Fairness Certificates for Differentially Private Classification

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

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

Abstract

In this work, we theoretically study the impact of differential privacy on fairness in binary 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 the probability that a decision function makes a negative prediction 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, such models can be used in sensitive contexts, and impact people's lives. It is thus crucial that users can trust in 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 study 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 the binary classification setting.

Fairness takes various forms (depending on the task and context), and several definitions have been proposed. 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 four popular group fairness 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), its implications for fairness are not yet well understood.

Contributions. In this work, we theoretically quantify the difference in fairness levels between private and non-private models. More precisely, 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 probability that a decision function makes a negative prediction, conditioned on an arbitrary event (such as membership to a sensitive group), is pointwise Lipschitz continuous with respect to the model. This pointwise Lipschitzness is inherited by many popular group fairness notions, such as Equalized Odds and Demographic Parity. Consequently, two sufficiently close models will have similar fairness levels. We then recall that, given enough training records, 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)) learn models that are close to the optimal non-private model. Combining these two results, we derive high probability bounds showing 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. Both results can be seen as certificates on the fairness loss due to privacy.

Related work. Among the prior work that studied fairness and privacy in a joint fashion, one may identify three main research directions. First, it has been empirically observed that privacy can exacerbate unfairness (Pujol et al., 2020; Bagdasaryan et al., 2019) and, conversely, that enforcing a fair model can lead to more privacy leakage for the unprivileged group (Chang and Shokri, 2020). Unfortunately, these observations are not supported by theoretical results. It is thus not clear if they hold for all datasets and private training methods. 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) 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 behaviour over the possible private models while our results are model-specific. Furthermore, their analysis is based on a second-order Taylor approximation, which might be loose even for popular loss functions. Finally and most importantly, loss-based fairness does not always guarantee the fairness of the actual decisions taken by the model. In contrast, our work provides guarantees in terms of 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 binary classification problem with a feature space $\mathcal{X}$, binary labels $\mathcal{Y} = \{-1, 1\}$, 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 $D = \{(x_1, s_1, y_1), \dots, (x_n, s_n, y_n)\}$ be a training set of n examples drawn i.i.d. from $\mathcal{D}$. Let $\mathcal{H}$ be a space of real-valued functions $h : \mathcal{X} \to \mathbb{R}$ equipped with a norm $\|\cdot\|_{\mathcal{H}}$. Given a decision function $h \in \mathcal{H}$, a binary prediction may be obtained using the sign function, that is $\text{sign}(h(x))$ with the convention $\text{sign}(0) = 1$. 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 will be evaluated through its accuracy $\text{Acc}(h) = \mathbb{P}(\text{sign}(h(X)) = Y)$, and its fairness level (as defined in Section 2.1). Furthermore, given a non-private algorithm $\mathcal{A}$, our goal will be to compare 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:

$$F_k(h, D) = C_k^0 + \sum_{k'=1}^{K} C_k^{k'} \, \mathbb{P}\left(h(X) < 0 | D_{k'}\right) \quad , \tag{1}$$

where the C_k's are group specific constants 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 (Calders et al., 2009), Equality of Opportunity (Hardt et al., 2016), Equalized Odds (Hardt et al., 2016), and Accuracy Parity (Zafar et al., 2017) are all particular cases of the general form in Equation (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:

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

where $\mathrm{Fair}(h, D) = 0$ indicates that h is fair and $\mathrm{Fair}(h, D) > 0$ that it is unfair.

2.2 Differential Privacy

We measure the privacy of machine learning models with differential privacy (see Definition 1 below). Differential privacy 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. One of the key properties of differential privacy guarantees is that it still holds after post-processing of the algorithm's outcome (Dwork, 2006), as long as this post-processing is independent on 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 1 (Differential Privacy (Dwork, 2006)). *Let $\mathcal{A}^{\mathrm{priv}} : (\mathcal{X} \times \mathcal{S} \times \mathcal{Y})^n \to \mathcal{H}$ be a randomized algorithm. We say that $\mathcal{A}^{\mathrm{priv}}$ 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 $H \subseteq \mathcal{H}$,*

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

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 impacts the output of $\mathcal{A}$. This is typically measured using the (L2) sensitivity of $\mathcal{A}$, which is defined as

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

The value of $\mathcal{A}$ on a dataset $D \in (\mathcal{X} \times \mathcal{S} \times \mathcal{Y})^n$ can 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 a building blocks in more complex algorithms. In particular, they can be composed, as described by 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

In this section, we show that several popular group fairness notions are pointwise Lipschitz with respect to the decision function. To this end, we first prove a more general result (which might be of independent interest) on the pointwise Lipschitzness of the probability for a decision function to make a negative prediction, conditionally on an arbitrary event.

