Optimal Fair Learning Robust to Adversarial Distribution Shift

Anonymous Authors[1]

Abstract

Previous work in fair machine learning has characterised the Fair Bayes Optimal Classifier (BOC) on a given distribution for both deterministic and randomized classifiers. We study the robustness of the Fair BOC to adversarial noise in the data distribution. Kearns & Li (1988) implies that the accuracy of the deterministic BOC without any fairness constraints is robust (Lipschitz) to malicious noise in the data distribution. We demonstrate that their robustness guarantee breaks down when we add fairness constraints. Hence, we consider the randomized Fair BOC, and our central result is that its accuracy is robust to malicious noise in the data distribution. Our robustness result applies to various fairness constraints—Demographic Parity, Equal Opportunity, Predictive Equality. Beyond robustness, we demonstrate that randomization leads to better accuracy and efficiency. We show that the randomized Fair BOC is nearly-deterministic, and gives randomized predictions on at most one data point, hence availing numerous benefits of randomness, while using very little of it.

1. Introduction

Machine learning models learn from the available training data and then are deployed to classify or rank new inputs. The effectiveness of machine learning models has resulted in improved efficiency across multiple domains but has also raised concerns about their fairness and possible amplification of biases in their training data (Barocas et al., 2019). When machine learning models are used to make decisions that skew the distribution of important economic resources or reinforce stereotypes, they compound disparities to cause social and economic harm. Fair classification has been an important topic of research, and binary fair classification where the model makes yes/no decisions algorithmically is a simple yet challenging setting to study foundational questions in optimal fair classification (Menon & Williamson, 2018b). In group-fair classification, each data point has certain sensitive attributes indicating the demographic group(s) to which it belongs (e.g., race, gender). Popular notions of group-fairness such as statistical or demographic parity, equal opportunity, equalized odds, and predictive parity are all motivated by the binary fair classification setting. Demographic parity prescribes the positivity rates to be equal across different groups (e.g., race, gender), whereas equal opportunity prescribes the true positive rates to be equal across different groups (Dwork et al., 2012; Hardt et al., 2016). Previous work has looked at various trade-offs between accuracy and fairness as well as the difficulty in satisfying multiple fairness constraints simultaneously (Celis et al., 2020). Previous work has also mathematically characterized the fair Bayes Optimal Classifier (BOC), namely, the optimal deterministic classifiers for maximizing accuracy subject to group-fairness constraints based such as demographic parity and equal opportunity (Menon & Williamson, 2018a; Chzhen et al., 2019; Celis et al., 2021; Zeng et al., 2022). Pre-processing or re-weighing for training data imbalances, in-processing by fairness-constrained training loss, and post-processing a model's predictions for balanced outcomes are three known ways to realize fair and accurate classifiers in practice (Kamiran & Calders, 2012; Agarwal et al., 2018; Barocas et al., 2019).

Biased or corrupted training data is a primary cause of unfairness in model predictions or outcomes. Moreover, robustness of a machine learning model under bias or corruption in the data distribution has been a more pragmatic concern that predates the research on fair machine learning. Learning robust classifiers is important because training and test distributions are not always identical and the training data may contain noise and malicious corruptions during data collection, curation, and annotation.

A particularly compelling and illustrative practical example for fair binary classification with maliciously corrupted training data is that of hate speech classifiers. Hate speech classifiers are known to exhibit biases against the same vulnerable demographics they were supposed to protect in online forums. For example, text in African American English (AAE) has higher likelihood of being misreported as

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

hate speech and even proper mentions of group identifiers such as ‘gay’ or ‘black’ get misreported as toxic or prejudiced. Moreover, the training data taken from online forums that is used to train hate speech classifiers contains societal biases of novice human annotators as well as malicious attempts made to bypass existing classifiers or filters used in data collections and annotation process (Davani et al., 2023; Davidson, 2023). Maliciously corrupted training data makes it difficult to train fair hate speech classifiers with robust accuracy and fairness guarantees that would be retained after real-world deployment (Davani et al., 2023; Davidson, 2023; Hartvigsen et al., 2022; Harris et al., 2022).

Classification under malicious noise is a theoretically challenging direction on its own, even without any fairness constraints. Balcan & Haghtalab (2020) survey research directions that originate from the work of Kearns & Li (1988) but focus on the hardness of learning linear classifiers under malicious noise and recent results that get around it. Unlike previous works on learning from malicious noise that consider any hypothesis class or a specific one such as linear classifiers, we consider the hypothesis class of all binary classifiers, deterministic as well as randomized.

Although previous work in fair machine learning has extensively studied Fair BOC and fair pre-/in-/post-processing methods to achieve best possible fairness-accuracy tradeoffs, their fairness and accuracy guarantees may not hold when training data is biased or contaminated and does not match test data. Adversarial or unknown bias in data makes it important to study the robustness of fairness and accuracy guarantees of Fair BOC. The seminal work of Kearns & Li (1988) implies the robustness to malicious noise of the deterministic BOC (without fairness constraints) in terms of a Lipschitz condition, i.e., given two similar distributions, the accuracy of the BOC on each distribution is also similar. In contrast, more recent findings by Konstantinov & Lampert (2022) reveal a concerning vulnerability: incorporating fairness constraints can render certain deterministic hypothesis classes non-robust to adversarial noise. This gap in understanding necessitates an investigation into the robustness of Fair BOC’s under adversarial distribution shift, which in turn is the focus of this paper.

1.1. Overview of Our Results

We summarize our key contributions.

- We demonstrate in Claim 1 (Section 3.1) that the deterministic Fair BOC is not robust to adversarial noise, corroborating Konstantinov & Lampert (2022).

Our main results concern the robustness of randomized Fair BOC’s.

- We prove in Theorems 1 (Section 3.2), 2 and 3 (Appendix B) that the randomized Fair BOC is robust to malicious noise across three popular fairness notions (Demographic Parity, Equal Opportunity, and Predictive Equality). This robustness is characterized by a (local) Lipschitz property, where the Lipschitz constant depends on the distribution (Yang et al., 2020). Toward this end, we first prove in Claims 2, 5, and 6 (Sections 3.2 and Appendix B) that a fixed hypothesis maintains comparable accuracy and fairness across two similar distributions. This, however, does not imply our main results since the Fair BOC may change significantly for neighboring distributions. We establish the Lipschitz property using a more sophisticated analysis of the specific structure of randomized fair BOC.

In addition to robustness, randomization confers multiple advantages.

- Claim 1 demonstrates that the Randomized Fair BOC can outperform its deterministic counterpart in accuracy by $0.5 - \epsilon$ (for any $\epsilon > 0$). We complement this with a tightness result in Claim 4 (Appendix A).
- The Randomized Fair BOC can be computed in polynomial time, whereas we prove in Claim 3 (Appendix A) that computing the deterministic Fair BOC is NP-complete.

Randomization is a very natural and useful resource for fairness as ties are often broken by a random coin toss. However, when it brings arbitrariness to critical decisions and causes societal harms, it needs to be used judiciously and sparingly (Creel & Hellman, 2021; Rosenblatt & Witter, 2024; Cooper et al., 2024). A key property of randomized Fair BOC is that it is *nearly deterministic*, being randomized at most on a single point in the domain and deterministic elsewhere. Thus, in a sense, we have the best of both worlds, preserving the benefits of randomization, while using very little of it.

Before presenting the problem formulation in Section 2, we present a geometric framework that provides intuition for the problem and our results in a unified context. More detailed comparison with most relevant previous work is given in Section 4, and we conclude in Section 5.

1.2. Geometric Intuition Behind our Results

The results in our paper are technical and involved as they quantitatively characterize the concerned Lipschitz constants. In what follows, we present a geometric intuition to help visualize how the arguments are structured. A learning algorithm takes as input a probability distribution and outputs a classifier. Consider a discrete setting for binary classification of n data points, where i-th data point has probability p_i and its (randomized) classifier prediction is $f_i \in [0,1]$, representing the probability that item i will be classified as 1. The space of all distributions is the simplex $\{(p_1,\ldots,p_n) \,:\, \Sigma_{i=1}^{n} p_i = 1 \text{ and } p_i \geq 0 \text{ for } 1 \leq i \leq n\}$, and the space of classifiers is the hypercube $[0,1]^n$. Thus, a learning algorithm is a map from the space of probability distribution to the space of classifiers. Now, we come describe three observations that form the crux of our intuition: fairness constraint is a hyperplane, accuracy is a linear objective, and robustness is continuity. Observe that a fairness criterion is a linear constraint (this may not be true of all fairness criteria, but is true of the well-known ones that we study in this paper). For example, demographic parity means the same proportion of both classes (A, D) of data points is classified as 1, which can be represented by a hyperplane H defined as $\Sigma_{i \in A} p_i f_i = \Sigma_{i \in D} p_i f_i$. Next, observe that accuracy is a linear objective, specifically $\Sigma_{i=1}^{n} p_i [f_i \mathcal{S}_i + (1-f_i)(1-\mathcal{S}_i)]$ where $\mathcal{S}_i$ (formally defined later as score) is the true probability that the i-th data point has class label 1. These two observations together imply that the most accurate classifier (subject to fairness) is a vertex of the intersection $H \cap [0,1]^n$. Since vertices of $H \cap [0,1]^n$ lie on the edges of $[0,1]^n$, it follows that the optimal fair classifier lies on the 1-skeleton of the hypercube, i.e., the vertices and edges of the hypercube (Kalai, 1994). This explains why the Fair BOC is deterministic on all but one of the data points - namely the data point corresponding to the edge of the hypercube on which the optimal vertex on the cut face lies. Note that these observations can be generalized from the entire hypercube to arbitrary hypothesis classes via 1-skeletons of their convex closures.

Lastly, we come to the observation that robustness is continuity. For the hypercube, so long as the input distribution moves continuously on the simplex, the output, the resulting optimal classifier, will be continuous on the hypercube. Note that the output may (will) not be continuous on the 1-skeleton of the hypercube as it may move along an edge of the cut face which is in a level set of the accuracy functional. Thus the accuracy (a linear functional) of the optimal will also be continuous. It is clear that continuity of the map is sufficient for robustness of the learning algorithm. Even though the fairness hyperplane and accuracy direction are (may be) correlated, continuity is necessary for robustness. In fact, the convex hull of a set of classifiers is the minimal body whose 1-skeleton contains all robust fair classifiers, i.e. the minimal robustification of the set of classifiers.

Though the geometric viewpoint is useful for understanding the robustness of fair classifiers it does not give us exact quantitative estimates of the sensitivity of output accuracy to input. When the input is close to the boundary of the simplex, where small perturbations can send the probability mass of a data point or an entire demographic class (A, D) to 0, it is easy to see that the sensitivity or Lipschitz constant will be very high. The focus of the paper is to get tight bounds on this Lipschitz constant and we expect that some of the techniques (such as flow-based decomposition) will find more general applications in the context of fairness and robustness. We remark that any fairness notion expressible as a linear function will satisfy the same robustness guarantee as a result of the reasoning above, however we only explicitly calculate Lipschitz constants for three of the most popular ones.

2. Problem Formulation

Consider a learning problem, where we are given a distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Z} \times \mathcal{Y}$, where $\mathcal{Z} = \{A, D\}$ represents the protected group membership (A denotes the advantaged group, and D denotes the disadvantaged group)[1], $\mathcal{X}$ represents all the other features, and $\mathcal{Y} = \{0,1\}$ represents the binary label set (we adopt the standard convention of associating the label 1 with success or acceptance). A randomized classification rule f is a function $f : \mathcal{X} \times \mathcal{Z} \to [0,1]$, where $f(x,z)$ denotes the probability of $(x,z) \in \mathcal{X} \times \mathcal{Z}$ being mapped to 1. A deterministic classifier is defined similarly, however the output of $f(x,z)$ is restricted to $\{0,1\}$. We consider the standard 0-1 loss function ℓ_{0-1}[2], whose expected value is given by $\mathcal{L}(f,\mathcal{P}) = \mathbb{E}[\ell_{0-1}(f)] = \Pr[f(X,Z) \neq Y]$, where the probability is over $(X,Z,Y) \sim \mathcal{P}$[3]. As is standard, we define accuracy as $\mathrm{Acc}(f,\mathcal{P}) = 1 - \mathcal{L}(f,\mathcal{P})$.

In many situations we may want to consider a fairness-aware learning problem, and find an accurate classifier on a given distribution that also satisfies some fairness constraints. Our work considers 3 of the most popular notions of fairness (Demographic Parity, Equal Opportunity, Predictive Equality). We present our results for Demographic Parity in the main body, and defer the results of the other 2 notions to Appendix B due to space constraints. We state the Demographic Parity definition below (Dwork et al., 2012).

