A. Supplementary Material

A.1. Proofs

Lemma 3.1. Given an hypothesis class $\mathcal{H}$ and a finite alphabet $\mathcal{A} : |\mathcal{A}| \geq 2$, problems 3 and 4 have the same minimum worst-group risk solution R^* if $\rho \leq \frac{1}{|\mathcal{A}|}$.

Proof. For any $h \in \mathcal{H}$, let $l_h = \ell(h(X), Y)$ be the random variable associated with the loss distribution of h induced by the randomness of X, Y. Let $\hat{\ell}_{h,\rho} = F_{l_h}^{-1}(1-\rho)$ be the $100 * (1-\rho)\%$ percentile of l_h, where $F_{l_h}^{-1}(\alpha) = \inf\{l \in \mathbb{R} : P(l_h \leq l) \geq \alpha\}$ is the inverse cdf of l_h. It is easy to observe that any distribution $p(A \mid X, Y), A \in \mathcal{A}$, that satisfies

$$\begin{aligned} &p(A = a' \mid X, Y) = \begin{cases} 1 & \text{if } \ell(h(X), Y) > \hat{\ell}_{h,\rho}, \\ \alpha(X, Y) \in [0, 1] & \text{if } \ell(h(X), Y) = \hat{\ell}_{h,\rho}, \\ 0 & \text{if } \ell(h(X), Y) < \hat{\ell}_{h,\rho}, \end{cases} \\ &p(A = a) \geq \rho, \ \forall a \in \mathcal{A}, \\ &p(A = a') = \rho, \ a' \in \mathcal{A}. \end{aligned} \tag{10}$$

is a solution to

$$\begin{array}{cc} \max\limits_{\substack{p(A|X,Y) \\ s.t.\ p(A) \succeq \rho,}} & \max\limits_{a \in \mathcal{A}} r_a(h), \end{array}$$

attaining the maximum risk at $r_{a'}(h)$. Here $\alpha(X, Y) \in [0, 1]$ is any tie-breaking assignment such that $p(A = a') = \rho$ and $p(A = a) \geq \rho$. That is, the worst-case partition greedily assigns $A = a'$ to all high loss samples until the budget $p(A = a') = \rho$ is satisfied, the tie-breaker assignment $\alpha(X, Y)$ simply indicates that for loss values exactly equal to $\hat{\ell}_{h,\rho}$, we can make any assignment we wish to as long as $p(A = a') = \rho$.

Furthermore, by applying the same reasoning as above, we observe that the simplified distribution $\hat{p}(A \mid X, Y), A \in \{0, 1\}, \ \hat{p}(A = 1 \mid X, Y) = p(A = a' \mid X, Y)$ is also a solution to

$$\begin{array}{cc} \max\limits_{\substack{a \in \{0,1\} \\ p(A|X,Y) \\ s.t.\ p(A) \succeq \rho,}} & r_a(h), \end{array}$$

with both achieving the same maximum risk. At this point we have proved the following equivalence:

$$\min_{h \in \mathcal{H}} \max_{\substack{p(A|X,Y) \\ s.t.\ p(A) \succeq \rho,}} \max_{a \in \mathcal{A}} r_a(h) = \min_{h \in \mathcal{H}} \max_{\substack{a \in \{0,1\} \\ p(A|X,Y) \\ s.t.\ p(A) \succeq \rho,}} r_a(h).$$

We now prove that, **in terms of worst case risk**, minimizing over $h \in \mathcal{H}$ is equivalent to minimizing over its respective Pareto efficient set $h \in \mathcal{H}_{P_{\mathcal{A}}}$ for the left side of the equation and $h \in \mathcal{H}_{P_{\mathcal{A}=\{0,1\}}}$ for the right side.

Looking at the left side equation, we note that for all $\bar{h} \in \mathcal{H}$ and $\bar{p}(A|X, Y) : p(A) \succeq \rho$, we have a corresponding risk vector $\{r_a(\bar{h})\}_{a \in \mathcal{A}}$. Let $a' = \arg\max_a r_a(\bar{h})$ be the worst group; by the properties of Pareto optimality, we know that there exists a model $\hat{h}$ such that

$$\hat{h} \in \mathcal{H}_{P_{\mathcal{A}}} : r_{a'}(\hat{h}) = r_{a'}(\bar{h}), r_a(\hat{h}) \leq r_a(\bar{h}) \ \forall a \in \mathcal{A} \setminus \{a'\}.$$

That is, there exists a Pareto efficient model that achieves the same risk on a' but less or equal risk in all other coordinates (note that if $\bar{h} \in \mathcal{H}_{P_{\mathcal{A}}}$ then $\bar{h} = \hat{h}$). Applying this property we observe that

$$\min_{h \in \mathcal{H}_{P_{\mathcal{A}}}} \max_{\substack{p(A|X,Y) \\ s.t.\ p(A) \succeq \rho,}} \max_{a \in \mathcal{A}} r_a(h) = \min_{h \in \mathcal{H}} \max_{\substack{p(A|X,Y) \\ s.t.\ p(A) \succeq \rho,}} \max_{a \in \mathcal{A}} r_a(h).$$

Using similar reasoning, we have that

$$\min_{h\in\mathcal{H}_{P_{\mathcal{A}=\{0,1\}}}} \max_{\substack{a\in\{0,1\}\\ p(A|X,Y)\\ s.t.\ p(A)\succeq\rho,}} r_a(h) = \min_{h\in\mathcal{H}} \max_{\substack{a\in\{0,1\}\\ p(A|X,Y)\\ s.t.\ p(A)\succeq\rho,}} r_a(h),$$

and thus,

$$\min_{h\in\mathcal{H}_{P_{\mathcal{A}}}} \max_{\substack{p(A|X,Y)\\ s.t.\ p(A)\succeq\rho,}} \max_{a\in\mathcal{A}} r_a(h) = \min_{h\in\mathcal{H}_{P_{\mathcal{A}=\{0,1\}}}} \max_{\substack{a\in\{0,1\}\\ p(A|X,Y)\\ s.t.\ p(A)\succeq\rho,}} r_a(h),$$

We want to restate that the equalities are valid in terms of worst case risk, there may be minimax models $h \in \mathcal{H}$ that do not belong to the Pareto set $h \in \mathcal{H}_{P_{\mathcal{A}}}$

□

Lemma 3.2. Given Problem 4 with $p(Y|X) > 0\ \forall X, Y$, and let the classification loss be cross-entropy or Brier score. Let $\bar{h}(X) : \bar{h}_i(X) = \frac{1}{|\mathcal{Y}|}\forall X, \forall i \in \{0, ..., |\mathcal{Y}| - 1\}$ be the uniform classifier, and let $\bar{h} \in \mathcal{H}$.

