v)}},(\sigma_p^{(v)}/\sqrt{d^{(v)}})\mathbf{I}_{d^{(v)}})$ is the isotropic Gaussian noise.

In the above assumption, because $\mathbb{E}\|\boldsymbol{\epsilon}_i^{(v)}\|$ is $\Theta(\sqrt{d^{(v)}})$, we normalize the scale of covariance $\sigma_p^{(v)}$ by dividing $\sqrt{d^{(v)}}$.

4. Motivation and Method

In this section, we begin with a theoretical analysis of the expectation kernel matrix, followed by the presentation of our method and its associated optimization algorithm.

4.1. Theoretical Analysis of the Expectation Kernel Matrix

We can compute the expectation of the kernel matrix as follows with our assumptions.

Theorem 4.1. *Under Assumption 3.1, if the kernel matrix is computed by Gaussian kernel function*

$$K(\mathbf{x}_i,\mathbf{x}_j)=\exp\left(-\|\mathbf{x}_i-\mathbf{x}_j\|^2/(2(\delta^{(v)})^2)\right),$$

the v-th expectation of the kernel matrix is represented by

$$[\mathbb{E}(\mathbf{K}_v)]_{ij}=\left(\frac{\delta^{(v)}}{\sqrt{(\delta^{(v)})^2+(\sigma_p^{(v)})^2/d^{(v)}+(\sigma_q^{(v)})^2/d^{(v)}}}\right)^{d^{(v)}}\cdot \exp\left(-\frac{\|\boldsymbol{\mu}_p^{(v)}-\boldsymbol{\mu}_q^{(v)}\|^2}{2(\delta^{(v)})^2+2(\sigma_p^{(v)}/)^2/d^{(v)}+2(\sigma_q^{(v)})^2/d^{(v)}}\right), \tag{1}$$

for the i-th and j-th samples belonged to the p-th and q-th cluster.

Remark. It is easy to see that $\mathbb{E}(\mathbf{K}_v)$ is also a kernel matrix. For two different clusters p and q, if their means $\boldsymbol{\mu}_p^{(v)}$ and $\boldsymbol{\mu}_q^{(v)}$ are well separated, then $\mathbb{E}(\mathbf{K}_v)$ can approach to a low rank block matrix with rank k. In this sense, $\mathbb{E}(\mathbf{K}_v)$ has an explicit clustering structure.

Denote that the normalized version of $\mathbf{K}_v$ as $\mathbf{K}_v/n$. The next theorem shows that the normalized kernel matrix can converge to its expectation as $d^{(v)}\to\infty$.

Theorem 4.2. *The following inequality holds with probability at least* $1-\exp(-\xi^2)$,

$$\left\|\frac{1}{n}\mathbf{K}_v-\frac{1}{n}\mathbb{E}\,\mathbf{K}_v\right\|_{\mathrm{F}}\precsim\frac{\sigma_\infty^{(v)}}{\delta^{(v)}\sqrt{d^{(v)}}}(1+\xi/\sqrt{n}),$$

where $\sigma_\infty^{(v)}=\max\{\sigma_1^{(v)},\cdots,\sigma_k^{(v)}\}$.

Remark. Theorem 4.2 means that the empirical kernel matrix is near its expectation. Meanwhile, according to Theorem 4.1, the rank of the expectation kernel matrix is close to k. The proposed method aims to learn a rank-k matrix from each base kernel matrix and fuse them into a unified one for clustering.

4.2. Proposed Method

According to Theorem 4.1, we know that the expectation, denoted as $\mathbb{E}\mathbf{K}_v$, exhibits a noiseless quality and manifests a distinct clustering structure. However, direct acquisition of $\mathbb{E}\mathbf{K}_v$ is unattainable. As stated in Theorem 4.2, the empirical kernel matrix closely approximates its expectation. Our objective is to extract information regarding $\mathbb{E}\mathbf{K}_v$ from $\mathbf{K}_v$. Utilizing the representation theorem, we express $K(\mathbf{x},\mathbf{y})$

for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ as an inner product $\langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$ within a Hilbert space, where $\phi(\mathbf{x})$ is defined as the feature map for $\mathbf{x}$. Consequently, we can represent $\mathbf{K}_v$ by $\mathbf{\Phi}\mathbf{\Phi}^\top$, where the rows of $\mathbf{\Phi}$ correspond to the feature maps of the training set. Initially, we employ kernel principal component analysis (KPCA) to denoise $\mathbf{K}_v$, as follows:

$$\min_{\mathbf{U}^\top \mathbf{U} = \mathbf{I}_k} \|\mathbf{\Phi} - \mathbf{U}\mathbf{U}^\top \mathbf{\Phi}\|_{\mathrm{F}}^2. \tag{2}$$

The solution of Eq. (2) is the first k left singular vectors of $\mathbf{\Phi}$. Thus, Eq. (2) is equivalent to

$$\min_{\mathbf{U}^\top \mathbf{U} = \mathbf{I}_k} \|\mathbf{K}_v - \mathbf{U}\mathbf{U}^\top \mathbf{K}_v\|_{\mathrm{F}}^2. \tag{3}$$

According to Section 2 of (Cohen et al., 2015), the problem in Eq. (3) can be reformed as

$$\min_{\widetilde{\mathbf{K}}_v \in \mathcal{M}_k} \|\widetilde{\mathbf{K}}_v - \mathbf{K}_v\|_{\mathrm{F}}^2. \tag{4}$$

From Theorem 4.1, the expectation of each base kernel shows a clear clustering structure. To unify the clustering information of all views, we aim to minimize the distance between $\widetilde{\mathbf{K}}_v$ and a rank-k consensus kernel matrix $\mathbf{K}^*$, i.e., the following optimization problem:

$$\min_{\widetilde{\mathbf{K}}_v, \mathbf{K}^* \in \mathcal{M}_k} \sum_{v=1}^{V} \|\widetilde{\mathbf{K}}_v - \mathbf{K}^*\|_{\mathrm{F}}^2. \tag{5}$$

Combining Eq. (4) and Eq. (5), we can obtain the following objective function:

$$\min_{\widetilde{\mathbf{K}}_v, \mathbf{K}^* \in \mathcal{M}_k} \sum_{v=1}^{V} \|\widetilde{\mathbf{K}}_v - \mathbf{K}_v\|_{\mathrm{F}}^2 + \|\widetilde{\mathbf{K}}_v - \mathbf{K}^*\|_{\mathrm{F}}^2, \tag{6}$$

where $\mathcal{M}_k$ denotes the set of all matrices with rank k.

Obviously, the objective function (6) has a critical issue, i.e., high complexity. It needs to compute V base kernel matrices which occupy $\mathcal{O}(Vn^2)$ space and are certain to cost more time than $\mathcal{O}(Vn^2)$. Inspired by the Nyström based kernel clustering (Wang et al., 2019) and the anchor-based methods (Yu et al., 2024a;b), we can effectively reveal the clustering structure from a small portion of the whole kernel matrix. Specifically, for each view, we can randomly select s ($s \ll n$) anchors $\mathbf{A}^{(v)} = \{\mathbf{a}_t^{(v)}\}_{t=1}^{s}$ from the training set $\mathbf{X}^{(v)}$. Then, we construct V base kernel similarity matrices $\{\mathbf{G}_v\}_{v=1}^{V}$, where $\mathbf{G}_v(i,t) = K_v(\mathbf{x}_i^{(v)}, \mathbf{a}_t^{(v)})$. By replacing $\mathbf{K}_v$ with $\mathbf{G}_v$, we can obtain a scalable multiple kernel clustering method with the following objective function:

$$\min_{\widetilde{\mathbf{G}}_v, \mathbf{G}^* \in \mathcal{M}_k} \sum_{v=1}^{V} \|\widetilde{\mathbf{G}}_v - \mathbf{G}_v\|_{\mathrm{F}}^2 + \|\widetilde{\mathbf{G}}_v - \mathbf{G}^*\|_{\mathrm{F}}^2. \tag{7}$$

4.3. Optimization

We can alternately optimize $\widetilde{\mathbf{G}}_v, \mathbf{G}^*$ in Eq. (7) by the following procedures.

1. Minimizing $\widetilde{\mathbf{G}}_v$ with fixed $\mathbf{G}^*$ and $\mathbf{G}_u$ ($u \neq v$). With fixed $\mathbf{G}^*$, the optimization problem in Eq. (7) can be written as

$$\min_{\widetilde{\mathbf{G}}_v \in \mathcal{M}_k} \|\widetilde{\mathbf{G}}_v - \mathbf{G}_v\|_{\mathrm{F}}^2 + \|\widetilde{\mathbf{G}}_v - \mathbf{G}^*\|_{\mathrm{F}}^2. \tag{8}$$

It is equivalent to the following problem.

$$\min_{\widetilde{\mathbf{G}}_v \in \mathcal{M}_k} \|\widetilde{\mathbf{G}}_v - (\mathbf{G}_v + \mathbf{G}^*)/2\|_{\mathrm{F}}^2. \tag{9}$$

The problem in (9) is a best rank-k approximation problem of $(\mathbf{G}_v + \mathbf{G}^*)/2$. We can solve it by the following steps. First, we perform singular value decomposition on $(\mathbf{G}_v + \mathbf{G}^*)/2$, i.e., $(\mathbf{G}_v + \mathbf{G}^*)/2 = \mathbf{U}\mathbf{D}\mathbf{V}^\top$. Then, we let $\mathbf{G}_v = \mathbf{U}_k \mathbf{D}_k \mathbf{V}_k^\top$, where $\mathbf{U}_k, \mathbf{V}_k$ are respectively the first k columns of $\mathbf{U}, \mathbf{V}$, and $\mathbf{D}_k$ is a diagonal matrix composed of the first k diagonal elements of $\mathbf{D}$.

2. Minimizing $\mathbf{G}^*$ with fixed $\{\widetilde{\mathbf{G}}_v\}_{v=1}^{V}$. With fixed $\{\widetilde{\mathbf{G}}_v\}_{v=1}^{V}$, the optimization problem in Eq. (7) can be written as

$$\min_{\mathbf{G}^* \in \mathcal{M}_k} \sum_{v=1}^{V} \|\widetilde{\mathbf{G}}_v - \mathbf{G}^*\|_{\mathrm{F}}^2. \tag{10}$$

It can be converted to the following problem.

$$\min_{\mathbf{G}^* \in \mathcal{M}_k} \left\| \mathbf{G}^* - \left(\sum_{v=1}^{V} \widetilde{\mathbf{G}}_v \right) / V \right\|_{\mathrm{F}}^2. \tag{11}$$

We can optimize (11) with a similar method as the optimization of (9). The optimization algorithm and the corresponding initialization are listed in Algorithm 1. We provide the convergence analysis in Section A.4 of the appendix.