3.1 Pointwise Lipschitzness of Conditional Negative Predictions

We first link the difference of conditional probabilities that two decision functions make a negative prediction to the distance that separates the two functions. To prove this result, we need to assume that the associated predictor, which outputs a real valued prediction given a decision function and an example, is Lipschitz continuous with respect to its first argument. This is formalized in the next definition.

Definition 2 (Lipschitz continuity of the predictor associated with $\mathcal{H}$). *Let $\mathcal{H}$ be a set of real valued functions $h : \mathcal{X} \to \mathbb{R}$ equipped with a norm $\|\cdot\|_{\mathcal{H}}$. The predictor associated with $\mathcal{H}$ is the function $g : \mathcal{H} \times \mathcal{X} \to \mathbb{R}$ such that $g(h, x) = h(x)$. We say that the predictor is L-Lipschitz continuous in its first argument if it respects the following condition for a given $L > 0$:*

$$\forall x \in \mathcal{X}, \quad \forall h, h' \in \mathcal{H}, \quad |g(h, x) - g(h', x)| = |h(x) - h'(x)| \leq L \|h - h'\|_{\mathcal{H}} \ .$$

This constant gives a uniform bound on how a function behaves in comparison to another, depending on the distance that separates them. For instance, consider a bounded $\mathcal{X} \subseteq \mathbb{R}^p$, and $\mathcal{H}$ the set of linear functions from $\mathcal{X}$ to $\mathbb{R}$. As $\mathcal{H}$ is a set of linear operators, it can be equipped with the operator norm $\|h\|_{\mathcal{H}} = \sup_{\|x\|_2 \leq 1} h(x)$. We obtain $L = \sup_{x \in \mathcal{X}} \|x\|_2$, since for $h, h' \in \mathcal{H}$, $|h(x) - h'(x)| \leq \|h - h'\|_{\mathcal{H}} \|x\|_2 \leq \sup_{x \in \mathcal{X}} \|x\|_2 \|h - h'\|_{\mathcal{H}}$. We now state our main result, that we prove in Appendix B.

Theorem 1 (Pointwise Lipschitzness of Conditional Negative Predictions). *Let $\mathcal{H}$ be a set of real valued functions, and L the Lipschitz constant defined in Definition 2. 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 $\mathbb{E}\left({}^{1}/_{|h(X)|} \mid E\right) < +\infty$, then*

$$|\mathbb{P}(h(X) < 0 \mid E) - \mathbb{P}(h'(X) < 0 \mid E)| \leq \mathbb{E}\left(\frac{1}{|h(X)|}\middle| E\right) L \|h - h'\|_{\mathcal{H}} \ .$$

Proof. (Sketch) 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 $|\mathbb{P}(h(X) < 0 \mid E) - \mathbb{P}(h'(X) < 0 \mid E)| \leq \mathbb{P}(|h(X)| \leq L\|h - h'\|_{\mathcal{H}} \mid E)$. Then, taking the inverse and applying Markov's inequality gives the desired result. □

Remark 1 (Markov's inequality and Chernoff's bound). *In the last step of the proof of Theorem 1, a tighter result can be obtained using Chernoff's bound instead of Markov's inequality (as shown in Appendix B). However, the resulting bound is harder to parse and interpret. Hence, for the sake of readability, we focus on the result obtained with Markov's inequality.*

Remark 2 (Importance of $\mathbb{E}(1/|h(X)| \mid E)$). *Given $x \in \mathcal{X}$, $1/|h(x)|$ is as high as $h(x)$ is close to zero. This implies that, when the probability (given E) that a point is near the decision threshold is small, $\mathbb{E}(1/|h(X)||E)$ is also small. This can notably be the case for large margin classifiers.*

Theorem 1 shows that the function $h \mapsto \mathbb{P}(h(X) < 0 \mid E)$ is pointwise Lipschitz over $\mathcal{H}$. In the remainder of this section, we will use Theorem 1 to show that popular group fairness notions are also pointwise Lipschitz.