Definition 1 (Demographic Parity)**.** Denote the selection rate for group z by $\mathrm{r}_z(f,\mathcal{P}) = \Pr[f(X,Z) = 1 \mid Z = z]$. f satisfies Demographic Parity[4] if the selection rates are equal across both groups, i.e., $\mathrm{r}_A(f,\mathcal{P}) = \mathrm{r}_D(f,\mathcal{P})$. We quantify the unfairness of f as the difference in selection rates across groups , i.e., $\mathrm{Unf}_{\mathrm{DP}}(f,\mathcal{P}) = |\mathrm{r}_A(f,\mathcal{P}) - \mathrm{r}_D(f,\mathcal{P})|$.

[1]Our results also hold when there are multiple groups, but for ease of exposition, we restrict our analysis to the case of 2 groups.

[2]Using the same proof techniques, our results also hold for the more general loss function ℓ_α, known in literature as cost-sensitive risk (Menon & Williamson, 2018b) , that assigns a weight α to False Positive errors, and a weight $(1-\alpha)$ to False Negative errors. However, for ease of exposition, we restrict our analysis here to ℓ_{0-1}.

[3]Henceforth, all probabilities will be over $(X,Z,Y) \sim \mathcal{P}$, unless explicitly stated.

[4]Classifiers satisfying DP will be often be referred to as DP-fair.

2.1. Fair Bayes Optimal Classifier

Given a distribution $\mathcal{P}$, the optimal (accuracy-maximizing) classifier f^* (the BOC) is given by $f^*(x,z) = \mathcal{T}_{\frac{1}{2}}(\Pr[Y = 1 \mid X = x, Z = z])$, where $\mathcal{T}_\gamma(\beta)$ is the threshold function that outputs 1 if $\beta \geq \gamma$, and 0 otherwise. We call the term β in the expression above the score or success probability of a point (x,z), and formally define it below.

Definition 2 (Score)**.** The score $\mathcal{S}$ of a point (x,z) is the probability that it has label 1, i.e., $\mathcal{S}(x,z) = \Pr[Y = 1 \mid (X = x, Z = z)]$.

The BOC basically accepts a point if its score is $\geq \frac{1}{2}$, and rejects it otherwise. Note that the BOC as described above is deterministic, and allowed for randomized classifiers will not provide any increase in accuracy. However, when fairness constraints are involved, the picture is more complicated, and it turns out that allowing for randomization actually can lead to a big jump in accuracy, as shown in Agarwal & Deshpande (2022). To see how randomization can improve the accuracy of fair classifiers, let us look at an example.

Example 1 (Accuracy jump in Randomized Fair BOC's)**.** Consider the following distribution $\mathcal{P}$[5] over $\mathcal{X} \times \mathcal{Z} \times \mathcal{Y}$, where $\mathcal{X} = \{x_1, x_2\}$ ($\mathcal{P}, \mathcal{S}(x,z) = (p,q)$ denotes that $\mathcal{P}(x,z) = p$, and $\mathcal{S}(x,z) = q$).

$$\mathcal{P}, \mathcal{S}(x_1, A) = (0.5, 0.75) \quad \mathcal{P}, \mathcal{S}(x_1, D) = (0.25, 0.5)$$
$$\mathcal{P}, \mathcal{S}(x_2, A) = (0, 0) \qquad\quad \mathcal{P}, \mathcal{S}(x_2, D) = (0.25, 0)$$

There are only 2 deterministic classifiers satisfying DP, either the constant 1 classifier f_1, or the constant 0 classifier f_0, with $\mathcal{L}(f_1) = \mathcal{L}(f_0) = \frac{1}{2}$. On the other hand, consider the following randomized classifier f, where $f(x_1, A) = \frac{1}{2}, f(x_1, D) = 1, f(x_2, A) = f(x_2, D) = 0$. It is easy to see that f satisfies DP, and $\mathcal{L}(f) = \frac{3}{8}$, hence improving over the accuracy of the optimal fair deterministic classifiers f_0 and f_1.

Given a distribution $\mathcal{P}$, Agarwal & Deshpande (2022) characterize the DP-Fair BOC (the optimal classifier subject to DP constraints) on a given distribution, which we now describe. We first detail some of their terminology.

Definition 3 (Cell)**.** Consider a randomized partition of the feature space $\mathcal{X} \times \mathcal{Z}$ into multiple disjoint components. We call these components cells, and denote a cell by $\mathcal{C}$.

One can also define the score of a cell, in the same way as we had defind the score of a point. We have already seen the BOC that thresholds based on scores. Randomized classifiers give us the ability to threshold by probability mass, instead of just thresholding by scores. To explain this better, we introduce the notion of group-wise sorted cells.

[5]Note that specifying a distribution over $\mathcal{X} \times \mathcal{Z} \times \mathcal{Y}$ is equivalent to specifying a distribution over $\mathcal{X} \times \mathcal{Z}$ along with the scores for every element, i.e., $\mathcal{S}(x,z)\ \forall\ (x,z) \in \mathcal{X} \times \mathcal{Z}$.

Definition 4 (Group-wise Sorted Cells)**.** Define $\mathcal{C}_z = \bigcup_{x \in \mathcal{X}} \mathcal{C}_{x,z}$, where the component cells of $\mathcal{C}_A$ and $\mathcal{C}_D$ are arranged in descending order of scores $\mathcal{S}$. If two or more cells from the same group have the same score, any ordering within them is acceptable.

By $\mathcal{C}_z(t)$, denote the topmost cells of $\mathcal{C}_z$ comprising of t fraction of the total probability mass of $\mathcal{C}_z$. Note that this may involve splitting a cell into 2 parts randomly. For example, in Example 1, $\mathcal{C}_A(\frac{1}{2})$ would involve splitting $\mathcal{C}_{x_1,A}$ into two equal parts randomly. However, in the deterministic setting, only $\mathcal{C}_A(0)$ and $\mathcal{C}_A(1)$ are defined, and $\mathcal{C}_A(\frac{1}{2})$ does not make sense. By $\tilde{\mathcal{T}}_t$, we denote the mass threshold classifiers accepts exactly $\mathcal{C}_z(t)$ for $z \in \mathcal{Z}$. In Example 1, the randomized classifier f we constructed is actually the mass-threshold classifier $\tilde{\mathcal{T}}_{\frac{1}{2}}$.

Definition 5 (Score Boundaries)**.** Consider the component cells of groupwise sorted $\mathcal{C}_A$ and $\mathcal{C}_D$. Then, $\mathcal{I} = \mathcal{I}_A \cup \mathcal{I}_D$, where $\mathcal{I}_z$ consists of all the boundary points between component cells in $\mathcal{C}_z$.

Definition 6 (Merged Cells)**.** Consider any $r_i \in \mathcal{I}$ in sorted order, and define a merged cell $\mathcal{C}_i$ as $\mathcal{C}_i = \mathcal{A}(\tilde{\mathcal{T}}_{r_i}) - \mathcal{A}(\tilde{\mathcal{T}}_{r_{i-}})$, where $\mathcal{A}(f)$ denotes the accepted elements by f, and r_{i-} denotes the element in $\mathcal{I}$ preceding r_i.

Characterization Given a distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Z} \times \mathcal{Y}$, the DP-Fair BOC $f^{\mathrm{DP}}_{\mathcal{P}}$ is given by the mass-threshold classifier $\tilde{\mathcal{T}}_{r'}$, where $r' = r_i \in \mathcal{I}$ is the unique i such that $\mathcal{S}(C_i) \geq 0.5$, and $\mathcal{S}(C_{i+}) < 0.5$, where r_{i+} denotes the element in $\mathcal{I}$ after r_i. Note that the DP-Fair BOC needs to use randomization on only one cell in the whole domain, since the candidate r' values lie in $\mathcal{I}$.

3. Robustness to Adversarial Distribution Shift

We study the robustness of the DP-Fair BOC to adversarial distribution shift. We show that given 2 similar distributions $\mathcal{P}, \mathcal{P}'$, the accuracy of the DP-Fair BOC on the respective distributions is similar (satisfies local Lipschitzness). Note that DP-fair BOC in the deterministic case does not exhibit such a robustness property, as we demonstrate in this example.

3.1. Non-Robustness of the Deterministic Fair BOC

Claim 1 (Non-Robustness of Deterministic Fair BOC)**.** *There exist distributions $\mathcal{P}, \mathcal{P}'$ with $TV(\mathcal{P}, \mathcal{P}') = \epsilon$, such that the deterministic DP-fair BOC's f, f' on $\mathcal{P}, \mathcal{P}'$, respectively, satisfies $|Acc(f, \mathcal{P}) - Acc(f', \mathcal{P}')| \geq \Omega(1)$.*

Proof. Consider the following distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Z} \times$

$\mathcal{Y}$, where $\mathcal{X} = \{x_1, x_2\}$.

$$\mathcal{P}, \mathcal{S}(x_1, A) = (0.25, 1) \quad \mathcal{P}, \mathcal{S}(x_1, D) = (0.25, 1)$$
$$\mathcal{P}, \mathcal{S}(x_2, A) = (0.25, 0) \quad \mathcal{P}, \mathcal{S}(x_2, D) = (0.25, 0)$$

Consider the (deterministic) classifier f, with $f(x_1, A) = f(x_1, D) = 1, f(x_2, A) = f(x_2, D) = 0$. It is easy to see that f satisfies DP, and $\mathrm{Acc}(f) = 1$, implying that f is the DP-Fair BOC in both the deterministic and randomized settings. Consider the neighboring distribution $\mathcal{P}'$ as follows.

$$\mathcal{P}', \mathcal{S}(x_1, A) = (0.25, 1) \quad \mathcal{P}', \mathcal{S}(x_1, D) = (0.25 + \epsilon, 1)$$
$$\mathcal{P}', \mathcal{S}(x_2, A) = (0.25, 0) \quad \mathcal{P}', \mathcal{S}(x_2, D) = (0.25 - \epsilon, 0)$$

There are only 2 deterministic classifiers satisfying DP, either the constant 1 classifier f_1, or the constant 0 classifier f_0, with $\mathcal{L}(f_1) = \frac{1}{2} + \epsilon$, and $\mathcal{L}(f_0) = \frac{1}{2} - \epsilon$, implying that f_1 is the DP-Fair BOC in the deterministic setting. Hence, the difference in accuracy of the deterministic DP-fair BOC on arbitrarily close $\mathcal{P}, \mathcal{P}'$ is almost 0.5, demonstrating the non-robustness for deterministic classifiers to distribution shift. □

3.2. Robustness of the Randomized Fair BOC

Before proving the main result (Theorem 1), we prove Lemmas 1 and 2, and Claim 2 that will help us prove Theorem 1. Lemma 1 shows that one can decompose a transition from distribution $\mathcal{P}$ to distribution $\mathcal{P}'$ with distance ϵ into a sequence of elementary transitions from $\mathcal{P}_{i-1}$ to $\mathcal{P}_i$ with distance ϵ_i such that $\epsilon = \sum_i \epsilon_i$ and for every i, the only difference between $\mathcal{P}_{i-1}$ and $\mathcal{P}_i$ is that mass is transferred from exactly one element of the domain to another.

