On the Vulnerability of Fairness Constrained Learning to Malicious Noise

Avrim Blum TTIC
Princewill Okoroafor Cornell University
Aadirupa Saha Apple
Kevin Stangl TTIC

Abstract

We consider the vulnerability of fairness-constrained learning to small amounts of malicious noise in the training data. [Konstantinov and Lampert, 2021] initiated the study of this question and presented negative results showing there exist data distributions where for several fairness constraints, any proper learner will exhibit high vulnerability when group sizes are imbalanced. Here, we present a more optimistic view, showing that if we allow randomized classifiers, then the landscape is much more nuanced. For example, for Demographic Parity we show we can incur only a $\Theta(\alpha)$ loss in accuracy, where α is the malicious noise rate, matching the best possible even without fairness constraints. For Equal Opportunity, we show we can incur an $O(\sqrt{\alpha})$ loss, and give a matching $\Omega(\sqrt{\alpha})$ lower bound. In contrast, [Konstantinov and Lampert, 2021] showed for proper learners the loss in accuracy for both notions is $\Omega(1)$. The key technical novelty of our work is how randomization can bypass simple "tricks" an adversary can use to amplify his power. We also consider additional fairness notions including Equalized Odds and Calibration. For these fairness notions, the excess accuracy clusters into three natural regimes $O(\alpha)$,$O(\sqrt{\alpha})$, and $O(1)$. These results provide a more fine-grained view of the sensitivity of fairness-constrained learning to adversarial noise in training data.

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

1 Introduction

The widespread adoption of machine learning algorithms across various domains, including recidivism prediction [Flores et al., 2016, Dieterich et al., 2016], credit lending [Kozodoi et al., 2022], and predictive policing [Lum and Isaac, 2016], has raised significant concerns regarding biases and unfairness in these models. Consequently, substantial efforts have been devoted to developing approaches for learning fair classification models that exhibit effective performance across protected attributes such as race and gender.

One critical aspect of addressing fairness in machine learning is ensuring the robustness of models against small amounts of adversarial corruption present in the training data. This data corruption may arise due to flawed data collection or cleaning processes [Saunders et al., 2013], strategic misreporting [Hardt et al., 2016a], under-representation of certain subgroups [Blum and Stangl, 2019], or distribution shift over time [Schrouff et al., 2022].

Empirical studies have demonstrated that such data unreliability is often centered on sensitive groups e.g. [Gianfrancesco et al., 2018], emphasizing the need to understand the vulnerability of fair learning to adversarial perturbations. A concerning possibility is that fairness constraints might allow the adversary to amplify the effect of their corruptions by exploiting how these constraints require the classifier to have comparable performance on every relevant sub-group, even small ones.

Previous work by [Konstantinov and Lampert, 2021] and [Celis et al., 2021] have explored this topic from a theoretical perspective, considering different adversarial noise models. [Celis et al., 2021] focused on the η-Hamming model, where the adversary selectively perturbs a fraction of the dataset by modifying the protected attribute.

[Konstantinov and Lampert, 2021] on the other hand, investigated the Malicious Noise model, where an α fraction of the data-set (or distribution) is uniformly

chosen and those data points are arbitrarily perturbed by the adversary. We will focus on this Malicious Noise model. In our study, we extend the framework of fair learning in the presence of Malicious Noise [Konstantinov and Lampert, 2021] by considering a broader range of fairness constraints and introducing a way to bypass some of their negative results by randomizing the hypothesis class.

[Konstantinov and Lampert, 2021] present a pessimistic outlook, highlighting data distributions in which any proper learner, particularly in scenarios with imbalanced group sizes, exhibits high vulnerability to adversarial corruption when the learner is constrained by Demographic Parity [Calders et al., 2009] or Equal Opportunity [Hardt et al., 2016b]. These results demonstrate novel and concerning challenges to designing fair learning algorithms resilient to adversarial manipulation in the form of Malicious Noise.

The results of [Konstantinov and Lampert, 2021] indicate that fairness constrained learning is much less robust than unconstrained learning.

In this paper, we present a more optimistic perspective on the vulnerability of fairness-constrained learning to malicious noise by introducing randomized classifiers. By allowing randomized classifiers, we can explore alternative strategies that effectively mitigate the impact of malicious noise and enhance the robustness of fairness-constrained models. In addition, we extend the analysis beyond the fairness constraints examined in [Konstantinov and Lampert, 2021], providing a complete characterization of the robustness of each constraint and revealing a diverse range of vulnerabilities to Malicious Noise.

1.1 Our Contributions

We bypass the impossibility results in [Konstantinov and Lampert, 2021] by allowing the learner to produce a randomized improper classifier. This classifier is constructed from hypotheses in the base class $\mathcal{H}$ using our post-processing procedure, which we refer to as the (P, Q)-Randomized Expansion of a hypothesis class $\mathcal{H}$, or $\mathcal{PQ}(\mathcal{H})$

Definition 1 ($\mathcal{PQ}(\mathcal{H})$)**.** *For each classifier $h \in \mathcal{H}$, for $p, q \in [0, 1]$*

$$h_{p,q}(x) := \begin{cases} h(x) & \text{with probability } 1-p \\ y \sim \text{Bernoulli}(q) & \text{otherwise} \end{cases}$$

We define $\mathcal{PQ}(\mathcal{H})$ as the expanded hypothesis class created by the set of all possible $h_{p,q}(x)$.

$$\mathcal{PQ}(\mathcal{H}) := \{h_{p,q} \mid h \in \mathcal{H}, p, q \in [0, 1]\}$$

When clear from context we drop the dependence on p, q and simply refer to $\hat{h} \in \mathcal{PQ}(\mathcal{H})$.

Larger p means we ignore more of the information in the base classifier h and rely on the Bernoulli(q). The main technical questions we address in this paper are:

How susceptible and sensitive are fairness constrained learning algorithms to Malicious Noise and to what extent does this vulnerability depend on the specific fairness notion, especially if we allow improper learning?

We focus on proving the existence of $h' \in \mathcal{PQ}(\mathcal{H})$ that satisfies a given fairness constraint and exhibits minimal accuracy loss on the original data distribution. Recall that α is the fraction of the overall distribution that is corrupted by the adversary.

Our list of contributions is:

1. We propose a way to bypass lower bounds [Konstantinov and Lampert, 2021] in Fair-ERM with Malicious Noise by extending the hypothesis class using the $\mathcal{PQ}(\mathcal{H})$ notion.
2. For the Demographic Parity [Calders et al., 2009] constraint, our approach guarantees no more than $O(\alpha)$ loss in accuracy (which is *optimal* in the Malicious Noise model *without* fairness constraints [Kearns and Li, 1988a]). In other words, in contrast to the perspective in [Konstantinov and Lampert, 2021] which shows $\Omega(1)$ accuracy loss, we show that Demographic Parity constrained ERM can be made just as robust to Malicious Noise as unconstrained ERM.
3. For the Equal Opportunity [Hardt et al., 2016b] constraint, we guarantee no more than $O(\sqrt{\alpha})$ accuracy loss and show that this is tight, i.e no classifier can do better.
4. For the fairness constraints Equalized Odds [Hardt et al., 2016b], Minimax Error [Diana et al., 2020], Predictive Parity, and our novel fairness constraint Parity Calibration, we show strong negative results. Namely, for each constraint there exist natural distributions such that an adversary that can force any algorithm to return a fair classifier that has $\Omega(1)$ loss in accuracy.
5. For Calibration [Pleiss et al., 2017], we observe that the excess accuracy loss is at most $O(\alpha)$.

2 Related Work

[Kearns and Li, 1988b] introduced the notion of malicious noise which is analyzed in [Bshouty et al., 2002,

Auer and Cesa-Bianchi, 1998, Klivans et al., 2009, Long and Servedio, 2011, Awasthi et al., 2014]. [Balcan et al., 2022] considers a related adversary as a way to formalize data poisoning attacks in adversarial robustness [Goodfellow et al., 2014].

The interaction of fairness constraints with explicitly unreliable data is a critical research direction since issues of bias and fairness are often closely connected with data reliability concerns [Gianfrancesco et al., 2018]. Malicious noise is both a way to model an explicit adversary and a way to consider unknown natural issues with the data distribution.

[Celis et al., 2021], which also studies fairness with data corruptions. primarily focuses on the stronger Nasty Noise Model [Bshouty et al., 2002] combined with an assumption on the minimum size of groups/events. They do not consider how randomized post-processing improves robustness.

The most closely related work to ours, [Konstantinov and Lampert, 2021], explores the limits of fairness-aware PAC learning within the classic malicious noise model of [Valiant, 1985], where the adversary can replace a uniformly random fraction of the data points with arbitrary data, with full knowledge of the learning algorithm, the data distribution, and the remaining examples. [Konstantinov and Lampert, 2021] focuses on binary classification with just two popular group fairness constraints: Demographic Parity [Calders et al., 2009] and Equal Opportunity [Hardt et al., 2016b]. In addition to those constraints, we also consider Equalized Odds and multiple Calibration notions. Similarly to [Konstantinov and Lampert, 2021], a key aspect of our results is how the size of the smaller group makes it more vulnerable to data corruption.

2.1 Group Fairness Discussion

Note that group fairness constraints [Chouldechova, 2017, Kleinberg et al., 2016, Hardt et al., 2016b] are relatively simple to evaluate and provide relatively weak guarantees in contrast to fairness notions in [Calders et al., 2009, Dwork et al., 2021, Hébert-Johnson et al., 2018], among others. However, despite this weakness, these group notions are used in practice [Madaio et al., 2021] to check model performance, so continuing to investigate them in parallel to the stronger fairness notions is worthwhile.