5. Theoretical Analysis

We will make a comprehensive theoretical analysis of the proposed scalable multiple kernel clustering (SMKC) from the following two perspectives: 1) storage and computational complexity and 2) the degree of approximation between $\mathbf{G}^*$ and $\mathbf{K}^*$ learned by optimizing (7) and (6), respectively.

1. Storage and computational complexity. From Algorithm 1, we need to store V base kernel similarity matrices (their sizes are $n \times s$) which occupy $\mathcal{O}(Vsn)$ space. The space to store other variables is less than $\mathcal{O}(Vsn)$. Thus, the storage complexity of SMKC is $\mathcal{O}(Vsn)$. In the initialization, we perform SVD on V matrices with size $n \times s$. This costs $\mathcal{O}(Vs^2n)$ times. Assume that the proposed SMKC can converge after T iterations. In each iteration, solving the problem in Eq. (11) consumes $\mathcal{O}(sn + s^2n)$ time

Algorithm 1 Scalable Multiple Kernel Clustering

Input: Training set V views $\{\mathbf{X}^{(v)}\}_{v=1}^{V}$; anchor sets $\{\mathbf{A}^{(v)}\}_{v=1}^{V}$ (sampling from $\mathbf{X}^{(v)}$ without replacement); Gaussian kernel functions $\{K_v(\cdot,\cdot)\}_{v=1}^{V}$ with different width $\{\delta^{(v)}\}_{v=1}^{V}$; number of clusters k.
Output: The clustering results.
Compute V base kernel similarity matrices $\{\mathbf{G}_v\}_{v=1}^{V}$ by $\mathbf{G}_v(i,t) = K_v(\mathbf{x}_i^{(v)}, \mathbf{a}_t^{(v)})$, for any $i \in [n], t \in [s]$.
Initialize $\widetilde{\mathbf{G}}_v = \mathbf{U}_k\mathbf{D}_k\mathbf{V}_k^\top$, where $\mathbf{U}_k, \mathbf{D}_k, \mathbf{V}_k$ are obtained by the rank-k truncated SVD of $\mathbf{G}_v$; sign $= 1$.
while sign=1 **do**
  Solve the problem in (11) to optimize $\mathbf{G}^*$.
  Solve the problem in (9) to optimize $\{\mathbf{G}_v\}_{v=1}^{V}$.
  if $\mathbf{G}^*$ converges **then**
    sign $= 0$.
  end if
end while
Perform k-means on the first k left singular vectors $\mathbf{G}^*$ for the final clustering results.

caused by the addition of two $n \times s$ matrices and the subsequent SVD. To optimize Eq. (9), we should add up V matrices with size $n \times s$ and perform SVD on their summation. This step will cost $\mathcal{O}((V-1)sn + s^2n)$ time. Above all, the total computational complexity can be bounded by $\mathcal{O}(TVsn + (V+T)s^2n)$. In the general setting, T, V can be regarded as constants, so the computational complexity is all basically $\mathcal{O}(s^2n)$. If we let the $s \ll n$, the proposed SMKC can efficiently handle large-scale datasets. We will verify its efficiency by experiments in Section 6.

2. The degree of approximation. Now, we prepare to make a theoretical analysis of the approximation degree between $\mathbf{G}^*$ and $\mathbf{K}^*$. Before we give the results, we first introduce an assumption about the eigen gap of kernel matrices.

Assumption 5.1. Assume that δ_k is the difference between the k-th and the $(k+1)$-th eigenvalue of some kernel matrix in the optimization of Algorithm 1. There exists some constant $c \geq 0$ such that $\delta_k \geq c$ holds.

Remark. In Assumption 5.1, we assume that the k-th eigen gap and the first k singular values of the matrices involved in the proposed algorithm are strictly larger than 0. For a base kernel matrix, when the sample number $n \to \infty$, the eigenvalues of the kernel matrix can converge to the eigenvalues of some integral operator. When the kernel function and sample space are fixed, the eigenvalues of the above integral operator can be regarded as constants. As the discussions about the spectral properties of the integral operator defined by Gaussian distribution and Gaussian kernel function (Shi et al., 2008), it can be easy to see that the eigen gap is larger than a constant. Thus, the assumption about the eigen gap is rational. The assumption of the eigengap can also be found in existing literature, such as (Von Luxburg et al., 2008; Mitz & Shkolnisky, 2022).

Theorem 5.2. *Under Assumption 3.1 and Assumption 5.1, when the inputs of Algorithm 1 are V whole base kernel matrices $\{\mathbf{K}_v\}_{v=1}^{V}$ ($\mathbf{K}_v \in \mathbb{R}^{n\times n}$), the output is $\mathbf{K}^* \in \mathbb{R}^{n\times n}$. When the inputs are $\{\mathbf{G}_v\}_{v=1}^{V}$ ($\mathbf{G}_v \in \mathbb{R}^{n\times s}$, and $\mathbf{G}_v$ is composed of the columns sampled from $\mathbf{K}_v$ uniformly without replacement), we assume the output is $\mathbf{G}^*$. In the above two situations, if the iteration numbers of Algorithm 1 are less than T, we have the following approximation bound*

$$\left\| \frac{1}{ns}\mathbf{G}^*(\mathbf{G}^*)^\top - \frac{1}{n^2}\mathbf{K}^*(\mathbf{K}^*)^\top \right\| \precsim (\xi + \sqrt{\log T})\sqrt{\frac{1}{s} - \frac{1}{n}}$$

holds with probability at least $1 - \exp(-\xi^2)$.

Remark. Theorem 5.2 gives an approximation bound between the solutions of Eq. (6) and Eq. (7). The bound is basically $\mathcal{O}(1/\sqrt{s})$. It shows that Eq. (7) can well approximate Eq. (6) with regard to the fusion of base kernels. In the clustering task, the clustering indicator matrix (i.e., the first k left singular vectors $\mathbf{G}^*$) plays an important role. Theorem 5.2 can also derive an upper bound about the clustering indicator matrices obtained from $\mathbf{G}^*$ and $\mathbf{K}^*$.

Corollary 5.3. *Under the same assumption in Theorem 5.2, assume the first k singular vectors of $\mathbf{G}^*$ and $\mathbf{K}^*$ are $\mathbf{U} \in \mathbb{R}^{n\times k}$ and $\mathbf{H} \in \mathbb{R}^{n\times k}$. The distance between the subspace spanned by $\mathbf{U}$ and $\mathbf{H}$ has an upper bound as*

$$\left\|\mathbf{U}\mathbf{U}^\top - \mathbf{H}\mathbf{H}^\top\right\|_{\mathrm{F}} \precsim (\xi + \sqrt{\log T})\sqrt{\frac{1}{s} - \frac{1}{n}}$$

holds with probability at least $1 - \exp(-\xi^2)$.

Remark. It is easy to prove Corollary 5.3 with Theorem 5.2 and Lemma A.7 in the appendix. From Corollary 5.3, we can see that $\mathbf{G}^*$ and $\mathbf{K}^*$ can produce similar clustering performance by performing k-means on their first k left singular vectors. We will empirically verify Corollary 5.3 in the real-world datasets in Section 6.

6. Experiments

The experiments are composed of four parts. 1) We verify the effectiveness of the proposed SMKC with 7 compared methods on 6 real-world datasets. 2) We empirically study the convergence of the objective function. 3) We perform SMKC on large-scale datasets to demonstrate its efficiency. 4) We learn the approximation degree of $\mathbf{U}$ and $\mathbf{H}$ in Corollary 5.3 with different numbers of anchors. All experiments are conducted on a desktop with Intel(R) Core(TM)-i7-10870H CPU.

6.1. Information of Datasets

We first introduce all the datasets used in the experiments. The detailed information is listed in Table 1. As seen, the

first 6 datasets are small-scale and middle-scale datasets. Their number of samples varies from 600 to 10800. We choose these datasets to compare the proposed method with other baseline multiple kernel methods. The remaining 5 large-scale datasets are used to verify the efficiency of the proposed SMKC. The smallest number of samples is 30475, and the largest is 325834. Existing multiple kernel clustering methods make it almost impossible to handle the datasets on such a large scale with a desktop. All the URLs of used datasets are listed in Section B of the appendix.

Table 1: Detailed information of the datasets used in experiments.

Dataset	Number of		
	Samples	Views	Clusters
Synthetic3d	600	3	3
100Leaves	1600	3	100
Mfeat	2000	6	10
Wiki	2866	2	10
Handwritten	10000	2	10
ALOI-100	10800	4	100
AwA	30475	6	50
Cifar10	50000	3	10
Cifar100	50000	3	100
YtVideo	101499	5	31
Winnipeg	325834	2	7

6.2. Comparison Experiments

In the comparison experiments, we select 7 multiple kernel clustering methods as baselines, and their detailed information is as follows.

- **Single best** is the best clustering performance by standard kernel k-means on each base kernels.
- **Multiple kernel k-means (MKKM)** (Huang et al., 2012). MKKM utilizes an alternative optimization method to update the clustering partitions and kernel weights to achieve the optimal result.
- **Optimal neighborhood kernel clustering (ONKC)** (Liu et al., 2017). ONKC selects the potentially optimal consensus kernel from the neighbourhood field formed by a linear combination of base kernels.
- **Multiple kernel k-means with matrix-induced regularization (MKKM-MiR)** (Liu et al., 2016). MKKM-MiR introduces a new regularization term to learn the optimal kernel weights, enhancing kernel diversity and reducing redundancy.
- **Mulitple kernel clustering with local alignment maximization (LKAM)** (Li et al., 2016). LKAM seeks to learn the ideal similarity matrix by aligning each sample with its k-nearest neighbours rather than considering all samples.
- **Localized multiple kernel k-means (LMKKM)** (Gönen & Margolin, 2014). LMMKM aims to unite multiple local kernels about the data samples.
- **Simple Multiple Kernel k-means (SMKKM)** (Liu, 2022). SMKKM proposes a min-max learning paradigm, minimizing kernel weights while maximizing the clustering partition matrix.

