

# Imbalanced Data Clustering using Equilibrium K-Means

Yudong He

**Abstract**—Traditional centroid-based clustering algorithms, such as hard K-means (HKM, or Lloyd’s algorithm) and fuzzy K-means (FKM, or Bezdek’s algorithm), display degraded performance when true underlying groups of data have varying sizes (i.e., imbalanced data). This paper introduces equilibrium K-means (EKM), a novel fuzzy clustering algorithm that has the robustness to imbalanced data by preventing centroids from crowding together in the center of large clusters. EKM is simple, alternating between two steps; fast, with the same time and space complexity as FKM; and scalable to large datasets. We evaluate the performance of EKM on two synthetic and ten real datasets, comparing it to other centroid-based algorithms, including HKM, FKM, maximum-entropy fuzzy clustering (MEFC), two FKM variations designed for imbalanced data, and the Gaussian mixture model. The results show that EKM performs competitively on balanced data and significantly outperforms other algorithms on imbalanced data. Deep clustering experiments on the MNIST dataset demonstrate the significance of making representation have an EKM-friendly structure when dealing with imbalanced data; In comparison to deep clustering with HKM, deep clustering with EKM obtains a more discriminative representation and a 35% improvement in clustering accuracy. Additionally, we reformulate HKM, FKM, MEFC, and EKM in a general form of gradient descent, where fuzziness is introduced differently and more simply than in Bezdek’s work, and demonstrate how the general form facilitates a uniform study of KM algorithms.

**Index Terms**—K-means, fuzzy clustering, imbalance learning, deep clustering

## I. INTRODUCTION

Imbalanced data refers to the true underlying groups of data having different sizes, which is common in datasets of medical diagnosis, fraud detection, and anomaly detection. Imbalanced data poses a challenge for learning algorithms because these algorithms tend to be biased towards the majority group [1]. While there is a considerable amount of research on supervised learning (e.g., classification) from imbalanced data [2]–[4], unsupervised learning has not been as thoroughly explored, because the unknown cluster sizes make the task more difficult [5]. Methods like resampling and boosting frequently used in supervised learning cannot be applied in unsupervised learning.

Clustering is an important unsupervised learning task involving grouping data into clusters based on similarity. K-means (KM) is the most popular clustering technique, valued for its simplicity, scalability, and effectiveness with real datasets. It can also be used as an initialization method for more advanced clustering techniques, such as the Gaussian

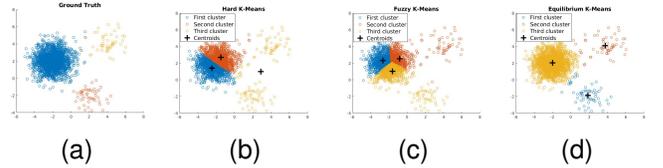


Fig. 1. Clustering results of a highly imbalanced dataset. (a) Ground truth. The colors represent the labels of the data points. (b) Clustering by hard K-means. (c) Clustering by fuzzy K-means. (d) Clustering by the proposed equilibrium K-means. They are all two-step alternating algorithms that iteratively compute centroids.

mixture model (GMM) [6], [7]. KM starts with an initial set of centroids (cluster centers) and iteratively refines them to increase cluster compactness. The hard KM (HKM, or Lloyd’s algorithm) [8], [9] and fuzzy KM (FKM, or Bezdek’s algorithm) [10] are the two most representative KM algorithms. HKM assigns a data point to only one cluster, while FKM assigns a data point to multiple clusters with varying degrees of membership.

Research on HKM and FKM has been ongoing. For example, the possibilistic FKM (PFKM) [11] was proposed to address the sensitivity of FKM to noise and outliers. FKM- $\sigma$  [12] was proposed to improve FKM’s performance on data points with uneven variations or non-spherical shapes in individual clusters. Fuzzy local information K-means (FLIKM) [13] was designed to promote FKM’s performance in image segmentation. Deep clustering, a technique that combines deep neural networks (DNNs) with HKM and FKM, was proposed to cluster high-dimensional data [14]–[17].

Although many variations based on HKM and FKM have been developed to deal with different situations, most of them display degraded effectiveness in cases of imbalanced data. This is due to the so-called “uniform effect” that causes the clusters generated by these algorithms to have similar sizes even when input data has highly varying group sizes [18]. An illustration of the uniform effect of HKM and FKM is given in Fig. 1 where we can observe that the centroids of HKM and FKM crowd together in the large true cluster.

### A. Existing Efforts to Overcome Uniform Effect

To the best of our knowledge, there are two popular methods to overcome the uniform effect in the field of fuzzy systems. The first method is to introduce more weight on the data points in small clusters at each updating iteration, biasing learning towards them, like a modified FKM called cluster-size insensitive FKM (csiFKM) developed by Noordam et al. [19]. Later, Lin et al. [20] proposed a size-insensitive integrity-based

Y. He is with the Department of Industry Engineering and Decision Analytics, The Hong Kong University of Science and Technology, Hong Kong (e-mail: yhebh@connect.ust.hk).

FKM (siibFKM) based on csiFKM to reduce the sensitivity of csiFKM to the distance between adjacent clusters. However, weighting based on cluster size inadvertently increases the influence of outliers, making these algorithms sensitive to noise [21].

The second method is called multiprototype clustering. This method first groups data into multiple subclusters with similar sizes and the final clusters are obtained by merging adjacent subclusters. Liang et al. [22] proposed a multiprototype clustering algorithm that employs FKM to generate subclusters. Later, Lu et al. [5] proposed a self-adaptive multiprototype clustering algorithm that automatically adjusts the number of subclusters. However, multiprototype clustering algorithms have a complex process and high time complexity of  $O(N^2)$ , where  $N$  is the number of data points in the dataset. Thus, they are computationally expensive for large datasets. We should additionally mention that Zeng et al. [23] recently proposed a soft multiprototype clustering algorithm with time complexity linear to  $N$ . However, their clustering process remains complex and is aimed at clustering high-dimensional and complex-structured data rather than imbalanced data.

### B. Our Contributions

In this paper, we propose equilibrium K-means (EKM) to address the issue of imbalanced data clustering and the aforementioned algorithms' limitations. Our contributions can be summarized in three main aspects.

First, we reformulate HKM, FKM, the maximum-entropy fuzzy clustering (MEFC) [24], [25], and EKM in a general form of gradient descent. We show that these algorithms aim to optimize different approximations of the same objective. This general form facilitates the uniform study of KM algorithms.

Second, we develop EKM based on the first contribution. EKM belongs to the family of fuzzy clustering and membership defined in EKM has a clear physical meaning. Repulsive forces appear among centroids of EKM, successfully reducing the uniform effect by preventing centroids from crowding together in a large cluster (see Fig. 1d for an example). Therefore, EKM is effective in dealing with imbalanced data and is not sensitive to noise. Similar to HKM and FKM, EKM is a two-step alternating algorithm that iteratively computes centroids. In addition to having the same time ( $O(N)$ ) and space complexity as FKM, EKM has a batch-learning version that can be applied to large datasets. We demonstrate the effectiveness of EKM on balanced and imbalanced data by conducting numerical experiments on two synthetic and ten real datasets from different domains. Due to space limitations, quantitative performance metrics are provided in the supplementary material. The scatter plots of partial clustering results we provide in the paper are sufficient to demonstrate the superiority of EKM on imbalanced data.

Finally, we investigate the joint learning of DNNs and EKM. We find that mapping high-dimensional data via DNNs to an EKM-friendly space can result in more discriminative low-dimensional representation than mapping to an HKM-friendly space. When tested on an imbalanced dataset derived from MNIST, joint learning of DNNs and EKM improves clustering

accuracy by 35% compared to joint learning of DNNs and HKM.

### C. Organization

We introduce HKM, FKM, and MEFC in Section II. In Section III, we derive their general form and the proposed EKM. The properties of EKM are studied in Section IV. We evaluate the performance of EKM on classic clustering tasks in Section V and on deep clustering in Section VI. Finally, we conclude in Section VII.

## II. K-MEANS AND ITS VARIATIONS

### A. The Hard K-Means Algorithm (Lloyd's Algorithm)

KM aims to partition  $N$  data points into  $K$  clusters, minimizing the sum of the variances within each cluster. Mathematically, the objective can be expressed as:

$$\arg \min_{\mathbb{S}_1, \dots, \mathbb{S}_K} \sum_{k=1}^K \sum_{\mathbf{x} \in \mathbb{S}_k} \|\mathbf{x} - \boldsymbol{\mu}_k\|_2^2, \quad (1)$$

where  $\mathbb{S}_k$  represents the set of data points in the  $k$ -th cluster,  $\mathbf{x} \in \mathbb{S}_k$  denotes a data point belonging to  $\mathbb{S}_k$ ,  $\boldsymbol{\mu}_k$  is the centroid of  $\mathbb{S}_k$ , expressed by

$$\boldsymbol{\mu}_k = \frac{1}{|\mathbb{S}_k|} \sum_{\mathbf{x} \in \mathbb{S}_k} \mathbf{x}, \quad (2)$$

$|\mathbb{S}_k|$  signifies the size of  $\mathbb{S}_k$ , and  $\|\cdot\|_2$  is the  $l_2$  norm. This optimization problem is NP-hard [26] and is commonly solved heuristically by Lloyd's algorithm [9]. Given  $K$  points  $(\mathbf{c}_1^{(1)}, \dots, \mathbf{c}_K^{(1)})$  as initial centroids, Lloyd's algorithm alternates between two steps:

- 1) Assign each data point to the nearest centroid, forming  $K$  clusters:

$$\mathbb{S}_k^{(\tau)} = \{\mathbf{x} : \|\mathbf{x} - \mathbf{c}_k^{(\tau)}\|_2^2 \leq \|\mathbf{x} - \mathbf{c}_i^{(\tau)}\|_2^2 \forall i, 1 \leq i \leq K\} \quad (3)$$

- 2) Recalculate the centroid of each cluster by taking the mean of all data points assigned to it:

$$\mathbf{c}_k^{(\tau+1)} = \frac{1}{|\mathbb{S}_k^{(\tau)}|} \sum_{\mathbf{x} \in \mathbb{S}_k^{(\tau)}} \mathbf{x}. \quad (4)$$

Lloyd's algorithm converges when the assignment of data points to clusters ceases to change, or when a maximum number of iterations is reached. The time complexity of one iteration of the above two steps is  $O(NK)$ . The HKM algorithm mentioned in this paper refers to Lloyd's algorithm.

### B. The Uniform Effect of K-Means

The uniform effect refers to the propensity to generate clusters of similar sizes, which is implicitly implied in the objective (1) of KM. For simplicity, kindly consider a case with two clusters (i.e.,  $K = 2$ ). Minimizing the objective function (1) is equivalent to maximizing the following objective function [27]:

$$\max_{\mathbb{S}_1, \mathbb{S}_2} N_1 N_2 \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2^2, \quad (5)$$

where  $N_1$  and  $N_2$  denote the sizes of  $\mathbb{S}_1$  and  $\mathbb{S}_2$ , respectively. By isolating the effect of  $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2^2$ , maximizing the above objective leads to  $N_1 = N_2 = N/2$ , indicating KM tends to produce equally sized clusters.

