# Exercise 12 - Solution

**Group Fairness**

Reliable and Trustworthy Artificial Intelligence
ETH Zurich

**Problem 1** (Group Fairness via Post-processing)**.** Consider the post-processing method described in the lecture (slide 6). In this task we will explore this method on a toy example. Consider the following dataset of 10 points from distribution $\mathcal{X}$, represented as tuples $(x, s, y)$, denoting respectively a 1D feature vector, binary sensitive group membership, and the target label:

$$D = \{(0.1, 0, 0), (0.2, 0, 0), (0.3, 0, 0), (0.8, 0, 1), (0.9, 0, 0), \\ (0.1, 1, 0), (0.3, 1, 1), (0.4, 1, 0), (0.5, 1, 0), (0.7, 1, 1)\}.$$

Further, assume our binary classifier $g$ is the identity function.

1. For standard thresholds of $t_0 = 0.5$ and $t_1 = 0.5$ (no post-processing), estimate the accuracy of $g$ using $D$?

2. Estimate the fairness of $g$ using $D$, i.e., calculate the values of demographic parity distance, equalized odds distance, and equal opportunity distance? Definitions of these fairness constraints are given in the first fairness lecture. To interpret them as a distance follow the example of DP-distance given in the group fairness lecture (for equalized odds, the distances for two cases should be averaged).

3. Assume we keep $t_0 = 0.5$ fixed and want to change $t_1$ as a way to apply post-processing to make $g$ more fair. For what value of $t_1$ is $g$ the most fair with respect to three distances? To answer this question, fill in the code in the provided `task1.py` to calculate the relevant metrics, and plot the dependence of accuracy and fairness metrics on $t_1$.

**Solution 1.** 1. Accuracy of $g$ is defined as $A(g) = P(y = g(x))$, and can be estimated with $n$ given samples $(x_i, s_i, y_i)$ as follows: $A(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{y_i = g(x_i)\}$. Applying this on $D$ gives $A(g) = 0.8$.

2. The distances that correspond to unfairness constraints can be defined as follows, and estimated on $D$ in a similar way as above:

   - **Demographic parity**:

   $$\Delta_{DP}(g) = |P(g(x) = 1|s = 0) - P(g(x) = 1|s = 1)|$$

   - **Equalized odds**:

   $$\Delta_{EO}(g) = \frac{1}{2}\sum_{c=0}^{1} |P(g(x) = 1|s = 0, y = c) - P(g(x) = 1|s = 1, y = c)|$$

   - **Equality of opportunity**:

   $$\Delta_{EOP}(g) = |P(g(x) = 1|s = 0, y = 1) - P(g(x) = 1|s = 1, y = 1)|$$

   The values on $D$ with $t_0 = t_1 = 0.5$ are: $\Delta_{DP}(g) = 0.2$, $\Delta_{EO}(g) = 0.375$, and $\Delta_{EOP}(g) = 0.5$.

3. See the solution file `task1_solution.py`. Demographic parity distance is minimized for $t_1 \in [0.4, 0.5)$, and the two other distances for $t_1 < 0.3$.

**Problem 2** (Bounding Unfairness with the Optimal Adversary)**.** Prove the key inequality used by FNF and FARE to upper bound the DP-distance of downstream classifiers:

$$\Delta^{DP}_{\mathcal{Z}_0,\mathcal{Z}_1}(g) \leq 2 \cdot BA_{\mathcal{Z}_0,\mathcal{Z}_1}(h^\star) - 1,$$

where we use the notation from the lecture. Namely, $\mathcal{Z}_0$ and $\mathcal{Z}_1$ are the conditional distributions of $z$ for $s = 0$ and $s = 1$, respectively. The DP-distance is defined as:

$$\Delta^{DP}_{\mathcal{Z}_0,\mathcal{Z}_1}(g) = \left| \mathop{\mathbb{E}}_{z\sim\mathcal{Z}_0} g(z) - \mathop{\mathbb{E}}_{z\sim\mathcal{Z}_1} g(z) \right|,$$

and the balanced accuracy of the adversary as:

$$BA_{\mathcal{Z}_0,\mathcal{Z}_1}(h) = \frac{1}{2}\left( \mathop{\mathbb{E}}_{z\sim\mathcal{Z}_0} (1 - h(z)) + \mathop{\mathbb{E}}_{z\sim\mathcal{Z}_1} h(z) \right).$$

