FAIR CLUSTERING VIA EQUALIZED CONFIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Fair clustering aims at eliminating effects of sensitive information in clustering assignment. Existing work on fair clustering addresses this problem as a vanilla clustering with constraints that the distribution of protected groups on each cluster should be similar. However, existing criteria for fair clustering does not take into account clustering accuracy, and may restrain the performance of clustering algorithms. To tackle this problem, in this work, we propose a novel metric, *equalized confidence*, for fair clustering based on the predicted clustering confidence. Instead of enforcing similar distribution of sensitive attributes across different clusters, equalized confidence requires similar predicted confidence across different sensitive groups, bypassing the problem of disparities in statistical features across demographic groups. In light of the new metric, we propose a fair clustering method to learn a fair and good representation for clustering. Compared with conventional methods on fair clustering which try to adjust clustering assignment, our method focuses on learning a fair representation for downstream tasks. Our method proposes to eliminate the disparities of predicted soft labels of samples in different demographic groups using Sinkhorn divergence, as well as to learn clustering-favorable representations for clustering. Experimental results show that our method performs better or comparably than state-of-the-art methods, and that our proposed metric fits better under clustering accuracy.

1 INTRODUCTION

While machine learning models have been shown to achieve remarkable performance in different fields, there have been concerns that machine learning models can reveal real-world discrimination unless proper regularization, especially in high-stakes social domains (Larson et al., 2016; Chouldechova, 2017; Dressel & Farid, 2018). Research on fairness in machine learning has received much attention, and different notions and metrics have been proposed to quantify disparities of automatic decision-making system, among which disparate impact (DI) and equalized odds (EOd) has been widely studied and adopted in fair classification literature as group fairness metrics.

In recent years, fair clustering has become an arising topic in algorithmic fairness, where the goal is to obtain a fair clustering assignment for unsupervised clustering tasks. Most existing work on fair clustering formulates the problem of fair clustering as achieving balance within each cluster, where samples from each sensitive group is expected to make up similar portions of the cluster. In the context of group fairness, this formulation naturally agrees with DI, and a clustering assignment with better balance also achieves lower DI. However, there are several problems with existing fair clustering methods: most works focus on adjusting clustering assignment, instead of learning fair representations for input features, limiting their applications in downstream tasks, and existing works are mostly validated on low-dimensional or linear-separable data, while their performance on high-dimensional data is not yet adequately discussed. Furthermore, existing metrics for fair clustering do not agree with post-hoc notions in group fairness, including EOd and equal opportunity (EOp). Furthermore, existing metrics for fair clustering do not agree with post-hoc notions in group fairness, including EOd and EOp, as they do not take into account the performance disparities across sensitive groups. Therefore we ask: *Can we define similar metrics for post-hoc performance disparity in the context of fair clustering?*

Due to the unsupervised nature of clustering, it is infeasible to directly adopt current post-hoc notions into fair clustering or incorporate relaxations of such notions with clustering process. Instead, inspired by previous relaxations in group fairness (Madras et al., 2018), we consider regularizing the predicted soft clustering assignment of each sensitive group, where the idea is to make sure that the predicted soft assignment within each cluster is equal for all sensitive subgroups. Following the formulation of our fair clustering metric, *equalized confidence*, we propose a novel fair clustering framework, where we use Sinkhorn divergence to measure the confidence difference in each cluster, along with clustering regularizers to preserve clustering structure and encourage high predicted confidence, and contrastive loss to learn linear-separable and clustering-favorable representations. We summarize our contribution as follows:

1. **Fairness metric**: We formulate a novel fair clustering metric, *equalized confidence*, to relate with post-hoc notions in group fairness.

2. **Fair clustering**: We propose a novel framework for fair clustering using Sinkhorn divergence to address confidence difference in fair clustering.

3. We validate the effectiveness of our method in improving fairness on four benchmark datasets, and we show the connection between notions in group fairness and fair clustering on color-reverse MNIST dataset.

2 Related Work

2.1 Fair Classification

Fair classification aims at eliminating the effect of sensitive attributes on machine learning models, and several notions have been proposed to quantify the disparities across senstive groups (Dwork et al., 2012; Hardt et al., 2016; Kusner et al., 2017). Our work is most closely related with group fairness notion in fair classification. In response, different methods have been proposed for obtaining classifiers with fairness guarantee. Generally, these methods can be divided into three categories: preprocessing (Sattigeri et al., 2018; Creager et al., 2019; Jiang & Nachum, 2020; Jang et al., 2021), where the goal is to map input feature or label to some latent space, such that the processed data obtains desired property in terms of fairness or disentanglement; inprocessing (Zafar et al., 2017; Agarwal et al., 2018; Roh et al., 2021; Wang et al., 2022), where relaxations of fairness metrics are incorporated during training as regularization terms or as means for sample reweighing; and postprocessing (Hardt et al., 2016; Kim et al., 2019; Jang et al., 2022), where the goal is to adjust the decision threshold in different sensitive groups to achieve expected fairness parity.

