Differential Privacy has Bounded Impact on Fairness in Classification

Paul Mangold, Michaël Perrot, Aurélien Bellet, Marc Tommasi

Differential Privacy has Bounded Impact on Fairness in Classification

Paul Mangold, Michaël Perrot, Aurélien Bellet, and Marc Tommasi

Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille, France

Abstract

We theoretically study the impact of differential privacy on fairness in classification. We prove that, given a class of models, popular group fairness measures are pointwise Lipschitz-continuous with respect to the parameters of the model. This result is a consequence of a more general statement on accuracy conditioned on an arbitrary event (such as membership to a sensitive group), which may be of independent interest. We use the aforementioned Lipschitz property to prove a high probability bound showing that, given enough examples, the fairness level of private models is close to the one of their non-private counterparts.

1 Introduction

The performance of machine learning models have mainly been evaluated in terms of utility, that is their ability to solve specific tasks. However, they can be used in sensitive contexts, and impact people's lives. It is thus crucial that users can trust these models. While trustworthiness encompasses multiple concepts, fairness and privacy have attracted a lot of interest in the past few years. Fairness requires models not to unjustly discriminate against specific individuals or subgroups of the population, and privacy preserves individual-level information about the training data from being inferred from the model. These two notions have been extensively studied in isolation: there exists numerous approaches to learn fair models (Caton and Haas, 2020; Mehrabi et al., 2021), or to preserve privacy (Dwork et al., 2014; Liu et al., 2021). However, only few works studied the interplay between privacy and fairness. In this paper, we take a step forward in this direction, proposing a new theoretical bound on the relative impact of privacy on fairness in classification.

Fairness takes various forms (depending on the task and context), and several definitions exist. On the one hand, the goal may be to ensure that similar individuals are treated similarly. This is captured by individual fairness (Dwork et al., 2012) and counterfactual fairness (Kusner et al., 2017). On the other hand, group fairness requires that decisions made by machine learning models do not unjustly discriminate against subgroups of the population. In this paper, we focus on group fairness and consider four popular definitions, namely Equalized Odds (Hardt et al., 2016), Equality of Opportunity (Hardt et al., 2016), Accuracy Parity (Zafar et al., 2017), and Demographic Parity (Calders et al., 2009).

Differential privacy (Dwork, 2006) has been widely adopted for controlling how much information the output of an algorithm may leak about its input data. It allows publishing machine learning models while preventing an adversary from guessing too confidently the presence (or absence) of an individual in the training data. To enforce differential privacy, one typically releases a noisy estimate of the true model (Dwork, 2006), so as to conceal any sensitive information contained in individual data points. This induces a trade-off between the strength of the protection and the utility of the learned model. While this trade-off has been extensively studied (Chaudhuri et al., 2011; Bassily et al., 2014; Liu et al., 2021), its implications for fairness are not yet well understood.

Contributions. In this work, we quantify the difference in fairness levels between private and non-private models in multi-class classification. We derive high probability bounds showing that this difference shrinks at a rate of $\widetilde{O}(\sqrt{p}/n)$. To obtain this result, we first prove that the accuracy of a model conditioned on an arbitrary event (such as membership to a sensitive group), is pointwise Lipschitz continuous with respect to the model parameters. This property is inherited

by many popular group fairness notions, such as Equalized Odds, Equality of Opportunity, Accuracy Parity and Demographic Parity. Consequently, two sufficiently close models will have similar fairness levels. We then upper-bound the distance between the optimal non-private model and the private models obtained with privacy preserving mechanisms like output perturbation (Chaudhuri et al., 2011; Lowy and Razaviyayn, 2021) or DP-SGD (Song et al., 2013; Bassily et al., 2014). These bounds hold for strongly convex empirical risk minimization formulations, potentially allowing explicit fairness-promoting convex regularization terms (Bechavod and Ligett, 2017; Huang and Vishnoi, 2019; Lohaus et al., 2020; Tran et al., 2021). Combining these two results, we derive high probability bounds on the fairness loss due to privacy. They show that, with enough training examples, (i) given an optimal non-private model, enforcing privacy will not harm fairness too much, and (ii) given a private model, the corresponding (unknown) non-private optimal model cannot be vastly fairer. Our results also highlight the role of the *confidence margin* of models in the disparate impact of differential privacy: notably, if the non-private model has high per-group confidence, then our bound on the loss in fairness due to privacy will be smaller. Our contributions can be summarized as follows:

- We prove that group fairness is pointwise Lipschitz, with a smaller constant for models with large margins.
- We bound the distance between private and optimal models, and show that the difference in their fairness levels decreases in $\widetilde{O}(\sqrt{p}/n)$.
- We show that this bound can be computed even when the optimal model is unknown, and numerically demonstrate that we obtain non-trivial guarantees.

Related work. The joint study of fairness and privacy in machine learning only goes back a few years, and has been the focus of a recent survey Fioretto et al. (2022). One may identify three main research directions. First, it has been empirically observed that privacy can exacerbate unfairness (Bagdasaryan et al., 2019; Pujol et al., 2020; Farrand et al., 2020; Uniyal et al., 2021) and, conversely, that enforcing fairness can lead to more privacy leakage for the unprivileged group (Chang and Shokri, 2020). These empirical results suggest that some properties of the dataset (such as group sizes and groupwise input norms) and the choice of the private training method may affect the extent of these disparate impacts. Unfortunately, these observations are not supported by theoretical results, and it is not clear why and when disparate impact occurs. Second, a few approaches have been proposed to learn models that are both fair and privacy preserving. However, these works either have limited theoretical guarantees on their performance (Kilbertus et al., 2018; Xu et al., 2019, 2020; Tran et al., 2020), or learn stochastic models which might not be usable in contexts where deterministic decisions are expected (Jagielski et al., 2019; Mozannar et al., 2020). Finally, a few works have shown that fairness and privacy are incompatible in some settings, in the sense that there exists data distributions where enforcing one prevents the other from being satisfied (Sanyal et al., 2022), or where enforcing both implies trivial utility (Cummings et al., 2019; Agarwal, 2020). While appealing at first glance, these results usually consider unrealistic cases that are hardly encountered in practice. In this paper, we also study fairness and privacy jointly but rather than studying whether they may be achieved simultaneously, we investigate the relative difference in fairness level between private and non-private models.

To the best of our knowledge, the work closest to ours is the one of Tran et al. (2021). They analyze the impact of privacy on fairness in Empirical Risk Minimization, where their notion of fairness is defined as the difference between the excess risk computed on the overall population and the excess risk computed on a subgroup of the population. They study the expected behavior over the possible private models while our results are model-specific. In line with our work, their results suggest that the distance to the decision boundary plays a key role in the disparate impact of differential privacy. However, the quantity appearing in their result is based on a second-order Taylor approximations which is loose for popular classification loss functions. In contrast, the quantity appearing in our bounds is precisely the confidence margin considered in prior work on multi-class margin-based classification (Cortes et al., 2013). Finally and most importantly, loss-based fairness does not necessarily imply that the actual decisions taken by the model are fair with respect to standard group-fairness notions (Lohaus et al., 2020). In contrast, our work provides guarantees in terms of these widely-accepted group fairness definitions.

2 Preliminaries

In this section, we present the fairness and privacy notions that will be used throughout the paper. We consider a multi-class classification setting with a feature space $\mathcal{X}$, a finite set of labels $\mathcal{Y}$, and a finite set $\mathcal{S}$ of values for the sensitive attribute. Let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{S} \times \mathcal{Y}$, and $D = \{(x_1, s_1, y_1), \ldots, (x_n, s_n, y_n)\}$ be a training set of n examples drawn i.i.d. from $\mathcal{D}$. Let $\mathcal{H}$ be a space of real-valued functions $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ equipped with a norm $\|\cdot\|_{\mathcal{H}}$. For an example $x \in \mathcal{X}$, the predicted label is the one with the highest value, that is $H(x) = \arg\max_{y \in \mathcal{Y}} h(x, y)$. In case of a tie, a random label among the most likely ones is predicted. The confidence margin of a model h for an example-label pair (x, y) is defined as $\rho(h, x, y) = h(x, y) - \max_{y' \neq y} h(x, y')$ (Cortes et al., 2013). This confidence margin is positive when the example x is classified as y by h and negative otherwise. In this paper, we assume that the margin is Lipschitz-continuous in the model h.

Assumption 2.1 (Lipschitzness of the margin). *We assume that ρ is Lipschitz-continuous in its first argument, that is for all $h, h' \in \mathcal{H}$ and $(x, y) \in \mathcal{X} \times \mathcal{Y}$,*

$$|\rho(h, x, y) - \rho(h', x, y)| \leq L_{x,y} \|h - h'\|_{\mathcal{H}} \quad ,$$

where $L_{x,y} < +\infty$ may depend on the example (x, y).

This assumption is not very restrictive. Typically, it is satisfied by any class of differentiable model with bounded gradients. As an illustration, consider linear models of the form $h(x, y) = W_y^T x$ where W is a real-valued matrix where each line is a vector W_y of label-specific parameters. Define $\|h - h'\|_{\mathcal{H}} = \|W - W'\|_2$.Then, we have $L_{x,y} = 2\|x\|_2$ since $|\rho(h, x, y) - \rho(h', x, y)| \leq |h(x, y) - h'(x, y)| + \max_{y' \neq y} |h(x, y') - h'(x, y')| \leq 2\|x\|_2 \|h - h'\|_{\mathcal{H}}$.

The goal of a learning algorithm $\mathcal{A} : (\mathcal{X} \times \mathcal{S} \times \mathcal{Y})^n \to \mathcal{H}$ is to find the best possible model to solve the task. In this work, the quality of a model h is evaluated through its accuracy $\mathrm{Acc}(h) = \mathbb{P}(H(X) = Y)$ but also its fairness level (as defined in Section 2.1). Furthermore, given a non-private algorithm $\mathcal{A}$, our goal will be to compare the quality of its output to that of a private version $\mathcal{A}^{\text{priv}}$ of $\mathcal{A}$ that guarantees differential privacy.

2.1 Fairness

In this paper, we focus on group fairness. These definitions are based on the idea that a group of individuals should not be discriminated against, compared to the overall population. Usually, these groups are defined by the sensitive attribute from $\mathcal{S}$. However, in some cases, it is necessary to consider more fine grained partitions. This is for example the case in Equalized Odds (Hardt et al., 2016), where a model is fair if its performance is the same on the overall population and on subgroups of individuals that share the same sensitive group and the same label. Thus, for the sake of generality, we assume that the data can be partitioned into K disjoint groups denoted by $D_1, \ldots, D_k, \ldots, D_K$. As in Maheshwari and Perrot (2022), we consider fairness definitions that, for each group k, can be written as:

$$F_k(h, D) = C_k^0 + \sum_{k'=1}^{K} C_k^{k'} \mathbb{P}(H(X) = Y \mid D_{k'}) \quad , \tag{1}$$

where the $C_k^{k'}$'s are group specific values independent of h, that typically depend on the size of the groups. In Appendix A, we show that usual group fairness notions such as Demographic Parity (with binary labels) (Calders et al., 2009), Equality of Opportunity (Hardt et al., 2016), Equalized Odds (Hardt et al., 2016), and Accuracy Parity (Zafar et al., 2017) can all be expressed in the form of (1). By convention, we consider that $F_k(h, D) > 0$ when the group k is advantaged by h compared to the overall population, $F_k(h, D) < 0$ when the group is disadvantaged and $F_k(h, D) = 0$ when h is fair for group k.

In some cases, rather than measuring fairness for each group k independently, it is interesting to summarize the information with an aggregate value. For example, we will use the mean of the absolute fairness level of each group:

$$\text{Fair}(h, D) = \frac{1}{K} \sum_{k=1}^{K} |F_k(h, D)| \quad , \tag{2}$$

which is 0 when h is fair and positive when it is unfair.

2.2 Differential Privacy

We measure the privacy of machine learning models with differential privacy (see Definition 2.2 below). Differential privacy (DP) guarantees that the outcomes of a randomized algorithm are similar when run on datasets that differ in at most one data point. It effectively preserves privacy by preventing an adversary observing the trained model from inferring the presence of an individual in the training set. A key property of differential privacy is that it still holds after post-processing of the algorithm's outcome (Dwork, 2006), as long as this post-processing is independent of the data. Let $D, D' \in (\mathcal{X} \times \mathcal{S} \times \mathcal{Y})^n$ be two datasets of n elements. We say that they are *neighboring* (denoted by $D \approx D'$) if they differ in at most one element.

Definition 2.2 (Differential Privacy – Dwork (2006)). *Let $\mathcal{A}^{\text{priv}} : (\mathcal{X} \times \mathcal{S} \times \mathcal{Y})^n \to \mathcal{H}$ be a randomized algorithm. We say that $\mathcal{A}^{\text{priv}}$ is (ϵ, δ)-differentially private if, for all neighboring datasets $D, D' \in (\mathcal{X} \times \mathcal{S} \times \mathcal{Y})^n$ and all subsets of hypotheses $\mathcal{H}' \subseteq \mathcal{H}$,*

$$\mathbb{P}(\mathcal{A}^{\text{priv}}(D) \in \mathcal{H}') \leq \exp(\epsilon)\, \mathbb{P}(\mathcal{A}^{\text{priv}}(D') \in \mathcal{H}') + \delta \ .$$

To design differentially private algorithms to estimate a function $\mathcal{A} : (\mathcal{X} \times \mathcal{S} \times \mathcal{Y})^n \to \mathbb{R}^p$, we need to quantify how much changing one point in a dataset can impact the output of $\mathcal{A}$. This is typically measured by (an upper bound on) the ℓ_2-sensitivity of $\mathcal{A}$, defined as

$$\Delta(\mathcal{A}) = \sup_{D \approx D'} \|\mathcal{A}(D) - \mathcal{A}(D')\|_2 \ .$$

The value of $\mathcal{A}$ on a dataset $D \in (\mathcal{X} \times \mathcal{S} \times \mathcal{Y})^n$ can then be released privately using the Gaussian mechanism (Dwork et al., 2014). Formally, to guarantee (ϵ, δ)-differential privacy, we add Gaussian noise to $\mathcal{A}(D)$, calibrated to its sensitivity and the desired level of privacy:

$$\mathcal{A}^{\text{priv}}(D) = \mathcal{A}(D) + \mathcal{N}\Big(0, \frac{2\Delta(\mathcal{A})^2 \log(1.25/\delta)}{\epsilon^2} \mathbb{I}_p\Big) \ ,$$

where $\mathcal{N}(0, \sigma^2 \mathbb{I}_p)$ is a sample from the normal distribution with mean zero and variance $\sigma^2 \mathbb{I}_p$. In many cases (*e.g.,* when the dataset D is large), $\mathcal{A}^{\text{priv}}$ is computed on a random subsample of D. Assuming $\mathcal{A}^{\text{priv}}$ is (ϵ, δ)-differentially private, applying $\mathcal{A}^{\text{priv}}$ to a randomly selected fraction q of D satisfies $(O(q\epsilon), q\delta)$-differential privacy, thereby amplifying privacy guarantees (Kasiviswanathan et al., 2011; Beimel et al., 2014). This privacy amplification by subsampling phenomenon, together with the Gaussian mechanism, serve as building blocks in more complex algorithms. In particular, they can be composed (Dwork et al., 2014), allowing the design of iterative private algorithms such as DP-SGD (Bassily et al., 2014; Abadi et al., 2016).

