Understanding the interactions between Fairness and Privacy in long-tailed data

Abstract

Fairness and Privacy are both desirable properties of machine learning algorithms. Using *Approximate differential privacy* [15] as the definition of privacy and minority group disparate accuracy [26] as the definition of fairness, we show that fairness and privacy can be at odds with each other especially when high accuracy is desired. We show theoretically that, in long-tailed data, enforcing privacy can result in poor fairness while allowing for high accuracy. In particular, the fairness gets worse as the number of subpopulations in the minority group increases. On the other hand, we show that a strict privacy regime can result in good fairness but at the cost of overall accuracy. We support our theoretical results with experimental evidence experiments using both synthetic as well as real world datasets like CelebA and CIFAR10. To summarise, our results show that i) in the regime of high accuracy, privacy is at odds with fairness and ii) in a strict privacy regime, fairness can be obtained by sacrificing accuracy.

1 INTRODUCTION

In recent years, questions about reliability of machine learning algorithms have become ever more important due to their widespread use in daily life. Fairness and privacy are two instances of such reliability traits that are desirable but often absent in modern machine learning algorithms [10, 12, 34]. As a result, there has been a flurry of recent works that aim to improve these properties in commonly used learning algorithms. However, most of these works look at these two properties individually with relatively less attention paid to how they affect each other.

Perhaps the most widespread statistical notion of privacy is that of *Differential Privacy* [15] and its slightly relaxed variant, *Approximate Differential Privacy* [16]. Despite its marginally weaker privacy guarantees, *Approximate Differential Privacy* enjoys better theoretical guarantees in terms of statistical complexity for learning [8, 20]. It is also more widely used in practice [1, 38]. Thus, we always use approximate differential privacy in this paper and for the sake of brevity, refer to it as differential privacy (DP). Intuitively, DP limits the amount of influence any single data point has on the output of the DP algorithm. This ensures that DP algorithms do not leak information about whether any particular data point was given as input to the algorithm. While, DP was initially popular as a theoretical construct, it has recently been put to practical use by large companies [18, 36] and governments [30] alike. Its popularity is largely due to its strong privacy guarantees, ease of implementation, and its quantitative nature.

There are many notions of fairness in machine learning [14, 17, 23, 25]. Minority or worst group accuracy [13, 26] and its difference from the overall accuracy is a common notion of fairness used in recent works. We define this difference as *disparate minority group accuracy* and use it to measure fairness in this paper. However, we expect our results to translate to other related definitions as well. Sagawa* et al. [32] observed that robust optimisation methods (under appropriate regularisations) obtain higher minority group accuracy but at the cost of a lower overall accuracy compared to vanilla training. Several other works [28, 35] have also observed this in practice thereby suggesting a possible trade-off between accuracy and fairness. Subsequent works including Goel et al. [22] and Menon et al. [31] have tried to minimise this trade-off.

While there has been a large body of work that aims to minimising the trade-off between accuracy and fairness and between accuracy and privacy, there is relatively little work that looks at the intersection of privacy, fairness, and accuracy. In this paper, we provide theoretical and experimental results to show that fairness and privacy can be at odds with each other especially when high accuracy is desired.

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

Contributions Our main contributions can be stated as —

- In Theorem 1 and 2, we provide asymptotic bounds for fairness and accuracy of DP algorithms showing that when high accuracy is desired, fairness is sacrificed.
- In Theorem 3, we show that in a very strict privacy regime, fairness can be achieved at the cost of accuracy.
- In Section 4, we provide experimental results using multiple architectures on synthetic as well as real world datasets (CelebA and CIFAR10) to validate our theory.

2 FAIRNESS AND PRIVACY

Differential privacy In the context of this paper, a differentially private (randomised) learning algorithm generates similar distributions over classifiers when trained over *neighbouring datasets*. Two datasets are neighbouring when they differ in one entry. However, the definition can be extended to difference in more entries using *group DP* which entails a gradual decline in the privacy guarantees. Formally,

Definition 1 (Approximate Differential Privacy [15, 16])**.** *Given any two neighbouring datasets S and S' and $\epsilon > 0, \delta \in (0,1)$ a randomised algorithm $\mathcal{A}$ is called (ϵ, δ)-differentially private if for all sets of outputs $\mathcal{Y}$, the following holds*

$$\mathbb{P}\left[\mathcal{A}\left(S\right) \in \mathcal{Y}\right] \leq e^{\epsilon}\mathbb{P}\left[\mathcal{A}\left(S'\right) \in \mathcal{Y}\right] + \delta.$$

It is now well understood that imposing this additional condition of DP while maintaining accuracy usually requires more data than is required for just achieving high accuracy. A string of theoretical works [3, 6, 9, 20] have shown that the sample complexity of learning certain concept classes privately and accurately can be arbitrarily larger than learning the same classes non-privately (i.e. with high accuracy but without privacy). On the other hand, it is easy to guarantee any arbitrary level of differential privacy if high accuracy is not desired. This can be achieved by simply composing the output of an accurate classifier with a properly calibrated *randomised response* mechanism [37]. This allows for a tradeoff between differential privacy and accuracy.

Disparate minority group accuracy Measuring worst group accuracy along with average accuracy is a common practice [13, 26, 32] for measuring fairness especially for deep learning algorithms. Consider a feature space X partitioned into a minority group X_1 [1] and a majority group X_2. Further, consider a label space Y, the ground truth label function $f : X \to \mathcal{Y}$, and a distribution Π on X.

Definition 2 (Disparate Minority Group Accuracy)**.** *Let the normalized distributions of Π on X_1 be Π_1. Then for any classifier $h : X \to \mathcal{Y}$, the disparate minority group accuracy of h is defined as*

$$\mathbb{E}_{x\sim\Pi_1}\left[h(x) \neq f(x)\right] - \mathbb{E}_{x\sim\Pi}\left[h(x) \neq f(x)\right]$$

Thus, the difference between the worst group accuracy and overall accuracy considered in previous works [13, 26] as well as *subgroup performance gap*, defined in Goel et al. [22], is equivalent to our definition of minority group disparate accuracy.

Tradeoffs between fairness and privacy Cummings et al. [11] and Agarwal [2] were one of the first to look at the impact of privacy on fairness. They constructed a distribution where any algorithm that is always fair and private will necessarily output a trivial constant classifier, thereby suggesting a tradeoff between fairness and privacy. However, there are multiple drawbacks with their work. First, they only looked at pure differential privacy which is not only theoretically more restrictive than approximate differential privacy [7, 8, 20] but also rarely used in practice. Second, their proof heavily relies on it being *pure* differential privacy and the algorithm being *always* fair; and their proofs are not amenable to relaxations of these assumptions. Further, they do not provide experiments to corroborate their theory perhaps due to the unrealistic requirements of pure DP. On the other hand, we look at approximate DP (which is a stronger result than pure DP), construct high probability bounds for both fairness and error, and provide experimental results to support our theory. Bagdasaryan et al. [5] provides some experimental evidence that DP-SGD [1] can have disparate impact on accuracy. However, unlike us, they do not discuss DP algorithms that are not DP-SGD and they do not provide any theoretical results to show the existence of the trade-off and its dependence on the subpopulation size and frequency. Perhaps, most closely related to our work is that of Feldman [19], who looks mainly at the impact of memorisation on test accuracy for long-tailed distributions. However, neither does their work foray into differential privacy nor into fairness in much detail.

3 THEORETICAL RESULTS

Before delving into the main results, we first discuss the data generation model which we will be necessary to understand the rest of this manuscript.

3.1 PRELIMINARIES

As defined above, we consider a set X to be our data domain and define the distribution $\Pi_{p,k,N}$ on the domain X .

Definition 3 ((p, N, k)-long-tailed distribution on X)**.** *Given $0 \leq p \leq 1,\ k > 1, N \in \mathbb{N}$, define two subsets i) $X_1 \subset X$ where $|X_1| = N$ and ii) $X_2 \subset X \setminus X_1$ where*

[1] X_1 is as a union of multiple smaller subpopulations

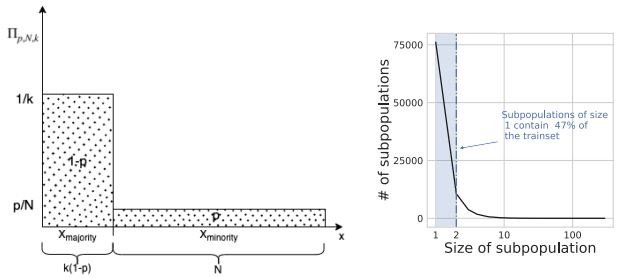

Figure 1: **(Left)**Illustration of $\Pi_{p,N,k}$. **(Right)** Frequency of occurrence of each sub-population of CelebA. Each sub-population corresponds to one combination of the 40 binary attributes of CelebA.

$|X_2| = (1-p)\,k$. [2] *Now, define the distribution* $\Pi_{p,N,k}$ *as*

$$\Pi_{p,N,k}(x) = \begin{cases} \frac{p}{N} & x \in X_1 \\ \frac{1}{k} & x \in X_2 \end{cases} \tag{1}$$

We also define the marginal distributions on the minority *subpopulation* X_1 *is defined as* $\Pi^1_{p,N}$

$$\forall \mathbf{x} \in X_1, \quad \Pi^1_{p,N}(\mathbf{x}) = \frac{\Pi_{p,N}(\mathbf{x})}{\sum_{x \in X_1} \Pi_{p,N}(\mathbf{x})} = \frac{\Pi_{p,N}(\mathbf{x})}{p} \tag{2}$$

$$\forall \mathbf{x} \notin X_1, \quad \Pi^1_{p,N}(\mathbf{x}) = 0$$

We provide an illustration of the distribution in Figure 1 (left). Note that $\Pi_{p,N,k}$ defines a valid probability distribution on X as $\frac{p}{N}N + \frac{1}{k}k\,(1-p) = 1$. Intuitively, p denotes the total probability mass of points in the minority group X_1 in the distribution $\Pi_{p,N,k}$ and N denotes the number of distinct sub-populations in the minority group. It is instructive to think of k as being much smaller than N. We will treat it as a constant throughout the paper and will, hence, remove it from the notation of the distribution for the sake of simplicity. Note that each element in X_1 has a probability mass of the order of $O\left(\frac{1}{N}\right)$ whereas elements in X_2 have a probability mass of $\frac{1}{k} = O(1) \gg O\left(\frac{1}{N}\right)$. Thus, intuitively elements in the set X_1 constitute the minority subpopulation and the set X_2 constitutes the majority subpopulation. In Figure 1, we use the 40 attributes of the CelebA dataset to divide the dataset into 2^{40} subpopulations. The blue shaded area shows that $\approx 47\%$ of the dataset is composed of examples belonging to subpopulations with density $O\left(\frac{1}{N}\right)$.