[Blum and Stangl, 2019] takes a somewhat converse perspective to this paper. Instead of considering worst case instances for how much fairness constraints force excess accuracy loss with an adversary, that paper asks how fairness constraints can help us recover from the biased data when the bias is more benign, but still complicates Empirical Risk Minimization. Future work could connect our results with theirs by considering an intermediate adversary model, such as [Massart and Nédélec, 2006].

Interestingly, there may be high level connections between our paper and work in federated learning with differing data quality levels, e.g [Chu et al., 2023].

3 Preliminaries

In fairness-constrained learning, the goal is to learn a classifier that achieves good predictive performance while satisfying certain fairness constraints that connect the performance of the classifier on multiple groups, to ensure effective performance on all groups.

Specifically, we start with a dataset consisting of examples with feature vectors ($x \in \mathcal{X}$), labels ($y \in \mathcal{Y}$), and group attributes ($z \in \mathcal{Z}$). We assume that each example is drawn i.i.d from a joint distribution $\mathcal{D}$ of random variables (X, Y, Z). There are multiple groups in the dataset, and we aim to ensure that the classifier's predictions do not unfairly favor or disfavor any particular group. We will denote $\mathcal{D}_z$ as the conditional distribution of random variables X and Y given $Z = z$. For simplicity, we will assume there are two disjoint groups: A and B in the dataset with B being the smaller and more vulnerable of the two. However, our results apply more broadly to any number of groups.

We aim to use the dataset to learn a classifier $f : \mathcal{X} \to \mathcal{Y}$ given a hypothesis class $\mathcal{H}$. However, in this paper we suppress sample complexity learning issues and focus on characterizing the accuracy properties of the best hypothesis in the expanded hypothesis class with a corrupted data distribution $\widetilde{D}$. The goal is to probe the fundamental sensitivity of Fair-ERM to unreliable data in the large sample limit.

To this end, we consider solving the standard risk minimization problem with fairness constraints, known as Fair-ERM.

$$\min_{h \in \mathcal{H}} \ \mathbb{E}_{(X,Y,Z)\sim\mathcal{D}} \left[\mathbf{1}(h(X) \neq Y)\right] \tag{1}$$

$$\text{subject to} \quad F_z(h) = F_{z'}(h) \qquad \forall z, z' \in \mathcal{Z}. \tag{2}$$

where $F_z(h)$ is some fairness statistic of h for group z given the true labels y, such as *true positive rate* : $(\text{TPR}) : F_z(h) = \mathbb{P}(h(X) = +1 | Y = +1, Z = z)$.

We make a mild *realizability assumption* that there exists a solution to this risk minimization problem. That is, there is at least one hypothesis in the class that satisfies the fairness constraint. This optimal solution is denoted as h^*.

As noted above, since we allow our hypothesis class to be group-aware, we can reason about h^*_z for all $z \in Z$, where h^*_z is the restriction of the optimal classifier h^* to members of group z. In other words, h^*_z is the optimal group-specific classifier for Group z.

3.1 Fairness Notions

Different formal notions of group fairness have previously been proposed in literature. These notions include, but are not limited to, Demographic Parity, Equal Opportunity, Equalized Odds, Minimax Fairness, and Calibration[Dwork et al., 2012, Calders et al., 2009, Hardt et al., 2016b, Kleinberg et al., 2016, Chouldechova, 2017].

Selecting the "right" fairness measure is, in general, application-dependent.[1] One of our goals in this work is to provide understanding of their implications under adversarial attack, which could aid in the selection process. For the convenience of the reader, we include a table in Appendix A summarizing the fairness notions we consider in this paper. Other than Calibration, these all are notions for binary classifiers. In Section 5.2 we will introduce a new variant of Calibration and will defer discussion of that notion until then.

3.2 Adversary Model

Throughout this paper, we focus on the Malicious Noise Model, introduced by [Kearns and Li, 1988a]. This model considers a worst-case scenario where an adversary has complete control over a uniformly chosen α proportion of the training data and can manipulate that fraction in order to move the learning algorithm towards their desired outcomes, i.e. increasing test time error [on un-corrupted data]. [•]

In [Kearns and Li, 1988a]'s model, the samples are drawn sequentially from a fixed distribution. With probability α and full knowledge of the learning algorithm, data distribution and all the samples that have been drawn so far, the adversary can replace sample (x, y) with an arbitrary sample $(\tilde{x}, \tilde{y})$.

At each time-step t,

1. The adversary chooses a distribution $\widetilde{\mathcal{D}}_t$ that is $\alpha-$close to the original distribution $\mathcal{D}$ in Total Variation distance.
2. The algorithm draws a sample (x_t, y_t) from $\widetilde{\mathcal{D}}_t$ instead of $\mathcal{D}$

[1] We would also note that these fairness constraints are imperfect measures of fairness that likely do not capture all of the normative properties relevant to a specific task or system.

Note that the adversary's choice at time t, $\widetilde{\mathcal{D}}_t$ can depend on the samples $\{x_1, y_1, \ldots, x_{t-1}, y_{t-1}\}$ chosen so far.

Reframing the Malicious Noise Model in this manner simplifies analysis and allows us to focus on the fundamental aspect of this model which is how the accuracy guarantees of fairness constrained learning change as a function of α.

3.3 Core Learning Problem

In the fair-ERM problem with Malicious Noise, our goal is to find the optimal classifier h^* subject to a fairness constraint. However, the presence of the Malicious Noise makes this objective challenging. Instead of observing samples from the true distribution $\mathcal{D}$, we observe samples from a corrupted distribution $\widetilde{\mathcal{D}}$.

In the standard ERM setting, [Kearns and Li, 1988b] show that the optimal classifier that can be learned using this corrupted data is one that is $O(\alpha)$-close to h^* in terms of accuracy [on the original distribution]. The fair-ERM problem with a Malicious Noise adversary introduces an additional layer of complexity, as we must also ensure fairness while achieving high accuracy.

Definition 2. *We say a learning algorithm for the fair-ERM problem is β-robust with respect to a fairness constraint F in the malicious adversary model with corruption fraction α, if it returns a classifier h such that $\widetilde{F}_z(h) = \widetilde{F}_{z'}(h)$ and*

$$|\mathbb{E}_{\mathcal{D}}\ [\mathbf{1}(h(X,Z) \neq Y)] - \mathbb{E}_{\mathcal{D}}\ [\mathbf{1}(h^*(X,Z) \neq Y)]| \leq \beta(\alpha)$$

where h^* is the optimal classifier for the fair-ERM problem on the true distribution $\mathcal{D}$ with respect to a hypothesis class $\mathcal{H}$ and β is a function of α.

This definition captures the desired properties of a learning algorithm that can perform well under the malicious noise model while achieving both accuracy and fairness, as measured by the fairness constraint F.

Thus, this is an agnostic learning problem [Haussler, 1992] with an adversary and fairness constraints. As referenced in the introduction, we will allow the learner to return $h' \in \mathcal{PQ}(\mathcal{H})$, where $\mathcal{PQ}(\mathcal{H})$ is a way to post-process each $h \in \mathcal{H}$ using randomness. In Sections 4 and 5.2 will characterize the optimal value of β given the relevant fairness constraint F and base hypothesis class $\mathcal{H}$.

4 Main Results: Demographic Parity, Equal Opportunity and Equalized Odds

We now present our technical findings for Demographic Parity, Equal Opportunity, and Equalized

Odds, and show how randomization enables better accuracy for Fair-ERM with Malicious Noise. [Konstantinov and Lampert, 2021] show impossibility results for Demographic Parity and Equal Opportunity where a *proper* learner is forced to return a classifier with $\Omega(1)$ excess unfairness and accuracy compared to h^* for a synthetic and finite hypothesis class/distribution.

To overcome this limitation, we propose a novel approach to make the hypothesis class $\mathcal{H}$ more robust, by injecting noise into each hypothesis $h \in \mathcal{H}$. In other words, we allow improper learning, and refer to the resulting expanded set of hypotheses as $\mathcal{PQ}(\mathcal{H})$. By injecting controlled noise into the hypotheses, we effectively "smooth out" the hypothesis class $\mathcal{H}$, making it more resilient against adversarial manipulation.

Since we allow group-aware classifiers, we learn two classifiers $h_A, h_B \in \mathcal{PQ}(\mathcal{H})$, typically distinct from each other. Our method minimizes fairness loss for any hypothesis class and true distribution $\mathcal{D}$, under the assumption that at least one classifier in the original hypothesis class $\mathcal{H}$ satisfies the fairness constraints. We aim to find a fair classifier $\hat{h} \in \mathcal{PQ}(\mathcal{H})$ that is as good as the best $h^* \in \mathcal{H}$.

4.1 Demographic Parity

Demographic Parity [Calders et al., 2009] requires that the decisions of the classifier are independent of the group membership; that is, $P_{(x,y)\sim\mathcal{D}_A}[h(x) = 1] = P_{(x,y)\sim\mathcal{D}_B}[h(x) = 1]$[2].

When the original distribution $\mathcal{D}$ is corrupted, a fair hypothesis on $\mathcal{D}$ may seem unfair to the learner. In order to analyze our approach it is important to understand how the fairness violation of a fixed hypothesis changes after the adversary corrupts an α proportion of the distribution.

