A. Proof

In this section we provide the detailed proof of Theorem 4.1 in the main paper. Before we give the proof, we first present a useful lemma that will be used during the proof.

**Lemma A.1** (log-sum inequality). Let \( \{a_i\}_{i \in [n]} \) and \( \{b_i\}_{i \in [n]} \) be two sequence of nonnegative numbers. Define \( a := \sum_{i \in [n]} a_i \) and \( b := \sum_{i \in [n]} b_i \), then the following inequality holds:

\[
a \log \frac{a}{b} \leq \sum_{i \in [n]} a_i \log \frac{a_i}{b_i},
\]

where the equality holds iff \( a_i/b_i \) are equal for all \( i \in [n] \).

Now we are ready to prove Theorem 4.1.

**Theorem 4.1.** Let \( X \sim \mathcal{D} \) be the input random variable of the clustering algorithm \( C \). If the clustering assignment \( C(X) \) is independent of the sensitive attribute \( G \), then

\[
I(Y; C(X)) \leq \sum_{g \in [M]} \Pr(G = g) \cdot I(Y_g; C(X_g)),
\]

where \( X_g \) is the input random variable in the \( g \)-th protected subgroup and \( Y_g \) is the corresponding external label of \( X_g \).

**Proof.** To simplify the notation, in what follows we use \( C = C(X) \) and drop the subscript \( \mathcal{D} \) from the symbol \( \Pr_\mathcal{D}(\cdot) \). By definition of the mutual information, we have:

\[
I(Y; C(X)) = \sum_{ij} \Pr(Y = i, C = j) \log \frac{\Pr(Y = i, C = j)}{\Pr(Y = i) \cdot \Pr(C = j)}
\]

\[
= \sum_{ij} \left( \sum_{g \in [M]} \Pr(Y = i, C = j, G = g) \right) \log \frac{\sum_{g \in [M]} \Pr(Y = i, C = j, G = g)}{\sum_{g \in [M]} \Pr(Y = i, G = g) \cdot \Pr(C = j)}
\]

To make the presentation uncluttered, we define \( p_{ijg} := \Pr(Y = i, C = j, G = g) \), \( p_i := \Pr(Y = i) \), \( p_j := \Pr(C = j) \) and \( p_g := \Pr(G = g) \).

\[
= \sum_{ij} \left( \sum_{g} p_{ijg} \right) \log \frac{\sum_{g} p_{ig} p_{ijg}}{\sum_{g} p_{ig}}.
\]

Now consider a fixed pair of \( i, j \), by the log-sum inequality in Lemma A.1, we can upper bound \( \left( \sum_{g} p_{ijg} \right) \log \frac{\sum_{g} p_{ig} p_{ijg}}{\sum_{g} p_{ig}} \) as

\[
\left( \sum_{g} p_{ijg} \right) \log \frac{\sum_{g} p_{ig} p_{ijg}}{\sum_{g} p_{ig}} = p_j \left( \sum_{g} p_{ig} p_{ijg} \right) \log \frac{\sum_{g} p_{ijg} p_{ig}}{\sum_{g} p_{ig}} \leq p_j \sum_{g} p_{ig} p_{ijg} \log \frac{p_{ijg} p_{ig}}{p_{ig}}.
\]
Furthermore, since $C(X)$ is independent of $G$, then $\forall g \in [M]$, the following equality holds:

$$\sum_g p_g p_{ij|g} \log \frac{p_{ij|g}}{p_{ij}} = \sum_g p_g p_{ij|g} \log \frac{p_{ij|g}}{p_{ij|g} \cdot p_j} = \sum_g p_g p_{ij|g} \log \frac{p_{ij|g}}{p_{ij|g} \cdot p_{ij|g}},$$

where in the last equality we use $p_j = p_{ij|g}$ since $C(X) = j$ is independent of $G = g$. Now substituting the above inequality back to (12) yields:

$$\sum_{ij} \left( \sum_g p_{ij|g} \right) \log \frac{\sum_g p_{ij|g}}{\sum_g p_{ij}} \leq \sum_{ij} \sum_g p_g p_{ij|g} \log \frac{p_{ij|g}}{p_{ij|g} \cdot p_j} = \sum_g p_g \left( \sum_{ij} p_{ij|g} \log \frac{p_{ij|g}}{p_{ij|g} \cdot p_{ij|g}} \right) = \sum_g \Pr(G = g) \cdot I(Y_g; C(X_g)),$$

which completes the proof.

\[ \square \]

**Proposition 4.1.** Let $X \sim \mathcal{D}$ be the input random variable of the clustering algorithm $C$. If the clustering assignment $C(X)$ is independent of the sensitive attribute $G$, then

$$H(C(X)) = \sum_{g \in [M]} \Pr(G = g) \cdot H(C(X_g)). \quad (10)$$

**Proof.** Under the assumption that $C(X)$ is independent of $G$ and by definition of conditional entropy, we have:

$$H(C(X)) = H(C(X) \mid G) = \sum_{g \in [M]} \Pr(G = g) \cdot H(C(X) \mid G = g) = \sum_{g \in [M]} \Pr(G = g) \cdot H(C(X_g)),$$

completing the proof. \[ \square \]

### B. Label Matching

In this section we provide the approximate calculation for clustering accuracy of the best and worst matching on each dataset. We firstly show the clustering accuracy we get on each protected subgroup in Table 5.

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>MNIST-USPS / MNIST</th>
<th>MNIST-USPS / USPS</th>
<th>Color Reverse MNIST / Original</th>
<th>Color Reverse MNIST / Reversed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.894</td>
<td>0.745</td>
<td>0.929</td>
<td>0.965</td>
</tr>
<tr>
<td>Subgroup</td>
<td>MTFL / With Glasses</td>
<td>MTFL / Without Glasses</td>
<td>Office-31 / Amazon</td>
<td>Office-31 / Webcam</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.633</td>
<td>0.760</td>
<td>0.690</td>
<td>0.823</td>
</tr>
</tbody>
</table>

To simplify the matching calculation, we assume that each cluster contains equivalent numbers of samples and have equal accuracy. We denote the size of a protected subgroup $m$ as $s_m$ and the accuracy as $Acc_m$. In MNIST-USPS, Color Reverse MNIST, and Office-31, the number of cluster $K$ is large, and the best and worst matching accuracy can be calculated as:

$$Acc_{\text{best}} = \sum_{m \in [M]} s_m \cdot Acc_m, \quad Acc_{\text{worst}} = \frac{\max \{ s_m \cdot Acc_m \}_{m=1}^M }{\sum_{m \in [M]} s_m}.$$  

However, when $K = 2$ and $M = 2$ in MTFL dataset, consider the final accuracy will be significantly contributed by the inconsistent samples within one cluster, we have the accuracy of worst matching for protected subgroup $m$ and $m'$ as:

$$Acc_{\text{worst}} = \frac{\max \{ s_m \cdot Acc_m + s_{m'} \cdot (1 - Acc_{m'}), s_m \cdot (1 - Acc_m) + s_{m'} \cdot Acc_{m'} \} }{\sum_{m \in [M]} s_m}.$$