We perform them on the first 6 datasets in Table 1, and record their clustering performance and consumed time. We use the default setting for all the comparison algorithms according to the corresponding papers. For the methods with hyper-parameters, we use grid search to select the optimal hyper-parameters according to clustering results. In fact, due to the lack of true labels, it is impossible to use clustering results to select the optimal hyper-parameters. The proposed SMKC has no hyper-parameters for tuning; thus, we can easily deploy it in real applications. For each view, the Gaussian kernel function is used to construct the kernel matrix, and the width δ^2 is specified as the mean of the pairwise squared distances. In SMKC, we fix the number of anchors as 1000 for sufficient anchors, i.e., 1000 columns from each base kernel matrix are sampled without replacement as the input of SMKC (except for the dataset termed Synthetic3d). For Synthetic3d, we choose all 600 columns. We use three commonly used clustering evaluating indicators to evaluate the clustering performance, i.e., accuracy (ACC), normalized mutual information (NMI), and purity. The records of experiments are listed in Table 2. "-" represents the consumed time that is longer than an hour.

As observed from Table 2, the proposed SMKC shows superior clustering performance compared with other 7 baseline methods. We can also find that:

1. SMKKM is a hyper-parameter-free method and demonstrates desirable clustering performance in many kernel datasets. (Liu, 2022). As a strong baseline, the proposed method outperforms SMKKM in most datasets. In "Handwritten", SMKC exceeds SMKKM 5.33%, 0.91% and 3.27% regarding ACC, NMI, and purity.

2. The proposed method has the advantage of high execution efficiency. As Table 2 shows, SMKC costs much less time than ONKC, MKKM-MiR, LKAM, LMKKM and SMKKM. Meanwhile, our method is comparable with MKKM regarding the running time but obtains much better clustering results.

Table 2: ACC, NMI, purity and time comparison of different clustering algorithms on 6 benchmark datasets. We use bold font to indicate the best performance among all algorithms. "-" represents the algorithm which consumes more than an hour.

Method	Single best	MKKM	ONKC	MKKM-MiR	LKAM	LMKKM	SMKKM	Proposed
				ACC (%)				
Synthetic3d	89.50	89.67	96.00	94.65	94.67	94.67	97.33	**97.67**
100Leaves	58.36	43.13	80.87	76.74	80.12	43.23	**81.14**	81.10
Mfeat	86.45	63.85	88.03	82.21	94.16	64.41	92.58	**94.95**
Wiki	53.15	53.15	48.45	45.23	38.57	52.98	48.97	**54.08**
Handwritten	47.93	70.25	-	70.88	69.38	63.86	72.11	**77.44**
ALOI-100	65.51	63.55	-	70.54	-	-	73.24	**74.46**
				NMI (%)				
Synthetic3d	69.13	67.37	84.53	80.58	81.38	80.69	88.36	**89.63**
100Leaves	78.08	69.83	91.96	89.61	90.04	69.88	92.39	**92.50**
Mfeat	75.79	65.17	78.82	73.85	**90.03**	65.61	86.44	89.48
Wiki	51.96	51.97	40.45	35.41	34.73	51.85	42.51	**52.73**
Handwritten	43.98	63.27	-	64.00	65.34	58.62	64.87	**65.78**
ALOI-100	79.19	78.10	-	81.35	-	-	83.16	**83.83**
				Purity (%)				
Synthetic3d	89.50	89.67	96.00	94.65	94.67	94.67	97.33	**97.67**
100Leaves	61.77	46.63	83.78	79.69	82.94	46.95	83.98	**84.12**
Mfeat	86.45	66.04	88.03	82.21	94.21	65.64	92.58	**94.95**
Wiki	61.29	61.29	51.65	47.92	47.49	61.22	53.31	**61.57**
Handwritten	52.28	70.67	-	73.49	72.61	68.39	74.17	**77.44**
ALOI-100	67.57	66.19	-	72.39	-	-	74.88	**76.14**
				Time (s)				
Synthetic3d	0.90	0.41	23.06	1.78	8.23	21.62	2.21	0.73
100Leaves	27.36	10.14	468.19	47.65	252.94	102.07	52.56	13.26
Mfeat	16.90	4.42	676.46	24.50	142.79	1225.06	25.61	11.47
Wiki	10.08	5.79	983.52	45.87	245.54	97.51	20.93	10.02
Handwritten	41.28	59.67	-	328.30	2045.23	1315.36	401.72	49.44
ALOI-100	372.23	455.59	-	945.41	-	-	1960.15	112.88

6.3. Approximation

In this subsection, we conduct experiments to verify the correctness of Corollary 5.3. The objective function Eq. (7) of the proposed SMKC is an approximation of Eq. (6). We derive two theoretical results as the approximation guarantee. Theorem 5.2 gives the approximation degree between the consensus kernel matrices $\mathbf{G}^*$ and $\mathbf{K}^*$. Notice that the final clustering indicator matrices $\mathbf{U}$ and $\mathbf{H}$ are produced by SVD on $\mathbf{G}^*$ or $\mathbf{K}^*$, respectively. Corollary 5.3 gives the approximation guarantee of $\mathbf{U}$ with regard to $\mathbf{H}$. We use two datasets for verification, i.e., Handwritten and ALOI-100. We first construct the whole kernel matrices as the input of SMKC to obtain $\mathbf{H}$. Then we let the anchor number s vary in the range of $\{100, 200, \cdots, 3000\}$, and uniformly sample s columns of the whole kernel matrices as the input for different $\mathbf{U}$. For each s, we record the subspace distance of $\mathbf{U}$ and $\mathbf{H}$, i.e., $\|\mathbf{U}\mathbf{U}^\top - \mathbf{H}\mathbf{H}\|_{\mathrm{F}}$. We illustrate the variation trend of the above distance in Figure 1. Meanwhile, we also record the corresponding execution time and clustering performance (i.e., NMI). To reduce the influence of randomness, we repeat the above sampling process for 20 times, and record the average values for different s. As the two subfigures on the left side of Figure 1 show, the error can decrease quickly as the number of anchors increases. As a reference, we also plot the curve of $f(s) = c\sqrt{1/s - 1/n}$, where c is a specified constant. We can see that the curve of true error can be upper bounded by $f(s)$ with different constant c. It testifies the validity of Corollary 5.3 empirically. The two subfigures in the middle of Figure 1 show that the execution time dramatically increases as s becomes large. Meanwhile, the NMI obtained by $\mathbf{U}$ first increases when s is smaller than 1000 and then fluctuates around the NMI obtained by $\mathbf{H}$. It shows that even if the number of anchors is much less than the sample number, the proposed method can also obtain similar clustering results acquired by performing SMKC on the whole kernel matrices. To summarize, the proposed SMKC can obtain desirable clustering performance in a short time.

6.4. Convergence

We illustrate the variation of the objective values in algorithm convergence in Figure 2. Four examples of the objective values of our algorithm at each iteration are recorded. We can observe that the objective function monotonically decreases and has a fast convergence rate.

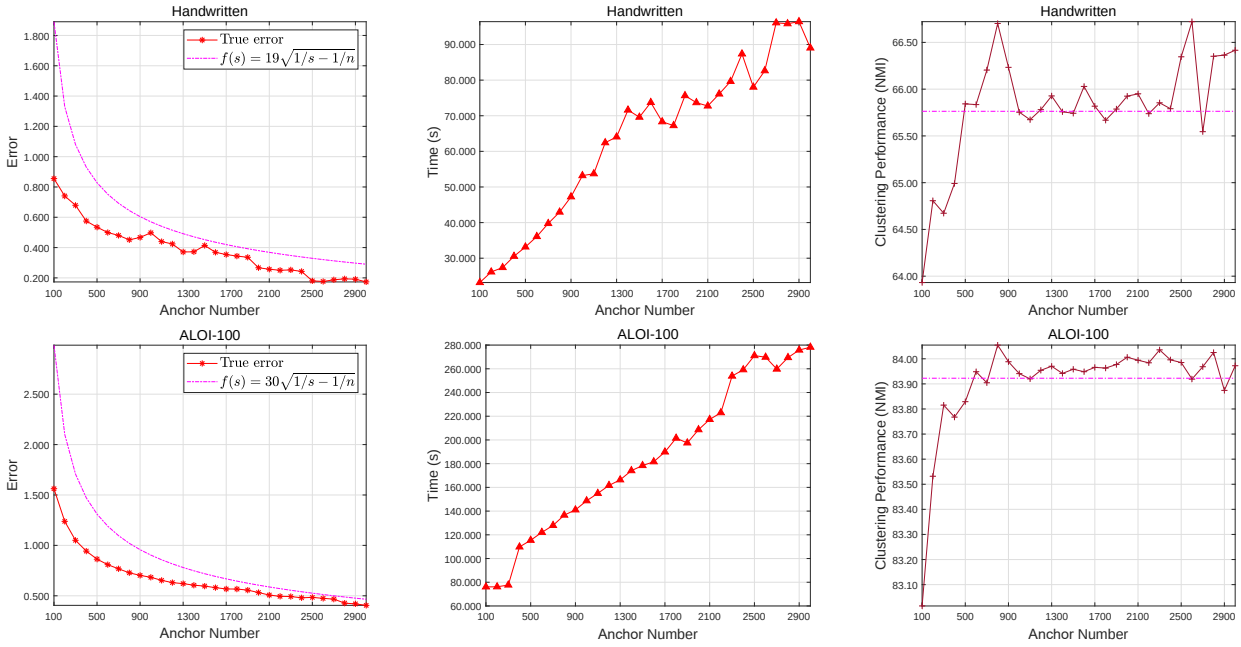

Figure 1: The experimental results for the verification of Corollary 5.3. Two subfigures on the left side show that the distance between $\mathbf{U}$ and $\mathbf{H}$ can be upper bounded by $c\sqrt{1/s - 1/n}$, where c is some constant. Two subfigures in the middle report the execution time of the proposed SMKC with the variation of sample number, while The subfigures on the right side record the corresponding NMI.

Table 3: Experimental results on AwA.

	ACC	NMI	Purity	Time (s)
View 1	7.64	8.38	9.49	
View 2	7.38	8.85	9.21	
View 3	7.22	7.51	8.95	106.22
View 4	7.61	8.90	9.25	
View 5	7.67	9.45	9.45	
View 6	8.02	9.40	10.21	
Proposed	**9.56**	**11.52**	**11.76**	96.86

Table 4: Experimental results on Cifar10.

	ACC	NMI	Purity	Time (s)
View 1	88.48	78.85	88.48	
View 2	85.50	72.09	85.50	16.51
View 3	86.37	73.58	86.37	
Proposed	**99.17**	**97.74**	**99.17**	33.78

Table 5: Experimental results on Cifar100.

