JMLR: Workshop and Conference Proceedings 1:1–25, 2021 Algorithmic Fairness through the Lens of Causality and Robustness

On the Impossibility of Fairness-Aware Learning from Corrupted Data

Nikola Konstantinov NIKOLA.KONSTANTINOV@IST.AC.AT
Christoph H. Lampert CHL@IST.AC.AT
IST Austria
Am Campus 1
3400 Klosterneuburg, Austria

Abstract

Addressing fairness concerns about machine learning models is a crucial step towards their long-term adoption in real-world automated systems. Many approaches for training fair models from data have been developed and an implicit assumption about such algorithms is that they are able to recover a fair model, despite potential historical biases in the data. In this work we show a number of impossibility results that indicate that there is no learning algorithm that can recover a fair model when a proportion of the dataset is subject to arbitrary manipulations. Specifically, we prove that there are situations in which an adversary can force any learner to return a biased classifier, with or without degrading accuracy, and that the strength of this bias increases for learning problems with underrepresented protected groups in the data. Our results emphasize on the importance of studying further data corruption models of various strength and of establishing stricter data collection practices for fairness-aware learning[1].

1. Introduction

Recent years have seen machine learning models greatly advancing the state-of-art performance of automated systems on many real-world tasks. As learned models become increasingly adopted in high-stake decision making, various fairness concerns arise. Indeed, it is now widely recognized that without addressing fairness issues during training, machine learning models can exhibit discriminatory behavior at prediction time (Barocas et al., 2019). Designing principled methods for certifying the fairness of a model is therefore key for increasing the trust in these methods among the general public.

To this end many ways of measuring and optimizing the fairness of learned models have been developed. The problem is perhaps best studied in the context of group fairness in classification, where the decisions of a binary classifier have to be nondiscriminatory with respect to a certain protected attribute, such as gender, race, etc. (Barocas et al., 2019). This is typically done by formulating a desirable fairness property for the task at hand and then optimizing for this property, alongside with accuracy, be it via a data preprocessing step, a modification of the training procedure, or by post-processing of a learned classifier on held-out data (Mehrabi et al., 2019). The underlying assumption is that by ensuring that

1. This paper is a shortened, workshop version of Konstantinov and Lampert (2021), https://arxiv.org/abs/2102.06004. For further results, including an analysis of algorithms achieving the lower bounds from this paper, we refer to the full version.

the fairness property holds exactly or approximately on the available data, one obtains a classifier whose decisions will also be fair at prediction time.

A major drawback of this framework is that for many real-world applications the training and validation data available are often times unreliable and biased (Biggio and Roli, 2018; Mehrabi et al., 2019). For example, demographic data collected via surveys or online polls is often difficult and expensive to verify. More generally any human-generated data is likely to contain various historical biases. Datasets collected via crowdsourcing or web crawing are also prone to both unwittingly created errors and conscious or even adversarially created biases.

These issues naturally raise concerns about the current practice of training and certifying fair models on such datasets. In fact, recent work has demonstrated that strong poisoning attacks can negatively impact the fairness of *specific learners* based on loss minimization. At the same time, little is known about the fundamental limits of fairness-aware learning from corrupted data. Previous work has only partially addressed the problem by studying weak data corruption models, for example by making specific label/attribute noise assumptions. However, these assumptions do not cover all possible (often unknown) problems that real-world data can possess. More generally, in order to avoid a cat-and-mouse game of designing defenses and attacks for fair machine learning models, one would need to be able to *certify fairness* as a property that holds when training under arbitrary, even adversarial, manipulations of the training data (Kearns and Li, 1993).

Contributions In our work, we address this gap by studying the effect of arbitrary data corruptions on fair learning algorithms. Our main contribution is a series of hardness results, which show that achieving fairness under worst-case data poisoning is provably impossible.

Specifically, we explore the fundamental limits of fairness-aware PAC learning within the classic *malicious adversary model* of Valiant (1985), where the adversary can replace a fraction of the data points with arbitrary data, with full knowledge of the learning algorithm, the data distribution and the remaining samples. We focus on binary classification with two popular group fairness constraints - demographic parity (Calders et al., 2009) and equal opportunity (Hardt et al., 2016).

First we show that learning under this adversarial model is provably impossible in a PAC sense - there is *no learning algorithm that can ensure convergence with high probability to a point on the accuracy-fairness Pareto front*, even in the limit of infinite training data. Furthermore, the irreducible error on the fairness measures we study is inversely proportional to the frequency of the rarer of the two protected attributes groups. This makes the robust learning problem especially hard when one of the protected subgroups in the data is underrepresented. These hardness results hold for *any learning algorithm* based on a corrupted dataset, including pre-, in- and post-processing methods in particular.