Lemma 1 (Decomposition into Elementary Transitions)**.** *Given distributions $\mathcal{P}, \mathcal{P}'$ with $TV(\mathcal{P}, \mathcal{P}') = \epsilon$, there exist distributions $\mathcal{P}_0, \mathcal{P}_1, \ldots, \mathcal{P}_n$ (for some n), such that the following two conditions hold:*

1. *Decomposability: $\mathcal{P} = \mathcal{P}_0$, $\mathcal{P}' = \mathcal{P}_n$, $TV(\mathcal{P}_{i-1}, \mathcal{P}_i) = \epsilon_i$, $\sum_{i=1}^n \epsilon_i = \epsilon$, and in the transition $\mathcal{P}_{i-1} \to \mathcal{P}_i$, ϵ_i mass moves from some element a_i to some b_i (all other elements remain constant).*
2. *Monotonicity: If $\mathcal{P}(A) \leq \mathcal{P}'(A)$, then for every $1 \leq i < n$, $\mathcal{P}_i(A) \leq \mathcal{P}_{i+1}(A)$ and $\mathcal{P}_i(D) \geq \mathcal{P}_{i+1}(D)$; otherwise, $\mathcal{P}_i(A) \geq \mathcal{P}_{i+1}(A)$ and $\mathcal{P}_i(D) \leq \mathcal{P}_{i+1}(D)$.*

Proof. We will prove the desired claim for n equal to number of elements q for which $\mathcal{P}(q) \neq \mathcal{P}(q')$. Our proof is by induction on n. For the base case, we have $n = 0$, in which case $\mathcal{P} = \mathcal{P}'$ and the claim trivially holds. For the induction step, let a be an element such that $\mathcal{P}(a) \neq \mathcal{P}'(a)$. Suppose $\mathcal{P}(a) > \mathcal{P}'(a)$ and a is in group A; the arguments for the other scenarios are analogous. We consider two cases. The first case is when there exists $b \in A$ such that $\mathcal{P}(b) < \mathcal{P}'(b)$. We define $\tilde{\mathcal{P}}$ as the same as $\mathcal{P}$ except that

$$\tilde{\mathcal{P}}(a) = \mathcal{P}(a) - \min\{\mathcal{P}(a) - \mathcal{P}'(a), \mathcal{P}'(b) - \mathcal{P}(b)\}$$
$$\tilde{\mathcal{P}}(b) = \mathcal{P}(b) + \min\{\mathcal{P}(a) - \mathcal{P}'(a), \mathcal{P}(b) - \mathcal{P}'(b)\}.$$

Note that either $\tilde{\mathcal{P}}(a) = \mathcal{P}'(a)$ or $\tilde{\mathcal{P}}(b) = \mathcal{P}'(b)$, which implies that the number of elements for which $\tilde{\mathcal{P}}$ and $\mathcal{P}'$ differ is less than n. Furthermore, $\mathcal{P}(A) = \tilde{\mathcal{P}}(A)$ and $\mathcal{P}(D) = \tilde{\mathcal{P}}(D)$. By induction, there exist a sequence of $m < n$ distributions $\tilde{\mathcal{P}} = \mathcal{P}_0, \mathcal{P}_1, \ldots, \mathcal{P}_m = \mathcal{P}'$ satisfying the decomposability and monotonicity properties. Appending the elementary transition $\mathcal{P} \to \tilde{\mathcal{P}}$ to the above sequence yields the desired sequence for $\mathcal{P}$ and $\mathcal{P}'$ with the decomposability and monotonicity properties.

The second case is when there does not exist any $b \in A$ such that $\mathcal{P}(b) < \mathcal{P}'(b)$. So, we have $\mathcal{P}(A) > \mathcal{P}'(A)$. Furthermore, there exists $b \in D$ such that $\mathcal{P}(b) < \mathcal{P}'(b)$. We define $\tilde{\mathcal{P}}$ in the same way as for the first case. Again, we have that either $\tilde{\mathcal{P}}(a) = \mathcal{P}'(a)$ or $\tilde{\mathcal{P}}(b) = \mathcal{P}'(b)$, which implies that the number of elements for which $\tilde{\mathcal{P}}$ and $\mathcal{P}'$ differ is less than n. Furthermore, $\mathcal{P}(A) > \tilde{\mathcal{P}}(A)$ and $\mathcal{P}(D) < \tilde{\mathcal{P}}(D)$. By induction, there exist a sequence of $m < n$ distributions $\tilde{\mathcal{P}} = \mathcal{P}_0, \mathcal{P}_1, \ldots, \mathcal{P}_m = \mathcal{P}'$ satisfying the decomposability and monotonicity properties. Again, appending the elementary transition $\mathcal{P} \to \tilde{\mathcal{P}}$ to the above sequence yields the desired sequence for $\mathcal{P}$ and $\mathcal{P}'$ with the decomposability and monotonicity properties. □

Claim 2 roughly states that given 2 similar distributions $\mathcal{P}, \mathcal{P}'$, the accuracy and DP-unfairness of any fixed hypothesis is similar on both $\mathcal{P}, \mathcal{P}'$. Such a property is useful when we want a guarantee that if we train a classifier on the corrupted distribution $\mathcal{P}'$, the performance of the classifier on the actual distribution $\mathcal{P}$ will be similar to that on $\mathcal{P}'$.

Claim 2 (Accuracy, DP Shift for Fixed Hypothesis)**.** *Given distributions $\mathcal{P}, \mathcal{P}'$, such that $TV(\mathcal{P}, \mathcal{P}') \leq \epsilon$, any hypothesis f satisfies the following two properties:*

1. *$|Acc(f, \mathcal{P}) - Acc(f, \mathcal{P}')|$ is at most ϵ.*
2. *$|Unf_{DP}(f, \mathcal{P}) - Unf_{DP}(f, \mathcal{P}')|$ is at most*[6]

$$\epsilon \left(\frac{1}{\min(\mathcal{P}(A), \mathcal{P}'(A))} + \frac{1}{\min(\mathcal{P}(D), \mathcal{P}'(D))} \right).$$

Proof. We first establish the desired statements for the special case where the transition from $\mathcal{P}$ to $\mathcal{P}'$ is elementary in that the only difference between the two distributions is that

[6]Note that the Lipschitz constant will blow up if the masses of either group becomes very small. Similar terms in the denominator will naturally feature in all our bounds. As such, robustness is not satisfied at such extremal points.

there are two elements a and b that have ϵ more mass and ϵ less mass, respectively, in $\mathcal{P}$ as compared to $\mathcal{P}'$ (all other elements have the same mass in the two distributions). At the end, we invoke Lemma 1 and transitivity to establish the general claim.

Accuracy: Divide the domain into 4 parts based on whether a point falls in categories TP, FP, TN, or FN according to f. Denote the probability mass of elements in category E under f by $\mathcal{P}(E)$. We know that $\mathrm{Acc}(f, \mathcal{P}) = \mathcal{P}(\mathrm{TP} \cup \mathrm{TN})$. In the worst case, a belongs to $\mathrm{TP} \cup \mathrm{TN}$, and b belongs to $\mathrm{FP} \cup \mathrm{FN}$. This transition leads to a loss in accuracy of ϵ, i.e., $\mathrm{Acc}(f, \mathcal{P}') = \mathrm{Acc}(f, \mathcal{P}) - \epsilon$. We note that it is enough to consider a loss in accuracy, since we can reverse the roles of the distributions and use the same argument for gain as that for loss.

Demographic Parity: First we notice the following

$$
\begin{aligned}
&|\mathrm{Unf}_{\mathrm{DP}}(f, \mathcal{P}) - \mathrm{Unf}_{\mathrm{DP}}(f, \mathcal{P}')| \\
&= ||\mathrm{r}_A(f, \mathcal{P}) - \mathrm{r}_D(f, \mathcal{P})| - |\mathrm{r}_A(f, \mathcal{P}') - \mathrm{r}_D(f, \mathcal{P}')|| \\
&\leq |(\mathrm{r}_A(f, \mathcal{P}) - \mathrm{r}_D(f, \mathcal{P})) - (\mathrm{r}_A(f, \mathcal{P}') - \mathrm{r}_D(f, \mathcal{P}'))| \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Triangle inequality)} \\
&= |(\mathrm{r}_A(f, \mathcal{P}) - \mathrm{r}_A(f, \mathcal{P}')) + (\mathrm{r}_D(f, \mathcal{P}') - \mathrm{r}_D(f, \mathcal{P}))| \\
&\leq |\mathrm{r}_A(f, \mathcal{P}) - \mathrm{r}_A(f, \mathcal{P}')| + |\mathrm{r}_D(f, \mathcal{P}') - \mathrm{r}_D(f, \mathcal{P})| \quad (1)
\end{aligned}
$$

The above argument breaks up the change in unfairness into two terms: (i) $\Delta\mathrm{r}_A \triangleq |\mathrm{r}_A(f, \mathcal{P}) - \mathrm{r}_A(f, \mathcal{P}')|$, which is the difference in selection rates of f for $\mathcal{P}$, and $\mathcal{P}'$ on A and (ii) $\Delta\mathrm{r}_D \triangleq |\mathrm{r}_D(f, \mathcal{P}) - \mathrm{r}_D(f, \mathcal{P}')|$, which is the difference in selection rates of f for $\mathcal{P}$ and $\mathcal{P}'$ on D.

We proceed to bound $\Delta\mathrm{r}_A$, and an identical argument can be used to bound $\Delta\mathrm{r}_D$. In our argument, we divide the domain into 4 parts based on the group membership and labeling according to f. Let the probability mass of elements in group z with label y under classifier f be denoted by $\mathcal{P}(z, f_y)$. If a, b lie in the same group z then $\mathcal{P}(A)$ remains unchanged, and it is easy to see that the maximum value of $\Delta\mathrm{r}_A$ is $\frac{\epsilon}{\mathcal{P}(A)}$, when $\mathcal{P}'(A, f_1) = \mathcal{P}(A, f_1) \pm \epsilon$. In case $a \in A$, and $b \in D$, then $\mathcal{P}'(A) = \mathcal{P}(A) - \epsilon$. We know that $\mathcal{P}'(A) = \mathcal{P}'(A, f_1) + \mathcal{P}'(A, f_0)$. We first consider the case where $\mathcal{P}'(A, f_1) = \mathcal{P}(A, f_1) - \epsilon$.

$$
\begin{aligned}
&|\mathrm{r}_A(f, \mathcal{P}) - \mathrm{r}_A(f, \mathcal{P}')| = \left|\frac{\mathcal{P}(A, f_1)}{\mathcal{P}(A)} - \frac{\mathcal{P}(A, f_1) - \epsilon}{\mathcal{P}(A) - \epsilon}\right| = \\
&= \left|\frac{\mathcal{P}(A)\epsilon - \mathcal{P}(A, f_1)\epsilon}{\mathcal{P}(A)\,(\mathcal{P}(A) - \epsilon)}\right| \leq \epsilon \left|\frac{1}{\mathcal{P}(A) - \epsilon}\right| \\
&\leq \epsilon \left(\frac{1}{\min(\mathcal{P}(A), \mathcal{P}'(A))}\right)
\end{aligned}
$$

We now consider the case where $\mathcal{P}'(A, f_0) = \mathcal{P}(A, f_0) - \epsilon$.

$$
\begin{aligned}
&|\mathrm{r}_A(f, \mathcal{P}) - \mathrm{r}_A(f, \mathcal{P}')| = \left|\frac{\mathcal{P}(A, f_1)}{\mathcal{P}(A)} - \frac{\mathcal{P}(A, f_1)}{\mathcal{P}(A) - \epsilon}\right| \\
&= \left|\frac{\mathcal{P}(A, f_1)\epsilon}{\mathcal{P}(A)\,(\mathcal{P}(A) - \epsilon)}\right| \leq \epsilon \left|\frac{1}{\mathcal{P}(A) - \epsilon}\right| \\
&\leq \epsilon \left(\frac{1}{\min(\mathcal{P}(A), \mathcal{P}'(A))}\right)
\end{aligned}
$$

Here we argued for when A loses mass. Using symmetry, we can similarly argue the case where A gains mass, i.e., $a \in D$, and $b \in A$, leading to $\mathcal{P}'(A) = \mathcal{P}(A) + \epsilon$. Hence, we conclude that

$$
|\mathrm{r}_A(f, \mathcal{P}) - \mathrm{r}_A(f, \mathcal{P}')| \leq \epsilon \left(\frac{1}{\min(\mathcal{P}(A), \mathcal{P}'(A))}\right) \quad (2)
$$

Also, here we argued for group A, and an identical argument for D shows that

$$
|\mathrm{r}_D(f, \mathcal{P}) - \mathrm{r}_D(f, \mathcal{P}')| \leq \epsilon \left(\frac{1}{\min(\mathcal{P}(D), \mathcal{P}'(D))}\right) \quad (3)
$$

Plugging Equations 2 and 3 into Equation 1, we get that

$$
\begin{aligned}
&|\mathrm{Unf}_{\mathrm{DP}}(f, \mathcal{P}) - \mathrm{Unf}_{\mathrm{DP}}(f, \mathcal{P}')| \leq \\
&\epsilon \left(\frac{1}{\min(\mathcal{P}(A), \mathcal{P}'(A))}\right) + \epsilon \left(\frac{1}{\min(\mathcal{P}(D), \mathcal{P}'(D))}\right)
\end{aligned}
$$