	ACC	NMI	Purity	Time (s)
View 1	86.10	96.80	89.98	
View 2	86.50	96.96	89.98	243.47
View 3	84.81	91.71	87.04	
Proposed	**91.45**	**98.16**	**92.77**	315.16

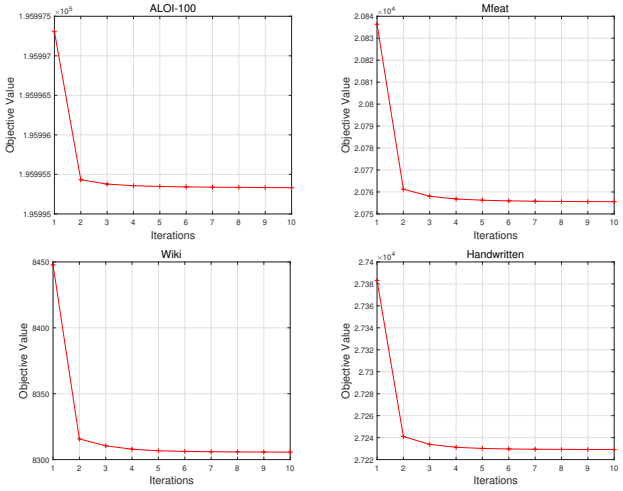

Figure 2: Illustration of the objective values of each iteration in the process of algorithm convergence.

6.5. Experiments on Large-Scale Datasets

We conduct experiments on five large-scale datasets to verify SMKC's ability to handle large-scale datasets. In these datasets, the least number of samples is 30475, and the largest is 325834. Existing multiple kernel clustering is almost impossible to deal with such scale datasets because they all need to construct the whole kernel matrix and the

consequent eigen (singular) decomposition. In the proposed SMKC, we only need to compute a few columns of the whole kernel matrix. As shown in the previous section, the complexity can be reduced to be linear with the sample number. In the large-scale experiments, we randomly sampled $s = 3 * [\sqrt{n}]$ anchors. We let the Gaussian width be the average squared distance between s selected anchors and n samples. For comparison, we perform kernel k-means with Nyström approximation (Kumar et al., 2012) on each single kernel and record the clustering results. To reduce the randomness, we repeat all experiments 20 times and report the average performance and consuming time from Table 3 to Table 7. From these tables, we can see that: 1. The proposed SMKC consistently outperforms the kernel k-means with Nyström in terms of clustering performance. SMKC can exceed the best single view by 1.54%, 10.69%, 4.95%, 2.49%, and 7.18% in terms of NMI on all 5 datasets. It shows that our method can effectively fuse the clustering structure of all single views. 2. In the aspect of execution time, SMKC can manage all large-scale datasets in the manner of kernel clustering method within hundreds of seconds. It shows the high efficiency of SMKC in handling large-scale datasets.

Table 6: Experimental results on YtVideo.

	ACC	NMI	Purity	Time (s)
View 1	8.95	5.55	26.86	148.64
View 2	17.52	16.91	29.67	
View 3	12.73	11.38	26.94	
View 4	17.57	15.69	29.26	
View 5	18.40	16.02	29.27	
Proposed	**20.89**	**17.64**	**33.02**	258.76

Table 7: Experimental results on Winnipeg.

	ACC	NMI	Purity	Time (s)
View 1	61.56	48.36	73.60	62.98
View 2	58.65	45.70	65.81	
Proposed	**68.74**	**55.07**	**80.84**	247.90

7. Conclusions and Limitations

In this paper, we obtain the expectation of kernel matrices under the assumption of a Gaussian-distributed data model. Then, we obtain an upper bound to measure how the empirical kernel matrix is concentrated in its expectation. Based on the theoretical results of the expectation kernel matrix, we then propose a scalable multiple kernel clustering algorithm with an anchor-based method. With different anchor numbers s, we give the approximation bounds for the consensus kernel and clustering indicator matrix. We provide strict and detailed proofs for all proposed theorems. Finally, we conduct experiments to test the effectiveness of the proposed method and the relevant theoretical results.

However, there are also some limitations existing in our work which can be summarized as follows:

1. Equal view weights: Our approach assumes that all views carry equal weight, neglecting to account for variations in the importance of individual views.

2. Sole use of Frobenius norm: We exclusively rely on the Frobenius norm for measuring the distance between matrices, potentially overlooking other metrics that might provide valuable insights.

3. Limited consideration of dataset distribution: The proposed method is tailored specifically for Gaussian-distributed datasets, thus failing to address potential variations present in datasets with alternative distributions.

We will address the above issues in the future.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NO. 62376252, 62325604, 62276271, 62376039, 62306324, 62376279), the Key Project of Natural Science Foundation of Zhejiang Province (LZ22F030003).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abbe, E., Fan, J., and Wang, K. An l_p theory of pca and spectral clustering. *The Annals of Statistics*, 50(4):2359–2385, 2022.

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

Cohen, M. B., Elder, S., Musco, C., Musco, C., and Persu, M. Dimensionality reduction for k-means clustering and low rank approximation. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing (STOC)*, pp. 163–172, 2015.

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*, pp. 551–556, 2004.

Ding, X. and Ma, R. Learning low-dimensional nonlinear structures from high-dimensional noisy data: an integral operator approach. *The Annals of Statistics*, 51(4):1744–1769, 2023.

Gönen, M. and Margolin, A. A. Localized data fusion for kernel k-means clustering with application to cancer biology. In *Advances in Neural Information Processing Systems*, pp. 1305–1313, 2014.

Han, X., Tong, X., and Fan, Y. Eigen selection in spectral clustering: a theory-guided practice. *Journal of the American Statistical Association*, 118(541):109–121, 2023.

Huang, H.-C., Chuang, Y.-Y., and Chen, C.-S. Multiple kernel fuzzy clustering. *IEEE Transactions on Fuzzy Systems*, 20(1):120–134, 2012.

Kumar, S., Mohri, M., and Talwalkar, A. Sampling methods for the nyström method. *The Journal of Machine Learning Research*, 13(1):981–1006, 2012.

Li, M., Liu, X., Wang, L., Dou, Y., Yin, J., and Zhu, E. Multiple kernel clustering with local kernel alignment maximization. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1704–1710, 2016.

Liang, W., Liu, X., Liu, Y., Huang, J.-J., Wang, S., Liu, J., Zhang, Y., Zhu, E., et al. Stability and generalization of kernel clustering: From single kernel to multiple kernel. *Advances in Neural Information Processing Systems*, 35: 33633–33645, 2022.

Liang, W., Liu, X., Liu, Y., Ma, C., Zhao, Y., Liu, Z., and Zhu, E. Consistency of multiple kernel clustering. In *International Conference on Machine Learning*, pp. 20650–20676. PMLR, 2023.

Liu, X. Simplemkkm: Simple multiple kernel k-means. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(4):5174–5186, 2022.

Liu, X. Hyperparameter-free localized simple multiple kernel k-means with global optimum. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Liu, X., Dou, Y., Yin, J., Wang, L., and Zhu, E. Multiple kernel k-means clustering with matrix-induced regularization. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pp. 1888–1894, 2016.

Liu, X., Zhou, S., Wang, Y., Li, M., Dou, Y., Zhu, E., and Yin, J. Optimal neighborhood kernel clustering with multiple kernels. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pp. 2266–2272, 2017.

Löffler, M., Zhang, A. Y., and Zhou, H. H. Optimality of spectral clustering in the gaussian mixture model. *The Annals of Statistics*, 49(5):2506–2530, 2021.

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

Ren, Z. and Sun, Q. Simultaneous global and local graph structure preserving for multiple kernel clustering. *IEEE transactions on neural networks and learning systems*, 32 (5):1839–1851, 2020.

Shi, T., Belkin, M., and Yu, B. Data spectroscopy: Learning mixture models using eigenspaces of convolution operators. In *Proceedings of the 25th international conference on Machine learning*, pp. 936–943, 2008.

Shi, T., Belkin, M., and Yu, B. Data spectroscopy: Eigenspaces of convolution operators and clustering. *The Annals of Statistics*, pp. 3960–3984, 2009.

Srivastava, P. R., Sarkar, P., and Hanasusanto, G. A. A robust spectral clustering algorithm for sub-gaussian mixture models with outliers. *Operations Research*, 71(1):224–244, 2023.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

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

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

Yang, M.-S., Lai, C.-Y., and Lin, C.-Y. A robust em clustering algorithm for gaussian mixture models. *Pattern Recognition*, 45(11):3950–3961, 2012.

Yu, S., Wang, S., Wen, Y., Wang, Z., Luo, Z., Zhu, E., and Liu, X. How to construct corresponding anchors for incomplete multiview clustering. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2023.

Yu, S., Wang, S., Dong, Z., Tu, W., Liu, S., Lv, Z., Li, P., Wang, M., and Zhu, E. A non-parametric graph clustering framework for multi-view data. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pp. 16558–16567, 2024a.

Yu, S., Wang, S., Zhang, P., Wang, M., Wang, Z., Liu, Z., Fang, L., Zhu, E., and Liu, X. Dvsai: Diverse view-shared

anchors based incomplete multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pp. 16568–16577, 2024b.

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 *Proceedings of the 2009 SIAM international conference on data mining*, pp. 638–649. SIAM, 2009.

A. Proof

A.1. The proof of Theorem 4.1

We need the following lemma to prove Theorem 4.1.

Lemma A.1. *For any two real numbers a, b and random variable $x \sim \mathcal{N}(0, \sigma^2)$, the equality*

$$\mathbb{E}_x\left[\exp\left(-\frac{(ax+b)^2}{2}\right)\right] = \frac{1}{\sqrt{1+a^2\sigma^2}} \cdot \exp\left(-\frac{b^2}{2+2a^2\sigma^2}\right)$$

holds.

Proof of Lemma A.1. Since $x \sim \mathcal{N}(0, \sigma^2)$, we have

$$\begin{aligned}
&\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) \exp\left(-\frac{(ax+b)^2}{2}\right) dx \\
=&\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2/\sigma^2 + a^2x^2 + 2abx + b^2}{2}\right) dx \\
=&\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(x\sqrt{\sigma^{-2}+a^2} + \frac{ab}{\sqrt{\sigma^{-2}+a^2}}\right)^2 + b^2 - \frac{a^2b^2}{\sigma^{-2}+a^2}}{2}\right) dx \\
=&\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(x\sqrt{\sigma^{-2}+a^2} + \frac{ab}{\sqrt{\sigma^{-2}+a^2}}\right)^2 + b^2 - \frac{a^2b^2}{\sigma^{-2}+a^2}}{2}\right) dx \\
=&\exp\left(-\frac{b^2}{2+2a^2\sigma^2}\right) \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(x\sqrt{\sigma^{-2}+a^2} + \frac{ab}{\sqrt{\sigma^{-2}+a^2}}\right)^2}{2}\right) dx \\
&\text{(Let } t = x\sqrt{\sigma^{-2}+a^2}. \text{)} \\
=&\exp\left(-\frac{b^2}{2+2a^2\sigma^2}\right) \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(t + \frac{ab}{\sqrt{\sigma^{-2}+a^2}}\right)^2}{2}\right) d\left(\frac{t}{\sqrt{\sigma^{-2}+a^2}}\right) \\
=&\exp\left(-\frac{b^2}{2+2a^2\sigma^2}\right) \cdot \frac{1}{\sqrt{1+a^2\sigma^2}}.
\end{aligned} \tag{12}$$

□

Now, we can proceed with the proof of Theorem 4.1.