Perhaps an even more concerning result from a practical perspective is that the adversary can also ensure that any learning algorithm will output a classifier that is *optimal in terms of accuracy, but exhibits a large amount of unfairness*. The bias of such a classifier might go unnoticed for a long time in production systems, especially in applications where sensitive attributes are not revealed to the system at prediction time for privacy reasons.

We conclude with a discussion on the implications of our hardness results, emphasizing on the need for developing and studying further data corruption models for fairness-aware

learning, as well as on the importance of strict data collection practices in the context of fair machine learning.

2. Related work

To the best of our knowledge, we are the first to investigate the information theoretic limits of fairness-aware learning against a malicious adversary. There is, however, related previous work on PAC analysis of fair learning algorithms, robust fair learning, and learning with poisoned training data, that we discuss in this section.

Fairness in classification. Fairness-aware learning has been widely studied in the context of classification. We refer to Mehrabi et al. (2019) for an exhaustive introduction to the field. In this paper we focus on two popular notions of group fairness - demographic parity (Calders et al., 2009) and equal opportunity (Hardt et al., 2016). A number of hardness results for fair learning are already known. In particular, Kleinberg et al. (2017) prove the incompatability of three fairness notions for a broad class of learning problems and Menon and Williamson (2018b) quantify fundamental trade-offs between fairness and accuracy. Both of these works, however, focus on learning with i.i.d. clean data.

Fairness and data corruption. Most relevant for our setup are a number of recent works that empirically study attacks and defenses on fair learners under adversarial data poisoning. In particular, Solans et al. (2020); Chang et al. (2020); Mehrabi et al. (2020) consider practical, gradient-based poisoning attacks against machine learning algorithms. All of these works demonstrate empirically that poisoned data can severely damage the performance of fair learners that are based on empirical loss minimization. In our work we go beyond this by proving a set of hardness results that hold for *arbitrary learning algorithms*.

Among works focusing on weaker adversarial models, a particularly popular topic is the one of fair learning with noisy or adversarially perturbed sensitive attributes (Lamy et al., 2019; Awasthi et al., 2020; Wang et al., 2020b; Celis et al., 2021a; Mehrotra and Celis, 2021; Celis et al., 2021b). Under the explicit assumption that the corruption does not effect the inputs and the labels, these works propose algorithms that can recover a fair model despite the data corruption. A related, but conceptually different topic is the one of fair learning without demographic information (Hashimoto et al., 2018; Kallus et al., 2020; Mozannar et al., 2020; Lahoti et al., 2020). Another commonly assumed type of corruption is label noise, which is shown to be overcomable under various assumptions by De-Arteaga et al. (2018); Jiang and Nachum (2020); Wang et al. (2020a); Fogliato et al. (2020). A distributionally robust approach for certifying fairness is taken by Taskesen et al. (2020), under the assumption that the real data distribution falls within a Wasserstein ball centered at the empirical data distribution. In Ignatiev et al. (2020) a formal methods framework for certifying fairness through unawareness, even in the presence of a specific type of data bias that targets their desired fairness measure, is provided. The vulnerability of fair learning algorithms to specific types of data corruption has also been demonstrated on real-world data by Calders and Žliobaitė (2013); Kallus and Zhou (2018).

An orthogonal line of work shows that imposing fairness constraints can neutralize the effects of corrupted data, under specific assumptions on the type of bias present (Blum and Stangl, 2020). Also related are the works of Tae et al. (2019); Li et al. (2021) who propose

procedures for data cleaning/outlier detection, without a specific adversarial model, that in particular improve fairness performance. Finally, the work of Lechner et al. (2021) also studies fairness-aware (representation) learning in the presence of a malicious opponent. However, in their setting the adversary can manipulate the classifier chosen on top of the representation, not the training data.

Learning against an adversary. Learning from corrupted training data is a field with long history, where both theoretical and practical aspects of attacking and defending ML models have been widely studied (Angluin and Laird, 1988; Kearns and Li, 1993; Cesa-Bianchi et al., 1999; Bshouty et al., 2002; Biggio et al., 2012; Charikar et al., 2017; Steinhardt et al., 2017; Chen et al., 2017; Diakonikolas et al., 2019). In this work we study fair learning within the so-called malicious adversary model, introduced by Valiant (1985). The fundamental limits of classic PAC learning in this context have been extensively explored by Kearns and Li (1993); Cesa-Bianchi et al. (1999). Our paper adds an additional dimension to this line of work, where fairness is considered alongside with accuracy as an objective for the learner.

3. Preliminaries

In this section we formalize the problem of fairness-aware learning against a malicious adversary, by giving precise definitions of the learning objectives and the studied data corruption model.

