Fair Kernel K-Means: from Single Kernel to Multiple Kernel

Peng Zhou *
School of Computer Science and Technology
Anhui University
Hefei, 230601
zhoupeng@ahu.edu.cn

Rongwen Li
School of Computer Science and Technology
Anhui University
Hefei, 230601
e22301284@stu.ahu.edu.cn

Liang Du
School of Computer and Information Technology
Shanxi University
Taiyuan, 237016
duliang@sxu.edu.cn

Abstract

Kernel k-means has been widely studied in machine learning. However, existing kernel k-means methods often ignore the *fairness* issue, which may cause discrimination. To address this issue, in this paper, we propose a novel Fair Kernel K-Means (FKKM) framework. In this framework, we first propose a new fairness regularization term that can lead to a fair partition of data. The carefully designed fairness regularization term has a similar form to the kernel k-means which can be seamlessly integrated into the kernel k-means framework. Then, we extend this method to the multiple kernel setting, leading to a Fair Multiple Kernel K-Means (FMKKM) method. We also provide some theoretical analysis of the generalization error bound, and based on this bound we give a strategy to set the hyper-parameter, which makes the proposed methods easy to use. At last, we conduct extensive experiments on both the single kernel and multiple kernel settings to compare the proposed methods with state-of-the-art methods to demonstrate their effectiveness. Our code is available at https://github.com/rongwenli/NeurIPS24-FMKKM.

1 Introduction

Clustering is a fundamental unsupervised machine learning task. In clustering, kernel methods, such as Kernel K-Means (KKM), can effectively separate nonlinear data into different clusters. Therefore, KKM has been widely studied in both the single kernel setting and multiple kernel setting [39, 50, 14, 15].

Notice that, in real-world applications, clustering is often used in some scenarios involving humans such as social networks [36] and crime analysis [32]. In these scenarios, since the humans are involved, we should guarantee the *fairness* of the clustering result, so that the clustering result will not cause discrimination to some specific groups. In the clustering task, we often consider the *group fairness*, where we have some pre-given groups that may suffer from the potential discrimination, called *protected groups*. Group fairness aims to partition data into some clusters and guarantee that no clusters contain a disproportionately small or large number of data in some specific protected

*Peng Zhou is the corresponding author. Peng Zhou and Rongwen Li are also with Anhui Provincial International Joint Research Center for Advanced Technology in Medical Imaging.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-3162

groups [6]. Although the above-mentioned kernel k-means and multiple kernel k-means methods show promising performance in the clustering task, none of them considers the fairness issue, and thus they may obtain some clustering results which cause discrimination to some groups.

To tackle this problem, in this paper, we propose a novel fair kernel k-means method and extend it from the single kernel setting to the multiple kernel setting. We follow a widely-used definition of fairness defined in [6], which is shown as Definition 1. By analyzing this definition, we carefully design a new fairness regularization term and prove that minimizing this term can lead to the optimal fairness defined in [6]. Besides, we observe that our fairness regularization term has a similar form of the loss function of KKM, and thus can be naturally and seamlessly plugged into the KKM framework, yielding an extremely simple and elegant Fair Kernel K-Means (FKKM) framework. This framework is so concise that we do not even need to modify the loss of KKM but just adjust the input kernel to our proposed *fair kernel*. This framework can also be easily extended to the Multiple Kernel K-Means (MKKM) task, leading to Fair Multiple Kernel K-Means (FMKKM). We also provide some theoretical analysis of its generalization error bound. Furthermore, based on the generalization error bound, we provide a strategy to set the hyper-parameter in our framework, which makes the method easy to use. Extensive experiments on single kernel clustering and multiple kernel clustering tasks show the effectiveness of our framework w.r.t. both the clustering accuracy and fairness.

The main contributions of our paper are summarized as follows:

- We propose a novel fairness regularization term and prove that minimizing this term can reach the optimal fairness defined in [6].
- Our proposed regularization term has a similar form to the KKM, and thus can be seamlessly integrated into the KKM and MKKM framework. To the best of our knowledge, this is the first work for fair kernel k-means and fair multiple kernel k-means.
- We provide a strategy to set the hyper-parameter based on the theoretical analysis, which makes the methods easy to use.
- Extensive experiments in both single and multiple kernel clustering show the effectiveness and superiority of our proposed methods compared with the state-of-the-art methods.

2 Related Work and Preliminaries

In this paper, we use a bold uppercase letter (e.g. $\mathbf{M}$) and a bold lowercase letter (e.g. $\mathbf{v}$) to denote a matrix and a vector, respectively. Given a matrix $\mathbf{M}$, we use M_{ij} to denote its (i,j)-th element.

2.1 Kernel K-means and Multiple Kernel K-means

Given a data matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$ with n instances and d features, let $\Phi(\cdot) : \mathbb{R}^d \mapsto \mathcal{H}$ represents a kernel mapping that maps $\mathbf{X}$ into a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$. The objective function of the kernel k-means with the sum-of-squares loss can be written as [39, 24]:

$$\min_{\mathbf{M},\mathbf{Y}\in Ind} \|\Phi(\mathbf{X}) - \mathbf{M}\mathbf{Y}^T\|_F^2, \tag{1}$$

where $\Phi(\mathbf{X}) = [\Phi(\mathbf{x}_1), \ldots, \Phi(\mathbf{x}_n)]$ and $\mathbf{M} = [\mathbf{m}_1, \ldots, \mathbf{m}_c]$ represents c clustering centroids in the RKHS $\mathcal{H}$. $\mathbf{Y} \in \{0,1\}^{n\times c}$ is an indicator matrix, which is denoted as Ind, and $Y_{ij} = 1$ if $\mathbf{x}_i$ is assigned to the j-th cluster, and otherwise $Y_{ij} = 0$. Setting the derivative of Eq.(1) w.r.t. $\mathbf{M}$ to zero, we can obtain the closed-form solution of $\mathbf{M}$. Taking it back to Eq.(1), it can be rewritten as [42]:

$$\min_{\mathbf{Y}\in Ind} \mathrm{Tr}\left(\mathbf{K}\right) - \mathrm{Tr}\left(\left(\mathbf{Y}^T\mathbf{Y}\right)^{-\frac{1}{2}}\mathbf{Y}^T\mathbf{K}\mathbf{Y}\left(\mathbf{Y}^T\mathbf{Y}\right)^{-\frac{1}{2}}\right), \tag{2}$$

where $\mathbf{K} = \Phi(\mathbf{X})^T\Phi(\mathbf{X}) \in \mathbb{R}^{n\times n}$ is a kernel matrix with $K_{ij} = \Phi(\mathbf{x}_i)^T\Phi(\mathbf{x}_j)$. For the convenience of optimization, we denote $\mathbf{H} = \mathbf{Y}\left(\mathbf{Y}^T\mathbf{Y}\right)^{-\frac{1}{2}}$. Since directly solving Eq.(2) is an NP-hard problem [16], previous works [26, 41, 21] substituted the constraints $\mathbf{Y} \in Ind$ with $\mathbf{H}^T\mathbf{H} = \mathbf{I}$, leading to:

$$\min_{\mathbf{H}^T\mathbf{H}=\mathbf{I}} \mathrm{Tr}\left(\mathbf{K}\left(\mathbf{I} - \mathbf{H}^T\mathbf{H}\right)\right). \tag{3}$$

The optimal $\mathbf{H}$ is formed by the c eigenvectors of $\mathbf{K}$ corresponding to the c largest eigenvalues. After obtaining $\mathbf{H}$, existing methods [54, 44, 35, 17] learn the final clustering results through some post-processing techniques such as k-means or spectral rotation on $\mathbf{H}$.

Multiple kernel k-means aims to fuse multiple base kernels to a consensus one for kernel k-means. Previous works assume that the ideal consensus kernel matrix is a combination of base kernel matrices i.e., $\mathbf{K}^* = \sum_{p=1}^{m} \gamma_p^2 \mathbf{K}^{(p)}$, where $\mathbf{K}^*$ is the consensus kernel matrix, and $\mathbf{K}^{(p)}$s are base kernels [27, 28, 19]. γ_p is the weight of the p-th base kernel. Replacing $\mathbf{K}$ in Eq.(3) with the consensus kernel $\mathbf{K}^*$, we can obtain the objective function of MKKM:

$$\min_{\mathbf{H},\boldsymbol{\gamma}} \operatorname{Tr}\left(\mathbf{K}^*\left(\mathbf{I}-\mathbf{H}^T\mathbf{H}\right)\right),\ \ s.t.\ \mathbf{H}^T\mathbf{H}=\mathbf{I},\ \boldsymbol{\gamma}^T\mathbf{1}=1,\ \gamma_p\geq 0,\ \mathbf{K}^*=\sum_{p=1}^{m}\gamma_p^2\mathbf{K}^{(p)}. \tag{4}$$

It can be solved by alternatively optimizing $\mathbf{H}$ and $\boldsymbol{\gamma}$.

2.2 Fair Clustering

Fair clustering considers the fairness in the clustering, which is an important problem in unsupervised machine learning. It was first introduced by Chierichetti et al., who proposed a fair decomposition method to avoid all members of a protected group being clustered into the same cluster [9]. However, this method can only handle two protected groups. To tackle this problem, Bera et al. further proposed a concept of fairness applicable to multiple protected groups in [6], which is defined as:

Definition 1 *(Fairness) [6] Given a data matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$ with n instances and d features, it is partitioned into c disjoint clusters $\mathcal{C} = \{\pi_1, \cdots, \pi_c\}$. Given t disjoint protected groups $\mathcal{G}_1, \mathcal{G}_2, \cdots, \mathcal{G}_t$, let $\eta_i = \frac{|\mathcal{G}_i|}{n}$ and $\eta_i(k) = \frac{|\pi_k \cap \mathcal{G}_i|}{|\pi_k|}$ denote the proportion of group $\mathcal{G}_i$ in the whole data and cluster π_k, respectively. The fairnesss of a cluster π_k is defined as:*

$$fairness\left(\pi_k\right) = \min\left(\frac{\eta_i}{\eta_i(k)}, \frac{\eta_i(k)}{\eta_i}\right),\ \forall i \in \{1, \cdots t\} \tag{5}$$

The fairness of the whole clustering result $\mathcal{C}$ is defined as:

$$fairness(\mathcal{C}) = \min_{k\in\{1,\cdots c\}} fairness(\pi_k) \tag{6}$$

Remark 1 *$fairness(\mathcal{C}) \in [0, 1]$, and the larger $fairness(\mathcal{C})$ is, the fairer the clustering result is. A fair clustering result requires that the proportion of $\mathcal{G}_i$ in each cluster, which is denoted as $\eta_i(k)$, should be close to the proportion of $\mathcal{G}_i$ in the whole data, which is denoted as η_i. When all $\eta_i(k) = \eta_i$, the $fairness$ will achieve its maximum value 1, which means it is perfectly fair.*

Based on Definition 1, many fair clustering methods have been proposed [4, 7, 1, 37]. For example, Ziko et al. proposed a variational fair clustering framework by integrating fairness term with a clustering objective [57]; Kleindessner et al. embedded fairness as a linear constraint into spectral clustering obtaining fair spectral clustering [18]; Ghadiri et al. introduced a fair k-means method that ensures all protected groups have equal cluster costs [12]; Li et al. proposed a deep fair clustering method [20]. Wang et al. embedded this fairness into deep clustering by learning a differentiated and fair clustering allocation function [40]; Chhabra et al. provided a robust deep fair clustering method by considering the fairness attack [8].

3 Methodology

3.1 Fairness Regularization Term

We first introduce our fairness regularization term. To control the fairness, according to Definition 1, we need to compute $|\pi_k \cap \mathcal{G}_i|$ and $|\pi_k|$ in $\eta_i(k)$. To this end, we introduce two indicator matrices $\mathbf{G} \in \{0, 1\}^{n\times t}$ and $\mathbf{Y} \in \{0, 1\}^{n\times c}$. $\mathbf{G}$ is a protected group indicator matrix, where $G_{ij} = 1$ if the i-th instance belongs to the j-th protected group, and $G_{ij} = 0$ otherwise. $\mathbf{Y}$ is a cluster indicator matrix, where $Y_{ij} = 1$ if the i-th instance belongs to the j-th cluster, and $Y_{ij} = 0$ otherwise. It is easy to verify that

$$\mathbf{G}^T\mathbf{Y} = \begin{bmatrix} |\pi_1 \cap \mathcal{G}_1| & |\pi_2 \cap \mathcal{G}_1| & \cdots & |\pi_c \cap \mathcal{G}_1| \\ |\pi_1 \cap \mathcal{G}_2| & |\pi_2 \cap \mathcal{G}_2| & \cdots & |\pi_c \cap \mathcal{G}_2| \\ \vdots & \vdots & \ddots & \vdots \\ |\pi_1 \cap \mathcal{G}_t| & |\pi_2 \cap \mathcal{G}_t| & \cdots & |\pi_c \cap \mathcal{G}_t| \end{bmatrix},\ and\ \ \mathbf{Y}^T\mathbf{Y} = \begin{bmatrix} |\pi_1| & 0 & \cdots & 0 \\ 0 & |\pi_2| & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & |\pi_c| \end{bmatrix} \tag{7}$$

 https://doi.org/10.52202/079017-3162

Notice that $\mathbf{G}$ is a constant matrix because the protected groups are often pre-given, while $\mathbf{Y}$ is a variable that needs to learn for clustering. Based on Eq.(7), we define a fair regularization term $\mathrm{Tr}\left(\mathbf{Y}^T\mathbf{G}\mathbf{G}^T\mathbf{Y}\left(\mathbf{Y}^T\mathbf{Y}\right)^{-1}\right)$ and provide the following Theorem, which shows that minimizing this regularization term leads to the maximum of the fairness defined in Definition 1.