3 Pointwise Lipschitzness and Group Fairness

Here, we show that several *group fairness notions are pointwise Lipschitz* with respect to the model. To this end, we first prove a more general result on the pointwise Lipschitzness of accuracy conditionally on an arbitrary event.

3.1 Pointwise Lipschitzness of Conditional Accuracy

We first relate the difference of conditional accuracy of two models to the distance that separates them. This is summarized in the next theorem.

Theorem 3.1 (Pointwise Lipschitzness of Conditional Accuracy). *Let $\mathcal{H}$ be a set of real-valued functions with $L_{X,Y}$ the Lipschitz constants defined in Assumption 2.1. Let $h, h' \in \mathcal{H}$ be two models, (X, Y, S) be a triple of random variables with distribution $\mathcal{D}$, and E be an arbitrary event. Assume that $\mathbb{E}\left({}^{L_{X,Y}}/_{|\rho(h', X, Y)|} \mid E\right) < +\infty$, then*

$$|\mathbb{P}(H(X) = Y \mid E) - \mathbb{P}(H'(X) = Y \mid E)| \leq \mathbb{E}\left(\frac{L_{X,Y}}{|\rho(h, X, Y)|} \,\middle|\, E\right) \|h - h'\|_{\mathcal{H}} \ . \tag{Lip}$$

Proof. (Sketch) The proof of this theorem is in two steps. First, we use the Lipschitzness of the margin (Assumption 2.1), the triangle inequality, and the union bound to show that $|\mathbb{P}(H(X) = Y \mid E) - \mathbb{P}(H'(X) = Y \mid E)| \leq \mathbb{P}\left({}^{L_{X,Y}}/_{|\rho(h,X,Y)|} \geq {}^{1}/_{\|h-h'\|_{\mathcal{H}}} \mid E\right)$. Then, applying Markov's inequality gives the desired result. The complete proof can be found in Appendix B. □

Theorem 3.1 shows the pointwise lipschitzness of the function $h \mapsto \mathbb{P}(H(X) = Y \mid E)$. Furthermore, it underlies the importance of having a large confidence margin $\rho(h, x, y)$ for a model h predicting label y for an example x. Hence, ${}^{L_{x,y}}/_{|\rho(h,x,y)|}$ is small when the model h is confident in its prediction for the true label y. This implies that, when the probability (given E) that a point has a small margin is small, $\mathbb{E}\left(\frac{L_{X,Y}}{|\rho(h,X,Y)|} \;\middle|\; E\right)$ is also small. This is notably the case for large margin classifiers.

It is worth noting that the bound presented Theorem 3.1 can be tightened (at the expense of readability) without affecting the quantities that need to be controlled, that is the margin $|\rho(h, x, y)|$ and the distance $\|h - h'\|_{\mathcal{H}}$. Hence, note that given $(x, y) \in \mathcal{X} \times \mathcal{Y}$, if $|\rho(h, x, y)| \geq L_{x,y} \|h - h'\|_{\mathcal{H}}$, then it means that h's margin is large enough to ensure that h and h' have the same prediction on x. The corresponding term in the expectation may then be accounted for as zero, improving the upper bound (Remark B.2). Interestingly, if all the examples are classified with such a large margin, our bound becomes 0, further hinting toward the importance of large margin classifiers. This result may be further tightened by using a Chernoff bound instead of Markov's inequality (Remark B.1), yielding $|\mathbb{P}(H(X)=Y|E)-\mathbb{P}(H'(X)=Y|E)| \leq \beta_{X,Y}(h)$, with

$$\beta_{X,Y}(h) = \inf_{t \geq 0} \left\{ e^{t\|h-h'\|_{\mathcal{H}}} \, \mathbb{E}\left(e^{-\frac{t|\rho(h,X,Y)|}{L_{X,Y}}} \middle| E\right)\right\} \ .$$

In the subsequent theoretical developments, we use the bound derived in Theorem 3.1 for the sake of readability. In the numerical experiments (Section 5), we use the version of the bound that yields the tightest results by combining both of the aforementioned techniques.

3.2 Pointwise Lipschitzness of Group Fairness Notions

We now use Theorem 3.1's general result to relate the fairness levels of two classifiers, based on their distance. In Theorem 3.2, we show that fairness notions in the form of (1) are pointwise Lipschitz.

Theorem 3.2 (Pointwise Lipschitzness of Fairness). *Let $h, h' \in \mathcal{H}$, and $L_{X,Y}$ defined as in Assumption 2.1. For any fairness notion of the form of* (1)*, we have, for all $k \in [K]$,*

$$|F_k(h, D) - F_k(h', D)| \leq \chi_k(h, D) \, \|h - h'\|_{\mathcal{H}} \ .$$

with $\chi_k(h, D) = \sum_{k'=1}^{K} \left|C_k^{k'}\right| \mathbb{E}\left(\frac{L_{X,Y}}{|\rho(h,X,Y)|} \;\middle|\; D_{k'}\right)$. Similarly, for the aggregate measure of fairness defined in (2)*,*

$$|\mathit{Fair}(h, D) - \mathit{Fair}(h', D)| \leq \frac{1}{K} \sum_{k=1}^{K} \chi_k(h, D) \, \|h - h'\|_{\mathcal{H}} \ .$$

Proof. (Sketch) To prove the first claim, we use the triangle inequality to show that, for each group, the absolute difference in fairness is bounded by a combination of absolute differences between conditional probabilities. We can then apply Theorem 3.1. The second claim follows by applying the first one to each group independently. The complete proof is provided in Appendix C. □

Theorem 3.2 implies that *classifiers that are sufficiently close have similar fairness levels*. This has two major consequences when studying a given model. On the one hand, we have an upper bound on the harm that can be done to fairness: small variations of the model cannot make it much more unfair. On the other hand, we have a lower bound on the distance needed to make a model fair: making the model significantly more fair requires to substantially alter it. In the next corollary, we instantiate Theorem 3.2 for various popular group fairness notions, and for accuracy.

Corollary 3.3. *Let $h, h' \in \mathcal{H}$, and $L_{X,Y}$ defined as in Assumption 2.1. The difference in fairness or accuracy between h and h' can be bounded as follows.*

***Equalized Odds (Hardt et al., 2016)**: the data is divided into $K = |\mathcal{Y} \times \mathcal{S}|$ groups such that for all $(y, r) \in \mathcal{Y} \times \mathcal{S}$,*

$$\chi_{(y,r)}(h, D) = \mathbb{E}\left(\frac{L_{X,Y}}{|\rho(h, X, Y)|} \,\middle|\, Y = y\right) + \mathbb{E}\left(\frac{L_{X,Y}}{|\rho(h, X, Y)|} \,\middle|\, Y = y, S = r\right) \quad .$$

***Equality of Opportunity (Hardt et al., 2016)**: we let $\mathcal{Y}' \subseteq \mathcal{Y}$ the set of desirable outcomes. The data is divided into $K = |\mathcal{Y} \times \mathcal{S}|$ such that for all $(y, r) \in \mathcal{Y} \times \mathcal{S}$,*

$$\chi_{(y,r)}(h, D) = \mathbb{E}\left(\frac{L_{X,Y}}{|\rho(h, X, Y)|} \,\middle|\, Y = y, S = r\right) + \mathbb{E}\left(\frac{L_{X,Y}}{|\rho(h, X, Y)|} \,\middle|\, Y = y\right) \quad ,$$

if y is a desired outcome, and $\chi_{(y,r)}(h, D) = 0$ otherwise.

***Accuracy Parity (Zafar et al., 2017)**: the data is divided into $K = |\mathcal{S}|$ groups such that for all $r \in \mathcal{S}$,*

$$\chi_{(r)}(h, D) = \mathbb{E}\left(\frac{L_{X,Y}}{|\rho(h, X, Y)|}\right) + \mathbb{E}\left(\frac{L_{X,Y}}{|\rho(h, X, Y)|} \,\middle|\, S = r\right) \quad .$$

***Demographic Parity (Binary Labels) (Calders et al., 2009)**: the data is divided into $K = |\mathcal{Y} \times \mathcal{S}|$ groups such that for all $(y, r) \in \mathcal{Y} \times \mathcal{S}$,*

$$\chi_{(y,r)}(h, D) = \mathbb{E}\left(\frac{L_{X,Y}}{|\rho(h, X, Y)|}\right) + \mathbb{E}\left(\frac{L_{X,Y}}{|\rho(h, X, Y)|} \,\middle|\, S = r\right) \quad .$$

***Accuracy**: the data is in a single group, such that*

$$\chi(h, D) = \mathbb{E}\left(\frac{L_{X,Y}}{|\rho(h, X, Y)|}\right) \quad .$$

Proof. This corollary follows from Theorem 3.2 by replacing the $C_k^{k'}$'s by their appropriate values (depending on the considered notion). See Appendix A for more details. □

Corollary 3.3 shows that our results are applicable to several *group fairness notions*, but also to *accuracy*. Note that the pointwise Lipschitz constant $\chi_k(h, D)$ depends on the considered notion. In Section 4, we use these results to quantify the relative fairness level between private and non-private models.

Finite sample analysis. In practice, it is often assumed that one does not have access to the true distribution $\mathcal{D}$ but rather to a finite sample $D = \{(x_1, s_1, y_1), \ldots, (x_n, s_n, y_n)\}$ of size n. An empirical estimate of the expectation from a finite sample is then defined as $\widehat{\mathbb{E}}\left(f(X)\right) = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$. The results presented in Theorem 3.1, Theorem 3.2, and Corollary 3.3 also hold in this finite sample setting. For instance, denoting by $\widehat{\chi}_k$ the empirical version of χ_k, we have that $\forall k \in [K]$,

$$\left|\widehat{F}_k(h, D) - \widehat{F}_k(h', D)\right| \leq \widehat{\chi}_k(h, D)\left\|h - h'\right\|_{\mathcal{H}} \quad .$$

One may then wonder whether it is possible to connect the true fairness of a model h to the empirical fairness of a second model h', that is bound $\left|F_k(h, D) - \widehat{F}_k(h', D)\right|$. In the next lemma, we show that such bound can indeed be obtained when h and h' were learned on D.

Lemma 3.4. *Let D be a finite sample of $n \geq \frac{8\log\left(\frac{2K+1}{\delta}\right)}{\min_{k'} p_{k'}}$ examples drawn i.i.d. from $\mathcal{D}$, where $p_{k'}$ is the true proportion of examples from group k'. Assume that $\mathbb{P}_{D\sim\mathcal{D}^n}\left(\sum_{k'=0}^{K}\left|C_k^{k'} - \widehat{C}_k^{k'}\right| > \alpha_C\right) \leq B_3\exp\left(-B_4\alpha_C^2 n\right)$. Let $\mathcal{H}$ be an*

hypothesis space and $d_{\mathcal{H}}$ be the Natarajan dimension of $\mathcal{H}$. With probability at least $1-\delta$ over the choice of D, $\forall h, h' \in \mathcal{H}$

$$\left|F_k(h,D) - \widehat{F}_k(h',D)\right| \leq \widehat{\chi}_k(h,D)\left\|h-h'\right\|_{\mathcal{H}} + \widetilde{O}\left(\sum_{k'=1}^{K}\left|\widehat{C}_k^{k'}\right|\sqrt{\frac{d_{\mathcal{H}} + \log\left({}^{K}/_{\delta}\right)}{np_{k'}}}\right) \ .$$

Proof. (Sketch) This lemma follows from bounding the two terms in the following inequality:

$$\left|F_k(h',D) - \widehat{F}_k(h,D)\right| \leq \left|\widehat{F}_k(h',D) - \widehat{F}_k(h,D)\right| + \left|F_k(h',D) - \widehat{F}_k(h',D)\right| \ .$$

The first term can be bounded using the empirical counterpart of Theorem 3.2. The second term is then bounded with high probability using standard uniform convergence bounds (Shalev-Shwartz and Ben-David, 2014). The complete proof can be found in Appendix D where the result is also extended to the simpler case where h and h' are fixed rather than learned on D. □

4 Bounding the Relative Fairness of Private Models

In this section, we quantify the difference of fairness between a private model and its non-private counterpart. Let $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{S} \times \mathcal{Y} \to \mathbb{R}$ be a loss function. Assume ℓ is Λ-Lipschitz, and μ-strongly-convex with respect to its first variable. Assume the norm $\|\cdot\|_{\mathcal{H}}$ is Euclidean, and that $\mathcal{H}$ is convex. We define the optimal model $h^* \in \mathcal{H}$ as

$$h^* = \operatorname*{arg\,min}_{h\in\mathcal{H}} f(h) = \frac{1}{n}\sum_{i=1}^{n}\ell(h; x_i, s_i, y_i) \ . \tag{3}$$

Two mechanisms are commonly used to find a differentially private approximation h^{priv} of h^*: output perturbation (Chaudhuri et al., 2011; Lowy and Razaviyayn, 2021), and DP-SGD (Bassily et al., 2014; Abadi et al., 2016). For both mechanisms, the distance $\left\|h^{\text{priv}} - h^*\right\|_{\mathcal{H}}$ can be upper bounded with high probability. In this section, we recall these two mechanisms and the corresponding high probability upper bounds. We then plug these bounds in Theorem 3.2 to bound the fairness level of the private solution h^{priv} relatively to the one of the true solution h^*.

4.1 Bounding the Distance between Private and Optimal Classifiers

Output perturbation. Output perturbation computes the non-private solution h^* of (3), and releases a private estimate by the Gaussian mechanism:

$$h^{\text{priv}} = \pi_{\mathcal{H}}(h^* + \mathcal{N}(\sigma^2\mathbb{I}_p)) \ ,$$

where $\pi_{\mathcal{H}}$ is the projection on $\mathcal{H}$. Let Δ be the sensitivity of the function $D \mapsto \arg\min_{w\in\mathcal{H}} f(w; D)$. In our setting, we have $\Delta = {}^{2\Lambda}/_{\mu n}$. Then, given $0 < \epsilon, \delta < 1$, h^{priv} is (ϵ,δ)-differentially private as long as $\sigma^2 \geq {}^{2\Delta^2\log(1.25/\delta)}/_{\epsilon^2}$. We bound the distance between h^{priv} and h^* with high probability in Lemma 4.1 (proved in Appendix E).