### C. The Fuzzy K-Means Algorithm (Bezdek's Algorithm)

FKM attempts to minimize the sum of weighted distances between data points and centroids, with the following objective and constraints [10]:

$$\begin{aligned} \min_{\mathbf{c}_1, \dots, \mathbf{c}_K, \{u_{kn}\}} & \sum_{n=1}^N \sum_{k=1}^K (u_{kn})^m \|\mathbf{x}_n - \mathbf{c}_k\|_2^2 \\ \text{subject to} & u_{kn} \in [0, 1] \forall k, n, \\ & \sum_{k=1}^K u_{kn} = 1 \forall n, \\ & 0 < \sum_{n=1}^N u_{kn} < N \forall k, \end{aligned} \quad (6)$$

where  $\mathbf{x}_n$  represents the  $n$ -th data point,  $\mathbf{c}_k$  denotes the  $k$ -th centroid,  $u_{kn}$  is a coefficient called membership that indicates the degree of  $\mathbf{x}_n$  belonging to the  $k$ -th cluster, and  $m \in (1, +\infty)$  is a hyperparameter controlling the degree of fuzziness level. Similar to the HKM algorithm, the FKM algorithm operates by alternating two steps:

- 1) Calculate the membership value of the  $n$ -th data point belonging to the  $k$ -th cluster:

$$u_{kn}^{(\tau)} = \frac{1}{\sum_{i=1}^K \left( \frac{\|\mathbf{x}_n - \mathbf{c}_k^{(\tau)}\|_2}{\|\mathbf{x}_n - \mathbf{c}_i^{(\tau)}\|_2} \right)^{\frac{2}{m-1}}}. \quad (7)$$

- 2) Recalculate the weighted centroid of the  $k$ -th cluster by:

$$\mathbf{c}_k^{(\tau+1)} = \frac{\sum_n (u_{kn}^{(\tau)})^m \mathbf{x}_n}{\sum_n (u_{kn}^{(\tau)})^m}. \quad (8)$$

The time complexity of one iteration of the above two steps is  $O(NK^2)$ . The higher time complexity of FKM than HKM is due to the extra membership calculation.

### D. Maximum-Entropy Fuzzy Clustering

Karayiannis [24] added an entropy term to the objective function of FKM, resulting in MEFC. The new objective function is given as follows:

$$\begin{aligned} \min_{\mathbf{c}_1, \dots, \mathbf{c}_K, \{u_{kn}\}} & \eta \sum_{n=1}^N \sum_{k=1}^K u_{kn} \ln u_{kn} \\ & + (1 - \eta) \frac{1}{N} \sum_{n=1}^N \sum_{k=1}^K u_{kn} \|\mathbf{x}_n - \mathbf{c}_k\|_2^2 \\ \text{subject to} & u_{kn} \in [0, 1] \forall k, n, \\ & \sum_{k=1}^K u_{kn} = 1 \forall n, \end{aligned} \quad (9)$$

where  $\eta \in (0, 1)$  is a hyperparameter, controlling the transition from maximization of the entropy to the minimization of

centroid-data distances. MEFC is similar to FKM but with a different definition of membership:

$$u_{kn}^{(\tau)} = \frac{\exp(-\lambda \|\mathbf{x}_n - \mathbf{c}_k^{(\tau)}\|_2^2)}{\sum_{i=1}^K \exp(-\lambda \|\mathbf{x}_n - \mathbf{c}_i^{(\tau)}\|_2^2)}, \quad (10)$$

where  $\lambda = \frac{1}{N} \frac{1-\eta}{\eta}$ . The time complexity of MEFC is the same as FKM, which is  $O(NK^2)$  for one iteration.

## III. SMOOTH K-MEANS - A UNIFIED FRAMEWORK

### A. Objective of Smooth K-Means

We propose a novel framework called smooth K-means (SKM) and demonstrate that the three KM algorithms introduced in the previous section are special cases of SKM.

Denote the squared Euclidean distance<sup>1</sup> between the  $k$ -th centroid and the  $n$ -th data point as

$$d_{kn} = \frac{1}{2} \|\mathbf{x}_n - \mathbf{c}_k\|_2^2, \quad (11)$$

and define the within-cluster sum of squares (WCSS) as the sum of squared Euclidean distances between data points and their nearest centroids, i.e.,

$$\text{WCSS} := \sum_{n=1}^N \min(d_{1n}, \dots, d_{Kn}). \quad (12)$$

The goal of SKM is to find  $K$  centroids that minimize an approximated WCSS, resulting in the following optimization problem:

$$\arg \min_{\mathbf{c}_1, \dots, \mathbf{c}_K} \sum_{n=1}^N h(d_{1n}, \dots, d_{Kn}), \quad (13)$$

where  $h(d_{1n}, \dots, d_{Kn})$  is a smooth approximation to  $\min(d_{1n}, \dots, d_{Kn})$  and is referred to as the smooth minimum function. Below we reveal the relationship between the SKM and the GMM.

### B. The Relationship Between SKM and GMM

The centroids obtained by minimizing WCSS are maximum likelihood estimators (MLEs) of parameters of "hard" GMM. The derivation is as follows. Assuming that the dataset  $(\mathbf{x}_1, \dots, \mathbf{x}_N)$  is sampled from  $K$  independent multivariate normal distributions. We denote the mean vector and covariance matrix of the  $k$ -th normal distribution as  $\boldsymbol{\mu}_k$  and  $\boldsymbol{\Sigma}_k$ , respectively. The standard GMM has the following well-known likelihood

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}, p | \mathbf{x}) & \propto \prod_{n=1}^N \sum_{k=1}^K p_{kn} \det(\boldsymbol{\Sigma}_k)^{-1/2} \\ & \exp\left(-\frac{1}{2} (\mathbf{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k)\right), \end{aligned} \quad (14)$$

where  $p_{kn}$  is the probability that the  $n$ -th data point is generated from the  $k$ -th normal distribution. The standard GMM assumes that data points are generated from multiple

<sup>1</sup>Other distance metrics, such as the absolute difference and the angle between points, can also be used to define  $d_{kn}$ . Considering the derivation process is the same, we use the Euclidean distance in the paper for simplicity.

normal distributions based on a certain probability distribution. On the other hand, the hard GMM assumes that a data point is generated from a single normal distribution. Accordingly, for all  $n$ ,  $p_{kn} = 1$  for one  $k \in \{1, \dots, N\}$  and  $p_{in} = 0$  if  $i \neq k$ . We further assume that all normal distributions have the same covariance matrix, specifically an identity matrix. Under these assumptions, the MLEs of  $\{p_{kn}\}$  are  $\hat{p}_{kn} = 1$  if  $k = \arg \min_{i \in \{1, \dots, K\}} \|\mathbf{x}_n - \boldsymbol{\mu}_i\|_2^2$ , and  $\hat{p}_{kn} = 0$  otherwise. By substituting  $\boldsymbol{\Sigma}_k = \mathbf{I}$  and  $\hat{p}_{kn}$  into (14) and taking the logarithmic value, we obtain the log-likelihood:

$$l(\boldsymbol{\mu} | \mathbf{x}) \propto - \sum_{n=1}^N \min(\|\mathbf{x}_n - \boldsymbol{\mu}_1\|_2^2, \dots, \|\mathbf{x}_n - \boldsymbol{\mu}_K\|_2^2). \quad (15)$$

Now it is clear that the MLEs of  $\{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K\}$  are the centroids minimizing WCSS.

The hard GMM model simplifies the standard GMM by only considering the impact of a data point on its closest centroid. However, as seen in the success of FKM, it can be more advantageous to consider the impact of a data point on all centroids. This can be accomplished in SKM by applying an approximation function to smooth WCSS. Hence, SKM can be viewed as a model between hard and standard GMM. In the following sections, we will introduce three common smooth minimum functions and explore the resulting distinct clustering algorithms.

### C. Three Common Smooth Minimum Functions

Assume a monotonically increasing and differentiable function  $f: [0, +\infty) \mapsto [0, +\infty)$  satisfies

$$\lim_{x \rightarrow +\infty} \frac{x}{f(x)} \rightarrow 0, \quad (16)$$

or equivalently

$$\lim_{x \rightarrow +\infty} \frac{1}{f'(x)} \rightarrow 0, \quad (17)$$

where  $f'$  is the first derivative of  $f$ . Let  $g(x) = 1/f(x)$ , a smooth minimum function  $h_1: [0, +\infty)^K \mapsto [0, +\infty)$  can be constructed by

$$h_1(x_1, \dots, x_K) = g^{-1}(g(x_1) + \dots + g(x_K)). \quad (18)$$

Let  $f(x)$  be defined as  $e^{\lambda x}$ , then the function  $h_1$  takes on a specific form known as LogSumExp:

$$\begin{aligned} h_1(x_1, \dots, x_K) &= \text{LSE}_\lambda(x_1, \dots, x_K) \\ &= -\frac{1}{\lambda} \ln(e^{-\lambda x_1} + \dots + e^{-\lambda x_K}), \end{aligned} \quad (19)$$

where  $\lambda$  is a parameter controlling the degree of approximation, with  $\text{LSE}_\lambda \rightarrow \min$  as  $\lambda \rightarrow +\infty$ .

Define  $f(x) = x^p$ , we can have another common smooth minimum function, called p-Norm, which has the following specific form

$$\begin{aligned} h_1(x_1, \dots, x_K) &= \text{PN}_p(x_1, \dots, x_K) \\ &= (x_1^{-p} + \dots + x_K^{-p})^{-1/p}, \end{aligned} \quad (20)$$

and converges to  $\min(x_1, \dots, x_K)$  as  $p \rightarrow +\infty$ . An additional definition is necessary for the domain of the p-Norm function

to legally include zeros:  $\text{PN}_p(x_1, \dots, x_K) = 0$  if some  $x_i = 0$  where  $1 \leq i \leq K$ .

Smooth minimum functions can also be constructed by

$$h_2(x_1, \dots, x_K) = \frac{x_1 g(x_1) + \dots + x_K g(x_K)}{g(x_1) + \dots + g(x_K)}. \quad (21)$$

A specific example is the Boltzmann operator, where  $g(x) = e^{-\alpha x}$ . The Boltzmann operator takes on the form of

$$\begin{aligned} h_2(x_1, \dots, x_K) &= \text{boltz}_\alpha(x_1, \dots, x_K) \\ &= \frac{\sum_{i=1}^K x_i e^{-\alpha x_i}}{\sum_{i=1}^K e^{-\alpha x_i}}, \end{aligned} \quad (22)$$

and converges to the minimum function as  $\alpha \rightarrow +\infty$ .