Theorem 1 *Given* $\mathbf{G}$ *and* $\mathbf{Y}$ *defined as mentioned before, we can obtain the maximum of fairness by optimizing the following objective function:*

$$\min_{\mathbf{Y}\in Ind} \mathrm{Tr}\left(\mathbf{Y}^T\mathbf{G}\mathbf{G}^T\mathbf{Y}\left(\mathbf{Y}^T\mathbf{Y}\right)^{-1}\right). \tag{8}$$

Proof 1 *We first have:*

$$\mathrm{Tr}\left(\mathbf{Y}^T\mathbf{G}\mathbf{G}^T\mathbf{Y}\left(\mathbf{Y}^T\mathbf{Y}\right)^{-1}\right) = \mathrm{Tr}\left(\left(\mathbf{Y}^T\mathbf{Y}\right)^{-\frac{1}{2}}\mathbf{Y}^T\mathbf{G}\mathbf{G}^T\mathbf{Y}\left(\mathbf{Y}^T\mathbf{Y}\right)^{-\frac{1}{2}}\right) = \left\|\mathbf{G}^T\mathbf{Y}\left(\mathbf{Y}^T\mathbf{Y}\right)^{-\frac{1}{2}}\right\|_F^2.$$

According to Eq,(7)*, we have*

$$\mathbf{G}^T\mathbf{Y}\left(\mathbf{Y}^T\mathbf{Y}\right)^{-\frac{1}{2}} = \begin{bmatrix} \frac{|\pi_1\cap\mathcal{G}_1|}{\sqrt{|\pi_1|}} & \frac{|\pi_2\cap\mathcal{G}_1|}{\sqrt{|\pi_2|}} & \cdots & \frac{|\pi_c\cap\mathcal{G}_1|}{\sqrt{|\pi_c|}} \\ \frac{|\pi_1\cap\mathcal{G}_2|}{\sqrt{|\pi_1|}} & \frac{|\pi_2\cap\mathcal{G}_2|}{\sqrt{|\pi_2|}} & \cdots & \frac{|\pi_c\cap\mathcal{G}_2|}{\sqrt{|\pi_c|}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{|\pi_1\cap\mathcal{G}_t|}{\sqrt{|\pi_1|}} & \frac{|\pi_2\cap\mathcal{G}_t|}{\sqrt{|\pi_2|}} & \cdots & \frac{|\pi_c\cap\mathcal{G}_t|}{\sqrt{|\pi_c|}}. \end{bmatrix} \tag{9}$$

Therefore, minimizing Eq.(8) *is equivalent to minimizing the following formula:*

$$\left\|\mathbf{G}^T\mathbf{Y}\left(\mathbf{Y}^T\mathbf{Y}\right)^{-\frac{1}{2}}\right\|_F^2 = \sum_{i=1}^{t}\sum_{k=1}^{c}\frac{|\pi_k\cap\mathcal{G}_i|^2}{|\pi_k|}. \tag{10}$$

According to Cauchy-Schwarz Inequality, we have:

$$\left(\sum_{k=1}^{c}\frac{|\pi_k\cap\mathcal{G}_i|^2}{|\pi_k|}\right)\left(\sum_{k=1}^{c}|\pi_k|\right) \geq \left(\sum_{k=1}^{c}|\pi_k\cap\mathcal{G}_i|\right)^2 = |\mathcal{G}_i|^2 \Rightarrow \sum_{k=1}^{c}\frac{|\pi_k\cap\mathcal{G}_i|^2}{|\pi_k|} \geq \frac{|\mathcal{G}_i|^2}{n}. \tag{11}$$

Summing Eq.(11) *w.r.t. i, we have*

$$\sum_{i=1}^{t}\sum_{k=1}^{c}\frac{|\pi_k\cap\mathcal{G}_i|^2}{|\pi_k|} \geq \sum_{i=1}^{t}\frac{|\mathcal{G}_i|^2}{n}. \tag{12}$$

The equation in Eq.(12) *holds if and only if $\frac{|\pi_1\cap\mathcal{G}_i|}{|\pi_1|} = \frac{|\pi_2\cap\mathcal{G}_i|}{|\pi_2|} = \cdots = \frac{|\pi_c\cap\mathcal{G}_i|}{|\pi_c|}$ for any i. It is easy to verify that $\frac{|\pi_1\cap\mathcal{G}_i|}{|\pi_1|} = \cdots = \frac{|\pi_c\cap\mathcal{G}_i|}{|\pi_c|} = \frac{\sum_k|\pi_k\cap\mathcal{G}_i|}{\sum_k|\pi_k|}$. Notice that π_k is a disjoint partition of all data, and thus we have $(\pi_1\cap\mathcal{G}_i)\cup\cdots\cup(\pi_c\cap\mathcal{G}_i) = \mathcal{G}_i$ and $(\pi_p\cap\mathcal{G}_i)\cap(\pi_q\cap\mathcal{G}_i) = \emptyset$ for any p, q. Therefore, we have $\sum_k|\pi_k\cap\mathcal{G}_i| = |\mathcal{G}_i|$. Similarly, we have $\sum_k|\pi_k| = n$. Taking them back to the condition of the equation holding, we have that the equation holds if and only if $\frac{|\pi_1\cap\mathcal{G}_i|}{|\pi_1|} = \frac{|\pi_2\cap\mathcal{G}_i|}{|\pi_2|} = \cdots = \frac{|\pi_c\cap\mathcal{G}_i|}{|\pi_c|} = \frac{|\mathcal{G}_i|}{n}$.*

Notice that $\frac{|\pi_k\cap\mathcal{G}_i|}{|\pi_k|} = \eta_i(k)$ and $\frac{|\mathcal{G}_i|}{n} = \eta_i$. Therefore, when we minimize Eq.(8)*, we have $\eta_i(k) = \eta_i$. According to Definition 1, it will lead to maximum fairness. This concludes the proof.*

According to Theorem 1, we provide a simple yet effective fair regularization term Eq.(8), and can easily plug it into the KKM and MKKM framework.

3.2 Fair Kernel K-means

Notice that the fairness regularization term $\mathrm{Tr}\left(\mathbf{Y}^T\mathbf{G}\mathbf{G}^T\mathbf{Y}\left(\mathbf{Y}^T\mathbf{Y}\right)^{-1}\right)$ has a similar form to KKM (i.e., Eq.(2)). Therefore, we can seamlessly integrate this term into the KKM framework, leading to a fair kernel k-means (FKKM):

$$\begin{aligned} &\min_{\mathbf{Y}\in Ind}\mathrm{Tr}\left(\mathbf{K}\right) - \mathrm{Tr}\left(\left(\mathbf{Y}^T\mathbf{Y}\right)^{-\frac{1}{2}}\mathbf{Y}^T\mathbf{K}\mathbf{Y}\left(\mathbf{Y}^T\mathbf{Y}\right)^{-\frac{1}{2}}\right) + \lambda\,\mathrm{Tr}\left(\mathbf{Y}^T\mathbf{G}\mathbf{G}^T\mathbf{Y}\left(\mathbf{Y}^T\mathbf{Y}\right)^{-1}\right) \\ \Longleftrightarrow\ &\max_{\mathbf{Y}\in Ind}\mathrm{Tr}\left(\mathbf{Y}^T\left(\mathbf{K}-\lambda\mathbf{G}\mathbf{G}^T\right)\mathbf{Y}\left(\mathbf{Y}^T\mathbf{Y}\right)^{-1}\right), \end{aligned} \tag{13}$$

where λ is a hyper-parameter to balance the trade-off between the clustering performance and the fairness. Larger λ will lead to a fairer clustering result. Of course, λ should not be too large, or it will dominate the loss function and the kernel k-means may not work. Comparing Eq.(13) with Eq.(2), we observe that if λ is small enough to make $\mathbf{K} - \lambda\mathbf{G}\mathbf{G}^T$ positive semi-definite (p.s.d.), we can regard $\mathbf{K} - \lambda\mathbf{G}\mathbf{G}^T$ as a new kernel matrix and Eq.(13) becomes a standard kernel k-means. In this case, we call $\mathbf{K} - \lambda\mathbf{G}\mathbf{G}^T$ a *fair kernel*.

However, in practice, to make $\mathbf{K}-\lambda\mathbf{G}\mathbf{G}^T$ be a valid kernel matrix, which means to make $\mathbf{K}-\lambda\mathbf{G}\mathbf{G}^T$ p.s.d., we should set a very small λ, which cannot guarantee the fairness. To address this issue, we find that we can add a large enough constant term $\alpha\,\mathrm{Tr}(\mathbf{I})$ to Eq.(13), to obtain a valid fair kernel matrix. In more detail, we have:

$$\mathrm{Tr}\left(\mathbf{Y}^T\left(\mathbf{K}-\lambda\mathbf{G}\mathbf{G}^T\right)\mathbf{Y}\left(\mathbf{Y}^T\mathbf{Y}\right)^{-1}\right)+\alpha\,\mathrm{Tr}(\mathbf{I}) = \mathrm{Tr}\left(\mathbf{Y}^T\left(\mathbf{K}+\alpha\mathbf{I}-\lambda\mathbf{G}\mathbf{G}^T\right)\mathbf{Y}\left(\mathbf{Y}^T\mathbf{Y}\right)^{-1}\right). \quad (14)$$

It shows that optimizing Eq.(14) is always exactly equivalent to optimizing Eq.(13), no matter how we set α. With a large enough α, we can easily set an appropriate λ to make $\tilde{\mathbf{K}} = \mathbf{K} + \alpha\mathbf{I} - \lambda\mathbf{G}\mathbf{G}^T$ be p.s.d., and thus be a valid kernel matrix. We will discuss how to set λ and α later.

In this way, we obtain an extremely simple yet elegant FKKM method. In this method, we do not even need to modify the loss of standard KKM. All we need is to modify the kernel by replacing $\mathbf{K}$ to a fair kernel $\tilde{\mathbf{K}} = \mathbf{K} + \alpha\mathbf{I} - \lambda\mathbf{G}\mathbf{G}^T$. It means that we realize the fairness on the data level rather than the model level.

3.3 Fair Multiple Kernel K-means

Eq.(14) can be naturally extended to a multiple kernel setting. Given a base kernel $\mathbf{K}^{(p)}$, we first construct its fair kernel $\tilde{\mathbf{K}}^{(p)} = \mathbf{K}^{(p)} + \alpha\mathbf{I} - \lambda\mathbf{G}\mathbf{G}^T$. Then similar to Eq.(4), we define the fair consensus kernel $\tilde{\mathbf{K}}^* = \sum_{p=1}^m \gamma_p^2\tilde{\mathbf{K}}^{(p)}$ and take it into Eq.(2) to obtain FMKKM:

$$\min_{\mathbf{Y},\boldsymbol{\gamma}} \mathrm{Tr}\left(\tilde{\mathbf{K}}^*\left(\mathbf{I}-\mathbf{Y}\left(\mathbf{Y}^T\mathbf{Y}\right)^{-1}\mathbf{Y}^T\right)\right) \;\; s.t.\ \mathbf{Y}\in Ind,\ \boldsymbol{\gamma}^T\mathbf{1}=1,\ \gamma_p\geq 0,\ \tilde{\mathbf{K}}^* = \sum_{p=1}^m \gamma_p^2\tilde{\mathbf{K}}^{(p)}. \quad (15)$$

Notice that since our fairness regularization term $\mathrm{Tr}\left(\mathbf{Y}^T\mathbf{G}\mathbf{G}^T\mathbf{Y}\left(\mathbf{Y}^T\mathbf{Y}\right)^{-1}\right)$ requires that $\mathbf{Y}$ should be a discrete indicator matrix, our FKKM (i.e., Eq.(14)) and FMKKM (i.e., Eq.(15)) directly solve the discrete $\mathbf{Y}$ instead of the conventional two-step methods which learn an orthogonal embedding $\mathbf{H}$ first and then obtain the discrete clustering result. As we know, in the two-step methods, the kernel k-means and the discretization post-processing are separated and when doing the discretization it cannot guarantee the clustering accuracy or fairness. Different from the two-step methods, we can directly learn the final clustering result $\mathbf{Y}$ by fully considering the clustering accuracy and fairness.

3.4 Optimization

3.4.1 Optimization of FKKM

When minimizing Eq.(14), we only need to solve one variable $\mathbf{Y}$. Notice that there is only one 1 in each row of $\mathbf{Y}$. Therefore, we can solve $\mathbf{Y}$ row by row. When solving the i-th row, we replace the i-th row with $[1, 0, \cdots, 0]$, $[0, 1, 0, \cdots, 0]$, ..., $[0, \cdots, 0, 1]$ respectively, and compute the values of the corresponding objective function to find the one which leads to the maximum. Then we set the i-th row as this row vector. Wang et al. propose an efficient method to compute these objective functions by reducing the computation redundancy [43].

3.4.2 Optimization of FMKKM

In Eq.(15), there are two groups of variables, i.e., $\mathbf{Y}$ and $\boldsymbol{\gamma}$. We solve them by a block coordinate descent method, which optimizes one variable when fixing the other.