There exists a critical partition size

$$\rho^* = |\mathcal{Y}| E_X[\min_y p(y \mid X)] \le 1$$

such that the solutions to Problem 4, $\forall \rho \le \rho^*$, are

$$\begin{aligned} h^* &= \bar{h}, \\ R^* = \bar{R} &= \begin{cases} \log|\mathcal{Y}| & \text{if } \ell = \ell_{CE} \\ \frac{|\mathcal{Y}|-1}{|\mathcal{Y}|} & \text{if } \ell = \ell_{BS} \end{cases}. \end{aligned}$$

That is, the solutions to all partitions with size smaller than ρ^* yield the uniform classifier with constant risk $\bar{R}$.

Proof. This proof is done in three steps, first we provide an upper bound of the solution of Problem 4, we then show that we can design a (potentially nonexistent) partition density that achieves this upper bound, and finally, we derive conditions under which the previously identified partition is guaranteed to exist.

We first prove that for any distributions $p(X, Y, A)$, it follows that

$$\min_{h\in\mathcal{H}} \max_{a\in\{0,1\}} r_a(h) \le \bar{R},$$

meaning that the solution to Problem 4 is upper bounded by the risk associated with the uniform classifier for cross-entropy and Brier score losses.

This is done by considering that for any distribution $p(X, Y, A)$, the conditional risk of the uniform classifier $\bar{h}(X)$: $\bar{h}_i(X) = \frac{1}{|\mathcal{Y}|}\forall X, \forall i \in \{0, ..., |\mathcal{Y}| - 1\}$ is

$$E_{X,Y|A}[\ell(\bar{h}(X), Y)] = \bar{R} = \begin{cases} \log|\mathcal{Y}| & \text{if } \ell = \ell_{CE} \\ \frac{|\mathcal{Y}|-1}{|\mathcal{Y}|} & \text{if } \ell = \ell_{BS} \end{cases},\ \forall p(X, Y, A),$$

Since $\bar{h} \in \mathcal{H}$, we have that $\min_{h\in\mathcal{H}} \max_{a\in\{0,1\}} r_a(h) \le \bar{R}\ \forall p(X, Y, A)$.

Then we show that if we can design $p(A|X,Y) : p(Y|X,A=1) = \frac{1}{|\mathcal{Y}|} \forall X,Y$ we have that, under this distribution, $\bar{h}, \bar{R} = \{\arg\} \min_{h\in\mathcal{H}} \max_{a\in\{0,1\}} r_a(h)$, which is the upper bound identified above.

For this assume that we have $p(Y|X,A=1) = \frac{1}{|\mathcal{Y}|} \forall X,Y$, it then follows that $\min_{h\in\mathcal{H}} r_1(h) = r_1(\bar{h}) = \bar{R}$ since $\bar{h}$ is, by design, the optimal classifier for group $a=1$ and $\bar{R}$ its best achievable risk. Then $r_{a=1}(h) \geq \bar{R} \,\forall h$ and since $r_{a=1}(\bar{h}) = r_{a=0}(\bar{h})$ it follows that

$$\bar{h}, \bar{R} = \{\arg\} \min_{h\in\mathcal{H}} \max_{a\in\{0,1\}} r_a(h).$$

Finally, we derive a necessary and sufficient condition for the existence of $p(A|X,Y) : p(Y|X,A{=}1) = \frac{1}{|\mathcal{Y}|} \;\forall X,Y$. Since we need

$$p(A{=}1|X,Y) = \frac{1}{|\mathcal{Y}|}\frac{p(A{=}1|X)}{p(Y|X)}$$

to be a well-defined distribution, the only degree of freedom available is $p(A{=}1 \mid X)$. Note that

$$p(Y|A{=}0,X) = \frac{p(Y|X)|\mathcal{Y}| - p(A{=}1|X)}{1-p(A{=}1|X)}\frac{1}{|\mathcal{Y}|} \geq 0 \;\forall X,Y,$$

therefore $p(A{=}1|X) \leq |\mathcal{Y}|p(Y|X),\; \forall Y,X \rightarrow p(A{=}1|X) \leq |\mathcal{Y}| \min_{y\in\mathcal{Y}} p(Y{=}y|X)$ and therefore

$$p(A{=}1) \leq E_X[|\mathcal{Y}| \min_{y\in\mathcal{Y}} p(y|X)] = \rho^*.$$

We also note that $\min_{y\in\mathcal{Y}} p(y|X) \leq \frac{1}{|\mathcal{Y}|}$, therefore $\rho^* \leq 1$

□

Note that the Lemma above can drop the hypothesis $p(Y|X) > 0 \;\forall X,Y$ by defining a new semi-uniform classifier $\bar{h}(X) : \bar{h}_i(X) = \frac{\mathbb{1}(i\in Y(X))}{|\mathcal{Y}(X)|} \forall X, \forall i \in \{0,...,|\mathcal{Y}|-1\}$, where $\mathcal{Y}(X)$ indicates the subset of labels y such that $p(y|X) > 0$. The proof proceeds similarly, with the resulting partition size $E_X[|\mathcal{Y}(X)| \min_{y\in\mathcal{Y}(X)} p(y|X)] = \rho^*$.