From elementary to arbitrary: Consider a general transition of distance ϵ from $\mathcal{P}$ to $\mathcal{P}'$. We invoke Lemma 1 to obtain intermediate distributions $\{\mathcal{P}_i\}$ with $TV(\mathcal{P}_{i-1}, \mathcal{P}_i) = \epsilon_i$ satisfying the decomposability and monotonocity properties. We apply the above proof for each elementary transition $\mathcal{P}_{i-1} \to \mathcal{P}_i$ of mass ϵ_i. For accuracy, we derive

$$
\begin{aligned}
|\mathrm{Acc}(f, \mathcal{P}) - \mathrm{Acc}(f, \mathcal{P}')| &\leq \sum_i |\mathrm{Acc}(f, \mathcal{P}_{i-1}) - \mathrm{Acc}(f, \mathcal{P}_i)| \\
&\leq \sum_i \epsilon_i = \epsilon.
\end{aligned}
$$

For Demographic Parity, we derive

$$
\begin{aligned}
&|\mathrm{Unf}_{\mathrm{DP}}(f, \mathcal{P}) - \mathrm{Unf}_{\mathrm{DP}}(f, \mathcal{P}')| \leq \sum_i |\mathrm{Unf}_{\mathrm{DP}}(f, \mathcal{P}_{i-1}) - \mathrm{Unf}_{\mathrm{DP}}(f, \mathcal{P}_i)| \\
&\leq \sum_i \epsilon_i \left(\frac{1}{\min(\mathcal{P}_{i-1}(A), \mathcal{P}_i(A))} + \frac{1}{\min(\mathcal{P}_{i-1}(D), \mathcal{P}_i(D))}\right) \\
&\leq \sum_i \epsilon_i \left(\frac{1}{\min(\mathcal{P}(A), \mathcal{P}'(A))} + \frac{1}{\min(\mathcal{P}(D), \mathcal{P}'(D))}\right) \\
&= \epsilon \left(\frac{1}{\min(\mathcal{P}(A), \mathcal{P}'(A))} + \frac{1}{\min(\mathcal{P}(D), \mathcal{P}'(D))}\right),
\end{aligned}
$$

where the second inequality follows from monotonocity and the last equation follows from decomposability. This completes the proof of the claim. □

We also use Lemma 2 (proof in Appendix A) for our main result .

Lemma 2 (Sensitivity of Accuracy). *Given any $\mathcal{P}, f$, and $\mathcal{P}', f'$ such that $TV(\mathcal{P}, \mathcal{P}') = \epsilon$, $f'(q)$ differs from $f(q)$ by $\Delta f(q)$ (and is identical elsewhere), then*

$$|Acc(f, \mathcal{P}) - Acc(f', \mathcal{P}')| \leq |\mathcal{P}(q)(2\mathcal{S}(q) - 1)\Delta f(q)| + \epsilon,$$

Now we state and prove our main result.

Theorem 1 (Robustness of DP-Fair BOC). *Given distributions $\mathcal{P}, \mathcal{P}'$ with $TV(\mathcal{P}, \mathcal{P}') = \epsilon$, $\left|Acc(f^{DP}_{\mathcal{P}}, \mathcal{P}) - Acc(f^{DP}_{\mathcal{P}'}, \mathcal{P}')\right|$ is at most*

$$\epsilon\left(1 + \frac{\max\left(\mathcal{P}(A), \mathcal{P}'(A)\right)}{\min(\mathcal{P}(A), \mathcal{P}'(A))} + \frac{\max\left(\mathcal{P}(D), \mathcal{P}'(D)\right)}{\min(\mathcal{P}(D), \mathcal{P}'(D))}\right).$$

Proof. We first establish the claim of the theorem for the special case where the transition from $\mathcal{P}$ to $\mathcal{P}'$ is elementary in that the only difference between the two distributions is that there are two elements a and b that have ϵ more mass and ϵ less mass, respectively, in $\mathcal{P}$ as compared to $\mathcal{P}'$ (all other elements have the same mass in the two distributions). At the end, we invoke Lemma 1 and transitivity to establish the general theorem statement.

Consider the transfer of ϵ mass from a to b in a continuous manner. During this process, either the cell corresponding to element a will monotonically increase in score or monotonically decrease in score. The same holds for the cell corresponding to element b. The scores of all other cells will remain the same. In the following argument, we assume that the score of the cell of a decreases monotonically and that of b increases monotonically. All of the arguments are analogous for the remaining three cases.

We break down the ϵ mass transfer into smaller increments. At any point, let $\tilde{\mathcal{P}}$ be the distribution at the start of this increment (so, $\tilde{\mathcal{P}} = \mathcal{P}$ initially) and $\tilde{\mathcal{P}}'$ be the distribution at the end of this increment (so, $\tilde{\mathcal{P}}' = \mathcal{P}'$ finally). For an incremental mass transfer, we analyze how the DP BOC changes from $f^{\text{DP}}_{\tilde{\mathcal{P}}}$ to $f^{\text{DP}}_{\tilde{\mathcal{P}}'}$. Since the mass transfer is from element a to b, it follows that both $\tilde{\mathcal{P}}(A)$ and $\tilde{\mathcal{P}}'(A)$ lie between $\mathcal{P}(A)$ and $\mathcal{P}'(A)$ while both $\tilde{\mathcal{P}}(D)$ and $\tilde{\mathcal{P}}'(D)$ lie between $\mathcal{P}(D)$ and $\mathcal{P}'(D)$. We consider the largest mass transfer $\delta\epsilon$ until one of the two following events occur.

1. Equal-score event: The cell of a has the same score as the adjacent cell lower in the sorted order or the cell of b has the same score as the adjacent cell higher in the sorted order.

2. Threshold event: The score of a merged cell containing a or b becomes exactly 0.5.

Bounding the accuracy change for $\delta\epsilon$: Note that by the choice of $\delta\epsilon$, during the transfer $\delta\epsilon$, all the cells remain in the same order in both groups; furthermore, all masses and scores of all cells other than the ones containing a or b remain the same during the transfer. By part 2 of Claim 2,

$$\begin{aligned}\delta\text{Unf}_{\text{DP}} &= \left|\text{Unf}_{\text{DP}}(f^{\text{DP}}_{\tilde{\mathcal{P}}}, \tilde{\mathcal{P}}) - \text{Unf}_{\text{DP}}(f^{\text{DP}}_{\tilde{\mathcal{P}}}, \tilde{\mathcal{P}}')\right| \\ &\leq \delta\epsilon\left(\frac{1}{\min(\tilde{\mathcal{P}}(A), \tilde{\mathcal{P}}'(A))} + \frac{1}{\min(\tilde{\mathcal{P}}(D), \tilde{\mathcal{P}}'(D))}\right) \\ &\leq \delta\epsilon\left(\frac{1}{\min(\mathcal{P}(A), \mathcal{P}'(A))} + \frac{1}{\min(\mathcal{P}(D), \mathcal{P}'(D))}\right)\end{aligned}$$

Since $\text{Unf}_{\text{DP}}(f^{\text{DP}}_{\tilde{\mathcal{P}}}, \tilde{\mathcal{P}}) = 0$, we know that $\delta\text{Unf}_{\text{DP}} = \text{Unf}_{\text{DP}}(f^{\text{DP}}_{\tilde{\mathcal{P}}}, \tilde{\mathcal{P}}') = \left|\text{r}_A(f^{\text{DP}}_{\tilde{\mathcal{P}}}, \tilde{\mathcal{P}}) - \text{r}_D(f^{\text{DP}}_{\tilde{\mathcal{P}}}, \tilde{\mathcal{P}}')\right|$. Consider the cell q that is split in the middle by the threshold corresponding to $f^{\text{DP}}_{\tilde{\mathcal{P}}}$ (for now, assume $q \in D$). Since neither the equal-score event nor the 0.5-score event occur, we see that after the transition, the boundary of $f^{\text{DP}}_{\mathcal{P}}$ intersecting q is $\delta\text{Unf}_{\text{DP}}$ away from the boundary in group A. To modify $f^{\text{DP}}_{\tilde{\mathcal{P}}} \to f^{\text{DP}}_{\tilde{\mathcal{P}}'}$, we therefore need to move to move the boundary at q by $\delta\text{Unf}_{\text{DP}}$ so that the boundaries in both groups align and DP is satisfied (the classifier remains the same apart from its action on q). The change in function value on element q, which we denote by $|\Delta f(q)|$, is bounded by $\delta\text{Unf}_{\text{DP}}\frac{\tilde{\mathcal{P}}(D)}{\tilde{\mathcal{P}}(q)}$, after scaling (since $\tilde{\mathcal{P}}(D)\delta\text{Unf}_{\text{DP}} = |\Delta f(q)|\,\tilde{\mathcal{P}}(q)$). At the end of the $\delta\epsilon$ mass transfer, by Lemma 2, the change in accuracy of the optimal fair classifier is given by

$$\begin{aligned}&\left|\text{Acc}(f^{\text{DP}}_{\tilde{\mathcal{P}}}, \tilde{\mathcal{P}}) - \text{Acc}(f^{\text{DP}}_{\tilde{\mathcal{P}}'}, \tilde{\mathcal{P}}')\right| \leq \left|\tilde{\mathcal{P}}(q)(2\mathcal{S}(q) - 1)\Delta f(q)\right| + \delta\epsilon \\ &\leq \delta\epsilon\left(1 + \frac{\tilde{\mathcal{P}}(D)\left|(2\mathcal{S}(q) - 1)\right|}{\min(\tilde{\mathcal{P}}(A), \tilde{\mathcal{P}}'(A))} + \frac{\tilde{\mathcal{P}}(D)\left|(2\mathcal{S}(q) - 1)\right|}{\min(\tilde{\mathcal{P}}(D), \tilde{\mathcal{P}}'(D))}\right) \\ &\leq \delta\epsilon\left(1 + \frac{\max\left(\mathcal{P}(D), \mathcal{P}'(D)\right)}{\min(\mathcal{P}(A), \mathcal{P}'(A))} + \frac{\max\left(\mathcal{P}(D), \mathcal{P}'(D)\right)}{\min(\mathcal{P}(D), \mathcal{P}'(D))}\right),\end{aligned}$$

where the last inequality follows from the facts that $|(2\mathcal{S}(q) - 1)| \leq 1$, $\tilde{\mathcal{P}}(A)$ and $\tilde{\mathcal{P}}'(A)$ both lie between $\mathcal{P}(A)$ and $\mathcal{P}'(A)$ and $\tilde{\mathcal{P}}(D)$ and $\tilde{\mathcal{P}}'(D)$ both lie between $\mathcal{P}(D)$ and $\mathcal{P}'(D)$.

In Appendix, we derive a better upper bound on $|(2\mathcal{S}(q) - 1)|$ and derive the following:

$$\begin{aligned}&\left|\text{Acc}(f^{\text{DP}}_{\tilde{\mathcal{P}}}, \tilde{\mathcal{P}}) - \text{Acc}(f^{\text{DP}}_{\tilde{\mathcal{P}}'}, \tilde{\mathcal{P}}')\right| \\ &\leq \delta\epsilon\left(1 + \frac{\max\left(\mathcal{P}(A), \mathcal{P}'(A)\right)}{\min(\mathcal{P}(A), \mathcal{P}'(A))} + \frac{\max\left(\mathcal{P}(A), \mathcal{P}'(A)\right)}{\min(\mathcal{P}(D), \mathcal{P}'(D))}\right).\end{aligned}$$

Putting the two upper bounds together yields the following:

$$\begin{aligned}&\left|\text{Acc}(f^{\text{DP}}_{\tilde{\mathcal{P}}}, \tilde{\mathcal{P}}) - \text{Acc}(f^{\text{DP}}_{\tilde{\mathcal{P}}'}, \tilde{\mathcal{P}}')\right| \\ &\leq \delta\epsilon\left(1 + \frac{\max\left(\mathcal{P}(A), \mathcal{P}'(A)\right)}{\min(\mathcal{P}(A), \mathcal{P}'(A))} + \frac{\max\left(\mathcal{P}(D), \mathcal{P}'(D)\right)}{\min(\mathcal{P}(D), \mathcal{P}'(D))}\right).\end{aligned}$$

Handling the equal-score and threshold events: We now describe how to handle the two events.

1. Equal-score event: If the cell of a has the same score as the adjacent cell lower in the sorted order, then we swap the two cells so that the cell of a is lower in the order. Similarly, if the cell of b has the same score as the adjacent cell higher in the order, then we swap the two cells so that the cell of b is higher in the order. We update the classifier f and note that this change has no impact on the accuracy of f.

2. Threshold event: The score of a merged cell containing a or b becomes exactly 0.5. We include the merged cell in the classifier f, again without changing accuracy.