When fixing $\boldsymbol{\gamma}$ to solve $\mathbf{Y}$, we have the following subproblem w.r.t $\mathbf{Y}$:

$$\max_{\mathbf{Y}\in Ind} \mathrm{Tr}\left(\mathbf{Y}^T\tilde{\mathbf{K}}^*\mathbf{Y}\left(\mathbf{Y}^T\mathbf{Y}\right)^{-1}\right), \quad (16)$$

where $\tilde{\mathbf{K}}^* = \sum_{p=1}^m \gamma_p^2\tilde{\mathbf{K}}^{(p)}$. It is the same as the optimization of FKKM.

https://doi.org/10.52202/079017-3162

When fixing $\mathbf{Y}$ to solve $\boldsymbol{\gamma}$, we have following subproblem w.r.t $\boldsymbol{\gamma}$:

$$\min_{\boldsymbol{\gamma}} \sum_{p=1}^{m} \gamma_p^2 h_p, \quad s.t. \sum_{p=1}^{m} \gamma_p = 1, \ \gamma_p \geq 0, \tag{17}$$

where $h_p = \operatorname{Tr}\left(\tilde{\mathbf{K}}^{(p)}\left(\mathbf{I} - \mathbf{Y}\left(\mathbf{Y}^T\mathbf{Y}\right)^{-1}\mathbf{Y}^T\right)\right)$. According to Cauchy-Schwarz Inequality, the closed-form solution of γ_p is:

$$\gamma_p = \frac{h_p^{-1}}{\sum_{j=1}^{m} h_j^{-1}}. \tag{18}$$

Appendix A shows the pseudo-codes of FKKM and FMKKM, respectively. When updating each row of $\mathbf{Y}$, the objective function of FKKM decreases and has a lower bound. Therefore, FKKM can always converge. Similarly, the convergence of FMKKM can also be guaranteed. Now, we analyze the time complexity. According to [43], optimizing the i-th row of $\mathbf{Y}$ has a time complexity of $O(nc)$. FKKM has a time complexity of $O\left(n^2c\right)$. Calculating $\boldsymbol{\gamma}$ has a time complexity of $O(n)$. Therefore, FMKKM also has a time complexity of $O\left(n^2c\right)$. According to [43], although the time complexity is square in the number of instances, it can be computed very efficiently in practice. Therefore, the time complexity of our method is comparable with the mainstream KKM and MKKM methods.

4 Theoretical Analysis

The generalization error bound of the k-means evaluates the expectation of distance between an unseen data and the clustering center it belongs to [30, 22, 21]. Since FKKM is a special case of FMKKM when $m = 1$, in this section, we derive the generalization error bound of our FMKKM. Before the derivation, we need the following two mild assumptions:

Assumption 1 *Each $\tilde{\mathbf{K}}^{(p)} = \mathbf{K}^{(p)} + \alpha\mathbf{I} - \lambda\mathbf{G}\mathbf{G}^T$ is a valid kernel matrix, i.e., $\tilde{\mathbf{K}}^{(p)}$ is symmetric and p.s.d.*

Remark 2 *This assumption is easy to satisfy. If $\tilde{\mathbf{K}}^{(p)}$ is not p.s.d., we can enlarge α to make the assumption hold.*

Assumption 2 *All $\mathbf{K}^{(p)}$ are upper bounded. We denote b as the maximum of elements in all $\mathbf{K}^{(p)}$.*

According to assumption 1, since all $\tilde{\mathbf{K}}^{(p)}$ are valid kernel matrices, $\tilde{\mathbf{K}}^*$ is also a valid kernel matrix. We define the corresponding kernel function of $\tilde{\mathbf{K}}^*$ as $\tilde{\mathcal{K}}^*(\cdot,\cdot)$, and its kernel mapping function is $\Phi_{\boldsymbol{\gamma}}(\mathbf{x}_i) = [\gamma_1\Phi_1(\mathbf{x}_i)^T, \ldots, \gamma_m\Phi_m(\mathbf{x}_i)^T]^T : \mathbb{R}^d \mapsto \mathcal{H}$, where $\Phi_1(\mathbf{x}_i), \ldots, \Phi_m(\mathbf{x}_i)$ are the induced kernel mapping function of $\tilde{\mathbf{K}}^{(1)}, \ldots, \tilde{\mathbf{K}}^{(m)}$, respectively. Let $\mathbf{M} = [\mathbf{m}_1, \ldots, \mathbf{m}_c]$ denote the learned centroids matrix in the RKHS $\mathcal{H}$, where $\mathbf{m}_i$ is the center of the i-th cluster in $\mathcal{H}$. FMKKM aims to minimize the error: $\mathbb{E}\left[\min_{\mathbf{y}\in\{\mathbf{e}_1,\ldots,\mathbf{e}_c\}} \|\Phi_{\boldsymbol{\gamma}}(\mathbf{x}) - \mathbf{M}\mathbf{y}\|_{\mathcal{H}}^2\right]$, where $[\mathbf{e}_1, \ldots, \mathbf{e}_c]$ are the standard orthonormal basis of $\mathbb{R}^c$ space, i.e., $\mathbf{e}_i$ is an all-zero vectors except that the i-th element is 1.

Then, we define a function class as our hypothesis space:

$$\mathcal{F} = \left\{ f : \mathbf{x} \mapsto \min_{\mathbf{y}\in\{\mathbf{e}_1,\ldots,\mathbf{e}_c\}} \|\Phi_{\boldsymbol{\gamma}}(\mathbf{x}) - \mathbf{M}\mathbf{y}\|_{\mathcal{H}}^2 \,\middle|\, \boldsymbol{\gamma}^T\mathbf{1} = 1, \gamma_p \geq 0, \mathbf{m}_k \in \mathcal{H} \right\}. \tag{19}$$

Similar to [21], we have the following Theorem to provide the generalization error bound:

Theorem 2 *Under Assumptions 1 and 2, given training data $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$, function class $\mathcal{F}$ defined in Eq.*(19)*, and any $\delta \geq 0$, with probability at least $1 - \delta$, the following inequality holds for all $f \in \mathcal{F}$:*

$$\begin{aligned}\mathbb{E}[f(\mathbf{x})] \leq & \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i) + \frac{2\sqrt{2\pi}}{\sqrt{n}}\left[(1+c^2)(b+\alpha) - (1+\frac{c^2}{t})\lambda + c\sqrt{2(b+\alpha-\lambda)\left(b+\alpha-\frac{\lambda}{t}\right)}\right] \\ & + \left(4(b+\alpha) - 2\left(1+\frac{1}{t}\right)\lambda\right)\sqrt{\frac{\log(1/\delta)}{2n}},\end{aligned} \tag{20}$$

where t and c are the number of protected groups and clusters, respectively.

Proof 2 *See Appendix B.*

The first term in Eq.(20) is the empirical error. Notice that, we have $\sum_{i=1}^{n} f(\mathbf{x}_i) = \mathrm{Tr}\left(\tilde{\mathbf{K}}^* \left(\mathbf{I} - \mathbf{Y}\left(\mathbf{Y}^T\mathbf{Y}\right)^{-1}\mathbf{Y}^T\right)\right)$, which means our loss function is to minimize exactly this empirical error. However, in the two-step methods, which apply $\mathbf{H}^T\mathbf{H} = \mathbf{I}$ where $\mathbf{H} = \mathbf{Y}\left(\mathbf{Y}^T\mathbf{Y}\right)^{-\frac{1}{2}}$ to replace $\mathbf{Y} \in Ind$, they only optimize a continual approximation of the empirical error.

Besides, the second and third terms represent the gap between the generalization and empirical errors. Intuitively, the gap is the smaller the better. To decrease the gap, we wish α to be as small as possible. However, Assumption 1 prevents α being too small because $\tilde{\mathbf{K}}^{(p)}$ should be p.s.d., or Theorem 2 will not hold anymore. Now we can derive the lower bound of α according to Assumption 1. Suppose σ_{min} as the smallest eigenvalue of $\mathbf{K}^{(1)}, \ldots, \mathbf{K}^{(p)}$. Then, the smallest eigenvalue of $\mathbf{K}^{(p)} + \alpha\mathbf{I}$ should be no smaller than $\sigma_{min} + \alpha$. Notice that we have the following Lemma:

Lemma 1 *Given two real symmetric matrices* $\mathbf{A}$ *and* $\mathbf{B}$ *with the same size, where the smallest eigenvalue of* $\mathbf{A}$ *is* σ_A *and the largest eigenvalue of* $\mathbf{B}$ *is* σ_B*. If* $\sigma_A \geq \sigma_B$*, then* $\mathbf{A} - \mathbf{B}$ *is p.s.d.*

Proof 3 *See Appendix C.*

Denoting σ_{max} as the largest eigenvalue of $\mathbf{G}\mathbf{G}^T$, it is easy to verify that $\sigma_{max} = |\mathcal{G}_{max}|$, where $\mathcal{G}_{max}$ is the protected group with the largest number of instances. According to Lemma 1, we have that if $\sigma_{min} + \alpha - \lambda * |\mathcal{G}_{max}| \geq 0$, $\tilde{\mathbf{K}}^{(p)}$ will be p.s.d. Therefore, α has a lower bound $\lambda * |\mathcal{G}_{max}| - \sigma_{min}$. In practice, σ_{min} is often very small and close to 0. To avoid the time consuming to compute the eigenvalues of the kernels, we can approximately set $\alpha = \lambda * |\mathcal{G}_{max}|$.

Take $\alpha = \lambda * |\mathcal{G}_{max}|$ back into the generalization error bound Eq.(20). We consider the gap between the generalization and empirical errors, i.e., the second and third terms:

$$
\begin{aligned}
&\frac{2\sqrt{2\pi}}{\sqrt{n}}\left[(1+c^2)(b+\alpha) - (1+\frac{c^2}{t})\lambda + c\sqrt{2(b+\alpha-\lambda)\left(b+\alpha-\frac{\lambda}{t}\right)}\right] + \left(4(b+\alpha) - 2\left(1+\frac{1}{t}\right)\lambda\right)\sqrt{\frac{\log(1/\delta)}{2n}} \\
\geq &\frac{2\sqrt{2\pi}}{\sqrt{n}}\left[(1+c^2)b + \left(|\mathcal{G}_{max}| - 1 + \frac{c^2(|\mathcal{G}_{max}|t-1)}{t}\right)\lambda + c\sqrt{2(b+(|\mathcal{G}_{max}|-1)\lambda)\left(b+\frac{|\mathcal{G}_{max}|t-1}{t}\lambda\right)}\right] \\
&+ (4b + 4(|\mathcal{G}_{max}|-1)\lambda)\sqrt{\frac{\log(1/\delta)}{2n}} \qquad (21)
\end{aligned}
$$

Notice that $|\mathcal{G}_{max}| - 1 \geq 0$ and $|\mathcal{G}_{max}|t - 1 \geq 0$, and thus we have that the gap decreases with λ decreases. It means that smaller λ leads to a lower gap. Therefore, λ is a trade-off between the clustering performance and fairness. Increasing λ may enlarge the error bound, but obtain a fairer result. Based on this theoretical analysis, we provide a strategy to set λ by observing a fairness metric, which can be computed without the ground truth. In more detail, we gradually enlarge λ from 0, set $\alpha = \lambda * |\mathcal{G}_{max}|$, and observe the fairness metric. If it gets stable good fairness, we stop enlarging λ and set λ as the current value. This strategy does not need the ground truth, which is appropriate for unsupervised learning, and can obtain an as small as possible λ to achieve a good fairness result.

5 Experiments

5.1 Data Sets and Experimental Setup

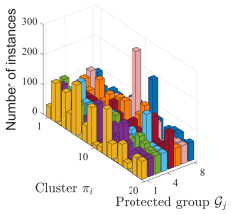
We conduct experiments on benchmark data sets which are widely used in fair clustering, including D&S [2], HAR [3], Jaffe [29], MNIST-USPS [20], Credit Card [52] and K1b [53]. D&S is a human daily and sports activities data set including 8 participants. HAR is a human action recognition data set including 30 participants. In both D&S and HAR data sets, the data of each participant form a protected group. Jaffe is a face image data set. Following [20], the face images with the same expressions are put into a protected group. MNIST-USPS is an image data set containing images of handwritten digits from the subsets of MNIST and USPS data sets. Following [20], we randomly sample 2000 images from MNIST to form one protected group and randomly sample 1800 images from USPS to form the other protected group. Credit card is a data set that describes the customers' default payments and the data of males and females form two protected groups respectively. K1b is a

https://doi.org/10.52202/079017-3162

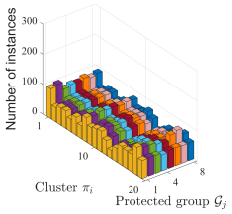
text data set. Following [48], we randomly assign each text to a protected group with a Bernoulli distribution whose $p = 0.5$ to form two protected groups. The statistical information of these data sets is shown in Appendix D.

Table 1: Comparison results on the single kernel setting. The best and second best results are denoted in **bold** and underlined, respectively.

Data sets		K-means	KKM	SC	FairSC	VFC	FFC	FKKM-f	FKKM
D&S	ACC	0.555	0.552	0.558	0.433	0.539	0.521	**0.648**	0.636
	NMI	0.650	0.602	0.652	0.575	0.617	0.583	**0.724**	0.683
	Bal	0	0	0	0	0.186	0.100	0	**0.559**
	MNCE	0.156	0.531	0.023	0	0.923	0.712	0.477	**0.991**
HAR	ACC	0.524	0.620	0.680	0.742	0.600	0.602	0.689	**0.771**
	NMI	0.596	0.609	0.618	0.703	0.654	0.490	0.625	**0.710**
	Bal	0	0	0	0	0.200	0.007	0	**0.250**
	MNCE	0.933	0.930	0.914	0	0.983	0.953	0.920	**0.989**
MNIST-USPS	ACC	0.363	0.396	0.406	**0.458**	0.360	0.437	0.403	0.432
	NMI	0.423	0.421	**0.435**	0.429	0.306	0.412	0.426	0.380
	Bal	0	0	0	0	0.142	0.217	0	**0.847**
	MNCE	0	0.003	0	0	0.544	0.684	0	**0.997**
Jaffe	ACC	0.927	0.948	0.901	0.957	0.981	0.901	0.954	**1**
	NMI	0.914	0.922	0.889	0.943	0.969	0.918	0.930	**1**
	Bal	0	0	0	0	0.400	0.250	0	**0.500**
	MNCE	0.808	0.900	0.765	0.827	0.983	0.924	0.897	**0.989**
Credit Card	ACC	0.362	0.381	0.311	0.351	0.381	0.364	0.400	**0.404**
	NMI	0.139	0.140	0.126	0.123	0.142	0.139	0.145	**0.148**
	Bal	0.510	0.550	0.567	0.603	0.586	0.550	0.536	**0.624**
	MNCE	0.953	0.961	0.967	0.973	0.970	0.969	0.956	**0.985**
K1b	ACC	0.742	0.669	0.667	**0.853**	0.778	0.663	0.826	0.809
	NMI	0.589	0.537	0.536	**0.666**	0.553	0.503	0.628	0.591
	Bal	0.666	0.775	0.763	0.667	0.794	0.773	0.703	**0.800**
	MNCE	0.971	0.989	0.987	0.971	0.990	0.989	0.978	**0.991**



(a) FMKKM-f



(b) FMKKM

Figure 1: Fairness visualization results of FMKKM-f and FMKKM on D&S.

