How Far Can Fairness Constraints Help Recover From Biased Data?

Mohit Sharma [1] **Amit Jayant Deshpande** [2]

Abstract

A general belief in fair classification is that fairness constraints incur a trade-off with accuracy, which biased data may worsen. Contrary to this belief, Blum & Stangl (2019) show that fair classification with equal opportunity constraints even on extremely biased data can recover optimally accurate and fair classifiers on the original data distribution. Their result is interesting because it demonstrates that fairness constraints can implicitly rectify data bias and simultaneously overcome a perceived fairness-accuracy trade-off. Their data bias model simulates under-representation and label bias in underprivileged population, and they show the above result on a stylized data distribution with i.i.d. label noise, under simple conditions on the data distribution and bias parameters.

We propose a general approach to extend the result of Blum & Stangl (2019) to different fairness constraints, data bias models, data distributions, and hypothesis classes. We strengthen their result, and extend it to the case when their stylized distribution has labels with Massart noise instead of i.i.d. noise. We prove a similar recovery result for arbitrary data distributions using fair reject option classifiers. We further generalize it to arbitrary data distributions and arbitrary hypothesis classes, i.e., we prove that for any data distribution, if the optimally accurate classifier in a given hypothesis class is fair and *robust*, then it can be recovered through fair classification with equal opportunity constraints on the biased distribution whenever the bias parameters satisfy certain simple conditions. Finally, we show applications of our technique to time-varying data bias in classification and fair machine learning pipelines.

Work done during internship at Microsoft Research India. [1]Indraprastha Institute of Information Technology, Delhi, India [2]Microsoft Research India. Correspondence to: Mohit Sharma <mohits@iiitd.ac.in>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

Fairness in machine learning has been an important qualitative and quantitative research topic for more than a decade (Barocas et al., 2019; Mehrabi et al., 2021). Several quantitative fairness metrics and various bias mitigation techniques developed over the years are available to practitioners as open-source fairness toolkits (Bellamy et al., 2018). Fairness-accuracy trade-offs often seem inevitable, and bias mitigation guarantees can fail when there is a mismatch between train and test data distributions; it is a double whammy when we get neither the desired accuracy nor the desired fairness on deployment. Hence, it is important to understand the role of fair classification and fairness constraints when the training data in machine learning pipelines has systematic biases (Blum & Stangl, 2019; Konstantinov & Lampert, 2022; Schrouff et al., 2022).

Blum & Stangl (2019) propose a model for data bias that simulates systematic under-representation of the underprivileged group and label bias/flip on the underprivileged positive population. Applying this data bias model to a stylized data distribution with i.i.d. label noise, they prove a recovery result that makes the following high-level point: careful choice of fairness constraints can implicitly rectify even extreme data bias and overcome a perceived fairness-accuracy trade-off. Formally, for their stylized distribution with i.i.d. label noise, they prove that the optimal fair classifier satisfying equal opportunity on a biased (or bias-induced) data distribution coincides with the Bayes optimal classifier (that also happens to be perfectly fair) on the original data distribution, if the bias parameters satisfy certain simple conditions. Their proof is tailored to equal opportunity constraints and uses a careful case analysis and iterative argument that is not amenable to arbitrary distributions or hypothesis classes. They also show a specific distribution where a similar result fails to hold if we use demographic parity constraints instead of equal opportunity constraints.

We take a significantly different approach and observe that the regression function for group-aware classification (i.e., the positive class probability conditioned on the input features and sensitive attribute) undergoes a linear fractional transformation when we apply various data bias models, including the one proposed by Blum & Stangl (2019). This observation plays an important role in our proofs because

the optimal group-aware fair classifier among the hypothesis class of all binary classifiers can be mathematically characterized by group-dependent thresholds on the regression function (Menon & Williamson, 2018; Chzhen et al., 2019). Under the conditions on the data distribution and bias parameters as in Blum & Stangl (2019), we show that the fairness-corrected thresholds applied to a linear fractional transformation recover the Bayes optimal classifier on the stylized distribution of Blum & Stangl (2019) with i.i.d. label noise. Our proof uncovers a stronger version of their result that these conditions on the bias parameters are both necessary and sufficient. Moreover, our proof readily extends to the case when the above distribution has Massart or malicious label noise (Massart & Nédélec, 2006; Rivest & Sloan, 1994; Sloan, 1996) instead of i.i.d. label noise. We extend our recovery result to arbitrary data distributions by considering reject option classifiers (Bartlett & Wegkamp, 2008; Cortes et al., 2016) that can abstain from prediction on a small fraction of inputs by paying a penalty. We further generalize our results to allow arbitrary data distribution as well as arbitrary hypothesis class. Generalizing our ideas beyond threshold-based arguments, we show that on arbitrary data distributions, if the optimal classifier in a given arbitrary hypothesis class is fair and *robust*, then it can be recovered through fair classification on the biased distribution, whenever the bias parameters satisfy certain simple conditions. Finally, we propose multi-step models to capture time-varying data bias and machine learning pipeline, and investigate necessary conditions on the data distribution and bias parameters for similar recovery results in the finite and infinite time horizons. Now we outline the organization of our results in the paper.

- Section 3 contains our theoretical setup for data bias and group-aware fair classification. In Subsection 3.1, we describe some recently studied data bias models in fair classification (Blum & Stangl, 2019; Dai & Brown, 2020; Wang et al., 2021; Biswas et al., 2019) (see Examples 3.1, 3.2, & 3.3), and show that all of them result in a linear fractional transformation of the regression function. Subsection 3.2 captures how fairness constraints and threshold-based characterizations of optimal fair classifiers change under data bias. We focus on equal opportunity constraints and the data bias model of Blum & Stangl (2019) (Examples 3.1) as an illustrative example running through the rest of our paper.

- In Section 4, we extend the result of Blum & Stangl (2019) to the case when their stylized distribution has Massart label noise instead of i.i.d. label noise (Theorem 4.1). We strengthen their result (Theorem 4.2) and prove its analog for demographic parity constraints replacing equal opportunity (Theorem 4.3).

- In Section 5, we show that our proof technique based on threshold classifiers extends to *arbitrary* data distributions, if we allow reject option classifiers that can abstain from prediction on a small fraction of inputs.

- In Section 6, we invent clever workarounds to generalize our results further to recover the optimal fair and *robust* hypothesis in an *arbitrary* hypothesis class simply by fair classification on the biased version of an *arbitrary* data distribution (Theorem 6.2).

- Finally, in Section 7, we propose time-varying data bias models (also applicable to multi-stage machine learning pipelines) and investigate necessary condition for extending our recovery results above to the finite and infinite time horizon (Theorems 7.2 & 7.3).

2. Related Work

There has been a plethora of recent work on fairness-accuracy trade-offs and data bias (Menon & Williamson, 2018; Wick et al., 2019; Blum & Stangl, 2019; Dutta et al., 2020; Maity et al., 2021), but the closest to our work is the result of Blum & Stangl (2019) that we strengthen and generalize in many ways. Though we take equal opportunity constraints (Hardt et al., 2016) and the data bias model of Blum & Stangl (2019) for under-representation and label bias as an illustrative example running through our paper, our techniques readily extend to other popular fairness constraints such as demographic parity (Dwork et al., 2012) and other recent data bias models (Dai & Brown, 2020; Wang et al., 2021; Biswas & Mukherjee, 2021); many possible extensions are covered in the Appendix. Our proof techniques in Section 4 & 5 lean heavily on threshold-based characterizations of optimal fair classifiers known in previous work (Menon & Williamson, 2018; Chzhen et al., 2019; Zeng et al., 2022a;b).

Now we describe recent related works that complement our approach to rectify data bias in fair classification. Feasibility of fair classification under data corruption and malicious noise in training data has been studied in Konstantinov & Lampert (2022); Blum et al. (2023). Recent work has studied fair classification with noisy sensitive attributes (Lamy et al., 2019; Ghosh et al., 2023; Celis et al., 2021), noisy labels (Fogliato et al., 2020), feature-dependent label bias (Jiang & Nachum, 2020), sample selection bias (Du & Wu, 2021; Zhu et al., 2023), subpopulation shift (Maity et al., 2021), and causal models of data bias (Plecko & Bareinboim, 2022; Madras et al., 2019; Cheong et al., 2023). All of them propose algorithmic modifications to the vanilla fair classification to rectify noisy or biased data. Complementing the theoretical aspects, recent work has also empirically investigated the effect of data bias and choice of fairness constraints on the accuracy and fairness of various fair clas-

sifiers (Islam et al., 2022; Akpinar et al., 2022; Sharma et al., 2023; Ghosh et al., 2023).

3. Data Bias Models & Fair Classification

Let (X, A, Y) be a random data point from the joint distribution D over $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$, where $\mathcal{X}, \mathcal{A}, \mathcal{Y}$ denote the set of features, the set of sensitive attributes, and the set of class labels, respectively. We assume the feature space $\mathcal{X}$ to be discrete. We consider group-aware classifiers $f : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}$ and assume, for simplicity, binary sensitive attributes $\mathcal{A} = \{0, 1\}$ and binary class labels $\mathcal{Y} = \{0, 1\}$. The binary classifier of maximum accuracy $h^* = \text{argmax}_h \Pr(h(X, A) = Y)$ is known as the Bayes optimal classifier, and is given by $h^*(x, a) = \mathbb{I}(\eta(x, a) \geq 1/2)$, where $\eta(x, a) = \Pr(Y = 1 | X = x, A = a)$ (Zeng et al., 2022a). This function η is known as the *regression function* in statistical machine learning; see Chapter 2 of Devroye et al. (1996). Binary classifiers that apply a threshold on the regression function η play a key role in our work as well as other recent works on classification beyond accuracy (Elkan, 2001; Singh & Khim, 2022) and fair classification (Menon & Williamson, 2018; Chzhen et al., 2019; Zeng et al., 2022b).

We use $\tilde{D}$ to denote the biased joint distribution over features, sensitive attributes and class labels, and use $(\tilde{X}, \tilde{A}, \tilde{Y})$ to denote a random data point from the biased distribution $\tilde{D}$. If X (or A) in the joint distribution remains unchanged when we go from D to $\tilde{D}$, then we simply use X (or A) in the place of $\tilde{X}$ (or $\tilde{A}$).

3.1. Regression Functions on Biased Data

Below we list some data bias models from recent works on fair classification from biased data (Blum & Stangl, 2019; Dai & Brown, 2020; Biswas & Mukherjee, 2021). We make a key observation that for all of these data bias models, the regression function on the biased distribution $\tilde{\eta}(x, a) = \Pr\left(\tilde{Y} = 1 | \tilde{X} = x, \tilde{A} = a\right)$ can be expressed as a linear fractional transformation of the regression function on the original distribution $\eta(x, a) = \Pr(Y = 1 | X = x, A = a)$. In other words,

$$\tilde{\eta}(x, a) = \frac{P\eta(x, a) + Q}{R\eta(x, a) + S}, \quad \text{for some } P, Q, R, S \in \mathbb{R}.$$

Please refer to Propositions A.1, A.2 and A.3 in Appendix A.2.

Example 3.1. (Blum & Stangl, 2019) Consider a biased distribution $\tilde{D}$ obtained from the original distribution D by the following process defined by under-representation and label bias parameters $\beta_p, \beta_n, \nu \in (0, 1)$. A random data point (X, A, Y) from D with $A = 1$ remains unchanged, the points with $A = 0, Y = 1$ survive independently with probability β_p, and the points with $A = 0, Y = 0$ survive independently with probability β_n. Finally, the survived points with $A = 0, Y = 1$ keep their class label 1 with probability $1 - \nu$, and it gets flipped to 0 with probability ν. For the privileged group $A = 1$, we have $\tilde{\eta}(x, 1) = \eta(x, 1)$, for all $x \in \mathcal{X}$, whereas for the underprivileged group $A = 0$, we prove that $\tilde{\eta}(x, 0) = \dfrac{(1 - \nu)\eta(x, 0)}{(1 - c)\eta(x, 0) + c}$, where $c = \dfrac{\beta_n}{\beta_p}$ in Proposition A.1.

Example 3.2. (Dai & Brown, 2020; Wang et al., 2021) Consider a biased distribution $\tilde{D}$ obtained from the original distribution D by introducing a group-dependent label flip from (X, A, Y) to $(X, A, \tilde{Y})$ where $\epsilon_a^1 = \Pr\left(\tilde{Y} = 0 | Y = 1, A = a\right)$ and $\epsilon_a^0 = \Pr\left(\tilde{Y} = 1 | Y = 0, A = a\right)$, with $0 \leq \epsilon_a^1 + \epsilon_a^0 < 1$. For this data bias model, we show that $\tilde{\eta}(x, a) = (1 - \epsilon_a^1 - \epsilon_a^0)\eta(x, a) + \epsilon_a^0$ in Proposition A.2.

Example 3.3. (Biswas & Mukherjee, 2021) Consider a biased distribution $\tilde{D}$ obtained from the original distribution D by introducing a group-dependent prior probability shift such that $\tilde{A} = A$ and $\Pr\left(\tilde{X} = x | \tilde{Y} = i, A = a\right) = \Pr(X = x | Y = i, A = a)$, for any $i, a \in \{0, 1\}$, but $\Pr\left(\tilde{Y} = i | A = a\right) \neq \Pr(Y = i | A = a)$. For this data bias model, we prove that $\tilde{\eta}(x, a) = \dfrac{\eta(x, a)}{(1 - \alpha)\eta(x, a) + \alpha}$, where $\alpha = \dfrac{\Pr\left(\tilde{Y} = 0 | A = a\right) \Pr(Y = 1 | A = a)}{\Pr\left(\tilde{Y} = 1 | A = a\right) \Pr(Y = 0 | A = a)}$ in Proposition A.3.

For classifiers that apply a threshold on $\eta(x, a)$ or $\tilde{\eta}(x, a)$, it is important to understand when the above linear fractional transformations are order-preserving.

Proposition 3.4. *Suppose $S \geq 0$, $R + S \geq 0$, and $PS - QR \geq 0$, then the transformation $\tilde{\eta}(x, a) = \dfrac{P\eta(x, a) + Q}{R\eta(x, a) + S}$ is order-preserving, i.e., $\eta(x_1, a) \leq \eta(x_2, a)$ iff $\tilde{\eta}(x_1, a) \leq \tilde{\eta}(x_2, a)$.*

The proof is provided in Appendix A.1. Note that all of the above data bias models satisfy the order-preservation property. For the rest of the paper, we exclusively focus on the under-representation and label bias model in Example 3.1 (Blum & Stangl, 2019) as an illustrative example. Our techniques are flexible and can be applied to obtain similar results for other data bias models.

3.2. Fair Classification on Biased Data

Demographic Parity (equal group-wise positivity rates) and Equal Opportunity (equal group-wise true positive rates)

are two most popular fairness constraints in classification. It is easy to see that the true positive rates (TPRs) and the true negative rates (TNRs) for a classifier on the biased data distribution can be expressed as linear combinations of its TPRs and TNRs on the original data distribution.

Proposition 3.5. *Let D be any distribution on $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$ and let $\tilde{D}$ be its biased version defined using any of the data bias models defined in Subsection 3.1. Let h be any hypothesis, and let $TPR_a(h)$ and $\widetilde{TPR}_a(h)$ be its true positive rates according to D and $\tilde{D}$, respectively, conditioned on the underprivileged group $A = a$. Let $TNR_a(h)$ and $\widetilde{TNR}_a(h)$ be the true negative rates defined similarly. Then*

1. $\widetilde{TPR}_a(h) = \Pr\left(Y = 1|\tilde{Y} = 1, A = a\right) TPR_a(h) + \Pr\left(Y = 0|\tilde{Y} = 1, A = a\right) FPR_a(h)$, *and*
2. $\widetilde{TNR}_a(h) = \Pr\left(Y = 0|\tilde{Y} = 0, A = a\right) TNR_a(h) + \Pr\left(Y = 1|\tilde{Y} = 0, A = a\right) FNR_a(h)$.

The proof of Proposition 3.5 is given in Appendix A.3, and we get the following interesting corollary for the data bias model in Example 3.1 (Blum & Stangl, 2019).

Corollary 3.6. *Let D be any distribution and $\tilde{D}$ be its biased version as in Example 3.1. Given any hypothesis class $\mathcal{H}$, let $\mathcal{H}_{fair,\ EO}$ be its subset that satisfies equal opportunity on the original distribution D, i.e., $\mathcal{H}_{fair,\ EO} = \{f \in \mathcal{H} : TPR_0(f) = TPR_1(f)\}$. Similarly, let $\widetilde{\mathcal{H}}_{fair,\ EO} = \{f \in \mathcal{H} : \widetilde{TPR}_0(f) = \widetilde{TPR}_1(f)\}$ be the subset of $\mathcal{H}$ satisfying equal opportunity on the biased distribution $\tilde{D}$. Then $\widetilde{TPR}_0(h) = TPR_0(h)$ and $\widetilde{TPR}_1(h) = TPR_1(h)$, and hence, $\mathcal{H}_{fair,\ EO} = \widetilde{\mathcal{H}}_{fair,\ EO}$.*

Remark 3.7. Corollary 3.6 can be easily extended to other fairness metrics such as demographic parity as well as the hypothesis class of *approximately* fair classifiers that satisfy fairness constraints up to a small additive or multiplicative error. However, we focus on exact equal opportunity as in Blum & Stangl (2019) for a direct, illustrative application. Please see Appendix A.3 for additional results.

The optimal fair classifier for equal opportunity (similarly, demographic parity) on a given data distribution can be expressed by group-dependent thresholds applied to the regression function (Menon & Williamson, 2018; Chzhen et al., 2019). Since the regression function $\tilde{\eta}(x, a)$ on the biased distribution $\tilde{D}$ is an order-preserving linear fractional transformation of $\eta(x, a)$, the optimal fair classifier for equal opportunity on $\tilde{D}$ is equivalent to applying group-dependent thresholds to $\eta(x, a)$.

