Journal of Machine Learning Research 23 (2022) 1-60 Submitted 10/21; Revised 3/22; Published 5/22

Fairness-Aware PAC Learning from Corrupted Data

Nikola Konstantinov[1] NIKOLAHRISTOV.KONSTANTINOV@INF.ETHZ.CH
Post-Doctoral Fellow, ETH AI Center
Universitätstrasse 6
8092 Zürich, Switzerland

Christoph H. Lampert CHL@IST.AC.AT
Institute of Science and Technology Austria (ISTA)
Am Campus 1
3400 Klosterneuburg, Austria

Editor: Pradeep Ravikumar

Abstract

Addressing fairness concerns about machine learning models is a crucial step towards their long-term adoption in real-world automated systems. While many approaches have been developed for training fair models from data, little is known about the robustness of these methods to data corruption. In this work we consider fairness-aware learning under worst-case data manipulations. We show that an adversary can in some situations force any learner to return an overly biased classifier, regardless of the sample size and with or without degrading accuracy, and that the strength of the excess bias increases for learning problems with underrepresented protected groups in the data. We also prove that our hardness results are tight up to constant factors. To this end, we study two natural learning algorithms that optimize for both accuracy and fairness and show that these algorithms enjoy guarantees that are order-optimal in terms of the corruption ratio and the protected groups frequencies in the large data limit.

Keywords: Fairness, robustness, data poisoning, trustworthy machine learning, PAC learning

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

1. Nikola Konstantinov conducted the work on this paper while being at the Institute of Science and Technology Austria (ISTA). A preliminary version of the work contained in this article appeared in the AFCR@NeurIPS workshop (Konstantinov and Lampert, 2021).

©2022 Nikola Konstantinov and Christoph H. Lampert.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v23/21-1189.html.

respect to a certain protected attribute, such as gender or race (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., 2022). The underlying assumption is that by ensuring that the fairness property holds exactly or approximately based 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., 2022). 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 empirically that strong poisoning attacks can negatively impact the fairness of *specific learners* based on loss minimization (Solans et al., 2020; Chang et al., 2020; Mehrabi et al., 2021). 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 the aforementioned issues by studying the effect of arbitrary data corruptions on fair learning algorithms. 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* on the set of all finite hypothesis spaces, even in the limit of infinite training data. Furthermore, the irreducible excess gap in 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 also show that our hardness results are tight up to constant factors, in terms of the corruption ratio and the protected group frequencies, by proving matching upper bounds. To this end we study the performance of two natural types of learning algorithms under the malicious adversary model. We show that both algorithms achieve order-optimal performance in the infinite data regime, *thereby providing tight upper and lower bounds on the irreducible error of fairness-aware statistical learning under adversarial data corruption.*

We conclude with a discussion on the implications of our hardness results, emphasizing 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 learning analysis of fair 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. (2022) 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). On the methodological side, our upper bounds analysis employs a technique for proving concentration of estimates of conditional probabilities that has previously been used in the context of group fairness by Woodworth et al. (2017) and Agarwal et al. (2018). A number of hardness results for fair learning are also 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) and Mehrabi et al. (2021) 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.* On the defense side, Roh et al. (2020) design and empirically study an adversarial training approach for dealing with data corruption when training fair models. Their defense is shown to be effective against specific poisoning attacks that aim to reduce the model accuracy. In

contrast, for our upper bounds we are interested in learners that provably work against any poisoning attack, including those that can target the fairness properties of the model as well.

Among works focusing on weaker adversarial models, a particularly popular topic is the one of fair learning with noisy or adversarially perturbed protected attributes (Lamy et al., 2019; Awasthi et al., 2020; Wang et al., 2020; Celis et al., 2021a; Mehrotra and Celis, 2021; Celis et al., 2021b). Under the explicit assumption that the corruption does not affect 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. (2021) and Fogliato et al. (2020). The concurrent work of Jo et al. (2022) studies the hardness of fairness-aware learning with adversarial corruptions of both the labels and the protected attributes (but not the input variables), also allowing for the adversary to choose the points it can manipulate. However, they focus on studying adversarial strategies for enforcing a fixed target model, while we focus on understanding the statistical limits on the performance of the learner in terms of both fairness and accuracy.

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) and 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) and Li et al. (2021) who propose procedures for data cleaning/outlier detection, without a specific adversarial model, that in particular improve fairness performance.

Learning against an adversary Learning from corrupted training data is a field with long history, where both the theoretical and the 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., 2019b). 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) and 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

Throughout the paper 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, being part of a majority/minority group). 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 $h \coloneqq \mathcal{L}(S^c)$. The performance of a learner can be measured via the expected 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 this definition, an implicit assumption is that $Y = 1$ corresponds to a beneficial outcome (for example, an applicant receiving a job), so that this fairness notion only considers instances where the correct outcome should be advantageous.

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) and Menon and Williamson (2018a) for demographic parity

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

and its analog for equal opportunity

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

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 assume without loss of generality that $P_{10} \leq P_{11}$ (note that we do not make any assumption about P_0 and P_1 in this case).

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 Learning against an adversary

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 not being partial to a specific type of data corruption, this worst-case approach has the advantage of 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* $\mathcal{A} : \cup_{n\in\mathbb{N}}(\mathcal{X} \times A \times \mathcal{Y})^n \to \cup_{n\in\mathbb{N}}(\mathcal{X} \times A \times \mathcal{Y})^n$ 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 $S^p = \{(x_i^p, a_i^p, y_i^p)\}_{i=1}^n$ *of the same size*. 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) and Cesa-Bianchi et al. (1999) [2]. 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 $\mathfrak{M} \subseteq [n]$.

2. Strictly speaking, the nasty noise model of Bshouty et al. (2002); Diakonikolas et al. (2019a), in which the adversary can even choose the marked points, is even stronger than the adversary model we consider here. However, we opted for studying the malicious noise model, because this weaker adversary is already sufficient for showing strong impossibility results on learning. We refer to Section 6 for further discussion on the choice of adversarial model.

- 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 $i \notin \mathfrak{M}$. That is, the adversary can replace all marked data points in an arbitrary manner, with *no assumptions whatsoever* about the points (x_i^p, a_i^p, y_i^p) for $i \in \mathfrak{M}$.

- 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 $|\mathfrak{M}| \sim Bin(n, \alpha)$.

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, which is in lines with our worst-case approach. Note that this is in contrast to the learner $\mathcal{L}$ that can only access the data points in S^p. We refer to Section 3.4 for a more formal treatment.

3.3 Multi-objective learning

Our goal is to study the performance of the classifier $\mathcal{L}(S^p)$ *learned on the corrupted data*, both in terms of its expected loss $\mathcal{R}(\mathcal{L}(S^p), \mathbb{P})$ and its fairness deviation $\mathcal{D}(\mathcal{L}(S^p), \mathbb{P})$ *on the clean (test) distribution* $\mathbb{P}$. We will be interested in the probabilities of these quantities being large or small, under the randomness of the sampling of S^p - that is the randomness of the clean data, the marked points and the adversary.

Note that it is not a priori clear how to trade-off the two metrics and that this is likely to be application-dependent. Therefore it is also unclear how to evaluate the quality of a hypothesis. In our work we study two possible ways to do so.

Weighted objective One approach is to assume that a (application dependent) trade-off parameter $\lambda \geq 0$ is predetermined, so that the learner has to approximately minimize

$$L_\lambda(h) = \mathcal{R}(h) + \lambda \mathcal{D}(h). \tag{6}$$

The value of λ will likely be application-dependent and to be determined by the entity issuing the learner. There are various legal and ethical considerations that may apply when determining a desired accuracy-fairness trade-off (Barocas et al., 2019). Therefore, we leave λ as an arbitrary, but fixed parameter, similarly to Menon and Williamson (2018a), and we assume that λ is known by both the learner and the adversary.

For a given value of λ, the quality of the hypothesis $\mathcal{L}(h^S)$ can be directly measured via $L_\lambda(\mathcal{L}(h^S)) - \min_{h \in \mathcal{H}} L_\lambda(h)$. We will use L_λ^{par} and L_λ^{opp} to denote the weighted objectives with $\mathcal{D}^{par}$ and $\mathcal{D}^{opp}$ respectively.

Element-wise comparisons Alternatively, one may want to consider the two objectives independently. Given a classifier $h \in \mathcal{H}$, denote by $\mathfrak{V}(h) = (\mathcal{R}(h), \mathcal{D}(h))$ the vector consisting of the values of the two objectives. Note that $\mathfrak{V}$ does not, in general, induce a total order on $\mathcal{H}$. Instead we can only compare two classifiers $h_1, h_2 \in \mathcal{H}$ if, say, h_1 dominates h_2 in the sense that both $\mathcal{R}(h_1) \leq \mathcal{R}(h_2)$ and $\mathcal{D}(h_1) \leq \mathcal{D}(h_2)$. We denote this relation by

$\mathfrak{V}(h_1) \preceq \mathfrak{V}(h_2)$. As we will see, these component-wise comparisons are still useful for understanding the limits of learning against an adversary.

Since $\mathcal{R}(h)$ and $\mathcal{D}(h)$ are two independent objectives, there is no clear notion of an "optimal" classifier under the $\preceq$ relation in general. Therefore, when studying the pairwise objective $\mathfrak{V}(h)$, we will assume there exists a classifier that is optimal both in terms of fairness and accuracy. Then this classifier is optimal also under the $\preceq$ relation and hence the quality of any other hypotheses can be measured against it.

Specifically, for our analysis of the $\mathfrak{V}$ objective in Section 5.2, we assume that there exists a $h^* \in \mathcal{H}$, such that $h^* \in \operatorname{argmin}_{h\in\mathcal{H}} \mathcal{R}(h)$ and $h^* \in \operatorname{argmin}_{h\in\mathcal{H}} \mathcal{D}(h)$, so that $\mathfrak{V}(h^*) \preceq \mathfrak{V}(h)$ for all $h \in \mathcal{H}$. Then the quality of $\mathcal{L}(S^p)$ can be measured as the $\mathbb{R}^2$ vector

$$\mathbf{L}(\mathcal{L}(S^p)) = \mathfrak{V}(\mathcal{L}(S^p)) - \mathfrak{V}(h^*). \tag{7}$$

As with the weighted objective, we use $\mathbf{L}^{par}(\mathcal{L}(S^p))$ and $\mathbf{L}^{opp}(\mathcal{L}(S^p))$ to denote the loss vector when demographic parity and equal opportunity are used respectively.

One particular situation that we will study in which a component-wise optimal classifier h^* exists, is within the realizable PAC learning model with equal opportunity fairness. Indeed, whenever a classifier $h^* \in \mathcal{H}$ satisfies $\mathbb{P}(h^*(X) = Y) = 1$, we have that both $\mathcal{R}(h^*) = 0$ and $\mathcal{D}^{opp}(h^*) = 0$ and so $\mathbf{L}^{opp}(\mathcal{L}(S^p)) = \mathfrak{V}^{opp}(\mathcal{L}(S^p))$. More generally, the existence of h^* is plausible whenever the equal opportunity fairness notion is considered, since it is known that this fairness notion generally aligns well with accuracy (Hardt et al., 2016). We expect that our analysis of the $\mathfrak{V}$ objective can be extended to situations where h^* is only ϵ-approximately optimal in both objectives, which would cover more real-world situations, and we deem that an interesting direction for future work.

3.4 The limits of fairness-aware learning against an adversary

Lower and upper bounds analysis Over the next sections we will be showing lower and upper bounds on $L_\lambda(\mathcal{L}(S^p))$ and $\mathbf{L}(\mathcal{L}(S^p))$, that is, the risk and the fairness deviation measure achieved by the learner when trained on the corrupted data. *Our lower 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. *Our upper bounds* complement the hardness results by constructing learners that recover a classifier with guarantees on fairness and accuracy that match the lower bounds, for a wide range of learning problems and adversaries.

Crucial in these results is the ordering of the quantifiers. These matter not only for the comparison between the upper and the lower bounds, but also 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. At the same time, our upper bounds read as:

There exists a learner $\mathcal{L}$, such that for any distribution $\mathbb{P}$, any adversary $\mathcal{A}$ and any $\delta \in (0,1)$, with probability at least $1-\delta$ ...

Since the learner is fixed before the distribution and the adversary are, it has to work for any such pair.

We note that all probability statements 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. For a fixed clean data distribution $\mathbb{P}$ and a fixed adversary $\mathcal{A}$, we denote the distribution of S^p as $\mathbb{P}^{\mathcal{A}}$.

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 excess 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 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).

On the other hand, the learners we construct for the upper bounds are shown to work for *any hypothesis space* that is finite or of finite VC dimension, in all cases matching the lower bounds.

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 understanding the effect of these quantities on the hardness of the learning problem. Comparing the lower bounds, which hold regardless of the sample size n, to the upper bounds in the limit of $n \to \infty$ allows us to reason about the absolute limits of fairness-aware learning against a malicious adversary. Indeed, in this large data limit, we find that our upper and lower

bounds match in terms of their dependence on α and P_0 or P_{10} up to constant factors. We note that designing algorithms that achieve *sample-optimal* guarantees in our context is beyond the scope of this work. However, we will also be interested in the *statistical rates of convergence* of the studied learners to the irreducible gap certified by the lower bounds. We refer to Section 5.2 for a formal treatment.

4. Lower bounds

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 in this section can be found in Appendix A.**

4.1 Pareto lower bounds

We begin by presenting two hardness results that intuitively show that for some hypothesis spaces $\mathcal{H}$ 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 of this theorem (as well as of the other hardness results presented in this section) 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, so that different classifiers perform well on each, yet can be made indistinguishable after the modifications of the adversary. Then no fixed learner with access only to the corrupted data can be "correct" with high probability on both distributions and so any learner will incur an excessively high loss and exhibit excessively high unfairness on at least one of them, regardless of the amount of available data.

Here we provide a sketch proof of Theorem 1, to illustrate the type of construction used. A complete proof can be found in Appendix A.

Proof (Sketch) Let $\eta = \frac{\alpha}{1-\alpha}$, so that $\alpha = \frac{\eta}{1+\eta}$. We assume here that $\eta = \frac{\alpha}{1-\alpha} \leq 2P_0(1-P_0)$, with the other case following from a similar construction, but with an adversary that uses a smaller value of α (so that it leaves some of the data points at its disposal untouched).

Take four distinct points $\{x_1, x_2, x_3, x_4\} \in \mathcal{X}$. We consider two distributions $\mathbb{P}_0$ and $\mathbb{P}_1$, where each $\mathbb{P}_i$ is defined as

$$\mathbb{P}_i(x, a, y) = \begin{cases} 1 - P_0 - \eta/2 & \text{if } x = x_1, a = 1, y = 1 \\ P_0 - \eta/2 & \text{if } x = x_2, a = 0, y = 0 \\ \eta/2 & \text{if } x = x_3, a = i, y = \neg i \\ \eta/2 & \text{if } x = x_4, a = \neg i, y = i \\ 0 & \text{otherwise} \end{cases}$$

Note that these are valid distributions, since $\eta \leq 2P_0(1 - P_0) \leq 2P_0 \leq 2(1 - P_0)$ by assumption and also that $P_0 = \mathbb{P}_i(A = 0)$ for both $i \in \{0, 1\}$. Consider the hypothesis space $\mathcal{H} = \{h_0, h_1\}$, with

$$h_0(x_1) = 1 \quad h_0(x_2) = 0 \quad h_0(x_3) = 1 \quad h_0(x_4) = 0$$

and

$$h_1(x_1) = 1 \quad h_1(x_2) = 0 \quad h_1(x_3) = 0 \quad h_1(x_4) = 1.$$

The point of this construction is as follows: there are only two points, x_3 and x_4, where the two distributions differ. This is also where the classifiers differ and, in fact, each classifier h_i is better performing on the distribution $\mathbb{P}_i$, in both accuracy and fairness, than the other classifier.

Indeed, it is easy to verify that

$$L(h_{\neg i}, \mathbb{P}_i) - L(h_i, \mathbb{P}_i) = \eta, \quad \text{for both } i = 0, 1. \tag{8}$$

Moreover,

$$\mathcal{D}^{par}(h_{\neg i}, \mathbb{P}_i) - \mathcal{D}^{par}(h_i, \mathbb{P}_i) = \frac{\eta}{2P_0(1 - P_0)}, \quad \text{for both } i = 0, 1. \tag{9}$$

Now what the adversary does is to use all of the marked data to insert points with inputs x_3 and x_4, but with flipped labels and protected attributes. Then, since the points with inputs x_3 and x_4 in the original data are sufficiently rare, the adversary manages to hide which one of the two distributions was the original one.

Specifically, consider a (randomized) malicious adversary $\mathcal{A}_i$ of power α, that given a clean distribution $\mathbb{P}_i$, changes every marked point to $(x_3, \neg i, i)$ with probability 0.5 and to $(x_4, i, \neg i)$ otherwise. Under a distribution $\mathbb{P}_i$ and an adversary $\mathcal{A}_i$, the probability of seeing a point $(x_3, i, \neg i)$ is $\frac{\eta}{2}(1 - \alpha) = \frac{\eta}{2}\frac{1}{1+\eta} = \alpha/2$, which is equal to the probability of seeing a point $(x_3, \neg i, i)$. Therefore, denoting the probability distribution of the corrupted dataset, under a clean distribution $\mathbb{P}_i$ and an adversary $\mathcal{A}_i$, by $\mathbb{P}'_i$, one can verify that $\mathbb{P}'_0 = \mathbb{P}'_1$, so the two initial distributions $\mathbb{P}_0$ and $\mathbb{P}_1$ become indistinguishable under the adversarial manipulation.

The proof concludes by formalizing the observation that any fixed learner $\mathcal{L} : \cup_{n \in \mathbb{N}}(\mathcal{X} \times A \times \mathcal{Y})^n \to \{h_0, h_1\}$ will perform poorly on at least one of the distribution-adversary pairs $(\mathbb{P}_i, \mathcal{A}_i)$, since the resulting corrupted data distributions are the same, but the optimal classifiers differ. ■

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. This is because the adversary can force the learner to return a hypothesis that is *a constant away from optimality* is both objectives, with a *non-vanishing probability*[3]. To prove the theorem we explicitly construct a hypothesis space that is not learnable against the malicious adversary. As discussed in Section 3.4, 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[4]

$$\mathcal{R}(\mathcal{L}(S^p)) - \mathcal{R}(h^*) \geq \Omega\left(\alpha\right) \tag{10}$$

and

$$\mathcal{D}^{par}(\mathcal{L}(S^p)) - \mathcal{D}^{par}(h^*) \geq \Omega\left(\frac{\alpha}{P_0}\right). \tag{11}$$

The lower bound on the excess loss (10) 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 excessively 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.*

Equations (10) and (11) immediately imply the following lower bounds on L_λ and $\mathbf{L}^{par}$:

$$L_\lambda^{par}(L(S^p)) - \min_{h\in\mathcal{H}} \mathcal{L}_\lambda^{par}(h) \geq \Omega\left(\alpha + \lambda\frac{\alpha}{P_0}\right). \tag{12}$$

$$\mathbf{L}^{par}(\mathcal{L}(S^p)) \succeq \left(\Omega\left(\alpha\right), \Omega\left(\frac{\alpha}{P_0}\right)\right) \tag{13}$$

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, so that $\mathcal{D}^{par}(\mathcal{L}(S^p)) = 1$. 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*

3. The constant 0.5 for the probability of the adversary succeeding is perfectly sufficient for proving the impossibility of PAC learnability. A more refined analysis may yield an even larger constant, although we have not explored this further.
4. 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

$$L_\lambda^{opp}(\mathcal{L}(S^p)) - \min_{h\in\mathcal{H}} L_\lambda^{opp}(h) \geq \Omega\left(\alpha + \lambda\frac{\alpha}{P_{10}}\right) \tag{14}$$

$$\mathbf{L}^{opp}(\mathcal{L}(S^p)) \succeq \left(\Omega(\alpha), \Omega\left(\frac{\alpha}{P_{10}}\right)\right) \tag{15}$$

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 $P_{10} = \mathbb{P}(Y = 1, A = 0)$ 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 unnecessarily 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. In Theorem 4 an additional multiplicative factor of $1 - \frac{P_{10}}{P_{11}}$ appears - while we believe this to be an artifact of the proof technique and not inherent, we do not currently have a lower bound construction that circumvents this term. However, considering the asymptotic behavior where $\alpha \to 0, P_{10} \to 0$, but $P_{11} = \Theta(1)$, this additional term is negligible.

5. Upper bounds

We now prove that the (sample-size-independent) lower bounds from the previous section are tight up to constant factors, by providing matching upper bounds for the same problem. We do so by studying the performance of two natural types of fairness-aware learning algorithms under the malicious adversary model. We find that these algorithms achieve order-optimal performance in the large data regime.

Complete proofs of all results in this section can be found in Appendix B. A sketch of the proofs is also presented in Section 5.3.

5.1 Upper bounds on the λ-weighted objectives

The first type of algorithms we study simply minimize an empirical estimate of the λ-weighted objective L_λ. We show that with high probability such learners achieve an order-optimal deviation from $\min_{h\in\mathcal{H}} L_\lambda(h)$ in the large data regime, as long as $\mathcal{H}$ has a finite VC dimension.

Throughout this section we assume that $\lambda > 0$ is an arbitrary, but fixed parameter, chosen depending on domain-specific considerations (see also Section 3.3).

Bound for demographic parity Let $h \in \mathcal{H}$ be a fixed hypothesis. We consider the following natural estimate of $\mathcal{D}^{par}(h)$, as given in equation (4), based on the corrupted dataset $S^p = \{(x_i^p, a_i^p, y_i^p)\}_{i=1}^n$:

$$\widehat{\mathcal{D}}^{par}(h) = \left| \frac{\sum_{i=1}^n \mathbb{1}\{h(x_i^p) = 1, a_i^p = 0\}}{\sum_{i=1}^n \mathbb{1}\{a_i^p = 0\}} - \frac{\sum_{i=1}^n \mathbb{1}\{h(x_i^p) = 1, a_i^p = 1\}}{\sum_{i=1}^n \mathbb{1}\{a_i^p = 1\}} \right|, \tag{16}$$

with the convention that $\frac{0}{0} = 0$ for the purposes of this definition. We also denote the empirical risk of h on S^p by $\widehat{R}^p(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{h(x_i^p) \neq y_i^p\}$.

Suppose that the learner $\mathcal{L}_\lambda^{par} : \cup_{n=1}^{\infty}(\mathcal{X} \times A \times \mathcal{Y})^n \to \mathcal{H}$ is such that

$$\mathcal{L}_\lambda^{par}(S^p) \in \operatorname*{argmin}_{h\in\mathcal{H}}(\widehat{R}^p(h) + \lambda\widehat{\mathcal{D}}^{par}(h)), \quad \text{for all } S^p.$$

That is, $\mathcal{L}_\lambda^{par}$ always returns a minimizer of the λ-weighted empirical objective. Then the following result holds.