3.2 Pointwise Lipschitzness and Group Fairness

We now use Theorem 1 to relate the fairness level of any two classifiers, based on the distance between their decision functions. Specifically, Theorem 2 bounds the relative fairness levels for fairness notions in the form of Equation (1).

Theorem 2 (Pointwise Lipschitzness of Fairness). *Let $h, h' \in \mathcal{H}$ be the decision functions of two binary classifiers, L be defined as in Definition 2, and $(X, S, Y) \sim \mathcal{D}$. For any fairness notion of the form of Equation* (1)*, we have:*

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

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:*

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

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 1. The second claim follows by using the triangle inequality and applying the first result to each group independently. The complete proof is provided in Appendix C □

Theorem 2 mainly implies that two binary classifiers with sufficiently close decision functions will have similar fairness levels. In the next corollary, we instantiate Theorem 2 for various popular group fairness notions.

Corollary 1 (Popular Group Fairness Notions). *Let $h, h' \in \mathcal{H}$ be the decision functions of two binary classifiers, L defined as in Definition 2. The following holds:*

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

$$\forall (y, r) \in \mathcal{Y} \times \mathcal{S}, \; \chi_{(y,r)}(h, D) = \mathbb{E}\left(\frac{1}{|h(X)|} \middle| Y = y \right) + \mathbb{E}\left(\frac{1}{|h(X)|} \middle| Y = y, S = r \right) \;\; .$$

***Equality of Opportunity (Hardt et al., 2016)**: the data is divided into $K = |\mathcal{S}|$ groups such that*

$$\forall (r) \in \mathcal{S}, \; \chi_{(r)}(h, D) = \mathbb{E}\left(\frac{1}{|h(X)|} \middle| Y = 1 \right) + \mathbb{E}\left(\frac{1}{|h(X)|} \middle| Y = 1, S = r \right) \;\; .$$

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

$$\forall (y, r) \in \mathcal{S}, \; \chi_{(y,r)}(h, D) = \mathbb{E}\left(\frac{1}{|h(X)|} \right) + \mathbb{E}\left(\frac{1}{|h(X)|} \middle| S = r \right) \;\; .$$

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

$$\forall (y, r) \in \mathcal{Y} \times \mathcal{S},\ \chi_{(y,r)}(h, D) = \mathbb{E}\left(\frac{1}{|h(X)|}\right) + \mathbb{E}\left(\frac{1}{|h(X)|}\middle| S = r\right) \ .$$

Proof. This corollary follows from Theorem 2 by replacing the constants $C_k^{k'}$ by appropriate values that depend on the fairness notion under consideration and are detailed in Appendix A. □

Corollary 1 shows that our results are applicable to several popular fairness notions and that the pointwise Lipschitz constant $\chi_k(h, D)L$ depends on the fairness notion under consideration. In Section 4, we will show that these results may be used to quantify the relative fairness level between private and non-private models.

Remark 3. *Interestingly, Theorem 1 is also applicable to accuracy. Thus, for two decision functions* $h, h' \in \mathcal{H}$*, and* L *defined as in Definition 2, accuracy satisfies*

$$|\mathrm{Acc}(h) - \mathrm{Acc}(h')| \leq \mathbb{E}\left(\frac{1}{|h(X)|}\right) L \left\|h - h'\right\|_{\mathcal{H}} \ .$$

4 Quantifying the Relative Fairness of Private Models

In this section, we aim at quantifying the fairness level of a privately learned model, compared to its non-private counterpart. Assume that the set of hypothesis functions $\mathcal{H}$ is parameterized by vectors from a closed convex subset of $\mathbb{R}^p$, where $h \in \mathcal{H}$ interchangeably denotes the function and its parameters. We have that $\|\cdot\|_{\mathcal{H}} = \|\cdot\|_2$. 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, β-smooth, μ-strongly-convex with respect to its first variable. We define $h^* \in \mathcal{H}$ as the parameters of the model learned by empirical risk minimization:

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

When clear from the context, we will omit the dependence of f on D. Let h^{priv} be a differentially private approximation of h^*. To guarantee differential privacy, the choice of h^{priv} must be randomized. Nevertheless, we can typically derive high-probability bounds on the distance between h^* and its private approximation h^{priv}. In the remainder of this section, we recall such bounds for two widely used private algorithms: output perturbation (Chaudhuri et al., 2011), and DP-SGD (Bassily et al., 2014). We then use these bounds together with our results of Section 3.2 to bound the fairness level of the private solution h^{priv} in comparison to the one of the true solution h^*.