A generalization of Lemma 3.2 is presented next in Lemma A.1 and relies on the following definition

Definition A.1. Given a loss function $\ell : \Delta^{|\mathcal{Y}|-1} \times \Delta^{|\mathcal{Y}|-1} \rightarrow \mathbb{R}^+$, we say that an hypothesis $\hat{h} : \mathcal{X} \rightarrow \Delta^{|\mathcal{Y}|-1}$ is **Uniformly Maximal** if it satisfies

$$\hat{h}(x), \hat{p}(Y|X=x) = \arg\min_{h(x)\in\Delta^{|\mathcal{Y}|-1}} \max_{p(Y|X=x)\in\Delta^{|\mathcal{Y}|-1}} \mathbb{E}_{Y|X=x}\ell(h(x),Y),\; \forall x \in \mathcal{X},$$
$$\hat{h}, \hat{R} = \{\arg\} \min_{h:\mathcal{X}\rightarrow\Delta^{|\mathcal{Y}|-1}} \mathbb{E}_{Y\sim\hat{p}(Y|X)}\ell(h(X),Y)\;.$$

In other words, a hypothesis $\hat{h}$ is uniformly maximal if it is minimax optimal and is the optimal solution for the maximal target distribution $\hat{p}(Y|X)$.

Lemma A.1. *Given Problem 4 with $p(Y|X) > 0 \;\forall X,Y$, and let the classification loss admit a uniformly maximal classifier $\hat{h}$ with corresponding distribution $\hat{p}(Y|X)$ and risk $\hat{R}$. Further, let $\hat{h} \in \mathcal{H}$*

There exists a critical partition size

$$\rho^* = \mathbb{E}_X\big[\min_y \frac{p(y|X)}{\hat{p}(y|X)}\big] \leq 1,$$

such that the solutions to Problem 4, $\forall\rho \leq \rho^*$, are $h^* = \hat{h}$ and $R^* = \hat{R}$.

That is, the solutions to all partitions with size smaller than ρ^* yield the uniformly maximal classifier $\hat{h}$ with risk $\hat{R}$.

Proof. Since $\hat{h}$ is a uniformly maximal classifier we have that for any distributions $p(X,Y,A)$, it follows that

$$\min_{h\in\mathcal{H}} \max_{a\in\{0,1\}} r_a(h) \leq \hat{R},$$

meaning that the solution to Problem 4 is upper bounded by the risk of the uniformly maximal classifier $\hat{R}$. This is immediate by the definition of a uniformly maximal classifier.

Now, in the same fashion as in Lemma 3.2) we show that if we can design $p(A|X,Y) : p(Y|X,A=1) = \hat{p}(Y|X)\ \forall X, Y$ (maximal distribution) we have that, under this distribution, $\hat{h}, \hat{R} = \{\arg\} \min\limits_{h \in \mathcal{H}} \max\limits_{a \in \{0,1\}} r_a(h)$, which is the upper bound identified above.

So lets assume that $p(Y|X,A=1) = \hat{p}(Y|X)\ \forall X, Y$, it then follows that $\min\limits_{h\in\mathcal{H}} r_1(h) = r_1(\hat{h}) = \hat{R}$ since $\hat{h}$ is, by design, the optimal classifier for group $a=1$ and $\hat{R}$ its best achievable risk. Then $r_{a=1}(h) \geq \hat{R}\ \forall h$ and since $r_{a=1}(\hat{h}) \geq r_{a=0}(\hat{h})$ it follows that

$$\hat{h}, \hat{R} = \{\arg\} \min_{h \in \mathcal{H}} \max_{a \in \{0,1\}} r_a(h).$$

Now we derive a necessary and sufficient condition for the existence of $p(A|X,Y) : p(Y|X,A{=}1) = \hat{p}(Y|X)\ \forall X, Y$. Since we need

$$p(A{=}1|X,Y) = \hat{p}(Y|X)\frac{p(A{=}1|X)}{p(Y|X)}$$

to be a well-defined distribution, the only degree of freedom available is $p(A{=}1 \mid X)$. Note that

$$\begin{aligned} p(Y|X) &= p(Y|A{=}0,X)(1-p(A{=}1|X)) + \hat{p}(Y|X)p(A{=}1|X)\ \forall X, Y, \\ p(Y|A{=}0,X) &= \tfrac{p(Y|X)-\hat{p}(Y|X)p(A{=}1|X)}{(1-p(A{=}1|X))} \geq 0,\ \forall X, Y,{}^{11}. \end{aligned}$$

Therefore, $p(A{=}1|X) \leq \frac{p(Y|X)}{\hat{p}(Y|X)},\ \forall Y, X \rightarrow p(A{=}1|X) \leq \min\limits_{y\in\mathcal{Y}} \frac{p(Y|X)}{\hat{p}(Y|X)}$ which results in

$$p(A{=}1) \leq \mathbb{E}_X[\min_{y\in\mathcal{Y}} \frac{p(y|X)}{\hat{p}(y|X)}] = \rho^*.$$

We also note that $\min\limits_{y\in\mathcal{Y}} \frac{p(y|X)}{\hat{p}(y|X)} \leq 1$, therefore $\rho^* \leq 1$.

□

Note that similar to Lemma 3.2 in Lemma A.1 above we can drop the hypothesis $p(Y|X) > 0\ \forall X, Y$ by defining $\mathcal{Y}(X)$ to be the subset of labels y such that $p(y|X) > 0$, and using the semi-uniformly minimax classifier $\hat{h}(X;\mathcal{Y}(X))$, where $\hat{h}_y(X;\mathcal{Y}(X))$ is the uniformly maximal classifier over the set $\mathcal{Y}(X)$, and is $0 \forall \mathcal{Y} \setminus \mathcal{Y}(X)$. Likewise, we can define the corresponding distribution $\hat{p}(y|x;\mathcal{Y}(x))$. The proof proceeds similarly, with the resulting partition size $\mathbb{E}_X[\min_{y\in\mathcal{Y}(X)} \frac{p(y|X)}{\hat{p}(Y|X;\mathcal{Y}(X))}] = \rho^*$.