Proposition 3.8. *For any distribution D and its biased version $\tilde{D}$ described in Example 3.1, let $\tilde{h}_{EO}$ be a classifier of the maximum accuracy among all binary classifiers that satisfy equal opportunity on $\tilde{D}$. Then there exists $\lambda^* \in \mathbb{R}$ such that $\tilde{h}_{EO}(x, a) = \mathbb{I}\left(\eta(x, a) \geq t_a\right)$, where*

$$t_a = \begin{cases} \dfrac{1}{1 + \dfrac{1-2\nu}{c} + \dfrac{\lambda^*}{\beta_n \Pr\left(Y=1, A=0\right)}}, & \text{for } a = 0 \\ \dfrac{1}{2 - \dfrac{\lambda^*}{\Pr\left(Y=1, A=1\right)}}, & \text{for } a = 1. \end{cases}$$

The Proof of Proposition is given in Appendix A.3. We can similarly derive the optimal threshold with the biased distribution for the Demographic parity constraint (Proposition A.4 in Appendix A.3).

4. Recovering Optimal Classifier from Biased Data for Massart Label Noise

Blum & Stangl (2019) consider a stylized distribution D with i.i.d. label noise and show that the optimal fair classifier $\tilde{h}_{EO}$ on the biased distribution $\tilde{D}$ (defined in Example 3.1) recovers the Bayes optimal (and fair) classifier h^* on the original distribution D, if the bias parameters satisfy certain simple conditions. Note that this does not require knowing, estimating, or correcting for data bias explicitly, and their result holds even for extreme under-representation and label bias in $\tilde{D}$. We first demonstrate the utility of our technique by generalizing the recovery result of Blum & Stangl (2019) to the case of Massart noise (Massart & Nédélec, 2006). We describe the distribution setup below, give a sketch of our proof, and point out the generality of our technique compared to Blum & Stangl (2019).

4.1. Generalizing Blum & Stangl (2019) Recovery Result for Massart Noise

Assume any arbitrary data distribution D on $\mathcal{X} \times \mathcal{A}$. Let $\Pr\left(A = 0\right) = r$ and $\Pr\left(A = 1\right) = 1 - r$, for some $0 < r < 1$. Let $h : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}$ be any hypothesis that satisfies $\Pr\left(h(X, A) = 1|A = 0\right) = \Pr\left(h(X, A) = 1|A = 1\right)$. Let $\delta < 1/2$, and extend the distribution to $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$ as follows. $Y|X = x, A = a$ takes value $h(x, a)$ with probability $1 - \delta(x, a)$, and $\neg h(x, a)$ with probability $\delta(x, a)$, for some $\delta(x, a) \leq \delta$. Let D be the resulting distribution on $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$. This type of bounded noise in class label is popularly known as Massart noise[1] in literature, based on a noise model proposed by Massart and Nédélec(Massart & Nédélec, 2006). We assume that the Massart noise is added in a way that equalizes the base rates on the two protected groups, i.e., $Pr(Y = 1|A = 0) = Pr(Y = 1|A = 1) = q$. Since $\delta < 1/2$, the Bayes optimal classifier h^* on the distri-

[1] Equivalently known as malicious classification noise in previous work (Rivest & Sloan, 1994; Sloan, 1996).

bution D coincides with h and satisfies Equal Opportunity.

Theorem 4.1. *For any distribution D defined as above and its biased version $\tilde{D}$ defined as in Example 3.1 using the bias parameters $\beta_p, \beta_n, \nu \in (0,1)$. If the data distribution and bias parameters satisfy*

$$(1-r)(1-2\delta) + r\left((1-\delta)\beta_p(1-2\nu) - \delta\beta_n\right) > 0$$

and

$$(1-r)(1-2\delta) + r\left((1-\delta)\beta_n - \delta\beta_p(1-2\nu)\right) > 0,$$

then the optimal equal opportunity classifier on the biased distribution $\tilde{D}$ recovers the Bayes optimal classifier on the original distribution D, i.e., $\tilde{h}_{EO} \equiv h^*$.

The proof of Theorem 4.1 is given in Appendix A.4. The same proof also works for group-dependent Massart noise, i.e., there exist $\delta_0, \delta_1 < 1/2$ such that $\delta(x,a) \leq \delta_a$, for all $(x,a) \in \mathcal{X} \times \mathcal{A}$.

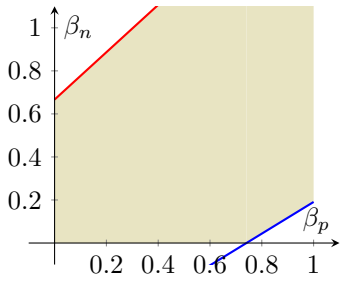

Figure 1. Recovery region for $\beta_p, \beta_n \in (0,1]$ given by the constraints $(1-r)(1-2\delta) + r\left((1-\delta)\beta_p(1-2\nu) - \delta\beta_n\right) > 0$ and $(1-r)(1-2\delta) + r\left((1-\delta)\beta_n - \delta\beta_p(1-2\nu)\right) > 0$ as in Theorem 4.1, when $r = 0.25$, $\nu = 0.05$, and $\delta = 0.45$. We can recover optimal and fair classifiers for a large range of data biases, including extreme under-representation, i.e., region close to the origin $(0,0)$, by applying just equal opportunity constraints.

The conditions in Theorem 4.1 above are identical to those in the recovery result of Blum & Stangl (2019) (Theorem 4.1 in their paper) that only works for the special case when $\delta(x,a) = \delta$, for all $(x,a) \in \mathcal{X} \times \mathcal{A}$. Their proof is arguably less flexible to other models of label noise and data bias, as it relies on clever, iterative modifications of an initial fair classifier until its accuracy cannot be improved further. For their special case $\delta(x,a) = \delta$, for all $(x,a) \in \mathcal{X} \times \mathcal{A}$, our technique gives a stronger statement that the conditions in Theorem 4.1 are in fact both necessary and sufficient. Figure 1 illustrates the recovery conditions in Theorem 4.1 for reasonably chosen group proportion parameter r, label bias ν and δ.

Theorem 4.2. *(a slightly stronger version of Theorem 4.1 in Blum & Stangl (2019)) For the data distribution D described above with $\delta(x,a) = \delta$, for all $(x,a) \in \mathcal{X} \times \mathcal{A}$, and its biased version $\tilde{D}$ as described in Example 3.1, the data distribution and bias parameters satisfy*

$$(1-r)(1-2\delta) + r((1-\delta)\beta_p(1-2\nu) - \delta\beta_n) > 0$$

and

$$(1-r)(1-2\delta) + r((1-\delta)\beta_n - (1-2\nu)\delta\beta_p) > 0,$$

if and only if the optimal equal opportunity classifier on $\tilde{D}$ recovers the Bayes optimal classifier on D, i.e., $\tilde{h}_{EO} \equiv h_{EO} \equiv h^*$.

Similarly, we can also obtain necessary and sufficient conditions for when the optimal demographic parity classifier on $\tilde{D}$ recovers h^*. Blum & Stangl (Blum & Stangl, 2019) only give a specific example where such a recovery is impossible via demographic parity constraints (see Subsection 3.1 of (Blum & Stangl, 2019)) but do not prove any analog of Theorem 4.2 (see Table 1 in (Blum & Stangl, 2019)).

Theorem 4.3. *For the data distribution D described above with $\delta(x,a) = \delta$, for all $(x,a) \in \mathcal{X} \times \mathcal{A}$, and its biased version $\tilde{D}$ as described in Example 3.1, the data distribution and bias parameters satisfy $\beta_p(1-\delta)(1-2\delta-2r(\nu-\delta)) + \delta\beta_n(1-2\delta-2r(1-\delta)) > 0$ and $\beta_p\delta(1-2r(1-\nu) - 2\delta(1-r)) + (1-\delta)\beta_n(1-2\delta(1-r)) > 0$, if and only if the optimal demographic parity classifier on $\tilde{D}$ recovers the Bayes optimal classifier on D, i.e., $\tilde{h}_{DP} \equiv h_{DP} \equiv h^*$.*

Appendix A.4 contains the proofs of Theorems 4.2 & 4.3.

4.2. Proof Sketches

We briefly outline our proof technique for Theorems 4.1, 4.2, 4.3 to explain an important technical contribution of our paper. We write fairness constrained accuracy maximization using a Lagrange multiplier λ. Proposition 3.8 characterizes $\tilde{h}_{EO}$ as $\tilde{h}_{EO}(x,a) = \mathbb{I}(\eta(x,a) \geq t_a)$ that applies group-dependent thresholds t_a on $\eta(x,a)$, where the threshold t_a is actually a function of the optimal Lagrange multiplier λ^*, the data distribution parameters, and the bias parameters. We show (see Lemma A.5) that as long as these thresholds t_a applied to $\eta(x,a)$ for both the groups $a=0$ and $a=1$ lie within the interval $(\delta, 1-\delta)$, we have $\tilde{h}_{EO} \equiv h^*$. We show that the possible choices of λ^* are narrowed down to allow only $\tilde{h}_{EO} \equiv h^*$ using the given conditions on the data distribution and bias parameters, and the fairness constraint on the resulting threshold classifier. We prove that the conditions in Theorem 4.2 are necessary and sufficient for the optimal λ^* parameter in Proposition 3.8 to satisfy that the group-dependent thresholds t_0 and t_1 lie in the interval $(\delta, 1-\delta)$, and equivalently, $\tilde{h}_{EO} \equiv h^*$.

5. Recovery of Optimal Reject Option Classifiers from Biased Data for Arbitrary Data Distributions

A major limitation of Theorems 4.1 and 4.2 is that they work only on stylized distributions, where the label noise is either i.i.d. or Massart. In this section, we remove this limitation

by proving a similar recovery result for arbitrary data distributions. Massart or i.i.d. label noise creates a clear separation between $\eta(x,a)$ values (or high-risk and low-risk), and allows a small interval margin for the group-wise thresholds to recover h^*. To mimic this in an arbitrary data distribution, we consider *reject option classifiers* that are allowed to abstain from prediction by paying a penalty (Bartlett & Wegkamp, 2008; Cortes et al., 2016; Charoenphakdee et al., 2021; Schreuder & Chzhen, 2021; Franc et al., 2023). Models that abstain from prediction play an important role in responsible machine learning, as predictions of high uncertainty can be overseen by a human-in-the-loop. As a result, many recent papers have studied reject option classifiers for fair classification (Madras et al., 2018; Schreuder & Chzhen, 2021; Shah et al., 2022).

Let D be an arbitrary distribution on $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$, with $\mathcal{A} = \{0,1\}$ and $\mathcal{Y} = \{0,1\}$. A reject option classifier $g : \mathcal{X} \times \mathcal{A} \to \{0,1,\perp\}$ either rejects or abstains from prediction on an input (x,a), denoted by $g(x,a) = \perp$, or it predicts $g(x,a) = h(x,a)$ using a binary classifier $h : \mathcal{X} \times \mathcal{A} \to \{0,1\}$. Let $\mathcal{H}$ denote the hypothesis class of all binary classifiers $h : \mathcal{X} \times \mathcal{A} \to \{0,1\}$, and let $\mathcal{H}^{\text{rej}}$ be the hypothesis class of all $g : \mathcal{X} \times \mathcal{A} \to \{0,1,\perp\}$. For rejection penalty given by $\delta > 0$, the optimal reject option classifier is defined as

$$h^{\text{rej}} = \operatorname*{argmin}_{g \in \mathcal{H}^{\text{rej}}} \Pr\left(g(X,A) \neq Y, g(X,A) \neq \perp\right) + \delta \Pr\left(g(X,A) = \perp\right).$$

Proposition 5.1 characterizes the optimal reject option classifier on any distribution D, which is a generalization of the known forms of Optimal reject option classifiers (Section 1 in Bartlett & Wegkamp (2008), Section 2 in Cortes et al. (2016)).

Proposition 5.1. *Let D be any distribution on $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$, with $\mathcal{A} = \{0,1\}$ and $\mathcal{Y} = \{0,1\}$. Let $\delta \in [0, 1/2)$ denote the rejection penalty and h^{rej} be the optimal reject option classifier defined as above. Then*

$$h^{rej}(x,a) = \begin{cases} 0, & \text{if } \eta(x,a) \leq \delta \\ \perp, & \text{if } \eta(x,a) \in (\delta, 1-\delta) \\ 1, & \text{if } \eta(x,a) \geq 1-\delta, \end{cases}$$

where $\eta(x,a) = \Pr(Y = 1 | X = x, A = a)$.

The proof of Proposition 5.1 is given in Appendix A.5. Proposition 5.1 shows how reject option induces a separation between high and low $\eta(x,a)$ values similar to the case of i.i.d. or Massart label noise. A larger separation has an obvious trade-off with a larger fraction of inputs being turned away for model prediction.

Now assume that the optimal reject option classifier h^{rej} satisfies equal opportunity on the non-rejected part of the distribution D, i.e., $\Pr\left(h^{\text{rej}}(X,A) = 1 | Y = 1, A = 0\right) = \Pr\left(h^{\text{rej}}(X,A) = 1 | Y = 1, A = 1\right)$. Theorem 5.2 shows that the optimal equal opportunity classifier on the biased distribution $\tilde{D}$ recovers h^{rej} on the non-rejected inputs, if the data distribution and bias parameters satisfy the same conditions as in Theorems 4.1 & 4.2.

Theorem 5.2. *Let D be an arbitrary distribution on $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$, with $\mathcal{A} = \{0,1\}$ and $\mathcal{Y} = \{0,1\}$. Let $\delta \in [0, 1/2)$ be the rejection penalty and suppose the optimal reject option classifier h^{rej} defined above satisfies equal opportunity on the non-rejected part of distribution D. Let $\tilde{D}$ be a biased version of D defined as in Example 3.1 using bias parameters $\beta_p, \beta_n, \nu \in (0,1)$, and let $\tilde{h}_{EO}$ be the optimal equal opportunity classifier on $\tilde{D}$. If the data distribution and bias parameters satisfy*

$$(1-r)(1-2\delta) + r\left((1-\delta)\beta_p(1-2\nu) - \delta\beta_n\right) > 0$$

and

$$(1-r)(1-2\delta) + r\left((1-\delta)\beta_n - \delta\beta_p(1-2\nu)\right) > 0,$$

then $\tilde{h}_{EO}(x,a) = h^{rej}(x,a)$ whenever $h^{rej}(x,a) \neq \perp$.

A complete proof of Theorem 5.2 can be found in Appendix A.5. Note that if we consider the entire distribution D instead of only the non-rejected inputs, then $\Pr\left(\tilde{h}_{EO}(X,A) \neq h^{\text{rej}}(X,A)\right) \leq \Pr\left(h^{\text{rej}}(X,A) = \perp\right)$. In other words, if $\Pr\left(h^{\text{rej}}(X,A) = \perp\right)$ is small, then $\tilde{h}_{EO}$ matches h^{rej} on distribution D with high probability.

6. Recovering Robust Hypothesis under Data Bias for Arbitrary Data Distributions and Arbitrary Hypothesis Classes

In this section, we remove the restriction on hypothesis class $\mathcal{H}$, assumed to be the class of all group-aware binary classifiers in Sections 4 & 5. Note that the characterization of optimal fair classifiers using group-aware thresholds on the regression function $\eta(x,a)$ plays an important role in our proofs from Sections 4 & 5. For an arbitrary hypothesis class $\mathcal{H}$, even the classifier $h^* \in \mathcal{H}$ that maximizes accuracy on D need not be a threshold classifier on $\eta(x,a)$. To work around this, we make an assumption that the optimal (and fair) classifier that we want to recover under data bias must be *robust* under small perturbations to the data distribution D. Our definition of ϵ-robustness is motivated by the linear fractional transformations of regression function observed in various data bias models earlier (see Section 3).

Let D be any distribution on $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$, with $\mathcal{A} = \{0,1\}$ and $\mathcal{Y} = \{0,1\}$. Let $\Pr(A = 0) = r$, $\Pr(A = 1) = 1 - r$, and let $\Pr(Y = 1 | A = 0) = \Pr(Y = 1 | A = 1) = q$. Let h^* be the most accurate classifier in $\mathcal{H}$, i.e.,

$$h^* = \operatorname*{argmax}_{h \in \mathcal{H}} \Pr\left(h(X,A) = Y\right)$$

$$= \underset{h \in \mathcal{H}}{\operatorname{argmax}} \underset{(X,A)}{\mathrm{E}} \left[h(X,A)(2\eta(X,A) - 1) \right].$$

Note that, for an arbitrary hypothesis class $\mathcal{H}$, the optimal h^* need not be a threshold classifier on $\eta(x,a)$. As in the previous sections, we assume that h^* satisfies equal opportunity on the distribution D, i.e., $\Pr(h^*(X,A) = 1 | Y = 1, A = 0) = \Pr(h^*(X,A) = 1 | Y = 1, A = 1)$.

Definition 6.1. We define $h^* \in \mathcal{H}$ to be *ϵ-robust* if, for any distribution D' with random data points (X, A, Y') s.t.

$$\eta'(x,a) = \Pr(Y' = 1 | X = x, A = a) = \frac{P_a \eta(x,a) + Q_a}{R_a \eta(x,a) + S_a},$$

with $S_a = 1, |R_a| \leq \epsilon, |Q_a| \leq \epsilon, 1 - \epsilon \leq P_a \leq 1 + \epsilon$, and $P_a S_a - Q_a R_a \geq 0$, the optimal classifier h^* remains unchanged, i.e., $h^* = h' = \underset{h \in \mathcal{H}}{\operatorname{argmax}} \Pr(h(X,A) = Y')$.

The above conditions give a scale-invariant proxy to say that the linear fractional transformation defined by P_a, Q_a, R_a, S_a is order-preserving and when it is appropriately scaled to make $S_a = 1$, it is ϵ-close to the identity transformation. Definition 6.1 says that the classifier h^* of maximum accuracy in $\mathcal{H}$ is robust to small near-identity perturbations of the data distribution D.