Proposition 3 (Parity after corruption)**.** *Let $\widetilde{\mathcal{D}}$ be any corrupted distribution chosen by the adversary, and h be a fixed hypothesis in $\mathcal{H}$. For a fixed group A, the following inequality bounds the change in the proportion of positive labels assigned by h:* $\left|P_{(x,y)\sim\widetilde{\mathcal{D}}_A}[h(x) = 1] - P_{(x,y)\sim\mathcal{D}_A}[h(x) = 1]\right| \leq \frac{\alpha}{(1-\alpha)r_A+\alpha}$ *where $r_A = P_{(x,y)\sim\mathcal{D}}[x \in A]$, i.e how prevalent the group is in the original distribution.*

This proposition provides an upper bound on the change in the proportion of positive labels assigned by a fixed hypothesis h in $\mathcal{H}$ after the distribution has been corrupted according to the Malicious Noise Model. The full proof can be found in the Appendix B. The proof shows that this change is bounded by a function of the corruption rate α and the proportion of the dataset in the fixed group A, denoted by r_A.

Intuitively, this means that the smaller a group is, the easier it is for the adversary to make a fair hypothesis seem unfair for members of that group.

Theorem 4. *For any hypothesis class $\mathcal{H}$ and distribution $\mathcal{D} = (\mathcal{D}_A, \mathcal{D}_B)$, a robust fair-ERM learner for the parity constraint in the Malicious Adversarial Model returns a hypothesis $\hat{h} \in \mathcal{PQ}(\mathcal{H})$ such that*

$$\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbf{1}(\hat{h}(x) \neq y)\right] - \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbf{1}(h^*(x) \neq y)\right]\right| \leq O(\alpha)$$

where h^* is the optimal classifier for the fair-ERM problem on the true distribution $\mathcal{D}$ with respect to hypothesis class $\mathcal{H}$.

This theorem states that a fair-ERM learner searching over the smoothed hypothesis class $\mathcal{PQ}(\mathcal{H})$ returns a classifier that is within α of the accuracy of the best fair classifier in the original class $\mathcal{H}$. The full constructive proof can be found in the appendix B.

The proof exhibits classifier $h \in \mathcal{PQ}(\mathcal{H})$ that satisfies the desired guarantee. This classifier mostly behaves identically to h^* but deviates with probability p_A on samples from group A (and with probability p_B on samples from group B). We give an explicit assignment of these probability values p_A, q_A, p_B, q_B in $[0, 1]$ so that h is perceived as fair by the learner. Then, we show that these values are small enough that the proportion of samples where $h(x) \neq h^*(x)$ is small ($O(\alpha)$). *This is the best possible outcome in the malicious adversary model without fairness constraints [Kearns and Li, 1988b].*

4.2 Equal Opportunity

Equal Opportunity [Hardt et al., 2016b] requires that the True Positive Rates of the classifier are equal across all the groups, that is, $P_{(x,y)\sim\mathcal{D}_A}[h(x) = 1 \mid y = 1] = P_{(x,y)\sim\mathcal{D}_B}[h(x) = 1 \mid y = 1]$. Similarly to Demographic Parity, we first provide bounds on how the fairness violation of a fixed hypothesis changes after the adversary corrupts an α proportion of the dataset. This is important because it gives an estimate of how much violation must be offset.

Proposition 5 (TPR after corruption)**.** *Let $\widetilde{\mathcal{D}}$ be any corrupted distribution chosen by the adversary, and h be a fixed hypothesis in $\mathcal{H}$. For a fixed group A, the following inequality bounds the change in True Positive Rate of h:*

$$\left|TPR_A(h, \widetilde{\mathcal{D}}) - TPR_A(h, \mathcal{D})\right| \leq \frac{\alpha}{(1-\alpha)r_A^+ + \alpha} \quad (3)$$

[2]Note there is no reference in the definition to the true labels, so a trivial hypothesis that flips a random coin for all examples would satisfy this notion, albeit at minimal accuracy.

where $TPR_A(h, \mathcal{D}) = P_{(x,y)\sim\mathcal{D}_A}[h(x) = 1|y = 1]$ *and* $r_A^+ = P_{(x,y)\sim\mathcal{D}}[y = 1 \cap x \in A]$

This proposition provides an upper bound on the change to the true positive rate in group A assigned by a fixed hypothesis h in $\mathcal{H}$ after the dataset has been corrupted according to the Malicious Noise Model. The full proof can be found in the appendix B.1.

Since $\alpha \in [0, 1]$, $O(\sqrt{\alpha})$ *means larger (meaning worse) accuracy loss, compared* $O(\alpha)$.

The function that bounds the change in True Positive rate is similar to that of Demographic Parity with the proportional size of group A r_A replaced with the proportion of the dataset that is positively labeled and in group A, r_A^+. We will see that this slight change in dependence makes the robust learning problem more difficult and leads to a worse dependence on α.

Theorem 6 (Upper Bound)**.** *For any hypothesis class* $\mathcal{H}$ *and distribution* $\mathcal{D} = (\mathcal{D}_A, \mathcal{D}_B)$*, a robust fair-ERM learner for the equal opportunity constraint in the Malicious Adversarial Model returns a hypothesis* $\hat{h}$ *such that* $\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbf{1}(\hat{h}(x) \neq y)\right] - \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbf{1}(h^*(x) \neq y)\right]\right| \leq O(\sqrt{\alpha})$ *where* h^* *is the optimal classifier for the fair-ERM problem on the true distribution* $\mathcal{D}$ *with respect to hypothesis class* $\mathcal{H}$.

This theorem states that a fair-ERM learner, when applied with the smoothed hypothesis class $\mathcal{PQ}(\mathcal{H})$, returns a classifier that is within $\sqrt{\alpha}$ of the accuracy of the best fair classifier in the original class $\mathcal{H}$. The full proof can be found in Appendix B.

In constructing a classifier $h \in \mathcal{PQ}(\mathcal{H})$, we aim for it to behave mostly identically to h^* but introduce deviations with probability p_A for samples from group A and probability p_B for samples from group B. However, in the case of the Equal Opportunity fairness constraint, this approach, as used for Demographic Parity, does not work effectively. We observe that the amount of correction required for each group depends inversely on the true positive rate, which presents challenges when the true positive rate (TPR) is close to 0 or 1.

For example, suppose the classifier achieves a 95% TPR for a fixed group. The adversary can manipulate the TPR to reach 100% by corrupting only a few samples. Correcting this change and bringing the TPR back down to 95% is an incredibly difficult task, similar to finding a *needle in a haystack*, since the learner essentially has to identify the corrupted samples to do so. In such cases, it might be easier for the learning algorithm to increase the TPR of the other groups from 95% to 100% instead.

The tradeoff lies in equalizing the corrections that only transform the TPR of a fixed group to its original value versus the corrections that transform the TPR of other groups to match the TPR of the group with the most corruptions.

Theorem 7 (Lower Bound)**.** *There exists a distribution* $\mathcal{D} = (\mathcal{D}_A, \mathcal{D}_B)$ *and a malicious adversary of power* α *that guarantees that any hypothesis,* $\hat{h}$*, returned by an improper learner for the fair-ERM problem with the equal opportunity constraint satisfies the following:* $\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbf{1}(\hat{h}(x) \neq y)\right] - \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbf{1}(h^*(x) \neq y)\right]\right| \geq \Omega(\sqrt{\alpha})$ *where* h^* *is the optimal classifier for the fair-ERM problem on the true distribution* $\mathcal{D}$ *with respect to a hypothesis class* $\mathcal{H}$.

In this lower bound, under the given conditions, no proper or improper learner can achieve an error rate lower than a threshold that scales with the square root of the adversary's power. In other words, as the adversary becomes more powerful (α increases), the error rate of the hypothesis returned by an improper learner will unavoidably be at least on the order of $\sqrt{\alpha}$.

The proof of this lower bound result sets up a scenario reflecting the *needle in the haystack* issue described earlier. We present a distribution with two groups, one of size $\sqrt{\alpha}$ and the other of size $1 - \sqrt{\alpha}$. We construct a hypothesis class where the optimal classifier has a high but not perfect true positive rate. Then we show that any improper learner must either suffer poor accuracy on the smaller group or lose $\Omega(\sqrt{\alpha})$ accuracy on the larger group. The full proof can be found in the Appendix B.1.

4.3 Equalized Odds

Equalized Odds [Hardt et al., 2016b] is a fairness constraint that requires equalizing True Positive Rates (TPRs) and False Positive Rates (FPRs) across different groups. This notion is very sensitive to the adversary's corrupted data and we exhibit a problematic lower bound, showing the adversary can force terrible performance.

The intuition is as follows; for a small group, the Adversary can set the Bayes Optimal TPR/FPRs rates of that group towards arbitrary values and so the learner must do the same on the larger group, regardless of their hypothesis class, forcing large error. The full proof is in Appendix C.

Theorem 8 (Lower Bound)**.** *For a learner seeking to maximize accuracy subject to satisfying Equalized Odds, an adversary with corruption fraction* α *can force an additional* $\Omega(1)$ *accuracy loss when compared to the performance of the optimal fair classifier on the true distribution.*

5 Main Results: Calibration

In this section, we explore various notions of calibration [Dawid, 1982] for our model. Calibration is a desirable property typically considered for classifiers, where predicted label probabilities should correspond to observed frequencies in the long run. For example, in weather forecasting, a well-calibrated predictor should have approximately 60% of days with rain when it forecasts a 60% chance of rain. This calibration requirement should hold for every predicted probability value output by the model.

Calibration has important fairness implications [Flores et al., 2016, Chouldechova, 2017, Pleiss et al., 2017, Hebert-Johnson et al., 2018] because a mis-calibrated predictor can lead to harmful actions in high-stakes settings, such as over-incarceration [Hamilton, 2019]. We show that varying the exact calibration requirements can substantially impact the model's accuracy loss when malicious noise is present in the training data.

In this section, we align closely with [Pleiss et al., 2017], where the learner seeks to maximize accuracy while ensuring the classifier is perfectly calibrated. Throughout this paper, we have focused on binary classifiers, so in Section 5.1 we consider a related notion called Predictive Parity [Chouldechova, 2017, Flores et al., 2016], before considering calibration notions for hypotheses with output in $[0, 1]$.