4.1 Two Private Mechanisms for DP-ERM

In this section, we describe two common differentially private mechanisms for solving problem (3). *Output perturbation* computes the true solution to (3), and releases a private estimate through the Gaussian mechanism. In contrast, *DP-SGD* uses iterative updates with privatized gradients. For each mechanism, we give a high-probability bound on the distance from the released private parameters to the optimal (non-private) ones.

Output Perturbation. The output perturbation mechanism, as proposed and analyzed by Chaudhuri et al. (2011); Lowy and Razaviyayn (2021), computes the non-private solution h^* of (3), and releases it privately using the Gaussian mechanism:

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

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

Lemma 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 $1 - \zeta$,*

$$\left\|h^{\text{priv}} - h^*\right\|_2^2 \leq 4p\sigma^2 \log(2/\zeta) = \frac{32p\Lambda^2 \log(1.25/\delta) \log(2/\zeta)}{\mu^2 n^2 \epsilon^2} \ .$$

DP-SGD. Differentially Private Stochastic Gradient Descent (DP-SGD) starts from a model $h^0 \in \mathcal{H}$ and updates them using stochastic gradients, computed on one data record at a time:

$$h^{t+1} = \pi_{\mathcal{H}}(h^t - \gamma(\nabla \ell(h^t; x_i, y_i) + \eta^t)), \qquad \text{with } i \sim [n], \eta^t \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_p) \ .$$

After $T > 0$ iterations, DP-SGD releases $h^{\text{priv}} = h^T$. Then, given $0 < \epsilon, \delta < 1$, h^{priv} is (ϵ, δ)-differentially private when $\sigma^2 \geq {64\Lambda^2 T^2 \log(3T/\delta) \log(2/\delta)}/{n^2 \epsilon^2}$. We bound the distance between h^{priv} and h^* (with high probability) in Lemma 2.

Lemma 2. *Let h^{priv} be the vector released by DP-SGD with appropriate noise to ensure (ϵ, δ)-differential privacy. 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).

Remark 4. *For clarity of exposition in Lemma 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, and strong convexity and smoothness assumptions can be relaxed.*

4.2 Bounding the Fairness of Private Models

We now use the bounds from Lemma 1 and Lemma 2 to control the loss in fairness levels due to privacy by plugging them into Theorem 2, which gives Theorem 3.

Theorem 3. *Let h^{priv} be a private estimation obtained through output perturbation of the optimal non-private solution h^* of (3), then given $h^{\text{ref}} \in \{h^*, h^{\text{priv}}\}$ we have, with probability at least $1 - \zeta$*

$$\left|\mathit{Fair}(h^{\text{priv}}, D) - \mathit{Fair}(h^*, D)\right| \leq \frac{L\Lambda\sqrt{32p \log(1.25/\delta) \log(2/\zeta)}}{K\mu n\epsilon} \sum_{k=1}^{K} \chi_k(h^{\text{ref}}, D) \ .$$

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

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

where $\widetilde{O}$ ignores logarithmic terms other than $\log(1/\delta)$.

This result shows that, when learning a private model, the unfairness due to privacy vanishes at a $\widetilde{O}({\sqrt{p}}/{n})$ rate. To our knowledge, our result is the first to quantify this rate. This result may be interpreted and used in two ways depending on the value of h^{ref}:

- **Assuming that $h^{\text{ref}} = h^{\text{priv}}$.** In this case, the private model is known *but its optimal non-private counterpart is not.* There, our result can certify, given enough examples, that 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.
- **Assuming that $h^{\text{ref}} = h^*$.** There, 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.

5 Conclusion

Our main result states that the unfairness incurred by privacy in differentially private binary classification vanishes at a rate of $\widetilde{O}(\sqrt{p}/n)$, where n is the number of training records, and p the number of parameters. This rate can be used as a certificate that a private model is not overly more unfair than the best non-private model, even when the latter is unknown. To obtain this result, we showed that several group fairness notions are pointwise Lipschitz with respect to the parameters of the model. This is the consequence of a more general result stating that the probability for a decision function to make a negative prediction conditioned on an arbitrary event (such as membership to a sensitive group) is pointwise Lipschitz continuous in the decision function.