Next, we define a label space $\mathcal{Y}$ and a function space $\mathcal{F} = \mathcal{Y}^X$ for labelling functions. With slightly abuse of notation, we also use $\mathcal{F}$ to represent a distribution on this function space and refer to it as the label prior. Our results do not restrict the size of $\mathcal{Y}$ and can hence, explain both binary and multi-class classification settings.

Finally, to generate a dataset of size m from a (p, N)-long-tailed distributions $\Pi_{p,N}$ on X, first sample m i.i.d. data points $S_m = \{\mathbf{x}_1, \cdots, \mathbf{x}_m\}$ from $\Pi_{p,N}$. Then sample a labelling function $f \sim \mathcal{F}$ and label each sampled covariate, $y_i = f(\mathbf{x}_i)$, to generate the labelled dataset $S_{m,f} = \{(\mathbf{x}_1, f(\mathbf{x}_1)), \ldots, (\mathbf{x}_m, f(\mathbf{x}_m))\}$. We use the notation $S_{\#,\ell}$ to denote the set of examples that appear exactly $\ell \in \mathbb{N}$ times in $S_{m,f}$. For all our theoretical results, we will look at an asymptotic regime where both the number of subpopulations N and the dataset size m goes to ∞ in a fixed ratio $\frac{N}{m} = c \geq 0$. This is common in high-dimensional statistics where the number of dimensions often grows to ∞. Intuitively, c quantifies the hardness of the learning problem. Next, we define the error and fairness measure for an algorithm on the distribution defined above. Consider a domain X, a label space $\mathcal{Y}$, a label prior $\mathcal{F}$, and a distribution $\Pi_{p,N}$ on X as defined above.

Error The error of a classifier $h \in \mathcal{Y}^X$ on the distribution $\Pi_{p,N}$ with respect to a labelling function $f \in \mathcal{F} \subset \mathcal{Y}^X$ is

$$\mathrm{err}\left(h, \Pi_{p,N}, f\right) = \mathbb{E}_{x \sim \Pi_{p,N}}\left[h(\mathbf{x}) \neq f(\mathbf{x})\right]. \tag{3}$$

For a randomised learning algorithm $\mathcal{A}$, the error of $\mathcal{A}$ fit on a dataset $S_m \sim \Pi^m_{p,N}$, labelled with $f \in \mathcal{F}$ is

$$\overline{\mathrm{err}}_{S_m,f}\left(\mathcal{A}, \Pi_{p,N}, f\right) = \mathbb{E}_{h \sim \mathcal{A}(S_{m,f})}\left[\mathrm{err}\left(h, \Pi_{p,N}, f\right)\right] \tag{4}$$

Definition 4 (Error measure on $\Pi_{p,N}$)**.** *Finally, the error of the algorithm on the distribution* $\Pi_{p,N}$ *with respect to a label prior* $\mathcal{F}$ *is*

$$\mathrm{err}\left(\mathcal{A}, \Pi_{p,N}, \mathcal{F}\right) = \mathbb{E}_{S_m \sim \Pi^m_{p,N}, f \sim \mathcal{F}}\left[\overline{\mathrm{err}}_{S_m,f}\left(\mathcal{A}, \Pi_{p,N}, f\right)\right] \tag{5}$$

While it is slightly unconventional to take an expectation over $\mathcal{F}$, this was previously used in Feldman [19] and is in fact a stronger notion than the worst case $f \in \mathcal{F}$.

Fairness Next, we define the *disparate minority group accuracy*, represented by Γ, that we use as a measure of fairness in this work. For sake of brevity, we will refer to this as *Disparate Accuracy* in the rest of the text.

Definition 5 (Fairness measure on $\Pi_{p,N}$)**.** *For* $X, \mathcal{F}, \Pi_{p,N}$, *and* $\Pi^1_{p,N}$ *as defined above, the fairness of algorithm* $\mathcal{A}$ *on the distribution* $\Pi_{p,N}$ *with respect to the label prior* $\mathcal{F}$ *is*

$$\Gamma\left(\mathcal{A}, \Pi_{p,N}, \mathcal{F}\right) = \mathrm{err}\left(\mathcal{A}, \Pi^1_{p,N}, \mathcal{F}\right) - \mathrm{err}\left(\mathcal{A}, \Pi_{p,N}, \mathcal{F}\right). \tag{6}$$

Assumptions Given $s_0 \in \mathbb{N}$ and $p_1, p_2 \in (0,1)$, we state that $\mathcal{A}$ satisfies the assumptions A1 and A2 if the following conditions are satisfied by the algorithm $\mathcal{A}$ for all datasets S.

[2] WLOG we will assume that k is such that $(1-p)\,k$ is an integer and if not, replace k with the closest number such that $(1-p)\,k$ is an integer.

A1 (Accuracy) For all $\ell > s_0$ and $x \in S_{\#,\ell}$,

$$\mathbb{P}_{h\sim\mathcal{A}(S,f),f\sim\mathcal{F}}[h(x) \neq f(x)] \leq p_1 \tag{7}$$

A2 (Privacy) For all $\ell \leq s_0$ and $x \in S_{\#,\ell}$,

$$\mathbb{P}_{h\sim\mathcal{A}(S,f),f\sim\mathcal{F}}[h(x) \neq f(x)] > 1 - p_2 \tag{8}$$

Intuitively, A1 states that the algorithm obtains low training error on typical data. More specifically, $\mathcal{A}$ is correct on training points that appear more than s_0 times in the dataset with a high probability. The second assumption A2 states that if a training data point occurs less than s_0 times in the training dataset S, the algorithm errs on it with a high probability. We refer to this assumption as the privacy assumption because for DP algorithms, this holds true for a certain relation between ϵ, s_0, and p_2 (shown in Lemma 1). We discuss the assumptions in further detail in Section 3.2.

3.2 PRIVACY AT THE COST OF FAIRNESS

The theoretical results below use the definitions and notations described in Section 2.

Accuracy and privacy comes at the cost of fairness First, we show how the error and disparate accuracy of a DP algorithm varies for a range of distributions with varying number of subpopulations N and weight of minority group p. For easier interpretation, we show a simplified version of Theorem 1 and prove a detailed version in Appendix A.1. Consider $X, \mathcal{F}$ as defined in Section 3.1. As discussed above, the dataset size m and N both simultaneously grows to ∞ in the ratio $c = \frac{N}{m}$. To keep the writing succinct, we will thus slightly abuse the notation by replacing the notation $\Pi_{p,N}$ with $\Pi_{p,c}$.

Theorem 1. *For any $p \in (0,1), c > 0$ such that $p/c \leq 1$, consider the distribution $\Pi_{p,c}$. Also, for any $\epsilon, \delta > 0$, consider an (ϵ, δ)-DP algorithm $\mathcal{A}$ that satisfies assumptions A1* (7) *and A2* (8) *for $s_0 = o(m)$ and $p_1, p_2 \in (0,1)$. Then, with high probability,*

$$\operatorname{err}(\mathcal{A}, \Pi_{p,c}, \mathcal{F}) \leq (1-p_1)p\alpha_0$$

and

$$\Gamma(\mathcal{A}, \Pi_{p,c}, \mathcal{F}) \geq (1-p)\gamma_0$$

where

$$\alpha_0 = \left(1 - \frac{\left(\frac{2c}{p}\left(s_0 - \frac{p}{2c}\right)^2 - 1\right) e^{-\frac{c}{p}\left(s_0 - \frac{p}{2c}\right)^2}}{\sqrt{2\pi}(s_0 - \frac{p}{2c})^3 \left(\sqrt{\frac{2c}{p}}\right)^3}\right) + p_1 \text{ and}$$

$$\gamma_0 = c_0 \left(1 - \frac{\sqrt{\frac{1+p}{2c}}}{\sqrt{2\pi}\left(s_0 - \frac{1+p}{2c}\right)} e^{-(s_0 - \frac{1+p}{2c})^2 \frac{c}{1+p}}\right) - p_1$$

where c_0 is some constant depending on $\epsilon, \delta, \mathcal{F}$.

Discussion of the results In this section, we discuss how the fairness and error changes with various properties — c, p, s_0 of the learning problem, data distribution, and algorithm respectively. To interpret the results of Theorem 1, we simplify the results by applying a change of variables $x = (s_0 - \frac{p}{2c})\sqrt{\frac{2c}{p}}, y = (s_0 - \frac{1+p}{2c})\sqrt{\frac{2c}{1+p}}$, and then writing α_0, γ_0 as $\alpha_0 = 1 - \left(\frac{1}{x} - \frac{1}{x^3}\right)\frac{e^{-x^2/2}}{\sqrt{2\pi}}$ and $\gamma_0 = 1 - \frac{e^{-y^2/2}}{\sqrt{2\pi}y}$. This gives the following implications of Theorem 1.

1. As $c \to \infty$, both $x, y = \Omega(\sqrt{c})$. Consequently both $\alpha_0, \gamma_0 = 1 - O\left(\frac{e^{-c}}{\sqrt{c}}\right)$. Thus, both the upper bound for error and lower bound for fairness increases with c. Intuitively, c is the ratio of the number of distinct subpopulations compared to the size of the observed dataset and it measures the difficulty of the problem. As c goes to infinity, most points in the minority group are not included in the training dataset. Thus both the error and disparate accuracy is very large.
2. As $s_0 \to \infty$, both $\alpha_0, \gamma_0 = 1 - O\left(\frac{e^{-s_0}}{s_0}\right)$. Thus, both the upper bound for error and lower bound for fairness increases with s_0. It is reasonable to consider s_0 as a proxy for privacy. A more private algorithm will have a larger s_0 (see Lemma 1). As s_0 increases, the algorithm makes more mistakes in the minority group, leading to a larger minority group error. In contrast, the error in the majority group remains low ($\approx p_1$) as most points in the majority group appear more than s_0 times due to the assumption $s_0 = o(m)$.
3. As $\frac{1}{p} \to \infty$, both $x, y = \Omega\left(\frac{1}{\sqrt{p}}\right)$. As a result, both $\alpha_0, \gamma_0 = 1 - O\left(\sqrt{p}e^{-1/p}\right)$. The upper bound for error decreases as $O\left(p - p^{3/2}e^{-1/p}\right) = O(p)$. On the other hand, the lower bound for fairness grows as $(1-p)\left(1 - O\left(\sqrt{p}e^{-1/p}\right)\right) = 1 - p - O\left(\sqrt{p}e^{-1/p}\right)$. p indicates the total mass of the minority group in the distribution. When p is small, the minority group occupies a small mass and thus making mistakes on the minority group has a small impact on the overall error. On the other hand, p does not impact the accuracy of the minority group as $\Pi^1_{p,c}$ is independent of p. Thus the overall effect of decreasing overall error by a factor of p and no change in minority group accuracy is a decrease of disparate accuracy by a factor of $(1-p)$.