Lemma 4.1. *Let h^{priv} be the vector released by output perturbation with noise $\sigma^2 = {}^{8\Lambda^2\log(1.25/\delta)}/_{\mu^2 n^2\epsilon^2}$, and $0 < \zeta < 1$, then with probability at least $1-\zeta$,*

$$\left\|h^{\text{priv}} - h^*\right\|_2^2 \leq \frac{32p\Lambda^2\log(1.25/\delta)\log(2/\zeta)}{\mu^2 n^2\epsilon^2} \ .$$

DP-SGD. DP-SGD starts from some $h^0 \in \mathcal{H}$ and updates it using stochastic gradients. That is, with $\gamma > 0$, $i \sim \mathcal{U}([n])$, and $\eta^t \sim \mathcal{N}(0, \sigma^2\mathbb{I}_p)$, we iteratively update

$$h^{t+1} = \pi_{\mathcal{H}}(h^t - \gamma(\nabla\ell(h^t; x_i, y_i) + \eta^t)) \ .$$

After $T > 0$ iterations, we release $h^{\text{priv}} = h^T$. Given $0 < \epsilon, \delta < 1$, h^{priv} is (ϵ, δ)-differentially private when $\sigma^2 \geq 64\Lambda^2 T^2 \log(3T/\delta) \log(2/\delta) / n^2 \epsilon^2$. Assuming the loss function is smooth in its first parameter, we bound the distance between h^{priv} and h^* with high probability in Lemma 4.2 (proved in Appendix F).

Lemma 4.2. *Let h^{priv} be the vector released by DP-SGD with $\sigma^2 = 64\Lambda^2 T^2 \log(3T/\delta) \log(2/\delta) / n^2 \epsilon^2$. Assume that $\sigma_*^2 = \mathbb{E}_{i \sim [n]} \left\| \nabla \ell(h^*; x_i, y_i) \right\|^2 \leq \sigma^2$. Let $0 < \zeta < 1$, then with probability at least $1 - \zeta$,*

$$\left\| h^{\text{priv}} - h^* \right\|_2^2 = \widetilde{O}\left(\frac{p \Lambda^2 \log(1/\delta)^2}{\zeta \mu^2 n^2 \epsilon^2} \right) \ ,$$

where $\widetilde{O}$ ignores logarithmic terms in n (the number of examples) and p (the number of model parameters).

Remark 4.3. *For clarity of exposition in Lemma 4.2, we did not use minimal assumptions and used the simplest variant of DP-SGD. Notably, the assumption on σ_* can be removed by using variance reduction schemes, and tighter bounds on σ can also be obtained using Rényi Differential Privacy (Mironov, 2017). Similarly, the assumption $\epsilon < 1$ is only used to give simple closed-form bounds. Strong convexity and smoothness assumptions can be relaxed as well.*

Table 1: Upper bound, with 99% probability, on the difference of fairness between private and non-private models for different fairness measures and accuracy. Privacy parameters are $\epsilon = 1$ and $\delta = 1/n^2$ where n is the number of samples in the training data.

Dataset	Equality of Opportunity	Equalized Odds	Demographic Parity	Accuracy Parity	Accuracy
celebA ($n = 182,339$)	0.1044	0.0975	0.0975	0.0975	0.0487
folktables ($n = 1,498,050$)	0.0017	0.0026	0.0026	0.0026	0.0013

4.2 Bounding the Fairness of Private Models

We now state our central result (Theorem 4.4), where we bound the fairness of h^{priv} relatively to the one of h^*.

Theorem 4.4. *Let h^* be the solution of* (3)*, and h^{priv} its private estimate obtained by output perturbation. Let $h^{\text{ref}} \in \{h^{\text{priv}}, h^*\}$, and $0 < \zeta < 1$. Then, the difference of fairness of group $k \in [K]$ satisfies, with probability at least $1 - \zeta$,*

$$\left| F_k(h^{\text{priv}}, D) - F_k(h^*, D) \right| \leq \frac{\chi_k(h^{\text{ref}}, D) L \Lambda \sqrt{32 p \log(1.25/\delta) \log(2/\zeta)}}{\mu n \epsilon} \ .$$

Similarly, if h^{priv} is estimated through DP-SGD, we have that, with probability at least $1 - \zeta$,

$$\left| F_k(h^{\text{priv}}, D) - F_k(h^*, D) \right| \leq \widetilde{O}\left(\frac{\chi_k(h^{\text{ref}}, D) L \Lambda \sqrt{p \log(1/\delta)}}{\sqrt{\zeta} \mu n \epsilon} \right) \ ,$$

where $\widetilde{O}$ ignores logarithmic terms in n (the number of examples) and p (the number of model parameters).

Proof. By Lemma 4.1 or Lemma 4.2, we control the distance $\left\| h^{\text{priv}} - h^* \right\|$. Plugging this bound in Theorem 3.2 gives the result. □

This result shows that, when learning a private model, *the unfairness due to privacy vanishes at a $\widetilde{O}(\sqrt{p}/n)$ rate*. To the best of our knowledge, our result is the first to quantify this rate. Importantly, *it highlights the role of the confidence margin* of the classifier on the impact of differential privacy on fairness. This is in line with previous empirical and theoretical work that identified the groupwise distances to the decision boundary as an important factor (Tran et al., 2021; Fioretto et al., 2022). However, our bounds are the first to quantify this impact through a classic notion of confidence margin studied in learning theory (Cortes et al., 2013).

Our result may be interpreted and used in various ways. A first example is the case where the private model is known but its optimal non-private counterpart is not. There, our result guarantees that, given enough examples, the fairness level of the private model is close to the one of the optimal non-private model. This allows the practitioner to give guarantees on the model, that the end user can trust. A second example is the case where the true model h^* is owned by someone who cannot share it, due to privacy concerns. Imagine that the model needs to be audited for fairness. Then, the model owner can compute a private estimate of their model, and send it to the (honest but curious) auditing company. The bound allows to obtain fairness bounds for the true model from the inspection of the private one, and thus acts as a certificate of correctness of the audit done on the private version of the model.

Remark 4.5. *The fairness guarantee for the private model given by Theorem 4.4 is relative to the fairness of the optimal model h^*, which may itself be quite unfair. A standard approach to promote fair models is to use convex relaxations of fairness as regularizers to the ERM problem (Bechavod and Ligett, 2017; Huang and Vishnoi, 2019; Lohaus et al., 2020). Interestingly, to be able to use output perturbation, we only require the objective function of* (3) *to be strongly convex and Lipschitz over $h \in \mathcal{H}$, which is the case for these relaxations when they are combined with a squared ℓ_2-norm. For binary classification with two sensitive groups, Lohaus et al. (2020) proved that, with a proper choice of regularization parameters, this approach can yield a fair h^* (see their Theorem 1 for more details). Combined with our results, this paves the way for the design of algorithms that learn provably private and fair classifiers. However, several crucial challenges remain to make this approach work in practice, such as (i) finding the appropriate regularization parameters privately, and (ii) providing guarantees on the resulting classifiers' accuracy. We leave this for future work.*

5 Numerical Experiments

In this section, we numerically illustrate the upper bounds from Section 4.2. We use the `celebA` (Liu et al., 2015) and `folktables` (Ding et al., 2021) datasets, which respectively contain $202,599$ and $1,664,500$ samples, with 39 and 10 features (including one sensitive attribute, sex, that is not not used for prediction), and binary labels. For each dataset, we use 90% of the records for training, and the remaining 10% for empirical evaluation of the bounds. We train ℓ_2-regularized logistic regression models, ensuring that the underlying optimization problem is 1-strongly-convex. This allows learning private models by output perturbation, for which the bound from Theorem 4.4 holds.

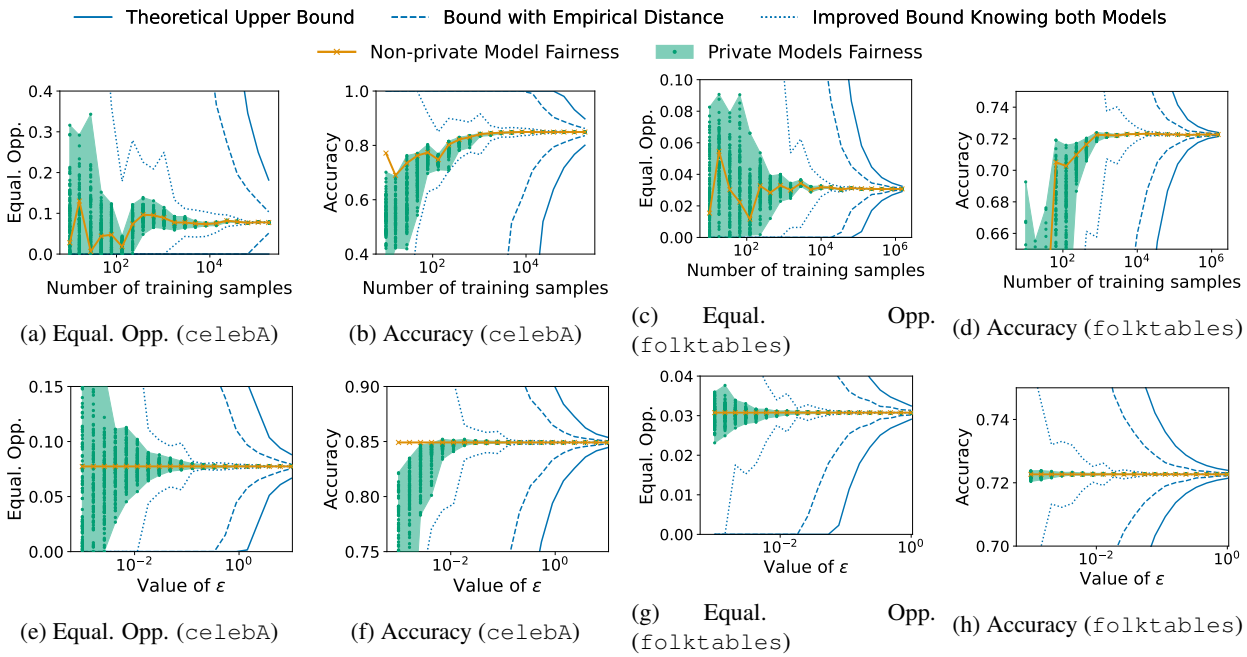
In Section 5.1, we show that we obtain non-trivial guarantees on the private model's fairness and accuracy. Then, we study the influence of the number of training samples and of the privacy budget ϵ in Section 5.2, and discuss the tightness of our result in Section 5.3.

5.1 Value of the Upper Bounds

In Table 1, we compute the value of Theorem 4.4's bounds. We learn a non-private ℓ_2-regularized logistic regression model, and use it to compute the bounds (averaged over the two groups) for multiple fairness and accuracy measures on two datasets. In all cases, our results give non-trivial guarantees on the difference of fairness: it is bounded by at most 0.105 for `celebA` and 0.0026 for `folktables`. This means that any $(1, 1/n^2)$-DP model learned by output perturbation will, with high probability, achieve a fairness level within this margin of that of the non-private model.

5.2 Influence of the Number of Training Samples and Privacy Budget

We now verify numerically the rate at which fairness and accuracy levels decrease when increasing the number of training records or privacy budget. In Figure 1, we plot the optimal model's equality of opportunity and accuracy, as a function of (i) in the first line, the number of samples n used for training, or (ii) in the second line, the privacy budget ϵ (see Appendix G for results with other fairness measures). For each value of n and ϵ, we plot Theorem 4.4's theoretical guarantees (solid blue line). With $\epsilon = 1$, our bounds give meaningful guarantees for $n \geq 100,000$ records on both `celebA` and `folktables` datasets (Figures 1a to 1d). When using all records, we obtain meaningful bounds for $\epsilon \geq 1$ for `celebA` and $\epsilon \geq 0.1$ for `folktables` (Figures 1e to 1h). Additionally, note that we obtain *both upper and lower bounds* on fairness and accuracy, confirming remarks from Section 3.2.



(a) Equal. Opp. (`celebA`) (b) Accuracy (`celebA`) (c) Equal. Opp. (`folktables`) (d) Accuracy (`folktables`)

(e) Equal. Opp. (`celebA`) (f) Accuracy (`celebA`) (g) Equal. Opp. (`folktables`) (h) Accuracy (`folktables`)

Figure 1: Equality of opportunity (Equal. Opp.) and Accuracy levels for optimal non-private model and random private ones as a function of the number of training records n (first line, with $\epsilon = 1$ and $\delta = 1/n^2$) and of the privacy budget ϵ (second line, using all available training records). For each value of n and ϵ, we sample 100 private models and take their minimum and maximum fairness/accuracy values to mark the area of attainable values. The solid blue line gives the theoretical guarantees from Theorem 4.4, while the dashed and dotted line give finer bounds when more information is available (see Section 5.3 for details).

We also report the fairness and accuracy levels of 100 private models computed by output perturbation (in green in Figure 1). As predicted by our theory, their fairness and accuracy converges towards the ones of their non-private counterparts as n and ϵ increase. Interestingly, our bounds seem to follow the same tendency as what we observe empirically (albeit with a larger multiplicative constant), suggesting that they capture the correct dependence in n and ϵ. We further discuss the tightness of our results in next section.

5.3 Tightness of the Bound

We now argue that the two major factors of looseness in our results are (i) the upper bound on $\left\|h^{\text{priv}} - h^*\right\|$ and (ii) the looseness of Assumption 2.1. While these cannot be improved in general, specific knowledge of h^{priv} and h^* (that is typically not available due to privacy) can lead to tighter bounds. First, when the distance $\left\|h^{\text{priv}} - h^*\right\|$ is known, we can use its actual value rather than the upper bounds of Section 4.1 (see dashed blue line in Figure 1). Second, when both h^{priv} and h^* are known, Assumption 2.1 can be substantially refined (see details in Appendix G.3). We evaluate this bound for the private model that is the farthest away from the non-private one (see dotted blue line in Figure 1). The resulting bound appears to be tight up to a small multiplicative constant. These two observations suggest that our bounds cannot be significantly tightened, unless one can obtain such knowledge through either private computation or additional assumptions on the data.

6 Conclusion

In this work, we proved that the fairness (and accuracy) costs induced by privacy in differentially private classification vanishes at a $\widetilde{O}(\sqrt{p}/n)$ rate, where n is the number of training records, and p the number of parameters. This rate follows

from a general statement on group fairness measures' regularity, that we prove to be pointwise Lipschitz with respect to the model. The pointwise Lipschitz constant explicitly depends on the confidence margin of the model, and we show it can be computed from a finite data sample. Importantly, our bounds does not require the knowledge of the optimal (non-private) model: they can thus be used in practical privacy-preserving scenarios. We numerically evaluate our bounds on real datasets, and highlight practical settings where non-trivial guarantees can be obtained.

Our results could help build more trustworthy machine learning models, by guaranteeing that their fairness and accuracy approximately match the one of the non-private model. We believe that our results are applicable to privacy-preserving methods beyond output perturbation and DP-SGD. Indeed, deriving high-probability bounds on the distance between the private and the non-private model is sufficient to apply them. Note however that our bounds crucially rely on the uniqueness of problem (3)'s solution, which is guaranteed by strong convexity. Relaxing this hypothesis is challenging, but would greatly broaden the scope of our results.

We stress that our results do not provide fairness guarantees *per se*, but only bound the difference of fairness between models. It is nonetheless a first step towards a more complete understanding of the interplay between privacy, fairness, and accuracy. We believe that our results can guide the design of fairer privacy-preserving machine learning algorithms. A first promising direction in this regard is to combine our bounds with fairness-promoting convex regularizers, as discussed in Remark 4.5. Another direction is the design of methods to privately learn models with large-margin guarantees, as recently considered by Bassily et al. (2022). Our results, which explicitly depend on the confidence margin of the model, suggest that better fairness guarantees could be obtained for these methods.