Proof of Theorem 4.1. For any two samples $\mathbf{x}_i^{(v)} \in \mathcal{C}_p$ and $\mathbf{x}_j^{(v)} \in \mathcal{C}_q$, we have

$$\begin{aligned}
[\mathbb{E}(\mathbf{K}_v)]_{ij} =& \mathbb{E}\exp\left(-\frac{\left\|\boldsymbol{\mu}_p^{(v)} + \boldsymbol{\epsilon}_i^{(v)} - \boldsymbol{\mu}_q^{(v)} - \boldsymbol{\epsilon}_j^{(v)}\right\|^2}{2(\delta^{(v)})^2}\right) \\
=& \prod_{t=1}^{d} \mathbb{E}\exp\left(-\frac{\left(\mu_{pt}^{(v)} - \mu_{qt}^{(v)} + \epsilon_{it}^{(v)} - \epsilon_{jt}^{(v)}\right)^2}{2(\delta^{(v)})^2}\right),
\end{aligned} \tag{13}$$

where $\mu_{pt}^{(v)}$ is the t-th component of $\boldsymbol{\mu}_p^{(v)}$, and the definitions of $\mu_{qt}^{(v)}, \epsilon_{it}^{(v)}, \epsilon_{jt}^{(v)}$ are similar to $\mu_{pt}^{(v)}$.

By the definition of $\boldsymbol{\epsilon}_i^{(v)}, \boldsymbol{\epsilon}_j^{(v)}$, we have $\boldsymbol{\epsilon}_i^{(v)} \sim \mathcal{N}(0, (\sigma_p^{(v)})^2/d)$ and $\boldsymbol{\epsilon}_j^{(v)} \sim \mathcal{N}(0, (\sigma_q^{(v)})^2/d)$. Thus, $\boldsymbol{\epsilon}_i^{(v)} - \boldsymbol{\epsilon}_j^{(v)} \sim \mathcal{N}(0, (\sigma_p^{(v)})^2/d + (\sigma_q^{(v)})^2/d)$. Letting $x = \boldsymbol{\epsilon}_i^{(v)} - \boldsymbol{\epsilon}_j^{(v)}$, $a = 1/\delta^{(v)}$ and $b = (\mu_{pt}^{(v)} - \mu_{qt}^{(v)})/\delta^{(v)}$, we have

$$
\begin{aligned}
&\mathbb{E} \exp\left(-\frac{\left(\mu_{pt}^{(v)} - \mu_{qt}^{(v)} + \epsilon_{it}^{(v)} - \epsilon_{jt}^{(v)}\right)^2}{2(\delta^{(v)})^2}\right) \\
=&\mathbb{E}_x \exp\left(-\frac{(ax+b)^2}{2}\right) \\
=&\frac{1}{\sqrt{1 + a^2((\sigma_p^{(v)})^2/d + (\sigma_q^{(v)})^2/d))}} \cdot \exp\left(-\frac{b^2}{2 + 2a^2((\sigma_p^{(v)})^2 + (\sigma_q^{(v)})^2)}\right) \\
=&\frac{\delta^{(v)}}{\sqrt{(\delta^{(v)})^2 + (\sigma_p^{(v)})^2/d + (\sigma_q^{(v)})^2/d}} \cdot \exp\left(-\frac{\left(\mu_{pt}^{(v)} - \mu_{qt}^{(v)}\right)^2}{2(\delta^{(v)})^2 + 2(\sigma_p^{(v)})^2/d + 2(\sigma_q^{(v)})^2/d}\right).
\end{aligned}
\tag{14}
$$

The second equality holds according to Lemma A.1. Combining Eq. (13) and Eq. (14), we can obtain Theorem 4.1.

□

A.2. The proof of Theorem 2

For ease of deduction, we omit the superscript (v) and subscript v in the proof, e.g., $\boldsymbol{\mu}_p$ is $\boldsymbol{\mu}_p^{(v)}$ and $K(\cdot,\cdot)$ is $K_v(\cdot,\cdot)$. The following lemma is an equivalent form of Theorem 5.2.2 of (Vershynin, 2018).

Lemma A.2. *(Gaussian concentration). Consider a random vector $\mathbf{z} \in \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ and a L-Lipschitz function $f : \mathbb{R}^d \mapsto \mathbb{R}$. Then,*

$$\Pr(f(\mathbf{z}) - \mathbb{E}\, f(\mathbf{z}) \geq Lt) \leq \exp(-ct^2)$$

holds with some constant c.

Then, we prove the following two Lipschitz properties of the Gaussian kernel function.

Lemma A.3. *By Assumption 3.1, the training set of some view can be written as $\mathbf{X} = \mathbf{U} + \mathbf{E}$, where $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, $\mathbf{U} = [\boldsymbol{\mu}_{1_j}, \cdots, \boldsymbol{\mu}_{n_j}]$ (for some sample i and $j \in [k]$, i_j means that $\mathbf{x}_i$ belongs to the j-th cluster), and $\mathbf{E} = [\boldsymbol{\epsilon}_1, \cdots, \boldsymbol{\epsilon}_n]$ is the matrix of Gaussian noise. We denote the kernel matrix constructed by $\mathbf{X}$ and Gaussian kernel function $K(\mathbf{x}, \mathbf{y}) = f_\delta(\|\mathbf{x} - \mathbf{y}\|) := (-\|\mathbf{x} - \mathbf{y}\|^2/2\delta^2)$ as $K(\mathbf{X}) := [K(\mathbf{X})]_{i\times j} = K(\mathbf{x}_i, \mathbf{x}_j)$. Then, the following two Lipschitz properties hold*

1. *$f_\delta(x) := \exp(-x^2/2\delta^2)$ is $\frac{1}{\sqrt{e}\delta}$-Lipschitz.*
2. *$K(\mathbf{X})$ is $\frac{2\sqrt{n}}{\sqrt{e}\delta}$-Lipschitz, w.r.t. F-norm.*

Proof of Lemma A.3. We first prove the first part of Lemma A.3. Notice that $f_\delta(x)$ is continuously differentiable. For any $x, y \in \mathbb{R}$, by Newton-Leibniz formula,

$$|f_\delta(y) - f_\delta(x)| = \left|\int_x^y f_\delta'(t)dt\right| = \left|\int_x^y \frac{t}{\delta^2}\exp\left(-\frac{t^2}{2\delta^2}\right)dt\right| \leq \frac{1}{\sqrt{e}\delta}|y - x|.$$

The last inequality holds due to the maximum of $\frac{t}{\delta^2}\exp\left(-\frac{t^2}{2\delta^2}\right)$ is $\frac{1}{\sqrt{e}\delta}$ for any $t \in \mathbb{R}$.

Then, we prove the latter part.

$$\begin{aligned}
&\|K(\mathbf{X}) - K(\mathbf{X}')\|_{\mathrm{F}}^2 \\
=&\sum_{i=1}^{n}\sum_{j=1}^{n}(K(\mathbf{x}_i,\mathbf{x}_j) - K(\mathbf{x}_i',\mathbf{x}_j'))^2 \\
=&\sum_{i=1}^{n}\sum_{j=1}^{n}(f_\delta(\|\mathbf{x}_i - \mathbf{x}_j\|) - f_\delta(\|\mathbf{x}_i' - \mathbf{x}_j'\|))^2 \\
\leq&\frac{1}{e\delta^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\|\mathbf{x}_i - \mathbf{x}_j\| - \|\mathbf{x}_i' - \mathbf{x}_j'\|)^2 \\
\leq&\frac{1}{e\delta^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\|\mathbf{x}_i - \mathbf{x}_i' + \mathbf{x}_j - \mathbf{x}_j'\|)^2 \\
\leq&\frac{2}{e\delta^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\|\mathbf{x}_i - \mathbf{x}_i'\|^2 + \|\mathbf{x}_j - \mathbf{x}_j'\|^2) \\
=&\frac{4n}{e\delta^2}\|\mathbf{X} - \mathbf{X}'\|_{\mathrm{F}}^2.
\end{aligned} \tag{15}$$

Thus, $K(\mathbf{X})$ is $\frac{2\sqrt{n}}{\sqrt{e}\delta}$-Lipschitz, w.r.t. F-norm. □

Lemma A.4. *The variance of $K(\mathbf{x}_i,\mathbf{x}_j)$ has the following upper bound*

$$\mathbb{E}(K(\mathbf{x}_i,\mathbf{x}_j) - \mathbb{E}\,K(\mathbf{x}_i,\mathbf{x}_j))^2 \precsim \frac{\sigma_{p,q}^2}{\delta^2 d},$$

where $\sigma_{p,q} = \max\{\sigma_p,\sigma_q\}$, and σ_p,σ_q are the noise scale of $\mathbf{x}_i,\mathbf{x}_j$, respectively. Furthermore,

$$\mathbb{E}\,\|K(\mathbf{X}) - \mathbb{E}\,K(\mathbf{X})\|_{\mathrm{F}} \precsim \frac{n\sigma_\infty}{\delta\sqrt{d}}.$$