### D. Clustering by Minimizing Smoothed WCSS

1) *The relationship between LogSumExp and Maximum-Entropy Fuzzy Clustering:* Approximating WCSS (12) by LogSumExp (19), we have the following objective:

$$\min_{\mathbf{c}_1, \dots, \mathbf{c}_K} J_{\text{LSE}}(\mathbf{c}_1, \dots, \mathbf{c}_K) = \sum_{n=1}^N -\frac{1}{\lambda} \ln \left( \sum_{k=1}^K e^{-\lambda d_{kn}} \right). \quad (23)$$

The differentiable objective function  $J_{\text{LSE}}$  has the first-order partial derivative:

$$\begin{aligned} \partial_{\mathbf{c}_k} J_{\text{LSE}} &= \frac{\partial J_{\text{LSE}}}{\partial \mathbf{c}_k} = \sum_{n=1}^N \frac{e^{-\lambda d_{kn}}}{\sum_{i=1}^K e^{-\lambda d_{in}}} \frac{\partial d_{kn}}{\partial \mathbf{c}_k} \\ &= - \sum_{n=1}^N \frac{e^{-\lambda d_{kn}}}{\sum_{i=1}^K e^{-\lambda d_{in}}} (\mathbf{x}_n - \mathbf{c}_k). \end{aligned} \quad (24)$$

The minimizer of  $J_{\text{LSE}}$  can be found by gradient descent iteration:

$$\mathbf{c}_k^{(\tau+1)} = \mathbf{c}_k^{(\tau)} - \gamma_k^{(\tau)} \partial_{\mathbf{c}_k} J_{\text{LSE}}(\mathbf{c}_k^{(\tau)}), \quad (25)$$

where  $\gamma_k^{(\tau)}$  is the learning rate at the  $\tau$ -th iteration. This updating procedure (25) is equivalent<sup>2</sup> to the MEFC algorithm if one set

$$\gamma_k^{(\tau)} = 1 / \sum_{n=1}^N \left( \frac{e^{-\lambda d_{kn}^{(\tau)}}}{\sum_{i=1}^K e^{-\lambda d_{in}^{(\tau)}}} \right). \quad (26)$$

Such learning rate value is related to the second-order partial derivative of  $J_{\text{LSE}}$ , which we will discuss later. The membership (10) of MEFC is identical to  $\partial J_{\text{LSE}} / \partial d_{kn}$ .

2) *Towards Lloyd's algorithm:* In the limit of  $\lambda \rightarrow \infty$ ,  $\frac{e^{-\lambda d_{kn}}}{\sum_{i=1}^K e^{-\lambda d_{in}}} \rightarrow 1$  if  $d_{kn} \leq d_{in} \forall i \in \{1, \dots, K\}$ , and  $\frac{e^{-\lambda d_{kn}}}{\sum_{i=1}^K e^{-\lambda d_{in}}} \rightarrow 0$  otherwise<sup>3</sup>. In this case, the learning rate  $\gamma_k^{(\tau)} \rightarrow 1/N_k^{(\tau)}$  where  $N_k^{(\tau)}$  is the number of data points closest to  $\mathbf{c}_k^{(\tau)}$ , and the updating procedure (25) approaches to Lloyd's algorithm.

<sup>2</sup>It should be noted that the equivalence mentioned in this paper is in the sense of the algorithm level, not in the criterion level.

<sup>3</sup>We assume  $\forall n$ , if  $k \neq i$ , then  $d_{kn} \neq d_{in}$ .

3) *Lloyd's algorithm and Newton's method*: The smooth objective function  $J_{\text{LSE}}$  possesses the second-order partial derivative given by

$$\begin{aligned} \partial_{\mathbf{c}_k}^2 J_{\text{LSE}} &= \frac{\partial^2 J_{\text{LSE}}}{\partial \mathbf{c}_k^2} = - \sum_{n=1}^N \left( \frac{e^{-\lambda d_{kn}}}{\sum_i e^{-\lambda d_{in}}} \frac{\partial^2 d_{kn}}{\partial \mathbf{c}_k^2} \right. \\ &\quad \left. + \frac{-\lambda e^{-\lambda d_{kn}} \sum_{i \neq k} e^{-\lambda d_{in}}}{(\sum_i e^{-\lambda d_{in}})^2} \frac{\partial d_{kn}}{\partial \mathbf{c}_k} \left( \frac{\partial d_{kn}}{\partial \mathbf{c}_k} \right)^\top \right) \\ &= \sum_{n=1}^N \frac{e^{-\lambda d_{kn}}}{\sum_i e^{-\lambda d_{in}}} (\mathbf{I} + \xi_{kn} \mathbf{D}_{kn}) \end{aligned} \quad (27)$$

where  $\xi_{kn} = \frac{\lambda \sum_{i \neq k} e^{-\lambda d_{in}}}{\sum_i e^{-\lambda d_{in}}}$  and  $\mathbf{D}_{kn} = (\mathbf{x}_n - \mathbf{c}_k)(\mathbf{x}_n - \mathbf{c}_k)^\top$ . By employing Newton's method, the minimizer of  $J_{\text{LSE}}$  can be iteratively found using

$$\mathbf{c}_k^{(\tau+1)} = \mathbf{c}_k^{(\tau)} - [\partial_{\mathbf{c}_k}^2 J_{\text{LSE}}(\mathbf{c}_k^{(\tau)})]^{-1} \partial_{\mathbf{c}_k} J_{\text{LSE}}(\mathbf{c}_k^{(\tau)}). \quad (28)$$

As  $\lambda \rightarrow +\infty$ ,  $\partial_{\mathbf{c}_k}^2 J_{\text{LSE}}(\mathbf{c}_k^{(\tau)}) \rightarrow N_k^{(\tau)} \mathbf{I}$ , and the updating procedure (28) approaches Lloyd's algorithm. This indicates that Lloyd's algorithm is essentially Newton's method, which aligns with the perspective presented in [28]. There is a close relationship between the learning rate value (26) and the second-order partial derivative of  $J_{\text{LSE}}$  (27): the gradient descent with the learning rate of (26) is equivalent to an approximated Newton's method in the sense that one term (the rank-1 matrix  $\mathbf{D}_{kn}$ ) in (27) is ignored. A discussion of the pros and cons of optimizing  $J_{\text{LSE}}$  using gradient descent or Newton's method would be valuable, but it is beyond the scope of this paper. In the following section, we will observe the same connection between Newton's method and FKM.

4) *The relationship between p-Norm and fuzzy K-means*:

By substituting the minimum function in WCSS with the p-Norm function (20), we have

$$\min_{\mathbf{c}_1, \dots, \mathbf{c}_K} J_{\text{PN}}(\mathbf{c}_1, \dots, \mathbf{c}_K) = \sum_{n=1}^N (d_{1n}^{-p} + \dots + d_{Kn}^{-p})^{-1/p}. \quad (29)$$

The objective  $J_{\text{PN}}$  has the first partial derivative of

$$\partial_{\mathbf{c}_k} J_{\text{PN}} = \frac{\partial J_{\text{PN}}}{\partial \mathbf{c}_k} = - \sum_{n=1}^N \frac{d_{kn}^{-p-1}}{(\sum_{i=1}^K d_{in}^{-p})^{1/p+1}} (\mathbf{x}_n - \mathbf{c}_k). \quad (30)$$

The minimizer of  $J_{\text{PN}}$  can be located using gradient descent iteration

$$\mathbf{c}_k^{(\tau+1)} = \mathbf{c}_k^{(\tau)} - \gamma_k^{(\tau)} \partial_{\mathbf{c}_k} J_{\text{PN}}(\mathbf{c}_k^{(\tau)}). \quad (31)$$

When setting  $p = 1/(m-1)$  and

$$\gamma_k^{(\tau)} = 1 / \sum_{n=1}^N \frac{(d_{kn}^{(\tau)})^{-p-1}}{(\sum_{i=1}^K (d_{in}^{(\tau)})^{-p})^{1/p+1}}, \quad (32)$$

this gradient descent iteration is equivalent to the FKM algorithm. The membership (7) of FKM is equivalent to a power exponent  $(1/m)$  of  $\partial J_{\text{PN}} / \partial d_{kn}$ .

The connection between FKM and Newton's method becomes evident by taking the second-order partial derivative of  $J_{\text{PN}}$ , which is

$$\partial_{\mathbf{c}_k}^2 J_{\text{PN}} = \frac{\partial^2 J_{\text{PN}}}{\partial \mathbf{c}_k^2} = \sum_{n=1}^N \frac{d_{kn}^{-p-1}}{(\sum_i d_{in}^{-p})^{1/p+1}} (\mathbf{I} + \zeta_{kn} \mathbf{D}_{kn}), \quad (33)$$

where  $\zeta_{kn} = (p+1) \frac{\sum_{i \neq k} d_{in}^{-p}}{d_{kn} \sum_i d_{in}^{-p}}$ . Hence, FKM can also be viewed as an approximated Newton's method that ignores the term of the rank-1 matrix  $\mathbf{D}_{kn}$ .

### E. From Boltzmann Operator to A Novel Clustering Algorithm

Employing the Boltzmann operator to smooth WCSS results in

$$\min_{\mathbf{c}_1, \dots, \mathbf{c}_K} J_B(\mathbf{c}_1, \dots, \mathbf{c}_K) = \sum_{n=1}^N \frac{\sum_{i=1}^K d_{in} e^{-\alpha d_{in}}}{\sum_{i=1}^K e^{-\alpha d_{in}}}. \quad (34)$$

The objective  $J_B$  possesses the first-order partial derivative of

$$\begin{aligned} \partial_{\mathbf{c}_k} J_B &= \frac{\partial J_B}{\partial \mathbf{c}_k} = - \sum_{n=1}^N \frac{e^{-\alpha d_{kn}}}{\sum_i e^{-\alpha d_{in}}} \left[ 1 - \alpha (d_{kn} \right. \\ &\quad \left. - \frac{\sum_i d_{in} e^{-\alpha d_{in}}}{\sum_i e^{-\alpha d_{in}}}) \right] (\mathbf{x}_n - \mathbf{c}_k). \end{aligned} \quad (35)$$

The minimizer of  $J_B$  can be found using gradient descent iteration

$$\mathbf{c}_k^{(\tau+1)} = \mathbf{c}_k^{(\tau)} - \gamma_k^{(\tau)} \partial_{\mathbf{c}_k} J_B(\mathbf{c}_1^{(\tau)}, \dots, \mathbf{c}_K^{(\tau)}), \quad (36)$$

where

$$\gamma_k^{(\tau)} = 1 / \left( \sum_{n=1}^N \frac{e^{-\alpha d_{kn}^{(\tau)}}}{\sum_i e^{-\alpha d_{in}^{(\tau)}}} \left[ 1 - \alpha (d_{kn}^{(\tau)} - \frac{\sum_i d_{in}^{(\tau)} e^{-\alpha d_{in}^{(\tau)}}}{\sum_i e^{-\alpha d_{in}^{(\tau)}}}) \right] \right). \quad (37)$$

This updating procedure can be reformulated into a two-step iteration procedure akin to FKM, that is

$$\mathbf{c}_k^{(\tau+1)} = \frac{\sum_n w_{kn}^{(\tau)} \mathbf{x}_n}{\sum_n w_{kn}^{(\tau)}}, \quad (38)$$

where weights  $w_{kn}^{(\tau)}$  are calculated by

$$w_{kn}^{(\tau)} = \frac{e^{-\alpha d_{kn}^{(\tau)}}}{\sum_{i=1}^K e^{-\alpha d_{in}^{(\tau)}}} \left[ 1 - \alpha (d_{kn}^{(\tau)} - \frac{\sum_{i=1}^K d_{in}^{(\tau)} e^{-\alpha d_{in}^{(\tau)}}}{\sum_{i=1}^K e^{-\alpha d_{in}^{(\tau)}}}) \right]. \quad (39)$$

The time complexity of one iteration is  $O(NK^2)$ . Although the sum of weights equals one, i.e.,

$$\sum_{k=1}^K w_{kn}^{(\tau)} = 1, \quad \forall n, \quad (40)$$

it is worth noting that these weights cannot be interpreted as probabilities or memberships since some weight values are negative. The membership of EKM is defined in Section III-G.

### F. Physical Interpretation of Equilibrium K-Means

The second law of thermodynamics asserts that in a closed system with constant external parameters (e.g., volume) and fixed entropy, the internal energy will reach its minimum value at the state of thermal equilibrium. The objective (34) of EKM follows this minimum energy principle. This connection can be established by envisioning data points as particles with discrete/quantized energy levels, where the number of energy levels is equivalent to the number of centroids, and the energy

value corresponds to the squared Euclidean distance between a data point and a centroid.