The statement in Lemma 3.2 is a direct corollary of this result, since it is straigthforward to observe that $\hat{h}(x) = \frac{1}{|\mathcal{Y}|}$ is a uniformly maximal distribution for Brier score and crossentropy; this solution is optimal when $\hat{p}(Y|X) = \frac{1}{|\mathcal{Y}|} \forall X, Y \in \mathcal{X} \times \mathcal{Y}$.

Other loss functions that place special emphasis on certain preferred outcomes can also be considered on Lemma 3.2. For example, weighted crossentropy is defined as

$$\mathbb{E}_{X,Y}[-\textstyle\sum_{y=1}^{|\mathcal{Y}|} w_y \mathbf{1}(Y{=}y) \log h_y(X)],$$

where $w_y > 0\, \forall y \in \mathcal{Y}$. This loss admits a uniformly maximal classifier $\hat{h}$ of the form

$$\hat{h}(x) = \{\hat{h}_y(x)\}_{y\in\mathcal{Y}}, \quad \hat{h}_y(x) = h_0^{\frac{w_0}{w_y}}, \forall x \in \mathcal{X}, \quad \hat{h}_0 : \textstyle\sum_y \hat{h}_0^{\frac{w_0}{w_y}} = 1,$$

with associated distribution $\hat{p}(y|x) = \frac{\hat{h}_y/w_y}{\sum_{y'\in\mathcal{Y}} \hat{h}_{y'}/w_{y'}}$. This reduces to $\hat{h}(x) = \frac{1}{|\mathcal{Y}|}$ for standard crossentropy.

[11] Note that requiring $p(Y|A{=}0,X) \leq 1$ is unnecessary since $\sum_Y p(Y|A{=}0,X) = 1$ holds for any choice of $p(A=1|X)$

Lemma 3.3. Given a distribution $p(X,Y)$ and any predefined partition group $p(A'|X,Y)$ with $A' \in \mathcal{A}'$, $|\mathcal{A}'|$ finite. Let $\hat{h}, \hat{R} = \{\arg\} \min\limits_{h\in\mathcal{H}} \max\limits_{a'\in\mathcal{A}'} r_{a'}(h)$ be the minimax fair solution for this partition and its corresponding minimax risk. Let h^* and R^* be the classifier and risk that solve Problem 4 with $\rho = \min_{a'} p(a')$. Then the price of minimax fairness can be upper bounded by

$$\max_{a'\in\mathcal{A}'} r_{a'}(h^*) - \hat{R} \leq R^* - \min_{h\in\mathcal{H}} r(h). \tag{11}$$

Proof. Observe that $\forall\, \mathcal{A}$, and for any distribution $p(A \mid X, Y)$, $A \in \mathcal{A}$ and $\forall h' \in \mathcal{H}$ we have

$$\min_{h\in\mathcal{H}} r(h) \leq r(h') = \sum_{a\in\mathcal{A}} p(a) r_a(h') \leq \max_{a\in\mathcal{A}} r_a(h').$$

We also have

$$h^*, p^*(A|X,Y), R^* = \{\arg\} \min_{h\in\mathcal{H}_{P_\mathcal{A}}} \max_{\substack{a\,\in\,\{0,1\}\\ p(A|X,Y)\\ s.t.\, p(A)\succeq\rho}} r_a(h).$$

Which, together with Lemma 3.1, implies

$$\max_{a'\in\mathcal{A}'} r_{a'}(h^*) \leq R^* \leq \max_{a^*\in\{0,1\}} r_{a^*}(h').$$

We combine the two and show

$$\begin{aligned}\max_{a'\in\mathcal{A}'} r_{a'}(h^*) - \hat{R} \quad &\leq R^* - \hat{R}\\ &\leq R^* - \min_{h\in\mathcal{H}} r(h).\end{aligned}$$

□

Lemma 4.1. Given Problem 4 with minimum group size $\rho \leq \frac{1}{2}$, the following problems are value equivalent:

$$\begin{aligned}
R^{\text{I}} &= \min_{h\in\mathcal{H}_{P_\mathcal{A}}} \max_{\substack{p(A|X,Y)\\ s.t. p(A)\succeq\rho}} \max_{a\in\{0,1\}} r_a(h),\\
R^{\text{II}} &= \min_{h\in\mathcal{H}_{P_\mathcal{A}}} \max_{\substack{p(A|X,Y)\\ s.t.\, p(A=1)=\rho}} r_1(h),\\
R^{\text{I}} &= R^{\text{II}}.
\end{aligned}$$

Proof. Following the arguments in the proof of Lemma 3.1 we observe that, for any $h \in \mathcal{H}$ and $\mathcal{A} = \{0,1\}$, we can consider the partition proposed in Equation 10 with $a' = 1$, which is a risk maximizing distribution for that particular h. This distribution satisfies $\max_{a\in\{0,1\}} r_a(h) = r_1(h)$, and also satisfies $p(A=1) = \rho$. Following the same reasoning as in the proof of Lemma 3.1, we can translate this equivalence in terms of worst case risk from the set $h \in \mathcal{H}$ to the set $h \in \mathcal{H}_{P_\mathcal{A}}$.

□

Lemma 4.2. Given the problem on the right hand side of Eq. 6, a convex hypothesis class $\mathcal{H}$, and a bounded loss function $0 \leq \ell(h(x), y) \leq C \;\forall x, y, h \in \mathcal{X}\times\mathcal{Y}\times\mathcal{H}$ that is strictly convex w.r.t its first input $h(x)$ the following problems are equivalent:

$$\begin{aligned}
\mathcal{H}^{\text{I}}, R^{\text{I}} &= \{\arg\} \min_{h\in\mathcal{H}_{P_\mathcal{A}}} \max_{\substack{p(A|X,Y)\\ s.t.\, p(A=1)=\rho}} r_1(h)\\
\mathcal{H}^{\text{II}}, R^{\text{II}} &= \{\arg\} \min_{h\in\mathcal{H}} \sup_{\substack{p(A|X,Y)\\ s.t.\, p(A=1)=\rho\\ p(A=1|X,Y)>0,\ \forall X,Y}} r_1(h),\\
R^{\text{I}} &= R^{\text{II}},\ \mathcal{H}^{\text{I}} \supseteq \mathcal{H}^{\text{II}},
\end{aligned}$$