3.1. Fairness-aware learning

We adopt the following standard group fairness classification framework. We consider a product space $\mathcal{X} \times A \times \mathcal{Y}$, where $\mathcal{X}$ is an input space, $\mathcal{Y} = \{0, 1\}$ is a binary label space and $A = \{0, 1\}$ is a set corresponding to a binary protected attribute (for example, race or gender). We assume that there is an unknown true data distribution $\mathbb{P} \in \mathcal{P}(\mathcal{X} \times A \times \mathcal{Y})$ from which the clean data is sampled. Denote by $\mathcal{H} \subseteq \{h : \mathcal{X} \to \mathcal{Y}\}$ the hypothesis space of all classifiers to be considered.

PAC learning Adopting a statistical PAC learning setup, we are interested in designing learning procedures that find a classifier based on training examples. Formally, a (statistical) fairness-aware learner $\mathcal{L} : \cup_{n\in\mathbb{N}}(\mathcal{X} \times A \times \mathcal{Y})^n \to \mathcal{H}$ is a function that takes a labeled dataset of an arbitrary size and outputs a hypothesis. Note that we consider learning in the purely statistical sense here, focusing on *any* procedure that outputs a hypothesis, regardless of its computational complexity, and seeking learners that are sample-efficient instead.

In a clean data setup, the learner is trained on a dataset $S^c = \{(x_i^c, a_i^c, y_i^c)\}_{i=1}^n$ sampled i.i.d. from $\mathbb{P}$ and outputs a hypothesis $h := \mathcal{L}(S^c)$. The performance of a learner can be measured via the standard 0/1 loss (a.k.a. the risk) with respect to the distribution $\mathbb{P}$

$$\mathcal{R}(h, \mathbb{P}) = \mathbb{P}(h(X) \neq Y). \tag{1}$$

Group fairness in classification In (group) fairness-aware learning, an additional desirable property of the classifier $h = \mathcal{L}(S^c)$ is that its decisions are fair in the sense that it does not exhibit discrimination with respect to one of the protected subgroups in the population. Many different formal notions of group fairness have previously been

proposed in the literature. The problem of selecting the "right" fairness measure is in general application-dependent and beyond the scope of this work.

Here we focus on the two arguably most widely adopted measures. The first one, *demographic parity* (Calders et al., 2009), requires that the decisions of the classifier are independent of the protected attribute, that is

$$\mathbb{P}(h(X) = 1|A = 0) = \mathbb{P}(h(X) = 1|A = 1). \tag{2}$$

The second one, *equal opportunity* (Hardt et al., 2016), states that the true positive rates of the classifier should be equal across the protected groups, that is

$$\mathbb{P}(h(X) = 1|A = 0, Y = 1) = \mathbb{P}(h(X) = 1|A = 1, Y = 1). \tag{3}$$

In practice, it is rarely the case that a classifier achieves perfect fairness. Therefore, we will instead be interested in controlling the *amount of unfairness* that h possesses, measured via corresponding fairness deviation measures $\mathcal{D}(h)$ (Woodworth et al., 2017; Menon and Williamson, 2018a; Williamson and Menon, 2019). Here we adopt the *mean difference score* measure of Calders and Verwer (2010); Menon and Williamson (2018a) for demographic parity

$$\mathcal{D}^{par}(h, \mathbb{P}) = |\mathbb{P}(h(X) = 1|A = 0) - \mathbb{P}(h(X) = 1|A = 1)| \tag{4}$$

and its analog for equal opportunity

$$\mathcal{D}^{opp}(h, \mathbb{P}) = |\mathbb{P}(h(X) = 1|A = 0, Y = 1) - \mathbb{P}(h(X) = 1|A = 1, Y = 1)|\,. \tag{5}$$

To avoid degenerate cases for these measures, we assume throughout the paper that $P_a = \mathbb{P}(A = a) > 0$ and $P_{1a} = \mathbb{P}(Y = 1, A = a) > 0$ for both $a \in \{0, 1\}$. For the rest of the paper, whenever we are interested in demographic parity fairness, we assume without loss of generality that $A = 0$ is the minority class, so that $P_0 \leq \frac{1}{2} \leq P_1$. Similarly, whenever the fairness notion is equal opportunity, we will assume that $P_{10} \leq P_{11}$.

Whenever the underlying distribution is clear from the context, we will drop the dependence of $\mathcal{R}(h, \mathbb{P})$ and $\mathcal{D}(h, \mathbb{P})$ on $\mathbb{P}$ and simply write $\mathcal{R}(h)$ and $\mathcal{D}(h)$.

3.2. Adversarial model

As argued in the introduction, machine learning models are often trained on unreliable datasets, where some of the points might be corrupted by noise, human biases and/or malicious agents. To model arbitrary manipulations of the data, we assume the presence of an adversary that can modify a certain fraction of the dataset and study fair learning in this context. In addition to its assumption-free nature, this worst-case approach can be seen as a tool for providing a *certificate for fairness*: if a system can work against a strong adversarial model, it will be effective under *any circumstances that are covered by the model*.