Thus, in a sense, between any two occurrences of these events, the change in accuracy is bounded by an amount proportional to the mass transfer; when we reach these occurrences, the mass transfer is paused, the BOC changes without any change in accuracy. Furthermore, at every occurrence of the event, one of these three events happen: the cell containing a moves down in the order, the cell containing b moves up in the order, or an additional merged cell is placed above the threshold. Since the number of times these events can occur is upper bounded by the number of cells in the two groups, this process is finite. Therefore, adding over all the $\delta\epsilon$ mass transfers, we obtain the desired upper bound on the change in accuracy between the BOC's for $\mathcal{P}$ and $\mathcal{P}'$.

$$\epsilon\left(1+\frac{\max\left(\mathcal{P}(A),\mathcal{P}'(A)\right)}{\min(\mathcal{P}(A),\mathcal{P}'(A))}+\frac{\max\left(\mathcal{P}(D),\mathcal{P}'(D)\right)}{\min(\mathcal{P}(D),\mathcal{P}'(D))}\right).$$

From elementary to arbitrary: Consider a general transition of distance ϵ from $\mathcal{P}$ to $\mathcal{P}'$. We invoke Lemma 1 to obtain intermediate distributions $\{\mathcal{P}_i\}$ with $TV(\mathcal{P}_{i-1},\mathcal{P}_i)=\epsilon_i$ satisfying the decomposability and monotonocity properties. We apply the above proof for each elementary transition $\mathcal{P}_{i-1}\to\mathcal{P}_i$ of mass ϵ_i. For accuracy, we derive

$$\begin{aligned}
&\left|\mathrm{Acc}(f^{\mathrm{DP}}_{\mathcal{P}},\mathcal{P})-\mathrm{Acc}(f^{\mathrm{DP}}_{\mathcal{P}'},\mathcal{P}')\right|\\
&\leq\sum_i\left|\mathrm{Acc}(f^{\mathrm{DP}}_{\mathcal{P}_{i-1}},\mathcal{P}_{i-1})-\mathrm{Acc}(f^{\mathrm{DP}}\mathcal{P}_i,\mathcal{P}_i)\right|\leq\sum_i\epsilon_i\cdot\\
&\left(1+\frac{\max(\mathcal{P}_{i-1}(A),\mathcal{P}_i(A))}{\min(\mathcal{P}_{i-1}(A),\mathcal{P}_i(A))}+\frac{\max\{\mathcal{P}_{i-1}(D),\mathcal{P}_i(D))\}}{\min(\mathcal{P}_{i-1}(D),\mathcal{P}_i(D))}\right)\\
&\leq\sum_i\epsilon_i\left(1+\frac{\max(\mathcal{P}(A),\mathcal{P}'(A))}{\min(\mathcal{P}(A),\mathcal{P}'(A))}+\frac{\max(\mathcal{P}(D),\mathcal{P}'(D))}{\min(\mathcal{P}(D),\mathcal{P}'(D))}\right)\\
&=\epsilon\left(1+\frac{\max(\mathcal{P}(A),\mathcal{P}'(A))}{\min(\mathcal{P}(A),\mathcal{P}'(A))}+\frac{\max(\mathcal{P}(D),\mathcal{P}'(D))}{\min(\mathcal{P}(D),\mathcal{P}'(D))}\right),
\end{aligned}$$

where the third inequality follows from monotonocity and the last equation follows from decomposability. This completes the proof of the theorem. □

We now state the following corollary, which follows from Claim 2 and Theorem 1. It roughly states that given 2 closeby distributions $\mathcal{P},\mathcal{P}'$, the accuracy of the respective DP-Fair BOC's is similar on $\mathcal{P}$. Such a property is useful when we want a guarantee that intuitively says that if we train on the corrupted distribution $\mathcal{P}'$, we get a similar outcome to what we would have gotten had we trained on the true distribution $\mathcal{P}$.

Corollary 1 (Stability of DP-Fair BOC)**.** *Given distributions $\mathcal{P},\mathcal{P}'$ with $TV(\mathcal{P},\mathcal{P}')=\epsilon$, $\left|Acc(f^{DP}_{\mathcal{P}},\mathcal{P})-Acc(f^{DP}_{\mathcal{P}'},\mathcal{P})\right|$ is at most*

$$\epsilon\left(2+\frac{\max\left(\mathcal{P}(A),\mathcal{P}'(A)\right)}{\min(\mathcal{P}(A),\mathcal{P}'(A))}+\frac{\max\left(\mathcal{P}(D),\mathcal{P}'(D)\right)}{\min(\mathcal{P}(D),\mathcal{P}'(D))}\right).$$

4. Comparison with Previous Work

Perhaps the most closely related work to ours is that of Blum et al. (2024). They aim to avoid the non-robustness phenomena highlighted in Konstantinov & Lampert (2022), in the following manner. Given any $\mathcal{H}$, and distributions $\mathcal{P},\mathcal{P}'$ over $\mathcal{X}\times\{0,1\}\times\{A,D\}$ with $TV(\mathcal{P},\mathcal{P}')=\epsilon$, they construct a randomized closure (robustification, see Section 1.2) of $\mathcal{H}$ called $PQ(\mathcal{H})$. Denote by f,f' the optimal classifiers (subject to DP constraints) on $\mathcal{P},\mathcal{P}'$ restricted to $\mathcal{H}$, $PQ(\mathcal{H})$ respectively. This satisfies a one-directional Lipschitzness constraint, i.e., $\mathrm{Acc}(f',\mathcal{P}')\geq\mathrm{Acc}(f,\mathcal{P})-O(\epsilon)$. They also show analogous results for EO and PE.

Our setup has some key differences. We do not consider any arbitrary $\mathcal{H}$, but the BOC setting which includes all deterministic classifiers (and the 1-skeleton of their convex closure, see Section 1.2). In fact, we can argue that for any $\mathcal{H}$, our proposed robustification (1-skeleton of the convex closure of $\mathcal{H}$) is more accurate, while having lower cardinality. The formal proof is out of the scope of our paper, but as an example we can verify that this holds for the case when $\mathcal{H}$ is the set of all deterministic classifiers. More crucially, our robustness guarantee is stronger, as their Lipschitzness guarantee is only one-directional. In addition, in most cases, their output hypothesis incorporates a lot of randomness, outputting a randomized decision on all points.

5. Conclusion

Some directions for further work include extending our results for binary classification to multi-class classification and regression. Another direction could be to look at relaxed or approximate versions of the fairness notions we considered. One could even look at other popular notions of fairness, or satisfying multiple fairness notions simultaneously. It would also be valuable to experimentally validate our theoretical claims. In addition, note that our results hold for adversarial noise, but it might be possible to strengthen the bounds if the noise came from a particular distribution.

Impact Statement

This paper presents work towards advancing theoretical understanding of fair machine learning under adversarial or malicious distribution shift. Our results provide normative principles in the debate on pros and cons of randomized algorithmic decisions with societal consequences. Our work underlines the need to compare and contrast different ways to rectify biases in algorithmic decision-making with societal consequences.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. A reductions approach to fair classification. In *International Conference on Machine Learning*, pp. 60–69. PMLR, 2018.

Agarwal, S. and Deshpande, A. On the power of randomization in fair classification and representation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 1542–1551, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533209. URL https://doi.org/10.1145/3531146.3533209.

Balcan, M.-F. and Haghtalab, N. Noise in classification. In *Beyond the Worst-Case Analysis of Algorithms*, 2020. URL https://arxiv.org/abs/2010.05080.

Barocas, S., Hardt, M., and Narayanan, A. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

Blum, A., Okoroafor, P., Saha, A., and Stangl, K. M. On the vulnerability of fairness constrained learning to malicious noise. In Dasgupta, S., Mandt, S., and Li, Y. (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4096–4104. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/blum24a.html.

Celis, L. E., Huang, L., Keswani, V., and Vishnoi, N. K. Classification with fairness constraints: A meta-algorithm with provable guarantees, 2020.

Celis, L. E., Huang, L., Keswani, V., and Vishnoi, N. K. Fair classification with noisy protected attributes: A framework with provable guarantees. In *International Conference on Machine Learning*, pp. 1349–1361. PMLR, 2021.

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. Leveraging labeled and unlabeled data for consistent fair binary classification. *Advances in Neural Information Processing Systems*, 32, 2019.

Cooper, A. F., Lee, K., Choksi, M. Z., Barocas, S., De Sa, C., Grimmelmann, J., Kleinberg, J., Sen, S., and Zhang, B. Arbitrariness and social prediction: the confounding role of variance in fair classification. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i20.30203. URL https://doi.org/10.1609/aaai.v38i20.30203.

Creel, K. and Hellman, D. The algorithmic leviathan: Arbitrariness, fairness, and opportunity in algorithmic decision making systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 816, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445942. URL https://doi.org/10.1145/3442188.3445942.

Davani, A. M., Atari, M., Kennedy, B., and Dehghani, M. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319, 03 2023. ISSN 2307-387X. doi: 10.1162/tacl_a_00550. URL https://doi.org/10.1162/tacl_a_00550.

Davidson, T. R. Hate speech detection and bias in supervised text classification. In *The Oxford Handbook of the Sociology of Machine Learning*. Oxford University Press, 2023. ISBN 9780197653609. doi: 10.1093/oxfordhb/9780197653609.013.16. URL https://doi.org/10.1093/oxfordhb/9780197653609.013.16.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science (ITCS) Conference*, pp. 214–226. Association for Computing Machinery, 2012. URL https://doi.org/10.1145/2090236.2090255.

Hardt, M., Price, E., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

Harris, C., Halevy, M., Howard, A., Bruckman, A., and Yang, D. Exploring the role of grammar and word choice in bias toward african american english (aae) in hate speech classification. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT'22, pp. 789–798, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533144. URL https://doi.org/10.1145/3531146.3533144.

Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., and Kamar, E. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3309–3326, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.234. URL https://aclanthology.org/2022.acl-long.234/.

Kalai, G. *Some Aspects of the Combinatorial Theory of Convex Polytopes*, pp. 205–229. Springer Netherlands, Dordrecht, 1994. ISBN 978-94-011-0924-6. doi: 10.1007/978-94-011-0924-6_9. URL https://doi.org/10.1007/978-94-011-0924-6_9.

Kamiran, F. and Calders, T. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

Kearns, M. and Li, M. Learning in the presence of malicious errors. In *Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing*, pp. 267–280, 1988.

Konstantinov, N. and Lampert, C. H. On the impossibility of fairness-aware learning from corrupted data. In Schrouff, J., Dieng, A., Rateike, M., Kwegyir-Aggrey, K., and Farnadi, G. (eds.), *Proceedings of The Algorithmic Fairness through the Lens of Causality and Robustness*, volume 171 of *Proceedings of Machine Learning Research*, pp. 59–83. PMLR, 13 Dec 2022. URL https://proceedings.mlr.press/v171/konstantinov22a.html.

Menon, A. K. and Williamson, R. C. The cost of fairness in binary classification. In *Conference on Fairness, accountability and transparency*, pp. 107–118. PMLR, 2018a.

Menon, A. K. and Williamson, R. C. The cost of fairness in binary classification. In Friedler, S. A. and Wilson, C. (eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 107–118. PMLR, 23–24 Feb 2018b. URL https://proceedings.mlr.press/v81/menon18a.html.

Rosenblatt, L. and Witter, R. T. Fairlyuncertain: A comprehensive benchmark of uncertainty in algorithmic fairness, 2024. URL https://arxiv.org/abs/2410.02005.

Yang, Y.-Y., Rashtchian, C., Zhang, H., Salakhutdinov, R., and Chaudhuri, K. A closer look at accuracy vs. robustness. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Zeng, X., Dobriban, E., and Cheng, G. Fair bayes-optimal classifiers under predictive parity. *Advances in Neural Information Processing Systems*, 35:27692–27705, 2022.

A. Missing Results from Section 3

A.1. NP-Completeness of Deterministic Fair Bayes Optimal Classifiers

Claim 3 (NP-Completeness of Deterministic DP-Fair BOC)**.** *The problem of computing a deterministic DP-Fair Bayes Optimal Classifier is NP-complete.*

Proof. We formalize the deterministic DP-Fair BOC decision problem as follows: Given a probability distribution $\mathcal{P}$ and a score function $\mathcal{S}$ over a domain $\mathcal{X} \times \{A/D\}$ and an accuracy α, determine whether there exists a deterministic Fair Bayes-Optimal classifier with accuracy at least α.