2.2 Fair Clustering

The concepts and notions of fair clustering are first proposed in (Chierichetti et al., 2017), where fairness in clustering is formulated as balanced distribution of sensitive attributes in each cluster. Chierichetti et al. (2017) proposes to first perform fair decomposition which separates input data into smaller fairlets with balance guarantee, and uses a K-center algorithm to achieve fairness. Since then, several works have been proposed to obtain balanced fair clustering (Backurs et al., 2019; Bera et al., 2019; Abraham et al., 2019; Ahmadian et al., 2020; Liu & Vicente, 2021). These methods usually focus on adjusting the clustering assignment to improve balance in each cluster and formulate the problem as clustering with relaxed balance guarantee, and generally deals with conventional clustering problems including K-means and K-center. Instead, Li et al. (2020) propose a deep fair clustering method to learn fair embedding for clustering, where the latent representation is expected to be independent of sensitive attributes. Other notions for fair clustering includes individual fairness (Mahabadi & Vakilian, 2020) and proportional fairness (Chen et al., 2019), where the idea is to ensure bounded distance between samples and their corresponding clustering centers, without accessing or assuming sensitive information.

3 Problem Definition

3.1 Balance

We begin by discussing the notion of balance in fair clustering. For simplicity, we discuss the case with binary sensitive attribute. Generally, given a set of samples $\{(x_i, a_i), 1 \leq i \leq N\}$ with $x_i \in \mathbb{R}^P$ the input feature and $a_i \in \{0, 1\}$ the binary sensitive attribute, let $f : \mathbb{R}^P \rightarrow \mathbb{R}^d$ be the function of encoder, $g : \mathbb{R}^d \to [0,1]^K$ be the function of clustering assignment which divides the samples into K clusters, and let $c : [0,1]^K \to \{0,1\}^K$ be the hard clustering assignment. A clustering algorithm is said to achieve balance if the distribution of the protected group membership is approximately equal in all the clusters:

$$\frac{1}{N}\sum_{i=1}^{N} a_i = \frac{1}{|\mathbb{C}_k|}\sum_{i\in\{j|c(g(f(x_j)))=k\}} a_i, \forall k \in [K],$$

where $\mathbb{C}_k$ refers to the set of samples in the k-th cluster. And a clustering algorithm is said to be ϵ-balance if the portion of protected group membership is lower-bounded by ϵ in each cluster:

$$\inf_{k,a} \frac{\sum_i \mathbb{1}(c(g(f(x_i))) = k, a_i = a)}{\sum_i \mathbb{1}(c(g(f(x_i))) = k)} \geq \epsilon. \tag{1}$$

Balance is naturally related to the content of DI in group fairness. Consider binary clustering assignment under binary ground truth label for simplicity, suppose the clustering algorithm achieve ϵ-balance and without loss of generality, assume $a = 1$ to be the major group, it is easy to see that the predicted positive rates of different sensitive groups are approximately equal, as each cluster follows similar distribution over sensitive information:

$$\begin{aligned} DI &= \left|\frac{\sum_i \mathbb{1}(c(g(f(x_i))) = 1, a_i = 0)}{|\mathbb{N}_0|} - \frac{\sum_j \mathbb{1}(c(g(f(x_j))) = 1, a_j = 1)}{|\mathbb{N}_1|}\right| \\ &= \left|\frac{\sum_i \mathbb{1}(c(g(f(x_i))) = 1, a_i = 0)}{|\mathbb{N}_1|}\delta - \frac{\sum_j \mathbb{1}(c(g(f(x_j))) = 1, a_j = 1)}{|\mathbb{N}_1|}\right| \\ &= \left|\frac{\delta|\mathbb{C}_1 \cap \mathbb{N}_0| - |\mathbb{C}_1 \cap \mathbb{N}_1|}{|\mathbb{N}_1|}\right| \\ &\leq \left|\frac{(1-\epsilon)\delta|\mathbb{C}_1| - \epsilon|\mathbb{C}_1|}{|\mathbb{N}_1|}\right| = \left|\frac{\delta|\mathbb{C}_1| - \epsilon(1+\delta)|\mathbb{C}_1|}{|\mathbb{N}_1|}\right|, \end{aligned} \tag{2}$$

where $\mathbb{N}_j$ refers to the set of samples in j-th protected group, $\delta = \frac{|\mathbb{N}_1|}{|\mathbb{N}_0|}$ is the empirical ratio, and the last inequality is obtained based on the definition of ϵ-balance in equation 1. Given δ, we can observe that the upper bound in equation 2 is monotonically decreasing w.r.t. ϵ, that is, greater balance level ϵ indicates lower upper-bound of DI. As shown in Fig. 1, when the clustering algorithm achieves perfect balance, the DI between blue and orange groups are also eliminated.