Formally a *fairness-aware adversary* is any procedure for manipulating a dataset, that is a *possibly randomized function* $\mathcal{A} : (\mathcal{X} \times A \times \mathcal{Y})^n \to (\mathcal{X} \times A \times \mathcal{Y})^n$ that takes in a clean dataset $S^c = \{(x_i^c, a_i^c, y_i^c)\}_{i=1}^n$ sampled i.i.d. from $\mathbb{P}$ and outputs a new, corrupted, dataset $S^p = \{(x_i^p, a_i^p, y_i^p)\}_{i=1}^n$. Depending on the type of restrictions that are imposed on the adversary, various adversarial models can be obtained.

In this work we adopt the powerful *malicious adversary model*, first introduced by Valiant (1985) and extensively studied by Kearns and Li (1993); Cesa-Bianchi et al. (1999). The formal data generating procedure is as follows:

- An i.i.d. *clean dataset* $S^c = \{(x_i^c, a_i^c, y_i^c)\}_{i=1}^n$ is sampled from $\mathbb{P}$.
- Each index/point $i \in \{1, 2, \ldots, n\}$ is *marked* independently with probability α, for a fixed constant $\alpha \in [0, 0.5)$. Denote all marked indexes by $\mathfrak{P} \subseteq [n]$.
- The *malicious adversary* computes, in a possibly randomized manner, a corrupted dataset $S^p = \{(x_i^p, a_i^p, y_i^p)\}_{i=1}^n \in (\mathcal{X} \times A \times \mathcal{Y})^n$, with the only restriction that $(x_i^p, a_i^p, y_i^p) = (x_i^c, a_i^c, y_i^c)$ for all $i \notin \mathfrak{P}$. That is, the adversary can replace all marked data points in an arbitrary manner. Note that *no assumptions whatsoever* can be made about the points (x_i^p, a_i^p, y_i^p) for $i \in \mathfrak{P}$.
- The corrupted dataset S^p is then passed on to the learner, who computes $\mathcal{L}(S^p)$.

For a fixed $\alpha \in [0, 0.5)$, we say that $\mathcal{A}$ is a malicious adversary of power α. Note that the number of marked points is $|\mathfrak{P}| = Bin(n, \alpha)$. Since no assumptions are made on the corrupted data points, they can, in particular, depend on the learner $\mathcal{L}$, the data distribution $\mathbb{P}$, the clean data S^c and all other parameters of the learning problem. That is, the adversary acts with full knowledge of the learning setup and without any computational constraints.

3.3. Fairness-aware learning against an adversary

Structure of the hardness results In the next section we will be showing lower bounds on $\mathcal{R}(\mathcal{L}(S^p))$ and $\mathcal{D}(\mathcal{L}(S^p))$, that is, the risk and the fairness deviation measure achieved by the learner when trained on the corrupted data. Our bounds can be thought of as hardness results that describe a limit on how well the learner can perform against the adversary. These are based on explicit constructions of hard learning problems and adversaries that demonstrate these limitations.

Crucial in these results is the ordering of the predicates. These matter for the sake of formalizing the powers of the adversary and the learner. Recall that the learner only operates with knowledge of the corrupted dataset. At the same time, the adversary is assumed to know not only the clean data, but also the target distribution and the learner. Therefore, our lower bounds are structured as follows:

For any learner $\mathcal{L}$ there exists a distribution $\mathbb{P}$ and an adversary $\mathcal{A}$, such that with constant probability ...

Note in particular that the adversary can be chosen after the learner is constructed and together with the distribution and it can therefore be tailored to their choice. On the other hand, the learner is fixed before the distribution and the adversary are, so it has to work for any such pair.

We note that all probability statements in our theorems refer to the randomness in the full generation process of the dataset S^p, that is the randomness of the clean data, the marked points and the adversary.

Role of the hypothesis space Learnability in our setup can be studied either as a property of any fixed hypothesis space, or as a property of a class of hypothesis spaces, for example the hypothesis spaces of finite size or finite VC dimension. However, one can easily see that for certain hypothesis spaces fairness can be satisfied trivially. For example, whenever $\mathcal{H}$ contains a classifier that is constant on the whole input space (that is, always predicts 1 or always predicts 0), a learner that returns this constant classifier, regardless of the observed data, will always be perfectly fair with respect to both fairness notions, under any distribution and against any adversary. We therefore opt to study the learnability of *classes of hypothesis spaces.*

In particular, our hardness results demonstrate the *existence of a finite hypothesis space*, such that a certain amount of inaccuracy and/or unfairness is unavoidable. Therefore, no learner can achieve better guarantees on the class of all finite hypothesis spaces, even in the infinite training data limit. This is in contrast to, for example, classic PAC learning with clean data, where the ERM algorithm is a PAC learner for all finite hypothesis spaces and more generally all spaces of finite VC dimension (Shalev-Shwartz and Ben-David, 2014).