5.1 Predictive Parity Lower Bound

Definition 9 (Predictive Parity [Chouldechova, 2017])**.** *A binary classifier $h : \mathcal{X} \to \{0, 1\}$ satisfies predictive parity if for groups A and B, $P_{x\sim\mathcal{D}_A}[h(x) = 1] > 0$, $P_{x\sim\mathcal{D}_B}[h(x) = 1] > 0$ and*

$$P_{(x,y)\sim\mathcal{D}_A}[y = 1|h(x) = 1] = P_{(x,y)\sim\mathcal{D}_B}[y = 1|h(x) = 1]$$

In later sections we consider other calibration notions. Here we consider an adversary who is attacking a learner constrained by equal predictive parity when group sizes are *imbalanced*.

Theorem 10. *For a malicious adversary with corruption fraction α, for Fair-ERM constrained to satisfy Predictive Parity, then there is no $h \in \mathcal{PQ}(\mathcal{H})$ with less than $\Omega(1)$ error.*

The intuition for this statement is that imbalanced group size will allow the adversary to change the conditional mean substantially. Below, we have an informal proof:

Proof Sketch: Suppose $P(x \in A) = 1 - \alpha$ and $P(x \in B) = \alpha$. Observe that whatever the initial value of $P_{(x,y)\sim\mathcal{D}_B}[y = 1|h(x) = 1]$, the adversary can drive this value $P_{(x,y)\sim\widetilde{\mathcal{D}}_B}[y = 1|h(x) = 1]$ to 50% or below by adding a duplicate copy of every natural example in group B with the opposite label.

Since all of these points are information-theoretically indistinguishable, any hypothesis for group B that makes any positive predictions incurs at least 50% error and $1/2 = P_{(x,y)\sim\widetilde{\mathcal{D}}_B}[y = 1|h(x) = 1]$ calibration error. Any classifier for group A satisfying Predictive Parity will have to do the same, yielding our $\Omega(1)$ error. □

5.2 Extension to Finer Grained Hypothesis Classes

A criticism of this lower bound might be that these calibration notions are very coarse and calibration is intended for fine-grained predictors, meaning those that have a finer grained discretization of the probabilities in $[0, 1]$. We now provide extensions for these lower bounds to real valued $\mathcal{H}$. Interestingly, we show if the learner can modify their 'binning strategy', the learner can 'decouple' the classifiers for the groups in the population and thus only suffer $O(\alpha)$ accuracy loss. We adopt the version of calibration from [Pleiss et al., 2017].

Definition 11 (Calibration)**.** *A classifier $h : \mathcal{X} \to [0, 1]$ is Calibrated with respect to distribution $\mathcal{D}$ if*

$$\forall r \in [0, 1], r = \mathbb{E}_{(x,y)\sim\mathcal{D}}[y = 1|h(x) = r]$$

We will primarily focus on the discretized version of this definition where the classifier assigns every data point to one of R bins, each with a corresponding label r, that partition $[0, 1]$ dis-jointly. We will refer to this partition as $[R]$ with $r \in [R]$ corresponding to the prediction of a bin.

$$\forall r \in [R], r = \mathbb{E}_{(x,y)\sim\mathcal{D}}[y = 1|h(x) = r]$$

Calibration as a fairness requirements with demographic groups requires that the classifier h is calibrated with respect to the group distributions $\mathcal{D}_A$ and $\mathcal{D}_B$ simultaneously. In the sections that follow when we say 'calibrated' this always refers to calibration with respect to $\mathcal{D}_A$ and $\mathcal{D}_B$.

Theorem 12. *The learner wants to maximize accuracy subject to using a calibrated classifier, $h : \mathcal{X} \to [R]$ where $[R]$ is a partition of $[0, 1]$ into bins.*

The learner may modify the binning strategy after the adversary commits to a corruption strategy. Then an adversary with corruption fraction α can force at most $O(\alpha)$ excess accuracy loss over the non-corrupted optimal classifier.

5.3 Parity Calibration

Motivated by Theorem 12, we introduce a *novel* fairness notion we call *Parity Calibration*[3] Informally, this notion is a generalization of Statistical/Demographic parity [Dwork et al., 2012] for the case of classifier with R bins partitioning $[0, 1]$.

Definition 13 (Parity Calibration)**.** *Classifier* $h : \mathcal{X} \to [R]$*, where* $[R]$ *is a partition of* $[0, 1]$ *into labelled bins, satisfies* Parity Calibration *if the classifier is Calibrated (Definition 11)* and

$$\forall r \in [R], P_{(x,y)\sim\mathcal{D}_A}[h(x) = r] = P_{(x,y)\sim\mathcal{D}_B}[h(x) = r]$$

Theorem 14. *Consider a learner maximizing accuracy subject to satisfying Parity Calibration. The learner may modify the binning strategy after the adversary commits to a corruption strategy. Then an adversary with corruption fraction* α *can force* $\Omega(1)$ *excess accuracy loss over the non-corrupted optimal classifier.*

If the size of Group B is $O(\alpha)$, then following a similar duplication strategy for Predictive Parity Theorem 10, then the adversary can force Group B to have an expected label of 50%, i.e. $\forall x \in B, \mathbb{E}_{x\sim\mathcal{D}_B}[y|x] = 50\%$. Thus, any classifier that is calibrated must assign all of Group B to a 50% bucket. In order to satisfy *Parity Calibration*, the classifier must do the same to Group A, yielding 50% error on Group A.

6 Discussion

We study Fair-ERM in the Malicious Noise model, and in some cases allow the learner to maintain optimal overall accuracy despite the signal in Group B being almost entirely washed out. In particular, we show that different fairness constraints have fundamentally different behavior in the presence of Malicious Noise, in terms of the amount of accuracy loss that a given level of Malicious Noise could cause a fairness-constrained learner to incur. The key to achieving our results, which are more optimistic than those in [Konstantinov and Lampert, 2021], is allowing for improper learners using the (P,Q)-randomized expansions of the given class $\mathcal{H}$. The type of smoothness we create by using $\mathcal{PQ}(\mathcal{H})$ seems to be a natural property that is likely shared by many natural hypothesis classes.

Fairness notions are motivated as a response to learned disparities when there is systemic error affecting one group. Fairness notions are supposed to mitigate this by ruling out classifiers that have worse performance on a sub-group. This can peg both classifiers at a lower level of performance in order to *motivate* [Hardt et al., 2016b] improving the data collection or labelling process to obtain more reliable performance. However, it is also desirable that fairness constraints perform gracefully when subject to Malicious Noise, because fairness constraints will be used in contexts where the data is unreliable and noisy. This tension, exposed by our work, motivates ongoing work studying the sensitivity level of fairness constraints.

[3]We would note that this is initial discussion of a novel fairness constraint that arose naturally from considering Theorem 12. The idea is in some cases it might be more desirable to have a more sensitive calibration notion, hence we define Parity Calibration. This notion requires further study and analysis before deployment in sensitive contexts.

This work was supported in part by the National Science Foundation under grant CCF-2212968, by the Simons Foundation under the Simons Collaboration on the Theory of Algorithmic Fairness, by the Defense Advanced Research Projects Agency under cooperative agreement HR00112020003. The views expressed in this work do not necessarily reflect the position or the policy of the Government and no official endorsement should be inferred. Approved for public release; distribution is unlimited.

References

[Auer and Cesa-Bianchi, 1998] Auer, P. and Cesa-Bianchi, N. (1998). On-line learning with malicious noise and the closure algorithm. *Annals of mathematics and artificial intelligence*, 23:83–99.

[Awasthi et al., 2014] Awasthi, P., Balcan, M. F., and Long, P. M. (2014). The power of localization for efficiently learning linear separators with noise. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 449–458.

[Balcan et al., 2022] Balcan, M.-F., Blum, A., Hanneke, S., and Sharma, D. (2022). Robustly-reliable learners under poisoning attacks. In *Conference on Learning Theory*, pages 4498–4534. PMLR.

[Blum and Stangl, 2019] Blum, A. and Stangl, K. (2019). Recovering from biased data: Can fairness constraints improve accuracy? *arXiv preprint arXiv:1912.01094*.

[Bshouty et al., 2002] Bshouty, N. H., Eiron, N., and Kushilevitz, E. (2002). Pac learning with nasty noise. *Theoretical Computer Science*, 288(2):255–275.

[Calders et al., 2009] Calders, T., Kamiran, F., and Pechenizkiy, M. (2009). Building classifiers with independency constraints. In *2009 IEEE international conference on data mining workshops*, pages 13–18. IEEE.

[Celis et al., 2021] Celis, L. E., Mehrotra, A., and Vishnoi, N. (2021). Fair classification with adversarial

perturbations. *Advances in Neural Information Processing Systems*, 34:8158–8171.

[Chouldechova, 2017] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.

[Chu et al., 2023] Chu, W., Xie, C., Wang, B., Li, L., Yin, L., Zhao, H., and Li, B. (2023). Focus: Fairness via agent-awareness for federated learning on heterogeneous data.

[Dawid, 1982] Dawid, P. (1982). The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610.

[Diana et al., 2020] Diana, E., Gill, W., Kearns, M., Kenthapadi, K., and Roth, A. (2020). Convergent algorithms for (relaxed) minimax fairness. *CoRR*, abs/2011.03108.

[Dieterich et al., 2016] Dieterich, W., Mendoza, C., and Brennan, T. (2016). Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 7(7.4):1.

[Dwork et al., 2012] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

[Dwork et al., 2021] Dwork, C., Kim, M. P., Reingold, O., Rothblum, G. N., and Yona, G. (2021). Outcome indistinguishability. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1095–1108.