However, one drawback regarding DI is that it does not take into account the difference in base rate between different demographic groups. If the distribution of training samples is uneven or biased against certain demographic groups, strictly enforcing DI on the predictor could result in good decisions on major group but poor or even random decisions on minor group. Besides, achieving zero DI could be against an optimal classifier when statistical features of different demographic groups vary. Similar problems also arises in fair clustering content when applying balance as fairness metric. As in Fig. 1, the cluster assignment shown achieves perfect 50%-balance (that is, the ratio of data from different sensitive groups in each cluster is exactly $1 : 1$). However, the decision boundary implied by current clustering induces severe disparities in both true positive rate (TPR) and true negative rate (TNR), with EOd 1.027, while the optimal decision boundary under current embedding reduces EOd by over 80%.

3.2 Equalized Confidence

In light of the aforementioned limitations of balance, we seek to define alternative metrics for fair clustering. Inspired by (Hardt et al., 2016), we can define equality of error rates across different sensitive groups in clustering:

$$p(c(g(f(x))) \neq y|y, a) = p(c(g(f(x))) \neq y|y), \forall a.$$

Similarly, we can define the EOd in fair clustering as follows:

$$EOd = \sum_{y,a}\sum_{a'\neq a} |p(c(g(f(x))) \neq y|y, a) - p(c(g(f(x))) \neq y|y, a')|.$$

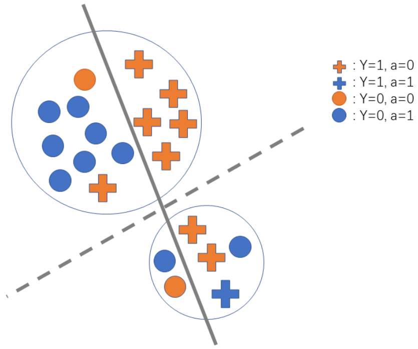

Figure 1: Demonstration of how balance does not ensure fairness in terms of clustering accuracy. The circle/cross represent negative/positive samples, and different colors show different sensitive groups. The dashed line represents the decision boundary obtained through fair clustering (w.r.t. balance), and the solid line indicates the optimal linear decision boundary under current embedding. Although the clustering assignment achieves perfect balance, the disparities in TPR and TNR are still large.

However, since equality of error rates is a post-hoc criteria and we do not have access to label information during training or validation, it is hard to incorporate such fairness metrics with clustering tasks. Instead, we consider a relaxed measure, where our goal is to achieve equalized predicted confidence across all the clusters. We define equalized confidence as follows:

$$\mathbb{E}[g(f(x))|c, a] = \mathbb{E}[g(f(x))|c].$$

And the confidence difference can be formulated as

$$\text{confidence difference} = \sum_{c,a} \sum_{a' \neq a} |\mathbb{E}[g(f(x))|c, a] - \mathbb{E}[g(f(x))|c, a']|.$$

4 Method

In this section, we discuss our fair clustering framework, bringing together the expectation of equalized confidence and clustering-favorable network. Specifically, we formulate the requirement of equalized confidence as a regularization term during training using Sinkhorn divergence, with several additional regularizer for ensuring accuracy. We start by introducing the formulation of Sinkhorn divergence.

4.1 Sinkhorn divergence

Sinkhorn divergence interpolates between Wasserstein distance and Maximum Mean Discrepancy (MMD) (Feydy et al., 2019). Consider binary sensitive attribute for simplicity, let $X^a \in \mathbb{R}^{|\mathbb{N}_a| \times P}$ be the set of elements of the a-th group, the pairwise kernel distances between elements of X^0 and X^1 can be defined through the cost matrix:

$$M := \left[-k\left(X_i^0, X_j^1\right)\right]_{ij} \in \mathbb{R}^{|\mathbb{N}_0| \times |\mathbb{N}_1|},$$

where k is some characteristic positive definite kernel. And the smoothed Wasserstein distance can be defined as

$$W_\lambda(X^0, X^1) := \min_{T \in U_{|\mathbb{N}_0|, |\mathbb{N}_1|}} \lambda \langle T, M \rangle - E(T),$$

where $\langle T, M\rangle$ is the Frobenius dot-product of T and M, $E(T) := -\sum_{ij} T_{ij}\log(T_{ij})$ is the entropy of T, and $U_{|\mathbb{N}_0|,|\mathbb{N}_1|}$ is the set of empirical joint distributions of X^0 and X^1 such that their row and column marginals equal $\frac{\mathbf{1}_{|\mathbb{N}_0|}}{|\mathbb{N}_0|}$and $\frac{\mathbf{1}_{|\mathbb{N}_1|}}{|\mathbb{N}_1|}$ respectively:

