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

$$\begin{aligned} p(A = a' \mid X, Y) &= \begin{cases} 1 & \text{if } \ell(h(X), Y) > \hat{\ell}_{h,\rho}, \\ \alpha(X, Y) \in [0, 1] & \text{if } \ell(h(X), Y) = \hat{\ell}_{h,\rho}, \\ 0 & \text{if } \ell(h(X), Y) < \hat{\ell}_{h,\rho}, \end{cases} \\ p(A = a) &\geq \rho, \ \forall a \in \mathcal{A}, \\ p(A = a') &= \rho, \ a' \in \mathcal{A}. \end{aligned} \tag{9}$$

is a solution to

$$\max_{\substack{p(A|X,Y) \\ s.t.\ p(a) \geq \rho,\ \forall a \in \mathcal{A}}} \max_{a \in \mathcal{A}} R_a(h),$$

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

$$\max_{\substack{p(A|X,Y) \\ s.t.\ p(A=a) \geq \rho, \forall a \in \{0,1\}}} \max_{a \in \{0,1\}} R_a(h),$$

with both achieving the same maximum risk.

At this point we have proved the following equivalence:

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

Now we want to prove that, **in terms of worst case risk**, minimizing over $h \in \mathcal{H}$ is equivalent to minimizing over its respective Pareto classifiers sets, $h \in \mathcal{P}_{\mathcal{A},\mathcal{H}}$ for the left side of the equation and $h \in \mathcal{P}_{\mathcal{A}=\{0,1\},\mathcal{H}}$ for the right side.

Looking at the left side equation, we note that for all $\bar{h} \in \mathcal{H}$ and $\bar{p}(A|X, Y) : P(A = a) \geq \rho\ \forall a \in \mathcal{A}$, 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

$$\hat{h} \in \mathcal{P}_{\mathcal{A},\mathcal{H}} : R_{a'}(\hat{h}) = R_{a'}(\bar{h}), R_a(\hat{h}) \leq R_a(\bar{h})\ \forall a \in \mathcal{A} \setminus \{a'\}.$$

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 $\bar{h} \in \mathcal{P}_{\mathcal{A},\mathcal{H}}$ then $\bar{h} = \hat{h}$). Applying this property we observe that

$$\min_{h \in \mathcal{P}_{\mathcal{A},\mathcal{H}}} \max_{\substack{p(A|X,Y) \\ s.t.\ p(A=a) \geq \rho,\ \forall a \in \mathcal{A}}} \max_{a \in \mathcal{A}} R_a(h) = \min_{h \in \mathcal{H}} \max_{\substack{p(A|X,Y) \\ s.t.\ p(A=a) \geq \rho,\ \forall a \in \mathcal{A}}} \max_{a \in \mathcal{A}} R_a(h).$$

Using similar reasoning, we have that

$$\begin{array}{c}\min\limits_{h\in\mathcal{P}_{\mathcal{A}=\{0,1\},\mathcal{H}}} \max\limits_{p(A|X,Y)} \max\limits_{a\in\{0,1\}} R_a(h)\\ s.t.\ p(A{=}a)\geq\rho,\ \forall a\in\{0,1\}\end{array} = \begin{array}{c}\min\limits_{h\in\mathcal{H}} \max\limits_{p(A|X,Y)} \max\limits_{a\in\{0,1\}} R_a(h),\\ s.t.\ p(A{=}a)\geq\rho,\ \forall a\in\{0,1\}\end{array}$$

and thus,

$$\begin{array}{c}\min\limits_{h\in\mathcal{P}_{\mathcal{A},\mathcal{H}}} \max\limits_{p(A|X,Y)} \max\limits_{a\in\mathcal{A}} R_a(h)\\ s.t.\ p(A{=}a)\geq\rho,\ \forall a\in\mathcal{A}\end{array} = \begin{array}{c}\min\limits_{h\in\mathcal{P}_{\mathcal{A}=\{0,1\},\mathcal{H}}} \max\limits_{p(A|X,Y)} \max\limits_{a\in\{0,1\}} R_a(h),\\ s.t.\ p(A{=}a)\geq\rho,\ \forall a\in\{0,1\}\end{array}$$

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 $h\in\mathcal{P}_{\mathcal{A},\mathcal{H}}$

$\square$

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

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

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

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

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 $\min\limits_{h\in\mathcal{H}}\max\limits_{a\in\{0,1\}} R_a(h)\leq\bar{R}\,\forall p(X,Y,A)$.