Boltzmann's law tells that at the state of thermodynamic equilibrium, the probability of a particle occupying a specific energy level decreases exponentially with the increase of the energy value of that level. Hence, the objective function (34) equals the expectation of the entire system's energy, and EKM seeks centroids to minimize this energy expectation. Due to this connection, we refer to the proposed algorithm (alternating between (38) and (39)) as equilibrium K-means.

### G. Membership Defined in Equilibrium K-Means

According to the physical interpretation of EKM, the exponential term  $e^{-\alpha d_{kn}}$  can be interpreted as the unnormalized probability of the  $n$ -th data point belonging to the  $k$ -th clusters. Hence, the membership of the  $n$ -th data point to the  $k$ -th cluster can be defined as

$$u_{kn} = \frac{e^{-\alpha d_{kn}}}{\sum_{i=1}^K e^{-\alpha d_{in}}}. \quad (41)$$

Note that, although the membership formula of EKM is the same as that (10) of MEFC, the values are different because centroids are calculated using distinct formulas. We have no intention in this paper to compare membership defined in different fuzzy clustering algorithms; therefore, we only define membership in EKM and leave the further discussion to subsequent research.

### H. Convergence of Smooth K-Means

HKM, FKM, MEFC, and EKM are special cases of SKM, which can be generalized as the following gradient descent algorithm:

$$\mathbf{c}_k^{(\tau+1)} = \mathbf{c}_k^{(\tau)} - \gamma_k^{(\tau)} \partial_{\mathbf{c}_k} J(\mathbf{c}_1^{(\tau)}, \dots, \mathbf{c}_K^{(\tau)}), \quad (42)$$

where  $J(\mathbf{c}_1^{(\tau)}, \dots, \mathbf{c}_K^{(\tau)}) = \sum_{n=1}^N h(d_{1n}^{(\tau)}, \dots, d_{Kn}^{(\tau)})$ ,  $h$  is a smooth minimum function,  $d_{kn}^{(\tau)} = \frac{1}{2} \|\mathbf{x}_n - \mathbf{c}_k^{(\tau)}\|_2^2$ , and the learning rate  $\gamma_k^{(\tau)}$  is given by

$$\gamma_k^{(\tau)} = 1 / \sum_{n=1}^N \left( \frac{\partial h}{\partial d_{kn}}(d_{1n}^{(\tau)}, \dots, d_{Kn}^{(\tau)}) \right). \quad (43)$$

Different KM algorithms can be obtained by taking  $h$  the corresponding explicit form (refer to Section III-D to III-E). This general form facilitates the uniform study of KM algorithms. Below we give a convergence guarantee conditioning on the properties of  $h$ :

**Theorem 1** (Convergence Condition): *The centroid sequence obtained by (42) converges to a (local) minimizer or saddle point of the objective function  $J$  if the following conditions can be satisfied:*

- 1) (Concavity) *The function  $h$  is a concave function at its domain  $[0, +\infty)^K$ .*
- 2) (Boundness) *The function  $h$  has a lower bound, i.e.,  $h > -\infty$ , and the learning rate set  $\{\gamma_k^{(\tau)}\}_{\tau,k}$  has a positive lower bound, i.e.,  $\exists \epsilon > 0$ , such that  $\gamma_k^{(\tau)} \geq \epsilon$  for all  $\tau$  and  $k$ .*

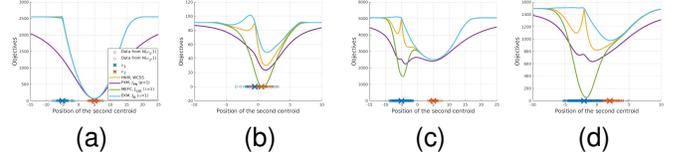


Fig. 2. Objectives as a function of the second centroid moving along the  $x$ -axis (the first centroid is fixed at  $\mu_1$ ). The yellow, purple, green, and blue curves are reformulated objective functions of HKM, FKM, MEFC, and EKM, respectively. (a), (b), (c), and (d) present objective functions under different data distributions.

*Proof:* See the Appendix for the proof, which generalizes the proof of convergence of fuzzy K-means in [29]. ■

It can be easily verified that when the smooth minimum function  $h$  is LogSumExp (19) and p-Norm (20) in which case SKM (42) is equivalent to MEFC and FKM, respectively, the above convergence condition can be satisfied with any initial centroids. Although the Boltzmann operator (22) is not concave, it is a smooth approximation of a concave function (the minimum function). Hence, EKM also exhibits good convergence behavior in numerical experiments.

## IV. COMPARISON OF DIFFERENT SMOOTHED OBJECTIVES

### A. Case Study

This section presents an empirical analysis of the behavior of different KM algorithms by examining their reformulated objective functions in some examples with well-designed data structures. Datasets comprising two classes of one-dimensional data points are generated by sampling from two normal distributions, drawing  $N_1$  samples from a distribution with mean  $\mu_1$  and unit variance, and  $N_2$  samples from another with mean  $\mu_2$  and unit variance. Using different parameter combinations, we generate four datasets: 1. A balanced, non-overlapping dataset ( $N_1 = N_2 = 50$ ,  $\mu_1 = -5$ , and  $\mu_2 = +5$ ; Fig. 2a); 2. A balanced, overlapping dataset ( $N_1 = N_2 = 50$ ,  $\mu_1 = -0.5$ , and  $\mu_2 = +0.5$ ; Fig. 2b); 3. An imbalanced, non-overlapping dataset ( $N_1 = 5000$ ,  $N_2 = 50$ ,  $\mu_1 = -5$ , and  $\mu_2 = +5$ ; Fig. 2c); 4. An imbalanced, overlapping dataset ( $N_1 = 2000$ ,  $N_2 = 50$ ,  $\mu_1 = -2$ , and  $\mu_2 = +2$ ; Fig. 2d).

We fix the first centroid at  $\mu_1$ , and plot the reformulated objectives of HKM (WCSS (12)), FKM ( $J_{PN}$  (29)), MEFC ( $J_{LSE}$  (23)), and EKM ( $J_B$  (34)) as a function of the position of the second centroid in Fig. 2. The four objective functions behave similarly on the first two balanced datasets, but it is worth noting that the last two imbalanced datasets. Fig. 2c and Fig. 2d show that the local and global minimum points of the objective functions of HKM, FKM, and MEFC are biased towards the center of the large cluster, i.e.,  $\mu_1$ . In contrast, EKM does not have an obvious local minimum and its global minimum point aligns with the true cluster center, i.e.,  $\mu_2$ , highlighting EKM's superiority in handling imbalanced data.

### B. The Analysis of EKM's Effectiveness on Imbalanced Data

We analyze the effectiveness of EKM on imbalanced data based on the gradient of its objective. The gradient of the

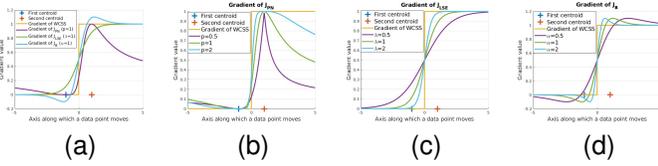


Fig. 3. Gradients as a function of a data point moving along the  $x$ -axis. (a) Gradients of WCSS,  $J_{PN}$ ,  $J_{LSE}$ , and  $J_B$  with fixed smoothing parameters. (b), (c), and (d) are gradients of  $J_{PN}$ ,  $J_{LSE}$ , and  $J_B$  with varied smoothing parameters, respectively.

smoothed WCSS with respect to the data-centroid distance (i.e.,  $\partial J/\partial d_{kn}$ ) can be interpreted as the force exerted by a spring. A positive gradient value represents an attractive force, while a negative value represents a repulsive force.

A simple example is given to plot the gradient values of different KM's objectives. We fix two centroids at  $-1$  and  $+1$ , respectively, and move a data point along the  $x$ -axis. In Fig. 3a we display the gradients of WCSS,  $J_{LSE}$ ,  $J_{PN}$ , and  $J_B$  with respect to the distance between the data point and the second centroid. As evident from the figure, data points on the side of the first centroid do not impact the second centroid of HKM, but they do attract the second centroids of FKM and MEFC. This finding supports the statement in [30] that FKM has a stronger uniform effect than HKM. On the other hand, data points near the first centroid have repulsive forces on the second centroid of EKM, which compensate for the attraction from other data points, effectively reducing the uniform effect. It is worth noting that data points near the second centroid of EKM have the strongest attraction forces to this centroid. Consequently, EKM's centroids are anchored by their surrounding data points, making their position less influenced by noise and outliers.

### C. The Choice of $\alpha$

The smoothing parameter  $\alpha$  impacts the performance of EKM, but the optimal choice remains unknown. This is not a difficulty unique to EKM. The FKM algorithm also struggles with selecting the optimal fuzzifier value,  $m$ . Despite numerous studies discussing the selection of  $m$ , a widely accepted solution has yet to be found [31].

As a rule of thumb, when the dimension of the data space is less than or equal to three, setting  $\alpha = 1$  appears effective after normalizing the data to have a zero mean and unit variance for each dimension. As the data space dimension increases, the data-centroid distance  $d_{kn}$  increases, necessitating a decrease in  $\alpha$  to ensure that the exponential term  $e^{-\alpha d_{kn}}$  falls within a normal range. Hence, in the case of a higher dimensional data clustering, we suggest initially setting  $\alpha$  to ten times the data variance and gradually reducing it until a sudden increase in centroid-centroid distance is observed. Because the increase in centroid distance implies the emergence of repulsive forces that reduce the uniform effect. This strategy is applied in deep clustering on the MNIST dataset in Section VI. An adaptive selection formula of  $\alpha$  and its validation are given in the supplemental material.

## V. NUMERICAL EXPERIMENTS

Numerical experiments are conducted to compare the performance of our proposed EKM algorithm with six other centroid-based algorithms: (1) HKM, (2) FKM, (3) MEFC, (4) csiFKM [19], (5) siibFKM [20], and (6) the GMM. Multiprototype-based algorithms such as [5], [22] are not appropriate as baseline models because they are too complex to be benchmarks for gauging the efficiency of the proposed EKM.

To avoid local optima convergence, each clustering algorithm is executed ten times, and the result with the lowest objective value is presented. Centroids are initialized independently for each replication by the K-means++ algorithm [32]. Convergence is achieved when the relative moving distance of centroids between successive iterations is less than  $1e-3$ , i.e.,

$$\frac{\left(\sum_{k=1}^K \|\mathbf{c}_k^{(\tau)} - \mathbf{c}_k^{(\tau-1)}\|_2^2\right)^{1/2}}{\left(\sum_{k=1}^K \|\mathbf{c}_k^{(\tau)}\|_2^2\right)^{1/2}} \leq 1e-3. \quad (44)$$

The maximum number of iterations is set to 100. FKM, csiFKM, and siibFKM employ a typical fuzzifier value of  $m = 2$  [33]. The fuzzifier value of MEFC is set to  $\lambda = 1$  and the smoothing parameter of EKM is set to  $\alpha = 1$ . We evaluate EKM's performance on 12 datasets, including two synthetic and ten real datasets. Besides, we also apply EKM to the application of image segmentation. Because of space limitations, experiment results from six datasets and the image segmentation are shown in the paper. Full experiment results can be found in the supplemental material. The six datasets and the test image are introduced below.