$$U_{|\mathbb{N}_0|,|\mathbb{N}_1|} := \left\{ T \in \mathbb{R}_+^{|\mathbb{N}_0|\times|\mathbb{N}_1|} : T\mathbf{1}_{|\mathbb{N}_1|} = \frac{\mathbf{1}_{|\mathbb{N}_0|}}{|\mathbb{N}_0|}, T^T\mathbf{1}_{|\mathbb{N}_0|} = \frac{\mathbf{1}_{|\mathbb{N}_1|}}{|\mathbb{N}_1|} \right\}.$$

And the Sinkhorn divergence between X^0 and X^1 is formulated as

$$S_\lambda(X^0, X^1) = 2W_\lambda(X^0, X^1) - W_\lambda(X^0, X^0) - W_\lambda(X^1, X^1)$$

As $\lambda \to \infty$, we have S_λ the W_2 distance projected by some positive definite kernel, and as $\lambda \to 0$, we have T the maximal entropy table in $U_{|\mathbb{N}_0|,|\mathbb{N}_1|}$, i.e., the outer product $\frac{(\mathbf{1}_{|\mathbb{N}_0|}\mathbf{1}_{|\mathbb{N}_1|}^T)}{|\mathbb{N}_0||\mathbb{N}_1|}$ of the marginals, and S_λ follows the unbiased estimator of MMD (Cuturi, 2013; Ramdas et al., 2017; Feydy et al., 2019):

$$\text{MMD} := \frac{1}{|\mathbb{N}_0|^2}\sum_{i,j=1}^{|\mathbb{N}_0|} k\left(X_i^0, X_j^0\right) + \frac{1}{|\mathbb{N}_1|^2}\sum_{i,j=1}^{N_1} k\left(X_i^1, X_j^1\right) - \frac{2}{|\mathbb{N}_0||\mathbb{N}_1|}\sum_{i=1}^{|\mathbb{N}_0|}\sum_{j=1}^{|\mathbb{N}_1|} k\left(X_i^0, X_j^1\right).$$

And the optimal empirical distribution T^* can be obtained via alternating optimization:

$$T^* = ue^{-\lambda M}v.$$

4.2 Objective Function

Sinkhorn loss. We use Sinkhorn divergence between soft predictions of different sensitive subgroups to quantify disparities in predicted confidence across different sensitive groups. Specifically, let $P_{jk} := \{g(f(x_i)) | a_i = j, \arg\max_k(g_k(f(x_i))) = k\}$ be the set of predicted soft assignment in the j,k-th group, we formulate the Sinkhorn loss as follows:

$$L_\lambda^f = \sum_{j,k}\sum_{j'\neq j} S_\lambda(G_{j'k}, G_{jk}). \tag{3}$$

Contrastive loss. Solving equation 3 along could lead to a degraded solution where the encoder reduces to constant assignment. In light of this, we apply contrastive loss during training, where the goal is to avoid model collapse, as well as to learn clustering-favorable representations for downstream tasks. For each training batch of size n, we apply random augmentation on each sample twice, resulting in a subset $\{\tilde{x}_i, 1 \leq i \leq 2n\}$. For each sample pair $(\tilde{x}_i, \tilde{x}_i^{\text{pos}})$, the contrastive loss is calculated by

$$L_{contr}(\tilde{x}_i) = -\log\frac{\exp(\text{sim}(f_\theta(\tilde{x}_i), f_\theta(\tilde{x}_i^{\text{pos}}))/\tau)}{\sum_{j\neq i}\exp\left(\text{sim}\left(f_\theta(\tilde{x}_i), f_\theta(\tilde{x}_j)\right)/\tau\right)},$$

where sim is the cosine similarity metric and τ is the scaling hyper-parameter. The overall contrastive loss on current batch takes the average over all the augmented samples:

$$L_{contr} = \sum_{i=1}^{2n} L_{contr}(\tilde{x}_i).$$

Clustering regularizer. Inspired by (Caron et al., 2018; Li et al., 2020), we use the following regularizer to encourage high predicted confidence:

$$L_{re} = \sum_i\sum_j -g_j(f(x_i))\log(g_j(f(x_i))).$$

We defer the detailed discussion regarding the choice of regularizers to appendix.