Acknowledgements

This work was supported by the Région Hauts de France (Projet STaRS Equité en apprentissage décentralisé respectueux de la vie privée) and by the Inria Exploratory Action FLAMED.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, pages 308–318, New York, NY, USA. Association for Computing Machinery. 1130 citations (Crossref) [2022-08-19].

Agarwal, S. (2020). Trade-offs between fairness and privacy in machine learning.

Bagdasaryan, E., Poursaeed, O., and Shmatikov, V. (2019). Differential privacy has disparate impact on model accuracy. In *Advances in Neural Information Processing Systems 32, 2019, Vancouver, BC, Canada*, pages 15453–15462.

Bassily, R., Mohri, M., and Suresh, A. T. (2022). Differentially private learning with margin guarantees. *arXiv preprint arXiv:2204.10376*.

Bassily, R., Smith, A., and Thakurta, A. (2014). Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473, Philadelphia, PA, USA. IEEE.

Bechavod, Y. and Ligett, K. (2017). Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*.

Beimel, A., Brenner, H., Kasiviswanathan, S. P., and Nissim, K. (2014). Bounds on the sample complexity for private learning and private data release. *Machine Learning*, pages 401–437.

Calders, T., Kamiran, F., and Pechenizkiy, M. (2009). Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE.

Caton, S. and Haas, C. (2020). Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*.

Chang, H. and Shokri, R. (2020). On the privacy risks of algorithmic fairness. *arXiv preprint arXiv:2011.03731*.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. (2011). Differentially Private Empirical Risk Minimization. *Journal of Machine Learning Research*, 12(29):1069–1109.

Cortes, C., Mohri, M., and Rostamizadeh, A. (2013). Multi-class classification with maximum margin multiple kernel. In *International Conference on Machine Learning*, pages 46–54. PMLR.

Cummings, R., Gupta, V., Kimpara, D., and Morgenstern, J. (2019). On the compatibility of privacy and fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pages 309–315.

Ding, F., Hardt, M., Miller, J., and Schmidt, L. (2021). Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34.

Dwork, C. (2006). Differential privacy. In *Encyclopedia of Cryptography and Security*.

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

Dwork, C., Roth, A., et al. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407.

Farrand, T., Mireshghallah, F., Singh, S., and Trask, A. (2020). Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy. *arXiv preprint arXiv:2009.06389*.

Fioretto, F., Tran, C., Hentenryck, P. V., and Zhu, K. (2022). Differential privacy and fairness in decisions and learning tasks: A survey. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5470–5477.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323.

Huang, L. and Vishnoi, N. (2019). Stable and fair classification. In *International Conference on Machine Learning*, pages 2879–2890. PMLR.

Jagielski, M., Kearns, M., Mao, J., Oprea, A., Roth, A., Sharifi-Malvajerdi, S., and Ullman, J. (2019). Differentially private fair learning. In *International Conference on Machine Learning*, pages 3000–3008. PMLR.

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith, A. (2011). What Can We Learn Privately? *SIAM Journal on Computing*, pages 793–826.

Kiefer, J. (1953). Sequential minimax search for a maximum. *Proceedings of the American mathematical society*, 4(3):502–506.

Kilbertus, N., Gascón, A., Kusner, M., Veale, M., Gummadi, K., and Weller, A. (2018). Blind justice: Fairness with encrypted sensitive attributes. In *International Conference on Machine Learning*, pages 2630–2639. PMLR.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. *Advances in neural information processing systems*, 30.

Liu, B., Ding, M., Shaham, S., Rahayu, W., Farokhi, F., and Lin, Z. (2021). When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Lohaus, M., Perrot, M., and Von Luxburg, U. (2020). Too relaxed to be fair. In *International Conference on Machine Learning*, pages 6360–6369. PMLR.

Lowy, A. and Razaviyayn, M. (2021). Output perturbation for differentially private convex optimization with improved population loss bounds, runtimes and applications to private adversarial training. *arXiv preprint arXiv:2102.04704*.

Maheshwari, G. and Perrot, M. (2022). Fairgrad: Fairness aware gradient descent. *arXiv preprint arXiv:2206.10923*.

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

Mironov, I. (2017). Renyi Differential Privacy. *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*.

Mozannar, H., Ohannessian, M., and Srebro, N. (2020). Fair learning with private demographic data. In *International Conference on Machine Learning*, pages 7066–7075. PMLR.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pujol, D., McKenna, R., Kuppam, S., Hay, M., Machanavajjhala, A., and Miklau, G. (2020). Fair decision making using privacy-protected data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 189–199, New York, NY, USA. Association for Computing Machinery.

Sanyal, A., Hu, Y., and Yang, F. (2022). How unfair is private learning? *arXiv preprint arXiv:2206.03985*.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Song, S., Chaudhuri, K., and Sarwate, A. D. (2013). Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 245–248, Austin, TX, USA. IEEE. 145 citations (Crossref) [2022-08-19].

Tran, C., Dinh, M., and Fioretto, F. (2021). Differentially private empirical risk minimization under the fairness lens. *Advances in Neural Information Processing Systems*, 34:27555–27565.

Tran, C., Fioretto, F., and Van Hentenryck, P. (2020). Differentially private and fair deep learning: A lagrangian dual approach. *arXiv preprint arXiv:2009.12562*.

Uniyal, A., Naidu, R., Kotti, S., Singh, S., Kenfack, P. J., Mireshghallah, F., and Trask, A. (2021). Dp-sgd vs pate: Which has less disparate impact on model accuracy? *arXiv preprint arXiv:2106.12576*.

Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N. (2017). Learning non-discriminatory predictors. In *Conference on Learning Theory*, pages 1920–1953. PMLR.

Xu, D., Du, W., and Wu, X. (2020). Removing disparate impact of differentially private stochastic gradient descent on model accuracy. *arXiv preprint arXiv:2003.03699*.

Xu, D., Yuan, S., and Wu, X. (2019). Achieving differential privacy and fairness in logistic regression. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 594–599.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180.

This appendix provides several examples of group fairness functions compatible with our framework (Appendix A), the proofs of the main theoretical results that were omitted in the main paper for the sake of readability (Appendices B to F), and additional experiments (Appendix G).

A Fairness functions

In this section we recall several well known fairness functions and show that they can be written in the form of Equation (1).

Example 1 (Equalized Odds (Hardt et al., 2016)). *A model h is fair for Equalized Odds when the probability of predicting the correct label is independent of the sensitive attribute, that is, $\forall (y, r) \in \mathcal{Y} \times \mathcal{S}$*

$$F_{(y,r)}(h, D) = \mathbb{P}\left(H(X) = Y \mid Y = y, S = r\right) - \mathbb{P}\left(H(X) = Y \mid Y = y\right).$$

We can then write $F_{(y,r)}(h, D)$ in the form of Equation (1) *as*

$$F_{(y,r)}(h, D) = C^0_{(y,r)} + \sum_{(y',r') \in \mathcal{Y} \times \mathcal{S}} C^{(y',r')}_{(y,r)} \, \mathbb{P}\left(H(x) = Y | Y = y', S = r'\right) \tag{4}$$

with

$$\begin{aligned} C^0_{(y,r)} &= 0 \\ C^{(y,r)}_{(y,r)} &= 1 - \mathbb{P}\left(S = r \mid Y = y\right) \\ \forall r' \neq r,\, C^{(y,r')}_{(y,r)} &= -\mathbb{P}\left(S = r' \mid Y = y\right) \\ \forall y' \neq y,\, \forall r' \in \mathcal{S},\, C^{(y',r')}_{(y,r)} &= 0 \end{aligned}$$

Proof. We have that

$$\begin{aligned} F_{(y,r)}(h, D) &= \mathbb{P}\left(H(X) = Y \mid Y = y, S = r\right) - \mathbb{P}\left(H(X) = Y \mid Y = y\right) \\ &= \mathbb{P}\left(H(X) = Y \mid Y = y, S = r\right) - \sum_{r' \in \mathcal{S}} \mathbb{P}\left(H(X) = Y \mid Y = y, S = r'\right) \mathbb{P}\left(S = r' \mid Y = y\right) \end{aligned}$$

which gives the result. □

Example 2 (Equality of Opportunity Hardt et al. (2016)). *A model h is fair for Equality of Opportunity when the probability of predicting the correct label is independent of the sensitive attribute for the set of desirable outcomes $\mathcal{Y}' \subset \mathcal{Y}$, that is $\forall (y, r) \in \mathcal{Y} \times \mathcal{S}$*

$$F_{(y,r)}(h, D) = \begin{cases} \mathbb{P}\left(H(X) = Y \mid Y = y, S = r\right) - \mathbb{P}\left(H(X) = Y \mid Y = y\right) & \text{if } y \in \mathcal{Y}', \\ 0 & \text{otherwise.} \end{cases}$$

We can then write $F_{(y,r)}(h, D)$ in the form of Equation (1) *as*

$$F_{(y,r)}(h, D) = C^0_{(y,r)} + \sum_{(y',r') \in \mathcal{Y} \times \mathcal{S}} C^{(y',r')}_{(y,r)} \, \mathbb{P}\left(H(X) = Y | Y = y', S = r'\right) \tag{5}$$

with, if $y \in \mathcal{Y}'$,

$$\begin{aligned} C^0_{(y,r)} &= 0 \\ C^{(y,r)}_{(y,r)} &= 1 - \mathbb{P}\left(S = r \mid Y = y\right) \end{aligned}$$

$$\forall r' \neq r,\, C^{(y,r')}_{(y,r)} = -\mathbb{P}\left(S = r' \mid Y = y\right)$$
$$\forall y' \neq y,\, \forall r' \in \mathcal{S},\, C^{(y',r')}_{(y,r)} = 0$$

and, if $y \in \mathcal{Y} \setminus \mathcal{Y}'$*,*

$$\forall y' \in \mathcal{Y},\, \forall r' \in \mathcal{S},\, C^{(y',r')}_{(y,r)} = 0.$$

Proof. We consider the two cases. On the one hand, when $y \in \mathcal{Y} \setminus \mathcal{Y}'$, then we have that

$$F_{(y,r)}(h, D) = 0$$

which gives the first part of the result. On the other hand, when $y \in \mathcal{Y}'$, then we have that

$$\begin{aligned} F_{(y,r)}(h, D) &= \mathbb{P}\left(H(X) = Y \mid Y = y, S = r\right) - \mathbb{P}\left(H(X) = Y \mid Y = y\right) \\ &= \mathbb{P}\left(H(X) = Y \mid Y = y, S = r\right) - \sum_{r' \in \mathcal{S}} \mathbb{P}\left(H(X) = Y \mid Y = y, S = r'\right) \mathbb{P}\left(S = r' \mid Y = y\right) \end{aligned}$$

which gives the second part of the result. □

Example 3 (Accuracy Parity Zafar et al. (2017)). *A model* h *is fair for Accuracy Parity when the probability of being correct is independent of the sensitive attribute, that is,* $\forall (r) \in \mathcal{S}$

$$F_{(r)}(h, D) = \mathbb{P}\left(H(X) = Y \mid S = r\right) - \mathbb{P}\left(H(X) = Y\right).$$

We can then write $F_{(r)}(h, D)$ *in the form of Equation* (1) *as*

$$F_{(r)}(h, D) = C^0_{(r)} + \sum_{(r') \in \mathcal{S}} C^{(r')}_{(r)}\, \mathbb{P}\left(H(X) = Y | S = r'\right) \tag{6}$$

with

$$\begin{aligned} C^0_{(r)} &= 0 \\ C^{(r)}_{(r)} &= 1 - \mathbb{P}\left(S = r\right) \\ \forall r' \neq r,\, C^{(r')}_{(r)} &= -\mathbb{P}\left(S = r'\right) \end{aligned}$$

Proof. We have that

$$\begin{aligned} F_{(r)}(h, D) &= \mathbb{P}\left(H(X) = Y \mid S = r\right) - \mathbb{P}\left(H(X) = Y\right) \\ &= \mathbb{P}\left(H(X) = Y \mid S = r\right) - \sum_{r' \in \mathcal{S}} \mathbb{P}\left(H(X) = Y \mid S = r'\right) \mathbb{P}\left(S = r'\right) \end{aligned}$$

which gives the result. □

Example 4 (Demographic Parity (Binary Labels) Calders et al. (2009)). *A model* h *is fair for Demographic Parity with binary labels when the probability of predicting a label is independent of the sensitive attribute, that is,* $\forall (y, r) \in \mathcal{Y} \times \mathcal{S}$

$$F_{(y,r)}(h, D) = \mathbb{P}\left(H(X) = y \mid S = r\right) - \mathbb{P}\left(H(X) = y\right).$$

Assuming that given a label y*, the second binary label is denoted* $\bar{y}$*, we can then write* $F_{(y,r)}(h, D)$ *in the form of Equation* (1) *as*

$$F_{(y,r)}(h, D) = C^0_{(y,r)} + \sum_{(y',r') \in \mathcal{Y} \times \mathcal{S}} C^{(y',r')}_{(y,r)}\, \mathbb{P}\left(H(X) = Y | Y = y', S = r'\right) \tag{7}$$

with

$$
\begin{aligned}
C^{0}_{(y,r)} &= \mathbb{P}\left(Y = y\right) - \mathbb{P}\left(Y = y \mid S = r\right) \\
C^{(y,r)}_{(y,r)} &= \mathbb{P}\left(Y = y \mid S = r\right) - \mathbb{P}\left(Y = y, S = r\right) \\
C^{(\bar{y},r)}_{(y,r)} &= \mathbb{P}\left(Y = \bar{y}, S = r\right) - \mathbb{P}\left(Y = \bar{y} \mid S = r\right) \\
\forall r' \neq r,\, C^{(y,r')}_{(y,r)} &= -\mathbb{P}\left(Y = y, S = r'\right) \\
\forall r' \neq r,\, C^{(\bar{y},r')}_{(y,r)} &= \mathbb{P}\left(Y = \bar{y}, S = r'\right)
\end{aligned}
$$

Proof. We have that

$$
\begin{aligned}
F_{(y,r)}(h, D) &= \mathbb{P}\left(H(X) = y \mid S = r\right) - \mathbb{P}\left(H(X) = y\right) \\
&= \mathbb{P}\left(H(X) = y \mid Y = y, S = r\right) \mathbb{P}\left(Y = y \mid S = r\right) + \mathbb{P}\left(H(X) = y \mid Y \neq y, S = r\right) \mathbb{P}\left(Y \neq y \mid S = r\right) \\
&\quad - \sum_{r' \in \mathcal{S}} \big(\mathbb{P}\left(H(X) = y \mid Y = y, S = r'\right) \mathbb{P}\left(Y = y, S = r'\right) \\
&\qquad + \mathbb{P}\left(H(X) = y \mid Y \neq y, S = r'\right) \mathbb{P}\left(Y \neq y, S = r'\right) \big) \\
&= \mathbb{P}\left(H(X) = y \mid Y = y, S = r\right) \mathbb{P}\left(Y = y \mid S = r\right) \\
&\quad + 1 - \mathbb{P}\left(H(X) \neq y \mid Y \neq y, S = r\right) \mathbb{P}\left(Y \neq y \mid S = r\right) \\
&\quad - \sum_{r' \in \mathcal{S}} \big(\mathbb{P}\left(H(X) = y \mid Y = y, S = r'\right) \mathbb{P}\left(Y = y, S = r'\right) \\
&\qquad + 1 - \mathbb{P}\left(H(X) \neq y \mid Y \neq y, S = r'\right) \mathbb{P}\left(Y \neq y, S = r'\right) \big).
\end{aligned}
$$