Parameters of the learning problem Our bounds will depend explicitly on the corruption ratio α and on the smaller of the protected class frequencies $P_0 = \mathbb{P}(A = 0)$ (for demographic parity) or on $P_{10} = \mathbb{P}(Y = 1, A = 0) \leq \mathbb{P}(Y = 1, A = 1)$ (for equal opportunity). To understand the limits of fairness-aware learning against a malicious adversary, we will analyze our bounds for small values of α and P_0 or P_{10}. Intuitively, the smaller the corruption rate α is, the easier it is for the learner to recover an accurate and fair hypothesis. On the other hand, a small value for P_0 or P_{10} implies that one of the subgroups is underrepresented in the population, and so intuitively the adversary can hide a lot of information about this group and thus prevent the learner from finding a fair hypothesis.

As we will see, this intuition is reflected in our bounds, which give a tool for comparing these quantities in terms of their effect on the hardness of the learning problem.

4. Impossibility results

We now present a series of hardness results that demonstrate that fair learning in the presence of a malicious adversary is provably impossible in a PAC learning sense. **Complete proofs of all results can be found in the supplementary material.**

4.1. Pareto lower bounds

We begin by presenting two hardness results that intuitively show that for some choices of $\mathcal{H}$ the adversary can prevent any learner from reaching the Pareto front of the accuracy-fairness optimization problem. We first demonstrate this for demographic parity:

Theorem 1 *Let $0 \leq \alpha < 0.5, 0 < P_0 \leq 0.5$. For any input set $\mathcal{X}$ with at least four distinct points, there exists a finite hypothesis space $\mathcal{H}$, such that for any learning algorithm $\mathcal{L} : \cup_{n\in\mathbb{N}}(\mathcal{X} \times A \times \mathcal{Y})^n \to \mathcal{H}$, there exists a distribution $\mathbb{P}$ for which $\mathbb{P}(A = 0) = P_0$, a malicious adversary $\mathcal{A}$ of power α and a hypothesis $h^* \in \mathcal{H}$, such that with probability at least* 0.5

$$\mathcal{R}(\mathcal{L}(S^p), \mathbb{P}) - \mathcal{R}(h^*, \mathbb{P}) \geq \min\left\{\frac{\alpha}{1-\alpha}, 2P_0P_1\right\}$$

and

$$\mathcal{D}^{par}(\mathcal{L}(S^p), \mathbb{P}) - \mathcal{D}^{par}(h^*, \mathbb{P}) \geq \min\left\{\frac{\alpha}{2P_0P_1(1-\alpha)}, 1\right\}.$$

The proof, which can be found in the supplementary material, is based on the so-called *method of induced distributions*, pioneered by Kearns and Li (1993). The idea is to construct two distributions that are sufficiently different, yet can be made indistinguishable by the adversary. Then, no learner can be "correct" with high probability on both distributions and so any learner will incur a high loss and exhibit high unfairness on at least one of them.

Discussion. Our hardness result implies that no learner can guarantee reaching a point on the Pareto front in a PAC learning sense, even for a simple family of hypothesis spaces, namely the finite ones. To prove the theorem we explicitly construct a hypothesis space that is not learnable against the malicious adversary. As discussed in Section 3.3, a constructive proof is necessary here, because fairness can be trivially satisfied on some hypothesis spaces, for example those that contain a constant classifier, which is fair under any distribution and against any adversary.

We now analyze the bounds and their behavior for small values of α and P_0. First assume that $\frac{\alpha}{1-\alpha} < 2P_0P_1$, which in particular is the case whenever $2\alpha < P_0$. Then under the conditions of the theorem, with probability at least 0.5[2]

$$\mathcal{R}(\mathcal{L}(S^p)) - \mathcal{R}(h^*) \geq \Omega(\alpha) \tag{6}$$

and

$$\mathcal{D}^{par}(\mathcal{L}(S^p)) - \mathcal{D}^{par}(h^*) \geq \Omega\left(\frac{\alpha}{P_0}\right). \tag{7}$$

The lower bound on the loss is known to hold for any hypothesis space as shown by Kearns and Li (1993). What Theorem 1 adds to this classic result is that for certain hypothesis spaces 1) the learner can at the same time be forced to produce an unfair classifier; 2) the fairness deviation measure $\mathcal{D}^{par}$ can be increased by $\Omega(\alpha/P_0)$. Note that *these results hold regardless of the sample size n, and so even in the infinite data limit.*

In the second case, when $\frac{\alpha}{1-\alpha} \geq 2P_0P_1$, the adversary can force a constant increase in the loss and make the classifier completely unfair, making $\mathcal{D}^{par}(\mathcal{L}(S^p)) = 1$. These observations, combined with the rates from the first case, indicate that unless $\alpha = o(P_0)$, the adversary can ensure that the resulting model's demographic parity deviation measure is constant. In particular, *if one of the protected groups is rare, even very small levels of data corruption can lead to a biased model.*

Next we show a similar result for equal opportunity.