It is easy to see that the above problem is is in NP since one can guess the 0-1 classification for each item in the domain and check in polynomial time that the resulting classifier is fair and satisfies the accuracy bound by verifying two linear inequalities. We now show that the problem is NP-hard via a polynomial-time reduction from the NP-complete Partition problem.

Partition: Given a set S of n positive integers $a_1, a_2, \ldots, a_n$ summing to $2s$, determine whether there exists a subset of S that sums to s.

The reduction from Partition to the deterministic DP-Fair BOC problem is as follows. Given an instance I of Partition, We create an instance I' of the deterministic fair BOC problem. Instance I' has $n+2$ items—(x_1, A) with mass $1/4$ and score 1, item (x_2, A) with mass $1/4$ and score 0, and then n items (y_i, D) with mass $a_i/(4s)$ and score 0.5—and ask whether there is a deterministic DP-Fair BOC with accuracy $\alpha = 3/4$. It is clear that there are at most 3 kinds of deterministic DP-fair classifiers: (i) the all-0 classifier that classifies all items as 0, (ii) the all-1 classifier that classifies all items as 1, and (iii) if and only f I is a yes-instance with S partitioned into S_1 and S_2 of equal sums, then the classifier that accepts exactly one of (x_1, A) or (x_2, A) and accepts all items in (y_i, D) with $a_i \in S_1$ and rejecting all items in (y_i, D) with $a_i \in S_2$. The first two classifiers have accuracy $1/4$ while the third, if it exists, has accuracy $3/4$ if (x_1, A) is accepted and less than $3/4$ otherwise. Thus, there exists a deterministic Fair BOC for instance I' with $3/4$ accuracy if and only if I is a yes-instance for the Partition problem. Clearly, the reduction is of time polynomial in the size of the deterministic DP-fair BOC instance, thus establishing its NP-completeness.

Since determining the existence of a deterministic DP-fair classifier with accuracy at least $3/4$ is NP-complete it follows immediately that finding a deterministic DP-fair BOC is also NP-complete. □

A.2. Maximal Accuracy Gain for Randomized Classifiers

Consider the example in Claim 1, and consider the following randomized classifier f', where

$$f'(x_1, A) = f'(x_1, D) = 1, f'(x_2, A) = 4\epsilon, f'(x_2, D) = 0.$$

It is easy to see that f' satisfies DP on $\mathcal{P}'$, and $\mathrm{Acc}(f') = 1 - \epsilon$, leading to a difference in accuracy bounded by ϵ. Hence, the DP-fair randomized BOC improves over the accuracy of its deterministic counterpart by $0.5 - 2\epsilon$, where $\epsilon > 0$ can be made arbitrarily small (so the gain in accuracy approaches 0.5). In the following, we argue that this example is tight, i.e., we cannot hope to achieve an improvement over 0.5.

Claim 4 (Bound in Accuracy Gain for Randomized classifiers)**.** *Given any distribution $\mathcal{P}$, the difference in accuracy of the Randomized and Deterministic DP-Fair BOC's on $\mathcal{P}$ is strictly lesser than* 0.5.

Proof. Note that the constant classifiers f_0, f_1 always satisfy DP, and $\mathrm{Acc}(f_0) = 1 - \mathrm{Acc}(f_1)$. Hence, the minimum accuracy of the optimal DP-fair deterministic classifier is 0.5. The maximum accuracy of its randomized counterpart is bounded by 1, hence bounding the difference in accuracy by 0.5. It suffices to show that these 2 events cannot occur simultaneously. Note that if some classifier has perfect accuracy, then all cells in the domain have score of either 0 or 1. In particular, this also holds if the optimal DP-fair randomized classifier has accuracy 1. However, observe that if we randomize over any cell with score of 0(1), we are accepting (rejecting) a part of it, leading to a loss in accuracy. This implies that any classifier with accuracy 1 has to be deterministic, concluding our proof. □

A.3. Completion of the Robustness Analysis for Demographic Parity

In this section, we present the argument that was deferred in the proof of Theorem 1. This argument concerns a better upper bound on $|2\mathcal{S}(q) - 1|$ than the vacuous bound of 1, where q is the element that is split by the threshold corresponding to the classifier f. Notice that since by assumption, f splits q in the middle, we know that there is a portion of q that is rejected. Hence, the weighted score of a merged cell involving q (say $\mathcal{C}_q$) has score below the threshold of 0.5. Let C_q contain some element t from group A. We are able to bound the score of $\mathcal{S}(q)$ by the following chain of inequalities.

$$\begin{aligned}\mathcal{S}(\mathcal{C}_q) \leq 0.5 \quad &\implies \quad \mathcal{S}(q)\mathcal{P}(D) + \mathcal{S}(t)\mathcal{P}(A) \leq 0.5\,(\mathcal{P}(D) + \mathcal{P}(A)) \\ &\implies \quad \mathcal{S}(q)\mathcal{P}(D) \leq 0.5\,(\mathcal{P}(D) + \mathcal{P}(A)) \\ &\implies \quad 2\mathcal{S}(q) - 1 \leq \frac{\mathcal{P}(A)}{\mathcal{P}(D)} \end{aligned} \tag{4}$$

Since f splits q in the middle, there is also a portion of q that is accepted. Hence, the weighted score of a merged cell involving q (say $\mathcal{C}_q$) has score above the threshold of 0.5. Let C_q contain some element t from group A. We are able to bound the score of $\mathcal{S}(q)$ by the following chain of inequalities.

$$\begin{aligned}\mathcal{S}(\mathcal{C}_q) \geq 0.5 \quad &\implies \quad \mathcal{S}(q)\mathcal{P}(D) + \mathcal{S}(t)\mathcal{P}(A) \geq 0.5\,(\mathcal{P}(D) + \mathcal{P}(A)) \\ &\implies \quad \mathcal{S}(q)\mathcal{P}(D) + \mathcal{P}(A) \geq 0.5\,(\mathcal{P}(D) + \mathcal{P}(A)) \\ &\implies \quad \mathcal{S}(q)\mathcal{P}(D) \geq 0.5\,(\mathcal{P}(D) - \mathcal{P}(A)) \\ &\implies \quad 2\mathcal{S}(q) - 1 \geq -\frac{\mathcal{P}(A)}{\mathcal{P}(D)} \end{aligned} \tag{5}$$

Combining Equations 4 and 5, we get

$$|2\mathcal{S}(q) - 1| \leq \frac{\mathcal{P}(A)}{\mathcal{P}(D)} \tag{6}$$

Using Equation 6, we get that

$$\begin{aligned}&\left|\mathrm{Acc}(f^{\mathrm{DP}}_{\tilde{\mathcal{P}}}, \tilde{\mathcal{P}}) - \mathrm{Acc}(f^{\mathrm{DP}}_{\tilde{\mathcal{P}}'}, \tilde{\mathcal{P}}')\right| \\ &\leq \delta\epsilon\left(\frac{1}{\min(\tilde{\mathcal{P}}(A), \tilde{\mathcal{P}}'(A))} + \frac{1}{\min(\tilde{\mathcal{P}}(D), \tilde{\mathcal{P}}'(D))}\right)\tilde{\mathcal{P}}(D)\frac{\mathcal{P}(A)}{\mathcal{P}(D)} + \delta\epsilon \\ &= \delta\epsilon\left(\frac{1}{\min(\tilde{\mathcal{P}}(A), \tilde{\mathcal{P}}'(A))} + \frac{1}{\min(\tilde{\mathcal{P}}(D), \tilde{\mathcal{P}}'(D))}\right)\tilde{\mathcal{P}}(A) + \delta\epsilon \\ &\leq \delta\epsilon\left(1 + \frac{\max\left(\mathcal{P}(A), \mathcal{P}'(A)\right)}{\min(\mathcal{P}(A), \mathcal{P}'(A))} + \frac{\max\left(\mathcal{P}(A), \mathcal{P}'(A)\right)}{\min(\mathcal{P}(D), \mathcal{P}'(D))}\right)\end{aligned} \tag{7}$$

The last equation follows by monotonicity. This completes the missing argument in the proof of Theorem 1.

A.4. Proof of Lemma 2

Proof. The contribution to accuracy of an element q is given by

$$\begin{aligned}&\mathcal{P}(q)\mathcal{S}(q)f(q) + \mathcal{P}(q)(1 - \mathcal{S}(q))(1 - f(q)) \\ &= 2\mathcal{P}(q)\mathcal{S}(q)f(q) + \mathcal{P}(q) - \mathcal{P}(q)\mathcal{S}(q) - \mathcal{P}(q)f(q)\end{aligned}$$

If $\mathcal{P}$ changes by ϵ and $f(q)$ changes by $\Delta f(q)$ (and remains constant elsewhere), then we can split the process into two parts: (i) $f(q)$ changes by $\Delta f(q)$ (and remains constant elsewhere) while $\mathcal{P}$ remains constant, and (ii) $\mathcal{P}$ changes by ϵ while f remains constant. We consider each of the parts.

If $\mathcal{P}$ remains fixed, and $f(q)$ changes by $\Delta f(q)$ to give $f'(q)$, then change in accuracy on q (and also overall accuracy) is given by

$$|2\mathcal{P}(q)\mathcal{S}(q)\Delta f(q) - \mathcal{P}(q)\Delta f(q)| = |\mathcal{P}(q)(2\mathcal{S}(q) - 1)\Delta f(q)|$$

If $\mathcal{P}$ changes by ϵ, and f' remains constant, then by Claim 2 the change in accuracy is bounded by ϵ. Thus, the total change in accuracy is bounded as follows.

$$|\text{Acc}(f, \mathcal{P}) - \text{Acc}(f', \mathcal{P}')| \leq |\mathcal{P}(q)(2\mathcal{S}(q) - 1)\Delta f(q)| + \epsilon.$$

□

B. Equal Opportunity and Predictive Equality

In the main body of the paper, we presented results for Demographic Parity. Our results also extend to the popular fairness notions of Equal Opportunity and Predictive Equality (Hardt et al., 2016; Barocas et al., 2019). However, we deferred these results to the appendix due to space constraints. Before presenting them, we first define these fairness notions.

Definition 7 (Equal TPR, or Equal Opportunity)**.** Denote the true positive rate of f on group z by

$$\text{TPR}_z(f, \mathcal{P}) = \Pr[f(X, Z) = 1 \mid Y = 1, Z = z].$$

f satisfies Equal Opportunity if the true positive rates are equal for both groups, i.e. $\text{TPR}_A(f, \mathcal{P}) = \text{TPR}_D(f, \mathcal{P})$. We quantify the unfairness of f as the difference in false negative rates across groups , i.e.,

$$\text{Unf}_{\text{EO}}(f, \mathcal{P}) = |\text{TPR}_A(f, \mathcal{P}) - \text{TPR}_D(f, \mathcal{P})|.$$

Definition 8 (Equal FPR, or Predictive Equality)**.** Denote the false positive rate of of f on group z by

$$\text{FPR}_z(f, \mathcal{P}) = \Pr[f(X, Z) = 1 \mid Y = 0, Z = z].$$

f satisfies Predictive Equality if the false positive rates are equal for both groups, i.e. $\text{FPR}_A(f, \mathcal{P}) = \text{FPR}_D(f, \mathcal{P})$. We quantify the unfairness of f as the difference in false positive rates across groups , i.e.,

$$\text{Unf}_{\text{PE}}(f, \mathcal{P}) = |\text{FPR}_A(f, \mathcal{P}) - \text{FPR}_D(f, \mathcal{P})|.$$

Remark. Classifiers satisfying these notions of fairness will be referred to as EO-fair, and PE-fair respectively. Our results for hold both fairness notions. The results for PE follow using the same proof techniques as that of EO (since we can just reverse the roles of the labels 0 and 1 in EO to get results for PE), we only present the proof of the results for EO. In addition, previous work has also considered equal False Negative rate (FNR) and equal True Negative rate (TNR) as notions of fairness. We observe that obtaining equal TPR is equivalent to obtaining equal FNR, and obtaining equal TNR is equivalent to obtaining equal FPR, and hence results for these notions of fairness also follow.

B.1. Fair Bayes Optimal Classifier

When discussing the DP-Fair BOC, we considered mass-threshold classifiers $\tilde{\mathcal{T}}_t$, that select $\mathcal{C}_z(t)$, and reject $\mathcal{C}_z - \mathcal{C}_z(t)$, for both $z = A$, and $z = D$. $\tilde{\mathcal{T}}_t$ applies the same threshold t to both groups A and D. In this section, we consider groupwise mass-threshold classifiers $\tilde{\mathcal{T}}_{t_A, t_D}$ that apply different thresholds t_A and t_D to groups A and D respectively.