Theorem 5 *Let $\mathcal{H}$ be any hypothesis space with $d = VC(\mathcal{H}) < \infty$. Let $\mathbb{P} \in \mathcal{P}(\mathcal{X} \times A \times \mathcal{Y})$ be a fixed distribution and let $\mathcal{A}$ be any malicious adversary of power $\alpha < 0.5$. Denote by $\mathbb{P}^{\mathcal{A}}$ the probability distribution of the corrupted data S^p, under the random sampling of the clean data, the marked points and the randomness of the adversary. Then for any $\delta \in (0,1)$ and $n \geq \max\left\{\frac{8\log(16/\delta)}{(1-\alpha)P_0}, \frac{12\log(12/\delta)}{\alpha}, \frac{d}{2}\right\}$, we have:*

$$\mathbb{P}^{\mathcal{A}}\left(L_\lambda^{par}\left(\widehat{h}\right) \leq \min_{h\in\mathcal{H}} L_\lambda^{par}(h) + \Delta_\lambda^{par}\right) > 1 - \delta,$$

where $\widehat{h} \coloneqq \mathcal{L}_\lambda^{par}(S^p)$ is the hypothesis returned by the learner, $L_\lambda^{par}(h) = \mathcal{R}(h) + \lambda\mathcal{D}^{par}(h)$ is the λ-weighted objective and [5]

$$\Delta_\lambda^{par} = 3\alpha + \lambda(2\Delta^{par}) + \widetilde{\mathcal{O}}\left(\sqrt{\frac{d}{n}} + \lambda\sqrt{\frac{d}{P_0 n}}\right)$$

and

$$\Delta^{par} = \frac{2\alpha}{\frac{P_0}{3} + \alpha} = \mathcal{O}\left(\frac{\alpha}{P_0}\right).$$

This result shows that for any $\mathcal{H}$ of finite VC dimension, any distribution $\mathbb{P}$ and against any malicious adversary $\mathcal{A}$ of power α, the learner $\mathcal{L}_\lambda^{par}$ is able, for sufficiently large values of the sample size $n \geq \Omega((P_0/\alpha)^2)$, to return with high probability a hypothesis $\widehat{h}$ such that

$$L_\lambda^{par}\left(\widehat{h}\right) - \min_{h\in\mathcal{H}} L_\lambda^{par}(h) \leq \mathcal{O}\left(\alpha + \lambda\frac{\alpha}{P_0}\right). \tag{17}$$

Note that these rates on the irreducible error term match our lower bound from Theorem 1 and Inequality (12). Indeed, the hardness result shows that no algorithm can guarantee better error rates than those in (17) on the family of finite hypothesis sets and hence also on the hypothesis sets with finite VC dimension.

5. The $\widetilde{\mathcal{O}}$-notation hides constant and logarithmic factors.

Bound for equal opportunity Similarly, we consider the following estimate for the equal opportunity deviation measure:

$$\widehat{\mathcal{D}}^{opp}(h) = \left| \frac{\sum_{i=1}^{n} \mathbb{1}\{h(x_i^p) = 1, a_i^p = 0, y_i^p = 1\}}{\sum_{i=1}^{n} \mathbb{1}\{a_i^p = 0, y_i^p = 1\}} - \frac{\sum_{i=1}^{n} \mathbb{1}\{h(x_i^p) = 1, a_i^p = 1, y_i^p = 1\}}{\sum_{i=1}^{n} \mathbb{1}\{a_i^p = 1, y_i^p = 1\}} \right|, \tag{18}$$

with the convention that $\frac{0}{0} = 0$ for the purposes of this definition. Suppose that a learner $\mathcal{L}_\lambda^{opp} : \cup_{n=1}^{\infty} (\mathcal{X} \times A \times \mathcal{Y})^n \to \mathcal{H}$ is such that

$$\mathcal{L}_\lambda^{opp}(S^p) \in \operatorname*{argmin}_{h \in \mathcal{H}} (\widehat{R}^p(h) + \lambda \widehat{\mathcal{D}}^{opp}(h)), \quad \text{for all } S^p,$$

that is, always returns a minimizer of the λ-weighted empirical objective. Then:

Theorem 6 *Let $\mathcal{H}$ be any hypothesis space with $d = VC(\mathcal{H}) < \infty$. Let $\mathbb{P} \in \mathcal{P}(\mathcal{X} \times A \times \mathcal{Y})$ be a fixed distribution and let $\mathcal{A}$ be any malicious adversary of power $\alpha < 0.5$. Then for any $\delta \in (0, 1)$ and $n \geq \max\left\{ \frac{8 \log(16/\delta)}{(1-\alpha) P_{10}}, \frac{12 \log(12/\delta)}{\alpha}, \frac{d}{2} \right\}$*

$$\mathbb{P}^{\mathcal{A}} \left(L_\lambda^{opp}\big(\widehat{h}\big) \leq \min_{h \in \mathcal{H}} L_\lambda^{opp}(h) + \Delta_\lambda^{opp} \right) > 1 - \delta,$$

where $\widehat{h} \coloneqq \mathcal{L}_\lambda^{opp}(S^p)$ is the hypothesis returned by the learner, $L_\lambda^{opp}(h) = \mathcal{R}(h) + \lambda \mathcal{D}^{opp}(h)$ and

$$\Delta_\lambda^{opp} = 3\alpha + \lambda(2\Delta^{opp}) + \widetilde{\mathcal{O}} \left(\sqrt{\frac{d}{n}} + \lambda \sqrt{\frac{d}{P_{10} n}} \right)$$

and

$$\Delta^{opp} = \frac{2\alpha}{\frac{P_{10}}{3} + \alpha} = \mathcal{O}\left(\frac{\alpha}{P_{10}} \right).$$

Again, for a sufficiently large sample size, this result implies an upper bound on the excess loss of the hypothesis $\widehat{h} \coloneqq \mathcal{L}_\lambda^{opp}(S^p)$ returned by the learner in terms of the weighted objective

$$L_\lambda^{opp}\big(\widehat{h}\big) - \min_{h \in \mathcal{H}} L_\lambda^{opp}(h) \leq \mathcal{O}\left(\alpha + \lambda \frac{\alpha}{P_{10}} \right), \tag{19}$$

which is again order optimal, according to Theorem 2 and Inequality (14).

5.2 Component-wise upper bounds

We now introduce a second type of algorithms, which return a hypothesis that achieves both a small loss and a small fairness deviation measure on the training data, or, if no such hypothesis exists, a random hypothesis. We show that, in the case when there exists a classifier that is optimal in both accuracy and fairness, with high probability such learners return a hypothesis $h \in \mathcal{H}$ that is order-optimal in both elements of the objective vector $\mathbf{L}(h)$, as long as $\mathcal{H}$ is of finite VC dimension and n is sufficiently large. Finally, in the

case of realizable PAC learning with equal opportunity fairness, we are able to provide an algorithm that achieves such order-optimal guarantees with *fast statistical rates*, for any finite hypothesis space.

Throughout the section only, we assume that there exists a classifier $h^* \in \mathcal{H}$, such that $\mathfrak{V}(h^*) \preceq \mathfrak{V}(h)$ for all $h \in \mathcal{H}$. That is, $\mathcal{R}(h^*) \leq \mathcal{R}(h)$ and $\mathcal{D}(h^*) \leq \mathcal{D}(h)$ for all $h \in \mathcal{H}$. We also assume that $d = VC(\mathcal{H}) < \infty$.

We note that the algorithms studied in this section require the knowledge of α and of P_0 and P_{10} for demographic parity and equal opportunity respectively, since they explicitly use these quantities when selecting a hypothesis. Even if these quantities are unknown in advance, estimates can often be obtained in practice, for example by having the quality of a small random subset of the data S^p verified by a trusted authority, or via conducting an additional survey/crowdsourcing experiment.

Bound for demographic parity Given a corrupted dataset $S^p = \{(x_i^p, a_i^p, y_i^p)\}$, let $\widehat{h}^r \in \operatorname{argmin}_{h\in\mathcal{H}} \widehat{\mathcal{R}}^p(h)$ and $\widehat{h}^{par} \in \operatorname{argmin}_{h\in\mathcal{H}} \widehat{\mathcal{D}}^{par}(h)$. Further, we define the sets

$$\mathcal{H}_1 = \left\{ h \in \mathcal{H} : \widehat{\mathcal{R}}^p(h) - \widehat{\mathcal{R}}^p(\widehat{h}^r) \leq 3\alpha + 4\sqrt{\frac{8d\log(\frac{en}{d}) + 2\log(16/\delta)}{n}} \right\}$$

$$\mathcal{H}_2 = \left\{ h \in \mathcal{H} : \widehat{\mathcal{D}}^{par}(h) - \widehat{\mathcal{D}}^{par}(\widehat{h}^{par}) \leq 2\Delta^{par} + 32\sqrt{\frac{2d\log(\frac{2en}{d}) + 2\log(96/\delta)}{(1-\alpha)P_0 n}} \right\}.$$

That is, $\mathcal{H}_1$ and $\mathcal{H}_2$ are the sets of classifiers that are not far from optimal on the train data, in terms of their risk and their fairness respectively. The upper bound terms are selected according to the concentration properties of the two measures and describe the amount of expected variability of those, due to the randomness of the training data. Now define the *component-wise learner*:

$$\mathcal{L}_{cw}^{par}(S^p) = \begin{cases} \text{any } h \in \mathcal{H}_1 \cap \mathcal{H}_2, & \text{if } \mathcal{H}_1 \cap \mathcal{H}_2 \neq \emptyset \\ \text{any } h \in \mathcal{H}, & \text{otherwise,} \end{cases}$$

that returns a classifier that is good in both metrics, if such exists, or an arbitrary classifier otherwise.

Intuitively, whenever a classifier is an element of $\mathcal{H}_1 \cap \mathcal{H}_2$, it performs relatively well on the data in terms of both accuracy and fairness, thereby being a good candidate for learning. On the other hand, situations where no such classifier exists are expected to be rare. This is because h^* is optimal in both metrics on the true data distribution and so it is also likely to perform close to optimal on the corrupted data, allowing for variations due to finite sample effects and data corruption. Therefore, with high probability $h^* \in \mathcal{H}_1 \cap \mathcal{H}_2$, so that the intersection is non-empty.

Formally, the following result holds.

Theorem 7 *Let $\mathcal{H}$ be any hypothesis space with $d = VC(\mathcal{H}) < \infty$. Let $\mathbb{P} \in \mathcal{P}(\mathcal{X} \times A \times \mathcal{Y})$ be a fixed distribution and let $\mathcal{A}$ be any malicious adversary of power $\alpha < 0.5$. Suppose that there exists a hypothesis $h^* \in \mathcal{H}$, such that $\mathfrak{V}(h^*) \preceq \mathfrak{V}(h)$ for all $h \in \mathcal{H}$. Then for any*

$\delta \in (0,1)$ and $n \geq \max\left\{\frac{8\log(16/\delta)}{(1-\alpha)P_0}, \frac{12\log(12/\delta)}{\alpha}, \frac{d}{2}\right\}$, with probability at least $1-\delta$:

$$\mathbf{L}^{par}\big(\widehat{h}\big) \preceq \left(6\alpha + \widetilde{\mathcal{O}}\left(\sqrt{\frac{d}{n}}\right), 4\Delta^{par} + \widetilde{\mathcal{O}}\left(\sqrt{\frac{d}{P_0 n}}\right)\right),$$

where $\widehat{h} \coloneqq \mathcal{L}_{cw}^{par}(S^p)$ is the hypothesis returned by the learner and

$$\mathbf{L}^{par}(\widehat{h}) = \Big(\mathcal{R}(\widehat{h}) - \mathcal{R}(h^*), \mathcal{D}^{par}(\widehat{h}) - \mathcal{D}^{par}(h^*)\Big).$$

Since $\Delta^{par} = \mathcal{O}\left(\frac{\alpha}{P_0}\right)$, in the large data limit we obtain that

$$\mathbf{L}^{par}(\widehat{h}) \preceq \left(\mathcal{O}(\alpha), \mathcal{O}\left(\frac{\alpha}{P_0}\right)\right). \tag{20}$$

Note that this bound is order-optimal for the class of finite hypothesis spaces, and hence also for the class of hypothesis spaces with finite VC dimension, according to Theorem 1 and Inequality (13).

Bound for equal opportunity Similarly, let $\widehat{h}^{opp} \in \operatorname{argmin}_{h\in\mathcal{H}} \widehat{\mathcal{D}}^{opp}(h)$. Further, we define the set

$$\mathcal{H}_3 = \left\{h \in \mathcal{H} : \widehat{\mathcal{D}}^{opp}(h) - \widehat{\mathcal{D}}^{opp}(\widehat{h}^{opp}) \leq 2\Delta^{opp} + 32\sqrt{\frac{2d\log(\frac{2en}{d}) + 2\log(96/\delta)}{(1-\alpha)P_{10}n}}\right\}.$$

That is, $\mathcal{H}_3$ is the set of classifiers that are not far from optimal on the train data, in terms of equal opportunity fairness. Now define the *component-wise learner* for equal opportunity:

$$\mathcal{L}_{cw}^{opp}(S^p) = \begin{cases} \text{any } h \in \mathcal{H}_1 \cap \mathcal{H}_3, & \text{if } \mathcal{H}_1 \cap \mathcal{H}_3 \neq \emptyset \\ \text{any } h \in \mathcal{H}, & \text{otherwise,} \end{cases}$$

that returns a classifier that is good in both metrics, if such exists, or an arbitrary classifier otherwise. Then the following result holds.

Theorem 8 *Let $\mathcal{H}$ be any hypothesis space with $d = VC(\mathcal{H}) < \infty$. Let $\mathbb{P} \in \mathcal{P}(\mathcal{X} \times A \times \mathcal{Y})$ be a fixed distribution and let $\mathcal{A}$ be any malicious adversary of power $\alpha < 0.5$. Suppose that there exists a hypothesis $h^* \in \mathcal{H}$, such that $\mathfrak{V}(h^*) \preceq \mathfrak{V}(h)$ for all $h \in \mathcal{H}$. Then for any $\delta \in (0,1)$ and $n \geq \max\left\{\frac{8\log(16/\delta)}{(1-\alpha)P_{10}}, \frac{12\log(12/\delta)}{\alpha}, \frac{d}{2}\right\}$, with probability at least $1-\delta$*

$$\mathbf{L}^{opp}(\widehat{h}) \preceq \left(6\alpha + \widetilde{\mathcal{O}}\left(\sqrt{\frac{d}{n}}\right), 4\Delta^{opp} + \widetilde{\mathcal{O}}\left(\sqrt{\frac{d}{P_{10} n}}\right)\right).$$

where $\widehat{h} \coloneqq \mathcal{L}_{cw}^{opp}(S^p)$ is the hypothesis returned by the learner and

$$\mathbf{L}^{opp}(\widehat{h}) = \Big(\mathcal{R}(\widehat{h}) - \mathcal{R}(h^*), \mathcal{D}^{opp}(\widehat{h}) - \mathcal{D}^{opp}(h^*)\Big).$$

Since $\Delta^{opp} = \mathcal{O}\left(\frac{\alpha}{P_{10}}\right)$, in the large data limit we obtain that

$$\mathbf{L}^{opp}(\widehat{h}) \preceq \left(\mathcal{O}(\alpha), \mathcal{O}\left(\frac{\alpha}{P_{10}}\right)\right). \tag{21}$$

Note that this bound is order-optimal for the class of finite hypothesis spaces, and hence also for the class of hypothesis spaces with finite VC dimension, according to Theorem 2 and Inequality (15).

Upper bound with fast rates Finally, we study learning with the equal opportunity fairness notion, in the realizable PAC learning framework, where a perfectly accurate classifier exists. Given this additional assumption, we are able to certify *convergence to an order-optimal error in both fairness and accuracy at fast statistical rates*. For simplicity we assume that $\mathcal{H}$ is finite here.

Specifically, note that while the results presented already achieve order-optimal guarantees in the limit as $n \to \infty$, for a finite amount of samples they incur an additional loss of $\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}}\right)$. Regarding P_0 (for demographic parity) or P_{10} (for equal opportunity) as fixed, all previous algorithms need $\widetilde{\mathcal{O}}\left(\frac{1}{\alpha^2}\right)$ samples to achieve an excess risk and fairness deviation measure of $\widetilde{\mathcal{O}}\left(\alpha\right)$. In contrast, the algorithm we present now only requires $\mathcal{O}\left(\frac{1}{\alpha}\right)$ samples.

Formally, assume that the underlying clean distribution $\mathbb{P}$ is such that there exists a $h^* \in \mathcal{H}$, for which $\mathbb{P}(h^*(X) = Y) = 1$. This implies that $L(h^*) = 0$ and $\mathcal{D}^{opp}(h^*) = 0$.

Key to the design of an algorithm that achieves fast statistical rates for the objective $\mathbf{L}$ are the following empirical estimates:

$$\bar{\gamma}^p_{1a}(h) = \frac{\sum_{i=1}^n \mathbb{1}\{h(x_i^p) = 0, a_i^p = a, y_1^p = 1\}}{\sum_{i=1}^n \mathbb{1}\{a_i^p = a, y_i^p = 1\}} \tag{22}$$

of $\bar{\gamma}_{1a}(h) \coloneqq \mathbb{P}(h(X) = 0 | A = a, Y = 1) = 0$ for $a \in \{0, 1\}$. The fact that $\bar{\gamma}_{1a}(h) = 0$ (as opposed to $\gamma_{1a}(h) = 1$) is crucial for obtaining the fast rates, since it allows for a concentration analysis based on the multiplicative Chernoff bounds only, rather than the additive ones and/or Hoeffding's inequality (Boucheron et al., 2013), which would lead to rates of $\mathcal{O}(\frac{1}{\alpha^2})$ again.

Given a (corrupted) training set S^p, denote by

$$\mathcal{H}^*(S^p) := \left\{ h \in \mathcal{H} \middle| \max_a \bar{\gamma}^p_{1a}(h) \leq \Delta^{opp} \wedge \widehat{\mathcal{R}}^p(h) \leq \frac{3\alpha}{2} \right\} \tag{23}$$

the set of all classifiers that have a small loss and small values of $\bar{\gamma}^p_{1a}$ for both $a \in \{0, 1\}$ on S^p. Consider the learner $\mathcal{L}^{fast}$ defined by

$$\mathcal{L}^{fast}(S^p) = \begin{cases} \text{any } h \in \mathcal{H}^*, & \text{if } \mathcal{H}^* \neq \emptyset \\ \text{any } h \in \mathcal{H}, & \text{otherwise.} \end{cases} \tag{24}$$

The intuition behind the construction is similar to before: hypotheses in $\mathcal{H}^*$ perform well on the training data and hence are good candidates. At the same time, we expect that $h^* \in \mathcal{H}^*$, so that $\mathcal{H}^*$ is non-empty.

Then the following result holds.

Theorem 9 *Let $\mathcal{H}$ be finite and $\mathbb{P} \in \mathcal{P}(\mathcal{X} \times A \times \mathcal{Y})$ be such that for some $h^* \in \mathcal{H}$, $\mathbb{P}(h^*(X) = Y) = 1$. Let $\mathcal{A}$ be any malicious adversary of power $\alpha < 0.5$. Then for any $\delta, \eta \in (0,1)$ and any*

$$n \geq \max\left\{\frac{8\log(16|\mathcal{H}|/\delta)}{(1-\alpha)P_{10}}, \frac{12\log(12/\delta)}{\alpha}, \frac{2\log(8|\mathcal{H}|/\delta)}{3\eta^2\alpha}, \frac{2\log(\frac{16|\mathcal{H}|}{\delta})}{3\eta^2(1-\alpha)P_{10}\alpha}\right\}$$
$$= \Omega\left(\frac{\log(|\mathcal{H}|/\delta)}{\eta^2 P_{10}\alpha}\right)$$

with probability at least $1-\delta$

$$\mathbf{L}^{opp}(\widehat{h}) \preceq \left(\frac{3\alpha}{1-\eta}, \frac{2\Delta^{opp}}{1-\eta}\right),$$

where $\widehat{h} \coloneqq \mathcal{L}^{fast}(S^p)$ is the hypothesis returned by the learner and

$$\mathbf{L}^{opp}(\widehat{h}) = \left(\mathcal{R}(\widehat{h}) - \mathcal{R}(h^*), \mathcal{D}^{opp}(\widehat{h}) - \mathcal{D}^{opp}(h^*)\right).$$

As an immediate consequence of Theorem 9, setting $\eta = \frac{1}{2}$, say, yields that for large n, with high probability

$$\mathbf{L}^{opp}(\widehat{h}) \preceq \left(\mathcal{O}(\alpha), \mathcal{O}\left(\frac{\alpha}{P_{10}}\right)\right). \tag{25}$$

Again, this bound is order-optimal for finite hypothesis sets, according to Theorem 2 and Inequality (15). In addition, regarding P_{10} as a constant, the number of samples needed for achieving this order-optimal element-wise error is indeed $\mathcal{O}(\frac{1}{\alpha})$, according to Theorem 9, which is faster than the $\widetilde{\mathcal{O}}(\frac{1}{\alpha^2})$ we obtained with the previous results.

5.3 Sketch of the upper bounds proofs

Here we present a sketch of the proofs of the upper bounds. The complete proofs can be found in Appendix B.

The proofs of Theorems 5, 6, 7, 8 rely on a series of results that describe the deviations of the corrupted fairness estimates $\widehat{\mathcal{D}}(h)$ from the true underlying population values $\mathcal{D}(h)$, uniformly over the hypothesis space $\mathcal{H}$. Key to this is bounding the effect of the data corruption, as expressed by the maximum achievable gap between the corrupted fairness estimates and the corresponding estimates based on the clean (but unknown) subset of the data. Then the large deviation properties of these clean data estimates are studied instead.

Here we make this specific for the case of demographic parity, with the analysis for equal opportunity being similar. We denote

$$\gamma_a^p(h) = \frac{\sum_{i=1}^n \mathbb{1}\{h(x_i^p) = 1, a_i^p = a\}}{\sum_{i=1}^n \mathbb{1}\{a_i^p = a\}}$$

and

$$\gamma_a(h) = \mathbb{P}(h(X) = 1 | A = a),$$

so that $\widehat{\mathcal{D}}^{par}(h) = |\gamma_0^p(h) - \gamma_1^p(h)|$ and $\mathcal{D}^{par}(h) = |\gamma_0(h) - \gamma_1(h)|$. Note that $\gamma_a^p(h)$ is an estimate of a conditional probability *based on the corrupted data*. We now introduce the corresponding estimate that only uses the *unknown clean subset* of the training set S^p

$$\gamma_a^c(h) = \frac{\sum_{i=1}^n \mathbb{1}\{h(x_i^p) = 1, a_i^p = a, i \notin \mathfrak{M}\}}{\sum_{i=1}^n \mathbb{1}\{a_i^p = a, i \notin \mathfrak{M}\}}.$$

Bounding the effect of the adversary First, we bound how far the corrupted estimates $\gamma_a^p(h)$ of $\gamma_a(h)$ are from the clean estimates $\gamma_a^c(h)$, uniformly over the hypothesis space $\mathcal{H}$:

Lemma 1 *If* $n \geq \max\left\{\frac{8\log(4/\delta)}{(1-\alpha)P_0}, \frac{12\log(3/\delta)}{\alpha}\right\}$, *we have*

$$\mathbb{P}^{\mathcal{A}}\left(\sup_{h\in\mathcal{H}}\left(|\gamma_0^p(h) - \gamma_0^c(h)| + |\gamma_1^p(h) - \gamma_1^c(h)|\right) \geq \frac{2\alpha}{\frac{P_0}{3} + \alpha}\right) < \delta.$$

Informally, this lemma allows us to connect the corrupted estimate $\widehat{\mathcal{D}}^{par}(h)$ with the corresponding ideal clean estimate $\widehat{\mathcal{D}}^{c}(h) = |\gamma_0^c(h) - \gamma_1^c(h)|$.