Theorem 2 *Let $0 \leq \alpha < 0.5$ and $P_{10} \leq P_{11} < 1$ be such that $P_{10} + P_{11} < 1$. For any input set $\mathcal{X}$ with at least five distinct points, there exists a finite hypothesis space $\mathcal{H}$, such that for any learning algorithm $\mathcal{L} : \cup_{n\in\mathbb{N}}(\mathcal{X} \times A \times \mathcal{Y})^n \to \mathcal{H}$, there exists a distribution $\mathbb{P}$ for which*

2. We use the Ω-notation for lower bounds on the growth rates of functions.

$\mathbb{P}(A = a, Y = 1) = P_{1a}$ *for* $a \in \{0, 1\}$*, a malicious adversary* $\mathcal{A}$ *of power* α *and a hypothesis* $h^* \in \mathcal{H}$*, such that with probability at least* 0.5

$$\mathcal{R}(\mathcal{L}(S^p), \mathbb{P}) - \mathcal{R}(h^*, \mathbb{P}) \geq \min\left\{\frac{\alpha}{1-\alpha}, 2P_{10}, 2(1 - P_{10} - P_{11})\right\}$$

and

$$\mathcal{D}^{opp}(\mathcal{L}(S^p), \mathbb{P}) - \mathcal{D}^{opp}(h^*, \mathbb{P}) \geq \min\left\{\frac{\alpha}{2(1-\alpha)P_{10}}, 1, \frac{1 - P_{10} - P_{11}}{P_{10}}\right\}.$$

Discussion. A similar analysis to the one after Theorem 1 applies here as well. In particular, whenever $\frac{\alpha}{1-\alpha} \leq 2\min\{P_{10}, 1 - P_{10} - P_{11}\}$, we obtain

$$\mathcal{R}(\mathcal{L}(S^p)) - \mathcal{R}(h^*) \geq \Omega(\alpha) \tag{8}$$

and

$$\mathcal{D}^{par}(\mathcal{L}(S^p)) - \mathcal{D}^{par}(h^*) \geq \Omega\left(\frac{\alpha}{P_{10}}\right). \tag{9}$$

The case when $\frac{\alpha}{1-\alpha} > 2\min\{P_{10}, 1 - P_{10} - P_{11}\}$ leads to a constant equal opportunity deviation measure. If in addition we have that $1 - P_{10} - P_{11} \geq P_{10}$, a completely unfair classifier will be returned. Consequently, if positive examples associated with one of the protected groups are rare (that is, if $\mathbb{P}(Y = 1, A = 0)$ is small), then even very small corruption ratios can lead to a biased model.

4.2. Hurting fairness without affecting accuracy

While the results above shed light on the fundamental limits of robust fairness-aware learning against an adversary, models that are inaccurate are often easy to detect in practice. On the other hand, a model that has good accuracy, but exhibits a bias with respect to the protected attribute can be much more problematic. This is especially true in applications where demographic data is not collected at prediction time for privacy reasons. In this case the model's bias might go unnoticed for a long time, thus adversely affecting one of the population subgroups and potentially extrapolating existing biases from the training data to future decisions.

We now show that such an unfortunate situation is indeed also possible under the malicious adversary model. The following results show that any learner will, in some situations, be forced by the adversary to return a model that is optimal in terms of accuracy, but exhibits high unfairness in terms of demographic parity.

Theorem 3 *Let* $0 \leq \alpha < 0.5, 0 < P_0 \leq 0.5$*. For any input set* $\mathcal{X}$ *with at least four distinct points, there exists a finite hypothesis space* $\mathcal{H}$*, such that for any learning algorithm* $\mathcal{L} : \cup_{n\in\mathbb{N}}(\mathcal{X} \times A \times \mathcal{Y})^n \to \mathcal{H}$*, there exists a distribution* $\mathbb{P}$ *for which* $\mathbb{P}(A = 0) = P_0$*, a malicious adversary* $\mathcal{A}$ *of power* α *and a hypothesis* $h^* \in \mathcal{H}$*, such that with probability at least* 0.5

$$\mathcal{R}(\mathcal{L}(S^p), \mathbb{P}) = \mathcal{R}(h^*, \mathbb{P}) = \min_{h\in\mathcal{H}} \mathcal{R}(h, \mathbb{P})$$

and

$$\mathcal{D}^{par}(\mathcal{L}(S^p), \mathbb{P}) - \mathcal{D}^{par}(h^*, \mathbb{P}) \geq \min\left\{\frac{\alpha}{2P_0}, 1\right\}.$$

We also present a corresponding result for equal opportunity.