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 $\bar{h}(X) : \bar{h}_i(X) = \frac{\mathbb{1}(i\in Y(X))}{|\mathcal{Y}(\mathcal{X})|} \forall X, \forall i \in \{0, ..., |\mathcal{Y}|-1\}$, 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 $E_X[|\mathcal{Y}(\mathcal{X})| \min_{y\in\mathcal{Y}(\mathcal{X})} p(y|X)] = \rho^*$.

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 $\hat{h}, \hat{R} = \{\arg\} \min_{h\in\mathcal{H}} \max_{a'\in\mathcal{A}'} R_{a'}(h)$ be the minimax fair solution for this partition and its corresponding minimax risk. Let h^* and R^* be the classifier and risks that solve Problem 4 with $\rho = \min_{a'} p(a')$. Then the price of minimax fairness can be upper bounded by

$$\max_{a'\in\mathcal{A}'} R_{a'}(h^*) - \hat{R} \leq R^* - \min_{h\in\mathcal{H}} R(h). \tag{10}$$

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

$$\min_{h\in\mathcal{H}} R(h) \leq R(h') = \sum_a p(a) R_a(h') \leq \max_a R_a(h').$$

We also have

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

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

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

□

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

$$\min_{h \in \mathcal{P}_{\mathcal{A},\mathcal{H}}} \max_{\substack{p(A|X,Y) \\ s.t.\ p(A=a) \geq \rho}} \max_{a \in \{0,1\}} R_a(h) = \min_{h \in \mathcal{P}_{\mathcal{A},\mathcal{H}}} \max_{\substack{p(A|X,Y) \\ s.t.\ p(A=1) = \rho}} R_1(h).$$

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 9 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 $h \in \mathcal{P}_{\mathcal{A},\mathcal{H}}$.

□

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:

$$\{\arg\} \min_{h \in \mathcal{P}_{\mathcal{A},\mathcal{H}}} \max_{\substack{p(A|X,Y) \\ s.t.\ p(A=1) = \rho}} R_{a=1}(h) = \{\arg\} \min_{h \in \mathcal{H}} \sup_{\substack{p(A|X,Y) \\ s.t.\ p(A=1) = \rho \\ p(A=1|X,Y) > 0 \, \forall X,Y}} R_{a=1}(h).$$

Proof. We present this proof in two steps. First, we show that, under the hypothesis class $\mathcal{P}_{\mathcal{A},\mathcal{H}}$, 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,

$$\{\arg\} \min_{h \in \mathcal{P}_{\mathcal{A},\mathcal{H}}} \max_{\substack{p(A|X,Y) \\ s.t.\ p(A=1) = \rho}} R_{a=1}(h) = \{\arg\} \min_{h \in \mathcal{P}_{\mathcal{A},\mathcal{H}}} \sup_{\substack{p(A|X,Y) \\ s.t.\ p(A=1) = \rho \\ p(A=1|X,Y) > 0 \, \forall X,Y}} R_{a=1}(h).$$

To prove this we start by defining the set of distributions complying with the restriction on the left hand side as

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

and the distribution subset on the right hand side as

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

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 $\hat{p}(A|X,Y) \in Q_{\rho,\geq}$, we can conclude

$$\{\arg\} \min_{h \in \mathcal{P}_{\mathcal{A},\mathcal{H}}} \max_{p(A|X,Y) \in Q_{\rho,\geq}} R_{a=1}(h) = \{\arg\} \min_{h \in \mathcal{P}_{\mathcal{A},\mathcal{H}}} \sup_{p(A|X,Y) \in Q_{\rho,>}} R_{a=1}(h).$$