In the single kernel setting, we compare our FKKM with K-means [13], Kernel K-means (KKM) [10], Spectral Clustering (SC) [33], and three state-of-the-art fair clustering methods, including SpFC [18], VFC [58], and FFC [34]. For the kernel methods (i.e., our FKKM and KKM), we use a Gaussian kernel with a bandwidth parameter fixing to $\sqrt{0.5} * D$, where D is the average distance between samples. In the multiple kernel setting, we compare our FMKKM with 9 state-of-the-art MKKM methods, including ONKC [25], MKCSS [55], DPMKKM [42], LFLKA [51], EMKC [38], OSLR [47], ASLR [46], CSAMKC [56], FAMKKM [41]. Detailed information of these compared methods is shown in Appendix E. Besides, for an ablation study, we also compare with the degeneration version of our method, which is without the fairness regularization term, denoted as FKKM-f (for single kernel version) and FMKKM-f (for multiple kernel version).

In the multiple kernel setting, following [11], we construct 12 kernels, including seven Gaussian kernels $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 / 2\epsilon^2\right)$ with $\epsilon = \sqrt{s} * D$, where s varies in the range of $\{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8\}$ and D is the average distance between samples; four polynomial kernels $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \left(a + \mathbf{x}_i^T \mathbf{x}_j\right)^b$ with $a = \{0, 1\}$ and $b = \{2, 4\}$; and a cosine kernel $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \left(\mathbf{x}_i^T \mathbf{x}_j\right) / \left(\|\mathbf{x}_i\| \cdot \|\mathbf{x}_j\|\right)$. Finally, all kernels have been normalized through $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) / \sqrt{\mathbf{K}(\mathbf{x}_i, \mathbf{x}_i)\mathbf{K}(\mathbf{x}_j, \mathbf{x}_j)}$ and then rescaled to $[0, 1]$. We use Accuracy (ACC) and Normalized Mutual Information (NMI) to evaluate the clustering performance. Besides, we also use balance (Bal) [20] and Minimal Normalized Conditional Entropy (MNCE) [49] to evaluate fairness. Specifically, Bal is defined as

$$\mathrm{Bal}(\mathcal{C}) = \min_k \left(\frac{N_k^{\min}}{N_k^{\max}} \right) \in [0, 1], \tag{22}$$

where N_k^{min} and N_k^{max} represent the number of instances in the smallest and the largest (in size) protected groups in cluster π_k, respectively. MNCE is defined as

$$\mathrm{MNCE} = \frac{\min_k \left(-\sum_i \frac{|\mathcal{G}_i \cap \pi_k|}{|\pi_k|} \log \frac{|\mathcal{G}_i \cap \pi_k|}{|\pi_k|} \right)}{-\sum_i \frac{|\mathcal{G}_i|}{n} \log \frac{|\mathcal{G}_i|}{n}} \in [0, 1]. \tag{23}$$

All metrics are the larger the better. Based on previous analysis of hyper-parameter setting, we search λ as $\lambda = 1, 2, \ldots$, by observing the corresponding MNCE. When the MNCE gets stable, i.e., the change of MNCE is smaller than 0.005, we stop the searching and use the current λ. For

Table 2: Comparison results on the multiple kernel setting. The best and second best results are denoted in **bold** and underlined, respectively.

Data sets		ONKC	MKCSS	DPMKKM	LFLKA	EMKC	OSLR	ASLR	CSAMKC	FAMKKM	FMKKM-f	FMKKM
D&S	ACC	0.505	0.543	0.614	**0.646**	0.491	0.590	0.508	0.598	0.601	0.645	0.616
	NMI	0.644	0.665	0.697	0.717	0.602	0.677	0.591	0.668	0.678	**0.718**	0.661
	Bal	0	0	0	0	0	0	0	0	0	0	**0.471**
	MNCE	0.333	0	0.333	0.622	0.501	0.649	0	0.598	0.476	0.585	**0.985**
HAR	ACC	0.526	0.646	0.692	0.695	0.732	0.717	0.574	0.668	0.705	0.697	**0.791**
	NMI	0.557	0.670	0.622	0.622	0.656	0.650	0.549	0.558	0.642	0.655	**0.752**
	Bal	0	0	0	0	0	0	0	0	0	0	**0.263**
	MNCE	0.933	0.920	0.905	0.914	0.939	0.917	0.520	0.928	0.923	0.888	**0.990**
MNIST-USPS	ACC	0.397	0.457	0.391	0.412	0.415	0.406	0.436	0.398	0.445	0.412	**0.495**
	NMI	0.400	0.442	0.359	0.407	0.406	0.406	0.449	0.382	0.402	0.416	**0.454**
	Bal	0	0	0	0	0	0	0	0.024	0	0	**0.808**
	MNCE	0	0	0	0	0	0	0	0.161	0	0	**0.993**
Jaffe	ACC	0.840	0.956	0.939	0.911	0.967	0.934	0.921	0.948	0.985	0.939	**0.995**
	NMI	0.848	0.958	0.924	0.887	0.964	0.903	0.936	0.925	0.971	0.914	**0.991**
	Bal	0	0.200	0	0	0	0	0	0	0.250	0	**0.500**
	MNCE	0.542	0.880	0	0.826	0.964	0.923	0.686	0.917	0.970	0.895	**0.989**
Credit Card	ACC	**0.402**	0.333	0.363	0.360	0.337	0.370	0.321	0.327	0.355	0.378	0.375
	NMI	0.141	0.139	0.126	0.135	0.119	0.138	0.103	0.091	0.123	**0.148**	0.147
	Bal	0.547	0.558	0.523	0.590	0.557	0.599	0.587	0.497	0.571	0.559	**0.641**
	MNCE	0.960	0.964	0.950	0.975	0.963	0.977	0.973	0.938	0.969	0.964	**0.989**
K1b	ACC	0.692	0.688	0.723	0.687	0.601	0.623	**0.850**	0.749	0.745	0.826	0.828
	NMI	0.435	0.535	0.286	0.545	0.436	0.523	**0.652**	0.554	0.581	0.632	0.601
	Bal	0.428	0.794	0.545	0.818	0.881	0.834	0.714	0.892	0.849	0.757	**0.935**
	MNCE	0.881	0.991	0.937	0.993	0.997	0.994	0.980	0.998	0.995	0.986	**1**

other comparison methods, we follow their recommended parameter configurations and search methodologies. All experiments are conducted on the 12th Gen Interl(R) Core(TM) i7-12700 with 32 GB RAM. All experiments are repeated 10 times and the average results are reported.

5.2 Experimental Results

Table 1 shows the comparison results in the single kernel setting, where the best and second best results are denoted in bold and underlined, respectively. It can be seen that FKKM exhibits better fairness compared to K-means, KKM, SC, our ablation version (i.e., FKKM-f), and even the fair clustering methods, indicating the effectiveness of our fairness regularization term. When comparing w.r.t. clustering performance (i.e., ACC and NMI), FKKM still often achieves the best or the second-best results.

Table 2 presents the comparison results in the multiple kernel setting. FMKKM easily achieves the best fairness, due to the effectiveness of our fairness regularization term. Moreover, FMKKM often achieves better or at least comparable ACC and NMI. Notice that our method just simply modifies the original MKKM and can achieve competitive clustering performance, demonstrating that our method is simple yet effective.

Figure 1 shows the visualization results. It shows the number of instances of each protected group $\mathcal{G}_j$ in each cluster π_i in the D&S data set obtained by FMKKM-f and FMKKM, respectively. As shown in Figure 1(a), in FMKKM-f, without the fairness regularization term, the numbers of data of each protected group in each cluster have a great difference, which means the result is unfair. Figure 1 (b) shows that the distribution of protected group in each cluster is more balanced, which means the result obtained by FMKKM is much fairer than FMKKM-f. It demonstrates the effectiveness of our fair regularization term.

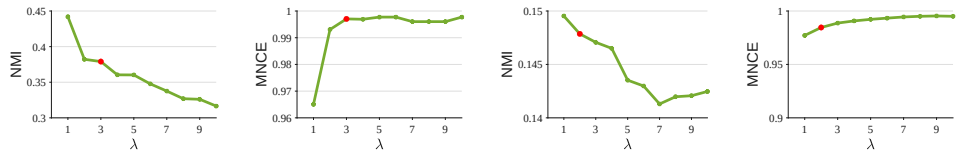
5.3 Efficiency Results

The convergence curves of our methods are shown in Appendix F. The results show that our methods often converge very fast. We also conduct experiments to compare the running time of our methods with other compared methods. Our method is faster than or at least comparable with other methods on many data sets. The detailed results are shown in Appendix F.

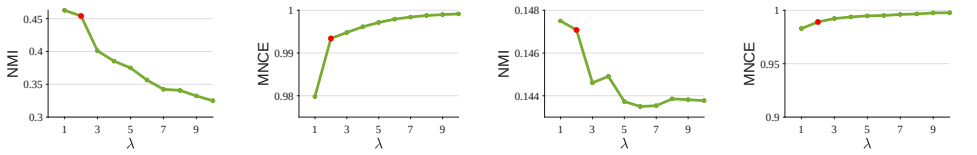
5.4 Parameter Study

Figure 2 shows the effects of λ of FKKM and FMKKM on MNIST-USPS and Credit Cards data sets. The other results are similar. The red points denote the λ selected by our strategy. We can see that

https://doi.org/10.52202/079017-3162



(a) NMI on MNIST-USPS (FKKM) (b) MNCE on MNIST-USPS (FKKM) (c) NMI on Credit Card (FKKM) (d) MNCE on Credit Card (FKKM)



(e) NMI on MNIST-USPS (FMKKM) (f) MNCE on MNIST-USPS (FMKKM) (g) NMI on Credit Card (FMKKM) (h) MNCE on Credit Card (FMKKM)

Figure 2: NMI and MNCE of our methods on MNIST-USPS and Credit Card data sets w.r.t. different values of λ. The red points represent the lambda that our algorithm automatically searches for.

with the increase of λ, the fairness grows and the clustering performance may decrease, which is consistent with our previous discussion. We can often achieve a good trade-off between fairness and performance at the red point, which shows the effectiveness of our hyper-parameter setting strategy.

6 Conclusion

In this paper, we focused on the fairness issue in KKM and MKKM. We carefully designed a novel fairness regularization term, which can be seamlessly plugged into the KKM and MKKM framework. Equipped with this fairness regularization term, we proposed a novel FKKM and FMKKM method. We also provided a hyper-parameter setting strategy based on the theoretical analysis to make the methods easy to use. Extensive experiments demonstrated the effectiveness and superiority of our proposed FKKM and FMKKM methods.

Although the proposed methods achieve promising performance on fairness, they still have some limitations. For example, in our methods, the protected groups must be pre-given or decided by humans. An interesting question is how to automatically decide the protected groups without human intervention. In the future, we will focus on this problem.

Acknowledgments

This work is supported by the National Natural Science Foundation of China grants 62176001 and 62376146, and Natural Science Project of Anhui Provincial Education Department grants 2023AH030004.

References

[1] M. Abbasi, A. Bhaskara, and S. Venkatasubramanian. Fair clustering via equitable group representations. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 504–514, 2021.

[2] K. Altun, B. Barshan, and O. Tunçel. Comparative study on classifying human activities with miniature inertial and magnetic sensors. *Pattern Recognition*, page 3605–3620, Oct 2010.

[3] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones. *The European Symposium on Artificial Neural Networks,The European Symposium on Artificial Neural Networks*, Jan 2013.

[4] A. Backurs, P. Indyk, K. Onak, B. Schieber, A. Vakilian, and T. Wagner. Scalable fair clustering. In *International Conference on Machine Learning*, pages 405–413. PMLR, 2019.

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

[6] S. Bera, D. Chakrabarty, N. Flores, and M. Negahbani. Fair algorithms for clustering. *Advances in Neural Information Processing Systems*, 32, 2019.

[7] I. O. Bercea, M. Groß, S. Khuller, A. Kumar, C. Rösner, D. R. Schmidt, and M. Schmidt. On the cost of essentially fair clusterings. *arXiv preprint arXiv:1811.10319*, 2018.

[8] A. Chhabra, P. Li, P. Mohapatra, and H. Liu. Robust fair clustering: A novel fairness attack and defense framework. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[9] F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii. Fair clustering through fairlets. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5029–5037, 2017.

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

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

[12] M. Ghadiri, S. Samadi, and S. Vempala. Socially fair k-means clustering. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Mar 2021.

[13] G. Hamerly and C. Elkan. Learning the k in k-means. *Advances in neural information processing systems*, 16, 2003.

[14] T. Handhayani and L. Hiryanto. Intelligent kernel k-means for clustering gene expression. *Procedia Computer Science*, 59:171–177, 2015.

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

[16] D. S. Hochba. Approximation algorithms for np-hard problems. *ACM Sigact News*, 28(2):40–52, 1997.

[17] J. Huang, F. Nie, and H. Huang. Spectral rotation versus k-means in spectral clustering. *Proceedings of the AAAI Conference on Artificial Intelligence*, page 431–437, Jun 2022.

[18] M. Kleindessner, S. Samadi, P. Awasthi, and J. Morgenstern. Guarantees for spectral clustering with fairness constraints. In *International conference on machine learning*, pages 3458–3467. PMLR, 2019.

[19] M. Li, Y. Zhang, S. Liu, Z. Liu, and X. Zhu. Simple multiple kernel k-means with kernel weight regularization. *Information Fusion*, 100:101902, 2023.

[20] P. Li, H. Zhao, and H. Liu. Deep fair clustering for visual learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9067–9076, June 2020.