Theorem 4 *Let $0 \leq \alpha < 0.5, P_{10} \leq P_{11} < 1$ be such that $P_{10} + P_{11} < 1$. For any input set $\mathcal{X}$ with at least five distinct points, there exists a finite hypothesis space $\mathcal{H}$, such that for any learning algorithm $\mathcal{L} : \cup_{n\in\mathbb{N}}(\mathcal{X} \times A \times \mathcal{Y})^n \to \mathcal{H}$, there exists a distribution $\mathbb{P}$ for which $\mathbb{P}(A = a, Y = 1) = P_{1a}$ for $a \in \{0, 1\}$, a malicious adversary $\mathcal{A}$ of power α and a hypothesis $h^* \in \mathcal{H}$, such that with probability at least* 0.5

$$\mathcal{R}(\mathcal{L}(S^p), \mathbb{P}) = \mathcal{R}(h^*, \mathbb{P}) = \min_{h\in\mathcal{H}} \mathcal{R}(h, \mathbb{P})$$

and

$$\mathcal{D}^{opp}(\mathcal{L}(S^p), \mathbb{P}) - \mathcal{D}^{opp}(h^*, \mathbb{P}) \geq \min\left\{ \frac{\alpha}{2(1-\alpha)P_{10}} \left(1 - \frac{P_{10}}{P_{11}}\right), 1 - \frac{P_{10}}{P_{11}} \right\}.$$

Once again the error terms on the fairness notions are inversely proportional to P_0 and P_{10} respectively, indicating that datasets in which one of the subgroups is underrepresented are particularly vulnerable to data manipulations.

5. Discussion

In this work we explored the statistical limits of fairness-aware learning algorithms on corrupted data, under the malicious adversary model. Our results show that data manipulations can have an inevitable negative effect on model fairness and that this effect is even more expressed for problems where a subgroup in the population is underrepresented.

While the strong adversarial model and the statistical PAC learning analysis we have considered are mostly of theoretical interest, we believe that the hardness results have several important implications. Indeed, crucial to increasing the trust in learned decision making systems is the ability to guarantee that they exhibit a high amount of fairness, regardless of any known or unforeseen biases in the training data. In contrast, we have shown that this is provably impossible under a strong adversarial model for the data corruption.

We believe that these results stress on the importance of developing and studying further data corruption models in the context of fairness-aware learning. As discussed in the related work, previous research has shown that it is possible to recover a fair model under corruptions of the labels or protected attributes only. While real-world data is likely to contain more subtle manipulations, one may hope that for certain applications there will be models of data corruption that are, on the one hand, sufficiently broad to cover the data issues and, on the other hand, specific enough so that fair learning becomes possible.

Our results can also be seen as a indication that strict data collection practices may in fact be necessary for designing provably fair machine learning models. Indeed, our bounds hold under the assumption that the learner can only access one dataset of unknown quality. In contrast, it has been shown that the use of even a small trusted dataset (that is, a certified clean subset of the data) can greatly improve the performance of machine learning models under corruption, both in the context of classic PAC learning (Hendrycks et al., 2018; Konstantinov and Lampert, 2019) and in the context of fairness-aware learning (Roh et al., 2020). Such data can also be helpful for the sake of validating the fairness of a model as a precautionary step before its real-world adoption.

In summary, understanding and accounting for the types of biases present in machine learning datasets is crucial for addressing the issues brought up in this work and for the development of certifiably fair learning models.

References

Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4): 343–370, 1988.

Pranjal Awasthi, Matthäus Kleindessner, and Jamie Morgenstern. Equalized odds postprocessing under imperfect group information. In *Conference on Uncertainty in Artificial Intelligence (AISTATS)*, 2020.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 2018.

Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *International Conference on Machine Learing (ICML)*, 2012.

Avrim Blum and Kevin Stangl. Recovering from biased data: Can fairness constraints improve accuracy? In *Foundations of Responsible Computing*, volume 156. Schloss Dagstuhl – Leibniz Center for Informatics, 2020.

Nader H Bshouty, Nadav Eiron, and Eyal Kushilevitz. Pac learning with nasty noise. *Theoretical Computer Science (TCC)*, 288(2):255–275, 2002.

Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery (DMKD)*, 2010.

Toon Calders and Indrė Žliobaitė. Why unbiased computational processes can lead to discriminative decision procedures. In *Discrimination and privacy in the information society*, pages 43–57. Springer, 2013.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *International Conference on Data Mining Workshops (IDCMW)*, 2009.

L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Fair classification with noisy protected attributes: A framework with provable guarantees. In *International Conference on Machine Learing (ICML)*, 2021a.

L Elisa Celis, Anay Mehrotra, and Nisheeth K Vishnoi. Fair classification with adversarial perturbations. *arXiv preprint arXiv:2106.05964*, 2021b.