[Flores et al., 2016] Flores, A. W., Bechtel, K., and Lowenkamp, C. T. (2016). False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation*, 80:38.

[Gianfrancesco et al., 2018] Gianfrancesco, M. A., Tamang, S., Yazdany, J., and Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine*, 178(11):1544–1547.

[Goodfellow et al., 2014] Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

[Hamilton, 2019] Hamilton, M. (2019). The sexist algorithm. *Behavioral Sciences and the Law*, 145.

[Hardt et al., 2016a] Hardt, M., Megiddo, N., Papadimitriou, C., and Wootters, M. (2016a). Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122.

[Hardt et al., 2016b] Hardt, M., Price, E., and Srebro, N. (2016b). Equality of opportunity in supervised learning. *CoRR*, abs/1610.02413.

[Haussler, 1992] Haussler, D. (1992). Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and Computation*, 100(1):78–150.

[Hébert-Johnson et al., 2018] Hébert-Johnson, U., Kim, M., Reingold, O., and Rothblum, G. (2018). Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR.

[Hebert-Johnson et al., 2018] Hebert-Johnson, U., Kim, M., Reingold, O., and Rothblum, G. (2018). Multicalibration: Calibration for the (Computationally-identifiable) masses. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1939–1948. PMLR.

[Kearns and Li, 1988a] Kearns, M. and Li, M. (1988a). Learning in the presence of malicious errors. In *Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing*, STOC '88, page 267–280, New York, NY, USA. Association for Computing Machinery.

[Kearns and Li, 1988b] Kearns, M. and Li, M. (1988b). Learning in the presence of malicious errors. In *Proceedings of the twentieth annual ACM symposium on Theory of computing*, pages 267–280.

[Kleinberg et al., 2016] Kleinberg, J. M., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *CoRR*, abs/1609.05807.

[Klivans et al., 2009] Klivans, A. R., Long, P. M., and Servedio, R. A. (2009). Learning halfspaces with malicious noise. *Journal of Machine Learning Research*, 10(12).

[Konstantinov and Lampert, 2021] Konstantinov, N. and Lampert, C. H. (2021). Fairness-aware learning from corrupted data. *CoRR*, abs/2102.06004.

[Kozodoi et al., 2022] Kozodoi, N., Jacob, J., and Lessmann, S. (2022). Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research*, 297(3):1083–1094.

[Long and Servedio, 2011] Long, P. and Servedio, R. (2011). Learning large-margin halfspaces with more malicious noise. *Advances in Neural Information Processing Systems*, 24.

[Lum and Isaac, 2016] Lum, K. and Isaac, W. (2016). To predict and serve? *Significance*, 13(5):14–19.

[Madaio et al., 2021] Madaio, M., Egede, L., Subramonyam, H., Vaughan, J. W., and Wallach, H. M. (2021). Assessing the fairness of AI systems: AI practitioners' processes, challenges, and needs for support. *CoRR*, abs/2112.05675.

[Massart and Nédélec, 2006] Massart, P. and Nédélec, É. (2006). Risk bounds for statistical learning.

[Pleiss et al., 2017] Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J. M., and Weinberger, K. Q. (2017). On fairness and calibration. *CoRR*, abs/1709.02012.

[Saunders et al., 2013] Saunders, C. L., Abel, G. A., El Turabi, A., Ahmed, F., and Lyratzopoulos, G. (2013). Accuracy of routinely recorded ethnic group information compared with self-reported ethnicity: evidence from the english cancer patient experience survey. *BMJ open*, 3(6):e002882.

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

[Valiant, 1985] Valiant, L. G. (1985). Learning disjunction of conjunctions. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'85, page 560–566, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**/No/Not Applicable]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Yes**/No/Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/**Not Applicable**]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**/No/Not Applicable]
   (b) Complete proofs of all theoretical results. [**Yes**/No/Not Applicable]
   (c) Clear explanations of any assumptions. [**Yes**/No/Not Applicable]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes/No/Not **Applicable**]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes/No/**Not Applicable**]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes/No/**Not Applicable**]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes/No/**Not Applicable**]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes/No/**Not Applicable**]

(b) The license information of the assets, if applicable. [Yes/No/**Not Applicable**]

(c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No//**Not Applicable**]

(d) Information about consent from data providers/curators. [Yes/No/**Not Applicable**]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Yes/No/**Not Applicable**]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/**Not Applicable**]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/**Not Applicable**]

A Fairness Notions

Fairness Constraints	
Demographic Parity [Dwork et al., 2012]	$P_{(x,y)\sim\mathcal{D}_A}[h(x)=1] = P_{(x,y)\sim\mathcal{D}_B}[h(x)=1]$
Equal Opportunity [Hardt et al., 2016b]	$P_{(x,y)\sim\mathcal{D}_A}[h(x)=1\|y=1] = P_{(x,y)\sim\mathcal{D}_B}[h(x)=1\|y=1]$
Equalized Odds [Hardt et al., 2016b]	$P_{(x,y)\sim\mathcal{D}_A}[h(x)=1\|y=1] = P_{(x,y)\sim\mathcal{D}_B}[h(x)=1\|y=1]$ and $P_{(x,y)\sim\mathcal{D}_A}[h(x)=1\|y=0] = P_{(x,y)\sim\mathcal{D}_B}[h(x)=1\|y=0]$
Predictive Parity [Chouldechova, 2017]	$P_{(x,y)\sim\mathcal{D}_A}[y=1\|h(x)=1] = P_{(x,y)\sim\mathcal{D}_B}[y=1\|h(x)=1]$
Calibration[4] [Kleinberg et al., 2016, Dawid, 1982]	$\forall r\in[0,1], \quad r = \mathbb{E}_{x,y\sim\mathcal{D}}[y\|h(x)=r]$

B Proofs

Proposition 3 (Parity after corruption)**.** *Let $\widetilde{\mathcal{D}}$ be any corrupted distribution chosen by the adversary, and h be a fixed hypothesis in $\mathcal{H}$. For a fixed group A, the following inequality bounds the change in the proportion of positive labels assigned by h:* $\left|P_{(x,y)\sim\widetilde{\mathcal{D}}_A}[h(x)=1] - P_{(x,y)\sim\mathcal{D}_A}[h(x)=1]\right| \leq \frac{\alpha}{(1-\alpha)r_A+\alpha}$ *where $r_A = P_{(x,y)\sim\mathcal{D}}[x\in A]$, i.e how prevalent the group is in the original distribution.*

Proof of Proposition 3. We want to bound the change in the proportion of positive labels assigned by h when we move from the original distribution $\mathcal{D}$ to the corrupted distribution $\widetilde{\mathcal{D}}$. For a fixed group A, we can express the proportion of positive labels assigned by h in $\widetilde{\mathcal{D}}$ in terms of the proportion of positive labels assigned by h in $\mathcal{D}$ as follows:

$$P_{(x,y)\sim\widetilde{\mathcal{D}}_A}[h(x)=1] = \frac{(1-\alpha)P_{(x,y)\sim\mathcal{D}_A}[h(x)=1]\cdot P_{(x,y)\sim\mathcal{D}}[x\in A] + E_A}{(1-\alpha)P_{(x,y)\sim\mathcal{D}}[x\in A] + \alpha_A} \tag{4}$$

where α_A is the proportion of the data set that is corrupted and in group A and E_A is the proportion of the data set that is corrupted, in group A and positively labeled by h.

Our goal is to obtain an upper bound on the difference between $P_{(x,y)\sim\widetilde{\mathcal{D}}_A}[h(x)=1]$ and $P_{(x,y)\sim\mathcal{D}_A}[h(x)=1]$. We use the fact that $E_A \leq \alpha$ and $\alpha_A \leq \alpha$ to obtain the following upper bound:

$$\left|P_{(x,y)\sim\widetilde{\mathcal{D}}_A}[h(x)=1] - P_{(x,y)\sim\mathcal{D}_A}[h(x)=1]\right| = \left|\frac{E_A - \alpha_A P_{(x,y)\sim\mathcal{D}_A}[h(x)=1]}{(1-\alpha)P_{(x,y)\sim\mathcal{D}}[x\in A]+\alpha_A}\right| \leq \frac{\alpha}{(1-\alpha)r_A+\alpha} \tag{5}$$

□

Theorem 4. *For any hypothesis class $\mathcal{H}$ and distribution $\mathcal{D} = (\mathcal{D}_A, \mathcal{D}_B)$, a robust fair-ERM learner for the parity constraint in the Malicious Adversarial Model returns a hypothesis $\hat{h}\in\mathcal{PQ}(\mathcal{H})$ such that*

$\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}\ [\mathbf{1}(\hat{h}(x)\neq y)] - \ \mathbb{E}_{(x,y)\sim\mathcal{D}}\ [\mathbf{1}(h^*(x)\neq y)]\right| \leq O(\alpha)$
where h^* is the optimal classifier for the fair-ERM problem on the true distribution $\mathcal{D}$ with respect to hypothesis class $\mathcal{H}$.

Proof of Theorem 4. For $z\in\{A,B\}$, let $F_z(h)$ and $\widetilde{F}_z(h)$ denote the proportions of positive labels assigned by h in group z in the original and corrupted distributions respectively. That is, for group A, $F_A(h) = P_{(x,y)\sim\mathcal{D}_A}[h(x)=1]$ and $\widetilde{F}_A(h) = P_{(x,y)\sim\widetilde{\mathcal{D}}_A}[h(x)=1]$. It suffices to show that there exists $h\in cl(\mathcal{H})$ that satisfies the guarantees above. Consider $h^*\in\mathcal{H}$. By the realizability assumption , h^* satisfies the parity constraint i.e $F_A(h^*) = F_B(h^*)$.