Here, we only consider binary labels, y and $\bar{y}$. Hence, $H(X) \neq y \Leftrightarrow H(X) = \bar{y}$ and $Y \neq y \Leftrightarrow Y = \bar{y}$. Thus, we obtain

$$
\begin{aligned}
&F_{(y,r)}(h, D) \\
&\quad = \mathbb{P}\left(H(X) = y \mid Y = y, S = r\right) \mathbb{P}\left(Y = y \mid S = r\right) + \left(1 - \mathbb{P}\left(H(X) = \bar{y} \mid Y = \bar{y}, S = r\right)\right) \mathbb{P}\left(Y = \bar{y} \mid S = r\right) \\
&\qquad - \sum_{r' \in \mathcal{S}} \big(\mathbb{P}\left(H(X) = y \mid Y = y, S = r'\right) \mathbb{P}\left(Y = y, S = r'\right) \\
&\qquad\quad + \left(1 - \mathbb{P}\left(H(X) = \bar{y} \mid Y = \bar{y}, S = r'\right)\right) \mathbb{P}\left(Y = \bar{y}, S = r'\right) \big) \\
&\quad = \mathbb{P}\left(H(X) = y \mid Y = y, S = r\right) \left[\mathbb{P}\left(Y = y \mid S = r\right) - \mathbb{P}\left(Y = y, S = r\right)\right] \\
&\qquad + \mathbb{P}\left(H(X) = \bar{y} \mid Y = \bar{y}, S = r\right) \left[\mathbb{P}\left(Y = \bar{y}, S = r\right) - \mathbb{P}\left(Y = \bar{y} \mid S = r\right)\right] \\
&\qquad + \sum_{r' \in \mathcal{S}, r' \neq r} \mathbb{P}\left(H(X) = y \mid Y = y, S = r'\right) \left(-\mathbb{P}\left(Y = y, S = r'\right)\right) \\
&\qquad + \sum_{r' \in \mathcal{S}, r' \neq r} \mathbb{P}\left(H(X) = \bar{y} \mid Y = \bar{y}, S = r'\right) \mathbb{P}\left(Y = \bar{y}, S = r'\right) \\
&\qquad + \mathbb{P}\left(Y = \bar{y} \mid S = r\right) - \mathbb{P}\left(Y = \bar{y}\right) \\
&\quad = \mathbb{P}\left(H(X) = Y \mid Y = y, S = r\right) \left[\mathbb{P}\left(Y = y \mid S = r\right) - \mathbb{P}\left(Y = y, S = r\right)\right] \\
&\qquad + \mathbb{P}\left(H(X) = Y \mid Y = \bar{y}, S = r\right) \left[\mathbb{P}\left(Y = \bar{y}, S = r\right) - \mathbb{P}\left(Y = \bar{y} \mid S = r\right)\right] \\
&\qquad + \sum_{r' \in \mathcal{S}, r' \neq r} \mathbb{P}\left(H(X) = Y \mid Y = y, S = r'\right) \left(-\mathbb{P}\left(Y = y, S = r'\right)\right) \\
&\qquad + \sum_{r' \in \mathcal{S}, r' \neq r} \mathbb{P}\left(H(X) = Y \mid Y = \bar{y}, S = r'\right) \mathbb{P}\left(Y = \bar{y}, S = r'\right) \\
&\qquad + \mathbb{P}\left(Y = y\right) - \mathbb{P}\left(Y = y \mid S = r\right)
\end{aligned}
$$

which gives the result. □

B Proof of Theorem 3.1

Theorem (Pointwise Lipschitzness of Conditional Negative Predictions). *Let $\mathcal{H}$ be a set of real vector-valued functions with $L_{X,Y}$ the Lipschitz constants defined in Assumption 2.1. Let $h, h' \in \mathcal{H}$ be two models, (X, Y, S) be a triple of random variables having distribution $\mathcal{D}$, and E be an arbitrary event. Assume that $\mathbb{E}\left(\frac{L_{X,Y}}{|\rho(h,X,Y)|} \;\middle|\; E\right) < +\infty$, then*

$$|\mathbb{P}(H(X) = Y \mid E) - \mathbb{P}(H'(X) = Y \mid E)| \leq \mathbb{E}\left(\frac{L_{X,Y}}{|\rho(h,X,Y)|} \;\middle|\; E\right) \|h - h'\|_{\mathcal{H}} \quad .$$

Proof. The proof of this theorem is in two steps. First, we use the Lipschitz continuity property associated with $\mathcal{H}$, the triangle inequality, and the union bound to show that $|\mathbb{P}\left(H(X) = Y \mid E\right) - \mathbb{P}\left(H'(X) = Y \mid E\right)| \leq \mathbb{P}\left(\frac{L_{X,Y}}{|\rho(h,X,Y)|} \leq \|h - h'\|_{\mathcal{H}} \;\middle|\; E\right)$. Then, applying Markov's inequality gives the desired result.

Bounding $|\mathbb{P}\left(H(X) = Y \mid E\right) - \mathbb{P}\left(H'(X) = Y \mid E\right)|$. We have that

$$\begin{aligned}
&\mathbb{P}\left(H(X) = Y \mid E\right) - \mathbb{P}\left(H'(X) = Y \mid E\right) \\
&\qquad \leq \mathbb{P}\left(\rho(h,X,Y) \geq 0 \mid E\right) - \mathbb{P}\left(\rho(h',X,Y) > 0 \mid E\right) \\
&\qquad = \mathbb{P}\left(\rho(h',X,Y) \leq 0 \mid E\right) - \mathbb{P}\left(\rho(h,X,Y) < 0 \mid E\right) \\
&\qquad = \mathbb{P}\left(\rho(h',X,Y) - \rho(h,X,Y) + \rho(h,X,Y) \leq 0 \mid E\right) - \mathbb{P}\left(\rho(h,X,Y) < 0 \mid E\right) \\
&\qquad = \mathbb{P}\left(\rho(h,X,Y) \leq \rho(h,X,Y) - \rho(h',X,Y) \mid E\right) - \mathbb{P}\left(\rho(h,X,Y) < 0 \mid E\right) \\
&\qquad \leq \mathbb{P}\left(\rho(h,X,Y) \leq |\rho(h,X,Y) - \rho(h',X,Y)| \mid E\right) - \mathbb{P}\left(\rho(h,X,Y) < 0 \mid E\right) \\
&\qquad\quad \downarrow \text{ Assumption 2.1.} \\
&\qquad \leq \mathbb{P}\left(\rho(h,X,Y) \leq L_{X,Y}\|h - h'\|_{\mathcal{H}} \mid E\right) - \mathbb{P}\left(\rho(h,X,Y) < 0 \mid E\right) \\
&\qquad = \mathbb{P}\left(\rho(h,X,Y) < 0 \bigcup 0 \leq \rho(h,X,Y) \leq L_{X,Y}\|h - h'\|_{\mathcal{H}} \;\middle|\; E\right) - \mathbb{P}\left(\rho(h,X,Y) < 0 \mid E\right) \\
&\qquad\quad \downarrow \text{ Union bound on disjoint events.} \\
&\qquad = \mathbb{P}\left(0 \leq \rho(h,X,Y) \leq L_{X,Y}\|h - h'\|_{\mathcal{H}} \mid E\right) \\
&\qquad \leq \mathbb{P}\left(|\rho(h,X,Y)| \leq L_{X,Y}\|h - h'\|_{\mathcal{H}} \mid E\right) \\
&\qquad = \mathbb{P}\left(\frac{|\rho(h,X,Y)|}{L_{X,Y}} \leq \|h - h'\|_{\mathcal{H}} \;\middle|\; E\right)
\end{aligned}$$

Similarly, we have that

$$\begin{aligned}
&\mathbb{P}\left(H'(X) = Y \mid E\right) - \mathbb{P}\left(H(X) = Y \mid E\right) \\
&\qquad \leq \mathbb{P}\left(\rho(h',X,Y) \geq 0 \mid E\right) - \mathbb{P}\left(\rho(h,X,Y) > 0 \mid E\right) \\
&\qquad = \mathbb{P}\left(\rho(h',X,Y) + \rho(h,X,Y) - \rho(h,X,Y) \geq 0 \mid E\right) - \mathbb{P}\left(\rho(h,X,Y) > 0 \mid E\right) \\
&\qquad = \mathbb{P}\left(\rho(h,X,Y) \geq -\left(\rho(h',X,Y) - \rho(h,X,Y)\right) \mid E\right) - \mathbb{P}\left(\rho(h,X,Y) > 0 \mid E\right) \\
&\qquad \leq \mathbb{P}\left(\rho(h,X,Y) \geq -|\rho(h',X,Y) - \rho(h,X,Y)| \mid E\right) - \mathbb{P}\left(\rho(h,X,Y) > 0 \mid E\right) \\
&\qquad\quad \downarrow \text{ Assumption 2.1} \\
&\qquad \leq \mathbb{P}\left(\rho(h,X,Y) \geq -L_{X,Y}\|h - h'\|_{\mathcal{H}} \mid E\right) - \mathbb{P}\left(\rho(h,X,Y) > 0 \mid E\right) \\
&\qquad = \mathbb{P}\left(\rho(h,X,Y) > 0 \bigcup 0 \geq \rho(h,X,Y) \geq -L_{X,Y}\|h - h'\|_{\mathcal{H}} \;\middle|\; E\right) - \mathbb{P}\left(\rho(h,X,Y) > 0 \mid E\right) \\
&\qquad\quad \downarrow \text{ Union bound on disjoint events.} \\
&\qquad = \mathbb{P}\left(0 \geq \rho(h,X,Y) \geq -L_{X,Y}\|h - h'\|_{\mathcal{H}} \mid E\right) \\
&\qquad \leq \mathbb{P}\left(-|\rho(h,X,Y)| \geq -L_{X,Y}\|h - h'\|_{\mathcal{H}} \mid E\right) \\
&\qquad = \mathbb{P}\left(\frac{|\rho(h,X,Y)|}{L_{X,Y}} \leq \|h - h'\|_{\mathcal{H}} \;\middle|\; E\right)
\end{aligned}$$

It implies that

$$|\mathbb{P}(H(X) = Y \mid E) - \mathbb{P}(H'(X) = Y \mid E)| \leq \mathbb{P}\left(\frac{|\rho(h, X, Y)|}{L_{X,Y}} \leq \|h - h'\|_{\mathcal{H}} \;\middle|\; E\right)$$

Bounding $\mathbb{P}\left(\frac{|\rho(h,X,Y)|}{L_{X,Y}} \leq \|h - h'\|_{\mathcal{H}} \;\middle|\; E\right)$. We use the Markov's Inequality and we assume that $\mathbb{E}\left(\frac{L_{X,Y}}{|\rho(h,X,Y)|} \;\middle|\; E\right) < +\infty$. Hence, we have that

$$\begin{aligned}\mathbb{P}\left(\frac{|\rho(h, X, Y)|}{L_{X,Y}} \leq \|h - h'\|_{\mathcal{H}} \;\middle|\; E\right) &= \mathbb{P}\left(\frac{L_{X,Y}}{|\rho(h, X, Y)|} \geq \frac{1}{\|h - h'\|_{\mathcal{H}}} \;\middle|\; E\right) \\ &\quad \downarrow \text{ Markov's inequality.} \\ &\leq \mathbb{E}\left(\frac{L_{X,Y}}{|\rho(h, X, Y)|} \;\middle|\; E\right) \|h - h'\|_{\mathcal{H}}\end{aligned}$$

It concludes the proof. □

Remark B.1. *In the last step of the proof of Theorem 3.1, we can also use the Chernoff bound:*

$$\begin{aligned}\mathbb{P}\left(\frac{|\rho(h, X, Y)|}{L_{X,Y}} \leq \|h - h'\|_{\mathcal{H}} \;\middle|\; E\right) &= \mathbb{P}\left(\exp\left(-t\frac{|\rho(h, X, Y)|}{L_{X,Y}}\right) \geq \exp\left(-t\,\|h - h'\|_{\mathcal{H}}\right) \;\middle|\; E\right) \\ &\leq \mathbb{E}\left(\exp\left(-t\frac{|\rho(h, X, Y)|}{L_{X,Y}}\right) \;\middle|\; E\right) \exp\left(t\,\|h - h'\|_{\mathcal{H}}\right) \quad .\end{aligned}$$

A correct choice of t would lead to potentially tighter bounds than the Markov's inequality at the expense of readability.

Remark B.2. *Before using Markov's inequality or Chernoff bound in Theorem 3.1, we can modify the probability as*

$$\mathbb{P}\left(\frac{|\rho(h, X, Y)|}{L_{X,Y}} \leq \|h - h'\|_{\mathcal{H}} \;\middle|\; E\right) = \mathbb{P}\left(\left[\frac{|\rho(h, X, Y)|}{L_{X,Y}}\right]^{\|h-h'\|_{\mathcal{H}}} \leq \|h - h'\|_{\mathcal{H}} \;\middle|\; E\right) \quad ,$$

where

$$\left[\frac{|\rho(h, X, Y)|}{L_{X,Y}}\right]^{\|h-h'\|_{\mathcal{H}}} = \begin{cases} \frac{|\rho(h,X,Y)|}{L_{X,Y}} & \text{if } |\rho(h, X, Y)| \leq L_{X,Y}\,\|h - h'\|_{\mathcal{H}} \quad , \\ +\infty & \text{otherwise} \quad . \end{cases}$$

This essentially means that, whenever the model's margin on a data record is large enough, its precise value is no more meaningful, as its prediction will not change whatsoever. The remaining of Theorem 3.1's proof is unchanged, except with $\left[\frac{|\rho(h,X,Y)|}{L_{X,Y}}\right]^{\|h-h'\|_{\mathcal{H}}}$ instead of $\frac{|\rho(h,X,Y)|}{L_{X,Y}}$.

Note that this can lead to much tighter bounds. Notably, when distance $\|h - h'\|_{\mathcal{H}}$ between h and h' is small enough, the difference of fairness may even become zero.