Nicolo Cesa-Bianchi, Eli Dichterman, Paul Fischer, Eli Shamir, and Hans Ulrich Simon. Sample-efficient strategies for learning in the presence of noise. *Journal of the ACM (JACM)*, 1999.

Hongyan Chang, Ta Duy Nguyen, Sasi Kumar Murakonda, Ehsan Kazemi, and Reza Shokri. On adversarial bias and the robustness of fair machine learning. *arXiv preprint arXiv:2006.08669*, 2020.

Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Symposium on Theory of Computing (STOC)*, 2017.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

Maria De-Arteaga, Artur Dubrawski, and Alexandra Chouldechova. Learning under selective labels in the presence of expert consistency. In *Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)*, 2018.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. In *International Conference on Machine Learing (ICML)*, 2019.

Riccardo Fogliato, Max G'Sell, and Alexandra Chouldechova. Fairness evaluation in presence of biased noisy labels. In *Conference on Uncertainty in Artificial Intelligence (AISTATS)*, 2020.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2016.

Tatsunori B Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learing (ICML)*, 2018.

Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. *Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

Alexey Ignatiev, Martin C Cooper, Mohamed Siala, Emmanuel Hebrard, and Joao Marques-Silva. Towards formal fairness in machine learning. In *International Conference on Principles and Practice of Constraint Programming (CP)*, 2020.

Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. In *Conference on Uncertainty in Artificial Intelligence (AISTATS)*, 2020.

Nathan Kallus and Angela Zhou. Residual unfairness in fair machine learning from prejudiced data. In *International Conference on Machine Learing (ICML)*, 2018.

Nathan Kallus, Xiaojie Mao, and Angela Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. In *Conference on Fairness, Accountability and Transparency (FAccT)*, 2020.

Michael Kearns and Ming Li. Learning in the presence of malicious errors. *SIAM Journal on Computing (SICOMP)*, 1993.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *Innovations in Theoretical Computer Science Conference (ITCS)*, 2017.

Nikola Konstantinov and Christoph Lampert. Robust learning from untrusted sources. In *International Conference on Machine Learing (ICML)*, 2019.

Nikola Konstantinov and Christoph H Lampert. Fairness-aware pac learning from corrupted data. *arXiv preprint arXiv:2102.06004*, 2021.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H Chi. Fairness without demographics through adversarially reweighted learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Alex Lamy, Ziyuan Zhong, Aditya K Menon, and Nakul Verma. Noise-tolerant fair classification. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

Tosca Lechner, Shai Ben-David, Sushant Agarwal, and Nivasini Ananthakrishnan. Impossibility results for fair representations. *arXiv preprint arXiv:2107.03483*, 2021.

Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

Ninareh Mehrabi, Muhammad Naveed, Fred Morstatter, and Aram Galstyan. Exacerbating algorithmic bias through fairness attacks. *arXiv preprint arXiv:2012.08723*, 2020.

Anay Mehrotra and L Elisa Celis. Mitigating bias in set selection with noisy protected attributes. In *Conference on Fairness, Accountability and Transparency (FAccT)*, 2021.

Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency (FAccT)*, 2018a.

Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)*, 2018b.

Hussein Mozannar, Mesrob I Ohannessian, and Nathan Srebro. Fair learning with private demographic data. In *International Conference on Machine Learing (ICML)*, 2020.

Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. FR-Train: A mutual information-based approach to fair and robust training. In *International Conference on Machine Learing (ICML)*, 2020.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

David Solans, Battista Biggio, and Carlos Castillo. Poisoning attacks on algorithmic fairness. In *European Conference on Machine Learning and Data Mining (ECML PKDD)*, 2020.

Jacob Steinhardt, Pang Wei W Koh, and Percy S Liang. Certified defenses for data poisoning attacks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

Ki Hyun Tae, Yuji Roh, Young Hun Oh, Hyunsu Kim, and Steven Euijong Whang. Data cleaning for accurate, fair, and robust models: A big data-AI integration approach. In *International Workshop on Data Management for End-to-End Machine Learning (DEEM)*, 2019.

Bahar Taskesen, Viet Anh Nguyen, Daniel Kuhn, and Jose Blanchet. A distributionally robust approach to fair classification. *arXiv preprint arXiv:2007.09530*, 2020.

Leslie G Valiant. Learning disjunction of conjunctions. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 1985.

Jialu Wang, Yang Liu, and Caleb Levy. Fair classification with group-dependent label noise. *arXiv preprint arXiv:2011.00379*, 2020a.

Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Michael I Jordan. Robust optimization for fairness with noisy protected groups. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020b.

Robert Williamson and Aditya Menon. Fairness risk measures. In *International Conference on Machine Learing (ICML)*, 2019.

Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Workshop on Computational Learning Theory (COLT)*, 2017.