Bounding the deviation of the clean data estimate Secondly, a technique used by Woodworth et al. (2017) and Agarwal et al. (2018) for proving concentration of fairness measures is used to derive a concentration result for the clean estimates $\gamma_a^c(h)$, around the true population values $\gamma_a(h)$. This, together with Lemma 1, allows us to bound the gap between the corrupted estimate $\widehat{\mathcal{D}}^{par}(h)$ and the true population value $\mathcal{D}^{par}(h)$, for a single hypothesis.

Making the bound uniform over $\mathcal{H}$ Finally, the bound obtained is made uniform over $\mathcal{H}$. For this, we use the classic symmetrization technique (Vapnik, 2013) for proving bounds uniformly over hypothesis spaces of finite VC dimension. However, since the objective is different from the 0-1 loss, care is needed to ensure that the argument goes through, so the proof is given in full detail in the supplementary material.

Once a uniform bound on the deviations of the corrupted fairness estimates from the true underlying population values is obtained, the results of Theorems 5, 6, 7, 8 follow similarly to most classic ERM results.

Proof of Theorem 9 Similarly to the other results, the proof of Theorem 9 first links the corrupted estimates $\bar{\gamma}_{1a}^p$ to their clean counterparts and then uses the clean data concentration to study the behavior of the corrupted estimates. However, an important tool that allows us to obtain the fast statistical rates, is a set of *multiplicative concentration bounds* on the $\bar{\gamma}_{1a}^p$ estimates. It is for this reason that $\mathcal{L}^{fast}$ learner uses the $\bar{\gamma}_{1a}^p$ estimates, instead of the γ_{1a}^p, see also the discussion after equation (51). Full details and a complete proof can be found in the supplementary material.

6. 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. We also provided upper bounds that match our hardness results up to constant factors, in the large data regime.

Below we outline several implications of our work and discuss some specific extensions that constitute interesting directions for future research.

Implications of our results 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 section, previous research has shown that it can be possible to recover a fair model under corruptions of the labels or the 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 an 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.

Extensions to other fairness notions We expect that our analysis can be extended to other group fairness measures. In particular, the work of Agarwal et al. (2018) has shown that a broad range of fairness notions based on conditional independence constraints are amendable to concentration of measure analysis, via an application of the proof technique of Woodworth et al. (2017). Since this technique is also at the core of the concentration arguments used in the proofs of our upper bounds, we expect that a similar analysis can be conducted for the broader class of fairness measures considered by Agarwal et al. (2018).

The lower bounds, however, require explicit constructions of hard learning problems to be designed and these constructions are necessarily tailored to the specific fairness notions being considered. Therefore, while the proof technique, namely the method of induced distributions (Kearns and Li, 1993), may be a useful tool for showing hardness results about other fairness measures, the key challenge of designing a hard learning problem instance for each measure remains open.

Extensions to other adversarial models In this work we have studied learning and fairness under the malicious adversary model (Kearns and Li, 1993). It will be interesting to analyze the limits of fairness-aware learning for other adversarial models as well. As mentioned above, studying weaker, application-specific adversaries may allow for PAC learnability.

On the other hand, fairness can be studied under the even stronger nasty noise model of Bshouty et al. (2002), which has also recently been analyzed in the context of robust mean estimation (Diakonikolas et al., 2019a). In this model the adversary does not get to manipulate a random subset of the data, but can instead choose the points that it alters. To our awareness, the only work that considers this adversarial model in the context of fairness is that of Celis et al. (2021b), who, however, only study manipulations of the protected attribute and not of the labels and features.

Since the nasty noise adversary is strictly stronger than the malicious one, our lower bounds hold for the nasty noise model as well. In particular, achieving optimal fairness remains impossible in this setup. Whether our lower bounds are order-optimal within the nasty adversary model as well, or stronger hardness results can be shown, is an interesting direction for future work.

Acknowledgments

The authors thank Eugenia Iofinova and Bernd Prach for providing feedback on early versions of this paper. This publication was made possible by an ETH AI Center postdoctoral fellowship to Nikola Konstantinov.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learing (ICML)*, 2018.

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.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Summer School on Machine Learning*, 2003.

Nader H Bshouty, Nadav Eiron, and Eyal Kushilevitz. Pac learning with nasty noise. *Theoretical Computer Science (TCC)*, 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. *Conference on Neural Information Processing Systems (NeurIPS)*, 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, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 2019a.

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)*, 2019b.

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.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58, 1963.

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.

Changhun Jo, Jy-yong Sohn, and Kangwook Lee. Breaking fair binary classification with optimal flipping attacks. *arXiv preprint arXiv:2204.05472*, 2022.

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. On the impossibility of fairness-aware learning from corrupted data. In *Algorithmic Fairness through the Lens of Causality and Robustness workshop at NeurIPS*, 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.

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

Ninareh Mehrabi, Muhammad Naveed, Fred Morstatter, and Aram Galstyan. Exacerbating algorithmic bias through fairness attacks. *Conference on Artificial Intelligence (AAAI)*, 2021.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 2022.

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.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer, 2013.

Jialu Wang, Yang Liu, and Caleb Levy. Fair classification with group-dependent label noise. *Conference on Fairness, Accountability and Transparency (FAccT)*, 2021.

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)*, 2020.

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.

Supplementary Material

The supplementary material is structured as follows.

- **Appendix A** contains the proofs of all lower bounds results. Section A.1 focuses on the Pareto lower bounds. Section A.2 contains the proofs for the lower bounds on fairness, given that accuracy is kept optimal.

- **Appendix B** contains the complete proofs of our upper bound results. In particular, Section B.1 explains the notation and introduces the classic concentration tools that we will use. In Section B.2 a number of concentration results under corrupted data for the demographic parity and equal opportunity fairness notions are shown. Finally, Section B.3 gives the formal proofs of all upper bound results, building on the concentration inequalities from the previous section.

Appendix A. Lower bounds proofs

In the proofs of our hardness results we use a technique from (Kearns and Li, 1993) called the *method of induced distributions*. The idea is to construct two distributions that are sufficiently different, so that different classifiers perform well on each, yet can be made indistinguishable after the modifications of the adversary. Then no fixed learner with access only to the corrupted data can be "correct" with high probability on both distributions and so any learner will incur excessively high loss and/or exhibit excessively high unfairness on at least one of the two distributions, regardless of the amount of available data.

The proofs of the four results are structured in a similar way, but use different constructions of the underlying learning problem, tailored to the fairness measure and the type of bound we want to show.

A.1 Pareto lower bounds proofs

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

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

and

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

Proof Let $\eta = \frac{\alpha}{1-\alpha}$, so that $\alpha = \frac{\eta}{1+\eta}$.

Case 1 Assume that $\eta = \frac{\alpha}{1-\alpha} \leq 2P_0(1-P_0)$. Take four distinct points $\{x_1, x_2, x_3, x_4\} \in \mathcal{X}$. We consider two distributions $\mathbb{P}_0$ and $\mathbb{P}_1$, where each $\mathbb{P}_i$ is defined as

$$\mathbb{P}_i(x,a,y) = \begin{cases} 1 - P_0 - \eta/2 & \text{if } x = x_1, a = 1, y = 1 \\ P_0 - \eta/2 & \text{if } x = x_2, a = 0, y = 0 \\ \eta/2 & \text{if } x = x_3, a = i, y = \neg i \\ \eta/2 & \text{if } x = x_4, a = \neg i, y = i \\ 0 & \text{otherwise} \end{cases}$$

Note that these are valid distributions, since $\eta \leq 2P_0(1-P_0) \leq 2P_0 \leq 2(1-P_0)$ by assumption and also that $P_0 = \mathbb{P}_i(A=0)$ for both $i \in \{0,1\}$. Consider the hypothesis space $\mathcal{H} = \{h_0, h_1\}$, with

$$h_0(x_1) = 1 \quad h_0(x_2) = 0 \quad h_0(x_3) = 1 \quad h_0(x_4) = 0$$

and

$$h_1(x_1) = 1 \quad h_1(x_2) = 0 \quad h_1(x_3) = 0 \quad h_1(x_4) = 1.$$

Note that $L(h_i, \mathbb{P}_i) = 0$ for both $i = 0, 1$. Moreover,

$$\begin{aligned} \mathcal{D}^{par}(h_0, \mathbb{P}_0) &= \left|\mathbb{P}_{(X,A,Y)\sim\mathbb{P}_0}(h_0(X) = 1|A=0) - \mathbb{P}_{(X,A,Y)\sim\mathbb{P}_0}(h_0(X) = 1|A=1)\right| \\ &= \left|\frac{\eta/2}{P_0 - \eta/2 + \eta/2} - \frac{1 - P_0 - \eta/2}{1 - P_0 - \eta/2 + \eta/2}\right| \\ &= \left|\frac{\eta}{2P_0} - \frac{2 - 2P_0 - \eta}{2(1-P_0)}\right| \\ &= \left|\frac{\eta}{2P_0(1-P_0)} - 1\right| \\ &= 1 - \frac{\eta}{2P_0(1-P_0)}, \end{aligned}$$

since $\eta \leq 2P_0(1-P_0)$ by assumption. Furthermore,

$$\begin{aligned} \mathcal{D}^{par}(h_1, \mathbb{P}_0) &= \left|\mathbb{P}_{(X,A,Y)\sim\mathbb{P}_0}(h_1(X) = 1|A=0) - \mathbb{P}_{(X,A,Y)\sim\mathbb{P}_0}(h_1(X) = 1|A=1)\right| \\ &= |0 - 1| \\ &= 1 \end{aligned}$$

Therefore, $\mathcal{D}^{par}(h_1, \mathbb{P}_0) - \mathcal{D}^{par}(h_0, \mathbb{P}_0) = \frac{\eta}{2P_0(1-P_0)}$. Similarly,

$$\begin{aligned} \mathcal{D}^{par}(h_1, \mathbb{P}_1) &= \left|\mathbb{P}_{(X,A,Y)\sim\mathbb{P}_1}(h_1(X) = 1|A=0) - \mathbb{P}_{(X,A,Y)\sim\mathbb{P}_1}(h_1(X) = 1|A=1)\right| \\ &= \left|\frac{\eta/2}{P_0 - \eta/2 + \eta/2} - \frac{1 - P_0 - \eta/2}{1 - P_0 - \eta/2 + \eta/2}\right| \\ &= 1 - \frac{\eta}{2P_0(1-P_0)} \end{aligned}$$

and

$$\mathcal{D}^{par}(h_0, \mathbb{P}_1) = \left|\mathbb{P}_{(X,A,Y)\sim\mathbb{P}_1}(h_0(X) = 1|A=0) - \mathbb{P}_{(X,A,Y)\sim\mathbb{P}_1}(h_0(X) = 1|A=1)\right|$$

$$= |0 - 1|$$
$$= 1,$$

so that $\mathcal{D}^{par}(h_0, \mathbb{P}_1) - \mathcal{D}^{par}(h_1, \mathbb{P}_1) = \frac{\eta}{2P_0(1-P_0)}$.

Consider a (randomized) malicious adversary $\mathcal{A}_i$ of power α, that given a clean distribution $\mathbb{P}_i$, changes every marked point to $(x_3, \neg i, i)$ with probability 0.5 and to $(x_4, i, \neg i)$ otherwise. Under a distribution $\mathbb{P}_i$ and an adversary $\mathcal{A}_i$, the probability of seeing a point $(x_3, i, \neg i)$ is $\frac{\eta}{2}(1-\alpha) = \frac{\eta}{2}\frac{1}{1+\eta} = \alpha/2$, which is equal to the probability of seeing a point $(x_3, \neg i, i)$. Therefore, denoting the probability distribution of the corrupted dataset, under a clean distribution $\mathbb{P}_i$ and an adversary $\mathcal{A}_i$, by $\mathbb{P}'_i$ (as a shorthand for $\mathbb{P}_i^{\mathcal{A}_i}$), we have

$$\mathbb{P}'_i(x, a, y) = \begin{cases} (1-\alpha)(1 - P_0 - \eta/2) & \text{if } x = x_1, a = 1, y = 1 \\ (1-\alpha)(P_0 - \eta/2) & \text{if } x = x_2, a = 0, y = 0 \\ \alpha/2 & \text{if } x = x_3, a = i, y = \neg i \\ \alpha/2 & \text{if } x = x_3, a = \neg i, y = i \\ \alpha/2 & \text{if } x = x_4, a = \neg i, y = i \\ \alpha/2 & \text{if } x = x_4, a = i, y = \neg i \\ 0 & \text{otherwise} \end{cases}$$

In particular, $\mathbb{P}'_0 = \mathbb{P}'_1$, so the two initial distributions $\mathbb{P}_0$ and $\mathbb{P}_1$ become indistinguishable under the adversarial manipulation.

Fix an arbitrary learner $\mathcal{L} : \cup_{n\in\mathbb{N}}(\mathcal{X} \times A \times \mathcal{Y})^n \to \{h_0, h_1\}$. Note that, if the clean distribution is $\mathbb{P}_0$, the events (in the probability space defined by the sampling of the poisoned train data)

$$\begin{aligned}\{L(\mathcal{L}(S^p), \mathbb{P}_0) - L(h_0, \mathbb{P}_0) \geq \eta\} &= \{\mathcal{L}(S^p) = h_1\} \\ &= \left\{\mathcal{D}^{par}(\mathcal{L}(S^p), \mathbb{P}_0) - \mathcal{D}^{par}(h_0, \mathbb{P}_0) \geq \frac{\eta}{2P_0(1-P_0)}\right\}\end{aligned}$$

are all the same. Similarly, if the clean distribution is $\mathbb{P}_1$

$$\begin{aligned}\{L(\mathcal{L}(S^p), \mathbb{P}_1) - L(h_1, \mathbb{P}_1) \geq \eta\} &= \{\mathcal{L}(S^p) = h_0\} \\ &= \left\{\mathcal{D}^{par}(\mathcal{L}(S^p), \mathbb{P}_1) - \mathcal{D}^{par}(h_1, \mathbb{P}_1) \geq \frac{\eta}{2P_0(1-P_0)}\right\}.\end{aligned}$$

Therefore, depending on whether we choose $\mathbb{P}_0$ or $\mathbb{P}_1$ as a clean distribution, we have

$$\begin{aligned}&\mathbb{P}_{S^p\sim\mathbb{P}'_0}\left((L(\mathcal{L}(S^p), \mathbb{P}_0) - L(h_0, \mathbb{P}_0) \geq \eta) \wedge \left(\mathcal{D}^{par}(\mathcal{L}(S^p), \mathbb{P}_0) - \mathcal{D}^{par}(h_0, \mathbb{P}_0) \geq \frac{\eta}{2P_0(1-P_0)}\right)\right) \\ &\quad = \mathbb{P}_{S^p\sim\mathbb{P}'_0}\left(\mathcal{L}(S^p) = h_1\right)\end{aligned}$$

and

$$\begin{aligned}&\mathbb{P}_{S^p\sim\mathbb{P}'_1}\left((L(\mathcal{L}(S^p), \mathbb{P}_1) - L(h_1, \mathbb{P}_1) \geq \eta) \wedge \left(\mathcal{D}^{par}(\mathcal{L}(S^p), \mathbb{P}_1) - \mathcal{D}^{par}(h_1, \mathbb{P}_1) \geq \frac{\eta}{2P_0(1-P_0)}\right)\right) \\ &\quad = \mathbb{P}_{S^p\sim\mathbb{P}'_1}\left(\mathcal{L}(S^p) = h_0\right)\end{aligned}$$

Finally, note that $\mathbb{P}'_0 = \mathbb{P}'_1$, so that either $\mathbb{P}_{S^p \sim \mathbb{P}'_0}(\mathcal{L}(S^p) = h_1) \geq 1/2$ or $\mathbb{P}_{S^p \sim \mathbb{P}'_1}(\mathcal{L}(S^p) = h_0) \geq 1/2$. Therefore, for at least one of $i = 0, 1$, both

$$L(\mathcal{L}(S^p), \mathbb{P}_i) - L(h_i, \mathbb{P}_i) \geq \eta = \frac{\alpha}{1-\alpha}$$

and

$$\mathcal{D}^{par}(\mathcal{L}(S^p), \mathbb{P}_i) - \mathcal{D}^{par}(h_i, \mathbb{P}_i) \geq \frac{\eta}{2P_0(1-P_0)} = \frac{\alpha}{2P_0(1-P_0)(1-\alpha)}$$

both hold with probability at least $1/2$ when the choice of distribution and adversary is $\mathbb{P}_i$ and $\mathcal{A}_i$ respectively. This concludes the proof in the first case.

Case 2 Now suppose that $\eta = \frac{\alpha}{1-\alpha} > 2P_0(1-P_0)$. Let $\alpha_1 \in (0, 0.5)$ be such that $\frac{\alpha_1}{1-\alpha_1} = 2P_0(1-P_0)$. Note that since $f(x) = \frac{x}{1-x}$ is monotonically increasing in (0,1), α_1 is unique and $\alpha_1 < \alpha$.

Now repeat the same construction as in Case 1, but with $\eta_1 = \frac{\alpha_1}{1-\alpha_1} = 2P_0(1-P_0)$. For every marked point, the adversary does the same as in Case 1 with probability α_1/α and does not change the point otherwise. Then the same argument as in Case 1 shows that for one $i \in \{0, 1\}$, both

$$L(\mathcal{L}(S^p), \mathbb{P}_i) - L(h_i, \mathbb{P}_i) \geq \eta_1 = \frac{\alpha_1}{1-\alpha_1} = 2P_0(1-P_0)$$

and

$$\mathcal{D}^{par}(\mathcal{L}(S^p), \mathbb{P}_i) - \mathcal{D}^{par}(h_i, \mathbb{P}_i) \geq \frac{\eta_1}{2P_0(1-P_0)} = 1$$

both hold with probability at least $1/2$. This concludes the proof of Theorem 1. ■

Theorem 2 *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

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

and

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

Proof Let $\eta = \frac{\alpha}{1-\alpha}$, so that $\alpha = \frac{\eta}{1+\eta}$.

Case 1 Assume that $\eta = \frac{\alpha}{1-\alpha} \leq 2\min\{P_{10}, 1 - P_{10} - P_{11}\}$. Take five distinct points $\{x_1, x_2, x_3, x_4, x_5\} \in \mathcal{X}$. We consider two distributions $\mathbb{P}_0$ and $\mathbb{P}_1$, where each $\mathbb{P}_i$ is defined as

$$\mathbb{P}_i(x, a, y) = \begin{cases} P_{11} & \text{if } x = x_1, a = 1, y = 1 \\ P_{10} - \eta/2 & \text{if } x = x_2, a = 0, y = 1 \\ \eta/2 & \text{if } x = x_3, a = i, y = \neg i \\ \eta/2 & \text{if } x = x_4, a = \neg i, y = i \\ 1 - P_{10} - P_{11} - \eta/2 & \text{if } x = x_5, a = 0, y = 0 \\ 0 & \text{otherwise} \end{cases}$$

Note that these are valid distributions, since $\eta \leq 2P_{10}, \eta \leq 2(1 - P_{10} - P_{11})$ by assumption, and that $P_{1a} = \mathbb{P}_i(A = a, Y = 1)$ for both $a \in \{0, 1\}, i \in \{0, 1\}$. Consider the hypothesis space $\mathcal{H} = \{h_0, h_1\}$, with

$$h_0(x_1) = 1 \quad h_0(x_2) = 1 \quad h_0(x_3) = 1 \quad h_0(x_4) = 0 \quad h_0(x_5) = 0$$

and

$$h_1(x_1) = 1 \quad h_1(x_2) = 1 \quad h_1(x_3) = 0 \quad h_1(x_4) = 1 \quad h_1(x_5) = 0$$

Note that $L(h_i, \mathbb{P}_i) = 0$ and $\mathcal{D}^{opp}(h_i, \mathbb{P}_i) = 0$ for both $i = 0, 1$. Note also that $L(h_1, \mathbb{P}_0) = L(h_0, \mathbb{P}_1) = \eta$. Moreover,

$$\begin{aligned} \mathcal{D}^{opp}(h_1, \mathbb{P}_0) &= \big|\mathbb{P}_{(X,A,Y)\sim\mathbb{P}_0}(h_1(X) = 1 | A = 0, Y = 1) \\ &\quad - \mathbb{P}_{(X,A,Y)\sim\mathbb{P}_0}(h_1(X) = 1 | A = 1, Y = 1)\big| \\ &= \left|\frac{P_{10} - \eta/2}{P_{10} - \eta/2 + \eta/2} - 1\right| \\ &= \frac{\eta}{2P_{10}} \end{aligned}$$

and similarly $\mathcal{D}^{opp}(h_0, \mathbb{P}_1) = \frac{\eta}{2P_{10}}$.

Consider a (randomized) malicious adversary $\mathcal{A}_i$ of power α, that given a clean distribution $\mathbb{P}_i$, changes every marked point to $(x_3, \neg i, i)$ with probability 0.5 and to $(x_4, i, \neg i)$ otherwise. Under a distribution $\mathbb{P}_i$ and an adversary $\mathcal{A}_i$, the probability of seeing a point $(x_3, i, \neg i)$ is $\frac{\eta}{2}(1 - \alpha) = \frac{\eta}{2}\frac{1}{1+\eta} = \alpha/2$, which is equal to the probability of seeing a point $(x_3, \neg i, i)$. Therefore, denoting the probability distribution of the corrupted dataset, under a clean distribution $\mathbb{P}_i$ and an adversary $\mathcal{A}_i$, by $\mathbb{P}'_i$, we have

$$\mathbb{P}'_i(x, a, y) = \begin{cases} (1 - \alpha)P_{11} & \text{if } x = x_1, a = 1, y = 1 \\ (1 - \alpha)(P_{10} - \eta/2) & \text{if } x = x_2, a = 0, y = 1 \\ \alpha/2 & \text{if } x = x_3, a = i, y = \neg i \\ \alpha/2 & \text{if } x = x_3, a = \neg i, y = i \\ \alpha/2 & \text{if } x = x_4, a = \neg i, y = i \\ \alpha/2 & \text{if } x = x_4, a = i, y = \neg i \\ (1 - \alpha)(1 - P_{10} - P_{11} - \eta/2) & \text{if } x = x_5, a = 0, y = 0 \\ 0 & \text{otherwise} \end{cases}$$

In particular, $\mathbb{P}'_0 = \mathbb{P}'_1$, so the two initial distributions $\mathbb{P}_0$ and $\mathbb{P}_1$ become indistinguishable under the adversarial manipulation.