C Proof of Theorem 3.2

Theorem (Pointwise Lipschitzness of Fairness). *Let $h, h' \in \mathcal{H}$, $L_{X,Y}$ be defined as in Assumption 2.1, and $(X, S, Y) \sim \mathcal{D}$. For any fairness notion of the form of Equation* (1)*, we have:*

$$\forall k \in [K], |F_k(h, D) - F_k(h', D)| \leq \chi_k(h, D)\,\|h - h'\|_{\mathcal{H}} \quad ,$$

with $\chi_k(h, D) = \sum_{k'=1}^{K} \left|C_k^{k'}\right| \mathbb{E}\left(\frac{1}{|h(X)|} \;\middle|\; D_{k'}\right)$. Similarly, for the aggregate measure of fairness defined in Equation (2)*, we have:*

$$|\mathit{Fair}(h, D) - \mathit{Fair}(h', D)| \leq \frac{1}{K}\sum_{k=1}^{K} \chi_k(h, D)\,\|h - h'\|_{\mathcal{H}} \quad .$$

Proof. The first part follows from the following derivation. For all k,

$$
\begin{aligned}
|F_k(h,D) - F_k(h',D)| &= \left| C_k^0 + \sum_{k'=1}^{K} C_k^{k'} \, \mathbb{P}\left(H(X) = Y | D_{k'}\right) - C_k^0 - \sum_{k'=1}^{K} C_k^{k'} \, \mathbb{P}\left(H'(X) = Y | D_{k'}\right) \right| \\
&= \left| \sum_{k'=1}^{K} C_k^{k'} \Big(\mathbb{P}\left(H(X) = Y | D_{k'}\right) - \mathbb{P}\left(H'(X) = Y | D_{k'}\right) \Big) \right| \\
&\quad \downarrow \text{ Triangle inequality.} \\
&\leq \sum_{k'=1}^{K} \left| C_k^{k'} \right| \left| \mathbb{P}\left(H(X) = Y | D_{k'}\right) - \mathbb{P}\left(H'(X) = Y | D_{k'}\right) \right| \\
&\quad \downarrow \text{ Theorem 3.1.} \\
&\leq \sum_{k'=1}^{K} \left| C_k^{k'} \right| \mathbb{E}\left(\frac{L_{X,Y}}{|\rho(h,X,Y)|} \,\middle|\, D_{k'} \right) \|h - h'\|_{\mathcal{H}} \quad .
\end{aligned}
$$

The second part is obtained thanks to the triangle inequality:

$$
\begin{aligned}
|\text{Fair}(h,D) - \text{Fair}(h',D)| &= \left| \frac{1}{K} \sum_{k=1}^{K} |F_k(h,D)| - \frac{1}{K} \sum_{k=1}^{K} |F_k(h',D)| \right| \\
&\quad \downarrow \text{ Triangle inequality.} \\
&\leq \frac{1}{K} \sum_{k=1}^{K} \big| |F_k(h,D)| - |F_k(h',D)| \big| \\
&\quad \downarrow \text{ Reverse triangle inequality.} \\
&\leq \frac{1}{K} \sum_{k=1}^{K} |F_k(h,D) - F_k(h',D)| \quad ,
\end{aligned}
$$

which gives the claim when combined with the first part of the theorem. □

D Proof of Lemma 3.4

Lemma (Finite Sample analysis). *Let D be a finite sample of $n \geq \frac{8 \log\left(\frac{2K+1}{\delta}\right)}{\min_{k'} p_{k'}}$ examples drawn i.i.d. from $\mathcal{D}$, where $p_{k'}$ is the true proportion of examples from group k'. Assume that $\mathbb{P}_{D \sim \mathcal{D}^n}\left(\sum_{k'=0}^{K} \left| C_k^{k'} - \widehat{C}_k^{k'} \right| > \alpha_C \right) \leq B_3 \exp\left(-B_4 \alpha_C^2 n\right)$. Let $\mathcal{H}$ be an hypothesis space and $d_{\mathcal{H}}$ be the Natarajan dimension of $\mathcal{H}$.*

- ***Assuming that** h **and** h' **are independent of** D. With probability $1 - \delta$ over the choice of D,*

$$
\left| F_k(h,D) - \widehat{F}_k(h',D) \right| \leq \widehat{\chi}_k(h,D) \, \|h - h'\|_{\mathcal{H}} + \widetilde{O}\left(\sum_{k'=1}^{K} \left| \widehat{C}_k^{k'} \right| \sqrt{\frac{\log\left(K/\delta\right)}{n p_{k'}}} \right)
$$

- ***Assuming that** h **and** h' **are dependent of** D. With probability $1 - \delta$ over the choice of D, $\forall h, h' \in \mathcal{H}$,*

$$
\left| F_k(h,D) - \widehat{F}_k(h',D) \right| \leq \widehat{\chi}_k(h,D) \, \|h - h'\|_{\mathcal{H}} + \widetilde{O}\left(\sum_{k'=1}^{K} \left| \widehat{C}_k^{k'} \right| \sqrt{\frac{d_{\mathcal{H}} + \log\left(K/\delta\right)}{n p_{k'}}} \right)
$$

Proof. First of all, notice that we have

$$
\left| F_k(h,D) - \widehat{F}_k(h',D) \right| = \left| F_k(h,D) - \widehat{F}_k(h,D) + \widehat{F}_k(h,D) - \widehat{F}_k(h',D) \right|
$$

$$
\begin{aligned}
&\leq \left|F_k(h,D) - \widehat{F}_k(h,D)\right| + \left|\widehat{F}_k(h,D) - \widehat{F}_k(h',D)\right| \\
&\quad \downarrow \text{ Theorem 3.2} \\
&\leq \left|F_k(h,D) - \widehat{F}_k(h,D)\right| + \widehat{\chi}_k(h',D)\left\|h - h'\right\|_{\mathcal{H}}
\end{aligned}
$$

Hence it remains to bound the first term. By definition of our fairness notions, notice that we have the following.

$$
\begin{aligned}
\left|F_k(h,D) - \widehat{F}_k(h,D)\right| &= \left|C_k^0 + \sum_{k'=1}^{K} C_k^{k'}\,\mathbb{P}\left(H(X) = Y | D_{k'}\right) - \widehat{C}_k^0 - \sum_{k'=1}^{K} \widehat{C}_k^{k'}\widehat{\mathbb{P}}\left(H(X) = Y | D_{k'}\right)\right| \\
&\leq \left|C_k^0 - \widehat{C}_k^0\right| + \left|\sum_{k'=1}^{K} C_k^{k'}\,\mathbb{P}\left(H(X) = Y | D_{k'}\right) - \widehat{C}_k^{k'}\widehat{\mathbb{P}}\left(H(X) = Y | D_{k'}\right)\right| \\
&\leq \left|C_k^0 - \widehat{C}_k^0\right| + \sum_{k'=1}^{K}\left|C_k^{k'}\,\mathbb{P}\left(H(X) = Y | D_{k'}\right) - \widehat{C}_k^{k'}\,\mathbb{P}\left(H(X) = Y | D_{k'}\right)\right. \\
&\quad \left. + \widehat{C}_k^{k'}\,\mathbb{P}\left(H(X) = Y | D_{k'}\right) - \widehat{C}_k^{k'}\widehat{\mathbb{P}}\left(H(X) = Y | D_{k'}\right)\right| \\
&\leq \left|C_k^0 - \widehat{C}_k^0\right| + \sum_{k'=1}^{K}\left|C_k^{k'}\,\mathbb{P}\left(H(X) = Y | D_{k'}\right) - \widehat{C}_k^{k'}\,\mathbb{P}\left(H(X) = Y | D_{k'}\right)\right| \\
&\quad + \sum_{k'=1}^{K}\left|\widehat{C}_k^{k'}\,\mathbb{P}\left(H(X) = Y | D_{k'}\right) - \widehat{C}_k^{k'}\widehat{\mathbb{P}}\left(H(X) = Y | D_{k'}\right)\right| \\
&\leq \left|C_k^0 - \widehat{C}_k^0\right| + \sum_{k'=1}^{K}\left|\widehat{C}_k^{k'}\right|\left|\mathbb{P}\left(H(X) = Y | D_{k'}\right) - \widehat{\mathbb{P}}\left(H(X) = Y | D_{k'}\right)\right| \\
&\quad + \sum_{k'=1}^{K}\left|C_k^{k'} - \widehat{C}_k^{k'}\right|\mathbb{P}\left(H(X) = Y | D_{k'}\right) \\
&\leq \left|C_k^0 - \widehat{C}_k^0\right| + \sum_{k'=1}^{K}\left|\widehat{C}_k^{k'}\right|\left|\mathbb{P}\left(H(X) = Y | D_{k'}\right) - \widehat{\mathbb{P}}\left(H(X) = Y | D_{k'}\right)\right| \\
&\quad + \sum_{k'=1}^{K}\left|C_k^{k'} - \widehat{C}_k^{k'}\right| \\
&\leq \sum_{k'=0}^{K}\left|C_k^{k'} - \widehat{C}_k^{k'}\right| + \sum_{k'=1}^{K}\left|\widehat{C}_k^{k'}\right|\left|\mathbb{P}\left(H(X) = Y | D_{k'}\right) - \widehat{\mathbb{P}}\left(H(X) = Y | D_{k'}\right)\right|
\end{aligned}
$$

We now need to consider two cases, depending on whether h depends on D or not.

Assuming that h is independent of D. In this case, our goal is to upper bound

$$
\underset{D\sim\mathcal{D}^n}{\mathbb{P}}\left(\left|F_k(h,D) - \widehat{F}_k(h,D)\right| > \alpha_C + \sum_{k'=1}^{K}\left|\widehat{C}_k^{k'}\right|\alpha_{k'}\right)
$$

Notice that, using the same trick that Woodworth et al. (2017) used to prove their Equation (38), we have that

$$
\underset{D\sim\mathcal{D}^n}{\mathbb{P}}\left(\left|F_k(h,D) - \widehat{F}_k(h,D)\right| > \alpha_C + \sum_{k'=1}^{K}\left|\widehat{C}_k^{k'}\right|\alpha_{k'}\right)
$$

$$
\begin{aligned}
&\leq \mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left(\sum_{k'=0}^{K}\left|C_k^{k'}-\widehat{C}_k^{k'}\right|+\sum_{k'=1}^{K}\left|\widehat{C}_k^{k'}\right|\left|\mathbb{P}\left(H(X)=Y|D_{k'}\right)-\widehat{\mathbb{P}}\left(H(X)=Y|D_{k'}\right)\right|>\alpha_C+\sum_{k'=1}^{K}\left|\widehat{C}_k^{k'}\right|\alpha_{k'}\right)\\
&\leq \mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left(\sum_{k'=0}^{K}\left|C_k^{k'}-\widehat{C}_k^{k'}\right|>\alpha_C\bigcup\left[\bigcup_{k'=1}^{K}\left|\mathbb{P}\left(H(X)=Y|D_{k'}\right)-\widehat{\mathbb{P}}\left(H(X)=Y|D_{k'}\right)\right|>\alpha_{k'}\right]\right)\\
&\leq \mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left(\sum_{k'=0}^{K}\left|C_k^{k'}-\widehat{C}_k^{k'}\right|>\alpha_C\right)+\sum_{k'=1}^{K}\mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left(\left|\mathbb{P}\left(H(X)=Y|D_{k'}\right)-\widehat{\mathbb{P}}\left(H(X)=Y|D_{k'}\right)\right|>\alpha_{k'}\right)\\
&\leq \mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left(\sum_{k'=0}^{K}\left|C_k^{k'}-\widehat{C}_k^{k'}\right|>\alpha_C\right)\\
&\quad+\sum_{k'=1}^{K}\sum_{i=0}^{n}\mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left(\left|\mathbb{P}\left(H(X)=Y|D_{k'}\right)-\widehat{\mathbb{P}}\left(H(X)=Y|D_{k'}\right)\right|>\alpha_{k'}\;\middle|\;|D_{k'}|=i\right)\mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left(|D_{k'}|=i\right)\\
&\quad\downarrow \text{ Let } p_{k'}=\mathbb{P}\left(D_{k'}\right)\\
&\leq \mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left(\sum_{k'=0}^{K}\left|C_k^{k'}-\widehat{C}_k^{k'}\right|>\alpha_C\right)\\
&\quad+\sum_{k'=1}^{K}\sum_{i=0}^{np_{k'}/2-1}\mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left(\left|\mathbb{P}\left(H(X)=Y|D_{k'}\right)-\widehat{\mathbb{P}}\left(H(X)=Y|D_{k'}\right)\right|>\alpha_{k'}\;\middle|\;|D_{k'}|=i\right)\mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left(|D_{k'}|=i\right)\\
&\quad+\sum_{k'=1}^{K}\sum_{i=np_{k'}/2}^{n}\mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left(\left|\mathbb{P}\left(H(X)=Y|D_{k'}\right)-\widehat{\mathbb{P}}\left(H(X)=Y|D_{k'}\right)\right|>\alpha_{k'}\;\middle|\;|D_{k'}|=i\right)\mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left(|D_{k'}|=i\right)\\
&\leq \mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left(\sum_{k'=0}^{K}\left|C_k^{k'}-\widehat{C}_k^{k'}\right|>\alpha_C\right)\\
&\quad+\sum_{k'=1}^{K}\sum_{i=0}^{np_{k'}/2-1}\mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left(|D_{k'}|=i\right)\\
&\quad+\sum_{k'=1}^{K}\sum_{i=np_{k'}/2}^{n}\mathop{\mathbb{P}}_{D_{k'}\sim\mathcal{D}_{k'}^i}\left(\left|\mathbb{P}\left(H(X)=Y|D_{k'}\right)-\widehat{\mathbb{P}}\left(H(X)=Y|D_{k'}\right)\right|>\alpha_{k'}\right)\mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left(|D_{k'}|=i\right)\\
&\leq \mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left(\sum_{k'=0}^{K}\left|C_k^{k'}-\widehat{C}_k^{k'}\right|>\alpha_C\right)\\
&\quad+\sum_{k'=1}^{K}\mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left(|D_{k'}|<\frac{np_{k'}}{2}\right)\\
&\quad+\sum_{k'=1}^{K}\sum_{i=np_{k'}/2}^{n}\mathop{\mathbb{P}}_{D_{k'}\sim\mathcal{D}_{k'}^i}\left(\left|\mathbb{P}\left(H(X)=Y|D_{k'}\right)-\widehat{\mathbb{P}}\left(H(X)=Y|D_{k'}\right)\right|>\alpha_{k'}\right)\mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left(|D_{k'}|=i\right)
\end{aligned}
$$

Using Hoeffding's inequality, we can show that

$$
\mathop{\mathbb{P}}_{D_{k'}\sim\mathcal{D}_{k'}^{n_{k'}}}\left(\left|\mathbb{P}\left(H(X)=Y|D_{k'}\right)-\widehat{\mathbb{P}}\left(H(X)=Y|D_{k'}\right)\right|>\beta\right)\leq 2\exp\left(-2\beta^2 n_{k'}\right)
$$

which implies

$$
\mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left(\left|F_k(h,D)-\widehat{F}_k(h,D)\right|>\alpha_C+\sum_{k'=1}^{K}\left|\widehat{C}_k^{k'}\right|\alpha_{k'}\right)
$$

$$
\begin{aligned}
&\quad \downarrow\ 2\exp\left(-2\beta^2 i\right) \geq 2\exp\left(-2\beta^2(i+1)\right). \\
&\leq \mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left(\sum_{k'=0}^{K}\left|C_k^{k'}-\widehat{C}_k^{k'}\right|>\alpha_C\right) \\
&\quad + \sum_{k'=1}^{K}\mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left(|D_{k'}|<\frac{np_{k'}}{2}\right) \\
&\quad + \sum_{k'=1}^{K} 2\exp\left(-\alpha_{k'}^2 np_{k'}\right) \\
&\quad \downarrow\ \text{By assumption, } \mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left(\sum_{k'=0}^{K}\left|C_k^{k'}-\widehat{C}_k^{k'}\right|>\alpha_C\right) \leq B_3\exp\left(-B_4\alpha_C^2 n\right). \\
&\leq B_3\exp\left(-B_4\alpha_C^2 n\right) \\
&\quad + \sum_{k'=1}^{K}\mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left(|D_{k'}|<\frac{np_{k'}}{2}\right) \\
&\quad + \sum_{k'=1}^{K} 2\exp\left(-\alpha_{k'}^2 np_{k'}\right) \\
&\quad \downarrow\ \text{Chernoff multiplicative bound.} \\
&\leq B_3\exp\left(-B_4\alpha_C^2 n\right) \\
&\quad + \sum_{k'=1}^{K}\exp\left(-\frac{np_{k'}}{8}\right) \\
&\quad + \sum_{k'=1}^{K} 2\exp\left(-\alpha_{k'}^2 np_{k'}\right)
\end{aligned}
$$