Training objective. The overall training objective can be written as follows:

$$L = \alpha_f L_\lambda^f + L_{contr} + \alpha_r L_{re}.$$

Clustering assignment. Following (Xie et al., 2016), we use student's t-distribution for clustering assignment:

$$g_k(f(x)) = \frac{\left(1 + \|f(x) - C_k\|^2 / \alpha\right)^{-\frac{\alpha+1}{2}}}{\sum_{k' \in [K]} \left(1 + \|f(x) - C_{k'}\|^2 / \alpha\right)^{-\frac{\alpha+1}{2}}},$$

where $[K]$ is the set indices of clustering center and α is the degree of freedom in student's t-distribution.

5 Experiments

5.1 Experimental Setup

We validate our method on four datasets:

- **Color reverse MNIST**: We reverse the pixel value of images in MNIST dataset by $p' = 255 - p$ and concatenate the reversed data with original dataset, and the sample source is set as sensitive attribute. We apply two different labelling methods: normal labelling $(0 - 9)$ and binary labelling (Krueger et al., 2021), where digital $0 - 4$ is mapped to class 0 and $5 - 9$ to class 1.
- **MNIST-USPS**: The MNIST-USPS dataset is constructed using all the training samples from MNIST and USPS dataset, and we set the data source as sensitive attribute. Similar to Color reverse MNIST, We consider two different labelling setting: normal labelling and binary labelling.
- **Office-31**: The Office-31 dataset contains 31 object categories in three domains: $Amazon$, $DSLR$ and $Webcam$. We use $Amazon$ and $Webcam$ for experiments which has the largest domain diversity, and the sample source is set as sensitive attribute.
- **MTFL**: Multi-task Facial Landmark (MTFL) dataset is annotated with gender, smiling, glasses and head pose annotations. We choose glasses as sensitive attribute and gender as ground truth label.

We implement our method in PyTorch 1.10.1 with one NVIDIA RTX-3090 GPU. We use ResNet-50 as the encoder structure, and α for the clustering assignment is set as 1. We use clustering accuracy and normalized mutual information (NMI) for performance evaluation, and we apply four different metrics for fairness evaluation: disparate impact (DI) (Rutherglen, 1987), balance, confidence difference 3.2 and equalized odds (EOd) (Hardt et al., 2016). Following our previous discussion, we consider balance as measure of the worst-case disparities in distribution of predicted soft assignments across different sensitive groups among all the clusters, and we calculate the metric as follows:

$$\text{Balance} = \min_{i,j} \frac{|\mathbb{C}_i \cap \mathbb{N}_j|}{|\mathbb{C}_i|},$$

where $\mathbb{C}_i$ refers to the set of samples in the i-th cluster, and $\mathbb{N}_j$ refers to the set of samples in j-th protected group. Larger balance value indicates smaller distributional disparities across sensitive groups. Due to distributional disparities, we have balance upper-bounded by the empirical ratio:

$$\text{Balance} \leq \min_{i} \frac{|\mathbb{N}_i|}{\sum_k |\mathbb{N}_k|}$$

We compare our method with four related clustering methods: **DFC** - Deep fair clustering method with adversarial training to improve balance (Li et al., 2020). **Falg** - Fair clustering with balance constraints (Bera et al., 2019). **K-means** - Vanilla Clustering by K-means. **ResNet** - Vanilla clustering with feature preprocessed by ResNet-50. We repeat experiments on each dataset three times and report the average results. In each repetition, we randomly split data into 64% training data, 16% validation data and 20% testing data. All the methods evaluated are trained and tested on the same data partitions each time. Hyperparameters of compared methods are set as suggested by the authors (Bera et al., 2019; Li et al., 2020). The hyperparameter of our method is tuned to find best confidence difference gap on validation data.

5.2 Experimental Results

Experimental results on the four datasets are shown in Tab. 1 - 6. Compared with baseline methods, our method achieves better accuracy on different datasets and under different labelling, and our method shows an improvement in confidence difference and EOd, which is in line with our discussion and formulation of equalized confidence. We also notice that on the four datasets, our method also improves balance and DI, and this shows that when the base rate difference is not extreme, it is possible to improve DI and EOd simultaneously, which is in line with previous empirical observation on fair classification. Compared with other fair clustering methods, our method achieve comparable performance in terms of balance and DI, with improvement in confidence difference and EOd, and we also notice that our method achieves better clustering accuracy on several datasets, owing to the introduction of contrastive loss. We show part of ablation study in Tab. 7-10 to validate the effectiveness of contrastive loss in improving clustering accuracy, and we defer full results to appendix.