Simplified theorem with proof sketch In the previous section, we considered assumptions that allow the algorithm to generate classifiers which makes mistakes on frequently occurring examples but with a low probability and makes mistakes on infrequent samples with a high probability. This is a reasonable assumption as Lemma 1 discusses. We also provide examples of algorithms (Lemma 2) that satisfy this.

However, in this section, we consider stricter assumptions to state a more simplified and interpretable result. We show

that if a DP algorithm (with additional stricter assumptions) intends to achieve low error, there are distributions where that low error comes at the cost of fairness. We first state the theorem and then discuss the proof sketch.

Theorem 2. *For $\epsilon \in (0,1)$ and $\delta \in (0, 0.01)$, consider any (ϵ, δ)-DP algorithm $\mathcal{A}$ that does not make mistakes on points occurring more than once in the dataset. Then, for any $\alpha \in (0,1)$, there exists $p \in (0,1), c > 0$ and label priors $\mathcal{F}$ such that with probability at least 0.99,*

$$\operatorname{err}(\mathcal{A}, \Pi_{p,c}, \mathcal{F}) \leq \alpha \quad \text{and} \quad \Gamma(\mathcal{A}, \Pi_{p,c}, \mathcal{F}) \geq 0.5.$$

Theorem 2 and its proof is included in Appendix A.2. Here we present a proof sketch and discuss the results. By Definition 3, the probability mass of each point in the majority group is $O(1)$ whereas the probability mass of each point in the minority group is $O\left(\frac{1}{m}\right)$. Thus, for a large enough dataset, we show that almost all points in the majority group appear more than once and consequently, the algorithm in Theorem 2 makes very less mistakes on the majority group. As a result, the error and the disparate accuracy are both consequences of mistakes, majorly, on the minority group. We compute the number of points that does not appear or appear just once in the minority group and use that to provide the upper bounds for error and lower bound for disparate accuracy. As these bounds are expressed in terms of p and c, the problem of showing that p, c exists to obtain any α and $\Gamma \geq 0.5$ then simply reduces to solving a system of two non-linear equations in p and c.

Understanding the assumptions Here, we look at the relation between s_0 used in the assumptions of Theorem 1, the privacy parameters ϵ and δ, and the parameter p_2 in Assumption A2. In fact, Lemma 1 shows that for all values of ϵ, δ and s_0, there exists a value of $p_2 = 1 - \beta$. For interpreting Lemma 1, consider $\beta = O(1)$, let $\tilde{f}$ be a labelling function, and S be any dataset

Lemma 1. *For an ϵ, δ-differentially private algorithm $\mathcal{A}$, and for all $s_0 \in \mathbb{N}$, there exists $p_2 \in (0,1)$ such that for all x that appears less than s_0 times in the dataset S*

$$\mathbb{P}_{h \sim \mathcal{A}(S, \tilde{f})}\left[h(\mathbf{x}) \neq \tilde{f}(\mathbf{x})\right] > 1 - p_2$$

with high probability, where $p_2 = 1 - \frac{(1 - s_0 e^{(s_0 - 1)\epsilon}\delta)e^{-s_0\epsilon}}{1 + e^{-s_0\epsilon}}$.

There are two important takeaways from the above lemma. First, note that if a differentially private algorithm produces a classifier with low overall error and if $\beta = O(1)$, then s_0 is necessarily small. This justifies the use of $s_0 = o(m)$ in Theorem 1 as we are interested in algorithms that performs with reasonably good accuracy. We look at the case of $s_0 = \Omega(m)$ in Theorem 3. Second, note that for a given β, s_0, and ϵ maintains an inverse relationship. Thus, a smaller ϵ implies a larger s_0. In other words, the assumptions in Theorem 1 implies that when ϵ is small, the algorithm makes mistakes on even very frequent examples and vice versa.

We further illustrate Lemma 1 for a group of algorithms. Consider an algorithm $\mathcal{A}_\eta$ that given a dataset S (as in Theorem 1), first flips the label of each point in the training set with probability η, stores the majority label for each unique point in the training set in a dictionary, and returns this dictionary. During prediction, the dictionary is queried to output the label and if the point is not in the dictionary, a random label is returned. Lemma 2 provides privacy and accuracy guarantees for this algorithm.

Lemma 2. *The algorithm A_η is $\left(O\left(\frac{1}{\eta}\right), 0\right)$- differentially private where $O(\cdot)$ is the big-Oh notation. Further, for any $\beta_1, \beta_2 \in (0,1)$,*
• if a point appears less than $\frac{\log(1/\beta_1)}{1 + \frac{1}{2}\log[1/4\eta(1-\eta)]}$ times in the training set, then $\mathbb{P}_{h \sim \mathcal{A}}[h(x) \neq f(x)] \geq \beta_1$ and
• if a point appears more than $-\frac{8(1-\eta)}{(1-2\eta)^2}\log\beta_2$ times, then $\mathbb{P}_{h \sim \mathcal{A}}[h(x) \neq f(x)] \leq \beta_1$.

Lemma 1 and 2 are proved in Appendix A.1. Lemma 2 shows that for all $\epsilon > 0$, we can find an $\eta = O\left(\frac{1}{\epsilon}\right)$ such that $\mathcal{A}_\eta$ is $(\epsilon, 0)$-differentially private. Further, this algorithm is more accurate on frequently occurring points and inaccurate on rare points, which aligns with Lemma 1. Hence, for any $\epsilon > 0$, there is an algorithm $\mathcal{A}_\eta$ that is $(O(1/\eta), 0)$-differentially private and is accurate on points appearing more than ℓ times with probability $1 - O\left(e^{-(1-2\eta)\ell}\right)$.

3.3 FAIRNESS AT THE COST OF ACCURACY

So far we have shown that under strict privacy requirements, if high overall accuracy is desired then it might be impossible to obtain good fairness. A natural question to ask is if it is possible to sacrifice accuracy for fairness. To understand this setting, we look at Theorem 3. The primary difference between this and the previous results is that now we consider larger values of s_0, in particular $s_0 = \Omega(m)$, which as Lemma 1 discussed, results in lower overall accuracy.

We show a simplified theorem here for easier interpretation and prove a detailed version formally in Appendix Appendix A.3 along with a discussion of its result. *Our result indicates that for very strict privacy parameters, fairness can be achieved at the cost of accuracy.*

Theorem 3. *For any $p \in (0,1), c > 0$ such that $p/c \leq 1$, consider the distribution $\Pi_{p,c}$. For any $\epsilon, \delta, \alpha > 0$, consider an ϵ, δ-DP algorithm $\mathcal{A}$ that satisfies assumptions A1 (7) and A2 (8) with $s_0 = \left(\frac{2-p}{2k(1-p)}\right)m + \alpha\sqrt{m}$ and some $p_1, p_2 \in (0,1)$. Then, with probability greater than 0.99,*

$$\operatorname{err}(\mathcal{A}, \Pi_{p,c}, \mathcal{F}) \geq \tilde{c}_1 p + (1 - p_2)(1 - p)\left(1 - e^{-\frac{4(1-p)\alpha^2}{(2-p)^2}}\right)$$

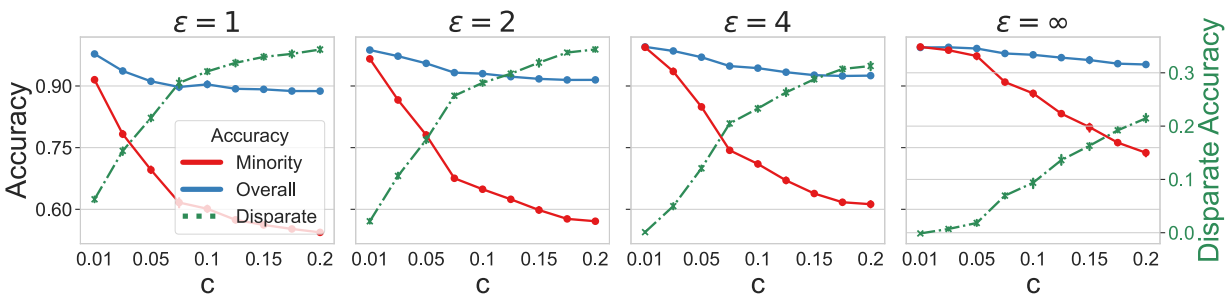

Figure 2: Each figure plots the disparate accuracy (Γ; higher is less fair) in green dashed line, the accuracy of the minority group with red, and the overall accuracy with blue on the y-axis and the parameter c in the X-axis. The left most ($\epsilon = 1$) achieves the strictest level of privacy and the right most ($\epsilon = \infty$) is vanilla training without any privacy constraints. The two figures in between achieve intermediate levels of privacy. Here $p = 0.2$. Experiment for $p = 0.5$ is in Appendix C.1

$$\Gamma\left(\mathcal{A}, \Pi_{p,c}, \mathcal{F}\right) \leq (1-p)\left[1-(1-e^{-\frac{4(1-p)\alpha^2}{(2-p)^2}})(1-p_2)\right]$$

where $\tilde{c}_1$ is a constant depending on ϵ, δ and $\mathcal{F}$.

As α increases, the lower bound for error grows and the upper bound for fairness reduces due to the term $\left(1-e^{-\alpha^2}\right)$. Note that s_0 also grows as $\Omega\left(\alpha\right)$. This shows that as examples of larger frequency gets misclassified due to increasing (s_0), the algorithm loses overall accuracy but achieves better fairness. With respect to p, the fairness upper bound decreases with increasing p, while the lower bound on error decreases with increasing p. The rate of change for both bounds is larger for larger α, where α essentially characterises how fast s_0 grows with respect to m.

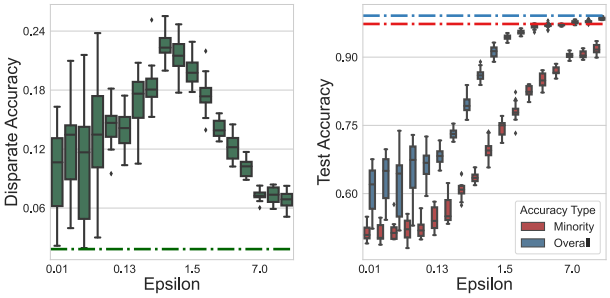

Figure 3: In the left figure, we plot the change in the disparate accuracy in green with the box plots highlighting the variance when run several times. In the right figure, the Blue boxes and the Red boxes show the overall and the minority accuracies respectively for varying ϵ. The dashed line of different colors show the respective metrics for vanilla training without privacy.