Fix an arbitrary learner $\mathcal{L} : \cup_{n\in\mathbb{N}}(\mathcal{X} \times A \times \mathcal{Y})^n \to \{h_0, h_1\}$. Note that, if the clean distribution is $\mathbb{P}_0$, the events (in the probability space defined by the sampling of the poisoned train data)

$$\begin{aligned}\{L(\mathcal{L}(S^p), \mathbb{P}_0) - L(h_0, \mathbb{P}_0) \geq \eta\} &= \{\mathcal{L}(S^p) = h_1\} \\ &= \left\{\mathcal{D}^{opp}(\mathcal{L}(S^p), \mathbb{P}_0) - \mathcal{D}^{opp}(h_0, \mathbb{P}_0) \geq \frac{\eta}{2P_{10}}\right\}\end{aligned}$$

are all the same. Similarly, if the clean distribution is $\mathbb{P}_1$

$$\begin{aligned}\{L(\mathcal{L}(S^p), \mathbb{P}_1) - L(h_1, \mathbb{P}_1) \geq \eta\} &= \{\mathcal{L}(S^p) = h_0\} \\ &= \left\{\mathcal{D}^{opp}(\mathcal{L}(S^p), \mathbb{P}_1) - \mathcal{D}^{opp}(h_1, \mathbb{P}_1) \geq \frac{\eta}{2P_{10}}\right\}.\end{aligned}$$

Therefore, depending on whether we choose $\mathbb{P}_0$ or $\mathbb{P}_1$ as a clean distribution, we have

$$\begin{aligned}&\mathbb{P}_{S^p\sim\mathbb{P}'_0}\left(L(\mathcal{L}(S^p), \mathbb{P}_0) - L(h_0, \mathbb{P}_0) \geq \eta \wedge \mathcal{D}^{opp}(\mathcal{L}(S^p), \mathbb{P}_0) - \mathcal{D}^{opp}(h_0, \mathbb{P}_0) \geq \frac{\eta}{2P_{10}}\right) \\ &\quad = \mathbb{P}_{S^p\sim\mathbb{P}'_0}\left(\mathcal{L}(S^p) = h_1\right)\end{aligned}$$

and

$$\begin{aligned}&\mathbb{P}_{S^p\sim\mathbb{P}'_1}\left(L(\mathcal{L}(S^p), \mathbb{P}_1) - L(h_1, \mathbb{P}_1) \geq \eta \wedge \mathcal{D}^{opp}(\mathcal{L}(S^p), \mathbb{P}_1) - \mathcal{D}^{opp}(h_1, \mathbb{P}_1) \geq \frac{\eta}{2P_{10}}\right) \\ &\quad = \mathbb{P}_{S^p\sim\mathbb{P}'_1}\left(\mathcal{L}(S^p) = h_0\right)\end{aligned}$$

Finally, note that $\mathbb{P}'_0 = \mathbb{P}'_1$, so that either $\mathbb{P}_{S^p\sim\mathbb{P}'_0}\left(\mathcal{L}(S^p) = h_1\right) \geq 1/2$ or $\mathbb{P}_{S^p\sim\mathbb{P}'_1}\left(\mathcal{L}(S^p) = h_0\right) \geq 1/2$. Therefore, for at least one of $i = 0, 1$, both

$$L(\mathcal{L}(S^p), \mathbb{P}_i) - L(h_i, \mathbb{P}_i) \geq \eta = \frac{\alpha}{1-\alpha}$$

and

$$\mathcal{D}^{opp}(\mathcal{L}(S^p), \mathbb{P}_i) - \mathcal{D}^{opp}(h_i, \mathbb{P}_i) \geq \frac{\eta}{2P_{10}} = \frac{\alpha}{2P_{10}(1-\alpha)}$$

both hold with probability at least $1/2$. This concludes the proof of the first case.

Case 2 Now assume that $\frac{\alpha}{1-\alpha} > 2\min\{P_{10}, 1 - P_{10} - P_{11}\}$. We distinguish two cases:

Case 2.1 Suppose that $P_{10} \leq 1 - P_{10} - P_{11}$. We have that $\frac{\alpha}{1-\alpha} > 2P_{10}$. Then, denote by α_1 the unique number between $(0, 0.5)$, such that $\frac{\alpha_1}{1-\alpha_1} = 2P_{10} = 2\min\{P_{10}, 1 - P_{10} - P_{11}\}$, and note that $\alpha_1 < \alpha$. Then repeat the same construction as in Case 1, but with $\eta_1 = \frac{\alpha_1}{1-\alpha_1}$ and an adversary that with probability α_1/α does the same as in Case 1 and leaves a marked point untouched otherwise.

Then the same argument as in Case 1 gives that for some $i \in \{0, 1\}$, with probability at least 0.5, both of the following hold

$$L(\mathcal{L}(S^p), \mathbb{P}_i) - L(h_i, \mathbb{P}_i) \geq \frac{\alpha_1}{1 - \alpha_1} = 2P_{10}$$

and

$$\mathcal{D}^{opp}(\mathcal{L}(S^p), \mathbb{P}_i) - \mathcal{D}^{opp}(h_i, \mathbb{P}_i) \geq \frac{\eta_1}{2P_{10}} = 1.$$

Case 2.2 In the case when $1 - P_{10} - P_{11} < P_{10}$ we have that $\frac{\alpha}{1-\alpha} > 2(1 - P_{10} - P_{11})$. Then, denote by α_2 the unique number between $(0, 0.5)$, such that $\frac{\alpha_2}{1-\alpha_2} = 2(1 - P_{10} - P_{11}) = 2\min\{P_{10}, 1 - P_{10} - P_{11}\}$, and note that $\alpha_2 < \alpha$. Then repeat the same construction as in Case 1, but with $\eta_2 = \frac{\alpha_2}{1-\alpha_2}$ and an adversary that with probability α_2/α does the same as in Case 1 and leaves a marked point untouched otherwise.

Then the same argument as in Case 1 gives that for some $i \in \{0, 1\}$, with probability at least 0.5, both of the following hold

$$L(\mathcal{L}(S^p), \mathbb{P}_i) - L(h_i, \mathbb{P}_i) \geq \frac{\alpha_2}{1 - \alpha_2} = 2(1 - P_{10} - P_{11})$$

and

$$\mathcal{D}^{opp}(\mathcal{L}(S^p), \mathbb{P}_i) - \mathcal{D}^{opp}(h_i, \mathbb{P}_i) \geq \frac{\eta_2}{2P_{10}} = \frac{1 - P_{10} - P_{11}}{P_{10}}.$$

This concludes the proof of Theorem 2. ■

A.2 Hurting fairness without affecting accuracy - proofs

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

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

and

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

Proof Let $\eta = \frac{\alpha}{1-\alpha}$, so that $\alpha = \frac{\eta}{1+\eta}$.

Case 1 First assume that $\eta = \frac{\alpha}{1-\alpha} \leq 2P_0(1-P_0)$. Take four distinct points $\{x_1, x_2, x_3, x_4\} \in \mathcal{X}$. We consider two distributions $\mathbb{P}_0$ and $\mathbb{P}_1$, where each $\mathbb{P}_i$ is defined as

$$\mathbb{P}_i(x, a, y) = \begin{cases} 1 - P_0 - \eta/2 & \text{if } x = x_1, a = 1, y = 1 \\ P_0 - \eta/2 & \text{if } x = x_2, a = 0, y = 0 \\ \eta/2 & \text{if } x = x_3, a = i, y = 1 \\ \eta/2 & \text{if } x = x_4, a = \neg i, y = 1 \\ 0 & \text{otherwise} \end{cases}$$

Note that these are valid distributions, since $\eta \leq 2P_0(1-P_0) \leq 2P_0 \leq 2(1-P_0)$ by assumption and also that $P_0 = \mathbb{P}_i(A=0)$ for both $i \in \{0,1\}$. Consider the hypothesis space $\mathcal{H} = \{h_0, h_1\}$, with

$$h_0(x_1) = 1 \quad h_0(x_2) = 0 \quad h_0(x_3) = 1 \quad h_0(x_4) = 0$$

and

$$h_1(x_1) = 1 \quad h_1(x_2) = 0 \quad h_1(x_3) = 0 \quad h_1(x_4) = 1.$$

Note that $L(h_i, \mathbb{P}_i) = L(h_{\neg i}, \mathbb{P}_i) = \eta/2$ for both $i = 0, 1$. Moreover,

$$\begin{aligned}
\mathcal{D}^{par}(h_0, \mathbb{P}_0) &= \left|\mathbb{P}_{(X,A,Y)\sim\mathbb{P}_0}(h_0(X)=1|A=0) - \mathbb{P}_{(X,A,Y)\sim\mathbb{P}_0}(h_0(X)=1|A=1)\right| \\
&= \left|\frac{\eta/2}{P_0 - \eta/2 + \eta/2} - \frac{1-P_0-\eta/2}{1-P_0-\eta/2+\eta/2}\right| \\
&= \left|\frac{\eta}{2P_0} - \frac{2-2P_0-\eta}{2(1-P_0)}\right| \\
&= \left|\frac{\eta}{2P_0(1-P_0)} - 1\right| \\
&= 1 - \frac{\eta}{2P_0(1-P_0)},
\end{aligned}$$

since $\eta \leq 2P_0(1-P_0)$ by assumption. Furthermore, $\mathcal{D}^{par}(h_1, \mathbb{P}_0) = 1$, so that $\mathcal{D}^{par}(h_1, \mathbb{P}_0) - \mathcal{D}^{par}(h_0, \mathbb{P}_0) = \frac{\eta}{2P_0(1-P_0)}$. Similarly,

$$\begin{aligned}
\mathcal{D}^{par}(h_1, \mathbb{P}_1) &= \left|\mathbb{P}_{(X,A,Y)\sim\mathbb{P}_1}(h_1(X)=1|A=0) - \mathbb{P}_{(X,A,Y)\sim\mathbb{P}_1}(h_1(X)=1|A=1)\right| \\
&= \left|\frac{\eta/2}{P_0 - \eta/2 + \eta/2} - \frac{1-P_0-\eta/2}{1-P_0-\eta/2+\eta/2}\right| \\
&= 1 - \frac{\eta}{2P_0(1-P_0)}
\end{aligned}$$

and $\mathcal{D}^{par}(h_0, \mathbb{P}_1) = 1$.

Consider a (randomized) malicious adversary $\mathcal{A}_i$ of power α, that given a clean distribution $\mathbb{P}_i$, changes every marked point to $(x_3, \neg i, 1)$ with probability 0.5 and to $(x_4, i, 1)$ otherwise. Under a distribution $\mathbb{P}_i$ and an adversary $\mathcal{A}_i$, the probability of seeing a point $(x_3, i, 1)$ is $\frac{\eta}{2}(1-\alpha) = \frac{\eta}{2}\frac{1}{1+\eta} = \alpha/2$, which is equal to the probability of seeing a point $(x_3, \neg i, 1)$. Therefore, denoting the probability distribution of the corrupted dataset, under a clean distribution $\mathbb{P}_i$ and an adversary $\mathcal{A}_i$, by $\mathbb{P}'_i$, we have

$$\mathbb{P}'_i(x,a,y) = \begin{cases}
(1-\alpha)(1-P_0-\eta/2) & \text{if } x = x_1, a = 1, y = 1 \\
(1-\alpha)(P_0-\eta/2) & \text{if } x = x_2, a = 0, y = 0 \\
\alpha/2 & \text{if } x = x_3, a = i, y = 1 \\
\alpha/2 & \text{if } x = x_3, a = \neg i, y = 1 \\
\alpha/2 & \text{if } x = x_4, a = \neg i, y = 1 \\
\alpha/2 & \text{if } x = x_4, a = i, y = 1 \\
0 & \text{otherwise}
\end{cases}$$

In particular, $\mathbb{P}'_0 = \mathbb{P}'_1$, so the two initial distributions $\mathbb{P}_0$ and $\mathbb{P}_1$ become indistinguishable under the adversarial manipulation.

Fix an arbitrary learner $\mathcal{L} : \cup_{n\in\mathbb{N}}(\mathcal{X} \times A \times \mathcal{Y})^n \to \{h_0, h_1\}$. Note that, if the clean distribution is $\mathbb{P}_0$, the events (in the probability space defined by the sampling of the poisoned train data)

$$\{\mathcal{L}(S^p) = h_1\} = \left\{\mathcal{D}^{par}(\mathcal{L}(S^p), \mathbb{P}_0) - \mathcal{D}^{par}(h_0, \mathbb{P}_0) \geq \frac{\eta}{2P_0(1-P_0)}\right\}$$

are all the same. Similarly, if the clean distribution is $\mathbb{P}_1$

$$\{\mathcal{L}(S^p) = h_0\} = \left\{\mathcal{D}^{par}(\mathcal{L}(S^p), \mathbb{P}_1) - \mathcal{D}^{par}(h_1, \mathbb{P}_1) \geq \frac{\eta}{2P_0(1-P_0)}\right\}.$$

Therefore, depending on whether we choose $\mathbb{P}_0$ or $\mathbb{P}_1$ as a clean distribution, we have

$$\mathbb{P}_{S^p\sim\mathbb{P}'_0}\left(\mathcal{D}^{par}(\mathcal{L}(S^p), \mathbb{P}_0) - \mathcal{D}^{par}(h_0, \mathbb{P}_0) \geq \frac{\eta}{2P_0(1-P_0)}\right) = \mathbb{P}_{S^p\sim\mathbb{P}'_0}\left(\mathcal{L}(S^p) = h_1\right)$$

and

$$\mathbb{P}_{S^p\sim\mathbb{P}'_1}\left(\mathcal{D}^{par}(\mathcal{L}(S^p), \mathbb{P}_1) - \mathcal{D}^{par}(h_1, \mathbb{P}_1) \geq \frac{\eta}{2P_0(1-P_0)}\right) = \mathbb{P}_{S^p\sim\mathbb{P}'_1}\left(\mathcal{L}(S^p) = h_0\right)$$

Finally, note that $\mathbb{P}'_0 = \mathbb{P}'_1$, so that either $\mathbb{P}_{S^p\sim\mathbb{P}'_0}\left(\mathcal{L}(S^p) = h_1\right) \geq 1/2$ or $\mathbb{P}_{S^p\sim\mathbb{P}'_1}\left(\mathcal{L}(S^p) = h_0\right) \geq 1/2$. Furthermore, $L(\mathcal{L}(S^p), \mathbb{P}_i) = \eta/2$ holds for both $i \in \{0, 1\}$, for any realization of the randomness. Therefore, for at least one of $i = 0, 1$, both

$$L(\mathcal{L}(S^p), \mathbb{P}_i) = L(h_i, \mathbb{P}_i) = \frac{\eta}{2}$$

and

$$\mathcal{D}^{par}(\mathcal{L}(S^p), \mathbb{P}_i) - \mathcal{D}^{par}(h_i, \mathbb{P}_i) \geq \frac{\eta}{2P_0(1-P_0)} = \frac{\alpha}{2P_0(1-P_0)(1-\alpha)}$$

both hold with probability at least $1/2$. This concludes the proof in the first case.

Case 2 Now suppose that $\eta = \frac{\alpha}{1-\alpha} > 2P_0(1-P_0)$. Let $\alpha_1 \in (0, 0.5)$ be such that $\frac{\alpha_1}{1-\alpha_1} = 2P_0(1-P_0)$. Note that since $f(x) = \frac{x}{1-x}$ is monotonically increasing in (0,1), α_1 is unique and $\alpha_1 < \alpha$.

Now repeat the same construction as in Case 1, but with $\eta_1 = \frac{\alpha_1}{1-\alpha_1} = 2P_0(1-P_0)$. For every marked point, the adversary does the same as in Case 1 with probability α_1/α and does not change the point otherwise. Then the same argument as in Case 1 shows that for one $i \in \{0, 1\}$, both

$$L(\mathcal{L}(S^p), \mathbb{P}_i) = L(h_i, \mathbb{P}_i) = \frac{\eta_1}{2} = P_0(1-P_0)$$

and

$$\mathcal{D}^{par}(\mathcal{L}(S^p), \mathbb{P}_i) - \mathcal{D}^{par}(h_i, \mathbb{P}_i) \geq \frac{\eta_1}{2P_0(1-P_0)} = 1$$

both hold with probability at least $1/2$. This concludes the proof of Theorem 3. ■

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

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

and

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

Proof Let $\eta = \frac{\alpha}{1-\alpha}$, so that $\alpha = \frac{\eta}{1+\eta}$.

Case 1 First assume that $\eta \leq 2P_{10}$. Take five distinct points $\{x_1, x_2, x_3, x_4, x_5\} \in \mathcal{X}$. We consider two distributions $\mathbb{P}_0$ and $\mathbb{P}_1$, where each $\mathbb{P}_i$ is defined as

$$\mathbb{P}_i(x, a, y) = \begin{cases} P_{11} - \eta/2 & \text{if } x = x_1, a = 1, y = 1 \\ P_{10} - \eta/2 & \text{if } x = x_2, a = 0, y = 1 \\ \eta/2 & \text{if } x = x_3, a = i, y = 1 \\ \eta/2 & \text{if } x = x_4, a = \neg i, y = 1 \\ 1 - P_{10} - P_{11} & \text{if } x = x_5, a = 0, y = 0 \\ 0 & \text{otherwise} \end{cases}$$

Note that these are valid distributions, since $\eta \leq 2P_{10} \leq 2P_{11}$ by assumption, and that $P_{1a} = \mathbb{P}_i(A = a, Y = 1)$ for both $a \in \{0, 1\}, i \in \{0, 1\}$. Consider the hypothesis space $\mathcal{H} = \{h_0, h_1\}$, with

$$h_0(x_1) = 1 \quad h_0(x_2) = 1 \quad h_0(x_3) = 1 \quad h_0(x_4) = 0 \quad h_0(x_5) = 0$$

and

$$h_1(x_1) = 1 \quad h_1(x_2) = 1 \quad h_1(x_3) = 0 \quad h_1(x_4) = 1 \quad h_1(x_5) = 0$$

Note that $L(h_i, \mathbb{P}_i) = L(h_{\neg i}, \mathbb{P}_i) = \eta/2$. Moreover,

$$\begin{aligned} \mathcal{D}^{opp}(h_0, \mathbb{P}_0) &= \left|\mathbb{P}_{(X,A,Y)\sim\mathbb{P}_0}(h_1(X) = 1 | A = 0, Y = 1)\right. \\ &\quad \left. - \mathbb{P}_{(X,A,Y)\sim\mathbb{P}_0}(h_1(X) = 1 | A = 1, Y = 1)\right| \\ &= \left|1 - \frac{P_{11} - \eta/2}{P_{11} - \eta/2 + \eta/2}\right| \\ &= \frac{\eta}{2P_{11}} \end{aligned}$$

and similarly $\mathcal{D}^{opp}(h_1, \mathbb{P}_0) = \frac{\eta}{2P_{10}}$. Since $P_{10} \leq P_{11}$, $\mathcal{D}^{opp}(h_0, \mathbb{P}_0) \leq \mathcal{D}^{opp}(h_1, \mathbb{P}_0)$ and

$$\mathcal{D}^{opp}(h_1, \mathbb{P}_0) - \mathcal{D}^{opp}(h_0, \mathbb{P}_0) = \frac{\eta}{2P_{10}}\left(1 - \frac{P_{10}}{P_{11}}\right).$$

Similarly $\mathcal{D}^{opp}(h_0, \mathbb{P}_1) = \frac{\eta}{2P_{10}}$ and $\mathcal{D}^{opp}(h_1, \mathbb{P}_1) = \frac{\eta}{2P_{11}}$, so that $\mathcal{D}^{opp}(h_1, \mathbb{P}_1) \leq \mathcal{D}^{opp}(h_0, \mathbb{P}_1)$ and

$$\mathcal{D}^{opp}(h_0, \mathbb{P}_1) - \mathcal{D}^{opp}(h_1, \mathbb{P}_1) = \frac{\eta}{2P_{10}} \left(1 - \frac{P_{10}}{P_{11}}\right).$$

Consider a (randomized) malicious adversary $\mathcal{A}_i$ of power α, that given a clean distribution $\mathbb{P}_i$, changes every marked point to $(x_3, \neg i, 1)$ with probability 0.5 and to $(x_4, i, 1)$ otherwise. Under a distribution $\mathbb{P}_i$ and an adversary $\mathcal{A}_i$, the probability of seeing a point $(x_3, i, 1)$ is $\frac{\eta}{2}(1-\alpha) = \frac{\eta}{2}\frac{1}{1+\eta} = \alpha/2$, which is equal to the probability of seeing a point $(x_3, \neg i, 1)$. Therefore, denoting the probability distribution of the corrupted dataset, under a clean distribution $\mathbb{P}_i$ and an adversary $\mathcal{A}_i$, by $\mathbb{P}'_i$, we have

$$\mathbb{P}'_i(x, a, y) = \begin{cases} (1-\alpha)(P_{11} - \eta/2) & \text{if } x = x_1, a = 1, y = 1 \\ (1-\alpha)(P_{10} - \eta/2) & \text{if } x = x_2, a = 0, y = 1 \\ \alpha/2 & \text{if } x = x_3, a = i, y = 1 \\ \alpha/2 & \text{if } x = x_3, a = \neg i, y = 1 \\ \alpha/2 & \text{if } x = x_4, a = \neg i, y = 1 \\ \alpha/2 & \text{if } x = x_4, a = i, y = 1 \\ (1-\alpha)(1 - P_{10} - P_{11}) & \text{if } x = x_5, a = 0, y = 0 \\ 0 & \text{otherwise} \end{cases}$$

In particular, $\mathbb{P}'_0 = \mathbb{P}'_1$, so the two initial distributions $\mathbb{P}_0$ and $\mathbb{P}_1$ become indistinguishable under the adversarial manipulation.

Fix an arbitrary learner $\mathcal{L} : \cup_{n \in \mathbb{N}} (\mathcal{X} \times A \times \mathcal{Y})^n \to \{h_0, h_1\}$. Note that, if the clean distribution is $\mathbb{P}_0$, the events (in the probability space defined by the sampling of the poisoned train data)

$$\{\mathcal{L}(S^p) = h_1\} = \left\{ \mathcal{D}^{opp}(\mathcal{L}(S^p), \mathbb{P}_0) - \mathcal{D}^{opp}(h_0, \mathbb{P}_0) \geq \frac{\eta}{2P_{10}} \left(1 - \frac{P_{10}}{P_{11}}\right) \right\}$$

are all the same. Similarly, if the clean distribution is $\mathbb{P}_1$

$$\{\mathcal{L}(S^p) = h_0\} = \left\{ \mathcal{D}^{opp}(\mathcal{L}(S^p), \mathbb{P}_1) - \mathcal{D}^{opp}(h_1, \mathbb{P}_1) \geq \frac{\eta}{2P_{10}} \left(1 - \frac{P_{10}}{P_{11}}\right) \right\}.$$

Therefore, depending on whether we choose $\mathbb{P}_0$ or $\mathbb{P}_1$ as a clean distribution, we have

$$\mathbb{P}_{S^p \sim \mathbb{P}'_0} \left(\mathcal{D}^{opp}(\mathcal{L}(S^p), \mathbb{P}_0) - \mathcal{D}^{opp}(h_0, \mathbb{P}_0) \geq \frac{\eta}{2P_{10}} \left(1 - \frac{P_{10}}{P_{11}}\right) \right) = \mathbb{P}_{S^p \sim \mathbb{P}'_0} \left(\mathcal{L}(S^p) = h_1\right)$$

and

$$\mathbb{P}_{S^p \sim \mathbb{P}'_1} \left(\mathcal{D}^{opp}(\mathcal{L}(S^p), \mathbb{P}_1) - \mathcal{D}^{opp}(h_1, \mathbb{P}_1) \geq \frac{\eta}{2P_{10}} \left(1 - \frac{P_{10}}{P_{11}}\right) \right) = \mathbb{P}_{S^p \sim \mathbb{P}'_1} \left(\mathcal{L}(S^p) = h_0\right)$$

Finally, note that $\mathbb{P}'_0 = \mathbb{P}'_1$, so that either $\mathbb{P}_{S^p \sim \mathbb{P}'_0} \left(\mathcal{L}(S^p) = h_1\right) \geq 1/2$ or $\mathbb{P}_{S^p \sim \mathbb{P}'_1} \left(\mathcal{L}(S^p) = h_0\right) \geq 1/2$. Moreover, $L(\mathcal{L}(S^p), \mathbb{P}_i) = L(h_i, \mathbb{P}_i) = \eta/2$ holds for both $i \in \{0, 1\}$, for any realization of the randomness. Therefore, for at least one of $i = 0, 1$, both

$$L(\mathcal{L}(S^p), \mathbb{P}_i) = L(h_i, \mathbb{P}_i) = \frac{\eta}{2}$$

and

$$\mathcal{D}^{opp}(\mathcal{L}(S^p), \mathbb{P}_i) - \mathcal{D}^{opp}(h_i, \mathbb{P}_i) \geq \frac{\eta}{2P_{10}}\left(1 - \frac{P_{10}}{P_{11}}\right) = \frac{\alpha}{2P_{10}(1-\alpha)}\left(1 - \frac{P_{10}}{P_{11}}\right)$$

both hold with probability at least $1/2$. This concludes the proof in the first case.