Now, by assumption that $n \geq \frac{8\log\left(\frac{2K+1}{\delta}\right)}{\min_{k'} p_{k'}}$ and setting

$$
\begin{aligned}
\alpha_C &= \sqrt{\frac{\log\left(\frac{B_3(2K+1)}{\delta}\right)}{B_4 n}} \\
\alpha_{k'} &= \sqrt{\frac{\log\left(\frac{2(2K+1)}{\delta}\right)}{np_{k'}}}
\end{aligned}
$$

yields that, with probability at least $1-\delta$,

$$
\left|F_k(h,D)-\widehat{F}_k(h,D)\right| \leq \sqrt{\frac{\log\left(\frac{B_3(2K+1)}{\delta}\right)}{B_4 n}} + \sum_{k'=1}^{K}\left|\widehat{C}_k^{k'}\right|\sqrt{\frac{\log\left(\frac{2(2K+1)}{\delta}\right)}{np_{k'}}}
$$

Assuming that h is dependent of D. In this case, our goal is to bound

$$
\mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left(\sup_{h\in\mathcal{H}}\left|F_k(h,D)-\widehat{F}_k(h,D)\right|>\alpha_C+\sum_{k'=1}^{K}\left|\widehat{C}_k^{k'}\right|\alpha_{k'}\right)
$$

Using similar arguments that in the independent case, we have that

$$
\mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left(\sup_{h\in\mathcal{H}}\left|F_k(h,D)-\widehat{F}_k(h,D)\right|>\alpha_C+\sum_{k'=1}^{K}\left|\widehat{C}_k^{k'}\right|\alpha_{k'}\right)
$$

$$
\begin{aligned}
&\leq \mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left(\sum_{k'=0}^{K}\left|C_k^{k'}-\widehat{C}_k^{k'}\right|>\alpha_C\right)\\
&\quad+\sum_{k'=1}^{K}\mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left(|D_{k'}|<\frac{np_{k'}}{2}\right)\\
&\quad+\sum_{k'=1}^{K}\sum_{i=np_{k'}/2}^{n}\mathop{\mathbb{P}}_{D_{k'}\sim\mathcal{D}_{k'}^i}\left(\sup_{h\in\mathcal{H}}\left|\mathbb{P}\left(H(X)=Y|D_{k'}\right)-\widehat{\mathbb{P}}\left(H(X)=Y|D_{k'}\right)\right|>\alpha_{k'}\right)\mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left(|D_{k'}|=i\right)
\end{aligned}
$$

Using the Multiclass Fundamental Theorem (Shalev-Shwartz and Ben-David, 2014, Theorem 29.3, Lemma 29.4) with $d_{\mathcal{H}}$ the Natarajan dimension of $\mathcal{H}$, we have that

$$
\mathop{\mathbb{P}}_{D_{k'}\sim\mathcal{D}_{k'}^{n_{k'}}}\left(\sup_{h\in\mathcal{H}}\left|\mathbb{P}\left(H(X)=Y|D_{k'}\right)-\widehat{\mathbb{P}}\left(H(X)=Y|D_{k'}\right)\right|>\beta\right)\leq 8n_{k'}^{d_{\mathcal{H}}}\left|\mathcal{Y}\right|^{2d_{\mathcal{H}}}\exp\left(-\frac{n_{k'}\beta^2}{32}\right)
$$

which implies

$$
\begin{aligned}
&\mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left(\sup_{h\in\mathcal{H}}\left|F_k(h,D)-\widehat{F}_k(h,D)\right|>\alpha_C+\sum_{k'=1}^{K}\left|\widehat{C}_k^{k'}\right|\alpha_{k'}\right)\\
&\qquad\leq \mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left(\sum_{k'=0}^{K}\left|C_k^{k'}-\widehat{C}_k^{k'}\right|>\alpha_C\right)\\
&\qquad\quad+\sum_{k'=1}^{K}\mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left(|D_{k'}|<\frac{np_{k'}}{2}\right)\\
&\qquad\quad+\sum_{k'=1}^{K}8\left(\frac{np_{k'}}{2}\right)^{d_{\mathcal{H}}}\left|\mathcal{Y}\right|^{2d_{\mathcal{H}}}\exp\left(-\frac{\left(\frac{np_{k'}}{2}\right)\alpha_{k'}^2}{32}\right)
\end{aligned}
$$

Now, by assumption that $n\geq\frac{8\log\left(\frac{2K+1}{\delta}\right)}{\min_{k'}p_{k'}}$ and setting

$$
\alpha_C=\sqrt{\frac{\log\left(\frac{B_3(2K+1)}{\delta}\right)}{B_4 n}}
$$

$$
\alpha_{k'}=\sqrt{\frac{64\log\left(\frac{8\left(\frac{np_{k'}}{2}\right)^{d_{\mathcal{H}}}\left|\mathcal{Y}\right|^{2d_{\mathcal{H}}}(2K+1)}{\delta}\right)}{np_{k'}}}=\sqrt{\frac{64\left(d_{\mathcal{H}}\left(\log\left(\frac{np_{k'}}{2}\right)+2\log\left(|\mathcal{Y}|\right)\right)+\log\left(\frac{8(2K+1)}{\delta}\right)\right)}{np_{k'}}}
$$

yields that, with probability at least $1-\delta$, $\forall h\in\mathcal{H}$

$$
\left|F_k(h,D)-\widehat{F}_k(h,D)\right|\leq\sqrt{\frac{\log\left(\frac{B_3(2K+1)}{\delta}\right)}{B_4 n}}+\sum_{k'=1}^{K}\left|\widehat{C}_k^{k'}\right|\sqrt{\frac{64\left(d_{\mathcal{H}}\left(\log\left(\frac{np_{k'}}{2}\right)+2\log\left(|\mathcal{Y}|\right)\right)+\log\left(\frac{8(2K+1)}{\delta}\right)\right)}{np_{k'}}}\ .
$$

This concludes the proof. □

E Bound for Output Perturbation (Proof of Lemma 4.1)

Lemma. *Let h^{priv} be the vector released by output perturbation with noise $\sigma^2=8\Lambda^2\log(1.25/\delta)/\mu^2n^2\epsilon^2$, and $0<\zeta<1$, then with probability at least $1-\zeta$,*

$$
\left\|h^{\mathrm{priv}}-h^*\right\|_2^2\leq\frac{32p\Lambda^2\log(1.25/\delta)\log(2/\zeta)}{\mu^2n^2\epsilon^2}\ .
$$

Proof. We prove this lemma in two steps. First, we show that for a given sensitivity, the distance $\left\|h^{\mathrm{priv}} - h^*\right\|$ is bounded. Second, we estimate the sensitivity.

Bounding the Error. Let Δ be the sensitivity of the function $D \to \arg\min_{w\in\mathcal{C}} f(w; D)$. Its value can be released under (ϵ, δ) differential privacy (Chaudhuri et al., 2011; Lowy and Razaviyayn, 2021) as follows:

$$h^{\mathrm{priv}} = h^* + \mathcal{N}(0, \sigma^2 \mathbb{I}_p) \ , \tag{8}$$

where $\sigma^2 = \frac{2\Delta^2 \log(1.25/\delta)}{\epsilon^2}$ and $h^* = \arg\min_{h\in\mathcal{C}} f(h)$. Then, Chernoff's bound gives, for $t, \alpha > 0$,

$$\mathbb{P}(\left\|h^{priv} - h^*\right\|^2 \geq \alpha) \leq \exp(-t\alpha)\, \mathbb{E}(\exp(t\left\|h^{priv} - h^*\right\|^2)) \tag{9}$$

$$= \exp(-t\alpha) \prod_{j=1}^{p} \mathbb{E}(\exp(t(h_j^{priv} - h_j^*)^2)) \ , \tag{10}$$

by independence of the noise's p coordinates. Since $h_j^{priv} - h_j^*$ is a Gaussian random variable of mean 0 and variance σ^2, we can compute $\mathbb{E}(\exp(t(h_j^{priv} - h_j^*))) = (1 - 2t\sigma^2)^{-1/2}$. We then obtain

$$\mathbb{P}(\left\|h^{priv} - h^*\right\|^2 \geq \alpha) \leq \exp(-t\alpha)(1 - 2t\sigma^2)^{-p/2} \ . \tag{11}$$

Let $t = 1/4p\sigma^2$, then it holds that $1 - 2t\sigma^2 = 1 - 1/2p \leq 1$ and

$$(1 - 2t\sigma^2)^{-p/2} = \exp\left(-\frac{p}{2}\log(1 - \frac{1}{2p})\right) \leq \exp\left(\frac{1}{2(1 - \frac{1}{p})}\right) \leq \exp(1/2) \leq 2 \ , \tag{12}$$

since $\frac{p}{2}\log(1 - \frac{1}{2p}) \geq \frac{p}{2}\frac{-1/2p}{1-1/2p} \geq -\frac{1}{2}$. Let $0 < \zeta < 1$, $t = 1/4p\sigma^2$ and $\alpha = 4p\sigma^2 \log(2/\zeta)$, we have proven

$$\mathbb{P}(\left\|h^{priv} - h^*\right\|^2 \geq \alpha) \leq 2\exp\left(-\frac{\alpha}{4p\sigma}\right) \leq \zeta \ . \tag{13}$$

The error obtained by output perturbation is thus upper bounded by $\left\|h^{priv} - h^*\right\|^2 \leq 4p\sigma^2 \log(2/\zeta) = \frac{8p\Delta^2 \log(1.25/\delta)\log(2/\zeta)}{\epsilon^2}$ with probability at least $1 - \zeta$.

Estimating the Sensitivity. Define $g(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(w; d_i')$ with $d_i' \in \mathcal{X} \times \mathcal{Y}$ such that $d_i' = d_i$ for all $i \neq 1$. By strong convexity, the two following inequalities hold for h, h',

$$f(h) \geq f(h') + \langle \nabla f(h'), h - h' \rangle + \frac{\mu}{2}\|h - h'\|^2 \ , \tag{14}$$

$$f(h') \geq f(h) + \langle \nabla f(h), h' - h \rangle + \frac{\mu}{2}\|h - h'\|^2 \ . \tag{15}$$

Summing these two inequalities give $\langle \nabla f(h) - \nabla f(h'), h - h' \rangle \geq \frac{\mu}{2}\left\|h - h'\right\|^2$. Let h_1^* and h_2^* be the respective minimizers of f and g over $\mathcal{C}$, taking $h = h_1^*$ and $h' = h_2^*$ gives

$$\frac{\mu}{2}\|h_1^* - h_2^*\|^2 \leq \langle \nabla f(h_1^*) - \nabla f(h_2^*), h_1^* - h_2^* \rangle \leq \|\nabla f(h_1^*) - \nabla f(h_2^*)\| \cdot \|h_1^* - h_2^*\| \ . \tag{16}$$

Now, if $\mathcal{C} = \mathbb{R}^p$, optimality conditions give

$$\nabla f(h_1^*) = 0 = \nabla g(h_2^*) = \nabla f(h_2^*) - \nabla F(h_2^*; d_1) + F(h_2^*; d_1') \ , \tag{17}$$

resulting in $\|\nabla f(h_1^*) - \nabla f(h_2^*)\| = \|\frac{1}{n}\nabla(h_2^*; d_1) - \frac{1}{n}\nabla(h_2^*; d_1')\| \leq \frac{2\Lambda}{n}$. Combined with (16), this shows that the sensitivity of $\arg\min_{h\in\mathcal{C}} f(h)$ is $\Delta = \frac{2\Lambda}{n\mu}$, which concludes the proof. □

F Convergence of DP-SGD (Proof of Lemma 4.2)

Lemma. *Let h^{priv} be the vector released by DP-SGD with $\sigma^2 = 64\Lambda^2 T^2 \log(3T/\delta)\log(2/\delta)/n^2\epsilon^2$. Assume that $\sigma_*^2 = \mathbb{E}_{i\sim[n]} \|\nabla \ell(h^*; x_i, y_i)\|^2 \leq \sigma^2$. Let $0 < \zeta < 1$, then with probability at least $1-\zeta$,*

$$\left\|h^{\text{priv}} - h^*\right\|_2^2 = \widetilde{O}\left(\frac{p\Lambda^2 \log(1/\delta)^2}{\zeta\mu^2 n^2 \epsilon^2}\right) \ ,$$

where $\widetilde{O}$ ignores logarithmic terms in n (the number of examples) and p (the number of model parameters).