We believe that such certificates can benefit privacy preserving machine learning by guaranteeing that learned models are satisfactory both in terms of accuracy and fairness. They are a step towards building machine learning models that the end user can trust, and could guide the design of fairer privacy-preserving machine learning algorithms. Conversely, they could be misused by malicious parties, who may try to trick uninformed users into thinking their models are fair and privacy preserving in situations where the number of available data records were insufficient to obtain meaningful certificates. Also, note that we do not address the problem of learning fair and private models but rather derive a guarantee that private models are not overly unfair *in comparison with non-private models.*

As of now, we only considered two algorithms for DP-ERM that require strongly-convex and smooth objective functions. While our results already cover a wide range of approaches, they are not compatible with some widely used models such as neural networks. Bridging this gap would make the derived results even more appealing. Similarly, these two mechanisms were not designed to impose fairness constraints. Thus, extending our results to fair methods for DP-ERM would be an important next step. Finally, our results theoretically show that, as the number of examples increases, the fairness loss due to the privacy preserving mechanism vanishes. However, we did not numerically assess the tightness of our results. All of these make very interesting directions for future work.

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., 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. 127 citations (Crossref) [2022-08-19].

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*, (3):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.

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.

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.

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

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. Publisher: Society for Industrial and Applied Mathematics.

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.

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)*, pages 263–275. 117 citations (Crossref) [2022-08-19] arXiv: 1702.07476.

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

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*.

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.

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) and the proofs of the main theoretical results that were omitted in the main paper for the sake of readability (Appendices B to E).

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, $\forall (y,r) \in \mathcal{Y} \times \mathcal{S}$*

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

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

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

with

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

Proof. We have that

$$\begin{aligned} &F_{(-1,r)}(h,D) \\ &= \mathbb{P}\left(\text{sign}(h(X)) = \text{sign}(Y) \mid Y = -1, S = r\right) - \mathbb{P}\left(\text{sign}(h(X)) = \text{sign}(Y) \mid Y = -1\right) \\ &= \mathbb{P}\left(h(X) < 0 \mid Y = -1, S = r\right) - \mathbb{P}\left(h(X) < 0 \mid Y = -1\right) \\ &= \mathbb{P}\left(h(X) < 0 \mid Y = -1, S = r\right) - \sum_{r' \in \mathcal{S}} \mathbb{P}\left(h(X) < 0 \mid Y = -1, S = r'\right) \mathbb{P}(S = r' \mid Y = -1) \end{aligned}$$

and

$$\begin{aligned} F_{(1,r)}(h,D) &= \mathbb{P}\left(\text{sign}(h(X)) = \text{sign}(Y) \mid Y = 1, S = r\right) - \mathbb{P}\left(\text{sign}(h(X)) = \text{sign}(Y) \mid Y = 1\right) \\ &= \mathbb{P}\left(h(X) \geq 0 \mid Y = 1, S = r\right) - \mathbb{P}\left(h(X) \geq 0 \mid Y = 1\right) \\ &= \sum_{r' \in \mathcal{S}} \mathbb{P}\left(h(X) < 0 \mid Y = 1, S = r'\right) \mathbb{P}(S = r' \mid Y = 1) - \mathbb{P}\left(h(X) < 0 \mid Y = 1, S = r\right) \end{aligned}$$

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 desirable outcome $Y = 1$, that $\forall (r) \in \mathcal{S}$*

$$F_{(r)}(h,D) = \mathbb{P}\left(h(X) \geq 0 \mid Y = 1, S = r\right) - \mathbb{P}\left(h(X) \geq 0 \mid Y = 1\right).$$

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

$$F_{(r)}(h, D) = C^0_{(r)} + \sum_{(r')\in\mathcal{S}} C^{(r')}_{(r)} \mathbb{P}\left(h(X) < 0 | Y = 1, S = r'\right) \tag{5}$$

with

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

Proof. We have that

$$\begin{aligned} F_{(r)}(h, D) &= \mathbb{P}\left(h(X) \geq 0 \mid Y = 1, S = r\right) - \mathbb{P}\left(h(X) \geq 0 \mid Y = 1\right) \\ &= \sum_{r'\in\mathcal{S}} \mathbb{P}\left(h(X) < 0 \mid Y = 1, S = r'\right) \mathbb{P}(S = r' \mid Y = 1) - \mathbb{P}\left(h(X) < 0 \mid Y = 1, S = r\right) \end{aligned}$$

which gives 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, $\forall (y, r) \in \mathcal{Y} \times \mathcal{S}$*