Case 2 Now assume that $\frac{\alpha}{1-\alpha} > 2P_{10}$. Then denote by α_1 the unique number between $(0, 0.5)$, such that $\frac{\alpha_1}{1-\alpha_1} = 2P_{10}$, and note that $\alpha_1 < \alpha$. Then repeat the same construction as in Case 1, but with $\eta_1 = \frac{\alpha_1}{1-\alpha_1}$ and an adversary that with probability α_1/α does the same as in Case 1 and leaves a marked point untouched otherwise.

Then the same argument as in Case 1 gives that for some $i \in \{0, 1\}$, with probability at least 0.5, both of the following hold

$$L(\mathcal{L}(S^p), \mathbb{P}_i) = L(h_i, \mathbb{P}_i) = \frac{\eta_1}{2} = P_{10}$$

and

$$\mathcal{D}^{opp}(\mathcal{L}(S^p), \mathbb{P}_i) - \mathcal{D}^{opp}(h_i, \mathbb{P}_i) \geq \frac{\eta_1}{2P_{10}} = \frac{\eta_1}{2P_{10}}\left(1 - \frac{P_{10}}{P_{11}}\right) = 1 - \frac{P_{10}}{P_{11}}.$$

This concludes the proof of Theorem 4. ■

Appendix B. Upper bounds proofs

We now present the complete proofs of our upper bounds. The main challenge lies in understanding the concentration properties of the empirical estimates of the fairness measures, as introduced in the main body of the paper. To this end, we first bound the effect that the data corruption may have on these estimates. We then leverage classic concentration techniques to relate the "ideal" clean data estimates to the corresponding population fairness measures.

B.1 Concentration tools and notation

We will use the following versions of the classic Chernoff bounds for large deviations of Binomial random variables, as they can be found, for example, in Kearns and Li (1993). Let $X \sim Bin(n, p)$. Then

$$\mathbb{P}(X \leq (1-\alpha)pn) \leq e^{-\alpha^2 np/2}$$

and

$$\mathbb{P}(X \geq (1+\alpha)pn) \leq e^{-\alpha^2 np/3},$$

for any $\alpha \in (0, 1)$. We will also use the Hoeffding's inequality (Hoeffding, 1963). Let $X_1, X_2, \ldots, X_n$ be independent random variables, such that each X_i is bounded in $[a_i, b_i]$ and let $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Then

$$\mathbb{P}\left(\left|\bar{X} - \mathbb{E}(\bar{X})\right| > t\right) \leq 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Throughout the section we denote the clean data distribution by $\mathbb{P} \in \mathcal{P}(\mathcal{X} \times A \times \mathcal{Y})$. As in the main body of the paper, we denote $P_a = \mathbb{P}(A = a)$ and $P_{1a} = \mathbb{P}(Y = 1, A = a)$ for both $a \in \{0, 1\}$. We assume without loss of generality that $0 < P_0 \leq \frac{1}{2} \leq P_1$ (when studying demographic parity) and $0 < P_{10} \leq P_{11}$ (when studying equal opportunity).

We will be interested in the concentration properties of certain empirical estimates based on the corrupted data S^p. Therefore, we denote the distribution that corresponds to all the randomness of the sampling of S^p, that is the randomness of the clean data, the marked points and the adversary, by $\mathbb{P}^{\mathcal{A}}$. Here we consider both $\mathbb{P}$ and $\mathcal{A}$ arbitrary, but fixed.

B.2 Concentration results

We study the concentration of the demographic parity and the equal opportunity fairness estimates in Sections B.2.1 and B.2.2 respectively.

B.2.1 Concentration for demographic parity

We use the notation $C_a = \sum_{i=1}^n \mathbb{1}\{a_i^p = a, i \notin \mathfrak{M}\}$ for the number of points in S^p that *were not* marked (that is, are *clean*) and contain a point from protected group a and $B_a = \sum_{i=1}^n \mathbb{1}\{a_i^p = a, i \in \mathfrak{M}\}$ for the number of points in S^p that *were* marked (that is, are potentially *bad*[6]) and contain a point from protected group a. Note that $B_0 + B_1 = |\mathfrak{M}|$ is the total number of poisoned points, which is $\mathrm{Bin}(n, \alpha)$, and $B_0 + B_1 = n - C_0 - C_1$. Similarly, denote by $C_a^1(h) = \sum_{i=1}^n \mathbb{1}\{h(x_i^p) = 1, a_i^p = a, i \notin \mathfrak{M}\}$ and $B_a^1(h) = \sum_{i=1}^n \mathbb{1}\{h(x_i^p) = 1, a_i^p = a, i \in \mathfrak{M}\}$.
Denote

$$\gamma_a^p(h) = \frac{\sum_{i=1}^n \mathbb{1}\{h(x_i^p) = 1, a_i^p = a\}}{\sum_{i=1}^n \mathbb{1}\{a_i^p = a\}}$$

and

$$\gamma_a(h) = \mathbb{P}(h(X) = 1 | A = a),$$

so that $\widehat{\mathcal{D}}^{par}(h) = |\gamma_0^p(h) - \gamma_1^p(h)|$ and $\mathcal{D}^{par}(h) = |\gamma_0(h) - \gamma_1(h)|$. Note that $\gamma_a^p(h)$ is an estimate of a conditional probability *based on the corrupted data*. We now introduce the corresponding estimate that only uses the clean (but unknown) subset of the training set S^p

$$\gamma_a^c(h) = \frac{C_a^1(h)}{C_a(h)} = \frac{\sum_{i=1}^n \mathbb{1}\{h(x_i^p) = 1, a_i^p = a, i \notin \mathfrak{M}\}}{\sum_{i=1}^n \mathbb{1}\{a_i^p = a, i \notin \mathfrak{M}\}}.$$

First we bound how far the corrupted estimates $\gamma_a^p(h)$ of $\gamma_a(h)$ are from the clean estimates $\gamma_a^c(h)$, uniformly over the hypothesis space $\mathcal{H}$:

Lemma 1 *If* $n \geq \max\left\{\frac{8\log(4/\delta)}{(1-\alpha)P_0}, \frac{12\log(3/\delta)}{\alpha}\right\}$, *we have*

$$\mathbb{P}^{\mathcal{A}}\left(\sup_{h \in \mathcal{H}} \left(|\gamma_0^p(h) - \gamma_0^c(h)| + |\gamma_1^p(h) - \gamma_1^c(h)|\right) \geq \frac{2\alpha}{P_0/3 + \alpha}\right) < \delta. \tag{26}$$

6. We use B_a with B for *bad* here, instead of P for *poisoned*, to avoid confusion with the protected group frequencies P_i.

Proof First we show that certain bounds on the random variables B_a and C_a hold with high probability. Then we show that the supremum in equation (26) is bounded when these bounds hold.

Step 1 Specifically, since $B_0 + B_1 \sim \text{Bin}(n, \alpha)$, by the Chernoff bounds and the assumption on n

$$\mathbb{P}^{\mathcal{A}}\left(B_0 + B_1 \geq \frac{3\alpha}{2}n\right) \leq e^{-\alpha n/12} \leq \frac{\delta}{3}.$$

Similarly, $C_0 \sim \text{Bin}(n, (1-\alpha)P_0)$ and $C_1 \sim \text{Bin}(n, (1-\alpha)P_1)$ and since $P_0 \leq P_1$ we get

$$\mathbb{P}^{\mathcal{A}}\left(C_0 \leq \frac{1-\alpha}{2}P_0 n\right) \leq e^{-(1-\alpha)P_0 n/8} \leq \frac{\delta}{4}$$

and

$$\mathbb{P}^{\mathcal{A}}\left(C_1 \leq \frac{1-\alpha}{2}P_1 n\right) \leq e^{-(1-\alpha)P_1 n/8} \leq \frac{\delta}{4}$$

Therefore, by a union bound

$$\mathbb{P}^{\mathcal{A}}\left(\left(B_0 + B_1 \geq \frac{3\alpha}{2}n\right) \vee \left(C_0 \leq \frac{1-\alpha}{2}P_0 n\right) \vee \left(C_1 \leq \frac{1-\alpha}{2}P_1 n\right)\right) \leq \frac{\delta}{3} + \frac{\delta}{4} + \frac{\delta}{4} < \delta.$$

Step 2 Now assume that all of $B_0 + B_1 < \frac{3\alpha}{2}n$, $C_0 > \frac{1-\alpha}{2}P_0 n$, $C_1 > \frac{1-\alpha}{2}P_1 n$ hold. This happens with probability at least $1-\delta$ according to Step 1. Let h be an arbitrary classifier. Since we consider h fixed, we will drop the dependence on h from the notation for the rest of this proof and write $\gamma_a^p = \gamma_a^p(h), C_a^1 = C_a^1(h)$, etc.

We now prove that for both $a \in \{0, 1\}$

$$\Delta_a := |\gamma_a^p - \gamma_a^c| \leq \frac{B_a}{C_a + B_a}. \tag{27}$$

For each $a \in \{0, 1\}$, this can be shown as follows. First, if $\sum_{i=1}^n \mathbb{1}\{a_i^p = a\} = B_a + C_a = 0$, then both $\gamma_a^p(h)$ and $\gamma_a^c(h)$ are equal to 0, because of the convention that $\frac{0}{0} = 0$. In addition, $B_a = C_a = 0$. Therefore, inequality (27) trivially holds.

Similarly, if $B_a = 0$, but $C_a > 0$, then $\gamma_a^p(h) = \gamma_a^c(h)$ and so $\Delta_a = 0$ and (27) holds.

Assume now that $B_a > 0$. Note that if $C_a = \sum_{i=1}^n \mathbb{1}\{a_i^p = a, i \notin \mathfrak{M}\} = 0$, then

$$\Delta_a = |\gamma_a^p(h) - \gamma_a^c(h)| = \left|\frac{B_a^1}{B_a} - 0\right| = \frac{B_a^1}{B_a} = \frac{B_a^1}{B_a + C_a} \leq \frac{B_a}{C_a + B_a}.$$

Finally, assume that both $C_a > 0$ and $B_a > 0$. Note that under any realization of the randomness of the data sampling and the adversary, for any $a \in \{0, 1\}$

$$\begin{aligned}\gamma_a^p(h) &= \frac{\sum_{i=1}^n \mathbb{1}\{h(x_i^p) = 1, a_i^p = a\}}{\sum_{i=1}^n \mathbb{1}\{a_i^p = a\}} \\ &= \frac{\sum_{i=1}^n \mathbb{1}\{h(x_i^p) = 1, a_i^p = a, i \notin \mathfrak{M}\} + \sum_{i=1}^n \mathbb{1}\{h(x_i^p) = 1, a_i^p = a, i \in \mathfrak{M}\}}{\sum_{i=1}^n \mathbb{1}\{a_i^p = a, i \notin \mathfrak{M}\} + \sum_{i=1}^n \mathbb{1}\{a_i^p = a, i \in \mathfrak{M}\}}\end{aligned}$$

$$= \frac{C_a^1 + B_a^1}{C_a + B_a}.$$

Therefore,

$$\Delta_a = |\gamma_a^p - \gamma_a^c| = \left| \frac{C_a^1 + B_a^1}{C_a + B_a} - \frac{C_a^1}{C_a} \right| = \frac{B_a}{C_a + B_a} \left| \frac{C_a^1}{C_a} - \frac{B_a^1}{B_a} \right| \leq \frac{B_a}{C_a + B_a}$$

and so (27) holds in all cases. Therefore, we can bound the sum $\Delta_0 + \Delta_1$ as follows:

$$\begin{aligned}
\Delta_0 + \Delta_1 &\leq \frac{B_0}{C_0 + B_0} + \frac{B_1}{C_1 + B_1} \\
&< \frac{B_0}{\frac{1-\alpha}{2} P_0 n + B_0} + \frac{B_1}{\frac{1-\alpha}{2} P_1 n + B_1} \\
&\leq \frac{B_0}{\frac{1-\alpha}{2} P_0 n + B_0} + \frac{B_1}{\frac{1-\alpha}{2} P_0 n + B_1} \\
&= \frac{B_0}{\frac{1-\alpha}{2} P_0 n + B_0} + 1 - \frac{\frac{1-\alpha}{2} P_0 n}{\frac{1-\alpha}{2} P_0 n - B_0 + (B_0 + B_1)} \\
&< \frac{B_0}{\frac{1-\alpha}{2} P_0 n + B_0} + 1 - \frac{\frac{1-\alpha}{2} P_0 n}{\frac{1-\alpha}{2} P_0 n - B_0 + \frac{3\alpha}{2} n} \\
&= 2 - (1-\alpha) P_0 n \left(\frac{1}{(1-\alpha) P_0 n + 2 B_0} + \frac{1}{(1-\alpha) P_0 n + 3\alpha n - 2 B_0} \right)
\end{aligned}$$

Studying the function $f(x) = \frac{1}{(1-\alpha)P_0 n + 2x} + \frac{1}{(1-\alpha)P_0 n + 3\alpha n - 2x}$, we see that

$$f'(x) = 2 \left(\frac{1}{((1-\alpha) P_0 n + 3\alpha n - 2x)^2} - \frac{1}{((1-\alpha) P_0 n + 2x)^2} \right).$$

Note that $B_0 \leq B_0 + B_1 < \frac{3\alpha}{2}$, so we may assume $0 \leq x < \frac{3\alpha}{2}$. Therefore, both $(1-\alpha)P_0 n + 3\alpha n - 2x > 0$ and $(1-\alpha)P_0 n + 2x > 0$. Therefore, $f'(x) = 0$ if and only if $(1-\alpha)P_0 n + 3\alpha n - 2x = (1-\alpha)P_0 n + 2x$, that is, $x = \frac{3\alpha}{4} n$. Moreover $f'(x) < 0$ if $x \in [0, \frac{3\alpha}{4} n)$ and $f'(x) > 0$ if $x \in (\frac{3\alpha}{4} n, \frac{3\alpha}{2} n)$. Therefore, $f(x)$ is minimized at $x = \frac{3\alpha}{4} n$ and so

$$\begin{aligned}
\Delta_0 + \Delta_1 &\leq 2 - (1-\alpha) P_0 n \left(\frac{1}{(1-\alpha) P_0 n + 2 B_0} + \frac{1}{(1-\alpha) P_0 n + 3\alpha n - 2 B_0} \right) \\
&\leq 2 - (1-\alpha) P_0 n \left(\frac{1}{(1-\alpha) P_0 n + \frac{3\alpha}{2} n} + \frac{1}{(1-\alpha) P_0 n + 3\alpha n - \frac{3\alpha}{2} n} \right) \\
&= \frac{6\alpha}{2(1-\alpha) P_0 + 3\alpha} \\
&\leq \frac{6\alpha}{P_0 + 3\alpha} = \frac{2\alpha}{P_0/3 + \alpha}
\end{aligned}$$

and hence (27) holds in this case as well. Since the derivations hold for any classifier $h \in \mathcal{H}$, the result follows. ■

For the rest of the section, we keep the notation $\Delta_a(h) = |\gamma_a^p(h) - \gamma_a^c(h)|$ for $a \in \{0,1\}$ and $\Delta^{par} = \frac{2\alpha}{P_0/3+\alpha}$.

Next we use the previous result and the technique of (Woodworth et al., 2017) for proving concentration results about conditional probability estimates to bound the probability of a large deviation of $\widehat{\mathcal{D}}^{par}(h)$ from $\mathcal{D}^{par}(h)$, for a fixed hypothesis $h \in \mathcal{H}$.

Lemma 2 *Let $h \in \mathcal{H}$ be a fixed hypothesis and $\mathbb{P} \in \mathcal{P}(\mathcal{X} \times A \times \mathcal{Y})$ be a fixed distribution. Denote $P_a = \mathbb{P}(A = a)$ for $a \in \{0,1\}$. Let $\mathcal{A}$ be any malicious adversary and denote by $\mathbb{P}^{\mathcal{A}}$ the probability distribution of the poisoned data S^p, under the random sampling of the clean data, the marked points and the randomness of the adversary. Then for any $n \geq \max\left\{\frac{8\log(8/\delta)}{(1-\alpha)P_0}, \frac{12\log(6/\delta)}{\alpha}\right\}$ and $\delta \in (0,1)$*

$$\mathbb{P}^{\mathcal{A}}\left(\left|\widehat{\mathcal{D}}^{par}(h) - \mathcal{D}^{par}(h)\right| \leq \Delta^{par} + 2\sqrt{\frac{\log(16/\delta)}{n(1-\alpha)P_0}}\right) \geq 1 - \delta. \tag{28}$$

Proof Again we write $\gamma_a^p = \gamma_a^p(h), C_a^1 = C_a^1(h)$, etc. since h is fixed. First we study the concentration of the clean estimate $\frac{C_a^1}{C_a}$ around γ_a. To this end, denote by $S_a^c = \{i : a_i^p = a, i \notin \mathfrak{M}\}$ the set of indexes of the poisoned data for which the protected group is a and the corresponding point was not marked for the adversary. Notice that S_a^c is a random variable and that $|S_a^c| = C_a$. Since $n \geq \frac{8\log(8/\delta)}{(1-\alpha)P_a}$ for both $a \in \{0,1\}$, we have

$$\begin{aligned}
\mathbb{P}^{\mathcal{A}}\left(|\gamma_a^c - \gamma_a| > t\right) &= \sum_{S_a^c} \mathbb{P}^{\mathcal{A}}\left(|\gamma_a^c - \gamma_a| > t | S_a^c\right) \mathbb{P}(S_a^c) \\
&\leq \mathbb{P}^{\mathcal{A}}\left(C_a \leq \frac{(1-\alpha)}{2} P_a n\right) \\
&+ \sum_{S_a^c : C_a > \frac{(1-\alpha)}{2} P_a n} \mathbb{P}^{\mathcal{A}}\left(\left|\frac{C_a^1}{C_a} - \gamma_a\right| > t \middle| S_a^c\right) \mathbb{P}^{\mathcal{A}}(S_a^c) \\
&\leq \exp\left(-\frac{(1-\alpha)P_a n}{8}\right) + \sum_{S_a^p : C_a > \frac{(1-\alpha)}{2} P_a n} 2\exp\left(-2t^2 C_a\right) \mathbb{P}^{\mathcal{A}}(S_a^c) \\
&\leq \frac{\delta}{8} + 2\exp\left(-t^2(1-\alpha)P_a n\right),
\end{aligned}$$

where the second inequality follows from Hoeffding's inequality. Note that this step crucially uses that the marked indexes are independent of the data. The triangle law gives

$$\begin{aligned}
||\gamma_0^p - \gamma_1^p| - |\gamma_0 - \gamma_1|| \leq |\gamma_0^p - \gamma_1^p - \gamma_0 + \gamma_1| &\leq |\gamma_0^p - \gamma_0| + |\gamma_1^p - \gamma_1| \\
&\leq |\gamma_0^p - \gamma_0^c| + |\gamma_0^c - \gamma_0| + |\gamma_1^p - \gamma_1^c| + |\gamma_1^c - \gamma_1| \\
&= |\gamma_0^c - \gamma_0| + |\gamma_1^c - \gamma_1| + \Delta_0 + \Delta_1.
\end{aligned}$$

Combining the previous two results (recall that we assume $P_0 \leq P_1$)

$$\mathbb{P}^{\mathcal{A}}(||\gamma_0^p - \gamma_1^p| - |\gamma_0 - \gamma_1|| > 2t + \Delta_0 + \Delta_1)$$

$$
\begin{aligned}
&\leq \mathbb{P}^{\mathcal{A}}\left(|\gamma_0^c - \gamma_0| + |\gamma_1^c - \gamma_1| + \Delta_0 + \Delta_1 > 2t + \Delta_0 + \Delta_1\right) \\
&\leq \mathbb{P}^{\mathcal{A}}\left((|\gamma_0^c - \gamma_0| > t) \vee (|\gamma_1^c - \gamma_1| > t)\right) \\
&\leq \mathbb{P}^{\mathcal{A}}\left(|\gamma_0^c - \gamma_0| > t\right) + \mathbb{P}^{\mathcal{A}}\left(|\gamma_1^c - \gamma_1| > t\right) \\
&\leq \frac{\delta}{4} + 4\exp(-t^2 n(1-\alpha)P_0).
\end{aligned}
$$

Setting $t = t_0 = \sqrt{\frac{\log(16/\delta)}{n(1-\alpha)P_0}}$ gives

$$
\mathbb{P}^{\mathcal{A}}\left(||\gamma_0^p - \gamma_1^p| - |\gamma_0 - \gamma_1|| > \Delta_0 + \Delta_1 + 2\sqrt{\frac{\log(16/\delta)}{n(1-\alpha)P_0}}\right) \leq \frac{\delta}{4} + 4\frac{\delta}{16} = \frac{\delta}{2}. \tag{29}
$$

In addition Lemma 1 gives

$$
\mathbb{P}^{\mathcal{A}}\left(\Delta_0 + \Delta_1 > \Delta^{par}\right) \leq \frac{\delta}{2}. \tag{30}
$$

Using (29) and (30) we obtain that:

$$
\begin{aligned}
&\mathbb{P}^{\mathcal{A}}\left(||\gamma_0^p - \gamma_1^p| - |\gamma_0 - \gamma_1|| \leq \Delta^{par} + 2\sqrt{\frac{\log(16/\delta)}{N(1-\alpha)P_0}}\right) \\
&\geq \mathbb{P}^{\mathcal{A}}\left(\left(||\gamma_0^p - \gamma_1^p| - |\gamma_0 - \gamma_1|| \leq \Delta_0 + \Delta_1 + 2\sqrt{\frac{\log(16/\delta)}{N(1-\alpha)P_0}}\right) \wedge \left(\Delta_0 + \Delta_1 \leq \Delta^{par}\right)\right) \\
&\geq 1 - \frac{\delta}{2} - \frac{\delta}{2} = 1 - \delta.
\end{aligned}
$$

■

Finally, we show how to extend the previous result to hold uniformly over the whole hypothesis space, provided that $\mathcal{H}$ has a finite VC-dimension $d := VC(\mathcal{H})$

Lemma 3 *Under the setup of Lemma 2, assume additionally that $\mathcal{H}$ has a finite VC-dimension d. Then for any $n \geq \max\left\{\frac{8\log(8/\delta)}{(1-\alpha)P_0}, \frac{12\log(6/\delta)}{\alpha}, \frac{d}{2}\right\}$ and $\delta \in (0,1)$*

$$
\mathbb{P}^{\mathcal{A}}_{S^p}\left(\sup_{h\in\mathcal{H}} |\widehat{\mathcal{D}}^{par}(h) - \mathcal{D}^{par}(h)| \leq \Delta^{par} + 16\sqrt{\frac{2d\log(\frac{2en}{d}) + 2\log(48/\delta)}{(1-\alpha)P_0 n}}\right) \geq 1 - \delta. \tag{31}
$$

Proof From Lemma 1, we have that whenever $n \geq \max\left\{\frac{8\log(8/\delta)}{(1-\alpha)P_0}, \frac{12\log(6/\delta)}{\alpha}\right\}$ and $\delta \in (0,1)$

$$
\mathbb{P}^{\mathcal{A}}\left(\sup_{h\in\mathcal{H}} \left(\Delta_0(h) + \Delta_1(h)\right) \geq \Delta^{par}\right) < \frac{\delta}{2}. \tag{32}
$$

Additionally, in the proof of Lemma 2 we showed that for a fixed classifier $h \in \mathcal{H}$ for any $\delta \in (0,1), t \in (0,1)$ and both $a \in \{0,1\}$, we have

$$
\mathbb{P}^{\mathcal{A}}\left(|\gamma_a^c(h) - \gamma_a(h)| > t\right) \leq \exp\left(-\frac{(1-\alpha)P_a n}{8}\right) + 2\exp\left(-t^2(1-\alpha)P_a n\right)
$$

$$\leq 3\exp\left(-\frac{t^2(1-\alpha)P_a n}{8}\right). \tag{33}$$

The proof consists of two steps. In Steps 1 and 2 we show how to extend inequality (33) to hold uniformly over $\mathcal{H}$. Then, we combine the two uniform bounds with a similar argument as in the proof of Lemma 2.