Now we are ready to state our result for recovering robust, fair hypothesis from biased data on arbitrary data distributions and arbitrary hypothesis classes.

Theorem 6.2. *For any distribution D and any hypothesis class $\mathcal{H}$, if the optimal classifier $h^* \in \mathcal{H}$ is ϵ-robust and the bias parameters β_p, β_n, ν satisfy $(1-\epsilon)\beta_n \leq \beta_p \leq (1+\epsilon)\beta_n$,*

$$r\left((1-\nu)\beta_p - (1-\epsilon)\beta_n\right) + \epsilon(1-r) \geq 0$$
$$\textit{and}$$
$$r\left((1+\epsilon)\beta_n - (1-\nu)\beta_p\right) + \epsilon(1-r) \geq 0,$$

then the optimal equal opportunity classifier from $\mathcal{H}$ on the biased distribution $\tilde{D}$ recovers the optimal classifier from $\mathcal{H}$ on the original distribution D, i.e., $\tilde{h}_{EO} \equiv h^*$.

We prove Theorem 6.2 in Appendix A.6. Our proof reuses the basic characterization of optimal fair classifiers using Lagrange multipliers, and although it uses the class probabilities $\eta(x,a)$'s in a crucial way, it circumvents the need for threshold-based arguments completely. Figure 2 illustrates the recovery conditions in Theorem 6.2, for reasonably chosen group proportion parameter r, label bias ν and hypothesis robustness parameter ϵ.

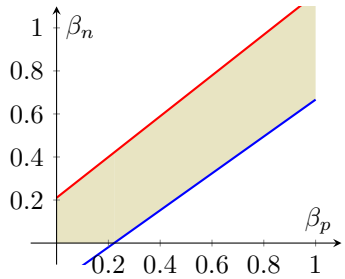

Figure 2. Recovery region for $\beta_p, \beta_n \in (0,1]$ given by the constraints $r\left((1-\nu)\beta_p - (1-\epsilon)\beta_n\right) + \epsilon(1-r) \geq 0$ and $r\left((1+\epsilon)\beta_n - (1-\nu)\beta_p\right) + \epsilon(1-r) \geq 0$ as in Theorem 6.2, when $r = 0.2$, $\nu = 0.1$, and $\epsilon = 0.05$. Even for arbitrary distributions and hypothesis classes, optimal and fair classifiers can be recovered from extreme under-representation and for a large range of data biases using just equal opportunity constraints.

7. Recovering from Time-Varying Data Bias

Data biases arise commonly in machine learning pipelines where data changes for downstream applications and over time. As an application of our techniques, we demonstrate how we can obtain recovery guarantees in data bias pipelines, i.e., when at every time step, we obtain a new shifted distribution. We model time-varying data bias as repeated applications of single-step data bias model (e.g., Example 3.1 used in previous sections).

Let $\tilde{D}_t$ be the biased data distribution obtained from an original distribution D, when the data bias model described in Example 3.1 gets applied repeatedly t times with possibly different bias parameters $(\beta_{p,i}, \beta_{n,i}, \nu_i)$ at i-th time step, and let $c_i = \dfrac{\beta_{n,i}}{\beta_{p,i}}$. Since the composition of linear fractional transformations remains a linear fractional transformation, we obtain the following generalization for how the regression function changes over time.

Proposition 7.1. *Let (X, A, Y) denote a random data point from any given distribution D and let $(X, A, \tilde{Y}_t)$ be a random data point from its corresponding biased distribution $\tilde{D}_t$ after applying the multi-stage time-varying data bias model described above. Let $\eta(x,a) = \Pr(Y = 1 | X = x, A = a)$ and $\tilde{\eta}_t(x,a) = \Pr\left(\tilde{Y}_t = 1 | X = x, A = a\right)$. Then*

$$\tilde{\eta}_t(x,0) = \frac{\eta(x,0)}{\displaystyle\sum_{i=1}^{t} \left(\frac{1-c_i}{1-\nu_i} \prod_{j=i+1}^{t} \frac{c_j}{1-\nu_j} \right) \eta + \prod_{i=1}^{t} \frac{c_i}{1-\nu_i}},$$

where $\displaystyle\prod_{j=t+1}^{t} \frac{c_j}{1-\nu_j} \stackrel{\text{def}}{=} 1.$

Note that $\tilde{D}_t$ conditioned on $A = 1$ remains unchanged, and therefore, $\tilde{\eta}_t(x,1) = \eta(x,1)$, since we are looking at the data bias model in Example 3.1. The proof of Proposition 7.1 is by simple induction on t. As a result of

Proposition 7.1, $\tilde{\eta}_t(x,a)$ can be expressed as another single-step data bias model that directly transforms $\eta(x,a)$ into $\tilde{\eta}_t(x,a)$ using a linear fractional transformation with $P = 0, Q = 0, R = \sum_{i=1}^{t}\left(\frac{1-c_i}{1-\nu_i}\prod_{j=i+1}^{t}\frac{c_j}{1-\nu_j}\right)$, and $S = \prod_{i=1}^{t}\frac{c_i}{1-\nu_i}$. The above P, Q, R, S obey the conditions in Proposition 3.4, so the corresponding linear fractional transformation is order-preserving.

7.1. Repeated Data Bias & Infinite Time Horizon

As a warm-up, we first study a simpler case where the bias parameters do not change at each time step, i.e., $\beta_{p,i} = \beta_p$, $\beta_{n,i} = \beta_n$, and $\nu_i = \nu$, for all $i \in [t]$). Equivalently, the same data bias model with the same bias parameters gets applied repeatedly t times. We work with the original distribution D described as in Theorem 4.2 for the ease of analysis, and assume that $\Pr(A=0) = r$, $\Pr(A=1) = 1-r$ and $\Pr(Y=1|A=0) = \Pr(Y=1|A=1) = q$. First, we show Theorem 7.2 about when the optimal equal opportunity classifier $\tilde{h}_{EO,t}$ on the biased distribution $\tilde{D}_t$ can recover h^* for the infinite time horizon as $t \to \infty$, and the necessary conditions on the data distribution and bias parameters to allow that. The proof of Theorem 7.2 is given in Appendix A.7.

Theorem 7.2. *Let $\tilde{D}_t$ be the biased distribution obtained by applying the bias model in Example 3.1 repeatedly t times with $\beta_{p,i} = \beta_p$, $\beta_{n,i} = \beta_n$, and $\nu_i = \nu$, for all $i \in [t]$ on a given distribution D defined as in Theorem 4.2. Let h^* be the Bayes optimal classifier on D and let $\tilde{h}_{EO,t}$ be the optimal equal opportunity classifier on $\tilde{D}_t$.*

- *If $\beta_n < 1$ then as $t \to \infty$, we have $\eta(x,0) \to 0$, for all $x \in \mathcal{X}$. Thus, we cannot have $\tilde{h}_{EO,t} \equiv h^*$ as $t \to \infty$.*
- *If $\beta_n = 1$ and $\tilde{h}_{EO,t} \equiv h^*$ as $t \to \infty$, then the data distribution and bias parameters must satisfy*

$$1 - \frac{(1-2\delta)}{(1-\delta)r} < \frac{\nu\beta_p}{1-\beta_p(1-\nu)} < 1 + \frac{(1-2\delta)}{\delta r}.$$

7.2. Time-Varying Data Bias Pipeline with Finite Steps

Now we generalize the necessary conditions in Theorem 4.2 for repeated application of the data bias model from Example 3.1, with possibly different bias parameters at each time step. We assume that the bias parameters can vary but are bounded.

Theorem 7.3. *Let D be a data distribution described as in Theorem 4.2 and $\tilde{D}_t$ be the resulting biased distribution after repeated application of the data bias model from Example 3.1 to D in t steps, with bounded but possibly different bias parameters in each step as $\beta_{n,t} \in [\beta_n, 1], \beta_{p,t} \in [\beta_p, 1]$ and $\nu_t \in [0, \nu]$, where $\nu < 1/2$. Let h^* be the Bayes optimal classifier on D and let $\tilde{h}_{EO,t}$ be the optimal equal opportunity classifier on $\tilde{D}_t$. If $\tilde{h}_{EO,t} \equiv h^*$, then the data distribution and bias parameters must satisfy*

$$\frac{(1-2\delta)(1-r)}{(1-\delta)r} - \frac{\delta}{1-\delta} > \frac{1}{\beta_n^t} - 2\beta_p^t(1-\nu)^t \quad \textit{and}$$

$$\frac{(1-2\delta)(1-r)}{\delta r} > (1-\nu)^t\beta_p^t(1-\beta_p)\frac{1-\beta_p^t(1-\nu)^t}{1-\beta_p(1-\nu)} - \frac{\beta_n^t}{\delta} - 2.$$

The proof of Theorem 7.3 is given in Appendix A.7. The first condition in Theorem 7.3 can be used to get the following upper bound on the time horizon up to which $\tilde{h}_{EO,t} \equiv h^*$ is possible.

Corollary 7.4. *The conditions in Theorem 7.3 above are satisfied only if $t < \frac{\log K_D}{\log(1/\beta_n)}$, where $K_D = \frac{(1-2\delta)(1-r)}{(1-\delta)r} - \frac{\delta}{1-\delta} + 2$.*

The above corollary can be obtained by noting that the term $2\beta_p^t(1-\nu)^t$ is upper bounded by 2 since it converges to 1, and any t greater than the value in the corollary violates the first inequality in Theorem 7.3. We derive similar time-varying bias recovery conditions for demographic parity in Theorem A.6 given in Appendix A.7.

8. Conclusion and Future Directions

In this paper, we investigate the phenomenon of using fair classification (in particular, equal opportunity constraints) to recover optimal and fair classifiers even from extremely biased version of the original data. We generalize the result of Blum & Stangl (2019) in many ways, for arbitrary distributions and arbitrary hypothesis classes, and develop techniques that are flexible and may be of independent interest in studying other fair classification problems and data bias models. Note that our approach based on Blum & Stangl (2019) does not require knowing, estimating, or correcting the data bias explicitly. Previous work has studied alternate approaches to get around data bias through reweighing and loss adjustment by estimating the extent of data bias (Biswas & Mukherjee, 2021; Dai & Brown, 2020; Wang et al., 2021). Rectifying data bias for fair classification in practice would require the best combination of both these approaches.

Given the flexibility of our technique, it would be interesting to investigate the applicability and limitations of our results to different data bias models (Dai & Brown, 2020; Wang et al., 2021; Biswas & Mukherjee, 2021; Konstantinov & Lampert, 2022; Blum et al., 2023) and a variety of fairness constraints (e.g., predictive parity (Zeng et al., 2022b)). A pragmatic future direction is to study the effect of data bias

on the sample complexity for fair classification (Donini et al., 2018; Tolbert & Diana, 2023).

An important question related to time-varying data bias models is to study the case when bias parameters at time t are a function of the model performance at time $t-1$, using ideas such as performative prediction (Perdomo et al., 2020; Brown et al., 2022)

Impact Statement

This paper presents work that aims to advance the field of Machine Learning. Our work has many potential societal consequences; pinpointing specific ones will be difficult here. Our work theoretically investigates the feasibility of using fair classification on extremely biased data as a method to recover optimal and fair classifiers on the original data. Historical, socio-cultural, implicit biases and other systematic biases in real-world data are results of complex interactions over time, and the simplistic data bias models studied in our work are insufficient to represent them truthfully. Our work underlines the need to study various possible ways to rectify data biases in algorithmic decision-making with societal consequences.

Acknowledgements

M.S. would like to thank Microsoft Research India for their support during the PhD through the Microsoft Research India Joint PhD fellowship.

References

Akpinar, N.-J., Nagireddy, M., Stapleton, L., Cheng, H.-F., Zhu, H., Wu, S., and Heidari, H. A sandbox tool to bias (stress)-test fairness algorithms. *arXiv preprint arXiv:2204.10233*, 2022.

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

Bartlett, P. L. and Wegkamp, M. H. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(8), 2008.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, October 2018. URL https://arxiv.org/abs/1810.01943.

Biswas, A. and Mukherjee, S. Ensuring fairness under prior probability shifts. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 414–424, 2021.

Biswas, A., Barman, S., Deshpande, A., and Sharma, A. Quantifying infra-marginality and its trade-off with group fairness. *arXiv preprint arXiv:1909.00982*, 2019.

Blum, A. and Stangl, K. Recovering from biased data: Can fairness constraints improve accuracy? In *Symposium on Foundations of Responsible Computing (FORC)*, 2019.

Blum, A., Okoroafor, P., Saha, A., and Stangl, K. On the vulnerability of fairness constrained learning to malicious noise. *arXiv preprint arXiv:2307.11892*, 2023.

Brown, G., Hod, S., and Kalemaj, I. Performative prediction in a stateful world. In *International Conference on Artificial Intelligence and Statistics*, pp. 6045–6061. PMLR, 2022.

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.

Charoenphakdee, N., Cui, Z., Zhang, Y., and Sugiyama, M. Classification with rejection based on cost-sensitive classification. In *International Conference on Machine Learning*, pp. 1507–1517. PMLR, 2021.

Cheong, J., Kalkan, S., and Gunes, H. Causal structure learning of bias for fair affect recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 340–349, 2023.

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.

Cortes, C., DeSalvo, G., and Mohri, M. Learning with rejection. In *Algorithmic Learning Theory: 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings 27*, pp. 67–82. Springer, 2016.

Dai, J. and Brown, S. M. Label bias, label shift: Fair machine learning with unreliable labels. In *NeurIPS 2020 Workshop on Consequential Decision Making in Dynamic Environments*, volume 12, 2020.

Devroye, L., Györfi, L., and Lugosi, G. *The Bayes Error*, pp. 9–20. Springer, 1996. ISBN 978-1-4612-0711-5. doi: 10.1007/978-1-4612-0711-5_2.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. S., and Pontil, M. Empirical risk minimization under fairness

constraints. *Advances in neural information processing systems*, 31, 2018.

Du, W. and Wu, X. Fair and robust classification under sample selection bias. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 2999–3003, 2021.

Dutta, S., Wei, D., Yueksel, H., Chen, P.-Y., Liu, S., and Varshney, K. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *International conference on machine learning*, pp. 2803–2813. PMLR, 2020.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pp. 214–226, 2012. ISBN 9781450311151.

Elkan, C. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pp. 973–978. Lawrence Erlbaum Associates Ltd, 2001.

Fogliato, R., Chouldechova, A., and G'Sell, M. Fairness evaluation in presence of biased noisy labels. In *International Conference on Artificial Intelligence and Statistics*, pp. 2325–2336. PMLR, 2020.

Franc, V., Prusa, D., and Voracek, V. Optimal strategies for reject option classifiers. *Journal of Machine Learning Research*, 24(11):1–49, 2023.

Ghosh, A., Kvitca, P., and Wilson, C. When fair classification meets noisy protected attributes. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 679–690, 2023.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 3323–3331, 2016. ISBN 9781510838819.

Islam, M. T., Fariha, A., Meliou, A., and Salimi, B. Through the data management lens: Experimental analysis and evaluation of fair classification. In *Proceedings of the 2022 International Conference on Management of Data*, pp. 232–246, 2022.

Jiang, H. and Nachum, O. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 702–712. PMLR, 2020.

Konstantinov, N. and Lampert, C. H. On the impossibility of fairness-aware learning from corrupted data. In *Algorithmic Fairness through the Lens of Causality and Robustness workshop*, pp. 59–83. PMLR, 2022.

Lamy, A., Zhong, Z., Menon, A. K., and Verma, N. Noise-tolerant fair classification. *Advances in Neural Information Processing Systems*, 32, 2019.

Madras, D., Pitassi, T., and Zemel, R. Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in Neural Information Processing Systems*, 31, 2018.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 349–358, 2019.

Maity, S., Mukherjee, D., Yurochkin, M., and Sun, Y. Does enforcing fairness mitigate biases caused by subpopulation shift? *Advances in Neural Information Processing Systems*, 34:25773–25784, 2021.

Massart, P. and Nédélec, É. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

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, 2018.

Perdomo, J., Zrnic, T., Mendler-Dünner, C., and Hardt, M. Performative prediction. In *International Conference on Machine Learning*, pp. 7599–7609. PMLR, 2020.

Plecko, D. and Bareinboim, E. Causal fairness analysis. *arXiv preprint arXiv:2207.11385*, 2022.

Rivest, R. and Sloan, R. A formal model of hierarchical concept-learning. *Information and Computation*, 114(1): 88–114, 1994.

Schreuder, N. and Chzhen, E. Classification with abstention but without disparities. In *Uncertainty in Artificial Intelligence*, pp. 1227–1236. PMLR, 2021.

Schrouff, J., Harris, N., Koyejo, O., Alabdulmohsin, I., Schnider, E., Opsahl-Ong, K., Brown, A., Roy, S., Mincu, D., Chen, C., et al. Maintaining fairness across distribution shift: do we have viable solutions for real-world applications? *arXiv preprint arXiv:2202.01034*, 2022.

Shah, A., Bu, Y., Lee, J. K., Das, S., Panda, R., Sattigeri, P., and Wornell, G. W. Selective regression under fairness criteria. In *International Conference on Machine Learning*, pp. 19598–19615. PMLR, 2022.

Sharma, M., Deshpande, A., and Shah, R. R. On testing and comparing fair classifiers under data bias. *arXiv preprint arXiv:2302.05906*, 2023.

Singh, S. and Khim, J. Optimal binary classification beyond accuracy. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=pm8Y8unXkkJ.

Sloan, R. H. *PAC Learning, Noise, and Geometry*, pp. 21–41. Birkhäuser Boston, 1996.

Tolbert, A. W. and Diana, E. Correcting underrepresentation and intersectional bias for fair classification. *arXiv preprint arXiv:2306.11112*, 2023.

Wang, J., Liu, Y., and Levy, C. Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 526–536, 2021.