Proof of Lemma A.4. For any sample $\mathbf{x}_i$ and its copy $\mathbf{x}_i'$, we assume that $\mathbf{x}_i = \boldsymbol{\mu}_p + \boldsymbol{\epsilon}_i$ and $\mathbf{x}_i' = \boldsymbol{\mu}_p + \boldsymbol{\epsilon}_i'$, where $\boldsymbol{\epsilon}_i,\boldsymbol{\epsilon}_i' \sim \mathcal{N}(\mathbf{0}_d, \frac{\sigma_p}{\sqrt{d}}\mathbf{I}_d)$. Similarly, we suppose $\mathbf{x}_j = \boldsymbol{\mu}_q + \boldsymbol{\epsilon}_j$, and let $\mathbf{x}_j'$ be its copy. Then, we have

$$\begin{aligned}
&|K(\mathbf{x}_i,\mathbf{x}_j) - K(\mathbf{x}_i',\mathbf{x}_j')| \\
=&|K(\boldsymbol{\mu}_p + \boldsymbol{\epsilon}_i, \boldsymbol{\mu}_q + \boldsymbol{\epsilon}_j) - K(\boldsymbol{\mu}_p + \boldsymbol{\epsilon}_i', \boldsymbol{\mu}_q + \boldsymbol{\epsilon}_j')| \\
=&|f_\delta(\|\boldsymbol{\mu}_p + \boldsymbol{\epsilon}_i - \boldsymbol{\mu}_q - \boldsymbol{\epsilon}_j\|) - f_\delta(\|\boldsymbol{\mu}_p + \boldsymbol{\epsilon}_i' - \boldsymbol{\mu}_q - \boldsymbol{\epsilon}_j'\|)| \\
\leq&\frac{1}{\sqrt{e}\delta}\left|\|\|\boldsymbol{\mu}_p + \boldsymbol{\epsilon}_i - \boldsymbol{\mu}_q - \boldsymbol{\epsilon}_j\| - \|\boldsymbol{\mu}_p + \boldsymbol{\epsilon}_i' - \boldsymbol{\mu}_q - \boldsymbol{\epsilon}_j'\|\right| \\
\leq&\frac{1}{\sqrt{e}\delta}\left(\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_i' + \boldsymbol{\epsilon}_j - \boldsymbol{\epsilon}_j'\|\right) \\
=&\frac{1}{\sqrt{e}\delta}\left(\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_i'\| + \|\boldsymbol{\epsilon}_j - \boldsymbol{\epsilon}_j'\|\right) \\
=&\frac{1}{\sqrt{e}\delta}\left(\frac{\sigma_p}{\sqrt{d}}\|\mathbf{z}_i - \mathbf{z}_i'\| + \frac{\sigma_q}{\sqrt{d}}\|\mathbf{z}_j - \mathbf{z}_j'\|\right) \\
\leq&\frac{2\sigma_{p,q}}{\delta\sqrt{ed}}\sqrt{\|\mathbf{z}_i - \mathbf{z}_i'\|^2 + \|\mathbf{z}_j - \mathbf{z}_j'\|^2} \\
=&\frac{2\sigma_{p,q}}{\delta\sqrt{ed}}\|[\mathbf{z}_i;\mathbf{z}_j] - [\mathbf{z}_i';\mathbf{z}_j']\|.
\end{aligned} \tag{16}$$

Thus, the function $[\mathbf{z}_i;\mathbf{z}_j] \mapsto K(\boldsymbol{\mu}_p + \frac{\sigma_p}{\sqrt{d}}\mathbf{z}_i, \boldsymbol{\mu}_q + \frac{\sigma_q}{\sqrt{d}}\mathbf{z}_j)$ is $\frac{2\sigma_{p,q}}{\delta\sqrt{ed}}$-Lipschitz. Notice that $[\mathbf{z}_i;\mathbf{z}_j] \sim \mathcal{N}(\mathbf{0}_{2d}, \mathbf{I}_{2d})$, according to Lemma A.2, we have

$$\Pr\left(|K(\mathbf{x}_i,\mathbf{x}_j) - \mathbb{E}\,K(\mathbf{x}_i,\mathbf{x}_j)|) \geq \frac{2\sigma_{p,q}}{\delta\sqrt{ed}}\right) \leq 2\exp(-ct^2),$$

where c is some constant. Assuming that $L = \frac{2\sigma_{p,q}}{\delta\sqrt{ed}}$, we have

$$\begin{aligned}
&\mathbb{E}(K(\mathbf{x}_i,\mathbf{x}_j) - \mathbb{E}\,K(\mathbf{x}_i,\mathbf{x}_j))^2 \\
=&\int_0^{+\infty} \Pr((K(\mathbf{x}_i,\mathbf{x}_j) - \mathbb{E}\,K(\mathbf{x}_i,\mathbf{x}_j))^2 > x)dx \\
=&\int_0^{+\infty} 2t\Pr(|K(\mathbf{x}_i,\mathbf{x}_j) - \mathbb{E}\,K(\mathbf{x}_i,\mathbf{x}_j)| > t)dt \\
=&2L^2\int_0^{+\infty} t\Pr(|K(\mathbf{x}_i,\mathbf{x}_j) - \mathbb{E}\,K(\mathbf{x}_i,\mathbf{x}_j)| > Lt)dt \\
\leq&2L^2\int_0^{+\infty} 2t\exp(-ct^2)dt \\
=&\frac{2L^2}{c}.
\end{aligned} \tag{17}$$

Above all, we know that

$$\mathbb{E}(K(\mathbf{x}_i,\mathbf{x}_j) - \mathbb{E}\,K(\mathbf{x}_i,\mathbf{x}_j))^2 \precsim \frac{\sigma_{p,q}^2}{\delta^2 d}.$$

We then prove the latter part of this lemma, and we have

$$\mathbb{E}\,\|K(\mathbf{X}) - \mathbb{E}\,K(\mathbf{X})\|_{\mathrm{F}}^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}(K(\mathbf{x}_i,\mathbf{x}_j) - \mathbb{E}\,K(\mathbf{x}_i,\mathbf{x}_j))^2 \precsim \frac{n^2\sigma_\infty^2}{\delta^2 d},$$

where $\sigma_\infty = \max\{\sigma_1,\cdots,\sigma_k\}$. Then, by Jensen's inequality, we have

$$\mathbb{E}\,\|K(\mathbf{X}) - \mathbb{E}\,K(\mathbf{X})\|_{\mathrm{F}} \leq \sqrt{\mathbb{E}\,\|K(\mathbf{X}) - \mathbb{E}\,K(\mathbf{X})\|_{\mathrm{F}}^2} \precsim \frac{n\sigma_\infty}{\delta\sqrt{d}}.$$

□

Lemma A.5. *the probability inequality*

$$\Pr(\|K(\mathbf{X}) - \mathbb{E}\,K(\mathbf{X})\|_{\mathrm{F}} - \mathbb{E}\,\|K(\mathbf{X}) - \mathbb{E}\,K(\mathbf{X})\|_{\mathrm{F}} \geq \frac{2\sqrt{n}\sigma_\infty t}{\delta\sqrt{ed}})) \leq \exp(-ct^2)$$

holds with some constant c.

We repeat to use Lemma A.2 to prove Lemma A.5. The detailed process is as follows.

Proof of Lemma A.5. Assuming that $F(\mathbf{X}) = \|K(\mathbf{X}) - \mathbb{E}\,K(\mathbf{X})\|_{\mathrm{F}}$ and $\mathbf{X}' = \mathbf{U} + \mathbf{E}'$ is another copy of $\mathbf{X}$, we can deduce that

$$\begin{aligned}
&|F(\mathbf{X}) - F(\mathbf{X}')| \\
=&|\|K(\mathbf{X}) - \mathbb{E}\,K(\mathbf{X})\|_{\mathrm{F}} - \|K(\mathbf{X}') - \mathbb{E}\,K(\mathbf{X}')\|_{\mathrm{F}}| \\
\leq&\|K(\mathbf{X}) - K(\mathbf{X}')\|_{\mathrm{F}} \\
\leq&\frac{2\sqrt{n}}{\sqrt{e}\delta}\|\mathbf{X} - \mathbf{X}'\|_{\mathrm{F}}. \text{ (By Lemma A.3.)}
\end{aligned} \tag{18}$$

Then, we can have

$$\|\mathbf{X} - \mathbf{X}'\|_{\mathrm{F}}^2 = \|\mathbf{E} - \mathbf{E}'\|_{\mathrm{F}}^2 = \sum_{i=1}^{n}\|\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_i'\|^2 \leq \frac{\sigma_\infty^2}{d}\sum_{i=1}^{n}\|\mathbf{z}_i - \mathbf{z}_i'\|^2 = \frac{\sigma_\infty^2}{d}\|[\mathbf{z}_1;\cdots;\mathbf{z}_n] - [\mathbf{z}_1';\cdots;\mathbf{z}_n']\|^2,$$

where $\sigma_\infty = \max\{\sigma_1, \cdots, \sigma_k\}$, and $\mathbf{z}_1, \cdots, \mathbf{z}_n, \mathbf{z}'_1, \cdots, \mathbf{z}'_n$ are standard Gaussian variables with distribution $\mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$. By the above deduction, we know that $[\mathbf{z}_1; \cdots; \mathbf{z}_n] \mapsto F\left(\mathbf{U} + \left[\frac{\sigma_{1_j}}{\sqrt{d}}\mathbf{z}_1, \cdots, \frac{\sigma_{n_j}}{\sqrt{d}}\mathbf{z}_n\right]\right)$ is $\frac{2\sqrt{n}\sigma_\infty}{\delta\sqrt{ed}}$-Lipschitz. Notice that $[\mathbf{z}_1; \cdots; \mathbf{z}_n] \sim \mathcal{N}(\mathbf{0}_{dn}, \mathbf{I}_{dn})$. According to Lemma A.2,

$$\Pr(\|K(\mathbf{X}) - \mathbb{E}\,K(\mathbf{X})\|_{\mathrm{F}} - \mathbb{E}\,\|K(\mathbf{X}) - \mathbb{E}\,K(\mathbf{X})\|_{\mathrm{F}} \geq \frac{2\sqrt{n}\sigma_\infty t}{\delta\sqrt{ed}}) \leq \exp(-ct^2)$$

with some constant c. □

Now, we can complete the proof of Theorem 4.2.

Proof of Theorem 4.2. Combining Lemma A.4 and Lemma A.5, we can obtain

$$\|K(\mathbf{X}) - \mathbb{E}\,K(\mathbf{X})\|_{\mathrm{F}} \precsim \mathbb{E}\,\|K(\mathbf{X}) - \mathbb{E}\,K(\mathbf{X})\|_{\mathrm{F}} + \frac{\sqrt{n}\sigma_\infty t}{\delta\sqrt{d}} \precsim \frac{\sigma_\infty}{\delta\sqrt{d}}(n + \sqrt{n}t)$$

holds with probability at least $1 - \exp(-t^2)$. □

A.3. The proof of Theorem 5.2

We need several lemmas and a definition to prove Theorem 5.2. The first lemma is the famous Weyl's bound.

Lemma A.6. *Suppose $\mathbf{A} \in \mathbb{R}^{n\times s}$ has rank r, let $\{\lambda_j\}_{j=1}^r$ be its singular values. Then, for an arbitrary matrix $\mathbf{E}$, assume that $\{\tilde{\lambda}_j\}_{j=1}^r$ are the singular values of $\mathbf{A} + \mathbf{E}$. Then, for any integer $1 \leq j \leq r$, the following inequality holds*

$$\max_{j\in[r]} |\lambda_j - \tilde{\lambda}_j| \leq \|\mathbf{E}\|.$$

The following lemma is about the perturbation of eigenvectors of Hermitian matrices.