Secondly, we show that, under these conditions, minimizing the supremum over $h \in \mathcal{P}_{\mathcal{A},\mathcal{H}}$ is the same as minimizing over $h \in \mathcal{H}$. That is,

$$\{\arg\} \min_{h \in \mathcal{P}_{\mathcal{A},\mathcal{H}}} \sup_{\substack{p(A|X,Y) \\ s.t.\ p(A=1) = \rho \\ p(A=1|X,Y) > 0\ \forall X,Y}} R_{a=1}(h) = \{\arg\} \min_{h \in \mathcal{H}} \sup_{\substack{p(A|X,Y) \\ s.t.\ p(A=1) = \rho \\ p(A=1|X,Y) > 0\ \forall X,Y}} R_{a=1}(h).$$

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

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

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

$$\begin{aligned} R_1(h) = &\int_{x \in \bar{\mathcal{X}}} \int_{y \in \mathcal{Y}} \frac{p(x,y)p(a=1|X,Y)}{\rho} \ell(h(X), Y) dx dy \\ &+ \int_{x \in \mathcal{X} \setminus \bar{\mathcal{X}}} \int_{y \in \mathcal{Y}} \frac{p(x,y)p(a=1|X,Y)}{\rho} \ell(h(X), Y) dx dy, \end{aligned}$$

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

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

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

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

From this we conclude that $\hat{h}$ and $\bar{h}$ can differ only in a zero-measure set, and thus $R_0(\hat{h}) = R_0(\bar{h})$, which implies that $\hat{h}, \bar{h} \in \arg\min_{\mathcal{P}_{\mathcal{A},\mathcal{H}}} R_1(h; p(A|X,Y))$ for any $p(A|X,Y) \in Q_{\rho,>}$.

□

Lemma 4.3. Consider the setting of Algorithm 1, with parameter $\epsilon > 0$ and $\eta = \max_{\alpha \in \{\alpha : \alpha_i \in [\epsilon, 1], \sum_i \frac{\alpha_i}{n} = \rho\}} \frac{||\alpha||_2}{\sqrt{2T}} \leq \sqrt{\frac{n\rho}{2T}}$, and L a 1-Lipschitz function w.r.t. α, 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. 8. Then we have

$$\max_{\alpha : \alpha_i \in [\epsilon, 1], \sum_i \frac{\alpha_i}{n} = \rho} \mathbb{E}_{h \sim P} L(h, \alpha) \leq \gamma R^* + \sqrt{\frac{2n\rho}{T}}.$$

Proof. We observe that loss function $L(h, \alpha)$ is concave (linear) w.r.t. α, and the set $\mathcal{Q}_{\rho,\epsilon} = \{\alpha \in \mathbb{R}^n : \alpha_i \in [\epsilon, 1], \sum_i \frac{\alpha_i}{n} = \rho\}$ is convex, with maximum norm $\max_{\alpha \in \mathcal{Q}_{\rho,\epsilon}} ||\alpha||_2 \leq \sqrt{n\rho}$. For each $\epsilon > 0$ we are therefore able to use Theorem 7 in Chen et al. (2017) to state

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

□

A.2 SYNTHETIC DATA

To design Figure 1 we used a simple synthetic dataset with covariates $X \in \{0, 1\}$ and target $Y \in \{0, 1\}$ were drawn from the following distribution

$$\begin{aligned} X &\sim \text{Ber}(0.5) \\ Y \mid X = 0 &\sim \text{Ber}(0.75) \\ Y \mid X = 1 &\sim \text{Ber}(0.9) \end{aligned}$$

We also designed a parametric family of partition functions $A \in \{0, 1\}$ to evaluate minimax fairness w.r.t. a known partition. For partition size $\rho < \frac{10}{16}$, the distribution $p(A \mid X, Y)$ follows

$$\begin{aligned} A \mid X = 0, Y = 0 &\sim \text{Ber}(1.6 * \rho) \\ A \mid X = 0, Y = 1 &\sim \text{Ber}(0.4 * \rho) \\ A \mid X = 1, Y = 0 &\sim \text{Ber}(1.6 * \rho) \\ A \mid X = 0, Y = 1 &\sim \text{Ber}(0.4 * \rho) \end{aligned}$$