[21] J. Liu, X. Liu, J. Xiong, Q. Liao, S. Zhou, S. Wang, and Y. Yang. Optimal neighborhood multiple kernel clustering with adaptive local kernels. *IEEE Transactions on Knowledge and Data Engineering*, 34(6):2872–2885, 2020.

[22] T. Liu, D. Tao, and D. Xu. Dimensionality-dependent generalization bounds for k-dimensional coding schemes. *Neural computation*, 28(10):2213–2249, 2016.

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

 https://doi.org/10.52202/079017-3162

[24] X. Liu, S. Zhou, L. Liu, C. Tang, S. Wang, J. Liu, and Y. Zhang. Localized simple multiple kernel k-means. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9293–9301, 2021.

[25] X. Liu, S. Zhou, Y. Wang, M. Li, Y. Dou, E. Zhu, and J. Yin. Optimal neighborhood kernel clustering with multiple kernels. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[26] X. Liu, X. Zhu, M. Li, L. Wang, E. Zhu, T. Liu, M. Kloft, D. Shen, J. Yin, and W. Gao. Multiple kernel k k-means with incomplete kernels. *IEEE transactions on pattern analysis and machine intelligence*, 42(5):1191–1204, 2019.

[27] Y. Lu, L. Wang, J. Lu, J. Yang, and C. Shen. Multiple kernel clustering based on centered kernel alignment. *Pattern Recognition*, 47(11):3656–3664, 2014.

[28] Y. Lu, X. Zheng, J. Lu, R. Wang, F. Nie, and X. Li. Self-paced and discrete multiple kernel k-means. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4284–4288, 2022.

[29] M. J. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, and J. Budynek. The japanese female facial expression (jaffe) database. In *Proceedings of third international conference on automatic face and gesture recognition*, pages 14–16, 1998.

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

[31] A. Maurer and M. Pontil. K -dimensional coding schemes in hilbert spaces. *IEEE Trans. Inf. Theory*, 56(11):5839–5846, 2010.

[32] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[33] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14, 2001.

[34] R. Pan and C. Zhong. Fairness first clustering: A multi-stage approach for mitigating bias. *Electronics*, page 2969, Jul 2023.

[35] Y. Pang, J. Xie, F. Nie, and X. Li. Spectral clustering by joint spectral embedding and spectral rotation. *IEEE Transactions on Cybernetics*, page 247–258, Jan 2020.

[36] A. Saxena, G. Fletcher, and M. Pechenizkiy. Fairsna: Algorithmic fairness in social network analysis. *ACM Computing Surveys*, 2022.

[37] M. Schmidt, C. Schwiegelshohn, and C. Sohler. Fair coresets and streaming algorithms for fair k-means. In *Approximation and Online Algorithms: 17th International Workshop, WAOA 2019, Munich, Germany, September 12–13, 2019, Revised Selected Papers 17*, pages 232–251. Springer, 2020.

[38] C. Tang, Z. Li, W. Yan, G. Yue, and W. Zhang. Efficient multiple kernel clustering via spectral perturbation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1603–1611, 2022.

[39] G. Tzortzis and A. Likas. The global kernel k-means clustering algorithm. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1977–1984. IEEE, 2008.

[40] B. Wang and I. Davidson. Towards fair deep clustering with multi-state protected variables. *arXiv preprint arXiv:1901.10053*, 2019.

[41] J. Wang, C. Tang, X. Zheng, X. Liu, W. Zhang, E. Zhu, and X. Zhu. Fast approximated multiple kernel k-means. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[42] R. Wang, J. Lu, Y. Lu, F. Nie, and X. Li. Discrete and parameter-free multiple kernel k-means. *IEEE Transactions on Image Processing*, 31:2796–2808, 2022.

[43] R. Wang, J. Lu, Y. Lu, F. Nie, and X. Li. Discrete and parameter-free multiple kernel k-means. *IEEE Transactions on Image Processing*, page 2796–2808, Jan 2022.

[44] D. Yan, L. Huang, and M. I. Jordan. Fast approximate spectral clustering. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 907–916, 2009.

[45] D. Yin, R. Kannan, and P. Bartlett. Rademacher complexity for adversarially robust generalization. In *International conference on machine learning*, pages 7085–7094. PMLR, 2019.

[46] J. You, Z. Ren, Q. Sun, Y. Sun, and X. Li. Approximate shifted laplacian reconstruction for multiple kernel clustering. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2862–2870, 2022.

[47] J. You, Z. Ren, F. R. Yu, and X. You. One-stage shifted laplacian refining for multiple kernel clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[48] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness constraints: A mechanism for fair classification. *CoRR*, abs/1507.05259, 2015.

[49] P. Zeng, Y. Li, P. Hu, D. Peng, J. Lv, and X. Peng. Deep fair clustering via maximizing and minimizing mutual information: Theory, algorithm and metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 23986–23995. IEEE, 2023.

[50] R. Zhang and A. I. Rudnicky. A large scale clustering scheme for kernel k-means. In *2002 International Conference on Pattern Recognition*, volume 4, pages 289–292. IEEE, 2002.

[51] T. Zhang, X. Liu, L. Gong, S. Wang, X. Niu, and L. Shen. Late fusion multiple kernel clustering with local kernel alignment maximization. *IEEE Transactions on Multimedia*, 25:993–1007, 2021.

[52] L. Zheng, Y. Zhu, and J. He. Fairness-aware multi-view clustering. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, pages 856–864. SIAM, 2023.

[53] S. Zhong and J. Ghosh. Generative model-based document clustering: a comparative study. *Knowl. Inf. Syst.*, 8(3):374–384, 2005.

[54] P. Zhou, L. Du, L. Shi, H. Wang, and Y. Shen. Recovery of corrupted multiple kernels for clustering. In Q. Yang and M. J. Wooldridge, editors, *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 4105–4111. AAAI Press, 2015.

[55] S. Zhou, X. Liu, M. Li, E. Zhu, L. Liu, C. Zhang, and J. Yin. Multiple kernel clustering with neighbor-kernel subspace segmentation. *IEEE transactions on neural networks and learning systems*, 31(4):1351–1362, 2019.

[56] S. Zhou, Q. Ou, X. Liu, S. Wang, L. Liu, S. Wang, E. Zhu, J. Yin, and X. Xu. Multiple kernel clustering with compressed subspace alignment. *IEEE Transactions on Neural Networks and Learning Systems*, 34(1):252–263, 2021.

[57] I. M. Ziko, J. Yuan, E. Granger, and I. B. Ayed. Variational fair clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11202–11209, 2021.

[58] I. M. Ziko, J. Yuan, E. Granger, and I. Ben Ayed. Variational fair clustering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):11202–11209, Sep 2022.

 https://doi.org/10.52202/079017-3162

A Pseudo-codes of FKKM and FMKKM

Algorithms 1 and 2 show the pseudo-codes of our FKKM and FMKKM, respectively.

Algorithm 1 Fair Kernel K-means

Input: Kernel matrix $\mathbf{K}$, protected groups $\mathcal{G}_1, \cdots, \mathcal{G}_t$, fairness hyper-parameter λ.

1: Construct protected group indicator matrix $\mathbf{G}$ and calculate α as $\alpha = |\mathcal{G}_{max}| * \lambda$.
2: Construct the fair kernel by $\tilde{\mathbf{K}} = \mathbf{K} + \alpha\mathbf{I} - \lambda\mathbf{G}\mathbf{G}^T$.
3: Initialize $\mathbf{Y}$ by running standard kernel k-means on $\mathbf{K}$.
4: **repeat**
5: Update $\mathbf{Y}$ row by row by maximize $\mathrm{Tr}\left(\mathbf{Y}^T\tilde{\mathbf{K}}\mathbf{Y}\left(\mathbf{Y}^T\mathbf{Y}\right)^{-1}\right)$.
6: **until** Converges

Output: The final partition matrix $\mathbf{Y}$.

Algorithm 2 Fair Multiple Kernel K-means

Input: Kernel matrices $\{\mathbf{K}^{(p)}\}_{p=1}^m$, protected groups $\mathcal{G}_1, \cdots, \mathcal{G}_t$, fairness hyper-parameter λ.

1: Construct protected group indicator matrix $\mathbf{G}$ and calculate α as $\alpha = |\mathcal{G}_{max}| * \lambda$.
2: Construct the corresponding fair kernel by $\tilde{\mathbf{K}}^{(p)} = \mathbf{K}^{(p)} + \alpha\mathbf{I} - \lambda\mathbf{G}\mathbf{G}^T$ for each base kernel matrix $\mathbf{K}^{(p)}$.
3: Initialize $\boldsymbol{\gamma} = \frac{1}{m}$ and $\mathbf{Y}$ by running standard kernel k-means on $\sum_{p=1}^m \gamma_p^2\mathbf{K}^{(p)}$.
4: **repeat**
5: Update $\mathbf{Y}$ row by row by solving Eq.(16).
6: Update $\boldsymbol{\gamma}$ by Eq.(18)
7: **until** Converges

Output: The final partition matrix $\mathbf{Y}$.

B Proof of Theorem 2

Denote $\hat{R}\left(\mathbf{M}, \boldsymbol{\gamma}\right) = \frac{1}{n}\sum_{i=1}^n \min_{\mathbf{y}\in\{\mathbf{e}_1,\ldots,\mathbf{e}_c\}} \left\|\Phi_{\boldsymbol{\gamma}}\left(\mathbf{x}_i\right) - \mathbf{M}\mathbf{y}\right\|_{\mathcal{H}}^2$. Our goal is to bound:

$$\sup_{f\in\mathcal{F}}\left(\mathbb{E}[f(\mathbf{x})] - \frac{1}{n}\sum_{i=1}^n f\left(\mathbf{x}_i\right)\right). \tag{24}$$

According to Assumption 2, we have the largest value of elements in $\mathbf{K}^{(p)}$ is b. Now, we consider $\tilde{\mathbf{K}}^{(p)} = \mathbf{K}^{(p)} + \alpha\mathbf{I} - \lambda\mathbf{G}\mathbf{G}^T$. Since the diagonal elements of $\mathbf{G}\mathbf{G}^T$ are 1, given any $\mathbf{x}_i$, we have that $\Phi_{\boldsymbol{\gamma}}^T\left(\mathbf{x}_i\right)\Phi_{\boldsymbol{\gamma}}\left(\mathbf{x}_i\right) = \sum_{p=1}^m \gamma_p^2\Phi_p^T\left(\mathbf{x}_i\right)\Phi_p\left(\mathbf{x}_i\right) \leq b + \alpha - \lambda$. Given two different instances $\mathbf{x}_i$ and $\mathbf{x}_j$, if they belong to the same protected group, we have that the (i,j)-th element in $\mathbf{G}\mathbf{G}^T$ is 1, and thus $\Phi_{\boldsymbol{\gamma}}^T\left(\mathbf{x}_i\right)\Phi_{\boldsymbol{\gamma}}\left(\mathbf{x}_j\right) = \sum_{p=1}^m \gamma_p^2\Phi_p^T\left(\mathbf{x}_i\right)\Phi_p\left(\mathbf{x}_j\right) \leq b - \lambda$. If $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to different protected groups, we have that the (i,j)-th element in $\mathbf{G}\mathbf{G}^T$ is 0, and thus $\Phi_{\boldsymbol{\gamma}}^T\left(\mathbf{x}_i\right)\Phi_{\boldsymbol{\gamma}}\left(\mathbf{x}_j\right) = \sum_{p=1}^m \gamma_p^2\Phi_p^T\left(\mathbf{x}_i\right)\Phi_p\left(\mathbf{x}_j\right) \leq b$.

Then, notice that

$$\min_{\mathbf{y}\in\{\mathbf{e}_1,\ldots,\mathbf{e}_c\}} \left\|\Phi_{\boldsymbol{\gamma}}\left(\mathbf{x}_i\right) - \mathbf{M}\mathbf{y}\right\|_{\mathcal{H}}^2 = \min\left\{\left\|\Phi_{\boldsymbol{\gamma}}\left(\mathbf{x}_i\right) - \mathbf{m}_1\right\|_{\mathcal{H}}^2, \cdots, \left\|\Phi_{\boldsymbol{\gamma}}\left(\mathbf{x}_i\right) - \mathbf{m}_c\right\|_{\mathcal{H}}^2\right\}, \tag{25}$$

where $\mathbf{m}_k$ denotes the k-th cluster centroid. Next, we denote $a_i = |\pi_k \cap \mathcal{G}_i|$ to represent the number of instances in all protected groups in cluster π_k. We have:

$$
\begin{aligned}
\left\|\Phi_{\boldsymbol{\gamma}}\left(\mathbf{x}_i\right)-\mathbf{m}_k\right\|_{\mathcal{H}}^2 &= \left\|\Phi_{\boldsymbol{\gamma}}(\mathbf{x}_i)-\frac{1}{|\pi_k|}\sum_{j\in\pi_k}\Phi_{\boldsymbol{\gamma}}\left(\mathbf{x}_j\right)\right\|_{\mathcal{H}}^2 \\
&\leq 2\left(\Phi_{\boldsymbol{\gamma}}^T(\mathbf{x}_i)\Phi_{\boldsymbol{\gamma}}(\mathbf{x}_i)+\frac{1}{|\pi_k|^2}\sum_{\mathbf{x}_p\in\pi_k}\sum_{\mathbf{x}_q\in\pi_k}\Phi_{\boldsymbol{\gamma}}(\mathbf{x}_p)^T\Phi_{\boldsymbol{\gamma}}(\mathbf{x}_q)\right) \\
&\leq 2\left((b+\alpha-\lambda)+\frac{\left(|\pi_k|\right)(b+\alpha-\lambda)+\left(\sum_{i=1}^t a_i^2-|\pi_k|\right)(b-\lambda)+\left(|\pi_k|^2-\sum_{i=1}^t a_i^2\right)b}{|\pi_k|^2}\right) \\
&= 2\left((b+\alpha-\lambda)+\frac{|\pi_k|\alpha-\lambda\sum_{i=1}^t a_i^2+b|\pi_k|^2}{|\pi_k|^2}\right) \\
&\leq 2\left(b+\alpha-\lambda+b-\frac{\lambda}{t}+\frac{\alpha}{|\pi_k|}\right) \\
&\leq 4\left(b+\alpha\right)-2\left(1+\frac{1}{t}\right)\lambda
\end{aligned}
\tag{26}
$$

The second to last inequality holds due to the Cauchy-Schwarz inequality $\sum_{i=1}^t a_i^2 \geq \frac{\left(\sum_{i=1}^t a_i\right)^2}{t}$ and $\sum_{i=1}^t a_i = |\pi_k|$. Therefore, we have:

$$
0 \leq f\left(x_i\right) \leq 4\left(b+\alpha\right)-2\left(1+\frac{1}{t}\right)\lambda. \tag{27}
$$

According to the Theorem 3.1 in [30], by utilizing McDiarmid's inequality, we have that for any $\delta \geq 0$, with probability at least $1-\delta$, for all $f \in \mathcal{F}$, the following inequality holds:

$$
\mathbb{E}[f(\mathbf{x})]-\frac{1}{n}\sum_{i=1}^n f\left(\mathbf{x}_i\right) \leq 2\mathfrak{R}n(\mathcal{F})+\left(4\left(b+\alpha\right)-2\left(1+\frac{1}{t}\right)\lambda\right)\sqrt{\frac{\log(1/\delta)}{2n}}, \tag{28}
$$

where:

$$
\mathfrak{R}_n(\mathcal{F})=\frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^n \sigma_i f\left(\mathbf{x}_i\right)\right], \tag{29}
$$

represents the Rademacher complexity of $\mathcal{F}$ [45]. $\sigma_1,\ldots,\sigma_n$ are Rademacher random variables uniformly distributed on $\{-1,1\}$.