Wick, M., Tristan, J.-B., et al. Unlocking fairness: a trade-off revisited. *Advances in neural information processing systems*, 32, 2019.

Zeng, X., Dobriban, E., and Cheng, G. Bayes-optimal classifiers under group fairness. *arXiv preprint arXiv:2202.09724*, 2022a.

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

Zhu, J., Galhotra, S., Sabri, N., and Salimi, B. Consistent range approximation for fair predictive modeling. *Proceedings of the VLDB Endowment*, 16(11):2925–2938, 2023.

A. Proofs

A.1. Proof of Proposition 3.4

Proof. $\tilde{\eta}(x_1,a) \leq \tilde{\eta}(x_2,a)$ iff $(P\eta(x_1,a)+Q)(R\eta(x_2,a)+S) \leq (P\eta(x_2,a)+Q)(R\eta(x_1,a)+S)$, using $R\eta(x,a)+S \geq 0$, for all $0 \leq \eta(x,a) \leq 1$. By canceling the common terms, the above inequality holds iff $(PS-QR)\eta(x_1,a) \leq (PS-QR)\eta(x_2,a)$, or equivalently $\eta(x_1,a) \leq \eta(x_2,a)$ because $PS-QR \geq 0$. □

A.2. Derivations for linear fractional transforms of Data Bias models

Proposition A.1. *Consider a biased distribution $\tilde{D}$ obtained from the original distribution D by the following process defined by bias parameters $\beta_p, \beta_n, \nu \in (0,1)$ (Blum & Stangl, 2019). A random data point (X,A,Y) from D with $A=1$ remains unchanged, the points with $A=0, Y=1$ have an independent survival probability of β_p, and the points with $A=0, Y=0$ have an independent survival probability of β_n. Finally, the survived points with $A=0, Y=1$ keep their class label 1 with probability $1-\nu$, and it gets flipped to 0 with probability ν. For the privileged group $A=1$, we have $\tilde{\eta}(x,1)=\eta(x,1)$, for all $x \in \mathcal{X}$, whereas for the underprivileged group $A=0$ we have $\tilde{\eta}(x,0) = \frac{(1-\nu)\eta(x,0)}{(1-c)\eta(x,0)+c}$, where $c = \frac{\beta_n}{\beta_p}$..*

Proof.

$$\begin{aligned}
\tilde{\eta}(x,0) &= \Pr\left(\tilde{Y}=1|X=x,A=0\right) \\
&= \frac{(1-\nu)\beta_p\eta(x,0)}{\beta_p\eta(x,0)+\beta_n(1-\eta(x,0))} \\
&= \frac{(1-\nu)\beta_p\eta(x,0)}{(\beta_p-\beta_n)\eta(x,0)+\beta_n} \\
&= \frac{(1-\nu)\eta(x,0)}{(1-c)\eta(x,0)+c}, \qquad \text{where } c = \frac{\beta_n}{\beta_p}.
\end{aligned}$$

Since no bias acts on group $A=1$, $\tilde{\eta}(x,1)=\eta(x,1)$. □

Proposition A.2. *Consider a biased distribution $\tilde{D}$ obtained form the original distribution D by introducing a group dependent label flip rate (Dai & Brown, 2020; Wang et al., 2021): $\epsilon_a^1 = \Pr\left(\tilde{Y}=0|Y=1,A=a\right)$ and $\epsilon_a^0 = \Pr\left(\tilde{Y}=1|Y=0,A=a\right)$, with $0 \leq \epsilon_a^1+\epsilon_a^0 < 1$. Then, the observed labels in $\tilde{D}$ obey the following relationship: $\tilde{y}_i = y_i$ with $1-\epsilon_{a_i}^{\mathbb{I}(y_i=1)}$ probability and $\tilde{y}_i = \neg y_i$ with probability $\epsilon_{a_i}^{\mathbb{I}(y_i=1)}$. In terms of a linear fractional transform:*

$$\tilde{\eta}(x,a) = (1-\epsilon_a^1-\epsilon_a^0)\eta(x,a)+\epsilon_a^0$$

Proof.

$$\begin{aligned}
\tilde{\eta}(x,0) &= \Pr\left(\tilde{Y}=1|X=x,A=0\right) \\
&= \frac{\Pr\left(\tilde{Y}=1,Y=0,X=x,A=a\right)+\Pr\left(\tilde{Y}=1,Y=1,X=x,A=a\right)}{\Pr\left(X=x,A=a\right)} \\
&= \Pr\left(\tilde{Y}=1|Y=0,A=a\right)(1-\eta(x,a))+\Pr\left(\tilde{Y}=1|Y=1,A=a\right)\eta(x,a) \\
&= (1-\epsilon_a^1-\epsilon_a^0)\eta(x,a)+\epsilon_a^0
\end{aligned}$$

□

Proposition A.3. *(Biswas & Mukherjee, 2021) Consider a biased distribution $\tilde{D}$ obtained from the original distribution D by introducing a group-dependent prior probability shift such that $\tilde{A}=A$ and $\Pr\left(\tilde{X}=x|\tilde{Y}=i,A=a\right) = \Pr\left(X=x|Y=i,A=a\right)$, for any $i,a \in \{0,1\}$, but $\Pr\left(\tilde{Y}=i|A=a\right) \neq \Pr\left(Y=i|A=a\right)$. For this data bias model, we prove that $\tilde{\eta}(x,a) = \frac{\eta(x,a)}{(1-\alpha)\eta(x,a)+\alpha}$, where $\alpha = \frac{\Pr\left(\tilde{Y}=0|A=a\right)\Pr\left(Y=1|A=a\right)}{\Pr\left(\tilde{Y}=1|A=a\right)\Pr\left(Y=0|A=a\right)}$.*

Proof.

$$\begin{aligned}\tilde{\eta}(x,a) &= \Pr\left(\tilde{Y}=1|\tilde{X}=x,A=a\right)\\ &= \frac{\Pr\left(\tilde{Y}=1,\tilde{X}=x,A=a\right)}{\Pr\left(\tilde{X}=x,A=a\right)}\end{aligned}$$

Since $\Pr\left(X=x|\tilde{Y}=1,A=a\right)=\Pr\left(X=x|Y=1,A=a\right)$ from the definition of the bias model, we can write:

$$\begin{aligned}&= \frac{\Pr\left(X=x|\tilde{Y}=1,A=a\right)\Pr\left(\tilde{Y}=1|A=a\right)}{\Pr\left(X=x|\tilde{Y}=1,A=a\right)\Pr\left(\tilde{Y}=1|A=a\right)+\Pr\left(X=x|\tilde{Y}=0,A=a\right)\Pr\left(\tilde{Y}=0|A=a\right)}\\ &= \frac{1}{1+\dfrac{\Pr\left(X=x|\tilde{Y}=0,A=a\right)\Pr\left(\tilde{Y}=0|A=a\right)}{\Pr\left(X=x|\tilde{Y}=1,A=a\right)\Pr\left(\tilde{Y}=1|A=a\right)}}.\end{aligned}$$

Thus,

$$\frac{1}{\tilde{\eta}(x,a)}-1=\frac{\Pr\left(X=x|\tilde{Y}=0,A=a\right)\Pr\left(\tilde{Y}=0|A=a\right)}{\Pr\left(X=x|\tilde{Y}=1,A=a\right)\Pr\left(\tilde{Y}=1|A=a\right)}.$$

Similarly,

$$\frac{1}{\eta(x,a)}-1=\frac{\Pr\left(X=x|Y=0,A=a\right)\Pr\left(Y=0|A=a\right)}{\Pr\left(X=x|Y=1,A=a\right)\Pr\left(Y=1|A=a\right)}.$$

Hence,

$$\frac{1}{\tilde{\eta}(x,a)}-1=\alpha\left(\frac{1}{\eta(x,a)}-1\right),\quad \text{where } \alpha=\frac{\Pr\left(\tilde{Y}=0|A=a\right)\Pr\left(Y=1|A=a\right)}{\Pr\left(\tilde{Y}=1|A=a\right)\Pr\left(Y=0|A=a\right)}.$$

In other words,

$$\tilde{\eta}(x,a)=\frac{\eta(x,a)}{(1-\alpha)\eta(x,a)+\alpha}.$$

□

A.3. Proofs for Section 3.2

Given any classifier $h:\mathcal{X}\times\mathcal{A}\to\mathcal{Y}$, let $TPR_a(h)$ and $\widetilde{TPR}_a(h)$ denote its true positive rates conditioned on $A=a$ for distributions D and $\tilde{D}$, respectively. Then

$$TPR_a(h)=\frac{\Pr\left(h(X,a)=1,A=a,Y=1\right)}{\Pr\left(Y=1,A=a\right)}=\frac{\underset{X|A=a}{\mathrm{E}}\left[h(X,a)\eta(X,a)\right]}{\underset{X|A=a}{\mathrm{E}}\left[\eta(X,a)\right]},$$

and

$$\widetilde{TPR}_a(h)=\frac{\Pr\left(h(X,a)=1,A=a,\tilde{Y}=1\right)}{\Pr\left(\tilde{Y}=1,A=a\right)}=\frac{\underset{X|A=a}{\mathrm{E}}\left[h(X,a)\tilde{\eta}(X,a)\right]}{\underset{X|A=a}{\mathrm{E}}\left[\tilde{\eta}(X,a)\right]}.$$

We can now prove Proposition 3.5.

Proof. **(Proof of Proposition 3.5)**

$$
\begin{aligned}
&\widetilde{TPR}_a(h) = \Pr\left(h(X,a)=1|\tilde{Y}=1, A=a\right)\\
&= \frac{\Pr\left(h(X,a)=1, \tilde{Y}=1|A=a\right)}{\Pr\left(\tilde{Y}=1|A=a\right)}\\
&= \frac{\Pr\left(Y=0|A=a\right)\Pr\left(h(X,a)=1, \tilde{Y}=1|Y=0, A=a\right)}{\Pr\left(\tilde{Y}=1|A=a\right)}\\
&\qquad + \frac{\Pr\left(Y=1|A=a\right)\Pr\left(h(X,a)=1, \tilde{Y}=1|Y=1, A=a\right)}{\Pr\left(\tilde{Y}=1|A=a\right)}\\
&= \frac{\Pr\left(Y=0|A=a\right)\Pr\left(\tilde{Y}=1|h(X,a)=1, Y=0, A=a\right)\Pr\left(h(X,a)=1|Y=0, A=a\right)}{\Pr\left(\tilde{Y}=1|A=a\right)}\\
&\qquad + \frac{\Pr\left(Y=1|A=a\right)\Pr\left(\tilde{Y}=1|h(X,a)=1, Y=1, A=a\right)\Pr\left(h(X,a)=1|Y=1, A=a\right)}{\Pr\left(\tilde{Y}=1|A=a\right)}\\
&= \frac{\Pr\left(Y=0|A=a\right)\Pr\left(\tilde{Y}=1|Y=0, A=a\right)\Pr\left(h(X,a)=1|Y=0, A=a\right)}{\Pr\left(\tilde{Y}=1|A=a\right)}\\
&\qquad + \frac{\Pr\left(Y=1|A=a\right)\Pr\left(\tilde{Y}=1|Y=1, A=a\right)\Pr\left(h(X,a)=1|Y=1, A=a\right)}{\Pr\left(\tilde{Y}=1|A=a\right)}\\
&\qquad\qquad \text{because } \tilde{Y} \text{ depends only on } A \text{ and } Y\\
&= \frac{\Pr\left(\tilde{Y}=1, Y=0|A=a\right)}{\Pr\left(\tilde{Y}=1|A=a\right)}\, FPR_a(h) + \frac{\Pr\left(\tilde{Y}=1, Y=1|A=a\right)}{\Pr\left(\tilde{Y}=1|A=a\right)}\, TPR_a(h)\\
&= \Pr\left(Y=0|\tilde{Y}=1, A=a\right)\, FPR_a(h) + \Pr\left(Y=1|\tilde{Y}=1, A=a\right)\, TPR_a(h).
\end{aligned}
$$

Similarly, we show that $\widetilde{TNR}_a(h)$ can be written as a linear combination of $TNR_a(h)$ and $FNR_a(h)$.

$$
\begin{aligned}
&\widetilde{TNR}_a(h)\\
&= \Pr\left(h(X,a)=a|\tilde{Y}=0, A=a\right)\\
&= \frac{\Pr\left(h(X,a)=0, \tilde{Y}=0|A=a\right)}{\Pr\left(\tilde{Y}=0|A=a\right)}\\
&= \frac{\Pr\left(Y=0|A=a\right)\Pr\left(h(X,a)=0, \tilde{Y}=0|Y=0, A=a\right)}{\Pr\left(\tilde{Y}=0|A=a\right)}\\
&\qquad + \frac{\Pr\left(Y=1|A=a\right)\Pr\left(h(X,a)=0, \tilde{Y}=0|Y=1, A=a\right)}{\Pr\left(\tilde{Y}=0|A=a\right)}
\end{aligned}
$$

$$= \frac{\Pr\left(Y=0|A=a\right)\Pr\left(\tilde{Y}=0|h(X,a)=0,Y=0,A=a\right)\Pr\left(h(X,a)=0|Y=0,A=a\right)}{\Pr\left(\tilde{Y}=0|A=a\right)}$$
$$+\frac{\Pr\left(Y=1|A=a\right)\Pr\left(\tilde{Y}=0|h(X,a)=0,Y=1,A=a\right)\Pr\left(h(X,a)=0|Y=1,A=a\right)}{\Pr\left(\tilde{Y}=0|A=a\right)}$$
$$= \frac{\Pr\left(Y=0|A=a\right)\Pr\left(\tilde{Y}=0|Y=0,A=a\right)\Pr\left(h(X,a)=0|Y=0,A=a\right)}{\Pr\left(\tilde{Y}=0|A=a\right)}$$
$$+\frac{\Pr\left(Y=1|A=a\right)\Pr\left(\tilde{Y}=0|Y=1,A=a\right)\Pr\left(h(X,a)=0|Y=1,A=a\right)}{\Pr\left(\tilde{Y}=0|A=a\right)}$$

because $\tilde{Y}$ depends only on A and Y

$$= \frac{\Pr\left(\tilde{Y}=0,Y=0|A=a\right)}{\Pr\left(\tilde{Y}=0|A=a\right)}\,TNR_a(h) + \frac{\Pr\left(\tilde{Y}=0,Y=1|A=a\right)}{\Pr\left(\tilde{Y}=0|A=a\right)}\,FNR_a(h)$$
$$= \Pr\left(Y=0|\tilde{Y}=0,A=a\right)\,TNR_a(h) + \Pr\left(Y=1|\tilde{Y}=0,A=a\right)\,FNR_a(h).$$

□

Proof. **(Proof of Proposition 3.8)** We begin with a known technique that uses Lagrange duality to give threshold-based characterization of optimal equal opportunity classifiers (Menon & Williamson, 2018; Chzhen et al., 2019). We define $\tilde{h}_{EO} = \underset{h\in\tilde{\mathcal{H}}_{\text{fair, EO}}}{\operatorname{argmax}} \Pr\left(h(X,A)=\tilde{Y}\right)$. By Lagrange duality, we can write

$$\tilde{h}_{EO} = \underset{h\in\mathcal{H}}{\operatorname{argmax}} \min_{\lambda\in\mathbb{R}} \Pr\left(h(X,A)=\tilde{Y}\right) + \lambda\Pr\left(h(X,A)=1|\tilde{Y}=1,A=0\right) - \lambda\Pr\left(h(X,A)=1|\tilde{Y}=1,A=1\right)$$
$$= \underset{h\in\mathcal{H}}{\operatorname{argmax}} \min_{\lambda\in\mathbb{R}} \sum_{a=0}^{1}\Pr\left(A=a\right)\underset{X|A=a}{\mathrm{E}}\left[h(X,a)(2\tilde{\eta}(X,a)-1)\right] + \frac{(-1)^{\mathbb{I}(a=1)}\,\lambda}{\underset{X|A=a}{\mathrm{E}}\left[\tilde{\eta}(X,a)\right]}\underset{X|A=a}{\mathrm{E}}\left[h(X,a)\tilde{\eta}(X,a)\right]$$
$$= \underset{h\in\mathcal{H}}{\operatorname{argmax}} \min_{\lambda\in\mathbb{R}} \sum_{a=0}^{1}\Pr\left(A=a\right)\underset{X|A=a}{\mathrm{E}}\left[h(X,a)\left(\left(2+\frac{(-1)^{\mathbb{I}(a=1)}\lambda}{\Pr\left(\tilde{Y}=1,A=a\right)}\right)\tilde{\eta}(X,a)-1\right)\right]$$
$$= \min_{\lambda\in\mathbb{R}} \underset{h\in\mathcal{H}}{\operatorname{argmax}} \sum_{a=0}^{1}\Pr\left(A=a\right)\underset{X|A=a}{\mathrm{E}}\left[h(X,a)\left(\left(2+\frac{(-1)^{\mathbb{I}(a=1)}\lambda}{\Pr\left(\tilde{Y}=1,A=a\right)}\right)\tilde{\eta}(X,a)-1\right)\right],$$

by Sion's minimax theorem as the objective is linear in λ and h. Thus, there exists $\lambda^*\in\mathbb{R}$ such that the objective on h splits group-wise with the optimal solution given by

$$\tilde{h}_{EO}(x,a) = \mathbb{I}\left(\tilde{\eta}(x,a) \geq \frac{1}{2+\frac{(-1)^{\mathbb{I}(a=1)}\lambda^*}{\Pr\left(\tilde{Y}=1,A=a\right)}}\right)$$

$$= \begin{cases} \mathbb{I}\left(\dfrac{(1-\nu)\eta(x,0)}{(1-c)\eta(x,0)+c} \geq \dfrac{1}{2+\dfrac{\lambda^*}{\Pr\left(\tilde{Y}=1,A=0\right)}}\right), & \text{for } a=0 \qquad \text{(Example 3.1)} \\ \mathbb{I}\left(\eta(x,1) \geq \dfrac{1}{2-\dfrac{\lambda^*}{\Pr\left(Y=1,A=1\right)}}\right), & \text{for } a=1 \end{cases}$$

$$= \begin{cases} \mathbb{I}\left(\eta(x,0) \geq \dfrac{c}{1-2\nu+c+\dfrac{\lambda^*(1-\nu)}{\Pr\left(\tilde{Y}=1,A=0\right)}}\right), & \text{for } a=0 \\ \mathbb{I}\left(\eta(x,1) \geq \dfrac{1}{2-\dfrac{\lambda^*}{\Pr\left(Y=1,A=1\right)}}\right), & \text{for } a=1 \end{cases}$$

Let $\Pr(A=0) = r$ and $\Pr(A=1) = 1-r$, and let $\Pr(Y=1|A=0) = \Pr(Y=1|A=1) = q$. Then $\Pr(Y=1,A=1) = \Pr(A=1)\Pr(Y=1|A=1) = (1-r)q$ and

$$\begin{aligned} \Pr\left(\tilde{Y}=1,A=0\right) &= \Pr(A=0)\Pr(Y=1|A=0)\Pr\left(\tilde{Y}=1|Y=1,A=0\right) \\ &\quad + \Pr(A=0)\Pr(Y=0|A=0)\Pr\left(\tilde{Y}=1|Y=0,A=0\right) \\ &= rq\beta_p(1-\nu). \end{aligned}$$

Therefore,

$$\tilde{h}_{EO}(x,a) = \begin{cases} \mathbb{I}\left(\eta(x,0) \geq \dfrac{c}{1-2\nu+c+\dfrac{\lambda^*}{\beta_p rq}}\right), & \text{for } a=0 \\ \mathbb{I}\left(\eta(x,1) \geq \dfrac{1}{2-\dfrac{\lambda^*}{(1-r)q}}\right), & \text{for } a=1 \end{cases}.$$

Equivalently, we can also write

$$\tilde{h}_{EO}(x,a) = \begin{cases} \mathbb{I}\left(\eta(x,0) \geq \dfrac{1}{1+\dfrac{1-2\nu}{c}+\dfrac{\lambda^*}{\beta_n rq}}\right), & \text{for } a=0 \\ \mathbb{I}\left(\eta(x,1) \geq \dfrac{1}{2-\dfrac{\lambda^*}{(1-r)q}}\right), & \text{for } a=1. \end{cases}.$$

□

We can derive similar results for demographic parity (Menon & Williamson, 2018; Chzhen et al., 2019).