and for $\frac{10}{16} \leq \rho \leq 1$

$$\begin{aligned} A \mid X = 0, Y = 0 &\sim \text{Ber}(1) \\ A \mid X = 0, Y = 1 &\sim \text{Ber}(0.4 * \rho') \\ A \mid X = 1, Y = 0 &\sim \text{Ber}(1) \\ A \mid X = 0, Y = 1 &\sim \text{Ber}(0.4 * \rho') \end{aligned}$$

where ρ' is scaled so that $p(A = 1) = \rho$. This partition function was chosen to ensure that the minimax classifier differs from the Bayes optimal classifier. For each value of $\rho \in [0, 1]$, we sampled a dataset X, Y, A from the joint distribution $p(X, Y, A)^{\otimes n}$, $n = 10K$ and ran both BPF and MMPF (Martinez et al. (2020)) on the resulting dataset. Results show empirical risks for both the predefined partition $p(A \mid X, Y)$, and also for the adversarial partition for each resulting classifier. A Jupyter notebook with the details will be provided.

A.3 ADDITIONAL RESULTS

Similar to Table 1, tables 3, 4 and 6 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 dataset; for MIMIC-III we considered gender and race with outcome (Mortality). 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 significant difference between DRO and BPF, moreover, in the case of Compas they do not seem to be deviating from the uniform classifier despite increasing the partition size, which could be due to a high level of noise.

Group/Outcome	Prop(%)	Baseline	ARL	DRO .15	BPF .15	DRO .3	BPF .3	DRO .5	BPF .5
female/0	8.6	41.5±0.2	37.2±2.3	50.3±0.2	50.1±0.7	50.7±0.6	51.1±1.3	48.5±1.0	48.2±1.4
female/1	34.9	86.5±0.5	79.9±1.7	50.1±0.1	51.0±1.2	51.9±0.8	51.8±0.3	70.5±1.3	73.2±0.2
male/0	11.1	44.1±0.6	37.1±2.0	50.3±0.3	50.2±0.3	51.0±0.8	50.9±1.0	50.3±1.0	49.1±1.3
male/1	45.4	87.1±0.1	81.1±1.4	50.1±0.0	51.3±1.3	52.1±0.9	52.6±0.4	71.3±1.0	74.0±0.1

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

Group/Outcome	Prop(%)	Baseline	ARL	DRO .15	BPF .15	DRO .3	BPF .3	DRO .5	BPF .5
African-American/0	24.8	51.1±0.6	53.5±0.4	50.2±0.0	50.4±0.8	50.2±0.0	49.8±0.7	50.3±0.1	50.4±0.4
African-American/1	27.9	60.1±0.2	63.8±0.2	50.2±0.1	50.7±0.9	50.2±0.1	50.2±0.2	50.4±0.2	49.5±0.5
Caucasian/0	19.2	64.1±0.6	65.0±0.5	50.4±0.1	50.6±0.8	50.5±0.1	50.4±0.8	50.7±0.3	49.9±0.6
Caucasian/1	14.1	45.0±0.4	49.0±0.5	50.0±0.1	51.5±0.9	49.9±0.1	49.6±0.7	49.9±0.1	50.3±0.7
Hispanic/0	4.5	63.8±0.4	70.0±1.0	50.7±0.2	50.4±0.4	50.7±0.3	50.4±0.7	51.1±0.6	49.9±0.5
Hispanic/1	3.7	42.6±0.6	43.2±0.7	49.8±0.2	50.6±0.9	48.7±0.2	50.8±0.8	49.4±0.2	50.1±0.5
Other/0	3.4	64.8±2.5	69.6±0.5	50.5±0.2	49.6±0.9	50.6±0.1	49.5±0.6	50.8±0.3	50.5±0.6
Other/1	2.4	43.0±1.3	43.7±1.2	50.0±0.3	49.5±0.8	50.0±0.2	49.9±1.2	50.0±0.3	51.0±0.8

Table 4: Accuracy on Compas dataset across ethnicity partitions (groups given no special consideration by the algorithms). Results shown for ARL, DRO and BPF models for varying partition sizes.