Next, we introduce the Gaussian complexity to provide an upper bound for $\mathfrak{R}_n(\mathcal{F})$ [5]:

$$
\mathfrak{G}_n(\mathcal{F})=\frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^n \beta_i f\left(\mathbf{x}_i\right)\right], \tag{30}
$$

where $\beta_1,\ldots,\beta_n$ are Gaussian random variables with zero mean and unit standard deviation. To bound the Rademacher complexity, we need the following two lemmas:

Lemma 2 *[23]* $\mathfrak{R}_n(\mathcal{F}) \leq \sqrt{\frac{\pi}{2}}\mathfrak{G}_n(\mathcal{F})$.

Lemma 3 *[23] Let* $G_f=\sum_{i=1}^n \beta_i G\left(\mathbf{x}_i,f\right)$ *and* $H_f=\sum_{i=1}^n \beta_i H\left(\mathbf{x}_i,f\right)$ *be two zero-mean separable Gaussian processes. If for all* $f_1, f_2 \in \mathcal{F}$*,*

$$
\mathbb{E}\left[\left(G_{f_1}-G_{f_2}\right)^2\right] \leq \mathbb{E}\left[\left(H_{f_1}-H_{f_2}\right)^2\right], \tag{31}
$$

then we have:

$$
\mathbb{E}\left[\sup_{f\in\mathcal{F}} G_f\right] \leq \mathbb{E}\left[\sup_{f\in\mathcal{F}} H_f\right]. \tag{32}
$$

https://doi.org/10.52202/079017-3162

In our setting, we define:

$$G\left(\mathbf{x}_i, f\right) = G_{\mathbf{M},\boldsymbol{\gamma}} \triangleq \sum_{i=1}^{n} \beta_i \left(\min_{\mathbf{y} \in \{\mathbf{e}_1, \ldots, \mathbf{e}_k\}} \left\| \Phi_{\boldsymbol{\gamma}}\left(\mathbf{x}_i\right) - \mathbf{M}\mathbf{y} \right\|_{\mathcal{H}}^2 \right). \tag{33}$$

Next, we aim to find H_f (i.e., $H_{\mathbf{M},\boldsymbol{\gamma}}$) such that:

$$\mathbb{E}_\beta \left[\left(G_{\mathbf{M}_1,\boldsymbol{\gamma}_1} - G_{\mathbf{M}_2,\boldsymbol{\gamma}_2} \right)^2 \right] \leq \mathbb{E}_\beta \left[\left(H_{\mathbf{M}_1,\boldsymbol{\gamma}_1} - H_{\mathbf{M}_2,\boldsymbol{\gamma}_2} \right)^2 \right]. \tag{34}$$

Specifically, for any $f_1, f_2 \in \mathcal{F}$, we have:

$$\begin{aligned}
&\left(\min_{\mathbf{y}} \left\| \Phi_{\boldsymbol{\gamma}_1}\left(\mathbf{x}_i\right) - \mathbf{M}_1\mathbf{y} \right\|_{\mathcal{H}}^2 - \min_{\mathbf{y}} \left\| \Phi_{\boldsymbol{\gamma}_2}\left(\mathbf{x}_i\right) - \mathbf{M}_2\mathbf{y} \right\|_{\mathcal{H}}^2 \right)^2 \\
\leq &\left(\max_{\mathbf{y}} \left\{ \left\| \Phi_{\boldsymbol{\gamma}_1}\left(\mathbf{x}_i\right) - \mathbf{M}_1\mathbf{y} \right\|_{\mathcal{H}}^2 - \left\| \Phi_{\boldsymbol{\gamma}_2}\left(\mathbf{x}_i\right) - \mathbf{M}_2\mathbf{y} \right\|_{\mathcal{H}}^2 \right\} \right)^2 \\
= &\left(\left(\left\| \Phi_{\boldsymbol{\gamma}_1}\left(\mathbf{x}_i\right) \right\|_{\mathcal{H}}^2 - \left\| \Phi_{\boldsymbol{\gamma}_2}\left(\mathbf{x}_i\right) \right\|_{\mathcal{H}}^2 \right) + \max_{\mathbf{y}} \left\{ 2 \left(\Phi_{\boldsymbol{\gamma}_2}^T\left(\mathbf{x}_i\right) \mathbf{M}_2 - \Phi_{\boldsymbol{\gamma}_1}^T\left(\mathbf{x}_i\right) \mathbf{M}_1 \right) \mathbf{y} + \mathbf{y}^T \left(\mathbf{M}_1^T\mathbf{M}_1 - \mathbf{M}_2^T\mathbf{M}_2 \right) \mathbf{y} \right\} \right)^2 \\
\leq &\left(\left(\left\| \Phi_{\boldsymbol{\gamma}_1}\left(\mathbf{x}_i\right) \right\|_{\mathcal{H}}^2 - \left\| \Phi_{\boldsymbol{\gamma}_2}\left(\mathbf{x}_i\right) \right\|_{\mathcal{H}}^2 \right) + \max_{\mathbf{y}} 2 \left(\Phi_{\boldsymbol{\gamma}_2}^T\left(\mathbf{x}_i\right) \mathbf{M}_2 - \Phi_{\boldsymbol{\gamma}_1}^T\left(\mathbf{x}_i\right) \mathbf{M}_1 \right) \mathbf{y} + \max_{\mathbf{y}} \mathbf{y}^T \left(\mathbf{M}_1^T\mathbf{M}_1 - \mathbf{M}_2^T\mathbf{M}_2 \right) \mathbf{y} \right)^2 \\
= &\left(\left(\left\| \Phi_{\boldsymbol{\gamma}_1}\left(\mathbf{x}_i\right) \right\|_{\mathcal{H}}^2 - \left\| \Phi_{\boldsymbol{\gamma}_2}\left(\mathbf{x}_i\right) \right\|_{\mathcal{H}}^2 \right) + \max_{\mathbf{y}} 2 \sum_{r=1}^{c} y_r \left(\Phi_{\boldsymbol{\gamma}_2}^T\left(\mathbf{x}_i\right) \mathbf{M}_2 - \Phi_{\boldsymbol{\gamma}_1}^T\left(\mathbf{x}_i\right) \mathbf{M}_1 \right) \mathbf{e}_r \right. \\
&\left. + \max_{\mathbf{y}} \sum_{r,s=1}^{c} y_r y_s \mathbf{e}_r^T \left(\mathbf{M}_1^T\mathbf{M}_1 - \mathbf{M}_2^T\mathbf{M}_2 \right) \mathbf{e}_s \right)^2 \\
\leq &4 \left(\left\| \Phi_{\boldsymbol{\gamma}_1}\left(\mathbf{x}_i\right) \right\|_{\mathcal{H}}^2 - \left\| \Phi_{\boldsymbol{\gamma}_2}\left(\mathbf{x}_i\right) \right\|_{\mathcal{H}}^2 \right)^2 + 2 \left(\max_{\mathbf{y}} 2 \sum_{r=1}^{c} y_r \left(\Phi_{\boldsymbol{\gamma}_2}^T\left(\mathbf{x}_i\right) \mathbf{M}_2 - \Phi_{\boldsymbol{\gamma}_1}^T\left(\mathbf{x}_i\right) \mathbf{M}_1 \right) \mathbf{e}_r \right)^2 \\
&+ 4 \left(\max_{\mathbf{y}} \sum_{r,s=1}^{c} y_r y_s \mathbf{e}_r^T \left(\mathbf{M}_1^T\mathbf{M}_1 - \mathbf{M}_2^T\mathbf{M}_2 \right) \mathbf{e}_s \right)^2 \\
\leq &4 \left(\left\| \Phi_{\boldsymbol{\gamma}_1}\left(\mathbf{x}_i\right) \right\|_{\mathcal{H}}^2 - \left\| \Phi_{\boldsymbol{\gamma}_2}\left(\mathbf{x}_i\right) \right\|_{\mathcal{H}}^2 \right)^2 + 8 \sum_{r=1}^{c} \left(\left(\Phi_{\boldsymbol{\gamma}_2}^T\left(\mathbf{x}_i\right) \mathbf{M}_2 - \Phi_{\boldsymbol{\gamma}_1}^T\left(\mathbf{x}_i\right) \mathbf{M}_1 \right) \mathbf{e}_r \right)^2 \\
&+ 4 \sum_{r,s=1}^{c} \left(\mathbf{e}_r^T \left(\mathbf{M}_1^T\mathbf{M}_1 - \mathbf{M}_2^T\mathbf{M}_2 \right) \mathbf{e}_s \right)^2.
\end{aligned} \tag{35}$$

The final two inequalities hold due to $(a+b+c)^2 \leq 4a^2 + 2b^2 + 4c^2$, $\sum_{r=1}^{c} y_r = 1$, and $\sum_{r,s=1}^{c} y_r y_s = 1$. Therefore, combining Eq.(33) and Eq.(35), we have:

$$\begin{aligned}
&\mathbb{E}_\beta \left[\left(G_{\mathbf{M}_1,\boldsymbol{\gamma}_1} - G_{\mathbf{M}_2,\boldsymbol{\gamma}_2} \right)^2 \right] \\
= &\mathbb{E}_\beta \left[\left(\sum_{i=1}^{n} \beta_i \left[\min_{\mathbf{y}} \left\| \Phi_{\boldsymbol{\gamma}_1}\left(\mathbf{x}_i\right) - \mathbf{M}_1\mathbf{y} \right\|_{\mathcal{H}}^2 - \min_{\mathbf{y}} \left\| \Phi_{\boldsymbol{\gamma}_2}\left(\mathbf{x}_i\right) - \mathbf{M}_2\mathbf{y} \right\|_{\mathcal{H}}^2 \right] \right)^2 \right] \\
= &\sum_{i=1}^{n} \left(\min_{\mathbf{y}} \left\| \Phi_{\boldsymbol{\gamma}_1}\left(\mathbf{x}_i\right) - \mathbf{M}_1\mathbf{y} \right\|_{\mathcal{H}}^2 - \min_{\mathbf{y}} \left\| \Phi_{\boldsymbol{\gamma}_2}\left(\mathbf{x}_i\right) - \mathbf{M}_2\mathbf{y} \right\|_{\mathcal{H}}^2 \right)^2 \\
\leq &\sum_{i=1}^{n} \left[4 \left(\left\| \Phi_{\boldsymbol{\gamma}_1}\left(\mathbf{x}_i\right) \right\|_{\mathcal{H}}^2 - \left\| \Phi_{\boldsymbol{\gamma}_2}\left(\mathbf{x}_i\right) \right\|_{\mathcal{H}}^2 \right)^2 + 8 \sum_{r=1}^{c} \left(\left(\Phi_{\boldsymbol{\gamma}_2}^T\left(\mathbf{x}_i\right) \mathbf{M}_2 - \Phi_{\boldsymbol{\gamma}_1}^T\left(\mathbf{x}_i\right) \mathbf{M}_1 \right) \mathbf{e}_r \right)^2 \right. \\
&\left. + 4 \sum_{r,s=1}^{c} \left(\mathbf{e}_r^T \left(\mathbf{M}_1^T\mathbf{M}_1 - \mathbf{M}_2^T\mathbf{M}_2 \right) \mathbf{e}_s \right)^2 \right] \\
= &\mathbb{E}_\beta \left[\left(H_{\mathbf{M}_1,\boldsymbol{\gamma}_1} - H_{\mathbf{M}_2,\boldsymbol{\gamma}_2} \right)^2 \right].
\end{aligned} \tag{36}$$

Then, we obtain $H_{\mathbf{M},\boldsymbol{\gamma}}$ as follows:

$$H_{\mathbf{M},\boldsymbol{\gamma}} = 2 \sum_{i=1}^{n} \beta_i \left\| \Phi_{\boldsymbol{\gamma}}^T\left(\mathbf{x}_i\right) \right\|_{\mathcal{H}}^2 + 2\sqrt{2} \sum_{i=1}^{n} \sum_{r=1}^{c} \beta_{ir} \Phi_{\boldsymbol{\gamma}}^T\left(\mathbf{x}_i\right) \mathbf{M}\mathbf{e}_r + 2 \sum_{i=1}^{n} \sum_{r,s=1}^{c} \beta_{irs} \mathbf{e}_r^T \mathbf{M}^T \mathbf{M} \mathbf{e}_s. \tag{37}$$