**Solution 2.** Suppose WLOG (other case is similar) that

$$\mathop{\mathbb{E}}_{z\sim\mathcal{Z}_0} g(z) \geq \mathop{\mathbb{E}}_{z\sim\mathcal{Z}_1} g(z),$$

i.e., the classifier $g$ predicts the positive outcome more often for sensitive group 0. Then, we can drop the absolute value in our definition of $\Delta^{DP}_{\mathcal{Z}_0,\mathcal{Z}_1}(g)$. Further, it holds that:

$$\mathop{\mathbb{E}}_{z\sim\mathcal{Z}_1} g(z) = 1 - \mathop{\mathbb{E}}_{z\sim\mathcal{Z}_1}(1-g(z)).$$

Combining these we can rewrite the DP distance as:

$$\Delta^{DP}_{\mathcal{Z}_0,\mathcal{Z}_1}(g) = \mathop{\mathbb{E}}_{z\sim\mathcal{Z}_0} g(z) + \mathop{\mathbb{E}}_{z\sim\mathcal{Z}_1}(1-g(z)) - 1.$$

Now consider some adversary $h$ which makes opposite predictions of $g$, i.e., $h = 1 - g$. Its balanced accuracy is (using the equations above):

$$\begin{aligned} BA_{\mathcal{Z}_0,\mathcal{Z}_1}(h) &= \frac{1}{2}\left(\mathop{\mathbb{E}}_{z\sim\mathcal{Z}_0} g(z) + \mathop{\mathbb{E}}_{z\sim\mathcal{Z}_1}(1-g(z))\right) \\ &= \frac{1}{2}(\Delta^{DP}_{\mathcal{Z}_0,\mathcal{Z}_1}(g)+1). \end{aligned}$$

For the optimal adversary $h^\star$ it by definition holds that $BA_{\mathcal{Z}_0,\mathcal{Z}_1}(h^\star) \geq BA_{\mathcal{Z}_0,\mathcal{Z}_1}(h)$. Combining with above we have $BA_{\mathcal{Z}_0,\mathcal{Z}_1}(h^\star) \geq \frac{1}{2}(\Delta^{DP}_{\mathcal{Z}_0,\mathcal{Z}_1}(g)+1)$, i.e.,

$$\Delta^{DP}_{\mathcal{Z}_0,\mathcal{Z}_1}(g) \leq 2BA_{\mathcal{Z}_0,\mathcal{Z}_1}(h^\star) - 1,$$

which recovers the inequality we have to prove.

**Problem 3** (FNF with Categorical Features)**.** Real-world datasets often contain categorical data, in which case the optimal bijections $f_0$ and $f_1$ can be directly computed, instead of using normalizing flows. Consider discrete samples $x$ coming from a probability distribution $q(x)$ where each component takes a value from a finite set $\{1, 2, \ldots, d_i\}$. This implies a finite $\mathcal{X} = \{x_1, \ldots, x_m\}$. As before, our goal is to find bijections $f_0 : \mathcal{X} \to \mathcal{Z}$ and $f_1 : \mathcal{X} \to \mathcal{Z}$ that minimize the statistical distance of the latent distributions, i.e., minimize the adversary advantage. For simplicity, you can assume $\mathcal{Z} = \{1, \ldots, m\}$.

1. Describe the procedure used to construct the optimal $f_0$ and $f_1$.

2. Provide a proof that this procedure minimizes the balanced accuracy of the optimal adversary.

3. Are such $f_0$ and $f_1$ always the best choice in practice? If not, why?

**Solution 3.**

1. Intuitively, the optimal solution w.r.t. fairness is obtained by "pairing up" inputs with similar probabilities in $q_0$ and $q_1$, i.e., mapping them to the same latent representation $z$. More precisely, we can let $i_1, i_2, ..., i_m$ and $j_1, ..., j_m$ denote the permutations of $\{1, 2, ..., m\}$ such that $q_0(x_{i_1}) \leq q_0(x_{i_2}) \leq ... \leq q_0(x_{i_m})$ and $q_1(x_{j_1}) \leq q_1(x_{j_2}) \leq ... \leq q_1(x_{j_m})$. Then we can set $f_0(x_{i_k}) = f_1(x_{j_k}) = k$ for every $k \in \{1, 2, ..., m\}$.