4 EXPERIMENTAL RESULTS

In this section, we look at experiments to support our theoretical arguments from Section 3. We conduct experiments on both synthetic (in Section 4.1), and real world computer vision datasets, (in Section 4.2), using deep neural networks.

4.1 SYNTHETIC EXPERIMENTS

First, we look at a synthetic data distribution that closely emulates the data distribution we use in our theoretical results in Theorem 1 and 2. Given $m \in \mathbb{N}, p \in (0, 0.5), c > 0$, and $k \in \mathbb{N}$, we construct a continuous version of the long-tailed distribution $\Pi_{p,c,m,k}$ (Definition 3) on a domain X. The domain X can be thought of as the vertices of a $O\left(\log\left(cm\right)\right)$-dimensional hypercube. The major difference is that instead of a discrete set of vertices X to sample from, here, we sample from a mixture of gaussians where each gaussian is placed on one of the randomly chosen vertices of X without replacement and the probability mass of the discrete object is equal to the total density of the gaussian placed on it. For the purpose of experiments, we choose $k = 64, m = 10^4$, vary c from 0.01 to 0.2, and choose $p \in \{0.2, 0.5\}$. We train a five-layered fully connected neural network with ReLU activations using DP-SGD [1] for varying levels of (ϵ, δ)-approximate differential privacy. We choose varying levels of ϵ while setting $\delta = 10^{-3}$. For a more detailed description of the data distribution and the training algorithm, please refer to Appendix B.

Fairness gets worse with increasing subpopulations As discussed in Section 3.1, a larger c makes the learning problem harder for the minority subpopulations, but not the majority sub-populations, which remains unaffected with the parameter c. In Figure 2, we show that increasing c hurts fairness by increasing the disparate accuracy, which is measured in the green dashed line. This corroborates our theoretical results from Theorem 1 regarding the dependence of disparate accuracy on c. Note that increasing c increases the disparate accuracy for all levels of ϵ but it hurts the most for smaller values of ϵ. It is further interesting to observe that the increase in fairness comes almost entirely from the drop in the minority accuracy, drawn in red, whereas the

overall accuracy, drawn in blue, stays relatively constant. This highlights our claim that, in the presence of strong privacy, fairness can be poor even when overall error is low.

Privacy constraints hurts fairness In this section, we analyse the dependence of fairness on ϵ for a fixed c in detail. In Figure 3 (left), we plot the disparate accuracies Γ for varying privacy parameter ϵ. There are two distinct phases in the development of the disparate accuracy with increasing ϵ. Somewhat counter-intuitively, first the disparate accuracy increases as privacy becomes less restrictive i.e. ϵ increases and then it decreases with further loosening privacy. While the second phase i.e. algorithm gets fairer as it becomes less private is expected, the first phase can be explained by looking at Figure 3 (right).

Figure 3 (right) plots the minority accuracy and the overall accuracy respectively with increasing ϵ. For a very small ϵ, the learned classifier is essentially a trivial classifier as evidenced by the very low overall accuracy ($\approx 60\%$). This is a trivial way of achieving fairness without learning an accurate classifier and is explained by Theorem 3 in our theoretical section. With loosening privacy, the classifier gets more accurate and less fairer in the first phase. The interesting regime is after the classifier obtains decent overall accuracy ($\approx 90\%$) with $\epsilon \geq 1.5$ in Figure 3 (right)). In this phase, when $\epsilon \geq 1.5$ Figure 3 (left) shows that *loosening privacy increases fairness or smaller ϵ implies larger disparate accuracy* and corresponds to Theorem 1.

4.2 EXPERIMENTS ON VISION DATASETS

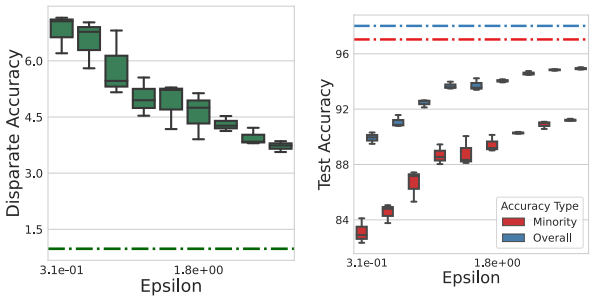

Figure 4: CelebA: **Left** Disparate accuracy decreases (low is fairer) with increasing ϵ (high is less privacy) i.e. fairness is worse for stricter privacy. **Right** Both minority and overall accuracy increases with ϵ. Dashed lines show the corresponding metric of a vanilla model (no privacy criterion).

In this section, we show that our claims are evident not just in theory and synthetic experiments but also in real world computer vision datasets. In particular, we conduct further experiments on two popular computer vision datasets — CelebA [29] and CIFAR10 [27]. CelebA is a dataset of approximately $160k$ training images of dimension 178×218 and another $20k$ of the same dimension for testing. CIFAR10 is a 10-class classification dataset where there are $50k$ training images and $10k$ test images of dimension $3 \times 32 \times 32$. For CIFAR10, we use a ResNet-18 [24] and for CelebA, we use a ResNet-50 architecture.

4.2.1 Minority and Majority groups

While the construction of the synthetic data distribution discussed above provides an easy definition of subpopulations, datasets like CelebA and CIFAR10 do not come with a label of what constitutes a subpopulations. In this section, we describe how we create the various subpopulations. Note that the creation of the sub-population is done before any private training and is thus agnostic to it.

CelebA Along with images, the CelebA dataset provides 40 attributes corresponding to each image which records characteristics like gender, hair color, facial hair etc. We create a binary classification problem by using the gender attribute as the target label. In addition, we use 11 of the remaining 39 binary attributes to create 2^{11} subpopulations and categorise each example into one of these 2^{11} sub-populations. Then, we create various versions of minority groups by aggregating the examples of all the unique sub-populations that have less than s examples in the test set, where $s \in \{5, 10, 20, 40, 60, 80, 100\}$. The remaining examples constitute the majority group. In this section, we run experiments using $s = 40$. We report results for rest of the subpopulation sizes in Appendix C.2.

CIFAR10 Unlike the synthetic distribution and CelebA as described above, CIFAR10 does not have a partition of majority and minority subpopulations. However, that does not mean that there aren't sub-populations in CIFAR10. In fact, recent works [33, 39] have shown the presence of these sub-populations in CIFAR-10 but in the context of influence functions and adversarial training respectively. We use the influence score estimates from Zhang and Feldman [39] to create the minority and majority groups.

First, we filter out examples from the trainset whose self-influence [19, 39] is larger than a threshold ρ. These examples constitute the different sub-populations in the trainset and together form the minority group in the trainset. Then, we filter out examples from the test set that are heavily influenced (influence score is greater than the threshold) by at least one of the minority examples in the trainset, as identified above. These examples constitute the minority group in the testset. In this section, we report results with $\rho = 0.1$. Other values of ρ show a similar trend and we plot results using $\rho = 0.01$ in Appendix C.3.

4.2.2 Stricter Privacy leads to worse fairness

In this section, we use the above definitions of minority and majority groups to measure the impact of privacy on

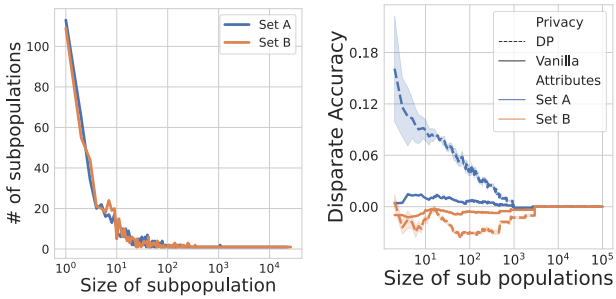

Figure 5: **Left** Both sets of attributes induces a similar distribution over sizes of subpopulations on CelebA. **Right** Set A has high disparate accuracy (worse fairness) for small sized subpopulations whereas Set B does not.

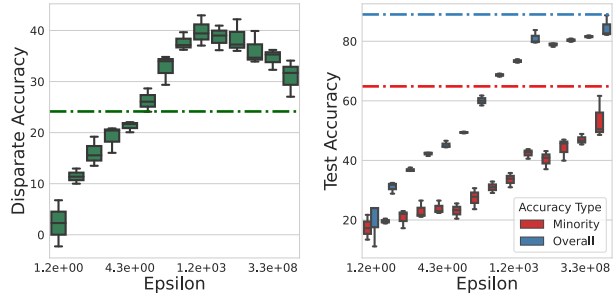

Figure 6: CIFAR10—**Left** Disparate accuracy decreases (low is fairer) with increasing ϵ (high is less privacy) i.e. fairness is worse for stricter privacy. **Right** figure shows that both minority and overall accuracy increases with ϵ. Dashed lines show the corresponding metric of a vanilla model (no privacy criterion).

fairness (using Definition 5). Like Section 4.1, we measure both the disparate accuracy and the individual minority and overall accuracies.

CelebA Figure 4 plots the change in fairness, measured with disparate accuracy, with respect to the ϵ parameter of differential privacy. To recall, a smaller ϵ indicates stricter privacy and a large disparate accuracy indicates less fairness. Figure 4 (left) show that smaller ϵ implies a larger disparate accuracy. This aligns with our theoretical results from Section 3. However, note that unlike Figure 3 (left), the disparate accuracy here monotonically decreases with increasing ϵ. Figure 4 (right) shows that the overall accuracy for CelebA is already very high ($\approx 85\%$) unlike Figure 3 (right) where the strictest privacy levels witnesses an overall accuracy close to the trivial accuracy of 50%.