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

Note that, in this case, we have $F_{(1,r)}(h, D) = F_{(-1,r)}(h, D)$, that is we use redundant constraints. This is a simple trick to fit the setting of Equation (1) that does not change the result of aggregate fairness defined in Equation (2). We can then write $F_{(y,r)}(h, D)$ in the form of Equation (1) as

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

with

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

Proof. We have that

$$\begin{aligned} F_{(y,r)}(h, D) &= \mathbb{P}\left(\text{sign}(h(X)) = \text{sign}(Y) \mid S = r\right) - \mathbb{P}\left(\text{sign}(h(X)) = \text{sign}(Y)\right) \\ &= \mathbb{P}\left(h(X) \geq 0 \mid Y = 1, S = r\right) \mathbb{P}\left(Y = 1 \mid S = r\right) \\ &\quad + \mathbb{P}\left(h(X) < 0 \mid Y = -1, S = r\right) \mathbb{P}\left(Y = -1 \mid S = r\right) \\ &\quad - \mathbb{P}\left(h(X) \geq 0 \mid Y = 1\right) \mathbb{P}\left(Y = 1\right) \\ &\quad - \mathbb{P}\left(h(X) < 0 \mid Y = -1\right) \mathbb{P}\left(Y = -1\right) \\ &= \mathbb{P}\left(Y = 1 \mid S = r\right) - \mathbb{P}\left(h(X) < 0 \mid Y = 1, S = r\right) \mathbb{P}\left(Y = 1 \mid S = r\right) \\ &\quad + \mathbb{P}\left(h(X) < 0 \mid Y = -1, S = r\right) \mathbb{P}\left(Y = -1 \mid S = r\right) \\ &\quad - \mathbb{P}\left(Y = 1\right) + \mathbb{P}\left(h(X) < 0 \mid Y = 1\right) \mathbb{P}\left(Y = 1\right) \\ &\quad - \mathbb{P}\left(h(X) < 0 \mid Y = -1\right) \mathbb{P}\left(Y = -1\right) \end{aligned}$$

$$
\begin{aligned}
&= \mathbb{P}(Y=1 \mid S=r) - \mathbb{P}(h(X)<0 \mid Y=1, S=r)\,\mathbb{P}(Y=1 \mid S=r) \\
&\quad + \mathbb{P}(h(X)<0 \mid Y=-1, S=r)\,\mathbb{P}(Y=-1 \mid S=r) \\
&\quad - \mathbb{P}(Y=1) + \sum_{(r')\in\mathcal{S}} \mathbb{P}(h(X)<0 \mid Y=1, S=r')\,\mathbb{P}(Y=1, S=r') \\
&\quad - \sum_{(r')\in\mathcal{S}} \mathbb{P}(h(X)<0 \mid Y=-1, S=r')\,\mathbb{P}(Y=-1, S=r') \\
&= \mathbb{P}(Y=1 \mid S=r) - \mathbb{P}(Y=1) \\
&\quad - \mathbb{P}(h(X)<0 \mid Y=1, S=r)\,\mathbb{P}(Y=1 \mid S=r) \\
&\quad + \sum_{(r')\in\mathcal{S}} \mathbb{P}(h(X)<0 \mid Y=1, S=r')\,\mathbb{P}(Y=1, S=r') \\
&\quad + \mathbb{P}(h(X)<0 \mid Y=-1, S=r)\,\mathbb{P}(Y=-1 \mid S=r) \\
&\quad - \sum_{(r')\in\mathcal{S}} \mathbb{P}(h(X)<0 \mid Y=-1, S=r')\,\mathbb{P}(Y=-1, S=r')
\end{aligned}
$$

which gives the result. □

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

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

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

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

with

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

Proof. We have that

$$
\begin{aligned}
F_{(y,r)}(h,D) &= \mathbb{P}(\text{sign}(h(X)) = y \mid S=r) - \mathbb{P}(\text{sign}(h(X)) = y) \\
&= y\,\mathbb{P}(h(X)<0) - y\,\mathbb{P}(h(X)<0 \mid S=r) \\
&= y \sum_{(r')\in\mathcal{S}} \mathbb{P}(h(X)<0 \mid S=r')\,\mathbb{P}(S=r') - y\,\mathbb{P}(h(X)<0 \mid S=r)
\end{aligned}
$$