### A. Datasets

1) *Synthetic 2-D Ball*: We generate data from three distinct normal distributions, each having different means:  $(-2, 2)$ ,  $(2, -2)$ , and  $(4, 4)$ , and the same identity covariance matrix. The dataset consists of 2,100 samples, of which 2000 samples from the normal distribution with means  $(-2, 2)$ , and 100 samples from the other two normal distributions (each distribution for 50 samples). The clustering results are shown in Fig. 4.

2) *Synthetic 2-D Noisy Ball*: We add 100 uniformly distributed noise data (around 5% of the number of clean samples) to the synthetic 2-D ball dataset in order to test the algorithms' robustness against noise. The clustering outcomes are displayed in Fig. 5.

3) *Fisher's Iris Data*: The Fisher's Iris dataset [34], [35] is a well-known dataset for evaluating the performance of clustering algorithms. It comprises 50 samples from each of three iris species (Iris setosa, Iris virginica, and Iris versicolor). Two features, the width and the length of the sepals, are utilized as clustering features and normalized to have zero means and unit variance. Fig. 6 depicts the clustering results.

4) *Imbalanced Fisher's Iris Data*: Because the instance distribution in the Iris dataset is balanced. To examine the capability of imbalanced data clustering, we generate an imbalanced Fisher's Iris dataset by removing the first 30 instances of Iris setosa, and merging the other two species (Iris virginica

and Iris versicolor) into a single class. The clustering results of this imbalanced dataset are presented in Fig. 7.

5) *Wisconsin Diagnostic Breast Cancer*: Wisconsin Diagnostic Breast Cancer (WDBC) [36] comprises 30 features of cell nuclei from diagnostic benign and malignant breast tumors, obtained from digitized images of fine-needle aspiration of breast masses. This database contains 569 instances, with an instance distribution of 357 benign and 212 malignant cases. We use the first three features (mean radius, mean texture, and mean perimeter) for clustering, each normalized with zero mean and unit variance. Fig. 8 presents the clustering results where the mean radius and the mean texture are used as coordinates for visualization.

6) *Imbalanced Wisconsin Diagnostic Breast Cancer*: The data distribution in WDBC is quite balanced. To examine the ability of imbalanced data clustering, we remove the first 200 malignant instances in WDBC, resulting in an imbalanced WDBC database with a distribution of 357 benign and 12 malignant instances. We still use the first three features for clustering, each normalized with zero mean and unit variance. The clustering results are shown in Fig. 9.

7) *Color-based image segmentation*: The test image *coloredChips.png* is chosen from Matlab's image gallery (shown in Fig. 10a). The number of clusters is set to five to expect the segmentation of the four colored chips and the background (i.e., the desk surface). Clustering is conducted in the RGB color space and the segmentation results can be viewed in Fig. 10b - 10e.

## B. Discussion of Results

In general, these results demonstrate that the proposed EKM algorithm performs competitively on balanced data and outperforms the other clustering algorithms on imbalanced data. In particular, HKM, FKM, and MEFC suffer from the uniform effect when data has an imbalanced distribution. The *csiFKM* and *siibFKM* algorithms are sensitive to noise data: Fig. 5c and Fig. 5d show that their centroids are highly deviated from the correct positions in the presence of noise. In comparison, EKM is more robust to noise because of the anchoring effect produced by the positive gradients (see Section IV-B), as seen in Fig. 5e.

The strong performance of EKM shown in Fig. 9 demonstrates its ability to identify clustered anomalies. This makes up a missing function of HKM and FKM in anomaly detection: they are typically used to detect isolated anomalies, not clustered anomalies [37]–[40].

The results of color-based image segmentation are shown in Fig. 10b - 10e. HKM, FKM, and MEFC erroneously separate the background (the desk surface) into two clusters. This happens because the background comprises most pixels, and splitting it into two clusters better balances the number of pixels within each clusters. On the other hand, EKM produces the expected segmentation result.

## VI. DEEP CLUSTERING

Clustering algorithms using Euclidean distance for defining similarity face a challenge when clustering high-dimensional

data, such as images, due to the small distance differences between various point pairs in high-dimensional space [41]. Deep clustering is a technique that addresses this issue by employing DNNs to map high-dimensional data to low-dimensional representation. Many deep clustering methods combine DNNs with HKM, e.g., [15]–[17]. However, this design is less favorable to imbalanced data [16]. In this section, we show the promise of EKM in the deep clustering of imbalanced data by replacing HKM with EKM in a popular deep clustering framework.

### A. Optimization Procedure

Deep clustering network (DCN) [15] is a popular deep clustering framework. As illustrated in Fig. 11, DCN maps high-dimensional data to low-dimensional representation through an autoencoder network. An autoencoder follows the autoencoder, mapping the representation back to the original high-dimensional space (i.e., reconstruction). To ensure that the low-dimensional representation maintains the primary information of the original data, the autoencoder and the autoencoder are jointly trained to minimize the reconstruction error. Additionally, to make the low-dimensional representation have a clustering-friendly structure, a clustering error is minimized along with the reconstruction error. DCN uses an alternating optimization algorithm to minimize the total error, and the optimization process is described below.

First, the autoencoder and the autoencoder are jointly trained to reduce the following loss for the incoming data  $\mathbf{x}_n$ :

$$\min_{\theta_e, \theta_d} L^n = l(\mathbf{g}(\mathbf{f}(\mathbf{x}_n)), \mathbf{x}_n) + \beta \|\mathbf{f}(\mathbf{x}_n) - \mathbf{C}\mathbf{s}_n\|_2^2, \quad (45)$$

where  $\mathbf{f}(\cdot)$  and  $\mathbf{g}(\cdot)$  are simplified symbols for autoencoder  $\mathbf{f}(\cdot; \theta_e)$  and autoencoder  $\mathbf{g}(\cdot; \theta_d)$ , respectively. The function  $l(\cdot)$  is the least-squares loss  $l(\hat{\mathbf{x}}, \mathbf{x}) = \|\hat{\mathbf{x}} - \mathbf{x}\|_2^2$  to measure the reconstruction error. The assignment vector  $\mathbf{s}_n \in \mathbb{R}^{K \times 1}$  has only one non-zero element and  $\mathbf{1}^T \mathbf{s}_n = 1$ , indicating which cluster the  $n$ -th data belongs to, and the  $k$ -th column of  $\mathbf{C} = [\mathbf{c}_1, \dots, \mathbf{c}_K]$  is the centroid of the  $k$ -th cluster. The parameter  $\beta$  balances the reconstruction error versus the clustering error. Then, the network parameters  $\{\theta_e, \theta_d\}$  are fixed, and the parameters  $\{\mathbf{s}_n\}$  are updated as follows:

$$s_{j,n} \leftarrow \begin{cases} 1, & \text{if } j = \arg \min_{k=\{1, \dots, K\}} \|\mathbf{f}(\mathbf{x}_n) - \mathbf{c}_k\|_2, \\ 0, & \text{otherwise,} \end{cases} \quad (46)$$

where  $s_{j,n}$  is the  $j$ -th element of  $\mathbf{s}_n$ . Finally,  $\mathbf{C}$  is updated by the batch-learning version of the HKM algorithm:

$$\mathbf{c}_k \leftarrow \mathbf{c}_k + (1/m_k^n)(\mathbf{f}(\mathbf{x}_n) - \mathbf{c}_k)s_{k,n}, \quad (47)$$

where  $m_k^n$  is the number of samples assigned to the  $k$ -th cluster before the incoming data  $\mathbf{x}_n$ , controlling the learning rate of the  $k$ -th centroid. Overall, the optimization procedure of DCN alternates between updating networks parameters  $\{\theta_e, \theta_d\}$  by solving (45) and updating HKM parameters  $\{\mathbf{C}, \{\mathbf{s}_n\}\}$  by (46) and (47).

However, the centroid updating rule (47) is problematic for imbalanced data due to the uniform effect. To address

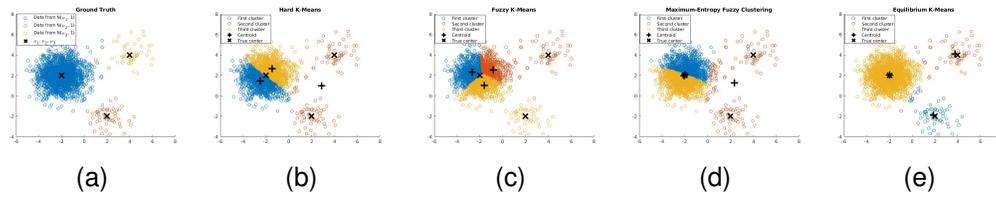


Fig. 4. (a) Synthetic 2-D ball data. (b) Clustering by HKM. (c) Clustering by FKM. (d) Clustering by MEFC. (e) Clustering by EKM.

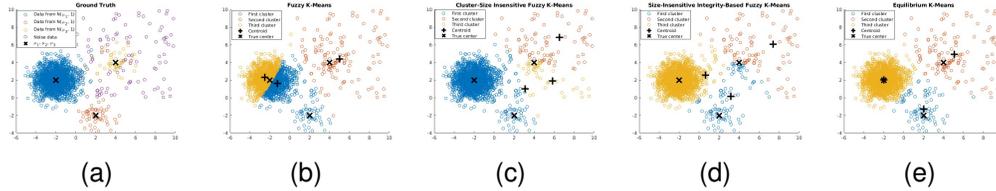


Fig. 5. (a) Synthetic 2-D noisy ball data. (b) Clustering by FKM. (c) Clustering by csifKM. (d) Clustering by siibFKM. (e) Clustering by EKM.

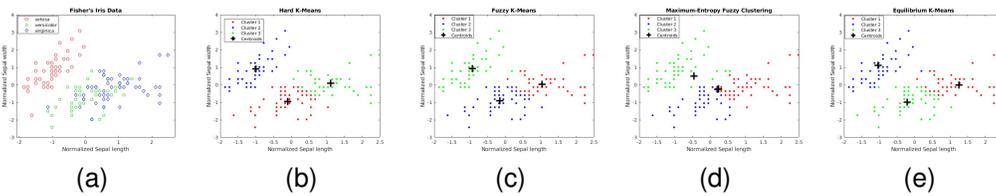


Fig. 6. (a) Fisher's Iris data. (b) Clustering by HKM. (c) Clustering by FKM. (d) Clustering by MEFC. (e) Clustering by EKM.

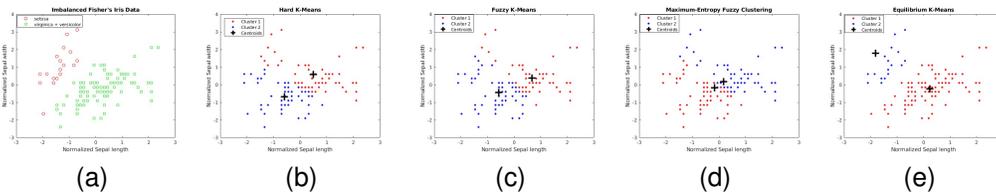


Fig. 7. (a) Imbalanced Fisher's Iris data. (b) Clustering by HKM. (c) Clustering by FKM. (d) Clustering by MEFC. (e) Clustering by EKM.