Are all subpopulations equal ? The previous section shows that when the minority group is composed of small sized subpopulations, differential privacy requirements hurt the fairness of the algorithm. Here, we highlight that not all small-sized sub-populations are hurt equally in this process. In Figure 5, we show that a different partition of CelebA composed of similar sized populations do not show similar behaviour in terms of how disparate accuracy changes with sizes of sub-populations. We will refer to the 11 attributes we chose to partition the testset for our experiments so far as *Set A* and here we choose another set of 11 attributes and refer to them as *Set B*. Figure 5 (left) shows that both Set A and Set B induces a very similar distribution over sizes of subpopulations on the test set. However, Figure 5 (right) shows that while the minority groups induced by Set A suffers very high disparate accuracy from private training compared to vanilla training, Set B shows very minor difference in disparate accuracy between private and vanilla training. Please see Appendix C.2 for more details on Set A and Set B. *This indicates that private training hurts fairness disproportionately more for some subpopulations compared to other subpopulations. In particular, this begs further research on where these minority groups that are worse-affected by private training intersects with the minority groups that are relevant for the specific domain.*

CIFAR10 Figure 6 (left) shows the change in fairness with respect to privacy. Similar to the synthetic experiments in Figure 3 and unlike CelebA, we observe two phases in the develop of fairness with respect to privacy. For small values of ϵ, Figure 6 (left) shows that the disparate accuracy increases with ϵ, indicating that fairness gets worse as privacy becomes more lenient. However, Figure 6 (right) shows that the overall accuracy in this regime is extremely poor. Very small values of ϵ heavily impacts the accuracy of the learnt classifier. It suggests that fairness can be achieved with privacy but at the cost of accuracy. We explore this theoretically in Section 3.3 providing experimental proof of teh regime discussed in Theorem 3 on a real world dataset.

However, if we only restrict ourselves to classifiers that have a high overall accuracy (say $\geq 70\%$), we can see that a small value of ϵ increases the disparate accuracy and thus, has a detrimental impact on fairness. This corresponds to the right side of both figures in Figure 6 and aligns with the empirical observations in Figure 3 and 4 and the theoretical results of Theorem 1.

5 CONCLUSION

In this paper, we have shown theoretically that when the minority group in the data is composed of multiple subpopulations, a DP algorithm can achieve very low error but will necessarily incur worse fairness. Further, we corroborate our theoretical results with experimental evidence on synthetic and real world computer vision datasets. However, as our theoretical results deal with a pessimistic worse case scenario, it is possible that in real world datasets, it is possible to achieve a more optimistic trade-off. This begs further research to develop fair and private algorithms that get nearer to the pareto optimal frontier.

REFERENCES

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, oct 2016. doi: 10.1145/2976749.2978318.

[2] S. Agarwal. Trade-offs between fairness and privacy in machine learning. 2021.

[3] N. Alon, R. Livni, M. Malliaris, and S. Moran. Private pac learning implies finite littlestone dimension. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 852–860, 2019.

[4] R. B. Ash. *Information theory*. Interscience tracts in pure and applied mathematics no. 19. Interscience Publishers, New York, NY [etc, [1st ed. 1965, 3rd printing 1967] edition, 1967 - 1965.

[5] E. Bagdasaryan, O. Poursaeed, and V. Shmatikov. Differential privacy has disparate impact on model accuracy. *Advances in Neural Information Processing Systems*, 32:15479–15488, 2019.

[6] A. Beimel, S. P. Kasiviswanathan, and K. Nissim. Bounds on the sample complexity for private learning and private data release. In *Theory of Cryptography*, pages 437–454. Springer Berlin Heidelberg, 2010. doi: 10.1007/978-3-642-11799-2_26.

[7] A. Beimel, K. Nissim, and U. Stemmer. Characterizing the sample complexity of private learners. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 97–110, 2013.

[8] A. Beimel, K. Nissim, and U. Stemmer. Private learning and sanitization: Pure vs. approximate differential privacy. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 363–378. Springer, 2013.

[9] M. Bun, R. Livni, and S. Moran. An equivalence between private classification and online prediction. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 389–402. IEEE, 2020.

[10] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[11] R. Cummings, V. Gupta, D. Kimpara, and J. Morgenstern. On the compatibility of privacy and fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization - UMAP'19 Adjunct*. ACM Press, 2019. doi: 10.1145/3314183.3323847.

[12] T. De Vries, I. Misra, C. Wang, and L. Van der Maaten. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 52–59, 2019.

[13] M. Du, S. Mukherjee, G. Wang, R. Tang, A. Awadallah, and X. Hu. Fairness via representation neutralization. *Advances in Neural Information Processing Systems*, 34, 2021.

[14] J. C. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

[15] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology - EUROCRYPT 2006*, pages 486–503. Springer Berlin Heidelberg, 2006. doi: 10.1007/11761679_29.

[16] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pages 265–284. Springer Berlin Heidelberg, 2006. doi: 10.1007/11681878_14.

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

[18] G. Fanti, V. Pihur, and Ú. Erlingsson. Building a rappor with the unknown: Privacy-preserving learning of associations and data dictionaries. *Proceedings on Privacy Enhancing Technologies (PoPETS)*, issue 3, 2016, 2016.

[19] V. Feldman. Does learning require memorization? a short tale about a long tail. *arXiv:1906.05271*, 2019.

[20] V. Feldman and D. Xiao. Sample complexity bounds on differentially private learning via communication complexity. In *Conference on Learning Theory*, pages 1000–1019. PMLR, 2014.

[21] V. Feller. *An introduction to probability theory and its applications. Vol. 1*. Feller, William «An» introduction to probability theory and its applications 1 (RERO)0141384-41slsp. J. Wiley, New York [etc, 3rd ed, revised print edition, 1970.

[22] K. Goel, A. Gu, Y. Li, and C. Ré. Model patching: Closing the subgroup performance gap with data augmentation. Aug. 2020.

[23] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. volume 29, pages 3315–3323, 2016.

[24] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, jun 2016. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.90. URL `http://ieeexplore.ieee.org/document/7780459/`.

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

[26] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[27] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[28] E. Z. Liu, B. Haghgoo, A. S. Chen, A. Raghunathan, P. W. Koh, S. Sagawa, P. Liang, and C. Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021.

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

[30] G. Long. Formal privacy methods for the 2020 census. 2020.

[31] A. K. Menon, A. S. Rawat, and S. Kumar. Overparameterisation and worst-case generalisation: friend or foe? In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=jphnJNOwe36`.

[32] S. Sagawa*, P. W. Koh*, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=ryxGuJrFvS`.

[33] A. Sanyal, P. K. Dokania, V. Kanade, and P. H. S. Torr. How benign is benign overfitting? In *International Conference on Learning Representations*, 2021.

[34] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017.

[35] N. Sohoni, M. Sanjabi, N. Ballas, A. Grover, S. Nie, H. Firooz, and C. Ré. Barack: Partially supervised group robustness with guarantees. Dec. 2021.

[36] J. Tang, A. Korolova, X. Bai, X. Wang, and X. Wang. Privacy loss in apple's implementation of differential privacy on macos 10.12. Sept. 2017.

[37] S. L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, mar 1965. doi: 10.1080/01621459.1965.10480775.

[38] A. Yousefpour, I. Shilov, A. Sablayrolles, D. Testuggine, K. Prasad, M. Malek, J. Nguyen, S. Ghosh, A. Bharadwaj, J. Zhao, G. Cormode, and I. Mironov. Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298*, 2021.

[39] C. Zhang and V. Feldman. What neural networks memorize and why: Discovering the long tail via influence estimation. In *Conference on Neural Information Processing Systems*, 2020.

A PROOFS

First, we look at a decomposition of error (Definition 4) and fairness (Definition 5) to measure the contribution of of the majority and the minority subpopulation. From Definition 4, we get

$$\begin{aligned}
\operatorname{err}(\mathcal{A}, \Pi_{p,c}, \mathcal{F}) =& \mathbb{E}_{h\sim\mathcal{A}(S_m,f), f\sim\mathcal{F}}\left[\sum_{x\in X} \Pi_{p,c}(x)\mathbb{1}\{h(x)\neq f(x)\}\right] \\
=& \mathbb{E}_{h\sim\mathcal{A}(S_m,f), f\sim\mathcal{F}}\left[\sum_{x\in X_1} p\Pi^1_{p,c}(x)\mathbb{1}\{h(x)\neq f(x)\}\right] \\
&+ \mathbb{E}_{h\sim\mathcal{A}(S_m,f), f\sim\mathcal{F}}\left[\sum_{x\in X_2} (1-p)\Pi^1_{p,c}(x)\mathbb{1}\{h(x)\neq f(x)\}\right] \qquad (9)
\end{aligned}$$

Then using Definition 5

$$\begin{aligned}
\Gamma(\mathcal{A}, \Pi_{p,c}, \mathcal{F}) =& \operatorname{err}\left(\mathcal{A}, \Pi^1_{p,c}, \mathcal{F}\right) - \operatorname{err}(\mathcal{A}, \Pi_{p,c}, \mathcal{F}) \\
=& \mathbb{E}_{h\sim\mathcal{A}(S_m,f), f\sim\mathcal{F}}\left[\sum_{x\in X_1} \Pi^1_{p,c}(x)\mathbb{1}\{h(x)\neq f(x)\}\right] \\
&- \mathbb{E}_{h\sim\mathcal{A}(S_m,f), f\sim\mathcal{F}}\left[\sum_{x\in X_1} p\Pi^1_{p,c}(x)\mathbb{1}\{h(x)\neq f(x)\}\right] \qquad \text{Using (9)} \\
&- \mathbb{E}_{h\sim\mathcal{A}(S_m,f), f\sim\mathcal{F}}\left[\sum_{x\in X_2} (1-p)\Pi^1_{p,c}(x)\mathbb{1}\{h(x)\neq f(x)\}\right] \\
=& (1-p)\left[\operatorname{err}\left(\mathcal{A}, \Pi^1_{p,c}, \mathcal{F}\right) - \operatorname{err}\left(\mathcal{A}, \Pi^2_{p,c}, \mathcal{F}\right)\right] \qquad (10)
\end{aligned}$$

We will state two existing results on concentration bounds of Bernoulli distribution and Gaussian distributions. First, we look at an anti-concentration bound of a Binomial random variable[4].

Inequality 1 (From Ash [4])**.** *Let X be a binomial random variable with parameters (m, p), then for any $k < mp$, the anti-concentration bound of X is*

$$\mathbb{P}[X \leq k] \geq \frac{1}{\sqrt{2m}} e^{-mD(\frac{k}{m}||p)}$$

where $D(a||p)$ represents the Kullback-Leibler divergence between two Bernoulli random variables with parameters a and p,

$$D(a||p) = a\log\frac{a}{p} + (1-a)\log\frac{1-a}{1-p}$$

Then, we look at the tail bound for Gaussian distribution[21].

Inequality 2 (From Feller [21])**.** *Let W be a random variable following $N(\mu, \sigma^2)$, then we have the following tail bound on X*

$$\left(\frac{1}{x} - \frac{1}{x^3}\right)\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} \leq \mathbb{P}[W \geq \mu + \sigma x] \leq \frac{e^{-\frac{x^2}{2}}}{x\sqrt{2\pi}}$$

We also use the following lemma frequently to lower bound the number of samples in the minority and the majority group with high probability.