Proposition A.4. *For any distribution D and its biased version $\tilde{D}$ described above, let $\tilde{h}_{DP}$ be a classifier of the maximum accuracy among all binary classifiers that satisfy demographic parity on $\tilde{D}$. Then there exists $\lambda^* \in \mathbb{R}$ such that*

$$\tilde{h}_{DP}(x,a) = \begin{cases} \mathbb{I}\left(\eta(x,0) \geq \dfrac{c(r-\lambda^*)}{2r(1-\nu)-(1-c)(r-\lambda^*)}\right), & \text{for } a=0 \\ \mathbb{I}\left(\eta(x,1) \geq \dfrac{1}{2}+\dfrac{\lambda^*}{2(1-r)}\right), & \text{for } a=1 \end{cases}$$

Proof. Similar to previous work that gives a threshold-based characterization of optimal demographic parity classifiers using Lagrange duality (Menon & Williamson, 2018), we use the same idea on the biased distribution. $\tilde{h}_{DP} = \underset{h\in\widetilde{\mathcal{H}}_{\text{fair, DP}}}{\operatorname{argmax}} \Pr\left(h(X,A)=\tilde{Y}\right)$. By Lagrange duality, we can write

$$\begin{aligned}
\tilde{h}_{DP} &= \underset{h\in\mathcal{H}}{\operatorname{argmax}} \min_{\lambda\in\mathbb{R}} \Pr\left(h(X,A)=\tilde{Y}\right) + \lambda \Pr\left(h(X,A)=1|A=0\right) - \lambda \Pr\left(h(X,A)=1|A=1\right) \\
&= \underset{h\in\mathcal{H}}{\operatorname{argmax}} \min_{\lambda\in\mathbb{R}} \sum_{a=0}^{1} \Pr\left(A=a\right) \mathop{\mathrm{E}}_{X|A=a}\left[h(X,a)(2\tilde{\eta}(X,a)-1)\right] + (-1)^{\mathbb{I}(a=1)}\,\lambda \mathop{\mathrm{E}}_{X|A=a}\left[h(X,a)\right] \\
&= \underset{h\in\mathcal{H}}{\operatorname{argmax}} \min_{\lambda\in\mathbb{R}} \sum_{a=0}^{1} \Pr\left(A=a\right) \mathop{\mathrm{E}}_{X|A=a}\left[h(X,a)\left(2\tilde{\eta}(X,a)-1+\frac{(-1)^{\mathbb{I}(a=1)}\,\lambda}{\Pr\left(A=a\right)}\right)\right] \\
&= \min_{\lambda\in\mathbb{R}} \underset{h\in\mathcal{H}}{\operatorname{argmax}} \sum_{a=0}^{1} \Pr\left(A=a\right) \mathop{\mathrm{E}}_{X|A=a}\left[h(X,a)\left(2\tilde{\eta}(X,a)-1+\frac{(-1)^{\mathbb{I}(a=1)}\,\lambda}{\Pr\left(A=a\right)}\right)\right],
\end{aligned}$$

by Sion's minimax theorem as the objective is linear in λ and h. Thus, there exists some optimal $\lambda^* \in \mathbb{R}$ such that the objective on h splits group-wise with the optimal solution given by

$$\begin{aligned}
\tilde{h}_{DP}(x,a) &= \mathbb{I}\left(\tilde{\eta}(x,a) \geq \frac{1}{2} - \frac{(-1)^{\mathbb{I}(a=1)}\lambda^*}{2\Pr\left(A=a\right)}\right) \\
&= \begin{cases} \mathbb{I}\left(\dfrac{(1-\nu)\eta(x,0)}{(1-c)\eta(x,0)+c} \geq \dfrac{1}{2} - \dfrac{\lambda^*}{2\Pr\left(A=0\right)}\right), & \text{for } a=0 \quad \text{(Example 3.1)} \\ \mathbb{I}\left(\eta(x,1) \geq \dfrac{1}{2}+\dfrac{\lambda^*}{2\Pr\left(A=1\right)}\right), & \text{for } a=1 \end{cases} \\
&= \begin{cases} \mathbb{I}\left(\eta(x,0) \geq \dfrac{c(r-\lambda^*)}{2r(1-\nu)-(1-c)(r-\lambda^*)}\right), & \text{for } a=0 \\ \mathbb{I}\left(\eta(x,1) \geq \dfrac{1}{2}+\dfrac{\lambda^*}{2(1-r)}\right), & \text{for } a=1 \end{cases}
\end{aligned}$$

□

A.4. Proofs for Section 4.1

For the distribution D described in Section 4.1, the following result holds:

Lemma A.5. *Let $h : \mathcal{X}\times\mathcal{A} \to \{0,1\}$ be a deterministic classifier. Let $Y|X=x, A=a$ be a random variable that takes value $h(x,a)$ with probability $1-\delta(x,a)$ and $\neg h(x,a)$ with probability $\delta(x,a)$, for some $\delta(x,a) \leq \delta < 1/2$. Let $\eta(x,a) = \Pr\left(Y=1|X=x,A=a\right)$ and let $g(x,a)$ be a threshold classifier given by $g(x,a) = \mathbb{I}\left(\eta(x,a) \geq t_a\right)$, using group-dependent thresholds t_0 and t_1, respectively. If $t_0, t_1 \in (\delta, 1-\delta)$ then $g \equiv h \equiv h^*$, where h^* is the Bayes optimal classifier for the joint distribution on $\mathcal{X}\times\mathcal{A}\times\mathcal{Y}$.*

Proof. It is folklore that the (group-aware) Bayes optimal classifier is given by $h^*(x,a) = \mathbb{I}\left(\eta(x,a) \geq 1/2\right)$ (Zeng et al., 2022b). One of the ways by which we can recover the Bayes Optimal classifier with the biased distribution $\tilde{D}$ is when the Bayes optimal classifier on the original distribution D does not change around the vicinity of $1/2$. Consider the following two cases for $h(x,a)$. If $h(x,a)=0$, then $Y|X=x,A=a$ takes value 0 with probability $\delta(x,a) \leq \delta$, and

hence, $\mathbb{I}(\eta(x,a) \geq t_a) = \mathbb{I}(\delta(x,a) \geq t_a) = \mathbb{I}(\delta(x,a) \geq 1/2) = 0$, for $\delta(x,a) \leq \delta < 1/2$ and $t_a \in (\delta, 1-\delta)$. On the other hand, if $h(x,a) = 1$, then $Y|X = x, A = a$ takes value 1 with probability $1 - \delta(x,a) \geq 1 - \delta$, and hence, $\mathbb{I}(\eta(x,a) \geq t_a) = \mathbb{I}(1 - \delta(x,a) \geq t_a) = \mathbb{I}(1 - \delta(x,a) \geq 1/2) = 1$, for $\delta(x,a) \leq \delta < 1/2$ and $t_a \in (\delta, 1-\delta)$. Therefore, $g \equiv h \equiv h^*$. □

We now describe the proof of Theorem 4.3. The proof for Theorem 4.2 is clubbed with the proof of Theorem 4.1 later in this section.

Proof. **(Proof of Theorem 4.3)** We use the characterization optimal demographic parity classifier obtained in Proposition A.4 for $\tilde{D}$, i.e., there exists $\lambda^* \in \mathbb{R}$ such that

$$\tilde{h}_{DP}(x,a) = \begin{cases} \mathbb{I}\left(\eta(x,0) \geq \dfrac{c(r-\lambda^*)}{2r(1-\nu) - (1-c)(r-\lambda^*)}\right), & \text{for } a = 0 \\ \mathbb{I}\left(\eta(x,1) \geq \dfrac{1}{2} + \dfrac{\lambda^*}{2(1-r)}\right), & \text{for } a = 1. \end{cases}$$

Step 1: If the given conditions on the bias parameters hold, then there exists a $\lambda \in \mathbb{R}$ such that $h_\lambda = h^*$. From Lemma A.5, we know that the threshold on $\eta(x,1)$ lies in the interval $(\delta, 1-\delta)$ if and only if $\delta < \dfrac{1}{2} + \dfrac{\lambda^*}{2(1-r)} < 1 - \delta$, or equivalently, $(2\delta - 1)(1-r) < \lambda^* < (1-2\delta)(1-r)$. Similarly, the threshold on $\eta(x,0)$ lies in the interval $(\delta, 1-\delta)$ if and only if $\delta < \dfrac{c(r-\lambda^*)}{2r(1-\nu) - (1-c)(r-\lambda^*)} < 1 - \delta$, or equivalently,

$$r\left(1 - \frac{2(1-\delta)(1-\nu)}{c + (1-c)(1-\delta)}\right) < \lambda^* < r\left(1 - \frac{2\delta(1-\nu)}{\delta(1-c) + c}\right).$$

There exists λ^* that simultaneously satisfies both sets of constraints given above if and only if

$$r\left(1 - \frac{2(1-\delta)(1-\nu)}{c + (1-c)(1-\delta)}\right) < (1-2\delta)(1-r) \text{ and } (2\delta-1)(1-r) < r\left(1 - \frac{2\delta(1-\nu)}{\delta(1-c) + c}\right).$$

Using $c = \beta_n/\beta_p$, the above conditions can be simplified as

$$\begin{aligned} &\beta_p(1-\delta)(1 - 2\delta - 2r(\nu - \delta)) + \delta\beta_n(1 - 2\delta - 2r(1-\delta)) > 0 \text{ and} \\ &\beta_p\delta(1 - 2r(1-\nu) - 2\delta(1-r)) + (1-\delta)\beta_n(1 - 2\delta(1-r)) > 0 \end{aligned}$$

Step 2: If the given conditions on the bias parameters hold, then there cannot exist any $\lambda \in \mathbb{R}$ such that both the group-wise thresholds applied to $\eta(x,a)$ in h_λ are at most δ, or both the thresholds are at least $1-\delta$. (Thresholds lying on the same side) The threshold on $\eta(x,1)$ is at most δ if and only if $\dfrac{1}{2} + \dfrac{\lambda}{2(1-r)} \leq \delta$, or equivalently, $\lambda \leq (2\delta - 1)(1-r)$. Similarly, the threshold on $\eta(x,0)$ is at most δ if and only if $\dfrac{c(r-\lambda)}{2r(1-\nu) - (1-c)(r-\lambda)} \leq \delta$, or equivalently, $\lambda \geq r\left(1 - \dfrac{2\delta\beta_p(1-\nu)}{\beta_n(1-\delta) + \delta\beta_p}\right)$. If both were to hold simultaneously, then $r\left(1 - \dfrac{2\delta\beta_p(1-\nu)}{\beta_n(1-\delta) + \delta\beta_p}\right) \leq (2\delta-1)(1-r)$, or equivalently, $\beta_p\delta(1 - 2r(1-\nu) - 2\delta(1-r)) + (1-\delta)\beta_n(1 - 2\delta(1-r)) \leq 0$, which would violate the second condition in our theorem.

On the other hand, the threshold on $\eta(x,1)$ is at least interval $1-\delta$ if and only if $\dfrac{1}{2} + \dfrac{\lambda}{2(1-r)} \geq 1 - \delta$, or equivalently, $\lambda \geq (1-2\delta)(1-r)$. Similarly, the threshold on $\eta(x,0)$ is at least $\dfrac{c(r-\lambda)}{2r(1-\nu) - (1-c)(r-\lambda)} \geq 1 - \delta$, or equivalently, $\lambda \leq r\left(1 - \dfrac{2(1-\delta)(1-\nu)}{c + (1-c)(1-\delta)}\right)$. If both were to hold simultaneously, then $(1-2\delta)(1-r) \leq r\left(1 - \dfrac{2(1-\delta)(1-\nu)}{c + (1-c)(1-\delta)}\right)$, or equivalently, $\beta_p(1-\delta)(1 - 2\delta - 2r(\nu - \delta)) + \delta\beta_n(1 - 2\delta - 2r(1-\delta)) \leq 0$, which would violate the first condition in our theorem.

Step 3: For any λ, if the group-wise thresholds on $\eta(x,a)$ in h_λ have one of them at most δ and another at least $1-\delta$, then h_λ cannot satisfy demographic parity unless $h_\lambda = h^*$. (Thresholds lying on opposite sides) We know that h^* satisfies demographic parity. Let $PR_a(h_\lambda) = \Pr(h_\lambda = 1 | A = a)$ (group positive rate). For demographic parity, we require $PR_0 = PR_1$. Suppose the threshold of h_λ on $\eta(x,0)$(call it t_0) and the threshold on $\eta(x,1)$(call it t_1) are separated by either δ or $1-\delta$. WLOG assume $t_0 < t_1$. Suppose $t_0 \leq \delta \leq t_1$. Since there is no input (x,a) with $\eta(x,a) \in (\delta, 1-\delta)$, we get $PR_0(\tilde{h}_\lambda) \geq PR_0(h^*)$ and $PR_1(h^*) \geq PR_1(\tilde{h}_\lambda)$. Since h^* satisfies demographic parity on the original distribution D, we have $PR_0(h^*) = PR_1(h^*)$. If h_λ satisfies demographic parity on the biased distribution $\tilde{D}$, we have $\widetilde{PR}_0(h_\lambda) = \widetilde{PR}_1(h_\lambda)$, and since $\widetilde{PR}_a(h_\lambda) = PR_a(h_\lambda)$ by Corollary 3.6, it implies $PR_0(h_\lambda) = PR_1(h_\lambda)$. Combining the two observations above, we get $PR_0(h_\lambda) = PR_0(h^*) = PR_1(h^*) = PR_1(h_\lambda)$, which is possible only if $h_\lambda \equiv h^*$ (as both are group-wise threshold classifiers). The other case $t_0 \leq 1-\delta \leq t_1$ can be argued similarly.

□

Now, we give a complete proof of Theorem 4.1. Note that Theorem 4.1 implies Theorem 4.2 as i.i.d. noise is a special case of Massart noise when we have $\delta(x,a) = \delta$, for all $(x,a) \in \mathcal{X} \times \mathcal{A}$. Thus, Theorem 4.1 is a stronger statement than Theorem 4.2 (the original recovery theorem of Blum & Stangl (Blum & Stangl, 2019)) while having the same necessary and sufficient conditions on the data and bias parameters.

Proof. **(Proof of Theorem 4.1)** Let

$$\tilde{h}_\lambda(x,a) = \begin{cases} \mathbb{I}\left(\eta(x,0) \geq \dfrac{1}{1 + \dfrac{1-2\nu}{c} + \dfrac{\lambda}{\beta_n r q}}\right), & \text{for } a = 0 \\ \mathbb{I}\left(\eta(x,1) \geq \dfrac{1}{2 - \dfrac{\lambda}{(1-r)q}}\right), & \text{for } a = 1. \end{cases}$$

From the characterization of optimal equal opportunity classifier shown earlier, there exists some optimal $\lambda^* \in \mathbb{R}$ such that $\tilde{h}_{EO} = h_{\lambda^*}$. We will show that if the conditions in Theorem 4.1 of Blum-Stangl hold, then the only possible choices left for $\lambda^* \in \mathbb{R}$ are the ones that give $h_{\lambda^*} = h^*$.