Lemma A.7. *(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*

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

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., $\arccos(\mathbf{h}_j^\top\hat{\mathbf{h}}_j)$.

The following lemma gives a perturbation bound of the best rank-k approximation of some matrix $\mathbf{A}$.

Lemma A.8. *For some real matrix $\mathbf{A} \in \mathbb{R}^{n\times s}$, denote its SVD is $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}$. Then, it is easy to check the best rank-k approximation of $\mathbf{A}$ is $\mathbf{U}_k\mathbf{D}_k\mathbf{V}_k$, where $\mathbf{U}_k, \mathbf{V}_k$ are respectively composed of the first columns of $\mathbf{U}, \mathbf{V}$, and $\mathbf{D}_k$ is a diagonal matrix whose diagonal elements are the first k singular values of $\mathbf{A}$. For another matrix $\mathbf{B} \in \mathbb{R}^{n\times s}$, we similarly assume the rank-k approximation of $\mathbf{B}$ as $\widetilde{\mathbf{U}}_k\widetilde{\mathbf{D}}_k\widetilde{\mathbf{V}}_k$. Then, we have*

$$\|\widetilde{\mathbf{U}}_k\widetilde{\mathbf{D}}_k^2\widetilde{\mathbf{U}}_k - \mathbf{U}_k\mathbf{D}_k^2\mathbf{U}_k\| \precsim \|\mathbf{A}\mathbf{A}^\top - \mathbf{B}\mathbf{B}^\top\| + \frac{\max\{\tilde{\lambda}_1^2, \lambda_1^2\}\|\mathbf{A}\mathbf{A}^\top - \mathbf{B}\mathbf{B}^\top\|}{\lambda_k - \lambda_{k+1}}.$$

Moreover, when $\lambda_1, \tilde{\lambda}_1 \leq 1$, we have

$$\|\widetilde{\mathbf{U}}_k\widetilde{\mathbf{D}}_k^2\widetilde{\mathbf{U}}_k^\top - \mathbf{U}_k\mathbf{D}_k^2\mathbf{U}_k^\top\| \precsim \|\mathbf{A}\mathbf{A}^\top - \mathbf{B}\mathbf{B}^\top\|.$$

Proof of Lemma A.8. We can make a decomposition as follows.

$$\begin{aligned}
&\|\widetilde{\mathbf{U}}_k\widetilde{\mathbf{D}}_k^2\widetilde{\mathbf{U}}_k - \mathbf{U}_k\mathbf{D}_k^2\mathbf{U}_k\| \\
=&\|\widetilde{\mathbf{U}}_k\widetilde{\mathbf{D}}_k^2\widetilde{\mathbf{U}}_k^\top - \mathbf{U}_k\widetilde{\mathbf{D}}_k^2\widetilde{\mathbf{U}}_k^\top + \mathbf{U}_k\widetilde{\mathbf{D}}_k^2\widetilde{\mathbf{U}}_k^\top - \mathbf{U}_k\mathbf{D}_k^2\widetilde{\mathbf{U}}_k^\top - \mathbf{U}_k\mathbf{D}_k^2\mathbf{U}_k^\top\| \\
\leq&\underbrace{\|\widetilde{\mathbf{U}}_k\widetilde{\mathbf{D}}_k^2\widetilde{\mathbf{U}}_k - \mathbf{U}_k\widetilde{\mathbf{D}}_k^2\widetilde{\mathbf{U}}_k^\top\|}_{\mathcal{A}} + \underbrace{\|\mathbf{U}_k\widetilde{\mathbf{D}}_k^2\widetilde{\mathbf{U}}_k^\top - \mathbf{U}_k\mathbf{D}_k^2\widetilde{\mathbf{U}}_k^\top\|}_{\mathcal{B}} + \underbrace{\|\mathbf{U}_k\mathbf{D}_k^2\widetilde{\mathbf{U}}_k^\top - \mathbf{U}_k\mathbf{D}_k^2\mathbf{U}_k^\top\|}_{\mathcal{C}}
\end{aligned} \tag{19}$$

For Item $\mathcal{A}$, we have

$$\mathcal{A} \leq \|\widetilde{\mathbf{U}}_k - \mathbf{U}_k\| \cdot \|\widetilde{\mathbf{D}}_k^2\| \cdot \|\widetilde{\mathbf{U}}_k\| \leq \tilde{\lambda}_1^2 \left\|\widetilde{\mathbf{U}}_k - \mathbf{U}_k\right\|_{\mathrm{F}} = \tilde{\lambda}_1^2 \sqrt{2k - 2\mathrm{Tr}\left(\widetilde{\mathbf{U}}_k^\top \mathbf{U}_k\right)} \precsim \tilde{\lambda}_1^2 \left\|\sin\theta(\mathbf{H}_{\boldsymbol{\gamma}}, \mathbf{H}_{\boldsymbol{\beta}})\right\|_{\mathrm{F}}.$$

By Lemma A.7, we have

$$\mathcal{A} \precsim \frac{\tilde{\lambda}_1^2 \|\mathbf{A}\mathbf{A}^\top - \mathbf{B}\mathbf{B}^\top\|}{\lambda_k - \lambda_{k+1}}.$$

Similarly, we can obtain

$$\mathcal{C} \precsim \frac{\lambda_1^2 \|\mathbf{A}\mathbf{A}^\top - \mathbf{B}\mathbf{B}^\top\|}{\lambda_k - \lambda_{k+1}}.$$

By Lemma A.6, we have

$$\mathcal{B} \leq \|\mathbf{U}_k\| \cdot \|\widetilde{\mathbf{D}}_k^2 - \mathbf{D}_k^2\| \cdot \|\widetilde{\mathbf{V}}_k\| \leq \max_{j\in[k]} |\tilde{\lambda}_j^2 - \lambda_j^2| \leq \|\mathbf{A}\mathbf{A}^\top - \mathbf{B}\mathbf{B}^\top\|.$$

Thus, we have

$$\|\widetilde{\mathbf{U}}_k \widetilde{\mathbf{D}}_k \widetilde{\mathbf{V}}_k - \mathbf{U}_k \mathbf{D}_k \mathbf{V}_k\| \precsim \|\mathbf{A}\mathbf{A}^\top - \mathbf{B}\mathbf{B}^\top\| + \frac{\max\{\tilde{\lambda}_1^2, \lambda_1^2\} \|\mathbf{A}\mathbf{A}^\top - \mathbf{B}\mathbf{B}^\top\|}{\lambda_k - \lambda_{k+1}}.$$

□

Lemma A.9. *(Bardenet & Maillard, 2015) Let $A_n = \{x_i\}_{i=1}^n$ be a finite sequence of real numbers, and $A_s = \{x_t\}_{t=1}^s$ are s points uniformly selected from it without replacement. Then, for any $t > 0$, the following probability inequality holds*

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

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

Definition A.10. By the reproducing property of kernel functions, for any two samples x, y and a kernel function $K(\cdot,\cdot)$ have the following property,

$$K(x, y) = \langle K_x, K_y \rangle_{\mathcal{H}},$$

where $\mathcal{H}$ represents Hilbert space, and K_x, K_y are two elements of $\mathcal{H}$. For some datasets $\{x_i\}_{i=1}^n$, we define an operator as follows,

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

Similarly, for some anchor set $\{a_t\}_{t=1}^s$, we define an operator as

$$T_s : \mathcal{H} \to \mathcal{H},\ T_s = \frac{1}{s}\sum_{t=1}^{s} \langle \cdot, K_{a_i} \rangle_{\mathcal{H}} K_{a_i}.$$

Lemma A.11. *For some Gaussian kernel matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$, we denote a kernel similarity kernel matrix composed of s (uniformly sampled without replacement) columns of $\mathbf{K}$ as $\mathbf{G}$. Then, we have*

$$\left\|\frac{1}{ns}\mathbf{G}\mathbf{G}^\top - \frac{1}{n^2}\mathbf{K}\mathbf{K}^\top\right\| \precsim t\sqrt{\frac{1}{s} - \frac{1}{n}} \tag{20}$$

holds with probability at least $1 - \exp(-t^2)$.

Proof of Lemma A.11. By the reproducing property of the kernel function, we can write $\mathbf{G} = \mathbf{\Phi}_n^\top \mathbf{\Phi}_s$, where $\mathbf{\Phi}_n = [K_{x_1}, \cdots, K_{x_n}]$ and $\mathbf{\Phi}_s = [K_{a_1}, \cdots, K_{a_s}]$ composed of the anchors corresponding sampled columns. We can also let $\mathbf{K} = \mathbf{\Phi}_n^\top \mathbf{\Phi}_n$. By Definition A.10, we have

$$\begin{aligned} &\|\frac{1}{ns}\mathbf{G}\mathbf{G}^\top - \frac{1}{n^2}\mathbf{K}\mathbf{K}^\top\| \\ =&\|\frac{1}{ns}\mathbf{\Phi}_n^\top\mathbf{\Phi}_s\mathbf{\Phi}_s^\top\mathbf{\Phi}_n - \frac{1}{n^2}\mathbf{\Phi}_n^\top\mathbf{\Phi}_n\mathbf{\Phi}_n^\top\mathbf{\Phi}_n\| \\ =&\|\frac{1}{n}\mathbf{\Phi}_n^\top(T_s - T_n)\mathbf{\Phi}_n\| \\ \leq&\|T_s - T_n\| \cdot \|T_n\|. \end{aligned} \tag{21}$$

Moreover, we have

$$\|T_n\| = \|\frac{1}{n}\mathbf{K}\| \leq \|\frac{1}{n}\mathbf{K}\|_{\mathrm{F}} = \frac{1}{n}\mathrm{Tr}(\mathbf{K}) = 1,$$

and

$$\|T_s - T_n\| \leq \sup_{\|f\|=1,\|g\|=1} \left|\frac{1}{n}\sum_{i=1}^{n} f(x_i)g(x_i) - \frac{1}{s}\sum_{t=1}^{s} f(x_t)g(x_t)\right| \leq 2t\sqrt{\frac{1}{s} - \frac{1}{n}}$$

holds with probability at least $1 - \exp(-t^2)$. The last inequality holds due to Lemma A.9. Combining all, we can derive Eq.(20).