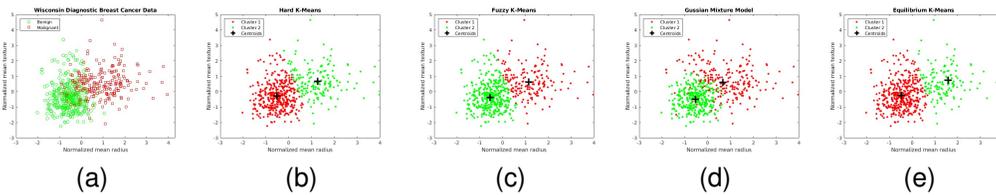


Fig. 8. (a) Wisconsin Diagnostic Breast Cancer dataset. Three features are used for clustering and two of them are used for visualization. (b) Clustering by HKM. (c) Clustering by FKM. (d) Clustering by GMM. (e) Clustering by EKM.

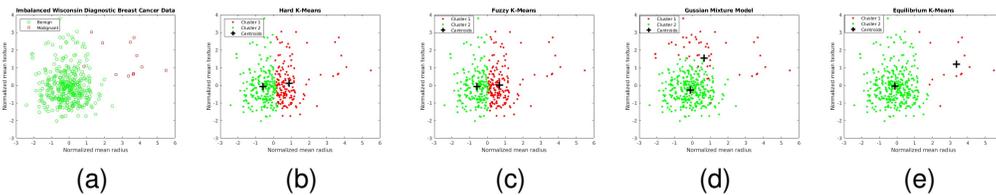


Fig. 9. (a) Imbalanced Wisconsin Diagnostic Breast Cancer dataset. (b) Clustering by HKM. (c) Clustering by FKM. (d) Clustering by GMM. (e) Clustering by EKM.

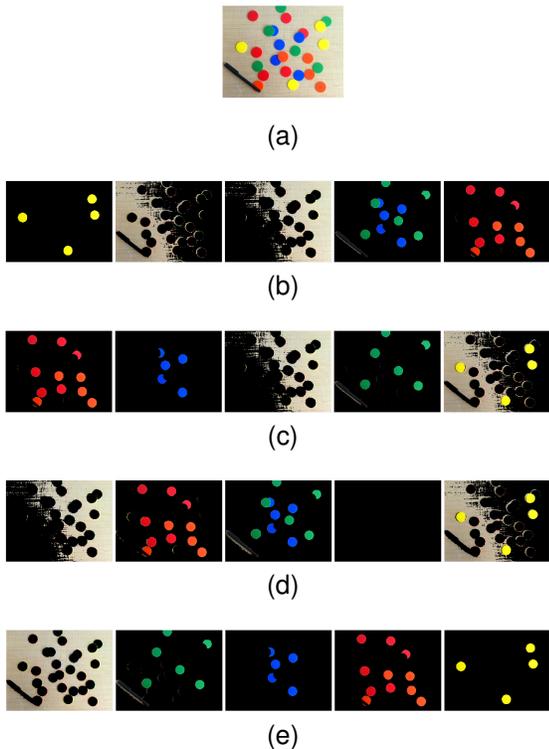


Fig. 10. (a) coloredChips.png. (b) Segmentation by HKM. (c) Segmentation by FKM. (d) Segmentation by MEFC. (e) Segmentation by EKM.

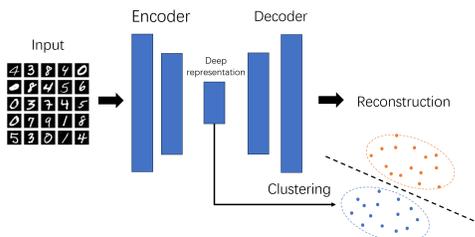


Fig. 11. Illustration of the DCN framework [15]. The parameters of the encoder, decoder, and clustering model are optimized jointly to minimize the reconstruction error and clustering error.

this issue, we propose to replace (47) with the batch-learning version of EKM:

$$\begin{aligned}
 \mathbf{c}_k &\leftarrow \mathbf{c}_k - (1/m_k^n) \partial_{\mathbf{c}_k} J_B(\mathbf{c}_1, \dots, \mathbf{c}_K) \\
 &= \mathbf{c}_k + (1/m_k^n) \frac{e^{-\alpha d_{kn}}}{\sum_{i=1}^K e^{-\alpha d_{in}}} [1 - \alpha(d_{kn} \\
 &\quad - \frac{\sum_{i=1}^K d_{in} e^{-\alpha d_{in}}}{\sum_{i=1}^K e^{-\alpha d_{in}})] (\mathbf{f}(\mathbf{x}_n) - \mathbf{c}_k),
 \end{aligned} \tag{48}$$

where  $d_{kn} = \frac{1}{2} \|\mathbf{f}(\mathbf{x}_n) - \mathbf{c}_k\|_2^2$ . There are other details and tricks to implement DCN, such as the initialization of the networks. We only introduce the part related to our contribution, and kindly refer to [15] for more implementation details.

### B. Clustering Performance on MNIST

To implement DCN, we refer to the code one of its authors provided, available at <https://github.com/boyangum/DCN-New>. We use the default neural network structure and

TABLE I  
EVALUATION ON FULL MNIST

Methods	SAE+HKM	DCN+HKM	SAE+EKM	DCN+EKM
NMI	0.725	0.798	0.711	<b>0.813</b>
ARI	0.667	<b>0.744</b>	0.642	0.731
ACC	0.795	<b>0.837</b>	0.782	0.808

hyperparameters. In particular, the dimension of the low-dimensional representation is set to ten, and the parameter  $\beta$  is set to one. The smoothing parameter  $\alpha$  is tuned on the representation obtained by the initialized DCN networks according to the strategy introduced in Section IV-C. We evaluate clustering performance using three widely accepted metrics: including normalized mutual information (NMI), adjusted Rand index (ARI), and clustering accuracy (ACC). NMI and ACC range from 0 to 1, with zero representing the worst and one representing the best. ARI ranges from -1 to 1, with minus one representing the worst and one representing the best.

We first evaluate the algorithm's performance on the full MNIST dataset [42], which contains 70,000 gray images of handwritten digits from 0 to 9. Each image has  $28 \times 28 = 784$  pixels. We set the smoothing parameter of EKM to  $\alpha = 5e-3$ . The clustering results are presented in TABLE I. We compare the proposed DCN+EKM with DCN+HKM and stacked auto-encoder (SAE). SAE is a specific version of DCN that only minimizes the reconstruction error. Thus, the learned representation by SAE does not have a clustering-friendly structure. The results show that DCN outperforms SAE, highlighting the importance of a clustering-friendly structure. We can also see that DCN+EKM performs similarly to DCN+HKM. We omit the results of DCN+FKM and DCN+MEFC due to space limitations. Their performance is not better than DCN+HKM.

MNIST contains 60,000 training images and 10,000 testing images. To evaluate the performance of imbalanced data clustering, we generate an imbalanced MNIST dataset by removing the training images of digits 1 to 9. This imbalanced dataset contains 15,923 images, of which approximately 6,900 are digit 0, while the remaining digits (1 to 9) each have about 1,000 images. We set the smoothing parameter of EKM to  $\alpha = 3.8e-3$ . The results are summarized in TABLE II. In conclusion, we find that an EKM-friendly representation structure is crucial for deep clustering of imbalanced data. We map the ten-dimensional representation obtained by DCN to a two-dimensional space by t-SNE [43] for visualization. The results are displayed in Fig. 12. The visualization result implies that the deep representation obtained by DCN+EKM is more discriminative than that obtained by DCN+HKM. We observe that EKM successfully identifies the large class (number 0) from other small classes without referring to the true labels, while HKM incorrectly divides the large class into four clusters.

## VII. CONCLUSION

This paper presents equilibrium K-means (EKM), a novel clustering algorithm effective for both balanced and imbalanced data. EKM is simple, interpretable, scalable to large

TABLE II  
EVALUATION ON IMBALANCED MNIST

Methods	SAE+HKM	DCN+HKM	SAE+EKM	DCN+EKM
NMI	0.551	0.584	0.583	<b>0.701</b>
ARI	0.317	0.325	0.396	<b>0.826</b>
ACC	0.413	0.434	0.497	<b>0.784</b>

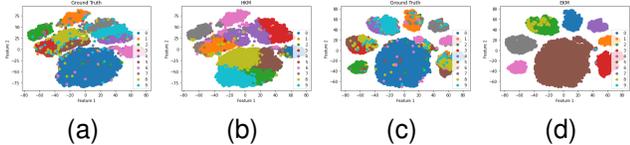


Fig. 12. Visualization of DCN representations by t-SNE. (a) Representations obtained by DCN+HKM, colored by true labels. (b) Representations obtained by DCN+HKM, colored by labels generated by HKM. (c) Representations obtained by DCN+EKM, colored by true labels. (d) Representations obtained by DCN+EKM, colored by labels generated by EKM.

datasets, and capable of creating unequal-sized clusters. Experimental results on real datasets from various domains show that EKM outperforms HKM, FKM, and other popular centroid-based algorithms on datasets with imbalanced data and performs comparably on datasets with balanced data. In the same deep clustering framework, compared with HKM and FKM, the clustering accuracy of imbalanced data using EKM has been noticeably improved. Furthermore, we reformulate HKM, FKM, and EKM in a general form of gradient descent. We encourage readers to study more properties of this general form to develop new K-means algorithms with non-heuristic memberships to address different issues.

#### APPENDIX A PROOF OF THEOREM 1

For concise and tidy, we denote the partial derivative  $\frac{\partial h}{\partial d_k}(d_{1n}^{(\tau)}, \dots, d_{Kn}^{(\tau)})$  as  $\frac{\partial h}{\partial d_k}|_{n,\tau}$ . Since  $h$  is a concave function at  $[0, +\infty)^K$ , we have

$$\begin{aligned} & h(d_{1n}, \dots, d_{Kn}) \\ & \leq h(d_{1n}^{(\tau)}, \dots, d_{Kn}^{(\tau)}) + \sum_{k=1}^K \frac{\partial h}{\partial d_k} \Big|_{n,\tau} \cdot (d_{kn} - d_{kn}^{(\tau)}), \end{aligned} \quad (49)$$

which holds for any  $n \in \{1, \dots, N\}$  and  $d_{1n}, \dots, d_{Kn} \in [0, +\infty)$ . Summing over  $n$ , it follows that

$$\begin{aligned} & J(\mathbf{c}_1, \dots, \mathbf{c}_K) \\ & \leq \sum_{n=1}^N \left[ h(d_{1n}^{(\tau)}, \dots, d_{Kn}^{(\tau)}) + \sum_{k=1}^K \frac{\partial h}{\partial d_k} \Big|_{n,\tau} \cdot (d_{kn} - d_{kn}^{(\tau)}) \right] \\ & = J(\mathbf{c}_1^{(\tau)}, \dots, \mathbf{c}_K^{(\tau)}) + \frac{1}{2} \sum_{k=1}^K \sum_{n=1}^N \left( \frac{\partial h}{\partial d_k} \Big|_{n,\tau} \cdot (\|\mathbf{x}_n - \mathbf{c}_k\|_2^2 - \|\mathbf{x}_n - \mathbf{c}_k^{(\tau)}\|_2^2) \right). \end{aligned} \quad (50)$$