Method	accuracy(%)	NMI(%)	Balance	DI	Conf. Dif.	EOd
Ours	81.46±2.15	77.82±1.26	0.05±0.03	0.10±0.02	0.02±0.01	0.09±0.02
DFC	82.51±2.27	78.24±1.86	0.09±0.01	0.07±0.03	0.06±0.02	0.14±0.03
K-means	59.54±2.31	69.63±1.54	0.02±0.02	0.17±0.03	0.05±0.02	0.13±0.03
ResNet	72.61±1.53	64.51±2.27	0.02±0.01	0.16±0.02	0.05±0.01	0.16±0.03
Falg	63.14±2.54	51.16±2.84	0.06±0.03	0.09±0.03	0.04±0.02	0.13±0.02

Table 1: Experimental results on MNIST-USPS dataset. Conf. Dif. refers to confidence difference. The dataset is constructed using all the samples from MNIST and USPS dataset, and the sample source is set as sensitive attribute. Higher accuracy/NMI indicate better clustering performance. Higher balance or lower DI/Confidence Difference/EOd correspond to better fairness.

Method	accuracy(%)	NMI(%)	Balance	DI	Conf. Dif.	EOd
Ours	73.67±1.86	70.23±1.36	0.07±0.03	0.07±0.02	0.02±0.01	0.08±0.02
DFC	72.26±1.17	70.41±1.21	0.09±0.04	0.06±0.03	0.06±0.01	0.16±0.03
K-means	63.21±1.84	52.27±1.87	0.04±0.03	0.12±0.03	0.10±0.03	0.17±0.02
ResNet	69.67±2.44	68.43±1.69	0.03±0.02	0.11±0.03	0.08±0.02	0.15±0.03
Falg	60.35±2.57	51.27±2.52	0.10±0.04	0.04±0.01	0.04±0.02	0.14±0.02

Table 2: Experimental results on MNIST-USPS dataset under binary labelling. The dataset is constructed using all the samples from MNIST and USPS dataset, and the sample source is set as sensitive attribute.

Method	accuracy(%)	NMI(%)	Balance	DI	Conf. Dif.	EOd
Ours	71.36±2.27	72.31±2.26	0.11±0.02	0.07±0.02	0.03±0.01	0.07±0.02
DFC	69.23±1.34	71.82±1.64	0.12±0.02	0.06±0.02	0.03±0.01	0.12±0.03
K-means	59.45±2.44	60.61±1.69	0.07±0.03	0.13±0.03	0.07±0.04	0.15±0.03
ResNet	64.45±1.58	69.25±1.21	0.06±0.02	0.15±0.04	0.06±0.03	0.17±0.03
Falg	67.73±2.24	71.18±1.85	0.11±0.02	0.06±0.03	0.04±0.02	0.13±0.02

Table 3: Experimental results on Office-31 dataset. The Amazon and Webcam domains are chosen for our experiments, and the domain source is set as the protected attribute.

5.3 Experimental results on color reverse MNIST with varying data distribution

We move on to discuss the connections and conflicts between metrics in group fairness and fair clustering. Specifically, we consider resampling the color reverse MNIST with binary labelling, and

Method	accuracy(%)	NMI(%)	Balance	DI	Conf. Dif.	EOd
Ours	65.47±1.86	74.41±2.54	0.31±0.04	0.09±0.02	0.03±0.01	0.07±0.02
DFC	57.73±1.64	67.73±2.25	0.34±0.04	0.08±0.03	0.05±0.01	0.12±0.02
K-means	42.27±1.87	41.17±2.43	0.11±0.03	0.17±0.04	0.09±0.03	0.14±0.03
ResNet	59.54±2.31	69.63±1.54	0.14±0.02	0.16±0.04	0.08±0.03	0.13±0.03
Falg	31.37±1.53	21.24±1.75	0.29±0.03	0.12±0.03	0.07±0.02	0.13±0.02

Table 4: Experimental results on color reverse MNIST dataset. The sample source from reversed or original dataset is set as the protected attribute.

Method	accuracy(%)	NMI(%)	Balance	DI	Conf. Dif.	EOd
Ours	62.27±2.31	69.61±2.54	0.34±0.03	0.08±0.02	0.02±0.01	0.06±0.02
DFC	52.27±2.46	62.34±2.26	0.37±0.02	0.07±0.03	0.05±0.01	0.13±0.02
K-means	39.25±2.23	41.12±2.26	0.14±0.02	0.18±0.04	0.10±0.03	0.14±0.03
ResNet	54.43±2.67	64.41±2.17	0.12±0.02	0.19±0.04	0.11±0.03	0.14±0.03
Falg	29.27±2.16	19.84±1.47	0.31±0.04	0.09±0.02	0.05±0.03	0.13±0.03

Table 5: Experimental results on color reverse MNIST dataset under binary labelling. The sample source from reversed or original dataset is set as the protected attribute.