which gives the result. □

B Proof of Theorem 1

Theorem (Pointwise Lipschitzness of Conditional Negative Predictions). *Let $\mathcal{H}$ be a set of real valued functions with L Lipschitz constant as defined in Definition 2. Let $h, h' \in \mathcal{H}$ be two models, (X,Y,S) be a triple of random variables sampled from $\mathcal{D}$, and E be an arbitrary event. Assume that $\mathbb{E}\left({}^{1}/_{|h(X)|}\right) < +\infty$, then*

$$
|\mathbb{P}(h(X)<0 \mid E) - \mathbb{P}(h'(X)<0 \mid E)| \leq \mathbb{E}\left(\frac{1}{|h(X)|}\middle| E\right) L\,\|h-h'\|_{\mathcal{H}}\quad .
$$

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 $|\mathbb{P}(h(X) < 0 \mid E) - \mathbb{P}(h'(X) < 0 \mid E)| \leq \mathbb{P}(|h(X)| \leq L\|h - h'\|_{\mathcal{H}} \mid E)$. Then, we use the Markov's inequality to obtain the desired result.

Bounding $|\mathbb{P}(h(X) < 0 \mid E) - \mathbb{P}(h'(X) < 0 \mid E)|$. We have that

$$
\begin{aligned}
&\mathbb{P}(h(X) < 0 \mid E) - \mathbb{P}(h'(X) < 0 \mid E) \\
&\qquad = \mathbb{P}(h'(X) \geq 0 \mid E) - \mathbb{P}(h(X) \geq 0 \mid E) \\
&\qquad = \mathbb{P}(h(X) + h'(X) - h(X) \geq 0 \mid E) - \mathbb{P}(h(X) \geq 0 \mid E) \\
&\qquad = \mathbb{P}(h(X) \geq -(h'(X) - h(X)) \mid E) - \mathbb{P}(h(X) \geq 0 \mid E) \\
&\qquad \leq \mathbb{P}(h(X) \geq -|h'(X) - h(X)| \mid E) - \mathbb{P}(h(X) \geq 0 \mid E) \\
&\qquad \leq \mathbb{P}(h(X) \geq -L\|h - h'\|_{\mathcal{H}} \mid E) - \mathbb{P}(h(X) \geq 0 \mid E) \\
&\qquad = \mathbb{P}(h(X) \geq 0 \vee 0 > h(X) \geq -L\|h - h'\|_{\mathcal{H}} \mid E) - \mathbb{P}(h(X) \geq 0 \mid E) \\
&\qquad\quad \downarrow \text{ Union bound.} \\
&\qquad = \mathbb{P}(h(X) \geq 0 \mid E) + \mathbb{P}(0 > h(X) \geq -L\|h - h'\|_{\mathcal{H}} \mid E) - \mathbb{P}(h(X) \geq 0 \mid E) \\
&\qquad = \mathbb{P}(0 > h(X) \geq -L\|h - h'\|_{\mathcal{H}} \mid E) \\
&\qquad \leq \mathbb{P}(-|h(X)| \geq -L\|h - h'\|_{\mathcal{H}} \mid E) \\
&\qquad \leq \mathbb{P}(|h(X)| \leq L\|h - h'\|_{\mathcal{H}} \mid E)
\end{aligned}
$$

Similarly, we have that

$$
\begin{aligned}
&\mathbb{P}(h'(X) < 0 \mid E) - \mathbb{P}(h(X) < 0 \mid E) \\
&\qquad = \mathbb{P}(h(X) + h'(X) - h(X) < 0 \mid E) - \mathbb{P}(h(X) < 0 \mid E) \\
&\qquad = \mathbb{P}(h(X) < h(X) - h'(X) \mid E) - \mathbb{P}(h(X) < 0 \mid E) \\
&\qquad \leq \mathbb{P}(h(X) < |h'(X) - h(X)| \mid E) - \mathbb{P}(h(X) < 0 \mid E) \\
&\qquad \leq \mathbb{P}(h(X) < L\|h - h'\|_{\mathcal{H}} \mid E) - \mathbb{P}(h(X) < 0 \mid E) \\
&\qquad = \mathbb{P}(h(X) < 0 \vee 0 \leq h(X) < L\|h - h'\|_{\mathcal{H}} \mid E) - \mathbb{P}(h(X) < 0 \mid E) \\
&\qquad\quad \downarrow \text{ Union bound.} \\
&\qquad = \mathbb{P}(h(X) < 0 \mid E) + \mathbb{P}(0 \leq h(X) < L\|h - h'\|_{\mathcal{H}} \mid E) - \mathbb{P}(h(X) < 0 \mid E) \\
&\qquad = \mathbb{P}(0 \leq h(X) < L\|h - h'\|_{\mathcal{H}} \mid E) \\
&\qquad \leq \mathbb{P}(|h(X)| < L\|h - h'\|_{\mathcal{H}} \mid E) \\
&\qquad \leq \mathbb{P}(|h(X)| \leq L\|h - h'\|_{\mathcal{H}} \mid E)
\end{aligned}
$$