Step 1: If the given conditions on the bias parameters hold, then there exists a $\lambda \in \mathbb{R}$ such that $\tilde{h}_\lambda = h^*$. The threshold on $\eta(x,1)$ lies in the interval $(\delta, 1-\delta)$ if and only if $\dfrac{1}{1-\delta} < 2 - \dfrac{\lambda}{(1-r)q} < \dfrac{1}{\delta}$, or equivalently, $\dfrac{-(1-2\delta)(1-r)q}{\delta} < \lambda < \dfrac{(1-2\delta)(1-r)q}{1-\delta}$. Similarly, the threshold on $\eta(x,0)$ lies in the interval $(\delta, 1-\delta)$ if and only if $\dfrac{1}{1-\delta} < 1 + \dfrac{1-2\nu}{c} + \dfrac{\lambda}{\beta_n r q} < \dfrac{1}{\delta}$, or equivalently,

$$\beta_n r q \left(\frac{\delta}{1-\delta} - \frac{1-2\nu}{c}\right) < \lambda < \beta_n r q \left(\frac{1-\delta}{\delta} - \frac{1-2\nu}{c}\right).$$

There exists λ that simultaneously satisfies both sets of constraints given above if and only if

$$\beta_n r q \left(\frac{\delta}{1-\delta} - \frac{1-2\nu}{c}\right) < \frac{(1-2\delta)(1-r)q}{1-\delta}$$
$$\frac{-(1-2\delta)(1-r)q}{\delta} < \beta_n r q \left(\frac{1-\delta}{\delta} - \frac{1-2\nu}{c}\right).$$

Using $c = \beta_n/\beta_p$, the above conditions can be rewritten as

$$(1-r)(1-2\delta) + r\left((1-\delta)\beta_p(1-2\nu) - \delta\beta_n\right) > 0 \quad \text{and}$$
$$(1-r)(1-2\delta) + r\left((1-\delta)\beta_n - \delta\beta_p(1-2\nu)\right) > 0.$$

Step 2: If the given conditions on the bias parameters hold, then there cannot exist any $\lambda \in \mathbb{R}$ such that both the group-wise thresholds applied to $\eta(x,a)$ in h_λ are at most δ, or both the thresholds are at least $1-\delta$. (Thresholds lying on the same side) The threshold on $\eta(x,1)$ is at most δ if and only if $2 - \frac{\lambda}{(1-r)q} \geq \frac{1}{\delta}$, or equivalently, $\frac{-(1-2\delta)(1-r)q}{\delta} \geq \lambda$. Similarly, the threshold on $\eta(x,0)$ is at most δ if and only if $1 + \frac{1-2\nu}{c} + \frac{\lambda}{\beta_n rq} \geq \frac{1}{\delta}$, or equivalently, $\lambda \geq \beta_n rq\left(\frac{1-\delta}{\delta} - \frac{1-2\nu}{c}\right)$. If both were to hold simultaneously, then $\frac{-(1-2\delta)(1-r)q}{\delta} \geq \beta_n rq\left(\frac{1-\delta}{\delta} - \frac{1-2\nu}{c}\right)$, or equivalently, $(1-r)(1-2\delta) + r\left((1-\delta)\beta_n - \delta\beta_p(1-2\nu)\right) \leq 0$, which would violate the second condition in Theorem 4.1 of Blum-Stangl.

On the other hand, the threshold on $\eta(x,1)$ is at least interval $1-\delta$ if and only if $\frac{1}{1-\delta} \geq 2 - \frac{\lambda}{(1-r)q}$, or equivalently, $\lambda \geq \frac{(1-2\delta)(1-r)q}{1-\delta}$. Similarly, the threshold on $\eta(x,0)$ is at least $1-\delta$ if and only if $\frac{1}{1-\delta} \geq 1 + \frac{1-2\nu}{c} + \frac{\lambda}{\beta_n rq}$, or equivalently, $\beta_n rq\left(\frac{\delta}{1-\delta} - \frac{1-2\nu}{c}\right) \geq \lambda$. If both were to hold simultaneously, then $\beta_n rq\left(\frac{\delta}{1-\delta} - \frac{1-2\nu}{c}\right) \geq \frac{(1-2\delta)(1-r)q}{1-\delta}$, or equivalently, $(1-r)(1-2\delta) + r\left((1-\delta)\beta_p(1-2\nu) - \delta\beta_n\right) \leq 0$, which would violate the first condition in Theorem 4.1 of Blum-Stangl.

Step 3: For any λ, if the group-wise thresholds on $\eta(x,a)$ in h_λ have one of them at most δ and another at least $1-\delta$, then h_λ cannot satisfy equal opportunity unless $h_\lambda = h^*$. (Thresholds lying on opposite sides) We know that h^* satisfies equal opportunity. Suppose the threshold of h_λ on $\eta(x,0)$(call it t_0) and the threshold on $\eta(x,1)$(call it t_1) are separated by either δ or $1-\delta$. WLOG assume $t_0 < t_1$. Suppose $t_0 \leq \delta \leq t_1$. Since there is no input (x,a) with $\eta(x,a) \in (\delta, 1-\delta)$, we get $TPR_0(h_\lambda) \geq TPR_0(h^*)$ and $TPR_1(h^*) \geq TPR_1(h_\lambda)$. Since h^* satisfies equal opportunity on the original distribution D, we have $TPR_0(h^*) = TPR_1(h^*)$. If h_λ satisfies equal opportunity on the biased distribution $\tilde{D}$, we have $\widetilde{TPR}_0(h_\lambda) = \widetilde{TPR}_1(h_\lambda)$, and since $\widetilde{TPR}_a(h_\lambda) = TPR_a(h_\lambda)$ by Corollary 3.6, it implies $TPR_0(h_\lambda) = TPR_1(h_\lambda)$. Combining the two observations above, we get $TPR_0(h_\lambda) = TPR_0(h^*) = TPR_1(h^*) = TPR_1(h_\lambda)$, which is possible only if $h_\lambda \equiv h^*$ (as both are group-wise threshold classifiers). The other case $t_0 \leq 1-\delta \leq t_1$ can be argued similarly.

Combing Steps 1, 2, and 3 to complete the proof: Finally, from Steps 1, 2, and 3 together, it is clear that if the conditions in Theorem 4.1 of Blum-Stangl are met then the only possible classifiers $\tilde{h}_\lambda$ are the ones for which $\tilde{h}_\lambda = h^*$, and they satisfy equal opportunity on $\tilde{D}$. So under these conditions, the optimal λ^* (whatever it may be) gives $\tilde{h}_{EO} = \tilde{h}_{\lambda^*} = h^*$.

□

A.5. Proofs for Section 5

Proof. **(Proof of Proposition 5.1)**

The optimal reject option classifier $g : \mathcal{X} \times \mathcal{A} \to \{0, 1, \perp\}$ with rejection penalty δ can be thought of as a pair of binary classifiers ρ and h in $\mathcal{H}$ such that:

$$\rho(x,a) = \begin{cases} 1, & \text{if } g(x,a) = \perp \\ 0, & \text{if } g(x,a) \neq \perp \end{cases}$$

and $g(x,a) = h(x,a)$ whenever $\rho(x,a) = 0$ (or equivalently, $g(x,a) \neq \perp$). The optimal reject option classifier is defined as

$$h^{\text{rej}} = \operatorname*{argmin}_{g \in \mathcal{H}^{\text{rej}}} \Pr\left(g(X,A) \neq Y, g(X,A) \neq \perp\right) + \delta \Pr\left(g(X,A) = \perp\right),$$

where $\mathcal{H}^{\text{rej}}$ be the hypothesis class of all reject option classifiers $g : \mathcal{X} \times \mathcal{A} \to \{0, 1, \perp\}$. Equivalently, it can be re-written as:

$$
\begin{aligned}
(\rho^*, h^*) &= \operatorname*{argmin}_{\rho\in\mathcal{H}} \operatorname*{argmin}_{h\in\mathcal{H}} \Pr\left(h(X,A)\neq Y, \rho(X,A)=0\right) + \delta \Pr\left(\rho(X,A)=1\right) \\
&= \operatorname*{argmin}_{\rho\in\mathcal{H}} \operatorname*{argmin}_{h\in\mathcal{H}} 1 - \Pr\left(\rho(X,A)=1\right) - \Pr\left(h(X,A)=Y, \rho(X,A)=0\right) + \delta \Pr\left(\rho(X,A)=1\right) \\
&= \operatorname*{argmax}_{\rho\in\mathcal{H}} \operatorname*{argmax}_{h\in\mathcal{H}} \Pr\left(h(X,A)=Y, \rho(X,A)=0\right) + (1-\delta) \Pr\left(\rho(X,A)=1\right) \\
&= \operatorname*{argmax}_{\rho\in\mathcal{H}} \operatorname*{argmax}_{h\in\mathcal{H}} \sum_{a=0}^{1} \Pr\left(A=a\right) \mathop{\mathrm{E}}_{X|A=a}\Big[\left(1-\rho(X,a)\right)\left(h(X,a)\eta(X,a) + (1-h(X,a))(1-\eta(X,a))\right) \\
&\qquad\qquad + (1-\delta)\rho(X,a)\Big] \\
&= \operatorname*{argmax}_{\rho\in\mathcal{H}} \sum_{a=0}^{1} \Pr\left(A=a\right) \mathop{\mathrm{E}}_{X|A=a}\Big[(1-\delta)\,\rho(X,a) \\
&\qquad\qquad + (1-\rho(X,a))\left(1-\eta(X,a) + \operatorname*{argmax}_{h\in\mathcal{H}} h(X,a)(2\eta(X,a)-1)\right)\Big].
\end{aligned}
$$

We can now focus on obtaining the optimal functions for each group. For any $\rho \in \mathcal{H}$, solving the inner optimization gives the optimal solution $h^{\text{rej}}(x,a) = \mathbb{I}\left(\eta(x,a) \geq \frac{1}{2}\right)$ whenever $\rho(x,a) \neq 1$. Thus, the outer optimization can be written as

$$
\rho^* = \operatorname*{argmax}_{\rho\in\mathcal{H}} \mathop{\mathrm{E}}_{X|A=a}\left[(1-\delta)\,\rho(X,a) + (1-\rho(X,a))\left(1-\eta(X,a) + \mathbb{I}\left(\eta(X,a)\geq\frac{1}{2}\right)(2\eta(X,a)-1)\right)\right].
$$

The above maximization can be solved point-wise for each $x \in \mathcal{X}$ in group $A=a$. Whenever $\eta(x,a) \geq 1/2$ for an input x from group a, the function inside the expectation becomes $(1-\delta)\rho(x,a) + (1-\rho(x,a))\eta(x,a)$, which is maximized by $\rho^*(x,a) = \mathbb{I}\left(\eta(x,a) \in [1/2, 1-\delta)\right)$. On the other hand, whenever $\eta(x,a) < 1/2$ for an input x, a, the function inside the expectation becomes $(1-\delta)\rho(x,a) + (1-\rho(x,a)(1-\eta(x,a))$, which is maximized by $\rho^*(x,a) = \mathbb{I}\left(\eta(x,a) \in (\delta, 1/2)\right)$. Thus, overall we can write $\rho^*(x,a) = \mathbb{I}\left(\eta(x,a) \in (\delta, 1-\delta)\right)$.

□

Proof. **(Proof of Theorem 5.2)** From Proposition 3.8, we can write:

$$
\tilde{h}_\lambda(x,a) = \begin{cases} \mathbb{I}\left(\eta(x,0) \geq \dfrac{1}{1+\dfrac{1-2\nu}{c}+\dfrac{\lambda}{\beta_n r q}}\right), & \text{for } a=0 \\ \mathbb{I}\left(\eta(x,1) \geq \dfrac{1}{2-\dfrac{\lambda}{(1-r)q}}\right), & \text{for } a=1. \end{cases}
$$

We will now attempt to recover a classifier that matches the predictions of h^{rej}, whenever $h^{\text{rej}}(x,a) \neq \perp$, but incurs some penalty while trying to label the points when $h^{\text{rej}}(x,a) = \perp$. From the characterization of the optimal equal opportunity classifier shown earlier, there exists some optimal $\lambda^* \in \mathbb{R}$ such that $\tilde{h}_{EO} = h^{\text{rej}}$, whenever $h^{\text{rej}}(x,a) \neq \perp$. Such a classifier will not change whenever it lies in the rejection interval $(\delta, 1-\delta)$.

Step 1: If the given conditions on the bias parameters hold, then there exists a $\lambda \in \mathbb{R}$ such that $\tilde{h}_\lambda = h^{\text{rej}}$, whenever $h^{\text{rej}}(x,a) \neq \perp$. The threshold on $\eta(x,1)$ lies in the interval $(\delta, 1-\delta)$ if and only if $\dfrac{1}{1-\delta} < 2 - \dfrac{\lambda}{(1-r)q} < \dfrac{1}{\delta}$, or equivalently, $\dfrac{-(1-2\delta)(1-r)q}{\delta} < \lambda < \dfrac{(1-2\delta)(1-r)q}{1-\delta}$. Similarly, the threshold on $\eta(x,0)$ lies in the interval $(\delta, 1-\delta)$ if and only if $\dfrac{1}{1-\delta} < 1 + \dfrac{1-2\nu}{c} + \dfrac{\lambda}{\beta_n r q} < \dfrac{1}{\delta}$, or equivalently,

$$\beta_n rq\left(\frac{\delta}{1-\delta}-\frac{1-2\nu}{c}\right)<\lambda<\beta_n rq\left(\frac{1-\delta}{\delta}-\frac{1-2\nu}{c}\right).$$

There exists λ that simultaneously satisfies both sets of constraints given above if and only if

$$\beta_n rq\left(\frac{\delta}{1-\delta}-\frac{1-2\nu}{c}\right)<\frac{(1-2\delta)(1-r)q}{1-\delta}$$
$$\frac{-(1-2\delta)(1-r)q}{\delta}<\beta_n rq\left(\frac{1-\delta}{\delta}-\frac{1-2\nu}{c}\right).$$

Using $c=\beta_n/\beta_p$, the above conditions can be rewritten as

$$(1-r)(1-2\delta)+r\left((1-\delta)\beta_p(1-2\nu)-\delta\beta_n\right)>0 \quad \text{and}$$
$$(1-r)(1-2\delta)+r\left((1-\delta)\beta_n-\delta\beta_p(1-2\nu)\right)>0.$$

Step 2: If the given conditions on the bias parameters hold, then there cannot exist any $\lambda\in\mathbb{R}$ such that both the group-wise thresholds applied to $\eta(x,a)$ in h_λ are at most δ, or both the thresholds are at least $1-\delta$. (Thresholds lying on the same side) The threshold on $\eta(x,1)$ is at most δ if and only if $2-\frac{\lambda}{(1-r)q}\geq\frac{1}{\delta}$, or equivalently, $\frac{-(1-2\delta)(1-r)q}{\delta}\geq\lambda$. Similarly, the threshold on $\eta(x,0)$ is at most δ if and only if $1+\frac{1-2\nu}{c}+\frac{\lambda}{\beta_n rq}\geq\frac{1}{\delta}$, or equivalently, $\lambda\geq\beta_n rq\left(\frac{1-\delta}{\delta}-\frac{1-2\nu}{c}\right)$. If both were to hold simultaneously, then $\frac{-(1-2\delta)(1-r)q}{\delta}\geq\beta_n rq\left(\frac{1-\delta}{\delta}-\frac{1-2\nu}{c}\right)$, or equivalently, $(1-r)(1-2\delta)+r\left((1-\delta)\beta_n-\delta\beta_p(1-2\nu)\right)\leq 0$, which would violate the second condition in Step 1.

On the other hand, the threshold on $\eta(x,1)$ is at least interval $1-\delta$ if and only if $\frac{1}{1-\delta}\geq 2-\frac{\lambda}{(1-r)q}$, or equivalently, $\lambda\geq\frac{(1-2\delta)(1-r)q}{1-\delta}$. Similarly, the threshold on $\eta(x,0)$ is at least $1-\delta$ if and only if $\frac{1}{1-\delta}\geq 1+\frac{1-2\nu}{c}+\frac{\lambda}{\beta_n rq}$, or equivalently, $\beta_n rq\left(\frac{\delta}{1-\delta}-\frac{1-2\nu}{c}\right)\geq\lambda$. If both were to hold simultaneously, then $\beta_n rq\left(\frac{\delta}{1-\delta}-\frac{1-2\nu}{c}\right)\geq\frac{(1-2\delta)(1-r)q}{1-\delta}$, or equivalently, $(1-r)(1-2\delta)+r\left((1-\delta)\beta_p(1-2\nu)-\delta\beta_n\right)\leq 0$, which would violate the first condition in Step 1.

Step 3: For any λ, if the group-wise thresholds on $\eta(x,a)$ in h_λ have one of them at most δ and another at least $1-\delta$, then h_λ cannot satisfy equal opportunity unless $h_\lambda=h^{\text{rej}}$, whenever $h^{\text{rej}}(x,a)\neq\perp$. (Thresholds lying on opposite sides) From our assumption, we know that h^{rej} satisfies equal opportunity whenever $h^{\text{rej}}(x,a)\neq\perp$. Suppose the threshold of h_λ on $\eta(x,0)$(call it t_0) and the threshold on $\eta(x,1)$(call it t_1) are separated by either δ or $1-\delta$. WLOG assume $t_0<t_1$. Suppose $t_0\leq\delta\leq t_1$. Since there is no input (x,a) with $\eta(x,a)\in(\delta,1-\delta)$, we get $TPR_0(h_\lambda)\geq TPR_0(h^*)$ and $TPR_1(h^*)\geq TPR_1(h_\lambda)$. Since h^* satisfies equal opportunity on the original distribution D, we have $TPR_0(h^*)=TPR_1(h^*)$. If h_λ satisfies equal opportunity on the biased distribution $\tilde{D}$, we have $\widetilde{TPR}_0(h_\lambda)=\widetilde{TPR}_1(h_\lambda)$, and since $\widetilde{TPR}_a(h_\lambda)=TPR_a(h_\lambda)$ by Corollary 3.6, it implies $TPR_0(h_\lambda)=TPR_1(h_\lambda)$. Combining the two observations above, we get $TPR_0(h_\lambda)=TPR_0(h^*)=TPR_1(h^*)=TPR_1(h_\lambda)$, which is possible only if $h_\lambda\equiv h^*$ (as both are group-wise threshold classifiers). The other case $t_0\leq 1-\delta\leq t_1$ can be argued similarly.