Method	accuracy(%)	NMI(%)	Balance	DI	Conf. Dif.	EOd
Ours	72.23±1.86	20.23±1.46	0.04±0.03	0.11±0.02	0.02±0.01	0.09±0.02
DFC	70.76±2.62	19.94±1.43	0.07±0.02	0.07±0.03	0.06±0.01	0.14±0.02
K-means	46.54±3.67	11.34±2.45	0.01±0.01	0.16±0.03	0.09±0.03	0.17±0.03
ResNet	64.83±1.47	18.14±1.44	0.02±0.02	0.19±0.04	0.09±0.02	0.18±0.04
Falg	65.63±2.11	18.87±1.25	0.05±0.02	0.09±0.02	0.06±0.03	0.14±0.03

Table 6: Experimental results on MTFL dataset. Glasses-wearing is chosen as sensitive attribute and gender is chosen as label.

Method	Accuracy	NMI
Our method	72.23±1.86%	20.23±1.46%
Our method (w/o contrastive loss)	69.62±1.36%	19.14±1.58%
Our method (w/o Sinkhorn divergence)	74.27±1.62%	20.84±2.31%
Our method (w/o regularization)	71.34±1.18%	19.83±2.64%

Table 7: Ablation study on MTFL dataset.

Method	Balance	DI	Conf. Dif.	EOd
Our method	0.13±0.04	0.09±0.02	0.02±0.01	0.09±0.02
Our method (w/o contrastive loss)	0.12±0.03	0.10±0.02	0.02±0.01	0.11±0.02
Our method (w/o Sinkhorn divergence)	0.08±0.04	0.12±0.02	0.05±0.01	0.16±0.03
Our method (w/o regularization)	0.13±0.04	0.09±0.02	0.03±0.01	0.09±0.02

Table 8: Ablation study on MTFL dataset.

Method	Clustering accuracy	NMI
Our method	71.36±2.27%	72.31±2.26%
Our method (w/o contrastive loss)	67.57±2.51%	70.52±2.64%
Our method (w/o Sinkhorn divergence)	73.51±1.72%	72.67±1.36%
Our method (w/o regularization)	71.23±1.87%	72.26±2.51%

Table 9: Ablation study on Office-31 dataset.

we vary the inner-class distribution w.r.t. sensitive attribute for both positive and negative samples to increase the disparities between distributions of the two subgroups, while keeping the number of samples from each group identical. As shown in Fig. 2, enforcing balance constraints during training leads to large increase in Eod while relatively small increase in DI as the degree of imbalance

Method	Balance	DI	Conf. Dif.	EOd
Our method	0.11±0.02	0.07±0.02	0.03±0.01	0.07±0.02
Our method (w/o contrastive loss)	0.11±0.02	0.07±0.02	0.04±0.02	0.09±0.02
Our method (w/o Sinkhorn divergence)	0.08±0.02	0.10±0.02	0.06±0.01	0.14±0.02
Our method (w/o regularization)	0.12±0.02	0.07±0.02	0.03±0.01	0.08±0.02

Table 10: Ablation study on Office-31 dataset.

becomes extreme. This shows the alignment between balance and DI and disaccord between balance and Eod under larger base rate difference. However, compared with other methods, our method shows a smaller increase in Eod and larger increase in DI and large decrease in balance as the degree of imbalance increases, which shows the alignment between confidence difference and EOd and disaccord between confidence difference and DI.

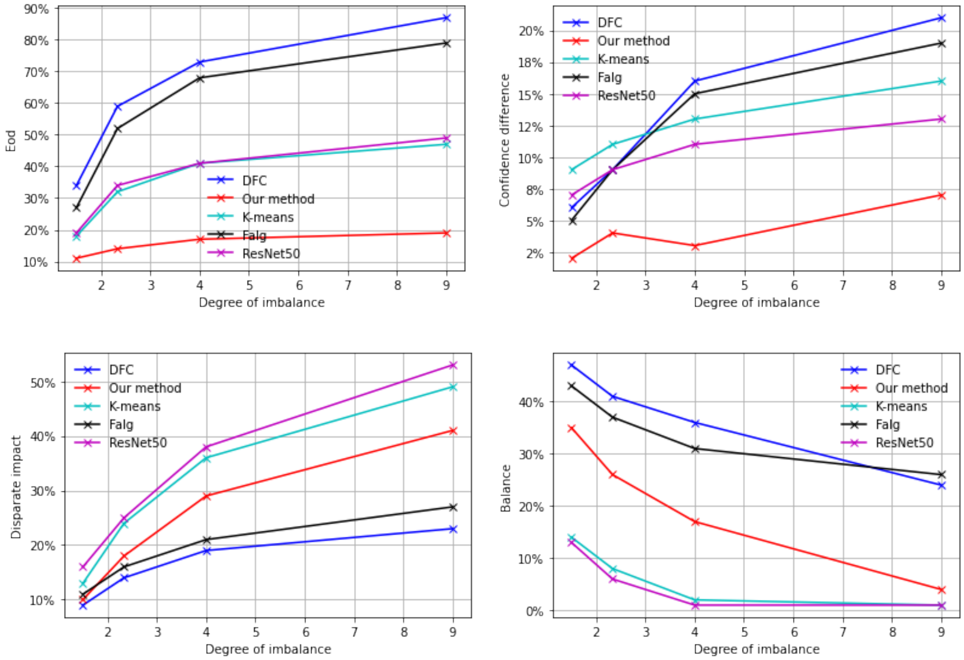

Figure 2: Results on color-reverse MNIST dataset under binary labelling. The number of reversed and original images are the same, but the number of positive samples in each group varies. The degree of imbalance is calculated by $\frac{N_{A=1,y=1}}{N_{A=0,y=1}}$.