□

Proof of Theorem 5.2. We denote the best rank-k approximation of some matrix $\mathbf{A}$ as $\mathcal{M}_k$. After the t-iteration ($t \in [T]$), we assume that $\mathbf{G}^*, \mathbf{K}^*$ are updated as $\mathbf{G}^{(t)}, \mathbf{K}^{(t)}$, respectively. Then, we have

$$\begin{aligned} &\left\|\frac{1}{ns}\mathbf{G}^{(t+1)}(\mathbf{G}^{(t+1)})^\top - \frac{1}{n^2}\mathbf{K}^{(t+1)}(\mathbf{K}^{(t+1)})^\top\right\| \\ =&\left\|\mathcal{M}_k\left(\frac{1}{V}\sum_{v=1}^{V}\frac{\widetilde{\mathbf{G}}_v^{(t+1)}}{\sqrt{ns}}\right)\mathcal{M}_k^\top\left(\frac{1}{V}\sum_{v=1}^{V}\frac{\widetilde{\mathbf{G}}_v^{(t+1)}}{\sqrt{ns}}\right) - \mathcal{M}_k\left(\frac{1}{V}\sum_{v=1}^{V}\frac{\widetilde{\mathbf{K}}_v^{(t+1)}}{n}\right)\mathcal{M}_k^\top\left(\frac{1}{V}\sum_{v=1}^{V}\frac{\widetilde{\mathbf{K}}_v^{(t+1)}}{n}\right)\right\| \\ \precsim&\left\|\left(\frac{1}{V}\sum_{v=1}^{V}\frac{\widetilde{\mathbf{G}}_v^{(t+1)}}{\sqrt{ns}}\right)\left(\frac{1}{V}\sum_{v=1}^{V}\frac{\widetilde{\mathbf{G}}_v^{(t+1)}}{\sqrt{ns}}\right)^\top - \left(\frac{1}{V}\sum_{v=1}^{V}\frac{\widetilde{\mathbf{K}}_v^{(t+1)}}{n}\right)\left(\frac{1}{V}\sum_{v=1}^{V}\frac{\widetilde{\mathbf{K}}_v^{(t+1)}}{n}\right)^\top\right\| \\ \precsim&\max_{v\in[V]}\left\|\left(\frac{\widetilde{\mathbf{G}}_v^{(t+1)}}{\sqrt{ns}}\right)\left(\frac{\widetilde{\mathbf{G}}_v^{(t+1)}}{\sqrt{ns}}\right)^\top - \left(\frac{\widetilde{\mathbf{K}}_v^{(t+1)}}{n}\right)\left(\frac{\widetilde{\mathbf{K}}_v^{(t+1)}}{n}\right)^\top\right\|, \end{aligned} \tag{22}$$

where the second last inequality holds due to $\mathcal{M}_k(\mathbf{A})\mathcal{M}_k^\top(\mathbf{A})$ is the best rank-k approximation of $\mathbf{A}\mathbf{A}^\top$ and Lemma A.8.

For the v-th item in the last line of Eq. (22), we have

$$\begin{aligned} &\left\|\left(\frac{\widetilde{\mathbf{G}}_v^{(t+1)}}{\sqrt{ns}}\right)\left(\frac{\widetilde{\mathbf{G}}_v^{(t+1)}}{\sqrt{ns}}\right)^\top - \left(\frac{\widetilde{\mathbf{K}}_v^{(t+1)}}{n}\right)\left(\frac{\widetilde{\mathbf{K}}_v^{(t+1)}}{n}\right)^\top\right\| \\ =&\left\|\mathcal{M}_k\left(\frac{\mathbf{G}_v + \mathbf{G}^{(t)}}{2\sqrt{ns}}\right)\mathcal{M}_k^\top\left(\frac{\mathbf{G}_v + \mathbf{G}^{(t)}}{2\sqrt{ns}}\right) - \mathcal{M}_k\left(\frac{\mathbf{K}_v + \mathbf{K}^{(t)}}{2n}\right)\mathcal{M}_k^\top\left(\frac{\mathbf{K}_v + \mathbf{K}^{(t)}}{2n}\right)\right\| \\ \precsim&\left\|\frac{1}{ns}\mathbf{G}_v\mathbf{G}_v^\top - \frac{1}{n^2}\mathbf{K}_v\mathbf{K}_v^\top\right\| + \left\|\frac{1}{ns}\mathbf{G}^{(t)}(\mathbf{G}^{(t)})^\top - \frac{1}{n^2}\mathbf{K}^{(t)}(\mathbf{K}^{(t)})^\top\right\|, \end{aligned} \tag{23}$$

where the last inequality can be derived by the similar way as Eq (22).

Combining Eq. (22), Eq. (23) and Lemma A.11, we have

$$\left\|\frac{1}{ns}\mathbf{G}^{(t+1)}(\mathbf{G}^{(t+1)})^\top - \frac{1}{n^2}\mathbf{K}^{(t+1)}(\mathbf{K}^{(t+1)})^\top\right\| \precsim \left\|\frac{1}{ns}\mathbf{G}^{(t)}(\mathbf{G}^{(t)})^\top - \frac{1}{n^2}\mathbf{K}^{(t)}(\mathbf{K}^{(t)})^\top\right\| + \xi\sqrt{\frac{1}{s} - \frac{1}{n}} \tag{24}$$

holds with probability at least $1-\exp(-\xi^2)$.

By the recurrence relation in (24), we have

$$\begin{aligned}
&\left\|\frac{1}{ns}\mathbf{G}^*(\mathbf{G}^*)^\top-\frac{1}{n^2}\mathbf{K}^*(\mathbf{K}^*)^\top\right\| \\
=&\left\|\frac{1}{ns}\mathbf{G}^T(\mathbf{G}^T)^\top-\frac{1}{n^2}\mathbf{K}^T(\mathbf{K}^T)^\top\right\| \\
\precsim&\left\|\frac{1}{ns}\mathbf{G}^{(T-1)}(\mathbf{G}^{(T-1)})^\top-\frac{1}{n^2}\mathbf{K}^{(T-1)}(\mathbf{K}^{(T-1)})^\top\right\|+\xi\sqrt{\frac{1}{s}-\frac{1}{n}} \\
\precsim&\cdots \\
\precsim&\left\|\frac{1}{ns}\mathbf{G}^1(\mathbf{G}^1)^\top-\frac{1}{n^2}\mathbf{K}^1(\mathbf{K}^1)^\top\right\|+\xi\sqrt{\frac{1}{s}-\frac{1}{n}}
\end{aligned} \tag{25}$$

holds with probability at least $(1-\exp(-\xi^2))^{(T-1)}$.

From the initialization of the proposed algorithm, we can obtain

$$\begin{aligned}
&\left\|\frac{1}{ns}\mathbf{G}^1(\mathbf{G}^1)^\top-\frac{1}{n^2}\mathbf{K}^1(\mathbf{K}^1)^\top\right\| \\
\precsim&\left\|\mathcal{M}_k\left(\frac{1}{V}\sum_{v=1}^{V}\frac{\mathbf{G}_v}{\sqrt{ns}}\right)\mathcal{M}_k^\top\left(\frac{1}{V}\sum_{v=1}^{V}\frac{\mathbf{G}_v}{\sqrt{ns}}\right)-\mathcal{M}_k\left(\frac{1}{V}\sum_{v=1}^{V}\frac{\mathbf{K}_v}{n}\right)\mathcal{M}_k^\top\left(\frac{1}{V}\sum_{v=1}^{V}\frac{\mathbf{K}_v}{n}\right)\right\| \\
\precsim&\xi\sqrt{\frac{1}{s}-\frac{1}{n}}
\end{aligned} \tag{26}$$

holds with probability at least $1-\exp(-\xi^2)$.

Combining Eq. (25) and Eq. (26), by union bound, we have

$$\left\|\frac{1}{ns}\mathbf{G}^*(\mathbf{G}^*)^\top-\frac{1}{n^2}\mathbf{K}^*(\mathbf{K}^*)^\top\right\|\precsim(\xi+\sqrt{\log T})\sqrt{\frac{1}{s}-\frac{1}{n}}$$

holds with probability at least $1-\exp(-\xi^2)$. □

A.4. Convergence Analysis

Now we prove the convergence of the objective function. Let the objective function be

$$f(\widetilde{\mathbf{G}}_v,\mathbf{G}^*)=\sum_{v=1}^{V}\|\widetilde{\mathbf{G}}_v-\mathbf{G}_v\|_{\mathrm{F}}^2+\|\widetilde{\mathbf{G}}_v-\mathbf{G}^*\|_{\mathrm{F}}^2.$$

According to (Cohen et al., 2015), the best rank-k approximation can be obtained by singular value decomposition. In the t-th iteration, when $(\widetilde{\mathbf{G}}_v)_{(t)}$ $(v\in[V])$ are updated, we know

$$f((\widetilde{\mathbf{G}}_v)_{(t+1)},(\mathbf{G}^*)_{(t)})\leq f((\widetilde{\mathbf{G}}_v)_{(t)},(\mathbf{G}^*)_{(t)}).$$

Subsequently, after updating $(\mathbf{G}^*)_{(t)}$, we know

$$f((\widetilde{\mathbf{G}}_v)_{(t+1)},(\mathbf{G}^*)_{(t+1)})\leq f((\widetilde{\mathbf{G}}_v)_{(t+1)},(\mathbf{G}^*)_{(t)}).$$

Then, we have $f((\widetilde{\mathbf{G}}_v)_{(t+1)},(\mathbf{G}^*)_{(t+1)})\leq f((\widetilde{\mathbf{G}}_v)_{(t)},(\mathbf{G}^*)_{(t)})$. We can see that f is a monotonically decreased function w.r.t. the iteration t. Obviously, $f((\widetilde{\mathbf{G}}_v)_{(t)},(\mathbf{G}^*)_{(t)})\geq 0$ for all t. By the monotone convergence theorem, we can ensure the convergence of the objective function by the proposed optimization method.

B. Datasets

1. 100Leaves: https://archive.ics.uci.edu/dataset/241/one+hundred+plant+species+leaves+data+set
2. ALOI-100: https://aloi.science.uva.nl/
3. Mfeat: https://archive.ics.uci.edu/dataset/72/multiple+features
4. Handwritten: https://cs.nyu.edu/roweis/data.html
5. Synthetic3d: https://cdinstitute.github.io/Building-Dataset-Generator/
6. Wiki: http://www.svcl.ucsd.edu/projects/crossmodal/
7. AwA: https://cvml.ista.ac.at/AwA/
8. Cifar10 & Cifar100: https://www.cs.toronto.edu/~kriz/cifar.html
9. YtVideo: http://archive.ics.uci.edu/ml/datasets/YouTube+Multiview+Video+Games+Dataset
10. Winnipeg: https://archive.ics.uci.edu/ml/datasets/Crop+mapping+using+fused+optical-radar+data+set