Lemma 3. *Let m_1, m_2 be two binomial random variables with parameters (m, p) and $(m, 1-p)$ respectively. Let m be the sum of m_1 and m_2, then, the probability that $m_1 \geq \frac{pm}{2}$ and $m_2 \geq \left(\frac{1-p}{2}\right)m$ is at least 0.99.*

Proof. As m_1 is a binomial distribution with parameter (m, p), it could be estimated with a Gaussian distribution with mean mp and variance $mp(1-p)$ by Central Limit Theorem (CLT). Then, the probability that $m_1 \geq \frac{pm}{2}$ and $m_2 \geq \left(\frac{1-p}{2}\right)m$ is lower bounded

$$
\begin{aligned}
\mathbb{P}\left[m_1 \geq \frac{mp}{2} \cap m_2 \geq \left(\frac{1-p}{2}\right) m\right] &= \mathbb{P}\left[\frac{mp}{2} \leq m_1 \leq \left(\frac{1+p}{2}\right) m\right] \\
&= 1 - \mathbb{P}\left[m_1 \leq \frac{mp}{2}\right] - \mathbb{P}\left[m_1 \geq \left(\frac{1+p}{2}\right) m\right] \\
&= 1 - \mathbb{P}\left[m_1 \leq \mathbb{E}(m_1) - \sqrt{\frac{mp}{4(1-p)}\text{Var}(m_1)}\right] \\
&\quad - \mathbb{P}\left[m_1 \geq \mathbb{E}(m_1) + \sqrt{\frac{m(1-p)}{4p}\text{Var}(m_1)}\right] \\
&= 1 - \mathbb{P}\left[|m_1 - \mathbb{E}(m_1)| \geq \sqrt{\frac{m(1-p)}{4(1-p)}\text{Var}(m_1)}\right] \quad \text{By Chebyshev's inequality} \\
&\geq 0.99
\end{aligned}
\tag{11}
$$

□

Next, we restate and prove Lemma 1.

Lemma 1. *For an ϵ, δ-differentially private algorithm $\mathcal{A}$, and for all $s_0 \in \mathbb{N}$, there exists $p_2 \in (0,1)$ such that for all x that appears less than s_0 times in the dataset S*

$$
\mathbb{P}_{h \sim \mathcal{A}(S, \tilde{f})}\left[h(\mathbf{x}) \neq \tilde{f}(\mathbf{x})\right] > 1 - p_2
$$

with high probability, where $p_2 = 1 - \frac{(1 - s_0 e^{(s_0-1)\epsilon}\delta)e^{-s_0\epsilon}}{1+e^{-s_0\epsilon}}$.

Proof of Lemma 1. Assume $\forall x \in S_{\#,\ell}, \mathbb{P}_{h\sim\mathcal{A}(S_m,f)}[h(x) \neq f(x)] < \beta$. For some $x_0 \in S_{\#,l}$, let $\tilde{f}$ be a function sampled from $\mathcal{F}_{|S_m \setminus \{x_0\}}$, the distribution over a set of labelling functions that have the same mapping as f at points in $S_m \setminus \{x_0\}$. That is, for all $x \in S_m \setminus \{x_0\}$, $\tilde{f}(x) = f(x)$ and $\tilde{f}(x_0) \neq f(x_0)$. Thus, data sets (S_m, f) and $(S_m, \tilde{f})$ are different at exactly ℓ points.

As the algorithm $\mathcal{A}$ is (ϵ, δ)-differential private, we can apply group privacy, for any $x \in X_{\#,\ell}$,

$$
\mathbb{P}_{h\sim\mathcal{A}(S_m,f)}[h(x) = \tilde{f}(x)] \geq \left(\mathbb{P}_{h\sim\mathcal{A}(S_m,\tilde{f})}[h(x) = \tilde{f}(x)] - le^{(l-1)\epsilon}\delta\right) e^{-l\epsilon} \geq \left(1 - \beta - le^{(l-1)\epsilon}\delta\right) e^{-l\epsilon} \tag{12}
$$

Also, following our assumption,

$$
\mathbb{P}_{h\sim\mathcal{A}(S_m,f)}[h(x) = \tilde{f}(x)] \leq \mathbb{P}_{h\sim\mathcal{A}(S_m,f)}[h(x) \neq f(x)] < \beta \tag{13}
$$

Combining the inequalities (12) and (13), we would reach a contradiction if

$$
\beta < \left(1 - \beta - le^{(l-1)\epsilon}\delta\right) e^{-l\epsilon} \tag{14}
$$

Then, we could get a contradiction by rearranging (14). Specifically, the assumption is rejected for ℓ satisfying the following inequality

$$
\beta \leq \frac{(1 - le^{(l-1)\epsilon}\delta)e^{-l\epsilon}}{1 + e^{-l\epsilon}} \tag{15}
$$

Note the right-hand side of the equation is decreasing in ℓ and never exceeds 0.5 for all $\epsilon, \delta > 0$ and $\ell \geq 1$. Thus, for any integer $s_0 \geq 1$, we can find some $\beta \leq \frac{(1-le^{(l-1)\epsilon}\delta)e^{-l\epsilon}}{1+e^{-l\epsilon}}$. That is, for all $\ell \leq s_0$, the inequality (15) holds and we can reach a contradiction.

Then we will show that $\tilde{f}$ that only differs from f at x_0 can be sampled from $\mathcal{F}_{|S_m \setminus \{x_0\}}$ with high probability.

$$
\mathbb{P}_{f\sim\mathcal{F}}\left[f(x_0) \neq \tilde{f}(x_0) | f(z) = \tilde{f}(z) \forall z \in S_m \setminus \{x_0\}\right] \geq 1 - ||\mathcal{F}_{|S_m \setminus \{x_0\}}||_\infty
$$

where $||\mathcal{F}_{|S_m \setminus \{x_0\}}||_\infty = \max_{y \in \mathcal{Y}} \mathbb{P}_{f \in \mathcal{F}_{|S_m \setminus \{x_0\}}} [f(x_0) = y]$.

□

Now, we provide detailed proofs for all the theorems in the main text.

A.1 PROOF FOR THEOREM 1

In this section, we will look at the proof of Theorem 1. First, we see the detailed version of the same theorem below.

Theorem 4 (Detailed version of Theorem 1)**.** *Consider $X, \mathcal{F}$ as defined above. Given a distribution as defined in Definition 3. For all $p \in (0,1), c > 0$ such that $p/c \leq 1$, for all $\epsilon, \delta > 0$, for any $s_0 \geq 1$ and $s_0 = o(m)$, if an algorithm $\mathcal{A}$ satisfies the following conditions*

1. *$\mathcal{A}$ is (ϵ, δ)-differential private*
2. *$\forall l > s_0$, $\mathbb{P}_{h \sim \mathcal{A}(S,f), f \sim \mathcal{F}}[h(x) \neq f(x)] \leq p_1$*
3. *$\forall l \leq s_0$, $\mathbb{P}_{h \sim \mathcal{A}(S,f), f \sim \mathcal{F}}[h(x) \neq f(x)] > 1 - p_2$*

Then, with probability greater than 0.99,

$$\text{err}\left(\mathcal{A}, \Pi_{p,c}, \mathcal{F}\right) \leq (1-p_1)p \left(1 - \frac{(s_0 - \frac{p}{2c})^2 \frac{2c}{p} - 1}{\sqrt{2\pi}(s_0 - \frac{p}{2c})^3 \left(\sqrt{\frac{2c}{p}}\right)^3} e^{-(s_0 - \frac{p}{2c})^2 \frac{c}{p}}\right) + p_1$$

and

$$\Gamma\left(\mathcal{A}, \Pi_{p,c}, \mathcal{F}\right) \geq c_0(1-p) \left(1 - \frac{\sqrt{\frac{1+p}{2c}}}{\sqrt{2\pi}\left(s_0 - \frac{1+p}{2c}\right)} e^{-(s_0 - \frac{1+p}{2c})^2 \frac{c}{1+p}}\right) - (1-p)p_1$$

where $c_0 = \min_{x \in X} 1 - e^\epsilon ||\mathcal{F}(x)||_\infty - \delta$ and $||\mathcal{F}(x)||_\infty = \max_{y \in \mathcal{Y}} [f(x) = y]$.

Proof of Theorem 4. Recall the definition of the minority and majority subpopulations X_1 and X_2 from Definition 5. Define S_1 to be the partition of S that belongs to X_1 and $m_1 = |S_1|$ to be the number of examples in S that belong to the minority subpopulation. Similarly, define S_2 and $m_2 = m - m_1$ as the set of majority examples and the size of the set respectively. Then, the following two sampling processes are equivalent

1. Repeatedly sample m points from $\Pi_{p,c}$.
2. First sample m_1 from a binomial distribution with n trials and success probability p. Then, repeatedly sample m_1 points from $\Pi^1_{p,c}$ and $m_2 = m - m_1$ points from $\Pi^2_{p,c}$.