6 Conclusion

Fair clustering with performance parity has yet received less attention. In this paper, we discuss the inherent connection between existing fair clustering notion and DI, and we consider a novel fair clustering metric, equalized confidence, which naturally adopts the property of error-based group fairness metrics. We propose a novel fair clustering method for learning fair and clustering-favorable representation, which utilizes Sinkhorn divergence to regulate the predicted confidence difference between different sensitive groups. We validate from experiments that our method achieves better or comparable performance in terms of both clustering accuracy and fairness, with an improvement in confidence difference and EOd compared with baseline fair clustering methods, and we show from experiments the connection and conflicts between notions in group fairness and fair clustering. Future directions of interest include alternative relaxations for other post-hoc fairness notions in fair clustering and robustness of existing fair clustering notions under noisy sensitive attribute or adversarial perturbation.

REFERENCES

Savitha Sam Abraham, Sowmya S Sundaram, et al. Fairness in clustering with multiple sensitive attributes. *arXiv preprint arXiv:1910.05113*, 2019.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pp. 60–69. PMLR, 2018.

Sara Ahmadian, Alessandro Epasto, Ravi Kumar, and Mohammad Mahdian. Fair correlation clustering. In *International Conference on Artificial Intelligence and Statistics*, pp. 4195–4205. PMLR, 2020.

Arturs Backurs, Piotr Indyk, Krzysztof Onak, Baruch Schieber, Ali Vakilian, and Tal Wagner. Scalable fair clustering. In *International Conference on Machine Learning*, pp. 405–413. PMLR, 2019.

Suman Bera, Deeparnab Chakrabarty, Nicolas Flores, and Maryam Negahbani. Fair algorithms for clustering. *Advances in Neural Information Processing Systems*, 32, 2019.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 132–149, 2018.

Xingyu Chen, Brandon Fain, Liang Lyu, and Kamesh Munagala. Proportionally fair clustering. In *International Conference on Machine Learning*, pp. 1032–1041. PMLR, 2019.

Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. Fair clustering through fairlets. *Advances in Neural Information Processing Systems*, 30, 2017.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, pp. 1436–1445. PMLR, 2019.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2681–2690. PMLR, 2019.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Taeuk Jang, Feng Zheng, and Xiaoqian Wang. Constructing a fair classifier with generated fair data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7908–7916, 2021.

Taeuk Jang, Pengyi Shi, and Xiaoqian Wang. Group-aware threshold adaptation for fair classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6988–6995, 2022.

Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 702–712. PMLR, 2020.

Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 247–254, 2019.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pp. 5815–5826. PMLR, 2021.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

J Larson, S Mattu, L Kirchner, and J Angwin. Compas analysis. *GitHub, available at: https://github.com/propublica/compas-analysis*, 2016.

Peizhao Li, Han Zhao, and Hongfu Liu. Deep fair clustering for visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9070–9079, 2020.

Suyun Liu and Luis Nunes Vicente. A stochastic alternating balance k-means algorithm for fair clustering. *arXiv preprint arXiv:2105.14172*, 2021.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pp. 3384–3393. PMLR, 2018.

Sepideh Mahabadi and Ali Vakilian. Individual fairness for k-clustering. In *International Conference on Machine Learning*, pp. 6586–6596. PMLR, 2020.

Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

Yuji Roh, Kangwook Lee, Steven Whang, and Changho Suh. Sample selection for fair and robust training. *Advances in Neural Information Processing Systems*, 34:815–827, 2021.

George Rutherglen. Disparate impact under title vii: an objective theory of discrimination. *Va. L. Rev.*, 73:1297, 1987.

Prasanna Sattigeri, Samuel C Hoffman, Vijil Chenthamarakshan, and Kush R Varshney. Fairness gan. *arXiv preprint arXiv:1805.09910*, 2018.

Jialu Wang, Xin Eric Wang, and Yang Liu. Understanding instance-level impact of fairness constraints. In *International Conference on Machine Learning*, pp. 23114–23130. PMLR, 2022.

Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pp. 478–487. PMLR, 2016.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pp. 1171–1180, 2017.