To bound the expectation of $H_{\mathbf{M},\boldsymbol{\gamma}}$, we introduce the following Lemma from [31]:

Lemma 4 *[31] Suppose that*

1) $(\mathbf{e}_r : 1 \leq r \leq c)$ *is an orthonormal basis of* $\mathbb{R}^c$ *;*

2) $\mathcal{M}$ *is the class of linear operators* $\mathbf{M} : \mathbb{R}^c \to H$ *with* $\|\mathbf{M}\mathbf{e}_r\|_{\mathcal{H}} \leq \omega$

3) $(\mathbf{x}_i : 1 \leq i \leq n)$ *is a sequence in* $H, \|\mathbf{x}_i\|_{\mathcal{H}} \leq \mu$ *;*

4) $(\beta_{ir} : 1 \leq i \leq n, 1 \leq r \leq c)$ *and* $(\beta_{irs} : 1 \leq i \leq n, 1 \leq r, s \leq r)$ *are orthogaussian (independent and* $N(0,1)$ *) sequences.*

Then the following three inequalities hold:

$$\mathbb{E}_\beta \sup_{\mathbf{M}\in\mathcal{M}} \sum_{i=1}^{n}\sum_{r=1}^{c} \beta_{ir} \langle x_i, \mathbf{M}\mathbf{e}_r\rangle \leq \omega\mu c\sqrt{n}, \tag{38}$$

$$\mathbb{E}_\beta \sup_{\mathbf{M}\in\mathcal{M}} \sum_{i=1}^{n}\sum_{r=1}^{c} \beta_{ir} \|\mathbf{M}\mathbf{e}_r\|_{\mathcal{H}}^2 \leq \omega^2 c\sqrt{n}, \tag{39}$$

$$\mathbb{E}_\beta \sup_{\mathbf{M}\in\mathcal{M}} \sum_{i=1}^{n}\sum_{r,s=1}^{c} \beta_{irs} \langle \mathbf{M}\mathbf{e}_r, \mathbf{M}\mathbf{e}_s\rangle \leq \omega^2 c^2\sqrt{n}, \tag{40}$$

where $\langle\cdot,\cdot\rangle$ *denotes the inner production.*

In our method, given an instance $\mathbf{x}_i$, we have that $\|\Phi_{\boldsymbol{\gamma}}(\mathbf{x}_i)\|_{\mathcal{H}} = \Phi_{\boldsymbol{\gamma}}^T(\mathbf{x}_i)\Phi_{\boldsymbol{\gamma}}(\mathbf{x}_i) \leq b + \alpha - \lambda$. Moreover, we also have $\|\mathbf{M}\mathbf{e}_r\|_{\mathcal{H}} \leq \sqrt{b+\alpha-\frac{\lambda}{t}}$ according to Eq.(26). As a result, according to Lemma 4, the expectation of $H_{\mathbf{M},\boldsymbol{\gamma}}$ can be be bounded as follows,

$$\begin{aligned}
&\mathbb{E}_\beta\left[\sup_{f\in\mathcal{F}} H_{\mathbf{M},\boldsymbol{\gamma}}\right]\\
=&\mathbb{E}_\beta\left[\sup_{f\in\mathcal{F}} 2\sum_{i=1}^{n}\beta_i\left\|\Phi_{\boldsymbol{\gamma}}^T(\mathbf{x}_i)\right\|_{\mathcal{H}}^2 + 2\sqrt{2}\sum_{i=1}^{n}\sum_{r=1}^{c}\beta_{ir}\Phi_{\boldsymbol{\gamma}}^T(\mathbf{x}_i)\mathbf{M}\mathbf{e}_r + 2\sum_{i=1}^{n}\sum_{r,s=1}^{c}\beta_{irs}\mathbf{e}_r^T\mathbf{M}^T\mathbf{M}\mathbf{e}_s\right]\\
\leq&2\mathbb{E}_\beta\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\beta_i\left\|\Phi_{\boldsymbol{\gamma}}^T(\mathbf{x}_i)\right\|_{\mathcal{H}}^2\right] + 2\sqrt{2}\mathbb{E}_\beta\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sum_{r=1}^{c}\beta_{ir}\Phi_{\boldsymbol{\gamma}}^T(\mathbf{x}_i)\mathbf{M}\mathbf{e}_r\right] + 2\mathbb{E}_\beta\left[\sup_{f\in\mathcal{F}}\beta_{irs}\mathbf{e}_r^T\mathbf{M}^T\mathbf{M}\mathbf{e}_s\right]\\
\leq&2(b+\alpha-\lambda)\sqrt{n} + 2c\sqrt{2(b+\alpha-\lambda)\left(b+\alpha-\frac{\lambda}{t}\right)n} + 2c^2\left(b+\alpha-\frac{\lambda}{t}\right)\sqrt{n}\\
=&2\sqrt{n}\left[(1+c^2)(b+\alpha) - (1+\frac{c^2}{t})\lambda + c\sqrt{2(b+\alpha-\lambda)\left(b+\alpha-\frac{\lambda}{t}\right)}\right]
\end{aligned} \tag{41}$$

Last, we can bound $\Re_n(\mathcal{F})$ with Lemma 2, Lemma 3, Eq.(29), Eq.(30), and Eq.(41):

$$\begin{aligned}
\Re_n(\mathcal{F}) \leq \frac{1}{n}\sqrt{\pi/2}\,\mathbb{E}_\beta\left[\sup_{f\in\mathcal{F}} G_{\mathbf{M},\boldsymbol{\gamma}}\right] \leq& \frac{1}{n}\sqrt{\pi/2}\,\mathbb{E}_\beta\left[\sup_{f\in\mathcal{F}} H_{\mathbf{M},\boldsymbol{\gamma}}\right]\\
\leq& \frac{\sqrt{2\pi}}{\sqrt{n}}\left[(1+c^2)(b+\alpha) - (1+\frac{c^2}{t})\lambda + c\sqrt{2(b+\alpha-\lambda)\left(b+\alpha-\frac{\lambda}{t}\right)}\right].
\end{aligned} \tag{42}$$

https://doi.org/10.52202/079017-3162

Substituting Eq.(42) into Eq.(28), we finally obtain for any $\delta \geq 0$, with probability at least $1-\delta$, for all $f \in \mathcal{F}$, the following holds:

$$\mathbb{E}[f(\mathbf{x})] \leq \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i) + \frac{2\sqrt{2\pi}}{\sqrt{n}}\left[(1+c^2)(b+\alpha) - (1+\frac{c^2}{t})\lambda + c\sqrt{2(b+\alpha-\lambda)\left(b+\alpha-\frac{\lambda}{t}\right)}\right] + \left(4(b+\alpha) - 2\left(1+\frac{1}{t}\right)\lambda\right)\sqrt{\frac{\log(1/\delta)}{2n}}. \tag{43}$$

This concludes the proof.

C Proof of Lemma 1

Since σ_A is the smallest eigenvalue of $\mathbf{A}$, and σ_B is the largest eigenvalue of $\mathbf{B}$, we have:

$$\mathbf{A} - \sigma_A \mathbf{I} \succeq 0, \tag{44}$$

$$\sigma_B \mathbf{I} - \mathbf{B} \succeq 0. \tag{45}$$

Summing up Eq.(44) and Eq.(45), we have:

$$\mathbf{A} - \mathbf{B} - (\sigma_A - \sigma_B)\mathbf{I} \succeq 0. \tag{46}$$

Considering the smallest eigenvalue of $\mathbf{A}-\mathbf{B}$, denoting as σ_{A-B}, and its corresponding eigenvector $\mathbf{v}_{A-B}$, we have $(\mathbf{A}-\mathbf{B})\mathbf{v}_{A-B} = \sigma_{A-B}\mathbf{v}_{A-B}$. Multiplying the left-hand side of Eq.(46) with $\mathbf{v}_{A-B}^T$ and $\mathbf{v}_{A-B}$, we have

$$\begin{aligned}
&\mathbf{v}_{A-B}^T(\mathbf{A}-\mathbf{B}-(\sigma_A-\sigma_B)\mathbf{I})\mathbf{v}_{A-B} \\
=&\mathbf{v}_{A-B}^T(\sigma_{A-B}-(\sigma_A-\sigma_B))\mathbf{v}_{A-B} \\
=&(\sigma_{A-B}-(\sigma_A-\sigma_B))\|\mathbf{v}_{A-B}\|_2^2.
\end{aligned} \tag{47}$$

Notice that $(\mathbf{A}-\mathbf{B}-(\sigma_A-\sigma_B)\mathbf{I})$ is positive semi-definite according to Eq.(46), which means $\mathbf{v}_{A-B}^T(\mathbf{A}-\mathbf{B}-(\sigma_A-\sigma_B)\mathbf{I})\mathbf{v}_{A-B} \geq 0$. Therefore, $\sigma_{A-B}-(\sigma_A-\sigma_B) \geq 0$, and thus $\sigma_{A-B} \geq \sigma_A - \sigma_B \geq 0$. This means that $\mathbf{A}-\mathbf{B}$ is positive semi-definite, which concludes the proof.

D Statistical Information of Data Sets

We conduct experiments on benchmark data sets which are widely used in fair clustering, including D&S [2], HAR [3], Jaffe [29], MNIST-USPS [20], Credit Card [52] and K1b [53]. D&S is a human daily and sports activities data set including 8 participants. HAR is a human action recognition data set including 30 participants. In both D&S and HAR data sets, the data of each participant form a protected group. Jaffe is a face image data set. Following [20], the face images with the same expressions are put into a protected group. MNIST-USPS is an image data set containing images of handwritten digits from the subsets of MNIST and USPS data sets. Following [20], we randomly sample 2000 images from MNIST to form one protected group and randomly sample 1800 images from USPS to form the other protected group. Credit card is a data set that describes the customers' default payments and the data of males and females form two protected groups respectively. K1b is a text data set. Following [48], we randomly assign each text to a protected group with a Bernoulli distribution whose $p = 0.5$ to form two protected groups. Details of the data sets are shown in Table 3.

E Introduction of Compared Methods

To show the effectiveness of our method on clustering performance and fairness, we compare our method with some state-of-the-art fair clustering and multiple kernel k-means methods, including:

- **SpFC** [18], which integrates fairness constraints into the Laplacian matrix of a graph.
- **VFC** [58], which is a universal variational fair clustering framework.

Table 3: Description of the data sets.

Data sets	# of Instances	# of Features	# of Cluster	Protected Groups
D&S	9120	5625	19	Person Identity (8)
HAR	10299	561	6	Person Identity (30)
MNIST-USPS	3800	256	10	Source of images (2)
Jaffe	213	676	10	Expression (7)
Credit Card	5000	22	5	Gender (2)
K1b	2340	21839	6	Synthetic Binary (2)

- **FFC** [34], which is a three-stage fair clustering method based on k-means method.
- **ONKC** [25], which is an optimal neighborhood kernel clustering algorithm to enhance the representability of the optimal kernel.
- **MKCSS** [55], which is a simple yet effective neighbor-kernel-based MKC algorithm to consider the intrinsic neighborhood structure among base kernels.
- **DPMKKM** [42], which is a novel discrete multiple kernel k-means by directly solving the clustering indicator matrix.
- **LFLKA** [51], whihc is a simple late fusion multiple kernel clustering with local kernel alignment maximisation approach.
- **EMKC** [38], which is effective multiple kernnel k-means by introducing spectral perturbation theory to laplacian matrix.
- **OSLR** [47], which is a one stage multiple kernel k-means by refining shifted laplacian matrix.
- **ASLR** [46], which is a effective multiple kernel k-means by reconstructing the laplacian matrix.
- **CSAMKC** [56], which is a fast multiple kernel k-means by adopting a novel sampling strategy to improve the performance of MKC.
- **FAMKKM** [41], which is fast and innovative multiple kernel k-means by incorporating two approximated partition matrices instead of the original individual partition matric for each base kernel.

F Efficiency Results

Figures 3 and 4 show the convergence curves of FKKM and FMKKM, respectively. We can see that our methods converge very fast and they often converge within 5 iterations.