Denote the function on the right side of the inequality as  $M(\mathbf{c}_1, \dots, \mathbf{c}_K)$ . With the boundness condition, we have  $\sum_{n=1}^N \frac{\partial h}{\partial d_k} \Big|_{n,\tau} > 0$  for any  $k$  and  $\tau$ , thus,  $M$  is a quadratic function and strictly convex, with the unique global

minimizer at  $(\mathbf{c}_1^{(\tau+1)}, \dots, \mathbf{c}_K^{(\tau+1)})$  defined by (42). Denote  $J(\mathbf{c}_1^{(\tau)}, \dots, \mathbf{c}_K^{(\tau)})$  as  $J^{(\tau)}$ ,  $M(\mathbf{c}_1^{(\tau)}, \dots, \mathbf{c}_K^{(\tau)})$  as  $M^{(\tau)}$ , and  $\partial_{\mathbf{c}_k} J(\mathbf{c}_1^{(\tau)}, \dots, \mathbf{c}_K^{(\tau)})$  as  $\partial_{\mathbf{c}_k} J^{(\tau)}$ . Each iteration of the centroid will reduce the objective function by

$$\begin{aligned} & J^{(\tau)} - J^{(\tau+1)} \\ & \geq J^{(\tau)} - M^{(\tau+1)} \\ & = \frac{1}{2} \sum_{k=1}^K \sum_{n=1}^N \frac{\partial h}{\partial d_k} \Big|_{n,\tau} \cdot (\|\mathbf{x}_n - \mathbf{c}_k^{(\tau)}\|_2^2 - \|\mathbf{x}_n - \mathbf{c}_k^{(\tau+1)}\|_2^2) \end{aligned}$$

substituting (42)

$$\begin{aligned} & = \frac{1}{2} \sum_{k=1}^K \sum_{n=1}^N \frac{\partial h}{\partial d_k} \Big|_{n,\tau} \cdot (\|\mathbf{x}_n - \mathbf{c}_k^{(\tau)}\|_2^2 - \|\mathbf{x}_n - \mathbf{c}_k^{(\tau)}\|_2^2) \\ & \quad + \gamma_k^{(\tau)} \partial_{\mathbf{c}_k} J^{(\tau)} \Big|_2^2 \\ & = \frac{1}{2} \sum_{k=1}^K \sum_{n=1}^N \frac{\partial h}{\partial d_k} \Big|_{n,\tau} \cdot (- (\gamma_k^{(\tau)})^2 \|\partial_{\mathbf{c}_k} J^{(\tau)}\|_2^2 \\ & \quad - 2\gamma_k^{(\tau)} (\mathbf{x}_n - \mathbf{c}_k^{(\tau)})^\top \partial_{\mathbf{c}_k} J^{(\tau)}) \end{aligned}$$

$$\text{using } \gamma_k^{(\tau)} = \frac{1}{\sum_{n=1}^N \frac{\partial h}{\partial d_k} \Big|_{n,\tau}}$$

$$\text{and } \partial_{\mathbf{c}_k} J^{(\tau)} = - \sum_{n=1}^N \frac{\partial h}{\partial d_k} \cdot (\mathbf{x}_n - \mathbf{c}_k^{(\tau)})$$

$$\begin{aligned} & = \sum_{k=1}^K -\frac{1}{2} \gamma_k^{(\tau)} \|\partial_{\mathbf{c}_k} J^{(\tau)}\|_2^2 + \sum_{k=1}^K \gamma_k^{(\tau)} \|\partial_{\mathbf{c}_k} J^{(\tau)}\|_2^2 \\ & = \sum_{k=1}^K \frac{1}{2} \gamma_k^{(\tau)} \|\partial_{\mathbf{c}_k} J^{(\tau)}\|_2^2 \end{aligned}$$

with the boundness condition of  $\gamma_k^{(\tau)}$

$$\geq \frac{1}{2} \epsilon \sum_{k=1}^K \|\partial_{\mathbf{c}_k} J^{(\tau)}\|_2^2,$$

(51)

where  $\epsilon$  is a positive number. Hence, the sequence  $(J^{(1)}, J^{(2)}, \dots)$  is non-increasing, and with the boundness condition that  $h > -\infty$ , we have  $\lim_{\tau \rightarrow +\infty} (J^{(\tau)} - J^{(\tau+1)}) \rightarrow 0$ . If the left side of the inequality (51) converges to zero, the right side of the inequality also converges to zero since it is non-negative. Consequently, we have  $\lim_{\tau \rightarrow +\infty} \partial_{\mathbf{c}_k} J^{(\tau)} \rightarrow 0$  for all  $k$ . Therefore, the sequence  $(\mathbf{c}_k^{(\tau)})$  converges to a stationary point of the objective function  $J$ . Because  $(J^{(1)}, J^{(2)}, \dots)$  is non-increasing, only (local) minimizers or saddle points appear as limit points.

#### REFERENCES

- [1] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016.
- [2] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on knowledge and data engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.
- [3] D. Ramyachitra and P. Manikandan, "Imbalanced dataset classification and solutions: a review," *International Journal of Computing and Business Research (IJCBR)*, vol. 5, no. 4, pp. 1–29, 2014.
- [4] J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, and M. Asadpour, "Boosting methods for multi-class imbalanced data classification: an experimental review," *Journal of Big Data*, vol. 7, pp. 1–47, 2020.

- [5] Y. Lu, Y.-M. Cheung, and Y. Y. Tang, "Self-adaptive multiprototype-based competitive learning approach: A k-means-type algorithm for imbalanced data clustering," *IEEE transactions on cybernetics*, vol. 51, no. 3, pp. 1598–1612, 2019.
- [6] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, "Gaussian mixture models and k-means clustering," *Numerical recipes: the art of scientific computing*, pp. 842–850, 2007.
- [7] E. Shireman, D. Steinley, and M. J. Brusco, "Examining the effect of initialization strategies on the performance of gaussian mixture modeling," *Behavior research methods*, vol. 49, pp. 282–293, 2017.
- [8] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.
- [9] S. Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.
- [10] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media, 2013.
- [11] N. R. Pal, K. Pal, J. M. Keller, and J. C. Bezdek, "A possibilistic fuzzy c-means clustering algorithm," *IEEE transactions on fuzzy systems*, vol. 13, no. 4, pp. 517–530, 2005.
- [12] D.-M. Tsai and C.-C. Lin, "Fuzzy c-means based clustering for linearly and nonlinearly separable data," *Pattern recognition*, vol. 44, no. 8, pp. 1750–1760, 2011.
- [13] S. Krinidis and V. Chatzis, "A robust fuzzy local information c-means clustering algorithm," *IEEE transactions on image processing*, vol. 19, no. 5, pp. 1328–1337, 2010.
- [14] A. Coates and A. Y. Ng, "Learning feature representations with k-means," in *Neural Networks: Tricks of the Trade: Second Edition*. Springer, 2012, pp. 561–580.
- [15] B. Yang, X. Fu, N. D. Sidiropoulos, and M. Hong, "Towards k-means-friendly spaces: Simultaneous deep learning and clustering," in *international conference on machine learning*. PMLR, 2017, pp. 3861–3870.
- [16] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 132–149.
- [17] M. M. Fard, T. Thonet, and E. Gaussier, "Deep k-means: Jointly clustering with k-means and learning representations," *Pattern Recognition Letters*, vol. 138, pp. 185–192, 2020.
- [18] H. Xiong, J. Wu, and J. Chen, "K-means clustering versus validation measures: a data distribution perspective," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 779–784.
- [19] J. Noordam, W. Van Den Broek, and L. Buydens, "Multivariate image segmentation with cluster size insensitive fuzzy c-means," *Chemometrics and intelligent laboratory systems*, vol. 64, no. 1, pp. 65–78, 2002.
- [20] P.-L. Lin, P.-W. Huang, C.-H. Kuo, and Y. Lai, "A size-insensitive integrity-based fuzzy c-means method for data clustering," *Pattern Recognition*, vol. 47, no. 5, pp. 2042–2056, 2014.
- [21] S. Askari, "Fuzzy c-means clustering algorithm for data with unequal cluster sizes and contaminated with noise and outliers: Review and development," *Expert Systems with Applications*, vol. 165, p. 113856, 2021.
- [22] J. Liang, L. Bai, C. Dang, and F. Cao, "The k-means-type algorithms versus imbalanced data distributions," *IEEE Transactions on Fuzzy Systems*, vol. 20, no. 4, pp. 728–745, 2012.
- [23] S. Zeng, X. Duan, J. Bai, W. Tao, K. Hu, and Y. Tang, "Soft multiprototype clustering algorithm via two-layer semi-nmf," *IEEE Transactions on Fuzzy Systems*, 2023.
- [24] N. B. Karayiannis, "Meca: Maximum entropy clustering algorithm," in *Proceedings of 1994 IEEE 3rd international fuzzy systems conference*. IEEE, 1994, pp. 630–635.
- [25] R.-P. Li and M. Mukaidono, "A maximum-entropy approach to fuzzy clustering," in *Proceedings of 1995 IEEE International Conference on Fuzzy Systems*, vol. 4. IEEE, 1995, pp. 2227–2232.
- [26] D. Aloise, A. Deshpande, P. Hansen, and P. Papat, "Np-hardness of euclidean sum-of-squares clustering," *Machine learning*, vol. 75, pp. 245–248, 2009.
- [27] J. Wu, *Advances in K-means clustering: a data mining thinking*. Springer Science & Business Media, 2012.
- [28] L. Bottou and Y. Bengio, "Convergence properties of the k-means algorithms," *Advances in neural information processing systems*, vol. 7, 1994.
- [29] L. Groll and J. Jakel, "A new convergence proof of fuzzy c-means," *IEEE Transactions on Fuzzy Systems*, vol. 13, no. 5, pp. 717–720, 2005.
- [30] K. Zhou and S. Yang, "Effect of cluster size distribution on clustering: a comparative study of k-means and fuzzy c-means clustering," *Pattern Analysis and Applications*, vol. 23, pp. 455–466, 2020.
- [31] A. Gupta, S. Datta, and S. Das, "Fuzzy clustering to identify clusters at different levels of fuzziness: An evolutionary multiobjective optimization approach," *IEEE transactions on cybernetics*, vol. 51, no. 5, pp. 2601–2611, 2019.
- [32] D. Arthur and S. Vassilvitskii, "K-means++ the advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 2007, pp. 1027–1035.
- [33] M. Huang, Z. Xia, H. Wang, Q. Zeng, and Q. Wang, "The range of the value for the fuzzifier of the fuzzy c-means algorithm," *Pattern Recognition Letters*, vol. 33, no. 16, pp. 2280–2284, 2012.
- [34] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.
- [35] E. Anderson, "The species problem in iris," *Annals of the Missouri Botanical Garden*, vol. 23, no. 3, pp. 457–509, 1936.
- [36] W. Wolberg, M. Olvi, N. Street, and W. Street, "Breast Cancer Wisconsin (Diagnostic)," UCI Machine Learning Repository, 1995, DOI: <https://doi.org/10.24432/C5DW2B>.
- [37] S. Chawla and A. Gionis, "k-means-: A unified approach to clustering and outlier detection," in *Proceedings of the 2013 SIAM international conference on data mining*. SIAM, 2013, pp. 189–197.
- [38] Z. Zhang, Q. Feng, J. Huang, Y. Guo, J. Xu, and J. Wang, "A local search algorithm for k-means with outliers," *Neurocomputing*, vol. 450, pp. 230–241, 2021.
- [39] P. Verma, M. Sinha, and S. Panda, "Fuzzy c-means clustering-based novel threshold criteria for outlier detection in electronic nose," *IEEE Sensors Journal*, vol. 21, no. 2, pp. 1975–1981, 2020.
- [40] H. Yadav, J. Singh, and A. Gosain, "Experimental analysis of fuzzy clustering techniques for outlier detection," *Procedia Computer Science*, vol. 218, pp. 959–968, 2023.
- [41] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "nearest neighbor" meaningful?" in *Database Theory—ICDT'99: 7th International Conference Jerusalem, Israel, January 10–12, 1999 Proceedings 7*. Springer, 1999, pp. 217–235.
- [42] Y. LeCun, "The mnist database of handwritten digits," <http://yann.lecun.com/exdb/mnist/>, 1998.
- [43] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. 11, 2008.