Proof. We present this proof in two steps. First, we show that, under the hypothesis class $\mathcal{H}_{P_{\mathcal{A}}}$, we can change the maximum over the set of distributions $P(A|X,Y) : P(A=1) = \rho$ for the supremum over the set of distributions $P(A|X,Y) : P(A=1) = \rho, P(A=1|X,Y) > 0\ \forall X, Y$. That is,

$$\begin{aligned}\mathcal{H}^{\text{III}}, R^{\text{III}} &= \{\arg\} \min_{h \in \mathcal{H}_{P_{\mathcal{A}}}} \sup_{\substack{p(A|X,Y) \\ s.t.\ p(A=1)=\rho \\ p(A=1|X,Y)>0\ \forall X,Y}} r_1(h), \\ R^{\text{I}} &= R^{\text{III}},\ \mathcal{H}^{\text{I}} \supseteq \mathcal{H}^{\text{III}}.\end{aligned}$$

To prove this, we start by defining the set of distributions satisfying $p(A=1)$ as

$$Q_{\rho,\geq} = \{p(A|X,Y) : \int p(A=1|x,y)p(x,y) = \rho \wedge p(A=1|X,Y) \geq 0\ \forall X, Y \in \mathcal{X} \times \mathcal{Y}\},$$

and the distribution subset on the above equation as

$$Q_{\rho,>} = \{p(A|X,Y) : \int p(A=1|x,y)p(x,y) = \rho \wedge p(A=1 \mid X,Y) > 0\ \forall X, Y \in \mathcal{X} \times \mathcal{Y}\}.$$

We can then observe that, for any model h and distributions $\hat{p}(A|X,Y) \in Q_{\rho,\geq}$ and $\bar{p}(A|X,Y) \in Q_{\rho,>}$, the distribution $p_\lambda(A|X,Y) = \lambda\bar{p}(A|X,Y) + (1-\lambda)\hat{p}(A|X,Y)$ satisfies $p_\lambda(A|X,Y) \in Q_{\rho,>}\ \forall\lambda \in (0,1]$. Furthermore, we have, by linearity of expectation

$$\begin{aligned}r_1(h; p_\lambda(A|X,Y)) &= \lambda r_1(h; \bar{p}(A|X,Y)) + (1-\lambda) r_1(h; \hat{p}(A|X,Y)) \\ &\leq \lambda C + (1-\lambda) r_1(h; \hat{p}(A|X,Y)), \\ r_1(h; p_\lambda(A|X,Y)) &\geq (1-\lambda) r_1(h; \hat{p}(A|X,Y))\end{aligned}$$

where we used explicit notation to indicate what distribution we are using to take expectation and the fact that the loss is upper bounded by C and lower bounded by 0. Therefore we conclude

$$\lim_{\lambda \to 0^+} p_\lambda(A|X,Y) = \hat{p}(A|X,Y)$$

and

$$\lim_{\lambda \to 0^+} r_1(h; p_\lambda(A|X,Y)) = r_1(h; \hat{p}(A|X,Y)).$$

Similarily for r_0

$$\lim_{\lambda \to 0^+} r_0(h; p_\lambda(A|X,Y)) = r_0(h; \hat{p}(A|X,Y)).$$

Since this transformation preserves the entire risk vector $r_0(h), r_1(h)$, and the results hold for any $h \in \mathcal{H}$ and $\hat{p}(A|X,Y) \in Q_{\rho,\geq}$, they hold in particular for any hypothesis $h \in \mathcal{H}^{\text{I}}$ and its corresponding distribution and group risk vector. From this we can conclude

$$\{\arg\} \min_{h \in \mathcal{H}_{P_{\mathcal{A}}}} \max_{p(A|X,Y) \in Q_{\rho,\geq}} r_1(h) \quad \supseteq \{\arg\} \min_{h \in \mathcal{H}_{P_{\mathcal{A}}}} \sup_{p(A|X,Y) \in Q_{\rho,>}} r_1(h).$$

Meaning $R^{\text{I}} = R^{\text{III}}$, $\mathcal{H}^{\text{I}} \supseteq \mathcal{H}^{\text{III}}$

Secondly, we show that, under these conditions, minimizing the supremum over $h \in \mathcal{H}_{P_{\mathcal{A}}}$ is the same as minimizing over $h \in \mathcal{H}$. That is, $R^{\text{III}} = R^{\text{II}}$, $\mathcal{H}^{\text{III}} = \mathcal{H}^{\text{II}}$.

We observe that, if ℓ is a strictly convex function w.r.t h, and $p(A|X,Y) \in Q_{\rho,>}$, we can write the following statements.

Let $\hat{h}, \bar{h} \in \arg\min_{h \in \mathcal{H}} R_1(h; p(A|X,Y))$ such that $\hat{h}(x) \neq \bar{h}(x)$ if and only if x in some set $\bar{\mathcal{X}} \subseteq \mathcal{X}$, and let $h_\lambda = \lambda\hat{h} + (1-\lambda)\bar{h} \in \mathcal{H}\ \forall \lambda \in [0,1]$. By the strict convexity of ℓ we have

$$\ell(h_\lambda(X), Y) = \lambda\ell(\hat{h}(X), Y) + (1-\lambda)\ell(\bar{h}(X), Y)\ \forall X, Y \in \mathcal{X} \setminus \bar{\mathcal{X}} \times \mathcal{Y},$$
$$\ell(h_\lambda(X), Y) < \lambda\ell(\hat{h}(X), Y) + (1-\lambda)\ell(\bar{h}(X), Y)\ \forall X, Y \in \bar{\mathcal{X}} \times \mathcal{Y}.$$