The first step uses the classic symmetrization technique (Vapnik, 2013) for proving bounds uniformly over hypothesis spaces of finite VC dimension. However, since the objective is different from the 0-1 loss, care is needed to ensure that the proof goes through, so we present it here in full detail.

Step 1 To make the dependence of the left-hand side of (33) on both h and the data S^p explicit, we set $\gamma_a^c(h, S^p) := \frac{C_a^1(h)}{C_a}$.

Introduce a ghost sample $S^1 = \{(x_i^1, a_i^1, y_i^1)\}_{i=1}^n$ also sampled in an i.i.d. manner from $\mathbb{P}^{\mathcal{A}}$, that is, S^1 is another, independent poisoned dataset [7]. Let $\gamma_a^c(h, S^1)$ be the empirical estimate of $\gamma_a(h)$ based on S^1.

First we show a symmetrization inequality for the γ_a measures

$$\mathbb{P}^{\mathcal{A}}_{S^p}\left(\sup_{h\in\mathcal{H}}|\gamma_a(h) - \gamma_a^c(h, S^p)| \geq t\right) \leq 2\mathbb{P}^{\mathcal{A}}_{S^p, S^1}\left(\sup_{h\in\mathcal{H}}\left|\gamma_a^c(h, S^1) - \gamma_a^c(h, S^p)\right| \geq t/2\right), \tag{34}$$

for any constant $1 > t \geq 2\sqrt{\frac{8\log(6)}{(1-\alpha)P_0 n}}$.

Indeed, let h^* be the hypothesis achieving the supremum on the left-hand side [8]. Note that

$$\begin{aligned}\mathbb{1}(|\gamma_a(h^*) - \gamma_a^c(h^*, S^p)| \geq t)&\mathbb{1}(\left|\gamma_a(h^*) - \gamma_a^c(h^*, S^1)\right| \leq t/2)\\ &\leq \mathbb{1}(\left|\gamma_a^c(h^*, S^1) - \gamma_a^c(h^*, S^p)\right| \geq t/2).\end{aligned}$$

Taking expectation with respect to S^1

$$\begin{aligned}\mathbb{1}(|\gamma_a(h^*) - \gamma_a^c(h^*, S^p)| \geq t)&\mathbb{P}^{\mathcal{A}}_{S^1}(\left|\gamma_a(h^*) - \gamma_a^c(h^*, S^1)\right| \leq t/2)\\ &\leq \mathbb{P}^{\mathcal{A}}_{S^1}(\left|\gamma_a^c(h^*, S^1) - \gamma_a^c(h^*, S^p)\right| \geq t/2).\end{aligned}$$

Now using Lemma 2

$$\begin{aligned}\mathbb{P}^{\mathcal{A}}_{S^1}\left(\left|\gamma_a(h^*) - \gamma_a^c(h^*, S^1)\right| \leq t/2\right) &\geq \mathbb{P}^{\mathcal{A}}_{S^1}\left(\left|\gamma_a(h^*) - \gamma_a^c(h^*, S^1)\right| \leq \sqrt{\frac{8\log(6)}{(1-\alpha)P_a n}}\right)\\ &\geq 1 - \frac{1}{2}\\ &= \frac{1}{2}.\end{aligned}$$

so

$$\frac{1}{2}\mathbb{1}(|\gamma_a(h^*) - \gamma_a^c(h^*, S^p)| \geq t) \leq \mathbb{P}^{\mathcal{A}}_{S^1}(\left|\gamma_a^c(h^*, S^1) - \gamma_a^c(h^*, S^p)\right| \geq t/2).$$

7. Formally, we associate S^1 also with a set $\mathfrak{M}_1$ of marked indexes.
8. If the supremum is not attained, this argument can be repeated for each element of a sequence of classifiers approaching the supremum

Taking expectation with respect to S^p

$$\begin{aligned}\mathbb{P}^{\mathcal{A}}_{S^p}(|\gamma_a(h^*) - \gamma^c_a(h^*, S^p)| \geq t) &\leq 2\mathbb{P}^{\mathcal{A}}_{S^p,S^1}(\left|\gamma^c_a(h^*, S^1) - \gamma^c_a(h^*, S^p)\right| \geq t/2) \\ &\leq 2\mathbb{P}^{\mathcal{A}}_{S^p,S^1}(\sup_{h\in\mathcal{H}}\left|\gamma^c_a(h, S^1) - \gamma^c_a(h, S^p)\right| \geq t/2).\end{aligned}$$

Step 2 Next we use the growth function of $\mathcal{H}$ and the symmetrization inequality (34) to bound the large deviations of $\gamma^c_a(h)$ uniformly over $\mathcal{H}$.

Specifically, gives n points $x_1, \ldots, x_n \in \mathcal{X}$, denote

$$\mathcal{H}_{x_1,\ldots,x_n}\{(h(x_1), \ldots, h(x_n)) : h \in \mathcal{H}\}.$$

Then define the growth function of $\mathcal{H}$ as

$$S_{\mathcal{H}}(n) = \sup_{x_1,\ldots,x_n} |\mathcal{H}_{x_1,\ldots,x_n}|.$$

We will use that well-known Sauer's lemma (see, for example, (Bousquet et al., 2003)), which states that whenever $n \geq d$, $S_{\mathcal{H}}(n) \leq \left(\frac{en}{d}\right)^d$

Notice that given the two datasets S^p, S^1 and the corresponding sets of marked indexes, the values of $\gamma^c_a(h, S^p)$ and $\gamma^c_a(h, S^1)$ depend only on the values of h on S^p and S^1 respectively. Therefore for any $1 > t \geq 2\sqrt{\frac{8\log(6)}{(1-\alpha)P_0 n}}$,

$$\begin{aligned}&\mathbb{P}^{\mathcal{A}}_{S^p}\left(\sup_{h\in\mathcal{H}}|\gamma_a(h) - \gamma^c_a(h, S^p)| \geq t\right) \\ &\quad\leq 2\mathbb{P}^{\mathcal{A}}_{S^p,S^1}\left(\sup_{h\in\mathcal{H}}\left|\gamma^c_a(h, S^1) - \gamma^c_a(h, S^p)\right| \geq t/2\right) \\ &\quad\leq 2S_{\mathcal{H}}(2n)\mathbb{P}^{\mathcal{A}}_{S^p,S^1}\left(\left|\gamma^c_a(h, S^1) - \gamma^c_a(h, S^p)\right| \geq t/2\right) \\ &\quad\leq 2S_{\mathcal{H}}(2n)\mathbb{P}^{\mathcal{A}}_{S^p,S^1}\left(\left|\gamma^c_a(h, S^1) - \gamma^c_a(h)\right| \geq t/4 \vee |\gamma^c_a(h, S^p) - \gamma^c_a(h)| \geq t/4\right) \\ &\quad\leq 4S_{\mathcal{H}}(2n)\mathbb{P}^{\mathcal{A}}_{S^p}\left(|\gamma^c_a(h, S^p) - \gamma^c_a(h)| \geq t/4\right) \\ &\quad\leq 12S_{\mathcal{H}}(2n)\exp\left(-\frac{t^2(1-\alpha)P_a n}{128}\right).\end{aligned}$$

Using $P_0 \leq P_1$ and Sauer's lemma, whenever $2n \geq d$ we have

$$\mathbb{P}^{\mathcal{A}}_{S^p}\left(\sup_{h\in\mathcal{H}}|\gamma_a(h) - \gamma^c_a(h, S^p)| \geq t\right) \leq 12\left(\frac{2en}{d}\right)^d \exp\left(-\frac{t^2(1-\alpha)P_0 n}{128}\right).$$

Using inversion, we get that

$$\mathbb{P}^{\mathcal{A}}_{S^p}\left(\sup_{h\in\mathcal{H}}|\gamma_a(h) - \gamma^c_a(h, S^p)| \geq 8\sqrt{\frac{2d\log(\frac{2en}{d}) + 2\log(48/\delta)}{(1-\alpha)P_0 n}}\right) \leq \frac{\delta}{4}, \tag{35}$$

whenever

$$1 > 8\sqrt{2\frac{d\log(\frac{2en}{d}) + 2\log(12/\delta)}{(1-\alpha)P_0 n}} \geq 2\sqrt{\frac{8\log(6)}{(1-\alpha)P_0 n}}.$$

It's easy to see that the right inequality holds whenever $\delta < 1$ and $2n \geq d$. In addition, inequality (35) trivially holds if the left inequality is not fulfilled. Therefore, (35) holds whenever $2n \geq d$.

Step 3 Finally, we use (32) and (35) to proof the lemma. Recall from the proof of Lemma 2 that

$$\begin{aligned}|\widehat{\mathcal{D}}^{par}(h) - \mathcal{D}^{par}(h)| &= ||\gamma_0^c(h, S^p) - \gamma_1^c(h, S^p)| - |\gamma_0(h) - \gamma_1(h)|| \\ &\leq |\gamma_0^c(h, S^p) - \gamma_0(h)| + |\gamma_1^c(h, S^p) - \gamma_1(h)| + \Delta_0(h) + \Delta_1(h).\end{aligned}$$

Therefore,

$$\begin{aligned}\sup_{h\in\mathcal{H}} |\widehat{\mathcal{D}}^{par}(h) - \mathcal{D}^{par}(h)| \leq \sup_{h\in\mathcal{H}} |\gamma_0^c(h, S^p) - \gamma_0(h)| + \sup_{h\in\mathcal{H}} |\gamma_1^c(h, S^p) - \gamma_1(h)| \\ + \sup_{h\in\mathcal{H}} (\Delta_0(h) + \Delta_1(h)).\end{aligned}$$

Now, using the union bound and inequalities (32) and (35), whenever

$$n \geq \max\left\{\frac{8\log(8/\delta)}{(1-\alpha)P_0}, \frac{12\log(6/\delta)}{\alpha}, \frac{d}{2}\right\}$$

we get

$$\begin{aligned}&\mathbb{P}_{S^p}^{\mathcal{A}}\left(\sup_{h\in\mathcal{H}} |\widehat{\mathcal{D}}^{par}(h) - \mathcal{D}^{par}(h)| \geq \Delta^{par} + 16\sqrt{\frac{2d\log(\frac{2en}{d}) + 2\log(48/\delta)}{(1-\alpha)P_0 n}}\right) \\ &\leq \mathbb{P}_{S^p}^{\mathcal{A}}\left(\sup_{h\in\mathcal{H}} |\gamma_0(h) - \gamma_0^c(h, S^p)| \geq 8\sqrt{\frac{2d\log(\frac{2en}{d}) + 2\log(48/\delta)}{(1-\alpha)P_0 n}}\right) \\ &+ \mathbb{P}_{S^p}^{\mathcal{A}}\left(\sup_{h\in\mathcal{H}} |\gamma_1(h) - \gamma_1^c(h, S^p)| \geq 8\sqrt{\frac{2d\log(\frac{2en}{d}) + 2\log(48/\delta)}{(1-\alpha)P_0 n}}\right) \\ &+ \mathbb{P}^{\mathcal{A}}\left(\sup_{h\in\mathcal{H}} (\Delta_0(h) + \Delta_1(h)) \geq \Delta^{par}\right) \\ &\leq \frac{\delta}{4} + \frac{\delta}{4} + \frac{\delta}{2} = \delta\end{aligned}$$

■

B.2.2 Concentration for equal opportunity

We introduce similar notation as in Section B.2.1, but tailored to the equal opportunity conditional probabilities.

We use the notation $C_{1a} = \sum_{i=1}^n \mathbb{1}\{i : a_i^p = a, y_i^p = 1, i \notin \mathfrak{M}\}|$ for the number of points in S^p that *were not* marked (are *clean*) and contain a point from protected group a and label $y = 1$ and $B_{1a} = \sum_{i=1}^n \mathbb{1}\{i : a_i^p = a, y_i^p = 1, i \in \mathfrak{M}\}|$ for the number of points in S^p that *were* marked (are potentially *bad*) and contain a point from protected group a and label $y = 1$. Note that $B_{10} + B_{11}$ is the total number of poisoned points for which $y = 1$ and so is at most $\text{Bin}(n, \alpha)$. Similarly, denote by $C_{1a}^1(h) = \sum_{i=1}^n \mathbb{1}\{i : h(x_i^p) = 1, a_i^p = a, y_1^p = 1, i \notin \mathfrak{M}\}|$ and $B_{1a}^1(h) = \sum_{i=1}^n \mathbb{1}\{i : h(x_i^p) = 1, a_i^p = a, y_i^p = 1, i \in \mathfrak{M}\}|$.

Denote

$$\gamma^p_{1a}(h) = \frac{\sum_{i=1}^n \mathbb{1}\{h(x_i^p) = 1, a_i^p = a, y_1^p = 1\}}{\sum_{i=1}^n \mathbb{1}\{a_i^p = a, y_i^p = 1\}}$$

and

$$\gamma_{1a}(h) = \mathbb{P}(h(X) = 1 | A = a, Y = 1),$$

so that $\widehat{\mathcal{D}}^{opp}(h) = |\gamma^p_{10}(h) - \gamma^p_{11}(h)|$ and $\mathcal{D}^{opp}(h) = |\gamma_{10}(h) - \gamma_{11}(h)|$. Note that $\gamma^p_{1a}(h)$ is an estimate of a conditional probability *based on the corrupted data*. We now introduce the corresponding estimate that only uses the clean (but unknown) subset of the training set S^p:

$$\gamma^c_a(h) = \frac{C^1_a(h)}{C_a(h)} = \frac{\sum_{i=1}^n \mathbb{1}\{h(x_i^p) = 1, a_i^p = a, y_i^p = 1, i \notin \mathfrak{M}\}}{\sum_{i=1}^n \mathbb{1}\{a_i^p = a, y_i^p = 1, i \notin \mathfrak{M}\}}.$$

Similarly to before, we first bound how far the corrupted estimates $\gamma^p_{1a}(h)$ of $\gamma_{1a}(h)$ are from the clean estimates $\gamma^c_{1a}(h)$, uniformly over the hypothesis space $\mathcal{H}$:

Lemma 4 *If* $n \geq \max\left\{\frac{8\log(4/\delta)}{(1-\alpha)P_0}, \frac{12\log(3/\delta)}{\alpha}\right\}$, *we have*

$$\mathbb{P}^{\mathcal{A}}\left(\sup_{h\in\mathcal{H}} (|\gamma^p_{10}(h) - \gamma^c_{10}(h)| + |\gamma^p_{11}(h) - \gamma^c_{11}(h)|) \geq \frac{2\alpha}{P_{10}/3 + \alpha}\right) < \delta. \tag{36}$$

Proof Similarly to the proof of Lemma 1, we first show that certain bounds on B_{1a} and C_{1a} hold with high probability. Then we show that the supremum in (36) is bounded whenever these bounds hold.

Step 1 Note that $B_{10} + B_{11} \leq B_0 + B_1 \sim \mathrm{Bin}(n, \alpha)$, and so

$$\mathbb{P}^{\mathcal{A}}\left(B_{10} + B_{11} \geq \frac{3\alpha}{2}n\right) \leq \mathbb{P}^{\mathcal{A}}\left(B_0 + B_1 \geq \frac{3\alpha}{2}n\right) \leq e^{-\alpha n/12} \leq \frac{\delta}{3}.$$

Similarly, $C_{10} \sim \mathrm{Bin}(n, (1-\alpha)P_{1a})$ and $C_{11} \sim \mathrm{Bin}(n, (1-\alpha)P_{11})$ and so

$$\mathbb{P}^{\mathcal{A}}\left(C_{10} \leq \frac{1-\alpha}{2}P_{10}n\right) \leq e^{-(1-\alpha)P_{10}n/8} \leq \frac{\delta}{4}$$

and

$$\mathbb{P}^{\mathcal{A}}\left(C_{11} \leq \frac{1-\alpha}{2}P_{11}n\right) \leq e^{-(1-\alpha)P_{11}n/8} \leq \frac{\delta}{4}.$$

Now since $n \geq \max\left\{\frac{8\log(4/\delta)}{(1-\alpha)P_{10}}, \frac{12\log(3/\delta)}{\alpha}\right\}$ and $P_{10} \leq P_{11}$

$$\begin{aligned}\mathbb{P}^{\mathcal{A}}\left(\left(B_{10} + B_{11} \geq \frac{3\alpha}{2}n\right) \vee \left(C_{10} \leq \frac{1-\alpha}{2}P_{10}n\right) \vee \left(C_{11} \leq \frac{1-\alpha}{2}P_{11}n\right)\right) &\leq \frac{\delta}{3} + \frac{\delta}{4} + \frac{\delta}{4} \\ &< \delta, \end{aligned} \tag{37}$$

Step 2 Now assume that all of $B_{10} + B_{11} < \frac{3\alpha}{2}n, C_{10} > \frac{1-\alpha}{2}P_{10}n, C_{11} > \frac{1-\alpha}{2}P_{11}n$ hold.

Consider an arbitrary, fixed $h \in \mathcal{H}$. Since h is fixed, we drop the dependence on h from the notation for the rest of the proof and write $\gamma^p_{1a} = \gamma^p_{1a}(h), C^1_{1a} = C^1_{1a}(h)$ etc.

We now prove that for both $a \in \{0, 1\}$

$$\Delta_{1a} \coloneqq |\gamma^p_a - \gamma^c_a| \leq \frac{B_{1a}}{C_{1a} + B_{1a}}. \tag{38}$$

For each $a \in \{0, 1\}$, this can be shown as follows. First, if $\sum_{i=1}^n \mathbb{1}\{a^p_i = a, y^p_i = 1\} = B_{1a} + C_{1a} = 0$, then both $\gamma^p_{1a}(h)$ and $\gamma^c_{1a}(h)$ are equal to 0, because of the convention that $\frac{0}{0} = 0$. In addition, $B_{1a} = C_{1a} = 0$. Therefore, inequality (38) trivially holds.

Similarly, if $B_{1a} = 0$ and $C_{1a} > 0$, then $\gamma^p_{1a}(h) = \gamma^c_{1a}(h)$ and so $\Delta_{1a} = 0$ and (38) holds.