## Supplemental Materials: Imbalanced Data Clustering using Equilibrium K-Means

### SUPPLEMENT

We evaluate the clustering performance of the proposed equilibrium K-means (EKM) on ten real datasets. Four datasets (Iris, Imbalanced Iris, Wisconsin Diagnostic Breast Cancer (WDBC), and Imbalanced WDBC) are introduced in detail in the main text. Herein we supplement the results for the remaining six datasets. They are from the UCI repository (available at <https://archive.ics.uci.edu/datasets>), which are: Wine, Ecoli, Image Segmentation (IS), Htru2, Rice, and Dry Bean. A detailed description of all ten datasets are summarized in TABLE SI, ordered according to the imbalance level measured by CV0, which we will define later. The dataset with a lower order has a higher CV0 value, indicating a higher imbalance level. In the ‘‘Feature’’ column of TABLE SI, the format ‘‘x (y)’’ indicates that the dataset has y features, of which x features are used for clustering. For instance, the Iris dataset has four features, of which two are used for clustering. Partial features are used for the sake of visualizing clustering results. In the ‘‘CV0’’ column, we calculate the coefficient of variation (CV) of the dataset to measure the level of data imbalance. CV is defined as the ratio of the standard deviation to the mean. Mathematically, given the number of instances in each class as  $N_1, \dots, N_K$ , we have

$$CV = \text{std}(\mathbf{N}) / \text{mean}(\mathbf{N}),$$

where

$$\text{mean}(\mathbf{N}) = \frac{\sum_{k=1}^K N_k}{K},$$

and

$$\text{std}(\mathbf{N}) = \sqrt{\frac{\sum_{k=1}^K (N_k - \text{mean}(\mathbf{N}))^2}{K - 1}}.$$

In general, the larger the CV value, the greater the imbalance of the data.

Each dataset is normalized so that each feature has zero mean and unit variance. The number of clusters is assumed to be equal to the number of classes for all datasets. For the supplementary six datasets, the smoothness parameter  $\alpha$  is chosen adaptively as follows:

$$\alpha = 2 / \bar{d}_{0n},$$

where  $\bar{d}_{0n} = \frac{1}{2} \sum_{n=1}^N \|\mathbf{x}_n\|_2^2 / N$ ,  $\mathbf{x}_n$  is the  $n$ -th data point, and  $N$  is the total number of data points. The value  $\bar{d}_{0n}$  can be interpreted as a data variance after the normalization of zero data mean. The normalized mutual information (NMI), adjusted Rand index (ARI), and clustering accuracy (ACC) of each algorithm on the ten datasets are presented in TABLE SII. Bold values indicate the best results. We also include three metrics: ‘‘CV1’’, ‘‘DCV’’ and ‘‘Running time’’ for readers’ interest. ‘‘CV1’’ is the CV value of the clustering result. ‘‘DCV’’ is the difference between CV0 and CV1, i.e.,  $\text{DCV} = \text{CV0} - \text{CV1}$ . A positive DCV value indicates that the data distribution of the true classes is more imbalanced than that of the generated clusters, while a negative DCV value

TABLE SI  
DETAILED DESCRIPTION OF REAL-WORLD DATASETS

Dataset	Instance	Feature	Class	CV0
Iris	150	2 (4)	3	0
IS	2310	19 (19)	7	0
Wine	178	13 (13)	3	0.1939
Rice	3810	7 (7)	2	0.2042
WDBC	569	3 (30)	3	0.3604
Dry Bean	13611	16 (16)	7	0.4951
Imbalanced Iris	120	2 (4)	2	0.9428
Ecoli	336	7 (7)	8	1.1604
Imbalanced WDBC	369	3 (30)	2	1.3222
Htru2	17898	8 (8)	2	1.4142

means the data distribution of the true classes is more uniform than that of the generated clusters. ‘‘Running time’’ is the sum of the running time of ten replicates.

We observe that the proposed EKM generally outperforms the other six algorithms on the last four datasets with CV values greater than 0.5. EKM outperforms other algorithms on two of the top six datasets with CV values less than 0.5. On the rest of the datasets, EKM still has a competitive performance. Overall, these results demonstrate that EKM is an efficient clustering algorithm for both balanced and imbalanced data.

TABLE SII  
EXPERIMENTAL RESULTS OF DIFFERENT ALGORITHMS ON REAL-WORLD DATASETS

Dataset	Measurement	HKM	FKM	MEFC	csiFKM	siibFKM	GMM	EKM
Iris	NMI	0.5945	0.5671	0.5194	0.5671	0.5675	<b>0.6775</b>	0.5457
	ARI	<b>0.5462</b>	0.5328	0.5291	0.5328	0.5281	0.5354	0.5134
	ACC	0.7867	0.7800	<b>0.7933</b>	0.7800	0.7800	0.7000	0.7733
	CV1	0.0346	0.0200	0.1217	0.0200	0.1442	0.8902	0.0346
	DCV	-0.0346	-0.0200	-0.1217	-0.0200	-0.1442	-0.8902	-0.0346
	Running time	0.0113	0.0111	0.0305	0.0142	0.1844	0.0715	0.0328
IS	NMI	0.5864	0.5953	0.6222	0.5971	0.5997	0.5657	<b>0.6463</b>
	ARI	0.4605	0.4973	0.4995	0.4962	0.3214	0.4456	<b>0.5161</b>
	ACC	0.5455	<b>0.6455</b>	0.5944	0.6433	0.5861	0.5641	0.5944
	CV1	0.8246	0.2450	0.6433	0.4617	1.0377	0.5037	0.6582
	DCV	-0.8246	-0.2450	-0.6433	-0.4617	-1.0377	-0.5037	-0.6582
	Running time	0.0321	0.4714	0.2616	0.7913	49.5369	1.0463	<b>0.7205</b>
Wine	NMI	0.8759	0.8759	0.8759	0.8610	0.3361	0.8466	<b>0.8920</b>
	ARI	0.8975	0.8975	0.8975	0.8804	0.1355	0.8649	<b>0.9134</b>
	ACC	0.9663	0.9663	0.9663	0.9607	0.5562	0.9551	<b>0.9719</b>
	CV1	0.1242	0.1242	0.1242	0.1219	1.1513	0.1070	0.1403
	DCV	0.0696	0.0696	0.0696	0.0720	-0.9574	0.0868	0.0535
	Running time	0.0075	0.0153	0.0088	0.0126	1.6469	0.0515	0.0094
Rice	NMI	0.5685	0.5688	<b>0.5699</b>	0.5554	0.5415	0.4891	0.5653
	ARI	0.6815	0.6824	<b>0.6833</b>	0.6642	0.6394	0.5649	0.6772
	ACC	0.9129	0.9131	<b>0.9134</b>	0.9076	0.9000	0.8759	0.9115
	CV1	0.2442	0.2183	0.2309	0.2977	0.3623	0.0163	0.2643
	DCV	-0.0401	-0.0141	-0.0267	-0.0935	-0.1581	0.1878	-0.0601
	Running time	0.0200	0.0438	0.0357	0.1032	2.1275	0.2303	0.0397
WBDC	NMI	0.4920	0.4870	<b>0.5005</b>	0.4423	0.2951	0.4618	0.4906
	ARI	0.5844	0.5963	<b>0.6257</b>	0.4402	0.2797	0.5722	0.5340
	ACC	0.8840	0.8875	<b>0.8963</b>	0.8366	0.7750	0.8787	0.8682
	CV1	0.5990	0.5393	0.4051	0.8227	0.9768	0.2461	0.7133
	DCV	-0.2386	-0.1790	-0.0447	-0.4623	-0.6164	0.1143	-0.3529
	Running time	0.0156	0.0226	0.0171	0.0385	1.7632	0.0894	0.0160
Dry Bean	NMI	0.7138	0.7055	0.7050	0.6990	0.6673	<b>0.7429</b>	0.6859
	ARI	0.6687	0.6676	0.6565	0.6607	0.6170	<b>0.6914</b>	0.5792
	ACC	0.7865	<b>0.8373</b>	0.7834	0.8001	0.8032	0.8160	0.7478
	CV1	0.5616	0.4697	0.5441	0.5154	0.5394	0.5480	0.6385
	DCV	-0.0666	0.0254	-0.0491	-0.0203	-0.0444	-0.0529	-0.1434
	Running time	0.2186	1.3271	0.6290	1.8517	139.5876	3.4712	0.8499
Imbalanced Iris	NMI	0.0001	0.0247	0.0061	0.6414	0.6870	<b>0.9101</b>	<b>0.9101</b>
	ARI	-0.0053	0.0049	-0.0023	0.7421	0.7865	<b>0.9582</b>	<b>0.9582</b>
	ACC	0.5250	0.5500	0.5250	0.9500	0.9583	<b>0.9917</b>	<b>0.9917</b>
	CV1	0.1179	0.0471	0.0236	1.0842	1.0607	0.9664	0.9664
	DCV	0.8250	0.8957	0.9192	-0.1414	-0.1179	-0.0236	-0.0236
	Running time	0.0057	0.0121	0.0142	0.0257	0.1474	0.0146	0.0115
Ecoli	NMI	0.6370	0.5837	0.6087	0.5999	0.6026	0.6015	<b>0.6530</b>
	ARI	0.4953	0.4113	0.4704	0.4176	0.4468	<b>0.6149</b>	0.5202
	ACC	0.6310	0.5863	0.6190	0.6012	0.6012	<b>0.6964</b>	0.6458
	CV1	0.6871	0.4671	0.7517	0.4482	0.7770	1.0893	0.7001
	DCV	0.4733	0.6933	0.4087	0.7123	0.3835	0.0711	0.4604
	Running time	0.0255	0.2672	0.0760	0.1770	5.3315	0.1939	0.0886
Imbalanced WBDC	NMI	0.0966	0.0848	0.0828	0.6846	0.3893	0.3038	<b>0.6907</b>
	ARI	0.0337	0.0184	0.0158	0.7855	0.5156	0.3323	<b>0.8308</b>
	ACC	0.6260	0.5827	0.5745	<b>0.9892</b>	0.9539	0.9051	<b>0.9892</b>
	CV1	0.2644	0.1418	0.1188	1.3529	1.2073	1.0540	1.3069
	DCV	1.0578	1.1804	1.2034	-0.0307	0.1150	0.2683	0.0153
	Running time	0.0141	0.0234	0.0094	0.0959	1.2488	0.0870	0.0535
Htru2	NMI	0.4068	0.4101	0.4076	0.0547	0.3854	0.2611	<b>0.5869</b>
	ARI	0.6071	0.5798	0.6076	0.0217	0.5967	0.3504	<b>0.7331</b>
	ACC	0.9366	0.9253	0.9366	0.9097	0.9428	0.8473	<b>0.9660</b>
	CV1	1.0891	1.0125	1.0882	1.4107	1.1813	0.7805	1.2342
	DCV	0.3251	0.4017	0.3260	0.0035	0.2329	0.6337	0.1800
	Running time	0.0825	0.2913	0.1094	2.0558	68.3254	0.4956	0.1119