Since for any $h \in \mathcal{H}$ we can write

$$\begin{aligned} r_1(h) &= \int_{x \in \bar{\mathcal{X}}} \int_{y \in \mathcal{Y}} \frac{p(x,y)p(a=1|X,Y)}{\rho} \ell(h(X),Y)dxdy \\ &+ \int_{x \in \mathcal{X} \setminus \bar{\mathcal{X}}} \int_{y \in \mathcal{Y}} \frac{p(x,y)p(a=1|X,Y)}{\rho} \ell(h(X),Y)dxdy, \end{aligned}$$

and we need $r_1(h_\lambda) \geq R_1(\bar{h}) = r_1(\hat{h})$, using the inequalities from the strict convexity of ℓ we note that $\bar{\mathcal{X}}$ must satisfy

$$\int_{x \in \bar{\mathcal{X}}} \int_{y \in \mathcal{Y}} \frac{p(x,y)p(a=1|X,Y)}{\rho} dxdy = 0,$$

or, equivalently, since $\{x, y : p(A=1|x,y) > 0)\} = \mathcal{X} \times \mathcal{Y}$ by hypothesis

$$\int_{x \in \bar{\mathcal{X}}} \int_{y \in \mathcal{Y}} p(x,y)dxdy = 0.$$

From this we conclude that $\hat{h}$ and $\bar{h}$ can differ only in a zero-measure set, and thus $r_0(\hat{h}) = r_0(\bar{h})$. Since all viable hypotheses $\hat{h}$ and $\bar{h}$ in $\mathcal{H}^{\text{II}}$ share the same risk vector (not only the group one risk value), then these hypotheses are all Pareto efficient. This then proves that $\hat{h}, \bar{h} \in \arg\min_{\mathcal{H}_{P_\mathcal{A}}} r_1(h; p(A|X,Y))$ for any $p(A|X,Y) \in Q_{\rho,>}$, and thus $\mathcal{H}^{\text{III}} = \mathcal{H}^{\text{II}}$.

□

Lemma 4.3. Consider the setting of Algorithm 1, with parameter $\epsilon > 0$, and $\eta = \max_{\boldsymbol{\alpha} \in \mathcal{U}_{\epsilon,\rho}} \frac{||\boldsymbol{\alpha}||_2}{\sqrt{2T}} \leq \sqrt{\frac{n\rho}{2T}}$ with $\mathcal{U}_{\epsilon,\rho} = \{\boldsymbol{\alpha} : \alpha_i \in [\epsilon, 1], \sum_i \frac{\alpha_i}{n} = \rho\}$, and L a 1-Lipschitz function w.r.t. $\boldsymbol{\alpha}$, let P be a uniform distribution over the set of models $\{h^1, \ldots, h^T\}$, and let R^* be the minimax solution to the loss presented in Eq. 9. Then we have

$$\max_{\boldsymbol{\alpha} \in \mathcal{U}_{\epsilon,\rho}} \mathbb{E}_{h \sim P} L(h, \boldsymbol{\alpha}) \leq \gamma R^* + \sqrt{\frac{2n\rho}{T}}.$$

Proof. We observe that loss function $L(h, \alpha)$ is concave (linear) w.r.t. α, and the set $\mathcal{U}_{\epsilon,\rho}$ is convex, with maximum norm $\max_{\alpha \in \mathcal{U}_{\epsilon,\rho}} ||\alpha||_2 \leq \sqrt{n\rho}$. For each $\epsilon > 0$ we are therefore able to use Theorem 7 in (Chen et al., 2017) to state

$$\begin{aligned} \max_{\alpha \in \mathcal{U}_{\epsilon,\rho}} \mathbb{E}_{h \sim P} L(h, \alpha) &\leq \gamma \min_{h \in \mathcal{H}} \max_{\alpha \in \mathcal{U}_{\epsilon,\rho}} L(h, \alpha) + \max_{\alpha \in \mathcal{U}_{\epsilon,\rho}} ||\alpha||_2 \sqrt{\frac{2}{T}}, \\ &\leq \gamma \min_{h \in \mathcal{H}} \max_{\alpha \in \mathcal{U}_{\epsilon,\rho}} L(h, \alpha) + \sqrt{\frac{2n\rho}{T}}. \end{aligned}$$

□

Lemma 4.4. Given $p(A{=}1|X,Y) \geq \epsilon, \forall X,Y$, $p(A{=}1){=}\rho$, and a bounded loss function $0 \leq \ell(h(X),Y) \leq C, \forall X,Y,h$ we denote the expected and empirical importance weighted risk as $r_1(h) = \mathbb{E}_{X,Y}[\frac{p(A=1|X,Y)}{p(A=1)}\ell(h(X),Y)]$ and $\hat{r}_1(h) = \sum_{i=1}^{n} \frac{p(A=1|x_i,y_i)}{np(A=1)}\ell(h(x_i),y_i)$ respectively. Where $h \in \mathcal{H}$ and $|\mathcal{H}|$ is the dimension of the hypothesis set. Under these conditions, the following PAC bound holds

$$P\left(\max_{h\in\mathcal{H}} |r_1(h) - \hat{r}_1(h)| \geq \frac{C}{\rho}\sqrt{\frac{\log\left(2|\mathcal{H}|/\delta\right)}{2n}}\right) \leq \delta$$

Proof. For a fixed h and a given distribution $p(A=1|X,Y)$, the random variable $\frac{p(A=1|X,Y)}{p(A=1)}\ell(h(X),Y) \in [0, \frac{C}{\rho}]$. Then, we can apply Hoeffding's concentration inequality and get

$$P(|r_1(h) - \hat{r}_1(h)| \geq \psi) \leq 2\exp\big(\frac{-2n\psi^2\rho^2}{C^2}\big).$$

Then, applying the union bound, we can extend this result for all $h \in \mathcal{H}$ we have

$$P(\max_{h\in\mathcal{H}} |r_1(h) - \hat{r}_1(h)| \geq \psi) \leq 2|\mathcal{H}|\exp\big(\frac{-2n\psi^2\rho^2}{C^2}\big) = \delta.$$

Writing ψ as a function of δ we have that

$$P\left(\max_{h\in\mathcal{H}} |r_1(h) - \hat{r}_1(h)| \geq \frac{C}{\rho}\sqrt{\frac{\log\left(2|\mathcal{H}|/\delta\right)}{2n}}\right) \leq \delta.$$

□

A.2. Comparing DRO and BPF on 0-loss distributions.