Assume now that $B_{1a} > 0$. Note that if $C_{1a} = \sum_{i=1}^n \mathbb{1}\{a^p_i = a, y^p_i = 1, i \notin \mathfrak{M}\} = 0$, then

$$\Delta_{1a} = |\gamma^p_{1a}(h) - \gamma^c_{1a}(h)| = \left|\frac{B^1_{1a}}{B_{1a}} - 0\right| = \frac{B^1_{1a}}{B_{1a}} = \frac{B^1_{1a}}{B_{1a} + C_{1a}} \leq \frac{B_{1a}}{C_{1a} + B_{1a}}.$$

Finally, assume that both $C_{1a} > 0$ and $B_{1a} > 0$. Note that under any realization of the randomness of the data sampling and the adversary, for any $a \in \{0, 1\}$

$$\begin{aligned}\gamma^p_{1a} &= \frac{\sum_{i=1}^n \mathbb{1}\{h(x^p_i) = 1, a^p_i = a, y^p_i = 1, i \notin \mathfrak{M}\} + \sum_{i=1}^n \mathbb{1}\{h(x^p_i) = 1, a^p_i = a, y^p_i = 1, i \in \mathfrak{M}\}}{\sum_{i=1}^n \mathbb{1}\{a^p_i = a, y^p_i = 1, i \notin \mathfrak{M}\} + \sum_{i=1}^n \mathbb{1}\{a^p_i = a, y^p_i = 1, i \in \mathfrak{M}\}} \\ &= \frac{C^1_{1a} + B^1_{1a}}{C_{1a} + B_{1a}}.\end{aligned}$$

Next we bound how far this quantity is from the clean estimator $\frac{C^1_{1a}}{C_{1a}}$

$$\Delta_{1a} \coloneqq |\gamma^p_{1a} - \gamma^c_{1a}| = \left|\frac{C^1_{1a} + B^1_{1a}}{C_{1a} + B_{1a}} - \gamma^c_{1a}\right| = \frac{B_{1a}}{C_{1a} + B_{1a}}\left|\gamma^c_{1a} - \frac{B^1_{1a}}{B_{1a}}\right| \leq \frac{B_{1a}}{C_{1a} + B_{1a}}.$$

Now since $B_{10} + B_{11} < \frac{3\alpha}{2}n, C_{10} > \frac{1-\alpha}{2}P_{10}n, C_{11} > \frac{1-\alpha}{2}P_{11}n$ hold, we get

$$\begin{aligned}\Delta_{10} + \Delta_{11} &\leq \frac{B_{10}}{C_{10} + B_{10}} + \frac{B_{11}}{C_{11} + B_{11}} \\ &< \frac{B_{10}}{\frac{1-\alpha}{2}P_{10}n + B_{10}} + \frac{B_{11}}{\frac{1-\alpha}{2}P_{11}n + B_{11}} \\ &\leq \frac{B_{10}}{\frac{1-\alpha}{2}P_{10}n + B_{10}} + \frac{B_{11}}{\frac{1-\alpha}{2}P_{10}n + B_{11}} \\ &= \frac{B_{10}}{\frac{1-\alpha}{2}P_{10}n + B_{10}} + 1 - \frac{\frac{1-\alpha}{2}P_{10}n}{\frac{1-\alpha}{2}P_{10}n - B_{10} + (B_{10} + B_{11})} \\ &< \frac{B_{10}}{\frac{1-\alpha}{2}P_{10}n + B_{10}} + 1 - \frac{\frac{1-\alpha}{2}P_{10}n}{\frac{1-\alpha}{2}P_{10}n - B_{10} + \frac{3\alpha}{2}n} \\ &= 2 - (1-\alpha)P_{10}n\left(\frac{1}{(1-\alpha)P_{10}n + 2B_{10}} + \frac{1}{(1-\alpha)P_{10}n + 3\alpha n - 2B_{10}}\right)\end{aligned}$$

The same argument as in Lemma 1 shows that this is maximized at $B_{10} = \frac{3\alpha}{4}n$ and so

$$\begin{aligned}\Delta_{10} + \Delta_{11} &\leq \frac{B_{10}}{C_{10} + B_{10}} + \frac{B_{11}}{C_{11} + B_{11}} \\ &< 2 - (1-\alpha)P_{10}n\left(\frac{1}{(1-\alpha)P_{10}n + \frac{3\alpha}{2}n} + \frac{1}{(1-\alpha)P_{10}n + 3\alpha n - \frac{3\alpha}{2}n}\right) \\ &\leq \frac{2\alpha}{P_{10}/3 + \alpha}.\end{aligned} \tag{39}$$

Since this holds for any arbitrary hypothesis $h \in \mathcal{H}$, the result follows. ■

Denote the irreducible error term for equal opportunity by $\Delta^{opp} = \frac{2\alpha}{P_{10}/3+\alpha}$. We then have the following bound for a fixed $h \in \mathcal{H}$:

Lemma 5 *Let $h \in \mathcal{H}$ be a fixed hypothesis and $D \in \mathcal{P}(\mathcal{X} \times A \times \mathcal{Y})$ be a fixed distribution. Denote $P_{1a} = \mathbb{P}(A = a, Y = 1)$ for $a \in \{0, 1\}$. Let $\mathcal{A}$ be any malicious adversary and denote by $\mathbb{P}^{\mathcal{A}}$ the probability distribution of the poisoned data S^p, under the random sampling of the clean data, the marked points and the randomness of the adversary. Then for any $n \geq \max\left\{\frac{8\log(8/\delta)}{(1-\alpha)P_{10}}, \frac{12\log(6/\delta)}{\alpha}\right\}$ and $\delta \in (0, 1)$*

$$\mathbb{P}^{\mathcal{A}}\left(\left|\widehat{\mathcal{D}}^{opp}(h) - \mathcal{D}^{opp}(h)\right| \leq \Delta^{opp} + 2\sqrt{\frac{\log(16/\delta)}{n(1-\alpha)P_{10}}}\right) \geq 1 - \delta \tag{40}$$

Proof The proof is exactly the same as the one of Lemma 2 , but with conditioning on $S^c_{1a} = \{i : a^p_i = a, y^p_i = 1, i \notin \mathfrak{M}\}$ (the set of indexes of the poisoned data for which the protected group is a, the label is 1 and the corresponding point was not marked for the adversary) instead. ■

The same argument as in Lemma 3 gives a uniform bound over the whole hypothesis space, provided that $\mathcal{H}$ has a finite VC-dimension $d := VC(\mathcal{H})$:

Lemma 6 *Under the setup of Lemma 5, assume additionally that $\mathcal{H}$ has a finite VC-dimension d. Then for any $n \geq \max\left\{\frac{8\log(8/\delta)}{(1-\alpha)P_{10}}, \frac{12\log(6/\delta)}{\alpha}, \frac{d}{2}\right\}$ and $\delta \in (0, 1)$*

$$\mathbb{P}^{\mathcal{A}}_{S^p}\left(\sup_{h\in\mathcal{H}} |\widehat{\mathcal{D}}^{opp}(h) - \mathcal{D}^{opp}(h)| \leq \Delta^{opp} + 16\sqrt{\frac{2d\log(\frac{2en}{d}) + 2\log(48/\delta)}{(1-\alpha)P_{10}n}}\right) \geq 1 - \delta. \tag{41}$$

Finally, we prove multiplicative bounds and claims in the case when $\mathbb{P}(h(X) = 1|A = 0, Y = 1) = \mathbb{P}(h(X) = 1|A = 1, Y = 1) = 1$ (which holds for example when $h(X) = Y$ almost surely). These will come in useful for proving the component-wise upper bound with fast rates.

We will be interested in the estimate

$$\bar{\gamma}^p_{1a}(h) = \frac{\sum_{i=1}^n \mathbb{1}\{h(x^p_i) = 0, a^p_i = a, y^p_1 = 1\}}{\sum_{i=1}^n \mathbb{1}\{a^p_i = a, y^p_i = 1\}}$$

of $\bar{\gamma}_{1a}(h) = \mathbb{P}(h(X) = 0|A = a, Y = 1)$. Again, we also introduce the corresponding clean data estimate $C^0_{1a}(h) \coloneqq \sum_{i=1}^n \mathbb{1}\{i : h(x_i^p) = 0, a_i^p = a, y_1^p = 1, i \notin \mathfrak{M}\}$ and

$$\bar{\gamma}^c_{1a}(h) = \frac{C^0_{1a}(h)}{C_{1a}} = \frac{\sum_{i=1}^n \mathbb{1}\{i : h(x_i^p) = 0, a_i^p = a, y_1^p = 1, i \notin \mathfrak{M}\}}{\sum_{i=1}^n \mathbb{1}\{i : a_i^p = a, y_1^p = 1, i \notin \mathfrak{M}\}}.$$

Denote also

$$\bar{\Delta}_{1a}(h) := \left|\bar{\gamma}^p_{1a}(h) - \gamma^c_{1a}(h)\right|,$$

We only show non-uniform bounds for a fixed $h \in \mathcal{H}$ here, so we omit the dependence of these quantities on h. We have:

Lemma 7 *Let $\mathbb{P} \in \mathcal{P}(\mathcal{X} \times A \times \mathcal{Y})$ be a fixed distribution and let $h \in \mathcal{H}$ be a fixed classifier. Denote $P_{1a} = \mathbb{P}(A = a, Y = 1)$ for $a \in \{0, 1\}$. Let $\mathcal{A}$ be any malicious adversary and denote by $\mathbb{P}^{\mathcal{A}}$ the probability distribution of the poisoned data S^p, under the random sampling of the clean data, the marked points and the randomness of the adversary. Then:*
(a) For any $n > 0$ and any $\eta, \delta \in (0, 1)$

$$\mathbb{P}^{\mathcal{A}}\left(\bar{\gamma}^p_{1a} \geq (1+\eta)\bar{\gamma}_{1a} + \bar{\Delta}_{1a}\right) \leq \exp\left(-\frac{(1-\alpha)P_{1a}n}{8}\right) + \exp\left(-\frac{1}{6}\eta^2(1-\alpha)P_{1a}\bar{\gamma}_{1a}n\right). \tag{42}$$

and

$$\mathbb{P}^{\mathcal{A}}\left(\bar{\gamma}^p_{1a} \leq (1-\eta)\bar{\gamma}_{1a} - \bar{\Delta}_{1a}\right) \leq \exp\left(-\frac{(1-\alpha)P_{1a}n}{8}\right) + \exp\left(-\frac{1}{4}\eta^2(1-\alpha)P_{1a}\bar{\gamma}_{1a}n\right). \tag{43}$$

(b) Assume further that $\mathbb{P}(h(X) = 0|A = 0, Y = 1) = \mathbb{P}(h(X) = 0|A = 1, Y = 1) = 0$. Then for any $\delta \in (0, 1)$ and $n \geq \max\left\{\frac{8\log(4/\delta)}{(1-\alpha)P_{10}}, \frac{12\log(3/\delta)}{\alpha}\right\}$

$$\mathbb{P}^{\mathcal{A}}\left(\bar{\gamma}^p_{10} + \bar{\gamma}^p_{11} \geq \Delta^{opp}\right) \leq \delta \tag{44}$$

Proof Let $S^c_{1a} = \{i : a_i^p = a, y_i^p = 1, i \notin \mathfrak{M}\}$. For any $a \in \{0, 1\}$ we have

$$\begin{aligned}
&\mathbb{P}^{\mathcal{A}}\left(\bar{\gamma}^p_{1a} \geq (1+\eta)\bar{\gamma}_{1a} + \bar{\Delta}_{1a}\right) \\
&= \sum_{S^c_{1a}} \mathbb{P}^{\mathcal{A}}\left(\bar{\gamma}^p_{1a} \geq (1+\eta)\bar{\gamma}_{1a} + \bar{\Delta}_{1a}\middle|S^c_a\right)\mathbb{P}(S^c_a) \\
&\leq \mathbb{P}^{\mathcal{A}}\left(C_{1a} \leq \frac{(1-\alpha)}{2}P_{1a}n\right) \\
&+ \sum_{S^c_{1a}: C_{1a} \geq \frac{(1-\alpha)}{2}P_{1a}n} \mathbb{P}^{\mathcal{A}}\left(\bar{\gamma}^p_{1a} \geq (1+\eta)\bar{\gamma}_{1a} + \bar{\Delta}_{1a}\middle|S^c_{1a}\right)\mathbb{P}^{\mathcal{A}}(S^c_{1a}) \\
&\leq \mathbb{P}^{\mathcal{A}}\left(C_{1a} \leq \frac{(1-\alpha)}{2}P_{1a}n\right) \\
&+ \sum_{S^c_{1a}: C_{1a} \geq \frac{(1-\alpha)}{2}P_{1a}n} \mathbb{P}^{\mathcal{A}}\left(\bar{\gamma}^p_{1a} - \frac{C^1_{1a}}{C_{1a}} + \frac{C^1_{1a}}{C_{1a}} \geq (1+\eta)\bar{\gamma}_{1a} + \bar{\Delta}_{1a}\middle|S^c_{1a}\right)\mathbb{P}^{\mathcal{A}}(S^c_{1a})
\end{aligned}$$

$$
\begin{aligned}
&\leq \mathbb{P}^{\mathcal{A}}\left(C_{1a} \leq \frac{(1-\alpha)}{2} P_{1a} n\right) \\
&+ \sum_{S^c_{1a}: C_{1a} \geq \frac{(1-\alpha)}{2} P_{1a} n} \mathbb{P}^{\mathcal{A}}\left(\frac{C^1_{1a}}{C_{1a}} \geq (1+\eta)\bar{\gamma}_{1a} \middle| S^c_{1a}\right) \mathbb{P}^{\mathcal{A}}(S^c_{1a}) \\
&\leq \exp\left(-\frac{(1-\alpha)P_{1a} n}{8}\right) + \sum_{S^p_a: C_{1a} \geq \frac{(1-\alpha)}{2} P_{1a} n} \exp\left(-\frac{\eta^2 C_{1a} \bar{\gamma}_{1a}}{3}\right) \mathbb{P}^{\mathcal{A}}(S^c_{1a}) \\
&\leq \exp\left(-\frac{(1-\alpha)P_{1a} n}{8}\right) + \exp\left(-\frac{1}{6}\eta^2 (1-\alpha) P_{1a} \bar{\gamma}_{1a} n\right).
\end{aligned}
$$

A similar argument, with the other direction of the Chernoff bounds, gives the other bound.

(b) Similarly to the argument in the proof of Lemma 4

$$
\bar{\Delta}_{1a} = \left|\bar{\gamma}^p_{1a} - \frac{C^0_{1a}}{C_{1a}}\right| \leq \frac{B_{1a}}{C_{1a} + B_{1a}}. \tag{45}
$$

Using the inequalities (37) and (39),

$$
\mathbb{P}^{\mathcal{A}}\left(\bar{\Delta}_{10} + \bar{\Delta}_{11} \geq \frac{2\alpha}{P_{10}/3 + \alpha}\right) \leq \mathbb{P}^{\mathcal{A}}\left(\frac{B_{10}}{C_{10} + B_{10}} + \frac{B_{11}}{C_{11} + B_{11}} \geq \frac{2\alpha}{P_{10}/3 + \alpha}\right) < \delta. \tag{46}
$$

Since also

$$
\mathbb{P}^{\mathcal{A}}\left(\frac{C^0_{1a}}{C_{1a}} > 0\right) = \sum_{S^c_{1a}} \mathbb{P}^{\mathcal{A}}\left(\frac{C^0_{1a}}{C_{1a}} > 0 \middle| S^c_{1a}\right) \mathbb{P}^{\mathcal{A}}\left(S^c_{1a}\right) = \sum_{S^c_{1a}} \mathbb{P}\left(\mathrm{Bin}(|S^c_{1a}|, 0) > 0\right) \mathbb{P}^{\mathcal{A}}(S^c_{1a}) = 0,
$$

we have that $0 \leq \bar{\gamma}^p_{1a} = \bar{\Delta}_{1a}$ almost surely, for both $a \in \{0, 1\}$. Therefore, $0 < \bar{\gamma}^p_{10} + \bar{\gamma}^p_{11} = \bar{\Delta}_{10} + \bar{\Delta}_{11}$ and the result follows. ■

B.3 Upper bound theorems - proofs

We are now ready to present the proofs of the upper bound results from the main body of the paper.

B.3.1 Upper bounds on the λ-weighted objective

First we prove the bounds for the λ-weighted objective.

Bound for demographic parity Let $\lambda \geq 0$ be fixed. Recall our notation for the λ-weighted objective:

$$
L^{par}_{\lambda}(h) = \mathcal{R}(h) + \lambda \mathcal{D}^{par}(h).
$$

Suppose that a learner $\mathcal{L}^{par}_{\lambda} : \cup_{n=1}^{\infty} (\mathcal{X} \times A \times \mathcal{Y})^n \to \mathcal{H}$ is such that

$$
\mathcal{L}^{par}(S^p) \in \operatorname*{argmin}_{h \in \mathcal{H}} (\widehat{R}^p(h) + \lambda \widehat{\mathcal{D}}^{par}(h)) \quad \text{for all } S^p.
$$

That is, $\mathcal{L}^{par}_{\lambda}$ always returns a minimizer of the λ-weighted empirical objective. Then we have the following:

Theorem 5 *Let $\mathcal{H}$ be any hypothesis space with $d = VC(\mathcal{H}) < \infty$. Let $\mathbb{P} \in \mathcal{P}(\mathcal{X} \times A \times \mathcal{Y})$ be a fixed distribution and $\mathcal{A}$ be any malicious adversary of power $\alpha < 0.5$. Denote by $\mathbb{P}^{\mathcal{A}}$ the probability distribution of the poisoned data S^p, under the random sampling of the clean data, the marked points and the randomness of the adversary. Then for any $\delta \in (0,1)$ and $n \geq \max\left\{\frac{8\log(16/\delta)}{(1-\alpha)P_0}, \frac{12\log(12/\delta)}{\alpha}, \frac{d}{2}\right\}$, we have*

$$\mathbb{P}^{\mathcal{A}}\left(L_\lambda^{par}(\mathcal{L}_\lambda^{par}(S^p)) \leq \min_{h\in\mathcal{H}} L_\lambda^{par}(h) + \Delta_\lambda^{par}\right) > 1-\delta,$$

where[9]

$$\Delta_\lambda^{par} = 3\alpha + 2\lambda\Delta^{par} + \widetilde{\mathcal{O}}\left(\sqrt{\frac{d}{n}} + \lambda\sqrt{\frac{d}{P_0 n}}\right)$$

and

$$\Delta^{par} = \frac{2\alpha}{P_0/3+\alpha} = \mathcal{O}\left(\frac{\alpha}{P_0}\right).$$

Proof By the standard concentrations results for the 0/1 loss (see, for example, Chapter 28.1 in (Shalev-Shwartz and Ben-David, 2014))

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}} |\widehat{\mathcal{R}}^c(h) - \mathcal{R}(h)| > 2\sqrt{\frac{8d\log(\frac{en}{d}) + 2\log(16/\delta)}{n}}\right) \leq \frac{\delta}{4},$$

where $\widehat{\mathcal{R}}^c(h) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(h(x_i^c) \neq y_i^c)$ is the loss of h on the clean data. Since the total number of poisoned points $|\mathfrak{M}| \sim \mathrm{Bin}(n,\alpha)$ and since $n > \frac{12\log(4/\delta)}{\alpha}$

$$\mathbb{P}^{\mathcal{A}}\left(\sup_{h\in\mathcal{H}} |\widehat{\mathcal{R}}^c(h) - \widehat{\mathcal{R}}^p(h)| > \frac{3\alpha}{2}\right) \leq \mathbb{P}^{\mathcal{A}}\left(|\mathfrak{M}| \geq \frac{3\alpha}{2}n\right) \leq e^{-\alpha n/12} \leq \frac{\delta}{4}.$$

Since $\sup_{h\in\mathcal{H}} |\widehat{\mathcal{R}}^p(h) - \mathcal{R}(h)| \leq \sup_{h\in\mathcal{H}} |\widehat{\mathcal{R}}^p(h) - \widehat{\mathcal{R}}^c(h)| + \sup_{h\in\mathcal{H}} |\widehat{\mathcal{R}}^c(h) - \mathcal{R}(h)|$, we obtain

$$\mathbb{P}^{\mathcal{A}}\left(\sup_{h\in\mathcal{H}} |\widehat{\mathcal{R}}^p(h) - \mathcal{R}(h)| > \frac{3\alpha}{2} + 2\sqrt{\frac{8d\log(\frac{en}{d}) + 2\log(16/\delta)}{n}}\right) \leq \frac{\delta}{2}. \tag{47}$$

In addition, Lemma 2 implies that

$$\mathbb{P}^{\mathcal{A}}\left(\sup_{h\in\mathcal{H}} \left|\widehat{\mathcal{D}}^{par}(h) - \mathcal{D}^{par}(h)\right| > \Delta^{par} + 16\sqrt{\frac{2d\log(\frac{2en}{d}) + 2\log(96/\delta)}{(1-\alpha)P_0 n}}\right) \leq \frac{\delta}{2}. \tag{48}$$

Now let $h_\lambda = \mathrm{argmin}_{h\in\mathcal{H}}(\widehat{R}^p(h) + \lambda\widehat{\mathcal{D}}^{par}(h))$ and let

$$\Delta_\lambda^{par} = 3\alpha + 4\sqrt{\frac{8d\log(\frac{en}{d}) + 2\log(16/\delta)}{n}} + 2\lambda\Delta^{par} + 32\lambda\sqrt{\frac{2d\log(\frac{2en}{d}) + 2\log(96/\delta)}{(1-\alpha)P_0 n}}$$

9. the $\widetilde{\mathcal{O}}$-notation hides constant and logarithmic factors

$$= \widetilde{\mathcal{O}}\left(\sqrt{\frac{d}{n}} + \lambda\sqrt{\frac{d}{P_0 n}}\right).$$

Then, using (47) and (50), we have that with probability at least $1-\delta$

$$\begin{aligned} L_\lambda^{par}(\mathcal{L}_\lambda^{par}(S^p)) &= \mathcal{R}(\mathcal{L}_\lambda^{par}(S^p)) + \lambda\mathcal{D}^{par}(\mathcal{L}_\lambda^{par}(S^p)) \\ &\leq \widehat{\mathcal{R}}^p(\mathcal{L}_\lambda^{par}(S^p)) + \lambda\widehat{\mathcal{D}}^{par}(\mathcal{L}_\lambda^{par}(S^p)) + \frac{1}{2}\Delta_\lambda^{par} \\ &= \min_{h\in\mathcal{H}}\left(\widehat{\mathcal{R}}^p(h) + \lambda\widehat{\mathcal{D}}^{par}(h)\right) + \frac{1}{2}\Delta_\lambda^{par} \\ &\leq \min_{h\in\mathcal{H}} L_\lambda^{par}(h) + \Delta_\lambda^{par}. \end{aligned}$$

■

Bound for equal opportunity We now show a similar result for the weighted-objective with the equal opportunity deviation measure

$$L_\lambda^{opp}(h) = \mathcal{R}(h) + \lambda\mathcal{D}^{opp}(h).$$

Let $\mathcal{L}_\lambda^{opp} : \cup_{n=1}^\infty (\mathcal{X}\times A\times\mathcal{Y})^n \to \mathcal{H}$ be such that

$$\mathcal{L}_\lambda^{opp}(S^p) \in \operatorname*{argmin}_{h\in\mathcal{H}}(\widehat{R}^p(h) + \lambda\widehat{\mathcal{D}}^{opp}(h)), \quad \text{for all } S^p.$$

That is, $\mathcal{L}_\lambda^{opp}$ always returns a minimizer of the λ-weighted empirical objective. Then:

Theorem 6 *Let $\mathcal{H}$ be any hypothesis space with $d = VC(\mathcal{H}) < \infty$. Let $\mathbb{P}\in\mathcal{P}(\mathcal{X}\times A\times\mathcal{Y})$ be a fixed distribution and $\mathcal{A}$ be any malicious adversary of power $\alpha < 0.5$. Denote by $\mathbb{P}^{\mathcal{A}}$ the probability distribution of the poisoned data S^p, under the random sampling of the clean data, the marked points and the randomness of the adversary. Then for any $\delta\in(0,1)$ and $n \geq \max\left\{\frac{8\log(16/\delta)}{(1-\alpha)P_{10}}, \frac{12\log(12/\delta)}{\alpha}, \frac{d}{2}\right\}$, we have*

$$\mathbb{P}^{\mathcal{A}}\left(L_\lambda^{opp}(\mathcal{L}_\lambda^{opp}(S^p)) \leq \min_{h\in\mathcal{H}} L_\lambda^{opp}(h) + \Delta_\lambda^{opp}\right) \leq \delta,$$

where

$$\Delta_\lambda^{opp} = 3\alpha + 2\lambda\Delta^{opp} + \widetilde{\mathcal{O}}\left(\sqrt{\frac{d}{n}} + \lambda\sqrt{\frac{d}{P_{10}n}}\right)$$

and

$$\Delta^{opp} = \frac{2\alpha}{P_{10}/3+\alpha} = \mathcal{O}\left(\frac{\alpha}{P_{10}}\right).$$

Proof Similarly to the proof of Theorem 5, we combine

$$\mathbb{P}^{\mathcal{A}}\left(\sup_{h\in\mathcal{H}}|\widehat{\mathcal{R}}^p(h)-\mathcal{R}(h)|>\frac{3\alpha}{2}+2\sqrt{\frac{8d\log(\frac{en}{d})+2\log(16/\delta)}{n}}\right)\leq\frac{\delta}{2}. \tag{49}$$

and Lemma 5

$$\mathbb{P}^{\mathcal{A}}_{S^p}\left(\sup_{h\in\mathcal{H}}|\widehat{\mathcal{D}}^{opp}(h)-\mathcal{D}^{opp}(h)|>\Delta^{opp}+16\sqrt{\frac{2d\log(\frac{2en}{d})+2\log(96/\delta)}{(1-\alpha)P_{10}n}}\right)\leq\frac{\delta}{2} \tag{50}$$

Now let $h_\lambda=\operatorname{argmin}_{h\in\mathcal{H}}(\widehat{R}^p(h)+\lambda\widehat{\mathcal{D}}^{opp}(h))$ and let

$$\Delta^{opp}_\lambda=3\alpha+4\sqrt{\frac{8d\log(\frac{en}{d})+2\log(16/\delta)}{n}}+2\lambda\Delta^{opp}+32\lambda\sqrt{\frac{2d\log(\frac{2en}{d})+2\log(96/\delta)}{(1-\alpha)P_{10}n}}.$$

Then we have that with probability at least $1-\delta$

$$\begin{aligned}
L^{opp}_\lambda(\mathcal{L}^{opp}_\lambda(S^p))&=\mathcal{R}(\mathcal{L}^{opp}_\lambda(S^p))+\lambda\mathcal{D}^{opp}(\mathcal{L}^{opp}_\lambda(S^p))\\
&\leq\widehat{\mathcal{R}}^p(\mathcal{L}^{opp}_\lambda(S^p))+\lambda\widehat{\mathcal{D}}^{opp}(\mathcal{L}^{par}_\lambda(S^p))+\frac{1}{2}\Delta^{opp}_\lambda\\
&=\min_{h\in\mathcal{H}}\left(\widehat{\mathcal{R}}^p(h)+\lambda\widehat{\mathcal{D}}^{opp}(h)\right)+\frac{1}{2}\Delta^{opp}_\lambda\\
&\leq\min_{h\in\mathcal{H}}L^{opp}_\lambda(h)+\Delta^{opp}_\lambda.
\end{aligned}$$

■

B.3.2 Component-wise upper bounds

We now prove the component-wise upper bound results.