Group/Outcome	prop(%)	BPF .15	BPF .30	BPF .40	BPF .50
Law school					
female/0	8.6	23.0±0.0	21.7±0.2	18.4±0.1	15.2±0.0
female/1	34.9	22.2±0.7	23.0±0.2	26.4±0.5	29.1±0.1
male/0	11.1	28.5±0.3	27.1±0.0	23.1±0.2	19.2±0.1
male/1	45.4	26.2±0.4	28.1±0.4	32.1±0.2	36.5±0.1
Compas					
African-American/0	24.8	25.1±0.5	26.8±0.6	27.7±0.3	28.4±0.0
African-American/1	27.9	23.5±0.4	23.8±0.7	24.5±0.2	24.6±0.1
Caucasian/0	19.2	15.4±0.0	15.5±0.3	15.4±0.3	15.3±0.3
Caucasian/1	14.1	19.6±0.4	18.5±0.3	17.9±0.3	17.8±0.4
Hispanic/0	4.5	4.0±0.1	3.5±0.1	3.4±0.0	3.2±0.1
Hispanic/1	3.7	5.4±0.2	5.7±0.1	5.4±0.1	5.1±0.1
Other/0	3.4	2.5±0.1	2.5±0.0	2.5±0.0	2.4±0.1
Other/1	2.4	4.5±0.2	3.8±0.0	3.3±0.0	3.2±0.1

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.

Group	Prop (%)	Baseline	DRO .05	BPF .05	DRO .25	BPF .25	DRO .45	BPF .45
Gender/Outcome (1 if passed away, 0 if survived)								
male/0	50.5	92.8±0.9	54.3±4.7	60.6±8.8	67.0±8.1	64.6±2.2	79.0±1.4	87.7±3.0
male/1	6.3	38.0±15.0	52.2±3.6	55.9±19.3	48.8±6.3	49.3±6.9	42.4±11.0	46.5±18.5
female/0	38.1	92.2±0.8	54.2±4.5	59.6±8.6	66.1±7.6	64.3±2.0	77.3±1.4	86.8±3.3
female/1	5.2	40.8±16.0	52.5±3.7	55.7±18.0	49.7±6.3	49.3±6.0	43.9±10.5	48.0±18.1
Ethnicity/Outcome (1 if passed away, 0 if survived)								
White/0	69.9	92.6±0.8	54.3±4.6	60.3±8.8	66.6±7.9	64.5±2.1	78.3±1.4	87.3±3.1
White/1	9.3	39.6±15.8	52.4±3.7	55.4±18.5	49.4±6.5	49.9±6.6	43.1±11.0	47.6±18.6
Black/0	7.5	91.9±1.0	54.3±4.6	59.8±9.9	66.0±7.6	64.3±2.2	77.3±1.6	86.8±3.4
Black/1	0.9	40.7±14.6	52.3±3.3	58.6±19.6	49.3±5.3	50.8±6.1	44.0±9.1	47.3±17.5
Hispanic/0	3.3	91.6±1.2	54.1±4.6	59.6±6.8	66.2±7.8	64.6±2.0	77.7±1.3	86.7±3.3
Hispanic/1	0.3	45.1±13.5	51.9±3.3	56.9±18.5	50.5±5.6	49.3±6.0	45.8±9.0	49.3±16.0
Asian/0	2.3	91.9±1.4	54.3±4.6	58.2±8.1	66.0±7.7	64.4±2.6	77.1±1.4	86.7±3.6
Asian/1	0.3	38.0±9.8	52.3±3.5	56.9±20.2	48.0±5.4	48.0±6.4	42.0±9.7	48.1±17.5
Other/0	5.4	94.1±0.6	54.5±4.7	60.4±7.2	68.0±8.1	65.1±2.0	80.5±1.3	89.0±2.4
Other/1	0.7	31.5±14.9	51.8±3.2	56.9±21.1	47.3±6.0	50.6±6.0	40.1±11.2	40.8±16.5

Table 6: Accuracy across gender and ethnicity partitions (groups given no special consideration by the algorithms) in the MIMIC-III dataset for DRO and BPF models for varying partition sizes.