After the corruption, the parity violation of h^*, $|\widetilde{F}_A(h^*) - \widetilde{F}_B(h^*)|$ may increase. Now we define the following parameters (p_z and q_z) for $z \in \{A, B\}$.

$$p_z = \begin{cases} \frac{F_z(h^*) - \widetilde{F}_z(h^*)}{1 - \widetilde{F}_z(h^*)} & \text{if } F_z(h^*) \geq \widetilde{F}_z(h^*) \\ \frac{\widetilde{F}_z(h^*) - F_z(h^*)}{\widetilde{F}_z(h^*)} & \text{otherwise} \end{cases} \qquad q_z = \begin{cases} 1 & \text{if } F_z(h^*) \geq \widetilde{F}_z(h^*) \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Now consider a hypothesis $\hat{h}$ that behaves as follows: Given a sample x:

- If $x \in A$, with probability p_A, return label q_A. Otherwise return $h^*(x)$
- Similarly, if $x \in B$, with probability p_B, return label q_B. Otherwise return $h^*(x)$

$\hat{h} \in \mathcal{PQ}(\mathcal{H})$ since it follows the definition of our closure model. We will now show that $\hat{h}$ satisfies the parity constraint in the corrupted distribution (i.e $\widetilde{F}_A(\hat{h}) = \widetilde{F}_B(\hat{h})$). First, observe that for $z \in \{A, B\}$, if $F_z(h^*) \geq \widetilde{F}_z(h^*)$, then $\widetilde{F}_z(\hat{h}) = F_z(h^*)$. This is because

$$\begin{aligned} \widetilde{F}_z(\hat{h}) &= (1 - p_z)\widetilde{F}_z(h^*) + p_z q_z \\ &= \widetilde{F}_z(h^*) + p_z(1 - \widetilde{F}_z(h^*)) \\ &= \widetilde{F}_z(h^*) + F_z(h^*) - \widetilde{F}_z(h^*) \\ &= F_z(h^*) \end{aligned}$$

Similarly, if $F_z(h^*) < \widetilde{F}_z(h^*)$, then $\widetilde{F}_z(\hat{h}) = F_z(h^*)$. This is because

$$\begin{aligned} \widetilde{F}_z(\hat{h}) &= (1 - p_z)\widetilde{F}_z(h^*) + p_z q_z \\ &= \widetilde{F}_z(h^*) + p_z(0 - \widetilde{F}_z(h^*)) \\ &= \widetilde{F}_z(h^*) + F_z(h^*) - \widetilde{F}_z(h^*) \\ &= F_z(h^*) \end{aligned}$$

Thus, $\widetilde{F}_A(\hat{h}) = F_A(h^*) = F_B(h^*) = \widetilde{F}_B(\hat{h})$. Therefore $\hat{h}$ satisfies the parity constraint in the corrupted distribution.

We will now show that $\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbf{1}(\hat{h}(x) \neq y)\right] - \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbf{1}(h^*(x) \neq y)\right]\right| \leq O(\alpha)$. Since $\hat{h}$ deviates from h^* with probability p_A on samples from A, and with probability p_B on samples from B, we only need to show that the proportion of samples such that $\hat{h}(x) \neq h^*(x)$ is small. Fix a group $z \in \{A, B\}$. If $F_z(h^*) \geq \widetilde{F}_z(h^*)$, then with probability $p_z = \frac{F_z(h^*) - \widetilde{F}_z(h^*)}{1 - \widetilde{F}_z(h^*)}$, $\hat{h}$ returns a positive label for samples in group z. Thus, the expected proportion of samples in group z such that $\hat{h}(x) \neq h^*(x)$ is p_z times the proportion of negative labelled samples (by h^*) in group z (since those get flipped to positive).

$$\begin{aligned} \mathbb{E}_{x\in z}[\mathbf{1}(\hat{h}(x) \neq h^*(x))] &= p_z \cdot P_{(x,y)\sim\mathcal{D}}[x \in z](1 - \widetilde{F}_z(h^*)) \\ &= \frac{F_z(h^*) - \widetilde{F}_z(h^*)}{1 - \widetilde{F}_z(h^*)} \cdot P_{(x,y)\sim\mathcal{D}}[x \in z](1 - \widetilde{F}_z(h^*)) \\ &= (F_z(h^*) - \widetilde{F}_z(h^*)) \cdot P_{(x,y)\sim\mathcal{D}}[x \in z] \end{aligned}$$

Similarly, if $\widetilde{F}_z(h^*) > F_z(h^*)$, then with probability $p_z = \frac{\widetilde{F}_z(h^*) - F_z(h^*)}{\widetilde{F}_z(h^*)}$, $\hat{h}$ returns a negative label. Thus, the expected proportion of samples in group z such that $\hat{h}(x) \neq h^*(x)$ is p_z times the proportion of positively labelled samples (by h^*) in group z (since those get flipped to negative).

$$\begin{aligned} \mathbb{E}_{x\in z}[\mathbf{1}(\hat{h}(x) \neq h^*(x))] &= p_z \cdot P_{(x,y)\sim\mathcal{D}}[x \in z] \cdot \widetilde{F}_z(h^*) \\ &= \frac{\widetilde{F}_z(h^*) - F_z(h^*)}{\widetilde{F}_z(h^*)} \cdot P_{(x,y)\sim\mathcal{D}}[x \in z] \cdot \widetilde{F}_z(h^*) \\ &= (\widetilde{F}_z(h^*) - F_z(h^*)) \cdot P_{(x,y)\sim\mathcal{D}}[x \in z] \end{aligned}$$

Therefore, the expected total number of samples such that $\hat{h}(x) \neq h^*(x)$ across the entire distribution is bounded as follows:

$$\begin{aligned}\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbf{1}(\hat{h}(x) \neq h^*(x))\right] &= \sum_{z\in\{A,B\}} |\widetilde{F}_z(h^*) - F_z(h^*)| \cdot P_{(x,y)\sim\mathcal{D}}[x \in z] \\ &\leq \sum_{z\in\{A,B\}} \frac{\alpha}{(1-\alpha)P_{(x,y)\sim\mathcal{D}}[x \in z] + \alpha} \cdot P_{(x,y)\sim\mathcal{D}}[x \in z]\end{aligned}$$

by proposition 3

$$\leq \frac{2\alpha}{(1-\alpha)}$$

Note that even though the adversary can choose a different distribution at each timestep, we can wlog assume the adversary chooses the same distribution $\widetilde{D}$ where the quantity $|\widetilde{F}_z(h^*) - F_z(h^*)|$ is maximized at every timestep, as in Proposition 3. Although the model in [Kearns and Li, 1988b] is slightly weaker than [Konstantinov and Lampert, 2021], this theorem holds in full generality for both models where we replace the difference $|\widetilde{F}_z(h^*) - F_z(h^*)|$ with the bounds from Lemma 2 of [Konstantinov and Lampert, 2021]. The dependence on α remains the same in both cases. □

B.1 Equal Opportunity

Proposition 5 (TPR after corruption)**.** *Let $\widetilde{\mathcal{D}}$ be any corrupted distribution chosen by the adversary, and h be a fixed hypothesis in $\mathcal{H}$. For a fixed group A, the following inequality bounds the change in True Positive Rate of h:*

$$\left| TPR_A(h, \widetilde{\mathcal{D}}) - TPR_A(h, \mathcal{D}) \right| \leq \frac{\alpha}{(1-\alpha)r_A^+ + \alpha} \tag{3}$$

where $TPR_A(h, \mathcal{D}) = P_{(x,y)\sim\mathcal{D}_A}[h(x) = 1 | y = 1]$ and $r_A^+ = P_{(x,y)\sim\mathcal{D}}[y = 1 \cap x \in A]$

Proof of Proposition 5. For a fixed group A, the TPR of h in $\widetilde{\mathcal{D}}$ can be expressed in terms of the TPR of h in the original distribution $\mathcal{D}$ as follows:

$$\mathrm{TPR}_A(h, \widetilde{\mathcal{D}}) = \frac{(1-\alpha)\mathrm{TPR}_A(h, \mathcal{D}) \cdot P_{(x,y)\sim\mathcal{D}}[x \in A] + E_A^+}{(1-\alpha)P_{(x,y)\sim\mathcal{D}}[x \in A] + \alpha_A^+} \tag{7}$$

where α_A is the proportion of the data set that is corrupted and in group A and E_A^+ is the proportion of the data set that is corrupted, in group A, is positive, and is predicted as positive by h. Thus,

$$\left| \mathrm{TPR}_A(h, \widetilde{\mathcal{D}}) - \mathrm{TPR}_A(h, \mathcal{D}) \right| = \left| \frac{E_A - \alpha_A \mathrm{TPR}_A(h, \mathcal{D})}{(1-\alpha)P_{(x,y)\sim\mathcal{D}}[x \in A] + \alpha_A} \right| \leq \frac{\alpha}{(1-\alpha)r_A^+ + \alpha} \tag{8}$$

since $E_A \leq \alpha$ and $\alpha_A \leq \alpha$ □

Theorem 6 (Upper Bound)**.** *For any hypothesis class $\mathcal{H}$ and distribution $\mathcal{D} = (\mathcal{D}_A, \mathcal{D}_B)$, a robust fair-ERM learner for the equal opportunity constraint in the Malicious Adversarial Model returns a hypothesis $\hat{h}$ such that $\left| \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbf{1}(\hat{h}(x) \neq y)\right] - \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbf{1}(h^*(x) \neq y)\right] \right| \leq O(\sqrt{\alpha})$ where h^* is the optimal classifier for the fair-ERM problem on the true distribution $\mathcal{D}$ with respect to hypothesis class $\mathcal{H}$.*