2. We will use the following definition of the balanced accuracy shown in the lecture:

$$BA_{\mathcal{Z}_0,\mathcal{Z}_1}(h) = \frac{1}{2}\int_{\mathcal{Z}} \Big(p_0(z)(1 - h(z)) + p_1(z)h(z)\Big)dz.$$

Replacing the integral with a sum (as $\mathcal{Z} = \{1, \ldots, m\}$) and plugging in the definition of the optimal adversary $h^\star(z) = \mathbb{1}\{p_1(z) \geq p_0(z)\}$ we obtain:

$$BA_{\mathcal{Z}_0,\mathcal{Z}_1}(h^\star) = \frac{1}{2}\sum_{k=1}^{m} \max(p_0(k), p_1(k)). \tag{1}$$

In the remainder of the proof we will use the following simple identity that holds for any $a, b, c \in \mathbb{R}$:

$$a \geq b \implies \max(a, c) \geq \max(b, c). \tag{2}$$

It is easy to check that this holds for all positions of $c$, i.e., $c \geq a$, $c \in [b, a)$, and $c < b$.

Assume WLOG that $\forall k \in \{1, \ldots, m\}$ we set $f_0(x_{i_k}) = k$. Now assume that for some $x, x'$ we set $f_1(x) = k$ and $f_1(x') = k'$, where $k < k'$ and thus $q_0(x_{i_k}) \leq q_0(x_{i_{k'}})$ but $q_1(x) > q_1(x')$. Let $S = \max(q_0(x_{i_k}), q_1(x)) + \max(q_0(x_{i_{k'}}), q_1(x'))$ denote the sum of the terms in eq. (1) corresponding to $k$ and $k'$. Similarly, let $S' = \max(q_0(x_{i_k}), q_1(x')) + \max(q_0(x_{i_{k'}}), q_1(x))$ denote the sum of the same terms if we swapped the mappings $f_1(x)$ and $f_1(x')$.

Now, fixing $q_0(x_{i_k})$ and $q_0(x_{i_{k'}})$, we consider two cases regarding the position of $q_1(x)$ and $q_1(x')$.

a) $q_1(x) \geq q_0(x_{i_{k'}})$: In this case we have $S = q_1(x) + \max(q_0(x_{i_{k'}}), q_1(x'))$ and $S' = \max(q_0(x_{i_k}), q_1(x')) + q_1(x)$. Applying eq. (2) with $a = q_0(x_{i_{k'}}), b = q_0(x_{i_k}), c = q_1(x')$ shows that $S \geq S'$.

b) $q_1(x) < q_0(x_{i_{k'}})$: In this case we have $S = \max(q_0(x_{i_k}), q_1(x)) + q_0(x_{i_{k'}})$ and $S' = \max(q_0(x_{i_k}), q_1(x')) + q_0(x_{i_{k'}})$. Again, applying eq. (2) with $a = q_1(x), b = q_1(x'), c = q_0(x_{i_k})$ shows that $S \geq S'$.

In all cases $S \geq S'$, meaning that if we swap the mappings $f_1(x)$ and $f_1(x')$, the balanced accuracy of the optimal adversary eq. (1) can never increase. As our

goal is to find mappings that minimize eq. (1), we can swap such pairs as long as they exist, and eventually we will arrive at the solution described above, where the mappings are sorted by $q_0$ and $q_1$.

3. No, as in practice we are not simply interested in maximizing fairness, but achieving some tradeoff between accuracy and fairness. The proposed sort may completely sacrifice the utility of representations for a given task, if points with different target labels get mapped to the same representation often. One possible solution could be to split inputs according to the target label $y$ and do the matching from above on two groups independently. We could further combine this with the original solution to achieve a desired tradeoff.