Proof. We start by recalling that in DP-SGD,

$$h^{t+1} = \pi_{\mathcal{H}}(h^t - \gamma(g^t + \eta^t)) \ . \tag{18}$$

Since $h^* \in \mathcal{H}$, and $\mathcal{H}$ is convex, we have

$$\begin{aligned}
\left\|h^{t+1} - h^*\right\|^2 &= \left\|\pi_{\mathcal{H}}(h^t - \gamma(g^t + \eta^t)) - h^*\right\|^2 &(19)\\
&= \left\|h^t - h^*\right\|^2 - 2\gamma\langle g^t + \eta^t, h^t - h^*\rangle + \gamma^2 \left\|g^t + \eta^t\right\|^2 &(20)\\
&\leq \left\|h^t - h^*\right\|^2 - 2\gamma\langle g^t + \eta^t, h^t - h^*\rangle + 2\gamma^2 \left\|g^t\right\|^2 + 2\gamma^2 \left\|\eta^t\right\|^2 \ , &(21)
\end{aligned}$$

where we developed the square and used $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ for $a, b \in \mathbb{R}^p$. Taking the expectation with respect to the stochastic gradient computation and noise, we obtain

$$\mathbb{E}\left\|h^{t+1} - h^*\right\|^2 \leq \left\|h^t - h^*\right\|^2 - 2\gamma\langle \nabla f(h^t), h^t - h^*\rangle + 2\gamma^2 \mathbb{E}\left\|g^t\right\|^2 + 2\gamma^2 \mathbb{E}\left\|\eta^t\right\|^2 \ , \tag{22}$$

since $\mathbb{E}(\eta^t) = 0$ and $\mathbb{E}(g^t) = \nabla f(h^t)$. Now recall that, by strong-convexity of f, we have

$$f(h^*) \geq f(h^t) + \langle \nabla f(h^t), h^* - h^t\rangle + \frac{\mu}{2}\left\|h^t - h^*\right\|^2 \ . \tag{23}$$

By reorganizing, we obtain $-2\gamma\langle \nabla f(h^t), h^t - h^*\rangle \leq -2\gamma(f(h^t) - f(h^*)) - \gamma\mu\|h^t - h^*\|^2$, which gives

$$\mathbb{E}\left\|h^{t+1} - h^*\right\|^2 \leq (1-\gamma\mu)\left\|h^t - h^*\right\|^2 - 2\gamma(f(h^t) - f(h^*)) + 2\gamma^2 \mathbb{E}\left\|g^t\right\|^2 + 2\gamma^2 \mathbb{E}\left\|\eta^t\right\|^2 \ . \tag{24}$$

Finally, remark that if $f = \frac{1}{n}\sum_{i=1}^n f_i$ with each f_i being β-smooth and $\mathbb{E}\, f_i = f$, we have, for $i \sim [n]$,

$$\begin{aligned}
\mathbb{E}\left\|\nabla f_i(h^t)\right\|^2 &= \mathbb{E}\left\|\nabla f_i(h^t) - \nabla f_i(h^*) + \nabla f_i(h^*)\right\|^2 &(25)\\
&\leq \mathbb{E}(2\left\|\nabla f_i(h^t) - \nabla f_i(h^*)\right\|^2 + 2\left\|\nabla f_i(h^*)\right\|^2) &(26)\\
&\leq \mathbb{E}(4\beta(f_i(h^t) - f_i(h^*) - \langle \nabla f_i(h^*), h^t - h^*\rangle) + 2\left\|\nabla f_i(h^*)\right\|^2) &(27)\\
&= 4\beta(f(h^t) - f(h^*)) + 2\,\mathbb{E}\left\|\nabla f_i(h^*)\right\|^2 \ , &(28)
\end{aligned}$$

since f_i is β-smooth, which implies, for all $w, v \in \mathbb{R}^p$,

$$\|\nabla f_i(w) - \nabla f_i(v)\|^2 \leq 2\beta(f_i(w) - f_i(v) - \langle \nabla f_i(v), w - v\rangle \ , \tag{29}$$

and $\mathbb{E}\,\nabla f_i(h^*) = 0$. Combined with the fact that $\mathbb{E}\,\|\nabla f_i(h^*)\|^2 \leq \sigma_*^2$ and $\mathbb{E}\,\|\eta^t\|^2 = p\sigma^2$, we obtained

$$\begin{aligned}
\mathbb{E}\left\|h^{t+1} - h^*\right\|^2 &\leq (1-\gamma\mu)\left\|h^t - h^*\right\|^2 + (4\beta\gamma^2 - 2\gamma)(f(h^t) - f(h^*)) + 2\gamma^2(\sigma_*^2 + \sigma^2) &(30)\\
&\leq (1-\gamma\mu)\left\|h^t - h^*\right\|^2 + 4\gamma^2\sigma^2 \ , &(31)
\end{aligned}$$

since $\gamma \leq 1/2\beta$, which implies $4\beta\gamma^2 - 2\gamma \leq 0$ and $\sigma^* \leq \sigma$. By induction, we obtain that, after T iterations,

$$\mathbb{E}\left\|h^T - h^*\right\|^2 \leq (1-\gamma\mu)^T \left\|h^0 - h^*\right\|^2 + 4\gamma^2 \sum_{t=0}^{T-1} (1-\gamma\mu)^{T-t}\sigma^2 \tag{32}$$

$$\leq (1-\gamma\mu)^T \left\|h^0 - h^*\right\|^2 + \frac{4\gamma\sigma^2}{\mu} \ . \tag{33}$$

Now, recall that DP-SGD is (ϵ, δ)-differentially private for $\sigma^2 = \frac{64\Lambda^2 T \log(3T/\delta)\log(2/\delta)}{n^2\epsilon^2}$ (following from the Gaussian mechanism, advanced composition theorem and amplification by subsampling). Thus, taking $\gamma = 1/2\beta$, and setting $T = \frac{2\beta}{\mu}\log(\mu\beta\left\|h^0 - h^*\right\|^2/2M^2)$, where $M^2 = \frac{64\Lambda^2 T\log(2/\delta)}{n^2\epsilon^2}$, yields

$$\mathbb{E}\left\|h^T - h^*\right\|^2 \leq \frac{2(T\log(3T/\delta)+1)M^2}{\beta\mu} \leq \frac{8M^2}{\mu^2}\log\Big(\frac{\mu\beta\left\|h^0-h^*\right\|^2}{2M^2}\Big)\log\Big(\frac{6\beta\log(\frac{\mu\beta\left\|h^0-h^*\right\|^2}{2M^2})}{\mu\delta}\Big) \ . \tag{34}$$

Using Markov inequality, we obtain

$$\mathbb{P}\left(\left\|h^T - h^*\right\|^2 \geq \frac{8M^2}{\zeta\mu^2}\log\Big(\frac{\mu\beta\left\|h^0-h^*\right\|^2}{2M^2}\Big)\log\Big(\frac{6\beta\log(\frac{\mu\beta\left\|h^0-h^*\right\|^2}{2M^2})}{\mu\delta}\Big)\right) \leq \zeta \ . \tag{35}$$

This results in the following upper bound, with probability at least $1-\zeta$,

$$\left\|h^T - h^*\right\|^2 \leq \frac{512\Lambda^2\log(3T/\delta)\log(2/\delta)}{\zeta\mu^2 n^2\epsilon^2}\log\Big(\frac{\mu\beta\left\|h^0-h^*\right\|^2}{2M^2}\Big)\log\Big(\frac{6\beta\log(\frac{\mu\beta\left\|h^0-h^*\right\|^2}{2M^2})}{\mu\delta}\Big) \tag{36}$$

$$= \widetilde{O}\left(\frac{G^2\log(1/\delta)}{\zeta\mu^2 n^2\epsilon^2}\right) \ , \tag{37}$$

which is the result of our lemma. □

G Additional Experimental Details

G.1 Experimental Setup

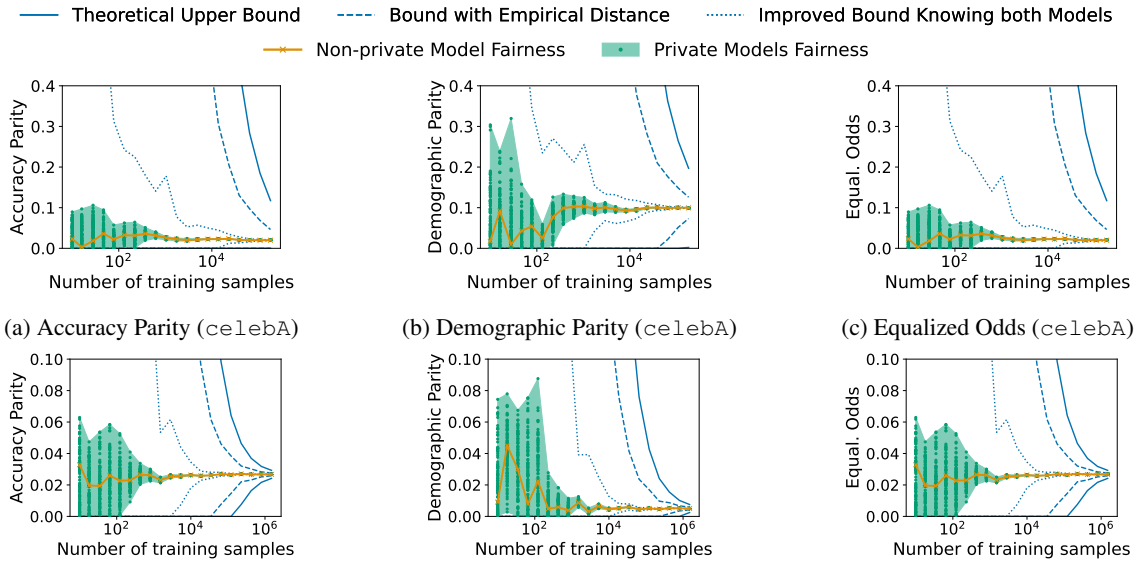
The `celebA` dataset (Liu et al., 2015) is a face attributes dataset, that can be downloaded at `http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html`, and the `folktables` dataset (Ding et al., 2021) is derived from US Census, and can be downloaded using a Python package available here `https://github.com/zykls/folktables`.

On each dataset, for each value of n, we train a ℓ_2-regularized logistic regression model using `scikit-learn` (Pedregosa et al., 2011). Private models are then learned using the output perturbation mechanism as described in Section 4.1. We then compute our bounds using the non-private model as reference, over a test set containing 10% of the data, that has not been used for training (containing $20,260$ records for `celebA` and $166,450$ records for `folktables`). The value of the bound is computed by minimizing the experession given by the Chernoff bound using the golden section search algorithm (Kiefer, 1953). The code is in the supplementary, and will be made public.

For the plots with different number of training records, we train 20 non-private models with a number of records logarithmically spaced between 10 and the number of records in the complete training set (that is, $182,339$ for `celebA` and $1,498,050$ for `folktables`). For the plots with different privacy budgets, we use 20 values logarithmically spaced between 10^{-3} and 10 for both datasets.

G.2 Results for Other Fairness Measures

Our bounds also hold for accuracy parity, demographic parity and equalized odds. The same plots as those presented in Figure 1 for these fairness notions are in Figure 2 and Figure 3. The comments from Section 5 on equality of opportunity and accuracy also hold for these three notions of fairness.



(a) Accuracy Parity (`celebA`) (b) Demographic Parity (`celebA`) (c) Equalized Odds (`celebA`)

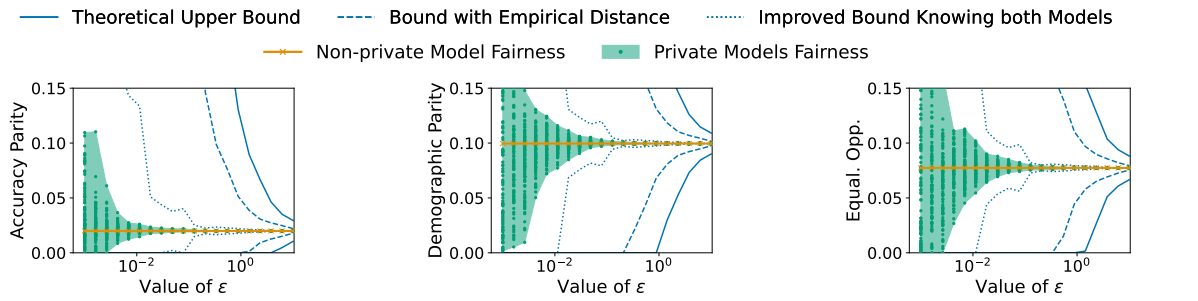
(d) Accuracy Parity (`folktables`) (e) Demographic Parity (`folktables`) (f) Equalized Odds (`folktables`)

Figure 2: Fairness and accuracy levels for optimal non-private model and random private ones as a function of the number n of training samples. For each value of n, we sample 100 private models and take their minimum and maximum fairness/accuracy values to mark the area of attainable values. The solid blue line and the dashed one give our guarantees, respectively from Theorem 4.4 with Lemma 4.1's bounds and with an empirical evaluation of $\left\|h^{\text{priv}} - h^*\right\|$.

G.3 Refined Bounds with Additional Knowledge of h^{priv} and h^*

In Assumption 2.1, we use a uniform Lipschitz bound for all $h, h' \in \mathcal{H}$. Let's consider the class $\mathcal{H}$ of linear models, where, for $h \in \mathcal{H}$, we denote by h_y the parameters of h associated with the label y, that is $h(x, y) = h_y^T x$. For linear models, we derived the bound $\|\rho(h, x, y) - \rho(h', x, y)\|_{\mathcal{H}} \leq 2 \|x\|_2 \|h - h'\|_{\mathcal{H}}$, as derived in Section 2. Note that this inequality can be very loose whenever x and $h_y - h'_y$ (for $y \in \mathcal{Y}$) are (close to) orthogonal. When they are orthogonal, this bounds only gives $0 = (h_y - h'_y)^T x \leq \left\|h_y - h'_y\right\|_2 \|x\|_2$. We can thus improve the inequality by remarking that we have

$$
\begin{aligned}
|\rho(h, x, y) - \rho(h', x, y)| &\leq |h(x, y) - h'(x, y)| + \max_{y' \neq y} |h(x, y') - h'(x, y')| \\
&= \left|h_y^T x - h_y'^T x\right| + \max_{y' \neq y} \left|h_{y'}^T x - h'_{y'} x\right| \\
&= \left|(h_y - h'_y)^T x\right| + \max_{y' \neq y} \left|(h_{y'} - h'_{y'}) x\right| \\
&= \left|(h_y - h'_y)^T p_{h_y - h'_y}(x)\right| + \max_{y' \neq y} \left|(h_{y'} - h'_{y'}) p_{h_y - h'_y}(x)\right| \\
&\leq 2 \max_{y' \in \mathcal{Y}} \left\|p_{h_y - h'_y}(x)\right\| \|h - h'\|_{\mathcal{H}} \quad ,
\end{aligned}
$$



(a) Accuracy Parity (`celebA`) (b) Demographic Parity (`celebA`) (c) Equality of Opportunity (`celebA`)

(d) Accuracy Parity (`folktables`) (e) Demographic Parity (`folktables`) (f) Equality of Opportunity (`folktables`)

Figure 3: Fairness and accuracy levels for optimal non-private model and random private ones as a function of privacy budget ϵ. For each value of ϵ, we sample 100 private models and take their minimum and maximum fairness/accuracy values to mark the area of attainable values. The solid blue line and the dashed one respectively give our guarantees, respectively from Theorem 4.4 with Lemma 4.1's bounds and with an empirical evaluation of $\left\|h^{\text{priv}} - h^*\right\|$.

where $p_{h_y - h'_y}(x)$ is the projection of x on the axis defined by $h_y - h'_y$. We can thus define a variant of $L_{X,Y}$ which depends on $h - h'$

$$L_{X,Y}^{h-h'} = 2 \max_{y \in \mathcal{Y}} \left\| p_{h_y - h'_y}(x) \right\| \quad . \tag{38}$$

Replacing Assumption 2.1 by this inequality in the proof of Theorem 3.1, we end up with the inequality

$$\left| \mathbb{P}\left(H(X) = Y \mid E\right) - \mathbb{P}\left(H'(X) = Y \mid E\right) \right| \leq \mathbb{P}\left(\frac{|\rho(h, X, Y)|}{L_{X,Y}^{h-h'}} \leq \|h - h'\|_{\mathcal{H}} \,\middle|\, E \right) \quad ,$$

where the probability is over $(X, S, Y) \sim \mathcal{D}$. We obtained the same bound as Theorem 3.1, except with $L_{X,Y}^{h-h'}$ instead of $L_{X,Y}$. Note that even if this gives a much tighter bound, this can generally not be computed, as one of h or h' is typically not known.