Proof of Theorem 6. It suffices to show that there exists $h \in \mathcal{PQ}(\mathcal{H})$ that satisfies the guarantees above. Consider $h^* \in \mathcal{H}$. By the realizability assumption, h^* satisfies the equal opportunity constraint i.e $\mathrm{TPR}_A(h^*, \mathcal{D}) = \mathrm{TPR}_B(h^*, \mathcal{D})$. After the corruption, the equal opportunity violation of h^*, $|\mathrm{TPR}_A(h^*, \widetilde{\mathcal{D}}) - \mathrm{TPR}_B(h^*, \widetilde{\mathcal{D}})|$ may increase. Now we define the following parameters (p_z^i and q_z^i) for $i, z \in \{A, B\}$.

$$p_z^i = \begin{cases} \frac{\widetilde{F}_i(h^*) - \widetilde{F}_z(h^*)}{1 - \widetilde{F}_z(h^*)} & \text{if } \widetilde{F}_i(h^*) \geq \widetilde{F}_z(h^*) \\ \frac{\widetilde{F}_z(h^*) - \widetilde{F}_i(h^*)}{\widetilde{F}_z(h^*)} & \text{otherwise} \end{cases} \quad q_z^i = \begin{cases} 1 & \text{if } \widetilde{F}_i(h^*) \geq \widetilde{F}_z(h^*) \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

One can think of the parameter p_z^i as the proportion of samples in group z whose outcomes needs to be changed in order to match the true positivity rate of group i. Now consider two hypotheses $\hat{h}_i$ for $i \in \{A, B\}$ that behave as follows: Given a sample x:

- If $x \in A$, with probability p_A^i, return label q_A^i. Otherwise return $h^*(x)$
- Similarly, if $x \in B$, with probability p_B^i, return label q_B^i. Otherwise return $h^*(x)$

One can think of $\hat{h}_i$ as a hypothesis that deviates from h^* on every other group to make their true positive rate on the corrupted distribution match that of group i. Observe that $\hat{h}_i \in \mathcal{PQ}(\mathcal{H})$ for $i \in \{A, B\}$ since it follows the definition of our closure model $\mathcal{PQ}(\mathcal{H})$. We will now show that $\hat{h}_i$ for $i \in \{A, B\}$ satisfies the True Positive Rate constraint on the corrupted distribution (i.e $\widetilde{F}_A(\hat{h}_i) = \widetilde{F}_B(\hat{h}_i)$ for fixed $i \in \{A, B\}$). First, observe that for $z \in \{A, B\}$, if $\widetilde{F}_i(h^*) \geq \widetilde{F}_z(h^*)$, then $\widetilde{F}_z(\hat{h}_i) = \widetilde{F}_i(h^*)$. This is because

$$\begin{aligned}
\widetilde{F}_z(\hat{h}_i) &= (1 - p_z)\widetilde{F}_z(h^*) + p_z q_z \\
&= \widetilde{F}_z(h^*) + p_z(1 - \widetilde{F}_z(h^*)) \\
&= \widetilde{F}_z(h^*) + \widetilde{F}_i(h^*) - \widetilde{F}_z(h^*) \\
&= \widetilde{F}_i(h^*)
\end{aligned}$$

Similarly, if $\widetilde{F}_i(h^*) < \widetilde{F}_z(h^*)$, then $\widetilde{F}_z(\hat{h}) = \widetilde{F}_i(h^*)$. This is because

$$\begin{aligned}
\widetilde{F}_z(\hat{h}) &= (1 - p_z)\widetilde{F}_z(h^*) + p_z q_z \\
&= \widetilde{F}_z(h^*) + p_z(0 - \widetilde{F}_z(h^*)) \\
&= \widetilde{F}_z(h^*) + \widetilde{F}_i(h^*) - \widetilde{F}_z(h^*) \\
&= \widetilde{F}_i(h^*)
\end{aligned}$$

Thus, $\widetilde{F}_A(\hat{h}_i) = \widetilde{F}_i(h^*) = \widetilde{F}_B(\hat{h}_i)$. Therefore $\hat{h}_i$ for $i \in \{A, B\}$ satisfies the Equal Opportunity Constraint on the corrupted distribution.

We will now show that the existence of at least one $\hat{h}_i$ for $i \in \{A, B\}$ satisfies $\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbf{1}(\hat{h}(x) \neq y)\right] - \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbf{1}(h^*(x) \neq y)\right]\right| \leq O(\sqrt{\alpha})$. Since $\hat{h}_i$ deviates from h^* with probability p_A^i on samples from A, and with probability p_B^i on samples from B, it suffices to show that $p_A^i \cdot r_A + p_B^i \cdot r_B$ is $O(\sqrt{\alpha})$ for $i \in \{A, B\}$.

We consider the following cases:

1. Suppose wlog $r_B \leq \frac{\sqrt{\alpha}}{1-\sqrt{\alpha}}$. Then $\hat{h}_B$ satisfies the guarantee. This is because $p_A^B = 0$ (by equation 9) and $p_B^B \leq 1$. Thus, $p_A^B \cdot r_A + p_B^B \cdot r_B$ is $O(\sqrt{\alpha})$.

2. If instead $\min(r_A, r_B) > \frac{\sqrt{\alpha}}{1-\sqrt{\alpha}}$. wlog let B be a group with the highest true positive rate greater than 0.5 or the smallest true positive rate less than 0.5. At least one group must satisfy this constraint. If B has the highest true positive rate greater than 0.5, then

$$\begin{aligned}
p_B^A &= \frac{\widetilde{F}_B(h^*) - \widetilde{F}_A(h^*)}{\widetilde{F}_B(h^*)} \\
&\leq \frac{\widetilde{F}_B(h^*) - F_B(h^*) + F_A(h^*) - \widetilde{F}_A(h^*)}{0.5}
\end{aligned}$$

since $\widetilde{F}_B(h^*) \geq 0.5$ and by realizability assumption $F_B(h^*) = F_A(h^*)$

$$\leq 2|\widetilde{F}_B(h^*) - F_B(h^*)| + 2|F_A(h^*) - \widetilde{F}_A(h^*)|$$

by proposition 5

$$\leq O(\sqrt{\alpha})$$

Thus, $p_A^A \cdot r_A + p_B^A \cdot r_B$ is at most $O(\sqrt{\alpha})$ The case where B has the smallest true positive rate follows similarly.

Similar to the proof of Theorem 4, we can assume wlog the adversary chooses the same distribution $\widetilde{D}$ where the quantity $|\widetilde{F}_z(h^*) - F_z(h^*)|$ is maximized at every timestep, as in Proposition 3. Although the model in [Kearns and Li, 1988b] is slightly weaker than [Konstantinov and Lampert, 2021], this theorem holds in full generality for both models where we replace the difference $|\widetilde{F}_z(h^*) - F_z(h^*)|$ with the bounds from Lemma 5 of [Konstantinov and Lampert, 2021]. The dependence on α remains the same in both cases.

□

Theorem 7 (Lower Bound)**.** *There exists a distribution $\mathcal{D} = (\mathcal{D}_A, \mathcal{D}_B)$ and a malicious adversary of power α that guarantees that any hypothesis, $\hat{h}$, returned by an improper learner for the fair-ERM problem with the equal opportunity constraint satisfies the following: $\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbf{1}(\hat{h}(x) \neq y)\right] - \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbf{1}(h^*(x) \neq y)\right]\right| \geq \Omega(\sqrt{\alpha})$ where h^* is the optimal classifier for the fair-ERM problem on the true distribution $\mathcal{D}$ with respect to a hypothesis class $\mathcal{H}$.*

Proof of Theorem 7. We will show a distribution and a malicious adversary of power α such that any hypothesis returned by a learner incurs at least $\sqrt{\alpha}$ expected excess error. The distribution $\mathcal{D}$ will be such that $P_{x\sim\mathcal{D}}[x \in B] = \Omega(\sqrt{\alpha})$. This distribution will be supported on exactly four points $x_1 \in A, x_2 \in A, x_3 \in B, x_4 \in B$ with labels $y_1 = +, y_2 = -, y_3 = +, y_4 = -$. We also have that

$$P_{x,y\sim\mathcal{D}}[x = x_1, y = +] = P_{x,y\sim\mathcal{D}}[x = x_2, y = -] = \frac{1-\sqrt{\alpha}}{2}$$

and

$$P_{x,y\sim\mathcal{D}}[x = x_3, y = +] = P_{x,y\sim\mathcal{D}}[x = x_4, y = -] = \frac{\sqrt{\alpha}}{2}$$

That is, each group has equal proportion of positives and negatives.