Combing Steps 1, 2, and 3 to complete the proof: Finally, from Steps 1, 2, and 3 together, it is clear that when the above conditions are met, then the only possible classifiers $\tilde{h}_\lambda$ are the ones for which $\tilde{h}_\lambda=h^{\text{rej}}$, whenever $h^{\text{rej}}(x,a)\neq\perp$, and they satisfy equal opportunity on $\tilde{D}$. Furthermore, by our assumption, h^{rej} is EO-fair.

□

A.6. Proofs for Section 6

Definition 6.1 of ϵ-robustness of h^* can be rewritten as

$$\begin{aligned}
h^*(x) &= \operatorname*{argmax}_{h\in\mathcal{H}} \Pr\left(h(X,A)=Y'\right)\\
&= \operatorname*{argmax}_{h\in\mathcal{H}} \mathop{\mathrm{E}}_{(X,A)}\left[h(X,A)(2\eta'(X,A)-1)\right]\\
&= \operatorname*{argmax}_{h\in\mathcal{H}} \sum_{a=0}^{1}\Pr\left(A=a\right)\mathop{\mathrm{E}}_{X|A=a}\left[h(X,a)(2\eta'(X,a)-1)\right]\\
&= \operatorname*{argmax}_{h\in\mathcal{H}} \sum_{a=0}^{1}\Pr\left(A=a\right)\mathop{\mathrm{E}}_{X|A=a}\left[h(X,a)\left(2\,\frac{P_a\eta(X,a)+Q_a}{R_a\eta(X,a)+S_a}-1\right)\right]\\
&= \operatorname*{argmax}_{h\in\mathcal{H}} \sum_{a=0}^{1}\Pr\left(A=a\right)\mathop{\mathrm{E}}_{X|A=a}\left[h(X,a)\frac{(2P_a-R_a)\eta(X,a)+(2Q_a-S_a)}{R_a\eta(X,a)+S_a}\right],
\end{aligned}$$

for any $S_a=1, |R_a|\le\epsilon, |Q_a|\le\epsilon, 1-\epsilon\le P_a\le 1+\epsilon$, and $P_aS_a-Q_aR_a\ge 0$.

Proof. **(Proof of Theorem 6.2)** Let $\tilde{h}_{EO}=\operatorname*{argmax}_{h\in\widetilde{\mathcal{H}}_{\text{fair}}}\Pr\left(h(X,A)=\tilde{Y}\right)$. By Lagrange duality, we can write

$$\begin{aligned}
\tilde{h}_{EO} &= \operatorname*{argmax}_{h\in\mathcal{H}}\min_{\lambda\in\mathbb{R}}\Pr\left(h(X,A)=\tilde{Y}\right)+\lambda\Pr\left(h(X,A)=1|\tilde{Y}=1,A=0\right)-\lambda\Pr\left(h(X,A)=1|\tilde{Y}=1,A=1\right)\\
&= \operatorname*{argmax}_{h\in\mathcal{H}}\min_{\lambda\in\mathbb{R}}\sum_{a=0}^{1}\Pr\left(A=a\right)\mathop{\mathrm{E}}_{X|A=a}\left[h(X,a)(2\tilde{\eta}(X,a)-1)\right]+\frac{(-1)^{\mathbb{I}(a=1)}\,\lambda}{\mathop{\mathrm{E}}_{X|A=a}\left[\tilde{\eta}(X,a)\right]}\mathop{\mathrm{E}}_{X|A=a}\left[h(X,a)\tilde{\eta}(X,a)\right]\\
&= \operatorname*{argmax}_{h\in\mathcal{H}}\min_{\lambda\in\mathbb{R}}\sum_{a=0}^{1}\Pr\left(A=a\right)\mathop{\mathrm{E}}_{X|A=a}\left[h(X,a)(2\tilde{\eta}(X,a)-1)\right]+\frac{(-1)^{\mathbb{I}(a=1)}\,\lambda}{\Pr\left(\tilde{Y}=1|A=a\right)}\mathop{\mathrm{E}}_{X|A=a}\left[h(X,a)\tilde{\eta}(X,a)\right]\\
&= \operatorname*{argmax}_{h\in\mathcal{H}}\min_{\lambda\in\mathbb{R}}\sum_{a=0}^{1}\Pr\left(A=a\right)\mathop{\mathrm{E}}_{X|A=a}\left[h(X,a)\left(\left(2+\frac{(-1)^{\mathbb{I}(a=1)}\lambda}{\Pr\left(\tilde{Y}=1,A=a\right)}\right)\tilde{\eta}(X,a)-1\right)\right]\\
&= \operatorname*{argmax}_{h\in\mathcal{H}}\min_{\lambda\in\mathbb{R}} r\mathop{\mathrm{E}}_{X|A=0}\left[h(X,0)\left(\left(2+\frac{\lambda}{\beta_p(1-\nu)rq}\right)\frac{(1-\nu)\eta(X,0)}{(1-c)\eta(X,0)+c}-1\right)\right]\\
&\qquad +(1-r)\mathop{\mathrm{E}}_{X|A=1}\left[h(X,1)\left(\left(2-\frac{\lambda}{(1-r)q}\right)\eta(X,1)-1\right)\right]\\
&= \operatorname*{argmax}_{h\in\mathcal{H}}\min_{\lambda\in\mathbb{R}} r\mathop{\mathrm{E}}_{X|A=0}\left[h(X,0)\,\frac{\left(2(1-\nu)+\frac{\lambda}{\beta_p rq}-(1-c)\right)\eta(X,0)-c}{(1-c)\eta(X,0)+c}\right]\\
&\qquad +(1-r)\mathop{\mathrm{E}}_{X|A=1}\left[h(X,1)\left(\left(2-\frac{\lambda}{(1-r)q}\right)\eta(X,1)-1\right)\right]\\
&= \operatorname*{argmax}_{h\in\mathcal{H}}\min_{\lambda\in\mathbb{R}} r\mathop{\mathrm{E}}_{X|A=0}\left[h(X,0)\,\frac{\left(1-2\nu+c+\frac{\lambda}{\beta_p rq}\right)\eta(X,0)-c}{(1-c)\eta(X,0)+c}\right]\\
&\qquad +(1-r)\mathop{\mathrm{E}}_{X|A=1}\left[h(X,1)\left(\left(2-\frac{\lambda}{(1-r)q}\right)\eta(X,1)-1\right)\right]\\
&= \operatorname*{argmax}_{h\in\mathcal{H}}\min_{\lambda\in\mathbb{R}} r\mathop{\mathrm{E}}_{X|A=0}\left[h(X,0)\,\frac{\left(1+\frac{1-2\nu}{c}+\frac{\lambda}{\beta_n rq}\right)\eta(X,0)-1}{\frac{1-c}{c}\eta(X,0)+1}\right]
\end{aligned}$$

$$+ (1-r) \mathop{\mathrm{E}}_{X|A=1}\left[h(X,1)\left(\left(2-\frac{\lambda}{(1-r)q}\right)\eta(X,1)-1\right)\right], \qquad \text{using } c=\frac{\beta_n}{\beta_p}.$$

Now plugging in

$$\frac{(2P_0-R_0)\eta(X,0)+(2Q_0-S_0)}{R_0\eta(X,0)+S_0} = \frac{\left(1+\frac{1-2\nu}{c}+\frac{\lambda}{\beta_n rq}\right)\eta(X,0)-1}{\frac{1-c}{c}\eta(X,0)+1} \qquad \text{and}$$
$$\frac{(2P_1-R_1)\eta(X,1)+(2Q_1-S_1)}{R_1\eta(X,1)+S_1} = \left(2-\frac{\lambda}{(1-r)q}\right)\eta(X,1)-1,$$

we derive

$$P_0=\frac{1-\nu}{c}+\frac{\lambda}{2\beta_n rq}, \qquad Q_0=0, \qquad R_0=\frac{1-c}{c}, \qquad S_0=1,$$
$$P_1=1-\frac{\lambda}{2(1-r)q}, \qquad Q_1=0, \qquad R_1=0, \qquad S_1=1.$$

Step 1: If the given conditions on the bias parameters hold, then there exists a $\lambda \in \mathbb{R}$ such that $1-\epsilon \le P_0, P_1 \le 1+\epsilon$, and therefore, $h_\lambda = h^*$. They satisfy our ϵ-robustness conditions if and only if

$$|R_0| = \left|\frac{1-c}{c}\right| = \left|\frac{\beta_p-\beta_n}{\beta_n}\right| \le \epsilon,$$
$$1-\epsilon \le P_0 = \frac{1-\nu}{c}+\frac{\lambda}{2\beta_n rq} \le 1+\epsilon$$
$$1-\epsilon \le P_1 = 1-\frac{\lambda}{2(1-r)q} \le 1+\epsilon.$$

The above conditions also imply $P_aS_a - Q_aR_a \ge 0$, for $a \in \{0,1\}$, so we do not need to include these separately. The first inequality is equivalent to $(1-\epsilon)\beta_n \le \beta_p \le (1+\epsilon)\beta_n$. The second and third inequalities above are equivalent to $2rq\left((1-\epsilon)\beta_n-(1-\nu)\beta_p\right) \le \lambda \le 2rq\left((1+\epsilon)\beta_n-(1-\nu)\beta_p\right)$ and $-2\epsilon(1-r)q \le \lambda \le 2\epsilon(1-r)q$, respectively. There exists a $\lambda \in \mathbb{R}$ that satisfies the above two conditions on λ simultaneously if and only if

$$2rq\left((1-\epsilon)\beta_n-(1-\nu)\beta_p\right) \le 2\epsilon(1-r)q \quad \text{and} \quad -2\epsilon(1-r)q \le 2rq\left((1+\epsilon)\beta_n-(1-\nu)\beta_p\right),$$

or equivalently,

$$r\left((1-\nu)\beta_p-(1-\epsilon)\beta_n\right)+\epsilon(1-r) \ge 0 \quad \text{and} \quad r\left((1+\epsilon)\beta_n-(1-\nu)\beta_p\right)+\epsilon(1-r) \ge 0.$$

Step 2: If the given conditions on the bias parameters hold, then there cannot exist any $\lambda \in \mathbb{R}$ such that $\max\{P_0,P_1\} < 1-\epsilon$ or $\min\{P_0,P_1\} > 1+\epsilon$. Suppose $\max\{P_0,P_1\} < 1-\epsilon$. Then

$$\frac{1-\nu}{c}+\frac{\lambda}{2\beta_n rq} < 1-\epsilon \quad \text{and} \quad 1-\frac{\lambda}{2(1-r)q} < 1-\epsilon.$$

Thus, $\lambda < 2rq\left((1-\epsilon)\beta_n-(1-\nu)\beta_p\right)$ and $\lambda > 2\epsilon(1-r)q$. This implies $2rq\left((1-\epsilon)\beta_n-(1-\nu)\beta_p\right) > 2\epsilon(1-r)q$, or equivalently, $r\left((1-\nu)\beta_p-(1-\epsilon)\beta_n\right)+\epsilon(1-r) < 0$, violating one of the conditions in the theorem. The case of $\min\{P_0,P_1\} > 1+\epsilon$ can be argued similarly.

Step 3: If the given conditions on the bias parameters hold, then an optimal $\lambda^* \in \mathbb{R}$ cannot correspond to $P_0 < 1-\epsilon$ and $P_1 > 1+\epsilon$ (or vice versa). Suppose $\lambda^* \in \mathbb{R}$ satisfies $P_0 < 1-\epsilon$ and $P_1 > 1+\epsilon$, then

$$P_0=\frac{1-\nu}{c}+\frac{\lambda^*}{2\beta_n rq} < 1-\epsilon \quad \text{and} \quad P_1=1-\frac{\lambda^*}{2(1-r)q} > 1+\epsilon,$$

which implies that $\lambda^* < 2rq\left((1-\epsilon)\beta_n-(1-\nu)\beta_p\right)$ and $\lambda^* < -2\epsilon(1-r)q$. We know that

$$\tilde{h}_{EO} = \mathop{\mathrm{argmax}}_{h\in\mathcal{H}} \Pr\left(h(X,A)=\tilde{Y}\right) + \lambda^* \Pr\left(h(X,A)=1|\tilde{Y}=1,A=0\right) - \lambda^* \Pr\left(h(X,A)=1|\tilde{Y}=1,A=1\right).$$

Since we consider group-aware classification, when we know the optimal λ^*, the optimal equal opportunity classifier $\tilde{h}_{EO}$ on the biased distribution $\tilde{D}$ can be obtained by optimizing separately on the two groups. Note that $\tilde{h}_{EO}$ maximizes $\Pr\left(h(X,A)=\tilde{Y}\right)+\lambda^*\Pr\left(h(X,A)=1|\tilde{Y}=1,A=0\right)-\lambda^*\Pr\left(h(X,A)=1|\tilde{Y}=1,A=1\right)$ over $h\in\mathcal{H}$, for the optimal $\lambda^*\in\mathbb{R}$. Since we consider group-aware classifiers, $\tilde{h}_{EO}$ maximizes the following objective on the underprivileged group $a=0$.

$$
\begin{aligned}
&\Pr\left(h(X,A)=\tilde{Y}|A=0\right)+\frac{\lambda^*}{\Pr\left(A=0\right)}\Pr\left(h(X,A)=1|\tilde{Y}=1,A=0\right)\\
&=\Pr\left(\tilde{Y}=1|A=0\right)\ \widetilde{TPR}_0(h)+\Pr\left(\tilde{Y}=0|A=0\right)\ \widetilde{TNR}_0(h)+\frac{\lambda^*}{r}\ \widetilde{TPR}_0(h)\\
&=\Pr\left(\tilde{Y}=1|A=0\right)\ TPR_0(h)+\Pr\left(\tilde{Y}=0,Y=0|A=0\right)\ TNR_0(h)+\Pr\left(\tilde{Y}=0,Y=1|A=0\right)\ FNR_0(h)\\
&\qquad\qquad+\frac{\lambda^*}{r}\ TPR_0(h)\qquad\qquad\text{using Proposition 3.5}\\
&=\frac{q\beta_p(1-\nu)}{q\beta_p+(1-q)\beta_n}\ TPR_0(h)+\frac{(1-q)\beta_n}{q\beta_p+(1-q)\beta_n}\ TNR_0(h)+\frac{q\beta_p\nu}{q\beta_p+(1-q)\beta_n}\ FNR_0(h)+\frac{\lambda^*}{r}\ TPR_0(h)\\
&=\left(\frac{q\beta_p(1-2\nu)}{q\beta_p+(1-q)\beta_n}+\frac{\lambda^*}{r}\right)\ TPR_0(h)+\frac{(1-q)\beta_n}{q\beta_p+(1-q)\beta_n}\ TNR_0(h)+\frac{q\beta_p\nu}{q\beta_p+(1-q)\beta_n}\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{because } FNR_0(h)=1-TPR_0(h).
\end{aligned}
$$

Thus, $\tilde{h}_{EO}$ essentially maximizes a weighted linear combination of $TPR_0(h)$ and $TNR_0(h)$ over $h\in\mathcal{H}$, where the ratios of the weights for $TPR_0(h)$ and $TNR_0(h)$, respectively, is

$$
\begin{aligned}
\frac{q\beta_p(1-2\nu)}{(1-q)\beta_n}+\frac{\lambda^*(q\beta_p+(1-q)\beta_n)}{r(1-q)\beta_n}&\le\frac{q(1-2\nu)}{(1-q)c}+\frac{\lambda^*}{r(1-q)\beta_n}\qquad\text{using }\beta_p,\beta_n\le 1\\
&=\frac{q}{1-q}\left(\frac{1-2\nu}{c}+\frac{\lambda^*}{\beta_n rq}\right)\\
&=\frac{q}{1-q}\left(\frac{2(1-\nu)}{c}+\frac{\lambda^*}{\beta_n rq}-\frac{1}{c}\right)\\
&<\frac{q}{1-q}\left(2(1-\epsilon)-\frac{1}{c}\right)\qquad\text{using } P_0=\frac{1-\nu}{c}+\frac{\lambda^*}{2\beta_n rq}<1-\epsilon\\
&\le\frac{q}{1-q}\left(1-\epsilon\right)\qquad\text{using }\left|\frac{1-c}{c}\right|\le\epsilon\text{, and hence, }\frac{1}{c}\ge 1-\epsilon
\end{aligned}
$$

Hence, $TPR_0(\tilde{h}_{EO})\le TPR_0(h^*)$. Similarly, $\tilde{h}_{EO}$ maximizes the following objective on the group $a=1$.

$$
\begin{aligned}
&\Pr\left(h(X,A)=\tilde{Y}|A=1\right)-\frac{\lambda^*}{\Pr\left(A=1\right)}\Pr\left(h(X,A)=1|\tilde{Y}=1,A=1\right)\\
&=\Pr\left(\tilde{Y}=1|A=1\right)\ \widetilde{TPR}_1(h)+\Pr\left(\tilde{Y}=0|A=1\right)\ \widetilde{TNR}_1(h)-\frac{\lambda^*}{1-r}\ \widetilde{TPR}_1(h)\\
&=q\left(1-\frac{\lambda^*}{(1-r)q}\right)\ TPR_1(h)+(1-q)\ TNR_1(h).
\end{aligned}
$$

Thus, $\tilde{h}_{EO}$ essentially maximizes a weighted linear combination of $TPR_1(h)$ and $TNR_1(h)$ over $h\in\mathcal{H}$, where the ratio of the weights for $TPR_1(h)$ and $TNR_1(h)$, respectively, is

$$
\frac{q}{1-q}\left(1-\frac{\lambda^*}{(1-r)q}\right)>\frac{q}{1-q}\left(1+2\epsilon\right)\qquad\text{using } P_1=1-\frac{\lambda^*}{2(1-r)q}>1+\epsilon.
$$

Hence, $TPR_1(\tilde{h}_{EO})\ge TPR_1(h^*)$.