Denote the True Positive rate of a classifier f restricted to a cell $\mathcal{C}$ by $\text{TPR}(f(\mathcal{C}))$. Given $r \in (0, 1]$, there is a unique classifier $\tilde{\mathcal{T}}_{t_A, t_D}$, such that $\text{TPR}(\tilde{\mathcal{T}}_{t_A, t_D}(\mathcal{C}_A)) = \text{TPR}(\tilde{\mathcal{T}}_{t_A, t_D}(\mathcal{C}_D)) = r$. Denote this classifier by f_r. Given $r = 0$, $\tilde{\mathcal{T}}_{t_A, t_D}$ need not be unique as there could exist cells with score 1. In that case, we define f_0 to be the unique groupwise mass-threshold classifier accepting exactly the cells with score 1. Denote the groupwise thresholds of f_r by r^A and r^D respectively, i.e., $f_r = \tilde{\mathcal{T}}_{r^A, r^D}$. We now introduce some terminology, before detailing the EO-Fair BOC as characterized in Agarwal & Deshpande (2022).

Definition 9 (TP-Boundaries)**.** Recall the set of score-boundaries $\mathcal{I}$. We then define the set of TP-boundaries $\mathcal{I}_{TP}$ as

$$\mathcal{I}_{TP} = \{r \mid r^A \in \mathcal{I}, \text{ or } r^D \in \mathcal{I}\}.$$

$\mathcal{I}_{TP}$ essentially consists of all the true positive rates r, such that, the corresponding groupwise threshold classifier $f_r = \tilde{\mathcal{T}}_{r^A, r^D}$ has a threshold at a point in the set of score boundaries $\mathcal{I}$.

As with DP, we define the notion of a merged cell, but notice that it differs from the notion of merged cell in the case of DP.

Definition 10 (Merged cell (EO))**.** Consider $r_i \in \mathcal{I}_{TP}$, and define a merged cell $\mathcal{C}_i$, where

$$\mathcal{C}_i = \mathcal{A}(f_{r_i}) - \mathcal{A}(f_{r_{i-}}),$$

where r_{i-} denotes the element in $\mathcal{I}_{TP}$ preceding r_i.

Characterization Given a distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Z} \times \mathcal{Y}$, the EO-Fair BOC $f_{\mathcal{P}}^{\text{EO}}$ is given by the mass-threshold classifier is given by the group wise mass-threshold classifier $f_{r'}$, where $r' = r_i \in \mathcal{I}$ is the unique i such that $\mathcal{S}(C_i) \geq 0.5$, and $\mathcal{S}(C_{i+}) < 0.5$, where r_{i+} denotes the element in $\mathcal{I}_{TP}$ after r_i.

B.2. Robustness to Adversarial Distribution Shift

We study the robustness of the EO-Fair BOC to adversarial distribution shift. We show that given two similar distributions $\mathcal{P}, \mathcal{P}'$, the accuracy of the EO-Fair BOC on the respective distributions is similar (satisfies local Lipschitzness). Before proving the main result (Theorem 2), we prove Claim 5, which analyzes the change in unfairness, with respect to EO, of a fixed classifier due to a distribution shift. Such a property is useful when we want a guarantee that if we train a classifier on the corrupted distribution $\mathcal{P}'$, the performance of the classifier on the actual distribution $\mathcal{P}$ will be similar to that on $\mathcal{P}'$.

Claim 5 (EO Shift for a Fixed Hypothesis)**.** *Given distributions $\mathcal{P}, \mathcal{P}'$, such that $TV(\mathcal{P}, \mathcal{P}') \leq \epsilon$, and any hypothesis f, it holds that*

$$|Unf_{EO}(f, \mathcal{P}) - Unf_{EO}(f, \mathcal{P}')| \leq \epsilon \left(\frac{1}{\min(\mathcal{P}(A,1), \mathcal{P}'(A,1))} + \frac{1}{\min(\mathcal{P}(D,1), \mathcal{P}'(D,1))} \right).$$

Proof. As in the proof of Claim 2, it follows from Lemma 1 and transitivity that it is enough to prove the statement of the claim for elementary transitions. We consider a transition $a \to b$ of mass ϵ. We first derive

$$\begin{aligned}
|\text{Unf}_{\text{EO}}(f, \mathcal{P}) - \text{Unf}_{\text{EO}}(f, \mathcal{P}')| &= ||\text{TPR}_A(f, \mathcal{P}) - \text{TPR}_D(f, \mathcal{P})| - |\text{TPR}_A(f, \mathcal{P}') - \text{TPR}_D(f, \mathcal{P}')|| \\
&\leq |(\text{TPR}_A(f, \mathcal{P}) - \text{TPR}_D(f, \mathcal{P})) - (\text{TPR}_A(f, \mathcal{P}') - \text{TPR}_D(f, \mathcal{P}'))| \\
&\qquad \text{(Triangle inequality)} \\
&= |(\text{TPR}_A(f, \mathcal{P}) - \text{TPR}_A(f, \mathcal{P}')) + (\text{TPR}_D(f, \mathcal{P}') - \text{TPR}_D(f, \mathcal{P}))| \\
&\leq |\text{TPR}_A(f, \mathcal{P}) - \text{TPR}_A(f, \mathcal{P}')| + |\text{TPR}_D(f, \mathcal{P}') - \text{TPR}_D(f, \mathcal{P})| \qquad (8)
\end{aligned}$$

This breaks up the change in unfairness into two terms, which correspond to the difference in true positive rates of f for $\mathcal{P}$ and $\mathcal{P}'$ on A, D respectively (denoted by $\Delta\text{TPR}_A, \Delta\text{TPR}_D$). Divide the domain into 8 parts based on the group membership and whether a point falls in TP, FP, TN, or FN according to f. Denote the probability mass of elements in group z in category E under f by $\mathcal{P}(E_z)$. We know that $\mathcal{P}(A) = \mathcal{P}(A,1) + \mathcal{P}(A,0) = (\mathcal{P}(\text{TP}_A) + \mathcal{P}(\text{FN}_A)) + (\mathcal{P}(\text{TN}_A) + \mathcal{P}(\text{FP}_A))$.

We proceed to bound ΔTPR_A, and an identical argument can be used to bound ΔTPR_D. If a, b lie in $\mathcal{P}(A,1)$, it remains unchanged, and it is easy to see that the maximum value of ΔTPR_A is $\frac{\epsilon}{\mathcal{P}(A,1)}$, when $\mathcal{P}'(\text{TP}_A) = \mathcal{P}(\text{TP}_A) \pm \epsilon$. In case $a \in (A,1)$, and $b \notin (A,1)$, then $\mathcal{P}'(A,1) = \mathcal{P}(A,1) - \epsilon$. We know that $\mathcal{P}'(A,1) = \mathcal{P}'(\text{TP}_A) + \mathcal{P}'(\text{FN}_A)$. We first consider the case where $\mathcal{P}'(\text{TP}_A) = \mathcal{P}(\text{TP}_A) - \epsilon$.

$$\begin{aligned}
|\text{TPR}_A(f, \mathcal{P}) - \text{TPR}_A(f, \mathcal{P}')| &= \left| \frac{\mathcal{P}(\text{TP}_A)}{\mathcal{P}(A,1)} - \frac{\mathcal{P}(\text{TP}_A) - \epsilon}{\mathcal{P}(A,1) - \epsilon} \right| \\
&= \left| \frac{\mathcal{P}(\text{TP}_A)\mathcal{P}(A,1) - \mathcal{P}(\text{TP}_A)\epsilon - \mathcal{P}(\text{TP}_A)\mathcal{P}(A,1) + \mathcal{P}(A,1)\epsilon}{\mathcal{P}(A,1)\left(\mathcal{P}(A,1) - \epsilon\right)} \right| \\
&= \left| \frac{\mathcal{P}(A,1)\epsilon - \mathcal{P}(\text{TP}_A)\epsilon}{\mathcal{P}(A,1)\left(\mathcal{P}(A,1) - \epsilon\right)} \right| \\
&= \epsilon \left| \frac{\mathcal{P}(\text{FN}_A)}{\mathcal{P}(A,1)\left(\mathcal{P}(A,1) - \epsilon\right)} \right| \\
&\leq \epsilon \left| \frac{1}{\mathcal{P}(A,1) - \epsilon} \right| \\
&= \epsilon \left| \frac{1}{\mathcal{P}'(A,1)} \right| \\
&\leq \epsilon \left(\frac{1}{\min(\mathcal{P}(A), \mathcal{P}'(A))} \right)
\end{aligned}$$

We now consider the case where $\mathcal{P}'(\text{FN}_A) = \mathcal{P}(\text{FN}_A) - \epsilon$.

$$\begin{aligned}
|\text{TPR}_A(f,\mathcal{P}) - \text{TPR}_A(f,\mathcal{P}')| &= \left|\frac{\mathcal{P}(\text{TP}_A)}{\mathcal{P}(A)} - \frac{\mathcal{P}'(\text{TP}_A)}{\mathcal{P}'(A,1)}\right| \\
&= \left|\frac{\mathcal{P}(\text{TP}_A)}{\mathcal{P}(A,1)} - \frac{\mathcal{P}(\text{TP}_A)}{\mathcal{P}(A,1)-\epsilon}\right| \\
&= \left|\frac{\mathcal{P}(\text{TP}_A)\mathcal{P}(A,1) - \mathcal{P}(\text{TP}_A)\epsilon - \mathcal{P}(\text{TP}_A)\mathcal{P}(A,1)}{\mathcal{P}(A,1)\left(\mathcal{P}(A,1)-\epsilon\right)}\right| \\
&= \left|\frac{\mathcal{P}(\text{TP}_A)\epsilon}{\mathcal{P}(A,1)\left(\mathcal{P}(A,1)-\epsilon\right)}\right| \\
&\le \epsilon\left|\frac{1}{\mathcal{P}(A,1)-\epsilon}\right| \\
&= \epsilon\left|\frac{1}{\mathcal{P}'(A,1)}\right| \\
&\le \epsilon\left(\frac{1}{\min(\mathcal{P}(A,1),\mathcal{P}'(A,1))}\right) \qquad (9)
\end{aligned}$$

Here we argued for when $(A,1)$ loses mass. We can similarly argue the case where $(A,1)$ gains mass, giving us an identical bound. Also, here we argued for group A, and an identical argument for D shows that

$$|\text{TPR}_A(f,\mathcal{P}) - \text{TPR}_A(f,\mathcal{P}')| \le \epsilon\left(\frac{1}{\min(\mathcal{P}(D,1),\mathcal{P}'(D,1))}\right) \qquad (10)$$

Plugging Equations 9 and 10 into Equation 8, we get that

$$|\text{Unf}_{\text{EO}}(f,\mathcal{P}) - \text{Unf}_{\text{EO}}(f,\mathcal{P}')| \le \epsilon\left(\frac{1}{\min(\mathcal{P}(D,1),\mathcal{P}'(D,1))}\right) + \epsilon\left(\frac{1}{\min(\mathcal{P}(D,1),\mathcal{P}'(D,1))}\right)$$

□

Now we state and prove our main result for EO-fair BOC.

Theorem 2 (Robustness of EO-Fair BOC)**.** *Given distributions $\mathcal{P},\mathcal{P}'$, such that $TV(\mathcal{P},\mathcal{P}') = \epsilon$, we have that*

$$\left|Acc(f^{EO}_{\mathcal{P}},\mathcal{P}) - Acc(f^{EO}_{\mathcal{P}'},\mathcal{P}')\right| \le \epsilon\left(1 + 2\max(\mathcal{P}(1),\mathcal{P}'(1))\left(\frac{1}{\min(\mathcal{P}(A,1),\mathcal{P}'(A,1))} + \frac{1}{\min(\mathcal{P}(D,1),\mathcal{P}'(D,1))}\right)\right),$$

where $f^{EO}_{\mathcal{P}}, f^{EO}_{\mathcal{P}'}$ are the EO-Fair BOC's on $\mathcal{P},\mathcal{P}'$ respectively

Proof. Following the proof of Theorem 1, by Lemma 1 and transitivity, it suffices to show the theorem statement for the case where the transition from $\mathcal{P}$ to $\mathcal{P}'$ is elementary in that the only difference between the two distributions is that there are two elements a and b that have ϵ more mass and ϵ less mass, respectively, in $\mathcal{P}$ as compared to $\mathcal{P}'$ (all other elements have the same mass in the two distributions). So, in the remainder of the proof, we only consider elementary transitions.