In this proof, we use the second process to sample the dataset S. The expectation is over $S \sim \Pi^m_{p,c}, f \sim \mathcal{F}$ and all probabilities is also over $f \sim \mathcal{F}$, but we omit them for simplicity. By Lemma 3, $m_1 \geq pm/2, m_2 \geq (1-p)m/2$ occurs with probability at least 0.99. Under this assumption on m_1, m_2, for all $\ell \leq s_0 = o(m)$,

$$\begin{aligned}
\mathbb{E}_{S_2 \sim \Pi^{2,m_2}_{p,c}}[|S_{2,\#l}|] &= \sum_{x \in X_2} \mathbb{I}\{x \text{ occurs } \ell \text{ times in } S\} \mathbb{P}_{S_2 \sim \Pi^{2,m_2}_{p,c}}[x \text{ occurs } \ell \text{ times in } S] \\
&= k(1-p)\binom{m_2}{\ell}\left(\frac{1}{k(1-p)}\right)^l \left(1 - \frac{1}{k(1-p)}\right)^{m_2 - l} \\
&\leq k(1-p)\left(\frac{m_2}{l}\right)^l \left(\frac{1}{k(1-p)}\right)^l \left(1 - \frac{1}{k(1-p)}\right)^{m_2 - l} \\
&\to 0 \qquad \because m_2 \geq \frac{(1-p)m}{2} \to \infty
\end{aligned} \tag{16}$$

As $s_0 = o(m)$ and $\mathbb{E}_{S_2 \sim \Pi_{p,c}^{2,m_2}}[|S_{2,\#l}|]$ goes to 0 exponentially for any ℓ, we have $\sum_{l=0}^{s_0} \mathbb{E}_{S_2 \sim \Pi_{p,c}^{2,m_2}}[|S_{2,\#l}|] \to 0$. Then, we will prove an upper bound for the error.

$$
\begin{aligned}
\operatorname{err}(\mathcal{A}, \Pi_{p,c}, \mathcal{F}) &= \mathbb{E}\left[\sum_{x \in X} \Pi_{p,c}(x) \mathbb{P}_{h \sim \mathcal{A}(S,f)}[h(x) \neq f(x)]\right] \\
&= \mathbb{E}\left[\sum_{l=0}^{s_0} \sum_{x \in S_{\#,l}} \Pi_{p,c}(x) \mathbb{P}_{h \sim \mathcal{A}(S,f)}[h(x) \neq f(x)]\right] \\
&+ \mathbb{E}\left[\sum_{l=s_0+1}^{m} \sum_{x \in S_{\#,l}} \Pi_{p,c}(x) \mathbb{P}_{h \sim \mathcal{A}(S,f)}[h(x) \neq f(x)]\right] \\
&\leq \sum_{l=0}^{s_0} \mathbb{E}\left[\sum_{x \in S_{\#,l}} \Pi_{p,c}(x)\right] + p_1 \mathbb{E}\left[\sum_{l=0}^{m} \sum_{x \in S_{\#,l}} \Pi_{p,c}(x)\right] - p_1 \sum_{l=0}^{s_0} \mathbb{E}\left[\sum_{x \in S_{\#,l}} \Pi_{p,c}(x)\right] \\
&= (1-p_1)\frac{p}{N} \sum_{l=0}^{s_0} \mathbb{E}\left[|S_{1,\#\ell}|\right] + p_1 && \text{By (16)} \\
&= (1-p_1)\frac{p}{N} \sum_{l=0}^{s_0} \sum_{x \in X_1} \mathbb{I}\{x \text{ occurs } \ell \text{ times in } S\} \mathbb{P}_{S_1 \sim \Pi_{p,c}^{1,m_1}}[x \text{ occurs } \ell \text{ times in } S] + p_1 \\
&= (1-p_1)\frac{p}{N} \sum_{l=0}^{s_0} \binom{m_1}{l} \left(\frac{1}{N}\right)^l \left(1 - \frac{1}{N}\right)^{m_1 - l} N + p_1 \\
&= (1-p_1) p \mathbb{P}_{\ell \sim \text{binom}(m_1, \frac{m_1}{N})}[\ell \leq s_0] + p_1 \\
&\approx (1-p_1) p \mathbb{P}_{\ell \sim N(\frac{m_1}{N}, \frac{m_1}{N}(1-\frac{1}{N}))}[\ell \leq s_0] + p_1 && \text{By CLT} \\
&= (1-p_1) p \Phi\left(\frac{s_0 - \frac{m_1}{N}}{\sqrt{\frac{1}{N}\left(1 - \frac{1}{N}\right) m_1}}\right) + p_1 \\
&\leq (1-p_1) p \Phi\left((s_0 - \frac{p}{2c})\sqrt{\frac{2c}{p}}\right) + p_1 && \because m_1 \geq \frac{p}{2} m \\
&\leq (1-p_1) p \left(1 - \frac{(s_0 - \frac{p}{2c})^2 \frac{2c}{p} - 1}{\sqrt{2\pi}(s_0 - \frac{p}{2c})^3 \left(\sqrt{\frac{2c}{p}}\right)^3} e^{-(s_0 - \frac{p}{2c})^2 \frac{c}{p}}\right) + p_1 && \text{By } \textit{Inequality } 2
\end{aligned}
\tag{17}
$$

where Φ is the cumulative distribution function of standard normal distribution. The last inequality follows from the lower tail bound of Gaussian distribution (Inequality 2).

Then, we will show that fairness is worse for an algorithm satisfying differential privacy with reasonable accuracy. We show this by obtaining the lower bound of the fairness using the fact that the error in the minority group mainly comes from infrequent samples and the error from the majority group would not exceed p_1 with high probability.

Let $\tilde{f}$ be sampled from a class of functions that only differ from f at x, then (S, f) and $(S, \tilde{f})$ are neighboring datasets as they only differ at x. As $\mathcal{A}$ is (ϵ, δ)-differentially private, we have the following inequality by the definition of differential privacy.

$$
\mathbb{P}_{h \sim \mathcal{A}(S,f)}[h(x) = f(x)] \leq e^{\epsilon} \mathbb{P}_{h \sim \mathcal{A}(S,\tilde{f})}[h(x) = f(x)] + \delta
$$

Thus, let $||\mathcal{F}(x)||_\infty = \max_{y\in\mathcal{Y}} \mathbb{P}_{f\sim\mathcal{F}}[f(x)=y]$ and $c_0 := \min_{x\in X} 1 - e^\epsilon ||\mathcal{F}(x)||_\infty - \delta$, then for all $x \in X$,

$$\begin{aligned}
\mathbb{P}_{h\sim\mathcal{A}(S,f)}[h(x)\neq f(x)] &= 1 - \mathbb{P}_{h\sim\mathcal{A}(S,f)}[h(x)=f(x)] \\
&\geq 1 - e^\epsilon \mathbb{P}_{h\sim\mathcal{A}(S,\tilde{f})}[h(x)=f(x)] - \delta \\
&\geq \min_{x\in X} 1 - e^\epsilon \max_{y\in\mathcal{Y}} \mathbb{P}_{f\sim\mathcal{F}}[f(x)=y] - \delta \\
&= \min_{x\in X} 1 - e^\epsilon ||\mathcal{F}(x)||_\infty - \delta = c_0
\end{aligned} \tag{18}$$

Using (18), we calculate a lower bound of the error in the minority group

$$\begin{aligned}
\operatorname{err}\left(\mathcal{A}, \Pi^1_{p,c}, \mathcal{F}\right) &= \mathbb{E}\left[\sum_{l=1}^{m_1} \sum_{x\in S_{1,\#l}} \Pi^1_{p,c}(x) \mathbb{P}_{h\sim\mathcal{A}(S,f)}[h(x)\neq f(x)]\right] \\
&\geq \sum_{l=1}^{s_0} \mathbb{E}\left[\sum_{x\in S_{1,\#l}} \Pi^1_{p,c}(x) \mathbb{P}_{h\sim\mathcal{A}(S,f)}[h(x)\neq f(x)]\right] \\
&\geq \frac{c_0}{N} \sum_{\ell=0}^{s_0} \mathbb{E}\left[|S_{1,\#\ell}|\right] && \text{By (18)} \\
&\geq c_0 \sum_{\ell=0}^{s_0} \binom{m_1}{\ell} \left(\frac{1}{N}\right)^\ell \left(1-\frac{1}{N}\right)^{m_1-\ell} \\
&= c_0 \mathbb{P}_{\ell\sim\text{binom}(m_1,\frac{m_1}{N})}\left[l \leq s_0\right] \\
&= c_0 \mathbb{P}_{\ell\sim N(\frac{m_1}{N},\frac{m_1}{N}(1-\frac{1}{N}))}\left[l \leq s_0\right] && \text{CLT} \\
&= c_0 \mathbb{P}_{Z\sim N(0,1)}\left[Z \leq \frac{s_0 - \frac{m_1}{N}}{\sqrt{m_1 \frac{1}{N}\left(1-\frac{1}{N}\right)}}\right] \\
&\geq c_0 \Phi\left(\frac{s_0 - \frac{1+p}{2c}}{\sqrt{\frac{1+p}{2c}}}\right) && \because m_1 \leq \frac{(1+p)m}{2} \\
&\geq c_0 \left(1 - \frac{\sqrt{\frac{1+p}{2c}}}{\sqrt{2\pi}\left(s_0 - \frac{1+p}{2c}\right)} e^{-(s_0-\frac{1+p}{2c})^2 \frac{c}{1+p}}\right) && \text{By } \textit{Inequality } 2
\end{aligned} \tag{19}$$

Similarly, as all points in the majority group appears infinitely many times as the sample size is sufficiently large, the error from majority group is upper bounded by p_1.

$$\begin{aligned}
\operatorname{err}\left(\mathcal{A}, \Pi^2_{p,c}, \mathcal{F}\right) &= \mathbb{E}\left[\sum_{l=s_0}^{m_2} \sum_{x\in S_{2,\#l}} \Pi^2_{p,c}(x) \mathbb{P}_{h\sim\mathcal{A}(S,f)}[h(x)\neq f(x)]\right] \\
&\leq \mathbb{E}\left[\sum_{l=s_0}^{m_2} \sum_{x\in S_{2,\#l}} \frac{1}{k(1-p)} p_1\right] \\
&\leq \mathbb{E}\left[\sum_{l=1}^{m_2} \sum_{x\in S_{2,\#l}} \frac{1}{k(1-p)} p_1\right] = p_1
\end{aligned} \tag{20}$$

Therefore, by the definition of fairness, we can show that

$$\begin{aligned}
\Gamma\left(\mathcal{A}, \Pi_{p,c}, \mathcal{F}\right) &= (1-p)\left(\operatorname{err}\left(\mathcal{A}, \Pi^1_{p,c}, \mathcal{F}\right) - \operatorname{err}\left(\mathcal{A}, \Pi^2_{p,c}, \mathcal{F}\right)\right) \\
&\geq (1-p)c_0 \left(1 - \frac{\sqrt{\frac{1+p}{2c}}}{\sqrt{2\pi}\left(s_0 - \frac{1+p}{2c}\right)} e^{-(s_0-\frac{1+p}{2c})^2 \frac{c}{1+p}}\right) - (1-p)p_1
\end{aligned} \tag{21}$$

□

Then, we restate and prove Lemma 2.

Lemma 2. *The algorithm A_η is $\left(O\left(\frac{1}{\eta}\right), 0\right)$- differentially private where $O(\cdot)$ is the big-Oh notation. Further, for any $\beta_1, \beta_2 \in (0,1)$,*
• *if a point appears less than $\frac{\log(1/\beta_1)}{1+\frac{1}{2}\log[1/4\eta(1-\eta)]}$ times in the training set, then $\mathbb{P}_{h\sim\mathcal{A}}[h(x) \neq f(x)] \geq \beta_1$ and*
• *if a point appears more than $-\frac{8(1-\eta)}{(1-2\eta)^2}\log\beta_2$ times, then $\mathbb{P}_{h\sim\mathcal{A}}[h(x) \neq f(x)] \leq \beta_1$.*

Proof. We first prove the first part that A_η is $(\epsilon, 0)$-differentially private.