Figures 5 and 6 show the running time of all methods on single kernel setting and multiple kernel setting, respectively. For better comparison, we report the logarithm of the time (in seconds). From Figures 5 and 6, we can see that our FKKM and FMKKM are faster than or at least comparable with many state-of-the-art methods, which well demonstrates the efficiency of our methods.

https://doi.org/10.52202/079017-3162

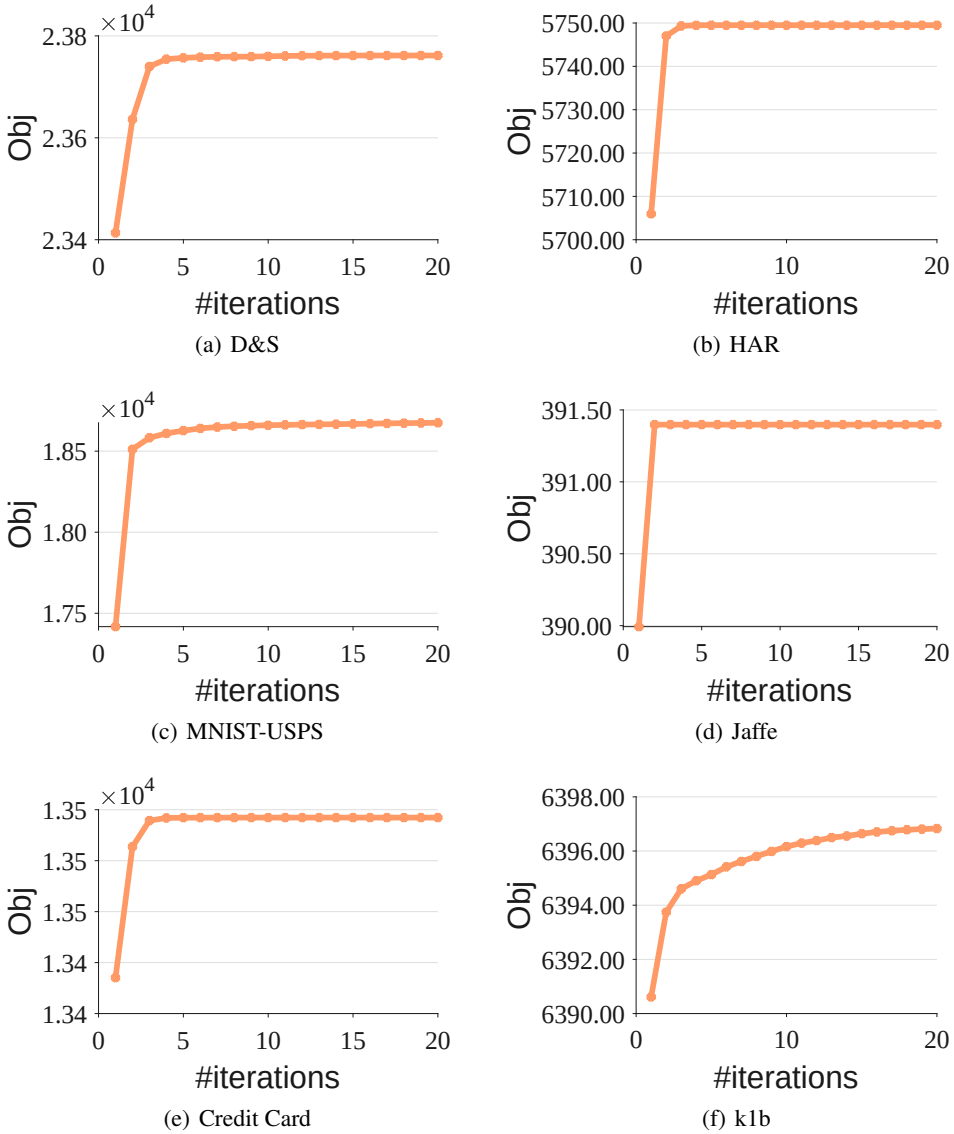


Figure 3: Convergence curves of all data sets on single kernel setting.

https://doi.org/10.52202/079017-3162

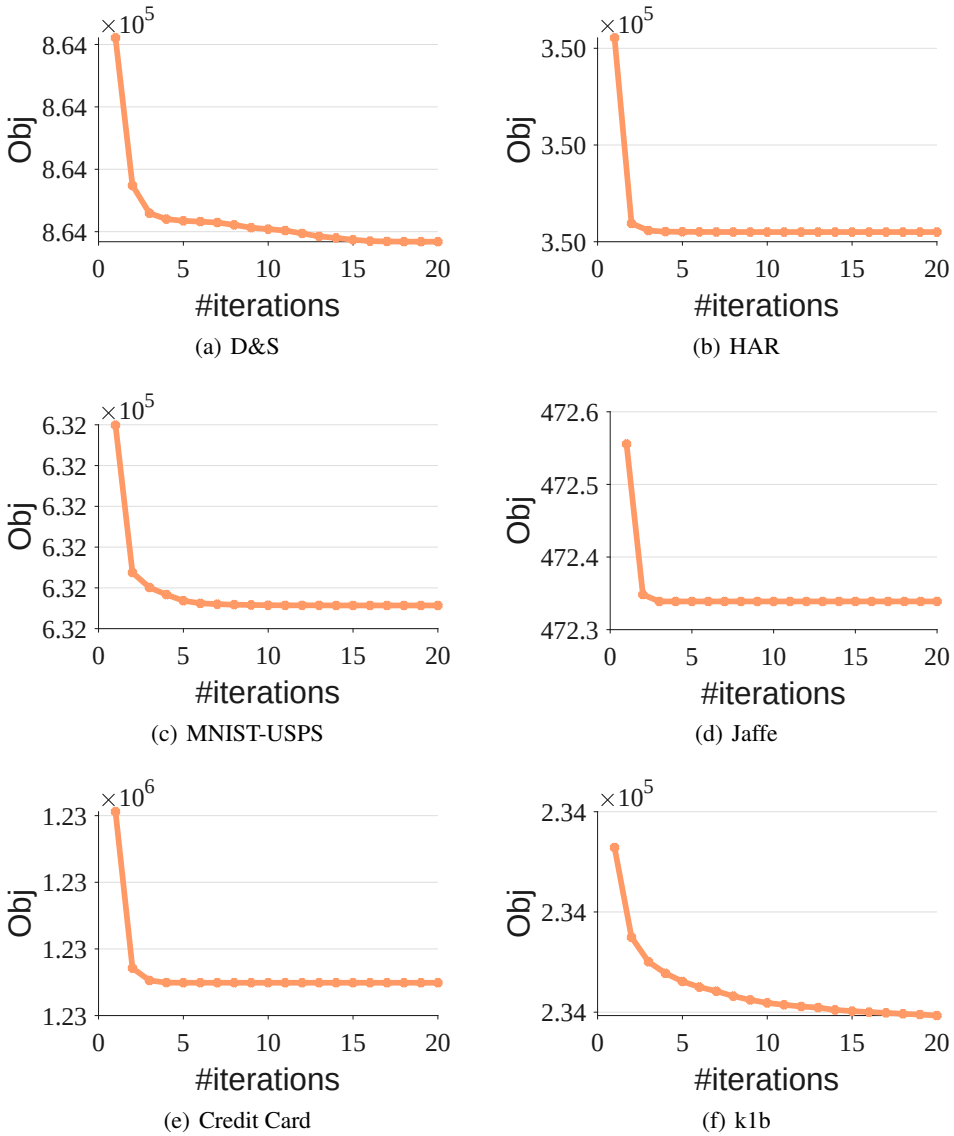
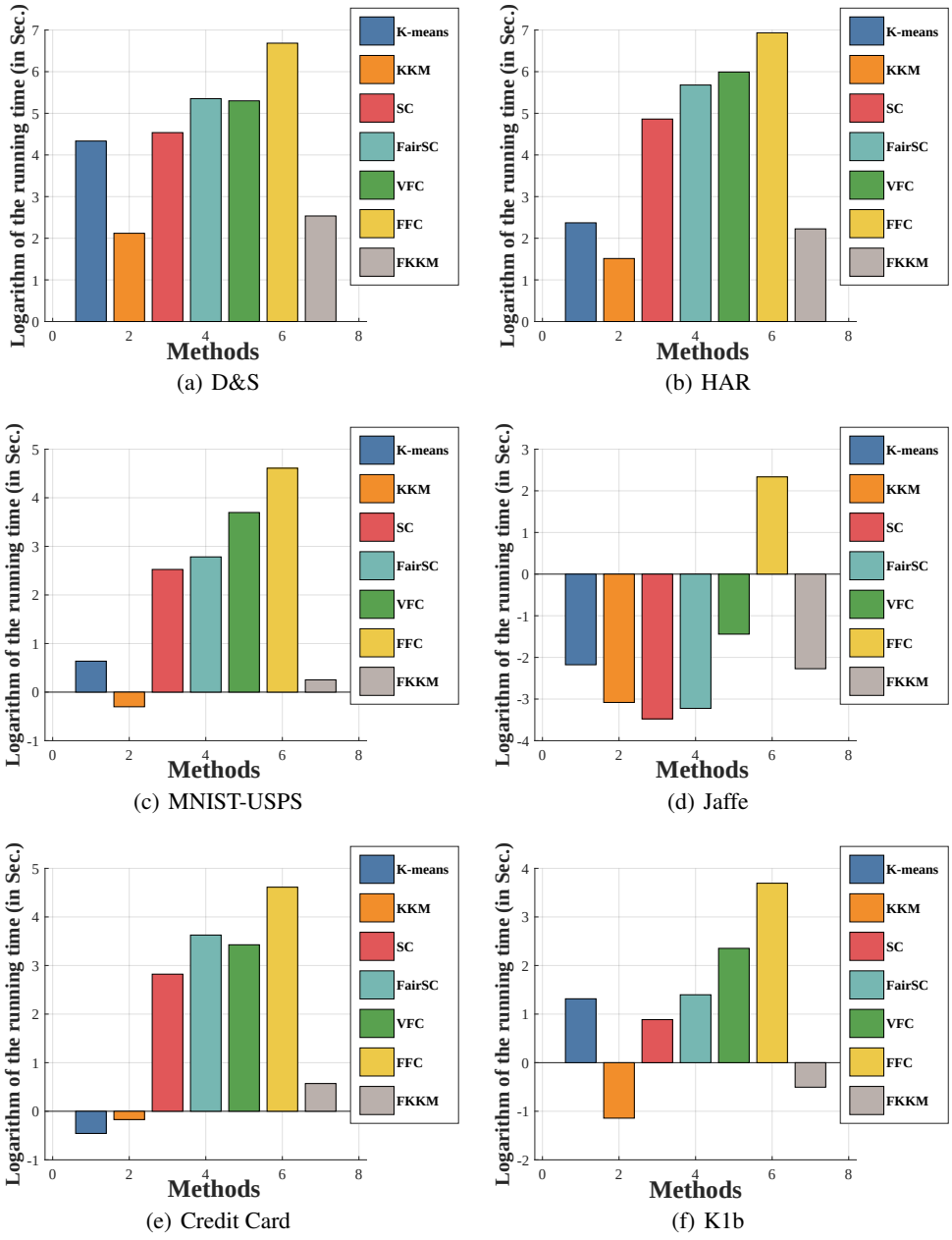


Figure 4: Convergence curves of all data sets on multiple kernel setting.

https://doi.org/10.52202/079017-3162



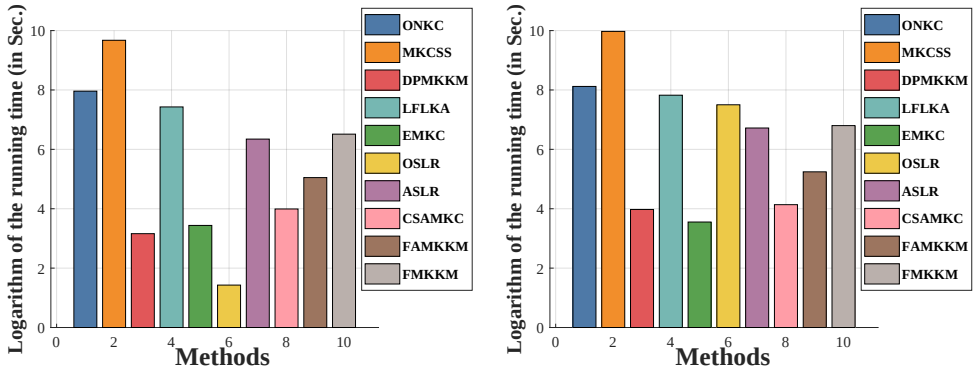
(a) D&S (b) HAR

(c) MNIST-USPS (d) Jaffe

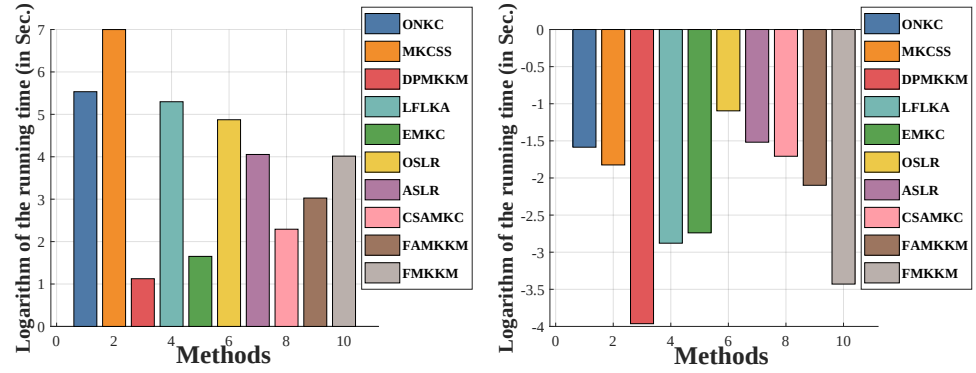
(e) Credit Card (f) K1b

Figure 5: Running time of all methods on the single kernel setting.

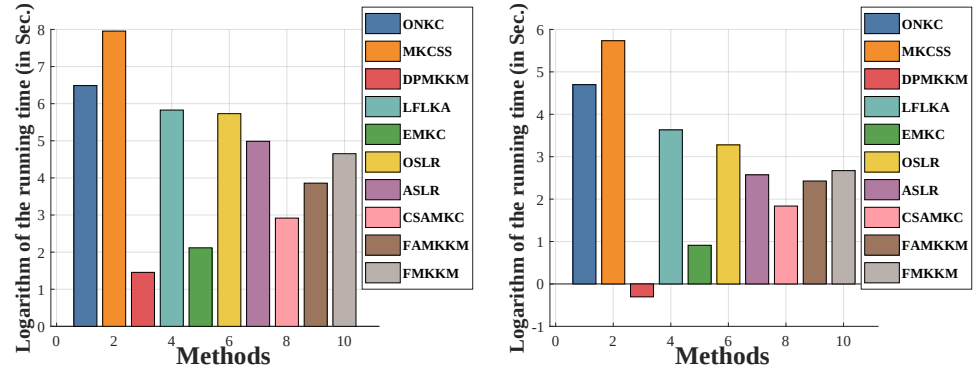
https://doi.org/10.52202/079017-3162



(a) D&S (b) HAR



(c) MNIST-USPS (d) Jaffe



(e) Credit Card (f) K1b

Figure 6: Running time of all methods on the multiple kernel setting.

https://doi.org/10.52202/079017-3162

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in Section Conclusion.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide the proofs of all theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the detailed experimental settings and the codes.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

https://doi.org/10.52202/079017-3162

Answer: [Yes]

Justification: We provide the codes.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide the detailed experimental setting.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We have computed the standard deviation but due to the space limit, we do not report this in the manuscript. We can provide it if requested.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the detailed experimental setting.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We conform with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: We do not find any societal impact of this work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

https://doi.org/10.52202/079017-3162

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All used assets are cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

 https://doi.org/10.52202/079017-3162