

## A APPENDIX

### A.1 DISCUSSION OF CLUSTERING REGULARIZER

The clustering regularizer by (Li et al., 2020) takes the following form:

$$\begin{aligned}
KL(P||Q) &= \sum_{i=1}^N \sum_{k \in [K]} p_{ki} \log \frac{p_{ki}}{q_{ki}} \\
&= \sum_{i=1}^N \sum_{k \in [K]} p_{ki} \log \frac{p_{ki}}{\frac{(p_{ki})^2 / \sum_{a_j=a} p_{kj}}{\sum_{k' \in [K]} ((p_{k'i})^2 / \sum_{a_i=a} p_{ki'})}} \\
&= \sum_{i=1}^N \sum_{k \in [K]} p_{ki} \log p_{ki} - p_{ki} \log \frac{(p_{ki})^2 / \sum_{a_j=a} p_{kj}}{\sum_{k' \in [K]} ((p_{k'i})^2 / \sum_{a_i=a} p_{ki'})} \\
&= \sum_{i=1}^N \sum_{k \in [K]} -p_{ki} \log p_{ki} + p_{ki} \log \sum_{a_j=a} p_{kj} + p_{ki} \log \sum_{k' \in [K]} \left( \frac{(p_{k'i})^2}{\sum_{a_i=a} p_{ki'}} \right) \\
&= \sum_{i=1}^N \left[ \sum_{k \in [K]} -p_{ki} \log p_{ki} + \sum_{k \in [K]} p_{ki} \log \sum_{a_j=a} p_{kj} + \log \sum_{k' \in [K]} \left( \frac{(p_{k'i})^2}{\sum_{a_i=a} p_{ki'}} \right) \right], \tag{4}
\end{aligned}$$

where  $p_{ki}$  is the predicted soft assignment for  $i$ -th sample regarding  $k$ -th cluster. For each  $p_{ki}$ , to minimize  $-p_{ki} \log p_{ki}$ , we have  $p_{ki} \rightarrow 1$  in its predicted cluster and  $p_{ki} \rightarrow 0$  in other clusters. Since the summation in equation 4 is taken over all samples and over all clusters, we can further simplify the second term as follows:

$$\sum_{i=1}^N \sum_{k \in [K]} p_{ki} \log \sum_{a_i=a} p_{ki} = \sum_{a \in [a]} \sum_{k \in [K]} P_{ka} \log P_{ka},$$

where  $P_{ka} := \sum_{a_i=a} p_{ki}$ . Consider  $P_{ka}$  of each group separately, for each  $P_{ka}$ , we have  $0 < P_{ka} < |\mathbb{N}_a|$ . Since the summation is taken over all cluters, to minimize this term, we need to make sure that  $\sum_{k \in [K]} P_{ka} \log P_{ka}$  is as small as possible, where the best possible assignment would be an equally distributed cluster assignment:

$$\sum_{k \in [K]} P_{ka} \log P_{ka} \leq |\mathbb{N}_a| \log \frac{|\mathbb{N}_a|}{K}.$$

It is easy to see that this term is in alignment with the balance notion. Similarly, we can further simplify the third term in equation 4 as

$$\begin{aligned}
&\sum_{i=1}^N \log \sum_{k' \in [K]} \left( \frac{(p_{k'i})^2}{\sum_{a'_i=a_i} p_{k'i'}} \right) \\
&\geq \sum_{i=1}^N \log \frac{\frac{1}{|K|^2} \times |K|}{\frac{N_a}{|K|}} \\
&= \sum_a N_a \log \frac{1}{N_a},
\end{aligned}$$

where the third term encourages samples of same sensitive attribute to have similar predicted soft assignment. Still, this term is not in accord with our expectation of enforcing high predicted confi-

dence or low confidence difference. Instead, we only keep the first term as our clustering regularizer:

$$L_{re} = \sum_{i=1}^N \sum_{k \in [K]} -p_{ki} \log p_{ki}.$$

## A.2 ABLATION STUDY

We include full results of ablation study in Tab [11](#), [18](#). We can see from results that contrastive loss and clustering regularizer help improve clustering accuracy, and Sinkhorn divergence helps reduce confidence disparities across different sensitive groups.

Method	Clustering accuracy	NMI
Our method	81.46±2.15%	77.82±1.26%
Our method (w/o contrastive loss)	76.49±2.21%	74.23±1.61%
Our method (w/o Sinkhorn divergence)	82.16±1.71%	78.31±1.22%
Our method (w/o regularization)	81.24±1.37%	77.21±1.26%

Table 11: Ablation study on MNIST-USPS dataset.

Method	Balance	DI	Conf. Dif.	EOd
Our method	0.36±0.03	0.10±0.02	0.02±0.01	0.09±0.02
Our method (w/o contrastive loss)	0.34±0.04	0.11±0.02	0.03±0.01	0.08±0.02
Our method (w/o Sinkhorn divergence)	0.21±0.03	0.17±0.04	0.09±0.03	0.14±0.03
Our method (w/o regularization)	0.35±0.03	0.10±0.02	0.02±0.01	0.09±0.02

Table 12: Ablation study on MNIST-USPS dataset.

Method	Clustering accuracy	NMI
Our method	65.47±1.86%	74.41±2.54%
Our method (w/o contrastive loss)	61.34±1.53%	69.63±1.67%
Our method (w/o Sinkhorn divergence)	67.64±1.45%	75.13±2.12%
Our method (w/o regularization)	65.13±1.42%	74.03±2.29%

Table 13: Ablation study on color reverse MNIST dataset.

Method	Balance	DI	Conf. Dif.	EOd
Our method	0.31±0.04	0.09±0.02	0.03±0.01	0.07±0.02
Our method (w/o contrastive loss)	0.29±0.04	0.10±0.02	0.03±0.01	0.08±0.02
Our method (w/o Sinkhorn divergence)	0.27±0.04	0.12±0.02	0.07±0.01	0.15±0.03
Our method (w/o regularization)	0.31±0.03	0.09±0.02	0.04±0.01	0.07±0.02

Table 14: Ablation study on color reverse MNIST dataset.

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 15: Ablation study on MTFD dataset.

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

Table 16: Ablation study on MTFD dataset.

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

Table 17: Ablation study on Office-31 dataset.

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

Table 18: Ablation study on Office-31 dataset.