Bound for demographic parity Recall our notation $\widehat{h}^r\in\operatorname{argmin}_{h\in\mathcal{H}}\widehat{\mathcal{R}}^p(h)$ and $\widehat{h}^{par}\in\operatorname{argmin}_{h\in\mathcal{H}}\widehat{\mathcal{D}}^{par}(h)$. Further, we define the sets

$$\mathcal{H}_1=\left\{h\in\mathcal{H}:\widehat{\mathcal{R}}^p(h)-\widehat{\mathcal{R}}^p(\widehat{h}^r)\leq3\alpha+4\sqrt{\frac{8d\log(\frac{en}{d})+2\log(16/\delta)}{n}}\right\}$$

$$\mathcal{H}_2=\left\{h\in\mathcal{H}:\widehat{\mathcal{D}}^{par}(h)-\widehat{\mathcal{D}}^{par}(\widehat{h}^{par})\leq2\Delta^{par}+32\sqrt{\frac{2d\log(\frac{2en}{d})+2\log(96/\delta)}{(1-\alpha)P_0n}}\right\}.$$

That is, $\mathcal{H}_1$ and $\mathcal{H}_2$ are the sets of classifiers that are not far from optimal on the train data, in terms of their risk and their fairness respectively. Define the *component-wise learner*:

$$\mathcal{L}^{par}_{cw}(S^p)=\begin{cases}\text{any } h\in\mathcal{H}_1\cap\mathcal{H}_2, & \text{if } \mathcal{H}_1\cap\mathcal{H}_2\neq\emptyset\\ \text{any } h\in\mathcal{H}, & \text{otherwise,}\end{cases}$$

that returns a classifier that is good in both metrics, if such exists, or an arbitrary classifier otherwise. Then we have the following:

Theorem 7 *Let $\mathcal{H}$ be any hypothesis space with $d = VC(\mathcal{H}) < \infty$. Let $\mathbb{P} \in \mathcal{P}(\mathcal{X} \times A \times \mathcal{Y})$ be a fixed distribution and let $\mathcal{A}$ be any malicious adversary of power $\alpha < 0.5$. Suppose that there exists a hypothesis $h^* \in \mathcal{H}$, such that $\mathfrak{V}(h^*) \preceq \mathfrak{V}(h)$ for all $h \in \mathcal{H}$. Then for any $\delta \in (0,1)$ and $n \geq \max\left\{\frac{8\log(16/\delta)}{(1-\alpha)P_0}, \frac{12\log(12/\delta)}{\alpha}, \frac{d}{2}\right\}$, with probability at least $1-\delta$:*

$$\mathbf{L}^{par}(\mathcal{L}^{par}_{cw}(S^p)) \preceq \left(6\alpha + \widetilde{\mathcal{O}}\left(\sqrt{\frac{d}{n}}\right), 4\Delta^{par} + \widetilde{\mathcal{O}}\left(\sqrt{\frac{d}{P_0 n}}\right)\right).$$

Proof From the proof of Theorem 5, we have that with probability at least $1-\delta$, both of the following hold:

$$\sup_{h\in\mathcal{H}} |\widehat{\mathcal{R}}^p(h) - \mathcal{R}(h)| \leq \frac{3\alpha}{2} + 2\sqrt{\frac{8d\log(\frac{en}{d}) + 2\log(16/\delta)}{n}},$$

$$\sup_{h\in\mathcal{H}} \left|\widehat{\mathcal{D}}^{par}(h) - \mathcal{D}^{par}(h)\right| \leq \Delta^{par} + 16\sqrt{\frac{2d\log(\frac{2en}{d}) + 2\log(96/\delta)}{(1-\alpha)P_0 n}}.$$

We show that under this event, $\mathcal{H}_1 \cap \mathcal{H}_2 \neq \emptyset$ and for any $h \in \mathcal{H}_1 \cap \mathcal{H}_2$,

$$\mathbf{L}^{par}(h) \preceq \left(6\alpha + 8\sqrt{\frac{8d\log(\frac{en}{d}) + 2\log(16/\delta)}{n}}, 4\Delta^{par} + 64\sqrt{\frac{2d\log(\frac{2en}{d}) + 2\log(96/\delta)}{(1-\alpha)P_0 n}}\right),$$

from which the result follows. Note that

$$\begin{aligned}
\widehat{\mathcal{R}}^p(h^*) &\leq \mathcal{R}(h^*) + \frac{3\alpha}{2} + 2\sqrt{\frac{8d\log(\frac{en}{d}) + 2\log(16/\delta)}{n}} \\
&\leq \mathcal{R}(\widehat{h}^r) + \frac{3\alpha}{2} + 2\sqrt{\frac{8d\log(\frac{en}{d}) + 2\log(16/\delta)}{n}} \\
&\leq \widehat{\mathcal{R}}^p(\widehat{h}^r) + 3\alpha + 4\sqrt{\frac{8d\log(\frac{en}{d}) + 2\log(16/\delta)}{n}}
\end{aligned}$$

and similarly

$$\widehat{\mathcal{D}}^{par}(h^*) \leq \widehat{\mathcal{D}}^{par}(\widehat{h}^r) + 2\Delta^{par} + 32\sqrt{\frac{2d\log(\frac{2en}{d}) + 2\log(96/\delta)}{(1-\alpha)P_0 n}}$$

Therefore, $h^* \in \mathcal{H}_1 \cap \mathcal{H}_2$ and so $\mathcal{H}_1 \cap \mathcal{H}_2 \neq \emptyset$.

Now take any $h \in \mathcal{H}_1 \cap \mathcal{H}_2$. We have that

$$\begin{aligned}
\mathcal{R}(h) &\leq \widehat{\mathcal{R}}^p(h) + \frac{3\alpha}{2} + 2\sqrt{\frac{8d\log(\frac{en}{d}) + 2\log(16/\delta)}{n}} \\
&\leq \widehat{\mathcal{R}}^p(\widehat{h}^r) + 3\frac{3\alpha}{2} + 6\sqrt{\frac{8d\log(\frac{en}{d}) + 2\log(16/\delta)}{n}} \\
&\leq \widehat{\mathcal{R}}^p(h^*) + 3\frac{3\alpha}{2} + 6\sqrt{\frac{8d\log(\frac{en}{d}) + 2\log(16/\delta)}{n}}
\end{aligned}$$

$$\leq \mathcal{R}(h^*) + 6\alpha + 8\sqrt{\frac{8d\log(\frac{en}{d}) + 2\log(16/\delta)}{n}}.$$

Similarly,

$$\mathcal{D}^{par}(h) \leq \mathcal{D}^{par}(h^*) + 4\Delta^{par} + 64\sqrt{\frac{2d\log(\frac{2en}{d}) + 2\log(96/\delta)}{(1-\alpha)P_0 n}}$$

and the result follows. ■

Bound for equal opportunity Similarly, let $\widehat{h}^{opp} \in \text{argmin}_{h\in\mathcal{H}} \widehat{\mathcal{D}}^{opp}(h)$. Further, we define the set

$$\mathcal{H}_3 = \left\{ h \in \mathcal{H} : \widehat{\mathcal{D}}^{opp}(h) - \widehat{\mathcal{D}}^{opp}(\widehat{h}^{opp}) \leq 2\Delta^{opp} + 32\sqrt{\frac{2d\log(\frac{2en}{d}) + 2\log(96/\delta)}{(1-\alpha)P_{10} n}} \right\}.$$

That is, $\mathcal{H}_3$ is the set of classifiers that are not far from optimal on the train data, in terms of equal opportunity fairness. Now define the *component-wise learner* for equal opportunity:

$$\mathcal{L}^{opp}_{cw}(S^p) = \begin{cases} \text{any } h \in \mathcal{H}_1 \cap \mathcal{H}_3, & \text{if } \mathcal{H}_1 \cap \mathcal{H}_3 \neq \emptyset \\ \text{any } h \in \mathcal{H}, & \text{otherwise,} \end{cases}$$

that returns a classifier that is good in both metrics, if such exists, or an arbitrary classifier otherwise. Then we have the following:

Theorem 8 *Let $\mathcal{H}$ be any hypothesis space with $d = VC(\mathcal{H}) < \infty$. Let $\mathbb{P} \in \mathcal{P}(\mathcal{X} \times A \times \mathcal{Y})$ be a fixed distribution and let $\mathcal{A}$ be any malicious adversary of power $\alpha < 0.5$. Suppose that there exists a hypothesis $h^* \in \mathcal{H}$, such that $\mathfrak{V}(h^*) \preceq \mathfrak{V}(h)$ for all $h \in \mathcal{H}$. Then for any $\delta \in (0,1)$ and $n \geq \max\left\{\frac{8\log(16/\delta)}{(1-\alpha)P_{10}}, \frac{12\log(12/\delta)}{\alpha}, \frac{d}{2}\right\}$, with probability at least $1-\delta$*

$$\mathbf{L}^{opp}(\mathcal{L}^{opp}_{cw}(S^p)) \preceq \left(6\alpha + \widetilde{\mathcal{O}}\left(\sqrt{\frac{d}{n}}\right), 4\Delta^{opp} + \widetilde{\mathcal{O}}\left(\sqrt{\frac{d}{P_{10}n}}\right)\right).$$

Proof From the proof of Theorem 6 we have that with probability at least $1-\delta$:

$$\sup_{h\in\mathcal{H}} |\widehat{\mathcal{R}}^p(h) - \mathcal{R}(h)| \leq \frac{3\alpha}{2} + 2\sqrt{\frac{8d\log(\frac{en}{d}) + 2\log(16/\delta)}{n}}$$

and Lemma 5

$$\sup_{h\in\mathcal{H}} |\widehat{\mathcal{D}}^{opp}(h) - \mathcal{D}^{opp}(h)| \leq \Delta^{opp} + 16\sqrt{\frac{2d\log(\frac{2en}{d}) + 2\log(96/\delta)}{(1-\alpha)P_{10}n}}.$$

The proof proceeds in an identical manner to that of Theorem 7. ■

Upper bound with fast rates Recall our notation:

$$\bar{\gamma}^p_{1a}(h) = \frac{\sum_{i=1}^n \mathbb{1}\{h(x_i^p) = 0, a_i^p = a, y_1^p = 1\}}{\sum_{i=1}^n \mathbb{1}\{a_i^p = a, y_i^p = 1\}} \tag{51}$$

as the empirical estimate of $\bar{\gamma}_{1a}(h) \coloneqq \mathbb{P}(h(X) = 0 | A = a, Y = 1) = 0$ for $a \in \{0, 1\}$. Given a (corrupted) training set S^p, denote by

$$\mathcal{H}^* := \left\{ h \in \mathcal{H} \middle| \max_a \bar{\gamma}^p_{1a}(h) \leq \Delta^{opp} \wedge \widehat{\mathcal{R}}^p(h) \leq \frac{3\alpha}{2} \right\} \tag{52}$$

the set of all classifiers that have a small loss and small values of $\bar{\gamma}^p_{1a}$ for both $a \in \{0, 1\}$ on S^p. Consider the learner $\mathcal{L}^{fast}$ defined by

$$\mathcal{L}^{fast}(S^p) = \begin{cases} \text{any } h \in \mathcal{H}^*, & \text{if } \mathcal{H}^* \neq \emptyset \\ \text{any } h \in \mathcal{H}, & \text{otherwise.} \end{cases} \tag{53}$$

We then have the following:

Theorem 9 *Let $\mathcal{H}$ be finite and $\mathbb{P} \in \mathcal{P}(\mathcal{X} \times A \times \mathcal{Y})$ be such that for some $h^* \in \mathcal{H}$, $\mathbb{P}(h^*(X) = Y) = 1$. Denote by $P_{1a} = \mathbb{P}(Y = 1, A = a)$ for $a \in \{0, 1\}$. Let $\mathcal{A}$ be any malicious adversary of power $\alpha < 0.5$. Then for any $\delta, \eta \in (0, 1)$ and any*

$$\begin{aligned} n &\geq \max\left\{ \frac{8\log(16|\mathcal{H}|/\delta)}{(1-\alpha)P_{10}}, \frac{12\log(12/\delta)}{\alpha}, \frac{2\log(8|\mathcal{H}|/\delta)}{3\eta^2\alpha}, \frac{2\log(\frac{16|\mathcal{H}|}{\delta})}{3\eta^2(1-\alpha)P_{10}\alpha} \right\} \\ &= \Omega\left(\frac{\log(|\mathcal{H}|/\delta)}{\eta^2 P_{10}\alpha} \right) \end{aligned}$$

with probability at least $1 - \delta$

$$\mathbf{L}^{opp}(\mathcal{L}^{fast}) \preceq \left(\frac{3\alpha}{1-\eta}, \frac{2\Delta^{opp}}{1-\eta} \right).$$

Proof Throughout the proof we will drop the dependence of $\mathcal{H}^*$ (and other subsets of $\mathcal{H}$) of the data S^p. We will be interested in the probability of certain events involving $\mathcal{H}^*$ under all randomness in the generation of S^p: the random sampling of the clean data, the marked point and the adversary (denoted by $\mathbb{P}^{\mathcal{A}}$ as elsewhere).

Step 1 First note that by Lemma 7(b), whenever $n \geq \max\left\{ \frac{8\log(16/\delta)}{(1-\alpha)P_{10}}, \frac{12\log(12/\delta)}{\alpha} \right\}$

$$\mathbb{P}^{\mathcal{A}}\left((\bar{\gamma}^p_{10}(h^*) > \Delta^{opp}) \vee (\bar{\gamma}^p_{11}(h^*) > \Delta^{opp}) \right) \leq \mathbb{P}^{\mathcal{A}}\left(\bar{\gamma}^p_{10}(h^*) + \bar{\gamma}^p_{11}(h^*) > \Delta^{opp} \right) \leq \frac{\delta}{4}$$

In addition, since $|\mathfrak{M}| \sim \mathrm{Bin}(n, \alpha)$

$$\mathbb{P}^{\mathcal{A}}\left(\widehat{\mathcal{R}}^p(h^*) > \frac{3\alpha}{2} \right) \leq \mathbb{P}^{\mathcal{A}}\left(|\mathfrak{M}| \geq \frac{3\alpha}{2} n \right) \leq \exp(-\frac{\alpha n}{12}) \leq \frac{\delta}{12}.$$

It follows that $\mathbb{P}^{\mathcal{A}}\left(h^* \notin \mathcal{H}^* \right) \leq \frac{\delta}{4} + \frac{\delta}{12} = \frac{\delta}{3}$.

Step 2 Next let $\mathcal{H}_1 \subset \mathcal{H}$ be the set $\left\{h \in \mathcal{H} \middle| \mathcal{R}(h, \mathbb{P}) > \frac{3\alpha}{1-\eta}\right\}$. For any $h \in \mathcal{H}_1$

$$\mathbb{P}^{\mathcal{A}}\left(\widehat{\mathcal{R}}^c(h) \leq 3\alpha\right) \leq \mathbb{P}^{\mathcal{A}}\left(\text{Bin}\left(n, \frac{3\alpha}{1-\eta}\right) \leq (1-\eta)\frac{3\alpha}{(1-\eta)}n\right) \leq \exp\left(-\eta^2\frac{3\alpha}{2(1-\eta)}n\right) \leq \frac{\delta}{8|\mathcal{H}|},$$

as long as $n \geq \frac{2\log(\frac{8|\mathcal{H}|}{\delta})}{3\eta^2\alpha} > \frac{2\log(\frac{8|\mathcal{H}|}{\delta})(1-\eta)}{3\eta^2\alpha}$. Taking a union bound over all $h \in \mathcal{H}_1$,

$$\mathbb{P}^{\mathcal{A}}\left(\min_{h\in\mathcal{H}_1} \widehat{\mathcal{R}}^c(h) \leq 3\alpha\right) \leq \frac{\delta}{8}$$

Since also $\mathbb{P}^{\mathcal{A}}(|\mathfrak{M}| \geq \frac{3\alpha}{2}) \leq \frac{\delta}{12}$ and $\widehat{\mathcal{R}}^p(h) \geq \widehat{\mathcal{R}}^c(h) - |\mathfrak{M}|$, we obtain

$$\mathbb{P}^{\mathcal{A}}\left(\min_{h\in\mathcal{H}_1} \widehat{\mathcal{R}}^p(h) \leq \frac{3\alpha}{2}\right) \leq \mathbb{P}^{\mathcal{A}}\left(\left(\min_{h\in\mathcal{H}_1} \widehat{\mathcal{R}}^c(h) \leq 3\alpha\right) \vee \left(|\mathfrak{M}| \geq \frac{3\alpha}{2}\right)\right) \leq \frac{\delta}{8} + \frac{\delta}{12} = \frac{5\delta}{24}. \tag{54}$$

Similarly, let $\mathcal{H}_2 = \left\{h \in \mathcal{H} \middle| \mathcal{D}^{opp}(h) > \frac{2}{1-\eta}\Delta^{opp}\right\}$. Fix any $h \in \mathcal{H}_2$. Assume without loss of generality that $\bar{\gamma}_{10} \geq \bar{\gamma}_{11} \geq 0$ (for this particular h only). Then $\bar{\gamma}_{10} \geq \bar{\gamma}_{10} - \bar{\gamma}_{11} = |\bar{\gamma}_{10} - \bar{\gamma}_{11}| = |\gamma_{10} - \gamma_{11}| > \frac{2}{1-\eta}\Delta^{opp}$ (note that the γ_{1a} are non-negative). At the same time, by Lemma 7(a),

$$\mathbb{P}^{\mathcal{A}}\left(\bar{\gamma}^p_{10} \leq (1-\eta)\bar{\gamma}_{10} - \bar{\Delta}_{10}\right) \leq \frac{\delta}{8|\mathcal{H}|},$$

whenever

$$n > \max\left\{\frac{8\log(\frac{16|\mathcal{H}|}{\delta})}{(1-\alpha)P_{10}}, \frac{4\log(\frac{16|\mathcal{H}|}{\delta})}{\eta^2(1-\alpha)P_{10}\bar{\gamma}_{10}}\right\}.$$

This is indeed the case since $n > \frac{8\log(\frac{16|\mathcal{H}|}{\delta})}{(1-\alpha)P_{10}}$ by assumption and also

$$n > \frac{2\log(\frac{16|\mathcal{H}|}{\delta})}{3\eta^2(1-\alpha)P_{10}\alpha} \geq \frac{4\log(\frac{16|\mathcal{H}|}{\delta})}{\eta^2(1-\alpha)P_{10}\bar{\gamma}_{10}}.$$

The last inequality is obtained by observing that $\bar{\gamma}_{10} \geq \frac{2}{1-\eta}\Delta^{opp} \geq 6\alpha$, which follows by using $P_{10} \leq 0.5, \alpha \leq 0.5, \eta > 0$.

Therefore, with probability at least $1 - \frac{\delta}{8|\mathcal{H}|}$, $\max_a \bar{\gamma}^p_{1a} = \bar{\gamma}^p_{10} > (1-\eta)\bar{\gamma}_{10} - \Delta_{10} \geq 2\Delta^{opp} - E_{10} \geq 2\Delta^{opp} - E_{10} - E_{11}$, with $E_{1a} = \frac{B_{1a}}{C_{1a}+B_{1a}}$, where we used inequality (45).

Crucially, $2\Delta^{opp} - E_{10} - E_{11}$ does not depend on h. Therefore, taking a union bound over all $h \in \mathcal{H}_2$,

$$\mathbb{P}^{\mathcal{A}}\left(\min_{h\in\mathcal{H}_2}\max_a \bar{\gamma}^p_{1a}(h) \leq 2\Delta^{opp} - E_{10} - E_{11}\right) \leq \frac{\delta}{8}.$$

Note also that since $n \geq \max\left\{\frac{8\log(16/\delta)}{(1-\alpha)P_{10}}, \frac{12\log(12/\delta)}{\alpha}\right\}$, using inequality (46),

$$\mathbb{P}^{\mathcal{A}}\left(E_{10} + E_{11} > \Delta^{opp}\right) \leq \frac{\delta}{4}.$$

Therefore

$$\begin{aligned}\mathbb{P}^{\mathcal{A}}\left(\min_{h\in\mathcal{H}_2}\max_a \bar{\gamma}^p_{1a}(h) \leq \Delta^{opp}\right) &\leq \mathbb{P}^{\mathcal{A}}\left(\min_{h\in\mathcal{H}_2}\max_a \bar{\gamma}^p_{1a} \leq 2\Delta^{opp} - E_{10} - E_{11}\right) \\ &\quad + \mathbb{P}^{\mathcal{A}}\left(E_{10} + E_{11} > \Delta^{opp}\right) \\ &\leq \frac{3\delta}{8}. \end{aligned} \tag{55}$$

Finally, using (54) and (55),

$$\begin{aligned}\mathbb{P}^{\mathcal{A}}\left(\mathcal{H}^* \cap (\mathcal{H}_1 \cup \mathcal{H}_2) \neq \emptyset\right) &= \mathbb{P}^{\mathcal{A}}\left(\left(\min_{h\in\mathcal{H}_1}\widehat{\mathcal{R}}^p(h) \leq \frac{3\alpha}{2}\right) \vee \left(\min_{h\in\mathcal{H}_2}\max_a \bar{\gamma}^p_{1a}(h) \leq \Delta^{opp}\right)\right) \\ &\leq \frac{5\delta}{24} + \frac{3\delta}{8} \\ &< \frac{2\delta}{3}.\end{aligned}$$

Step 3 Combining steps 1 and 2, we have that with probability at least $1-\delta$, $h^* \in \mathcal{H}^*$ (and so $\mathcal{H}^*$ is non-empty) and for any $h \in \mathcal{H}$, $\mathcal{R}(h,\mathbb{P}) \leq \frac{3\alpha}{1-\eta}$ and $\mathcal{D}^{opp}(h) \leq \frac{2}{1-\eta}\Delta^{opp}$ which completes the proof. ■