Consider the transfer of ϵ mass from a to b in a continuous manner. During this process, either the cell corresponding to element a will monotonically increase in score or monotonically decrease in score. The same holds for the cell corresponding to element b. The scores of all other cells will remain the same. In the following argument, we assume that the score of the cell of a decreases monotonically and that of b increases monotonically. All of the arguments are analogous for the remaining three cases.

Let f denote the EO BOC for the current distribution $\mathcal{P}$ at any instant in this mass transfer process ending in distribution $\mathcal{P}'$. As the mass transfer proceeds, we analyze how the EO BOC changes from $f^{\text{EO}}_{\mathcal{P}}$ to $f^{\text{EO}}_{\mathcal{P}'}$. We consider the largest mass transfer $\delta\epsilon$ until one of the two following events occur.

1. Equal-score event: The cell of a has the same score as the adjacent cell lower in the sorted order or the cell of b has the same score as the adjacent cell higher in the sorted order.

2. Threshold event: The score of a merged cell containing a or b becomes exactly 0.5.

Note that by the choice of $\delta\epsilon$, during the transfer $\delta\epsilon$, all the cells remain in the same order in both groups; furthermore, all masses and scores of all cells other than the ones containing a or b remain the same during the transfer. By Claim 2,

$$\begin{aligned}\delta\mathrm{Unf}_{\mathrm{EO}} &= \left|\mathrm{Unf}_{\mathrm{EO}}(f^{\mathrm{EO}}_{\mathcal{P}}, \mathcal{P}) - \mathrm{Unf}_{\mathrm{EO}}(f^{\mathrm{EO}}_{\mathcal{P}}, \mathcal{P}')\right| \\ &\leq \delta\epsilon\left(\frac{1}{\mathcal{P}(A,1)} + \frac{1}{\mathcal{P}(D,1)}\right).\end{aligned}$$

Since $\mathrm{Unf}_{\mathrm{EO}}(f^{\mathrm{EO}}_{\mathcal{P}}(\mathcal{P})) = 0$, we know that $\delta\mathrm{Unf}_{\mathrm{EO}} = \mathrm{Unf}_{\mathrm{EO}}(f^{\mathrm{EO}}_{\mathcal{P}}, \mathcal{P}') = \left|\mathrm{TPR}_A(f^{\mathrm{EO}}_{\mathcal{P}}, \mathcal{P}') - \mathrm{TPR}_D(f^{\mathrm{EO}}_{\mathcal{P}}, \mathcal{P}')\right|$. Consider the cell q that is split in the middle by the threshold corresponding to f (for now, assume $q \in D$). Since neither the equal-score event nor the 0.5-score event occur, we see that after the transition, $f^{\mathrm{EO}}_{\mathcal{P}}$ has $\delta\mathrm{Unf}_{\mathrm{EO}}$ difference in TPR between groups. To modify $f^{\mathrm{EO}}_{\mathcal{P}} \to f^{\mathrm{EO}}_{\mathcal{P}'}$, we therefore need to move to move the boundary at q so that TPR in both groups align and EO is satisfied (the classifier f remains the same apart from its action on q). The change in function ($|\Delta f(q)|$) of element q is bounded by $\delta\mathrm{Unf}_{\mathrm{EO}}\frac{\mathcal{P}(D,1)}{\mathcal{S}(q)\mathcal{P}(q)}$, after scaling (since $\mathcal{P}(D,1)\delta\mathrm{Unf}_{\mathrm{EO}} = |\Delta f(q)|\,\mathcal{P}(q)\mathcal{S}(q)$). If f' denotes the EO-fair BOC for the distribution at the end of the $\delta\epsilon$ mass transfer (just prior to any of the two events), then by Lemma 2, the change in accuracy of the optimal fair classifier is bounded by

$$\begin{aligned}|\mathrm{Acc}(f,\mathcal{P}) - \mathrm{Acc}(f',\mathcal{P}')| &\leq |\mathcal{P}(q)(2\mathcal{S}(q)-1)\Delta f(q)| + \delta\epsilon \\ &\leq \delta\epsilon\left(\frac{1}{\min(\mathcal{P}(A,1),\mathcal{P}'(A,1))} + \frac{1}{\min(\mathcal{P}(D,1),\mathcal{P}'(D,1))}\right)\frac{\mathcal{P}(D,1)\,|(2\mathcal{S}(q)-1)|}{\mathcal{S}(q)} \quad (11)\\ &\leq \delta\epsilon\left(\frac{1}{\min(\mathcal{P}(A,1),\mathcal{P}'(A,1))} + \frac{1}{\min(\mathcal{P}(D,1),\mathcal{P}'(D,1))}\right)\frac{\mathcal{P}(D)}{\mathcal{S}(q)} + \delta\epsilon, \quad (12)\end{aligned}$$

where the last equation follows by monotonicity. Note that $\frac{1}{\mathcal{S}(q)}$ can potentially blow up, and we would like to bound it. Notice that since by assumption, f splits q in the middle, we know that there is a portion of q that is accepted. Hence, the weighted score of a merged cell involving q (say $\mathcal{C}_q$) has score above the threshold of 0.5. Let C_q contain some element t from group A, and denote length of group z in $\mathcal{C}_q$ by l_z. Since the TPR of both components are equal, we know that

$$\frac{\mathcal{P}(A)l_A\mathcal{S}(t)}{\mathcal{P}(A,1)} = \frac{\mathcal{P}(D)l_D\mathcal{S}(q)}{\mathcal{P}(D,1)} \tag{13}$$

Also, since $\mathcal{S}(\mathcal{C}_q) \geq 0.5$, we know that

$$\mathcal{P}(A)l_A\mathcal{S}(t) + \mathcal{P}(D)l_D\mathcal{S}(q) \geq \frac{\mathcal{P}(A)l_A + \mathcal{P}(D)l_D}{2} \tag{14}$$

Combining Equations 13, and 14, and after a bunch of simplification, we get that

$$\begin{aligned}\frac{1}{\mathcal{S}(q)} &\leq \frac{2\mathcal{P}(1)}{\mathcal{P}(D,1)} - \frac{\mathcal{P}(A,1)}{\mathcal{P}(D,1)\mathcal{S}(t)} \quad (15)\\ &\leq \frac{2\mathcal{P}(1)}{\mathcal{P}(D,1)} \quad (16)\end{aligned}$$

Where the second equation follows because $\mathcal{S}(t) \geq 0$. Plugging Equation 16 into Equation 12, we get that

$$\begin{aligned}|\mathrm{Acc}(f,\mathcal{P}) - \mathrm{Acc}(f',\mathcal{P}')| &\leq \delta\epsilon\left(\frac{1}{\min(\mathcal{P}(A,1),\mathcal{P}'(A,1))} + \frac{1}{\min(\mathcal{P}(D,1),\mathcal{P}'(D,1))}\right)2\mathcal{P}(1) + \delta\epsilon \\ &\leq \delta\epsilon\left(\frac{1}{\min(\mathcal{P}(A,1),\mathcal{P}'(A,1))} + \frac{1}{\min(\mathcal{P}(D,1),\mathcal{P}'(D,1))}\right)2\max(\mathcal{P}(1),\mathcal{P}'(1)) + \delta\epsilon \quad \text{(monotonicity)}\end{aligned}$$

The handling of the equal-score and threshold events is identical to that in the proof of Theorem 1. We repeat here for convenience.

1. Equal-score event: If the cell of a has the same score as the adjacent cell lower in the sorted order, then we swap the two cells so that the cell of a is lower in the order. Similarly, if the cell of b has the same score as the adjacent cell higher in the order, then we swap the two cells so that the cell of b is higher in the order. We update the classifier f and note that this change has no impact on the accuracy of f.

2. Threshold event: The score of a merged cell containing a or b becomes exactly 0.5. We include the merged cell in the classifier f, again without changing accuracy.

Thus, between any two occurrences of these events, the change in accuracy is bounded by an amount proportional to the mass transfer; when we reach these occurrences, the mass transfer is paused, the BOC changes without any change in accuracy. Furthermore, at every occurrence of the event, one of these three events happen: the cell containing a moves down in the order, the cell containing b moves up in the order, or an additional merged cell is placed above the threshold. Since the number of times these events can occur is upper bounded by the number of cells in the two groups, this process is finite. Therefore, adding over all the $\delta\epsilon$ mass transfers, we obtain the desired bound on the change in accuracy between the BOC's for $\mathcal{P}$ and $\mathcal{P}'$, thus completing the proof of the theorem. □

We now state the following corollary, which follows from Claim 5 and Theorem 2. It roughly states that given 2 closeby distributions $\mathcal{P}, \mathcal{P}'$, the accuracy of the respective EO-Fair BOC's is similar on $\mathcal{P}$. Such a property is useful when we want a guarantee that intuitively says that if we train on the corrupted distribution $\mathcal{P}'$, we get a similar outcome to what we would have gotten had we trained on the true distribution $\mathcal{P}$.

Corollary 2 (Stability of EO-Fair BOC)**.** *Given distributions $\mathcal{P}, \mathcal{P}'$, such that $TV(\mathcal{P}, \mathcal{P}') = \epsilon$, we have that*

$$\left|Acc(f^{EO}_{\mathcal{P}}, \mathcal{P}) - Acc(f^{EO}_{\mathcal{P}'}, \mathcal{P})\right| \le 2\epsilon \left(1 + \max(\mathcal{P}(1), \mathcal{P}'(1)) \left(\frac{1}{\min(\mathcal{P}(A,1), \mathcal{P}'(A,1))} + \frac{1}{\min(\mathcal{P}(D,1), \mathcal{P}'(D,1))}\right)\right),$$

where $f^{EO}_{\mathcal{P}}, f^{EO}_{\mathcal{P}'}$ are the EO-Fair BOC's on $\mathcal{P}, \mathcal{P}'$ respectively

B.3. Predictive Equality

We can obtain analogous results for Predictive Equality from the same proof techniques as that of Equal Opportunity (since we can just reverse the roles of the labels 0 and 1 in EO to get results for PE). Hence, we only discuss the proofs for EO, and state the analogous results for PE below without proof.

Claim 6 (PE Shift for a Fixed Hypothesis)**.** *Given distributions $\mathcal{P}, \mathcal{P}'$, such that $TV(\mathcal{P}, \mathcal{P}') \le \epsilon$, and any hypothesis f, it holds that*

$$|Unf_{PE}(f, \mathcal{P}) - Unf_{PE}(f, \mathcal{P}')| \le \epsilon \left(\frac{1}{\min(\mathcal{P}(A,0), \mathcal{P}'(A,0))} + \frac{1}{\min(\mathcal{P}(D,0), \mathcal{P}'(D,0))}\right).$$

Theorem 3 (Robustness of PE-Fair BOC)**.** *Given distributions $\mathcal{P}, \mathcal{P}'$, such that $TV(\mathcal{P}, \mathcal{P}') = \epsilon$, we have that*

$$\left|Acc(f^{PE}_{\mathcal{P}}, \mathcal{P}) - Acc(f^{PE}_{\mathcal{P}'}, \mathcal{P}')\right| \le \epsilon \left(1 + 2\max(\mathcal{P}(0), \mathcal{P}'(0)) \left(\frac{1}{\min(\mathcal{P}(A,0), \mathcal{P}'(A,0))} + \frac{1}{\min(\mathcal{P}(D,0), \mathcal{P}'(D,0))}\right)\right),$$

where $f^{PE}_{\mathcal{P}}, f^{PE}_{\mathcal{P}'}$ are the PE-Fair BOC's on $\mathcal{P}, \mathcal{P}'$ respectively.

Corollary 3 (Stability of PE-Fair BOC)**.** *Given distributions $\mathcal{P}, \mathcal{P}'$, such that $TV(\mathcal{P}, \mathcal{P}') = \epsilon$, we have that*

$$\left|Acc(f^{PE}_{\mathcal{P}}, \mathcal{P}) - Acc(f^{PE}_{\mathcal{P}'}, \mathcal{P})\right| \le 2\epsilon \left(1 + \max(\mathcal{P}(0), \mathcal{P}'(0)) \left(\frac{1}{\min(\mathcal{P}(A,0), \mathcal{P}'(A,0))} + \frac{1}{\min(\mathcal{P}(D,0), \mathcal{P}'(D,0))}\right)\right),$$

where $f^{EO}_{\mathcal{P}}, f^{EO}_{\mathcal{P}'}$ are the EO-Fair BOC's on $\mathcal{P}, \mathcal{P}'$ respectively