Combining the two observation $TPR_0(\tilde{h}_{EO})\le TPR_0(h^*)$ and $TPR_1(\tilde{h}_{EO})\ge TPR_1(h^*)$ above and that h^* satisfies equal opportunity, we get $TPR_0(\tilde{h}_{EO})\le TPR_0(h^*)=TPR_1(h^*)\le TPR_1(\tilde{h}_{EO})$. However, $\tilde{h}_{EO}\in\widetilde{\mathcal{H}}_{\text{fair, EO}}=$

$\mathcal{H}_{\text{fair, EO}}$ by Corollary 3.6, so we must have $TPR_0(\tilde{h}_{EO}) = TPR_0(h^*) = TPR_1(h^*) = TPR_1(\tilde{h}_{EO})$. This means that $\tilde{h}_{EO}$ would also maximize accuracy on the original distribution D, and therefore, by the ϵ-robustness property, we must have $\tilde{h}_{EO} \equiv h^*$.

The case of $P_0 > 1 + \epsilon$ and $P_1 < 1 - \epsilon$ can be argued similarly. □

A.7. Proofs for Section 7

Proof. **(Proof for Theorem 7.2)** We start with the Bayes Optimal EO classifier at the time step t:

$$\tilde{h}^{(t)}_{EO}(x,a) = \mathbb{I}\left(\tilde{\eta}_t(x,a) \geq \frac{1}{2 + \frac{(-1)^{\mathbb{I}(a=1)}\lambda^*}{\Pr\left(\tilde{Y}_t = 1, A = a\right)}}\right)$$

From Proposition 7.1 we can obtain the transformed $\tilde{\eta}_t$ with the same bias parameters at each time step:

$$= \begin{cases} \mathbb{I}\left(\dfrac{\eta(x,0)}{\frac{1-c}{1-\nu-c}\left(1-\left(\frac{c}{1-\nu}\right)^t\right)\eta(x,0) + \left(\frac{c}{1-\nu}\right)^t} \geq \dfrac{1}{2 + \frac{\lambda^*}{\Pr\left(\tilde{Y}_t = 1, A = 0\right)}}\right), & \text{for } a = 0 \\ \mathbb{I}\left(\eta(x,1) \geq \dfrac{1}{2 - \frac{\lambda^*}{\Pr(Y=1, A=1)}}\right), & \text{for } a = 1 \end{cases}$$

$$= \begin{cases} \mathbb{I}\left(\eta(x,0) \geq \dfrac{1}{\frac{1-2\nu-c}{1-\nu-c}\left(\frac{1-\nu}{c}\right)^t + \frac{\lambda^*}{\beta_n^t rq} + \frac{1-c}{1-\nu-c}}\right), & \text{for } a = 0, \text{using } c = \beta_n/\beta_p \\ \mathbb{I}\left(\eta(x,1) \geq \dfrac{1}{2 - \frac{\lambda^*}{(1-r)q}}\right), & \text{for } a = 1. \end{cases}$$

So the threshold on $\eta(x,1)$ lies in the interval $(\delta, 1-\delta)$ if and only if $\delta < \left(2 - \frac{\lambda^*}{(1-r)q}\right)^{-1} < 1 - \delta$, which is equivalent to $\frac{-(1-2\delta)(1-r)q}{\delta} < \lambda^* < \frac{(1-2\delta)(1-r)q}{1-\delta}$. Given this, now let's analyze the threshold on $\eta(x,0)$. If $\beta_n < 1$, then as $t \to \infty$ we have $\lambda^*/(\beta_n^t rq) \to \infty$, and therefore, the threshold on $\eta(x,0)$ in the above expression tends to 0, i.e., it cannot remain within $(\delta, 1-\delta)$ interval. If $\beta_n = 1$ then $c = \beta_n/\beta_p > 1 - \nu$. Thus, $((1-\nu)/c)^t \to 0$ as $t \to \infty$. Using this and $\beta_n = 1$, the above expression becomes

$$\frac{1}{\frac{1-2\nu-c}{1-\nu-c}\left(\frac{1-\nu}{c}\right)^t + \frac{\lambda^*}{\beta_n^t rq} + \frac{1-c}{1-\nu-c}} \to \frac{1}{\frac{\lambda^*}{rq} + \frac{1-c}{1-\nu-c}} \quad \text{as } t \to \infty.$$

The threshold on $\eta(x,1)$ lies in the interval $(\delta, 1-\delta)$ if and only if $\frac{-(1-2\delta)(1-r)q}{\delta} < \lambda^* < \frac{(1-2\delta)(1-r)q}{1-\delta}$. Similarly, the limit expression as $t \to \infty$ for threshold on $\eta(x,0)$ lies in the interval $(\delta, 1-\delta)$ if and only if $\frac{1}{1-\delta} < \frac{\lambda^*}{rq} + \frac{1-c}{1-\nu-c} < \frac{1}{\delta}$, or equivalently, $rq\left(\frac{1}{1-\delta} - \frac{1-c}{1-\nu-c}\right) < \lambda^* < rq\left(\frac{1}{\delta} - \frac{1-c}{1-\nu-c}\right)$. There exists a λ^* that

simultaneously satisfies the constraints on both the thresholds if and only if

$$\frac{-(1-2\delta)(1-r)q}{\delta} < rq\left(\frac{1}{\delta} - \frac{1-c}{1-\nu-c}\right) \quad \text{and} \quad rq\left(\frac{1}{1-\delta} - \frac{1-c}{1-\nu-c}\right) < \frac{(1-2\delta)(1-r)q}{1-\delta}.$$

Thus, the necessary conditions to recover h^* using equal opportunity fair classification on the biased distribution after t steps as $t \to \infty$ can be written as $\beta_n = 1$ and

$$\frac{r-(1-2\delta)(1-r)}{(1-\delta)r} < \frac{1-c}{1-\nu-c} < \frac{r+(1-2\delta)(1-r)}{\delta r}.$$

The above conditions can be further simplified as $\beta_n = 1$ and $1 - \frac{(1-2\delta)}{(1-\delta)r} < \frac{\nu\beta_p}{1-\beta_p(1-\nu)} < 1 + \frac{(1-2\delta)}{\delta r}$. □

Proof. (Proof for Theorem 7.3) Similar to the proof of Theorem 4.2, we begin with $\tilde{\eta}(x,a)$:

$$\tilde{h}_{EO,t}(x,a) = \mathbb{I}\left(\tilde{\eta}_t(x,a) \geq \frac{1}{2 + \frac{(-1)^{\mathbb{I}(a=1)}\lambda^*}{\Pr\left(\tilde{Y}=1, A=a\right)}}\right)$$

Because the population in group 1 is unaffected, we can obtain the conditions on λ^*. $\Pr(Y=1, A=1) = (1-r)q$. From Lemma A.5, we know that the threshold on $\eta(x,1)$ lies in the interval $(\delta, 1-\delta)$ if and only if $\frac{1}{1-\delta} < 2 - \frac{\lambda^*}{(1-r)q} < \frac{1}{\delta}$, or equivalently, $\frac{-(1-2\delta)(1-r)q}{\delta} < \lambda^* < \frac{(1-2\delta)(1-r)q}{1-\delta}$. We will now work towards obtaining the set of conditions on λ^* by focusing on the quantities inside the indicator on $\tilde{\eta}_t(x,0)$. Using Proposition 7.1:

$$\tilde{\eta}_t(x,0) = \frac{\eta(x,0)}{\sum_{i=1}^{t}\left(\frac{1-c_i}{1-\nu_i}\prod_{j=i+1}^{t}\frac{c_j}{1-\nu_j}\right)\eta(x,0) + \prod_{i=1}^{t}\frac{c_i}{1-\nu_i}} \geq \frac{1}{2 + \frac{\lambda^*}{\Pr\left(\tilde{Y}=1, A=0\right)}}$$

From previous derivations, we know that $\Pr\left(\tilde{Y}=1, A=0\right) = rq\prod_{i=1}^{t}(\beta_{p,i}(1-\nu_i)$. Simplifying the above expression gives us the following:

$$\eta(x,0) \geq \frac{1}{2\prod_{i=1}^{t}\frac{1-\nu_i}{c_i} + \frac{\lambda^*}{rq\prod_{i=1}^{t}\beta_{n,i}} - \sum_{i=1}^{t}\left(\frac{1-c_i}{c_i}\prod_{j=1}^{i-1}\frac{1-\nu_j}{c_j}\right)}$$

where $\prod_{j=1}^{i-1}\frac{1-\nu_j}{c_j} = 1$ whenever $j > i$.

Using similar arguments as in Proposition 4.2, the denominator must lie in the range $((1-\delta)^{-1}, \delta^{-1})$, which gives us the following inequalities on λ^*:

$$\frac{rq\prod_{i=1}^{t}\beta_{n,i}}{1-\delta} - 2rq\prod_{i=1}^{t}[(1-\nu_i)\beta_{p,i}] + rq\left(\prod_{i=1}^{t}[(1-\nu_i)\beta_{p,i}]\right)\sum_{j=1}^{t}\left(\frac{1-c_j}{c_j}\prod_{k=1}^{j-1}\frac{1-\nu_k}{c_k}\right) < \lambda^* \tag{1}$$

and

$$\lambda^* < \frac{rq\prod_{i=1}^{t}\beta_{n,i}}{\delta} - 2rq\prod_{i=1}^{t}[(1-\nu_i)\beta_{p,i}] + rq\left(\prod_{i=1}^{t}[(1-\nu_i)\beta_{p,i}]\right)\sum_{j=1}^{t}\left(\frac{1-c_j}{c_j}\prod_{k=1}^{j-1}\frac{1-\nu_k}{c_k}\right) \tag{2}$$

To get a tight bound, we can obtain an upper bound on the left side of Equation 1 and a lower bound on the right side of Equation 2, using the assumed maximal bias of $\beta_{n,t} \in [\beta_n, 1], \beta_{p,t} \in [\beta_p, 1]$ and $\nu_t \in [0, \nu],$ where $\nu < \frac{1}{2}$. This gives us the following bounds on λ^*:

$$\frac{rq}{1-\delta} - 2rq(1-\nu)^t\beta_p^t + rq\frac{1-\beta_n^t}{\beta_n^t} < \lambda^* < \frac{rq\beta_n^t}{\delta} - 2rq + rq(1-\nu)^t\beta_p^t(\beta_p - 1)\frac{1-\beta_p^t(1-\nu)^t}{1-\beta_p(1-\nu)} \tag{3}$$

Comparing this with the conditions on λ^* found earlier: $\frac{-(1-2\delta)(1-r)q}{\delta} < \lambda^* < \frac{(1-2\delta)(1-r)q}{1-\delta}$, we get the following inequalities:

$$\frac{(1-2\delta)(1-r)}{(1-\delta)r} - \frac{\delta}{1-\delta} > \frac{1}{\beta_n^t} - 2\beta_p^t(1-\nu)^t$$

and

$$\frac{(1-2\delta)(1-r)}{\delta r} > (1-\nu)^t\beta_p^t(1-\beta_p)\frac{1-\beta_p^t(1-\nu)^t}{1-\beta_p(1-\nu)} - \frac{\beta_n^t}{\delta} - 2$$

□

Theorem A.6. *Assuming boundedness on data bias parameters at each time step:* $\beta_{n,t} \in [\beta_n, 1], \beta_{p,t} \in [\beta_p, 1]$ *and* $\nu_t \in [0, \nu],$ *where* $\nu < \frac{1}{2}$*; whenever the following relationships hold:*

$$(1-3\delta)(\beta_p - 1)\frac{1-\beta_p^t(1-\nu)^t}{1-\beta_p(1-\nu)} - 2\delta - (1-\delta)\left(\frac{1-\beta_n^t}{\beta_n^t} - 2\beta_p^t(1-\nu)^t\right) > 0 \quad \textit{and}$$
$$1 + \delta\left((\beta_p - 1)\frac{1-\beta_p^t(1-\nu)^t}{1-\beta_p(1-\nu)} - \frac{2}{\beta_n^t} - (2\delta - 1)\frac{1-\beta_n^t}{\beta_n^t}\right) - 2(\delta - 1) > 0,$$

we have $\tilde{h}_{DP,t}(x,a) = h_{DP}(x,a) = h^*$.

Proof. We again begin with the thresholding results on Demographic Parity from Proposition A.4.

$$\tilde{h}_{DP}(x,a) = \mathbb{I}\left(\tilde{\eta}(x,a) \geq \frac{1}{2} - \frac{-1^{\mathbb{I}(a=1)}\lambda^*}{2}\right)$$

$$\text{(Using Proposition 7.1)} = \begin{cases} \mathbb{I}\left(\dfrac{\eta(x,0)}{\sum_{i=1}^{t}\left(\frac{1-c_i}{1-\nu_i}\prod_{j=i+1}^{t}\frac{c_j}{1-\nu_j}\right)\eta(x,0) + \prod_{i=1}^{t}\frac{c_i}{1-\nu_i}} \geq \frac{1}{2} - \frac{\lambda^*}{2}\right), & \text{for } a = 0 \\ \mathbb{I}\left(\eta(x,1) \geq \frac{1}{2} + \frac{\lambda^*}{2}\right), & \text{for } a = 1 \end{cases}$$

$$= \begin{cases} \mathbb{I}\left(\eta(x,0) \geq \dfrac{(1-\lambda^*)\prod_{i=1}^{t}\frac{c_i}{1-\nu_i}}{2-(1-\lambda^*)\sum_{i=1}^{t}\left(\frac{1-c_i}{1-\nu_i}\prod_{j=i+1}^{t}\frac{c_j}{1-\nu_j}\right)}\right), & \text{for } a=0 \\ \mathbb{I}\left(\eta(x,1) \geq \dfrac{1}{2}+\dfrac{\lambda^*}{2}\right), & \text{for } a=1 \end{cases}$$

$$= \begin{cases} \mathbb{I}\left(\eta(x,0) \geq \dfrac{1}{\frac{2}{1-\lambda^*}\prod_{i=1}^{t}\frac{1-\nu_i}{c_i}-\sum_{i=1}^{t}\left(\frac{1-c_i}{c_i}\prod_{j=1}^{i-1}\frac{1-\nu_j}{c_j}\right)}\right), & \text{for } a=0 \\ \mathbb{I}\left(\eta(x,1) \geq \dfrac{1}{2}+\dfrac{\lambda^*}{2}\right), & \text{for } a=1 \end{cases}$$

where $\prod_{j=1}^{i-1}\frac{1-\nu_j}{c_j} = 1$ whenever $j > i$. From Lemma A.5, we know that the threshold on $\eta(x,1)$ lies in the interval $(\delta, 1-\delta)$ if and only if $\delta < \frac{1}{2}+\frac{\lambda^*}{2} < 1-\delta$, or equivalently, $2\delta - 1 < \lambda^* < 1-2\delta$. Similarly, the threshold on $\eta(x,0)$ lies in the interval $(\delta, 1-\delta)$ if and only if the denominator lies in the range $((1-\delta)^{-1}, \delta^{-1})$ which gives us the following inequalities on λ^*:

$$\frac{1+(1-\delta)\left(\sum_{i=1}^{t}\left(\frac{1-c_i}{c_i}\prod_{j=1}^{i-1}\frac{1-\nu_j}{c_j}\right)-2\prod_{i=1}^{t}\frac{1-\nu_i}{c_i}\right)}{1+(1-\delta)\sum_{i=1}^{t}\left(\frac{1-c_i}{c_i}\prod_{j=1}^{i-1}\frac{1-\nu_j}{c_j}\right)} < \lambda^* \tag{4}$$

$$\lambda^* < \frac{1+\delta\left(\sum_{i=1}^{t}\left(\frac{1-c_i}{c_i}\prod_{j=1}^{i-1}\frac{1-\nu_j}{c_j}\right)-2\prod_{i=1}^{t}\frac{1-\nu_i}{c_i}\right)}{1+\delta\sum_{i=1}^{t}\left(\frac{1-c_i}{c_i}\prod_{j=1}^{i-1}\frac{1-\nu_j}{c_j}\right)} \tag{5}$$

To get a tight bound, we can obtain an upper bound on Equation 4 and a lower bound on Equation 5, using the assumed maximal bias of $\beta_{n,t} \in [\beta_n, 1], \beta_{p,t} \in [\beta_p, 1]$ and $\nu_t \in [0, \nu],$ where $\nu < \frac{1}{2}$. This gives us the following bounds on λ^*:

$$\frac{1+(1-\delta)\left(\frac{1-\beta_n^t}{\beta_n^t}-2\beta_p^t(1-\nu)^t\right)}{1+(1-\delta)(\beta_p-1)\left(\frac{1-\beta_p^t(1-\nu)^t}{1-\beta_p(1-\nu)}\right)} < 1-2\delta \text{ , and}$$

$$2\delta - 1 < \frac{1+\delta\left((\beta_p-1)\left(\frac{1-\beta_p^t(1-\nu)^t}{1-\beta_p(1-\nu)}\right)-\frac{2}{\beta_n^t}\right)}{1+\delta\frac{1-\beta_n^t}{\beta_n^t}}$$

Therefore, to have a λ^* which satisfies both the sets of the inequalities, the following conditions must hold:

$$(1-3\delta)(\beta_p-1)\frac{1-\beta_p^t(1-\nu)^t}{(1-\beta_p(1-\nu)} - 2\delta - (1-\delta)\left(\frac{1-\beta_n^t}{\beta_n^t}-2\beta_p^t(1-\nu)^t\right) > 0 \tag{6}$$

and

$$1+\delta\left((\beta_p-1)\frac{1-\beta_p^t(1-\nu)^t}{(1-\beta_p(1-\nu)} - \frac{2}{\beta_n^t} - (2\delta-1)\frac{1-\beta_n^t}{\beta_n^t}\right) - 2(\delta-1) > 0 \tag{7}$$

□