It implies that

$$
|\mathbb{P}(h'(X) < 0 \mid E) - \mathbb{P}(h(X) < 0 \mid E)| \leq \mathbb{P}(|h(X)| \leq L\|h - h'\|_{\mathcal{H}} \mid E)
$$

Bounding $\mathbb{P}(|h(X)| \leq L\|h - h'\|_{\mathcal{H}} \mid E)$. We use the Markov's Inequality and we assume that $\mathbb{P}(|h(X)| = 0 \mid E) = 0$. Hence, we have that

$$
\begin{aligned}
\mathbb{P}(|h(X)| \leq L\|h - h'\|_{\mathcal{H}} \mid E) &= \mathbb{P}\left(\frac{1}{|h(X)|} \geq \frac{1}{L\|h - h'\|_{\mathcal{H}}} \;\middle|\; E\right) \\
&\qquad \downarrow \text{ Markov's inequality.} \\
&\leq \mathbb{E}\left[\frac{1}{|h(X)|} \;\middle|\; E\right] L\|h - h'\|_{\mathcal{H}}
\end{aligned}
$$

It concludes the proof. □

Remark 5. *In the last step of the proof of Theorem 1, we can also use the Chernoff bound:*

$$\begin{aligned}\mathbb{P}\left(|h(X)| \leq L\|h-h'\|_{\mathcal{H}} \mid E\right) &= \mathbb{P}\left(\exp\left(-t\,|h(X)|\right) \geq \exp\left(-tL\|h-h'\|_{\mathcal{H}}\right) \mid E\right) \\ &\leq \mathbb{E}\left[\exp\left(-t\,|h(X)|\right) \mid E\right] \exp\left(tL\|h-h'\|_{\mathcal{H}}\right)\end{aligned}$$

A correct choice of t would lead to potentially tighter bounds than the Markov's inequality at the expense of readability, hence we chose to focus on the latter.

C Proof of Theorem 2

Theorem (Pointwise Lipschitzness of Fairness). *Let $h, h' \in \mathcal{H}$ be the decision functions of two binary classifiers, L be defined as in Definition 2, and $(X, S, Y) \sim \mathcal{D}$. For any fairness notion of the form of Equation* (1)*, we have:*

$$\forall k, |F_k(h, D) - F_k(h', D)| \leq \chi_k(h, D) L \left\|h - h'\right\|_{\mathcal{H}} \quad .$$

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:*

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

Proof. The first part follows $\forall k$ from the following derivation.

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

The second part is obtained thanks to the triangle inequality:

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

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

D Bound for Output Perturbation

D.1 Error via Output Perturbation

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:

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

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 $1-\zeta$.

D.2 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 $\langle \nabla f(h)-\nabla f(h'), h-h'\rangle \geq \frac{\mu}{2}\|h-h'\|^2$. 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 $\Delta = \frac{2\Lambda}{n\mu}$.

E Convergence of DP-SGD

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 $w, v \in \mathbb{R}^p$,

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

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

$$\begin{aligned}
\mathbb{E}\left\|h^T - h^*\right\|^2 &\leq (1-\gamma\mu)^T\left\|h^0 - h^*\right\|^2 + 4\gamma^2 \sum_{t=0}^{T-1}(1-\gamma\mu)^{T-t}\sigma^2 &(32)\\
&\leq (1-\gamma\mu)^T\left\|h^0 - h^*\right\|^2 + \frac{4\gamma\sigma^2}{\mu} \ . &(33)
\end{aligned}$$

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 $1-\zeta$,

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

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