The adversary commits to a poisoning strategy that places positive examples from Group B into the negative region of the optimal classifier. That is, the adversary changes the original distribution $\mathcal{D}$ so that

$$P_{x,y\sim\mathcal{D}}[x = x_1, y = +] = P_{x,y\sim\mathcal{D}}[x = x_2, y = -] = \frac{(1-\alpha)(1-\sqrt{\alpha})}{2}$$

$$P_{x,y\sim\mathcal{D}}[x = x_3, y = +] = P_{x,y\sim\mathcal{D}}[x = x_4, y = -] = \frac{(1-\alpha)\sqrt{\alpha}}{2}$$

and $P_{x,y\sim\mathcal{D}}[x = x_4, y = +] = \alpha$

We assume the perfect classifier is in the hypothesis class. Now fix a classifier h returned by a learner. This classifier must satisfy equal opportunity. Let p_1, p_2, p_3, p_4 be the probability that h classifies x_1, x_2, x_3, x_4 as positive, respectively. Observe that $\widetilde{\text{TPR}}(h_A) = p_1$ and $\widetilde{\text{TPR}}(h_B) = 1-(1-p_4)\alpha' - (1-p_3)(1-\alpha')$ where $\alpha' = \frac{2\sqrt{\alpha}}{(1-\alpha)+2\sqrt{\alpha}}$. The latter is due to the samples $(x_4, +)$ which the adversary added to the distribution. The adversary added an α amount which turned out to be an α' proportion of the positives in B. Since this classifier satisfies equal opportunity on the corrupted distribution, it must be the case that $p_1 = 1-(1-p_4)\alpha' - (1-p_3)(1-\alpha')$. Thus, $(1-p_1) \geq (1-p_4)\alpha'$. The error of h on the original distribution is therefore

$$\begin{aligned}
&(1-p_1+p_2)\frac{(1-\sqrt{\alpha})}{2} + (1-p_3+p_4)\frac{\sqrt{\alpha}}{2} \\
&\geq (1-p_1)\frac{(1-\sqrt{\alpha})}{2} + p_4\frac{\sqrt{\alpha}}{2}
\end{aligned}$$

by the equal opportunity constraint

$$\begin{aligned}
&\geq (1-p_4)\alpha' \frac{(1-\sqrt{\alpha})}{2} + p_4 \frac{\sqrt{\alpha}}{2} \\
&= (1-p_4) \cdot \frac{2\sqrt{\alpha}}{(1-\alpha)+2\sqrt{\alpha}} \cdot \frac{(1-\sqrt{\alpha})}{2} + p_4 \frac{\sqrt{\alpha}}{2} \\
&\geq (1-p_4)\frac{\sqrt{\alpha}}{2} + p_4 \frac{\sqrt{\alpha}}{2} \geq \Omega(\sqrt{\alpha})
\end{aligned}$$

□

C Equalized Odds

Now we will consider Equalized Odds.

Equalized Odds Proof of $\Omega(1)$ *accuracy loss:* it suffices to exhibit a 'bad' distribution and matching corruption strategy; which we exhibit below.

1. Say Group A has $1-\alpha$ of the probability mass i.e. $P_{(x,y)\sim\mathcal{D}}[x \in A] \geq 1-\alpha$ and thus $P_{(x,y)\sim\mathcal{D}}[x \in B] \leq \alpha$.
2. The positive fraction for each group under distribution $\mathcal{D}$ is $P_{(x,y)\sim\mathcal{D}_A}[y=1] = P_{(x,y)\sim D_B}[y=1] = \frac{1}{2}$
3. Since $P_{(x,y)\sim\mathcal{D}}[x \in B] \leq \alpha$, the adversary has sufficient corruption budget such that they can inject a duplicate copy of each example in B but with the opposite label. That is, for each example x in Group B in the training set, the adversary adds another identical example but with the opposite label.

This adversarial data ensures that on Group B, any hypothesis h (of any form) will now satisfy

$$P_{x\sim\hat{\mathcal{D}}_B}[h(x)=1|y=1] = P_{x\sim\hat{\mathcal{D}}_B}[h(x)=1|y=0] = p$$

for some value $p \in [0,1]$ due to the indistinguishable duplicated examples; i.e. the hypothesis can choose how often to accept examples [e.g. increase or decrease p] but it cannot distinguish positive/negative examples in Group B.

Note that we can select p using some arbitrary h but that randomness does not help us. Observe that similarly, the True Negative/False Negative Rates on Groyp B must be $1-p$.

Since A is evenly split among positive and negative and we must satisfy Equalized Odds, this means that our error rate on group A is

$$\begin{aligned}
&P_{(x,y)\sim\mathcal{D}_A}[h(x) \neq y] = P_{(x,y)\sim\mathcal{D}_A}[h(x) \neq y \cap y=1] + P_{(x,y)\sim\mathcal{D}_A}[h(x) \neq y \cap y=0] \\
&= P_{(x,y)\sim\mathcal{D}_A}[h(x) \neq 1|y=1]P[y=1] + P_{(x,y)\sim\mathcal{D}_A}[h(x) \neq 0|y=0]P[y=0] \\
&= P_{(x,y)\sim\mathcal{D}_A}[h(x) \neq 0|y=1]P[y=1] + P_{(x,y)\sim\mathcal{D}_A}[h(x) = 1|y=0]P[y=0] \\
&= (1-TPR_A)\frac{1}{2} + FPR_A\frac{1}{2} \\
&= (1-p)(\frac{1}{2}) + p(\frac{1}{2}) = \frac{1}{2}
\end{aligned}$$

So, the adversary has forced us to have 50% error on group A which yeilds the result. □

D Calibration Proofs

Proof of Theorem 10, Predictive Parity Lower Bound. To show that Predictive Parity requires $\Omega(1)$ error when the adversary has corruption budget α, even with our hypothesis class $\mathcal{PQ}(\mathcal{H})$, it suffices to exhibit a 'bad' distribution and matching corruption strategy; which we exhibit below.

Recall that we require that $P_{x\sim\mathcal{D}_A}[h(x)=1] > 0$ and $P_{x\sim\mathcal{D}_B}[h(x)=1] > 0$. This is to avoid the case where the learner rejects all points from Group B.

1. Assume that group A has $1-\alpha$ of the probability mass i.e. $P_{(x,y)\sim\mathcal{D}}[x \in A] \geq 1-\alpha$ and thus $P_{(x,y)\sim\mathcal{D}}[x \in B] \leq \alpha$.

2. The positive fraction for each group under distribution $\mathcal{D}$ is $P_{(x,y)\sim\mathcal{D}_A}[y=1] = P_{(x,y)\sim D_B}[y=1] = \frac{1}{2}$

3. Since $P_{(x,y)\sim\mathcal{D}}[x \in B] \leq \alpha$, the adversary has sufficient corruption budget such that they can a duplicate copy of each example in B but with the opposite label. That is, for each example x in Group B in the training set, the adversary adds another identical example but with the opposite label.

This adversarial data ensures that on Group B, any hypothesis h (of any form) will now satisfy

$$P_{(x,y)\sim\hat{\mathcal{D}}_B}[y=1|h(x)=1] = P_{(x,y)\sim\hat{\mathcal{D}}_B}[y=0|h(x)=0] = \frac{1}{2}$$

due to the indistinguishable duplicated examples. So, for Group A, to satisfy Predictive Parity, both these terms must also equal $\frac{1}{2}$ and induce 50% error on Group A. □

Proof of Theorem 12, Calibration $O(\alpha)$. In order to prove this statement, we consider h^* which is the Bayes Predictor $h^* = \mathbb{E}[y|x]$, but using some finite binning scheme $[R]$. Clearly h^* is calibrated on natural data and $h^* : \mathcal{X} \to [R]$.

We will show how to modify h^* to still satisfy the fairness constraint on the corrupted data without losing too much accuracy, regardless of the adversarial strategy.

In the case of Calibration, we will do this by just separately re-calibrating each group.

Let $[\hat{R}] := [R]$. We will now modify $[\hat{R}]$ from $[R]$ to be calibrated on the malicious data.

That is; For each group z (i.e $z = A$ or $z = B$), for each bin $r \in [R]$ (i.e., $x : h^*(x) = r$), we create a new bin if there is no bin in $[R]$ with value $\hat{r} = E_{(x,y)\sim\mathcal{D}_z}[y|h^*(x) = r]$.

That is, we define $\hat{h}(x) = \hat{r}$ for all $x \in g$ such that $h^*(x) = r$.

Observe that by construction, $\hat{h}$ is calibrated separately for each group, so it is calibrated overall. We just need to analyze the excess error of $\hat{h}$ compared to h^*. We will show this is only $O(\alpha)$.

Observe that increase in expected error is how much that bin is shifted from the true probability $h^*(x)$.

For each bin $r \in [R]$, the shift in $|r - \hat{r}|$ is at most the fraction of points in the bin that are malicious noise. Let $x \in MAL$ mean point x is a corrupted point.

Then

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{D}}[h^*(x) - \hat{h}(x)] \leq \sum_{r\in[R]} P[x \in r]|r - \hat{r}| \\
&= \sum_{r\in[R]} P[x \in r]\frac{P[x \in r \cap x \in MAL]}{P[x \in r]} \\
&\leq \sum_{r\in[R]} P[x \in r \cap x \in MAL] = O(\alpha) \quad \text{Definition of Malicious Noise Model}
\end{aligned}$$

Note that this is considering $L1$ error, accuracy loss is less than for $L2$ error, immediate for since $\alpha \in [0, 1)$.

□

E Minimax Fairness

In this Section, we will briefly and informally consider Minimax Fairness. Introduced in [Diana et al., 2020] this notion optimizes for a different objective.

Using their notation ($\epsilon_k = \mathbb{E}_{(x,y)\sim\mathcal{D}_k}[h(x) \neq y]$ or group-wise error) with a groupwise max error bound of $1 > \gamma > 0$

$$h^* = \underset{h\in\Delta H}{\operatorname{argmin}} \quad \mathbb{E}_{(x,y)\sim\mathcal{D}}[h(x) \neq y]$$
$$\max_{1\leq k\leq K} \epsilon_k(h) \leq \gamma$$

Letting OPT refer to the value of solution of the optimization problem, the learning goal is to find an h that is ϵ-approximately optimal for the mini-max objective, meaning that h satisfies:

$$max_k \epsilon_k(h) \leq OPT + \epsilon$$

Observe that if the goal of the learner is compete with the value of OPT on the unmodified data, in our malicious noise model this objective is ineffective since if one group is of size $O(\alpha)$, the adversary can always drive the error rate on that group $\Omega(1)$.

This model seems incompatible with malicious noise due to the sensitivity of minimax fairness to small groups.

Observe that the Minimax Fairness framework includes Equalized Error rates as a special case.