The methodology proposed in DRO (Hashimoto et al., 2018; Duchi et al., 2020) produces (constrained) minimax solutions that are not guaranteed to be Pareto efficient. In the context of Lemma 4.2, for any value of ρ, we can interpret any solution to Problem 4 to be a special subset of the potential solutions to the DRO objective. In particular, let h^* be a solution to Problem 4 with parameter ρ, and let $l_h = \ell(h^*(X),Y)$ be the random variable associated with the loss distribution of h. Let $\hat{\ell}_{h,\rho} = F_{l_h}^{-1}(1-\rho)$ be the $100 * (1-\rho)\%$ percentile of l_h. In this scenario, the hypothesis h^* is a valid solution to the DRO objective with parameter $\eta = \hat{\ell}_{h,\rho}$.

The reverse is not true in general, not all DRO solutions are Pareto efficient solutions, consider a simple binary classification scenario $\mathcal{X} \times \mathcal{Y} = [0,1] \times \{0,1\}$, with joint distribution $p(x)p(y|X) = U[-1,1]\delta_{y=1}$ and crossentropy loss. It is simple to see that $h(x) = 1$ is the only solution to Problem 4 for any value of ρ, however, the solution to

$$\min_{h:\mathcal{X}\to[0,1]} \mathbb{E}_{X,Y}[(-Y\log(h(X)) - (1-Y)\log(1-h(X)) - \eta)^+ + \eta]$$

is any function that satisfies $h(x) \geq e^{\eta}$, $\forall x \in [0,1]$. Which may yield sub-optimal classifiers even when a simple, 0-loss classifier is available. A numerical example of this is presented in the Figure 4, where we have a 1 dimension input $X \sim U[-1,1]$ and binary target variable $Y \in \{0,1\}$. Note that we define $p(Y=1|X)$ such that there is zero conditional entropy for $X \leq -0.5$ or $X \geq 0.5$; the major differences between DRO and BPF are observed in this noise-free, which was also illustrated in Figure 1. In this case we can see that, even though perfect classification is possible in some regions of the data distribution, DRO does not achieve perfect loss but BPF does. Although in practice some datasets may not present these noise free regions, using BPF does not sacrifice performance on the worst group. It can also control for best/worst group tradeoffs via the ϵ parameter (in case one wishes to sacrifice less performance on the best group), and has the advantage of providing the worst empirical distribution on the training data.

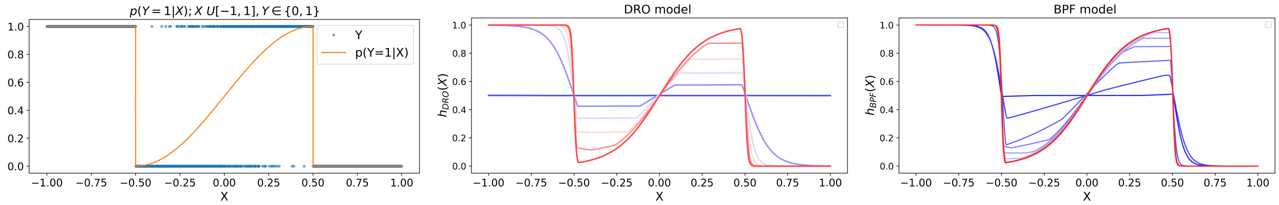

Figure 4. Synthetic data example with single-dimensional input $X \sim U[-1,1]$ and binary target variable $Y \in \{0,1\}$. Image on the left shows the conditional probability $p(Y=1|X)$ and some samples in blue. Images on the center and right show different solutions achieved by DRO and BPF respectively. It is worth noting that the classifiers recovered by each method are different, in particular, BPF is able to provide better service than DRO on noise-free regions of the input space without sacrificing performance elsewhere ($X \leq -0.5$ and $X \geq 0.5$)

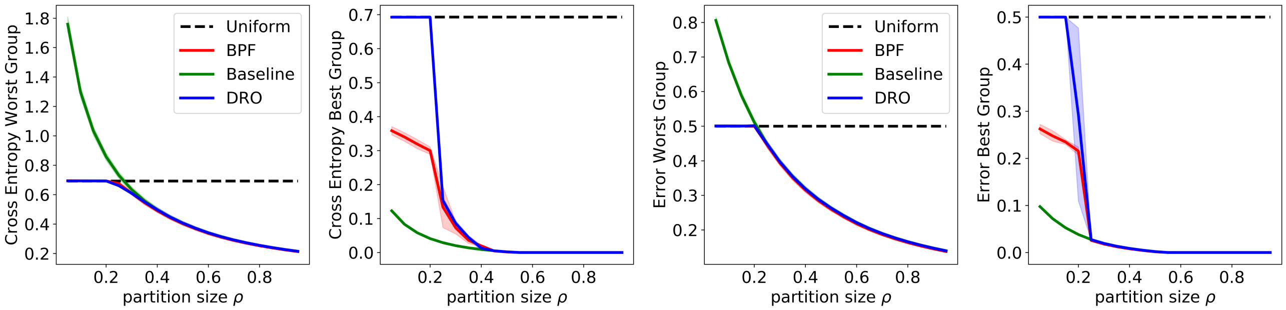

Figure 5. Performance comparison between DRO and BPF on the synthetic example presented in Figure 4. First two columns correspond to cross entropy performance on worst and best groups respectively, the latter two correspond to error rate on those same groups.

A.3. Additional Results

Similar to Table 1, tables 3 and 4 compare the performance of the competing methods (Baseline, ARL, DRO and BPF) on a predefined demographic. For the law school dataset we considered gender and outcome; race and outcome was considered for the Compas and MIMIC-III datasets. Table 5 show the demographic composition of worst groups based on the mentioned populations. It is worth noting that for these particular predefined groups there is no major difference between DRO and BPF.

Group	Prop(%)	ARL.15	DRO .15	BPF .15	ARL.25	DRO .25	BPF .25	ARL.4	DRO .4	BPF .4
White,Male	48.2%	68.9%	85.3%	86.8%	68.9%	94.1%	94.1%	68.9%	94.1%	94.1%
White,Female	35.5%	70.5%	85.2%	87.0%	70.5%	94.4%	94.5%	70.5%	94.4%	94.4%
Nonwhite,Male	7.8%	59.5%	76.6%	77.5%	59.5%	81.9%	82.0%	59.5%	81.9%	81.9%
Nonwhite,Female	8.5%	58.8%	75.9%	76.7%	58.8%	80.9%	80.9%	58.8%	80.9%	80.9%

Table 3. Accuracy on law school dataset across gender and ethnicity partitions (groups given no special consideration by the algorithms). Results shown for ARL, DRO and BPF models for varying partition sizes.

A.4. Experimental Details

To train the BPF classifier with standard SGD, we use the minibatch version of Eq.9

$$L(h,\boldsymbol{\alpha}) = \frac{1}{n_b}\frac{\sum_{i=1}^{n_b}\alpha_i \ell(h(x_i),y_i)}{\rho},$$

where n_b is the batch size, this loss is simply the minibatch equivalent of the importance-weighted loss

$$\mathbb{E}_{i\sim i|A=1}[\ell(h(x_i),y_i)] = \mathbb{E}_{i\sim U[\frac{\alpha_1}{N\rho},\dots,\frac{\alpha_N}{N\rho}]}[\ell(h(x_i),y_i)] = \mathbb{E}_{i\sim U[1,\dots,N]}[\frac{\alpha_i}{\rho}\ell(h(x_i),y_i)]$$

Group	Prop(%)	ARL.15	DRO .15	BPF .15	ARL.25	DRO .25	BPF .25	ARL.4	DRO .4	BPF .4
Black	8.5%	50.0%	52.8%	51.3%	50.0%	62.6%	60.6%	50.0%	74.7%	73.2%
White	79.3%	50.0%	52.8%	51.2%	50.0%	62.6%	60.6%	50.0%	74.5%	73.1%
Asian	2.7%	50.0%	52.7%	51.2%	50.0%	62.2%	60.2%	50.0%	73.8%	72.4%
Hispanic	3.5%	50.0%	53.1%	51.4%	50.0%	63.9%	61.7%	50.0%	77.3%	75.6%
Other	6.0%	50.0%	52.8%	51.2%	50.0%	62.3%	60.4%	50.0%	74.1%	72.6%

Table 4. Accuracy across gender and ethnicity partitions (groups given no special consideration by the algorithms) in the MIMIC-III dataset for ARL, DRO and BPF models for varying partition sizes.

Group/Outcome	prop(%)	BPF .15	BPF .30	BPF .40
Law school				
White,Male/0	1.5%	4.5%	2.6%	1.5%
White,Male/1	46.6%	36.9%	44.0%	46.6%
White,Female/0	1.2%	3.5%	2.0%	1.2%
White,Female/1	34.3%	27.6%	29.7%	34.3%
Nonwhite,Male/0	1.1%	3.2%	1.8%	1.1%
Nonwhite,Male/1	6.7%	9.2%	8.4%	6.7%
Nonwhite,Female/0	1.3%	3.8%	2.2%	1.3%
Nonwhite,Female/1	7.2%	11.4%	9.3%	7.2%
Compas				
African-American/0	24.7%	24.7%	24.7%	24.7%
African-American/1	27.1%	27.1%	27.1%	27.1%
Caucasian/0	20.9%	20.9%	20.9%	20.9%
Caucasian/1	13.4%	13.4%	13.4%	13.4%
Hispanic/0	5.2%	5.2%	5.2%	5.2%
Hispanic/1	3.1%	3.1%	3.1%	3.1%
Other/0	3.6%	3.6%	3.6%	3.6%
Other/1	2.0%	2.0%	2.0%	2.0%
MIMIC-III				
Black/0	7.6%	5.2%	5.2%	6.3%
Black/1	1.0%	3.4%	3.2%	2.2%
White/0	70.0%	48.4%	48.4%	58.7%
White/1	9.2%	31.0%	31.1%	20.7%
Asian/0	2.4%	1.7%	1.7%	2.0%
Asian/1	0.3%	1.1%	1.1%	0.8%
Hispanic/0	3.2%	2.2%	2.2%	2.7%
Hispanic/1	0.3%	0.9%	0.9%	0.6%
Other/0	5.3%	3.6%	3.6%	4.4%
Other/1	0.7%	2.4%	2.5%	1.6%

Table 5. Demographic composition of worst groups as a function of minimum partition size on the law school and Compas dataset. BPF homogenizes outcomes across partitions and protected attributes.

The training parameters used for all methods presented in the main paper are summarized in the following table

All methods are implemented using the same codebase on PyTorch. Experiments are run on a single GeForce RTX 2080 Ti and take less than 1 hour wall time each. The code is available at github.com/natalialmg/BlindParetoFairness.

Method	DRO	ARL	ARL HC	BPF
Learning rate	$1\exp-5$	$1\exp-5$	$1\exp-5$	$1\exp-5$
Batch size	128	128	128	128
Training Loss	cross entropy	cross entropy	cross entropy	cross entropy
Network architecture	(512,)	(512,)	(512,)	(512,)
Custom parameter	$\eta = \{0, 0.05, \ldots, 1\}$	(512,512,) , (512,), (256,256,), (256,)	(512,512,), (512,), (256,256,), (256,)	$\rho = \{0.05, ..., 0.9\}$, $\epsilon = 0.01$

Table 6. Summary of training parameters per method for all presented experiments. Network architecture is the number and size of the hidden layers of the classifer; for ARL and ARL HC, the adversarial network architecture is presented in the same format on the custom parameter row.