For any two neighboring datasets (S, f) and $(S, \tilde{f})$ that differ in one element, the classifier from a majority vote algorithm with random flips of the labels would changes the most if (S, f) and $(S, \tilde{f})$ differs in a singleton point x'.

$$\begin{aligned}
\frac{\mathbb{P}_{h\sim\mathcal{A}(S,f),f\sim\mathcal{F}}[h(x)=f(x),\forall x\in S]}{\mathbb{P}_{h\sim\mathcal{A}(S,\tilde{f}),\tilde{f}\sim\mathcal{F}_{|S\setminus x'}}[h(x)=f(x),\forall x\in S]} &= \frac{\prod_{x\in S}\mathbb{P}_{h\sim\mathcal{A}(S,f),f\sim\mathcal{F}}[h(x)=f(x)]}{\prod_{x\in S}\mathbb{P}_{h\sim\mathcal{A}(S,\tilde{f}),\tilde{f}\sim\mathcal{F}_{|S\setminus x'}}[h(x)=f(x)]} \\
&= \frac{\mathbb{P}_{h\sim\mathcal{A}(S,f),f\sim\mathcal{F}}[h(x')=f(x')]}{\mathbb{P}_{h\sim\mathcal{A}(S,\tilde{f}),\tilde{f}\sim\mathcal{F}_{|S\setminus x'}}[h(x')=f(x')]} \\
&= \frac{\mathbb{P}_{h\sim\mathcal{A}(S,f),f\sim\mathcal{F}}[x' \text{ is not flipped }]}{\mathbb{P}_{h\sim\mathcal{A}(S,\tilde{f}),\tilde{f}\sim\mathcal{F}_{|S\setminus x'}}[x' \text{ is flipped and flipped to } f(x')]} \\
&= \frac{1-\eta}{\eta\frac{1}{|\mathcal{Y}-1|}} = \frac{1-\eta}{\eta}(|\mathcal{Y}|-1)
\end{aligned} \tag{22}$$

Then, we show the inequality regarding the probability that A_η misclassifies a point occurring ℓ times in S.

For a point x occurring ℓ times in S, we denote all replicates of this point in the dataset as $[x_1, ..., x_\ell]$. As each point is flipped with probability η, let $k_i^x = \mathbb{I}\{x_i \text{ is not flipped}\}$, then $\sum_{i=1}^{\ell} k_i^x$ is a binomial random variable with ℓ trials and success probability $1-\eta$. The algorithm A_η would correctly classify x if $\sum_{i=1}^{\ell} k_i^x > \ell/2$.

The upper bound of the error probability is shown in Sanyal et al. [33], but we reproduce it here for completeness.

$$\begin{aligned}
\mathbb{P}\left[\sum_{i=1}^{\ell} k_i^x \leq \frac{\ell}{2}\right] &= \mathbb{P}\left[\sum_{i=1}^{\ell} k_i^x \leq \frac{\ell}{2\mu}\mu + \mu - \mu\right] \\
&= \mathbb{P}\left[\sum_{i=1}^{\ell}\left(1-\left(1-\frac{\ell}{2\mu}\right)\right)\mu\right] \\
&\leq e^{-\frac{(1-2\eta)^2}{8(1-\eta)^2}\mu} && \text{By Chernoff bound} \\
&= e^{-\frac{(1-2\eta)^2}{8(1-\eta)}\ell} && \mu = (1-\eta)\ell
\end{aligned} \tag{23}$$

Then, we show a lower bound for the error probability. Define $D(a||p)$ to be the Kullback-Leibler divergence between two Bernoulli random variables with parameter a and p, i.e. $D(a||p) = a\log\frac{a}{p} + (1-a)\log\frac{1-a}{1-p}$.

$$\begin{aligned}
\mathbb{P}\left[\sum_{i=1}^{\ell} k_i^x \leq \frac{\ell}{2}\right] &\geq \frac{1}{\sqrt{2\ell}}e^{-\ell D(\frac{\ell}{2\ell}||1-\eta)} && \text{By Inequality 1} \\
&= \frac{1}{\sqrt{2\ell}}e^{-\ell D(\frac{\ell}{2\ell}||1-\eta)} \\
&= \frac{1}{\sqrt{2\ell}}e^{-\ell\left(\frac{1}{2}\log\frac{1}{2(1-\eta)}+\frac{1}{2}\log\frac{1}{2\eta}\right)} \\
&= \frac{1}{\sqrt{2l}}\left[4\eta(1-\eta)\right]^{\frac{\ell}{2}}
\end{aligned} \tag{24}$$

Given the upper bound and lower bound of the error probability, we can show that for any $\beta_1 \in (0,1)$, if a point appears less than $\ell \leq \frac{-\log(\beta)}{1-\frac{1}{2}\log[4\eta(1-\eta)]}$ times, $\mathbb{P}_{h\sim\mathcal{A}}[h(x) \neq f(x)] \geq \beta_1$.

If

$$\ell \leq -\frac{\log(\beta)}{1-\frac{1}{2}\log[4\eta(1-\eta)]}$$

As $-1+\frac{1}{2}\log[4\eta(1-\eta)] < 0$,

$$-\ell + \frac{\ell}{2}\log[4\eta(1-\eta)] \geq \log\beta_1$$

As $\log(2\ell) \leq 2\ell$, we have

$$-\frac{1}{2}\log(2\ell) + \frac{\ell}{2}\log[4\eta(1-\eta)] \geq \log\beta_1$$

This is equivalent to

$$\mathbb{P}_{h\sim\mathcal{A}}[h(x) \neq f(x)] \geq \frac{1}{\sqrt{2\ell}}[4\eta(1-\eta)]^{\frac{\ell}{2}} \geq \beta_1$$

Similarly, we can also use the upper bound of the error probability to show that for any $\beta_2 \in (0,1)$, if a point appears more than $\ell \geq -\frac{8(1-\eta)}{(1-2\eta)^2}\log\beta_2$ times in the dataset, then $\mathbb{P}_{h\sim\mathcal{A}}[h(x) \neq f(x)] \leq \beta_2$.

If

$$\ell \geq -\frac{8(1-\eta)}{(1-2\eta)^2}\log\beta_2$$

We have

$$-\frac{(1-2\eta)^2\ell}{8(1-\eta)} \leq \log\beta_2$$

Thus,

$$\mathbb{P}_{h\sim\mathcal{A}}[h(x) \neq f(x)] \leq e^{-\frac{(1-2\eta)^2\ell}{8(1-\eta)}} \leq \beta_2$$

□

A.2 PROOF FOR THEOREM 2

Theorem 5 (Detailed version of Theorem 2)**.** *Consider X as defined above. Let $\mathcal{F}$ be a family of label priors such that $\max_{x\in X_1}||\mathcal{F}(x)||_\infty \leq 0.1$ and $\min_{x\in X_1}\min_{y\in\mathcal{Y}}\mathbb{P}_{f\sim\mathcal{F}}[f(x)=y] \geq 0.05$. For any $\alpha \in (0,1)$, $\epsilon \in (0,1)$, $\delta \in (0,0.01)$, if an algorithm $\mathcal{A}$ satisfies the following two conditions,*

1. *$\mathcal{A}$ does not make mistakes on points appearing more than once in the dataset input to $\mathcal{A}$,*
2. *$\mathcal{A}$ is (ϵ,δ)-differential private,*

then, there exists $p \in (0,1)$, $c > 0$ such that when X is distributed as $\Pi_{p,c}$ (Definition 3), with probability greater than 0.99

$$\mathrm{err}(\mathcal{A}, \Pi_{p,c}, \mathcal{F}) \leq \alpha$$

and

$$\Gamma(\mathcal{A}, \Pi_{p,c}, \mathcal{F}) \geq 0.5$$

where $||\mathcal{F}(x)||_\infty = \max_{y\in\mathcal{Y}}\mathbb{P}_{f\sim\mathcal{F}}[f(x)=y]$.

Proof. For a sample S of size m where each point is sampled from $\Pi_{p,c}$, denote the points from the minority group as S_1 and the points from the majority group as S_2. Denote the size of S_1, S_2 as m_1, m_2. We also use the notation $S_{1,\#\ell}$ and $S_{2,\#\ell}$ to represent the set of elements that occur ℓ times in S_1 and S_2 respectively. In this proof, the expectation is taken over $S \sim \Pi_{p,c}^m$, $f \sim \mathcal{F}$ and the probability is also over $f \sim \mathcal{F}$, but we will omit them for simplicity.

By Lemma 3 we have that with probability at least 0.99, $m_1 \geq \frac{pm}{2}$, $m_2 \geq \frac{(1-p)m}{2}$. For such m_1 and m_2, we can show that almost all points in the majority group occurs more than once, i.e. $\lim_{m\to\infty}\mathbb{E}[|S_{2,\#0}|] = 0$ and $\lim_{m\to\infty}\mathbb{E}[|S_{2,\#1}|] = 0$.

$$
\begin{aligned}
\mathbb{E}\left[|S_{2,\#0}|\right] &= \sum_{x\in X_2} \mathbb{I}\{x \text{ occurs 0 times in } S\}\mathbb{P}[x \text{ occurs 0 times in } S] \\
&= k(1-p)\binom{m_2}{0}\left(\frac{1}{k(1-p)}\right)^0\left(1-\frac{1}{k(1-p)}\right)^{m_2} \qquad \because m_2 = 1-m_1 \le \frac{pm}{2} \\
&\le k(1-p)\left(1-\frac{1}{k(1-p)}\right)^{\frac{p}{2}m} \\
&\xrightarrow{m\to\infty} 0
\end{aligned} \tag{25}
$$

Similarly, $\lim_{m\to\infty}\mathbb{E}[|S_{2,\#1}|] = 0$.

We then demonstrate with high probability, the error of an algorithm that is (ϵ,δ)-differential private and only makes mistakes on singleton examples is bounded.

Let $\tilde{c}_1 := \max_{x\in X_1}\max_{y\in\mathcal{Y}}\mathbb{P}_{f\sim\mathcal{F}}[f(x)\neq y]$, $\tilde{c}_2 := \max_{x\in X_2}\max_{y\in\mathcal{Y}}\mathbb{P}_{f\sim\mathcal{F}}[f(x)\neq y]$. Then, for all $x\in X_1$,

$$
\mathbb{P}_{h\sim A(S,f)}\left[h(x)\neq f(x)\right] \le \max_{x\in X_1}\max_{y\in\mathcal{Y}}\mathbb{P}_{f\sim\mathcal{F}}[f(x)\neq y] = \tilde{c}_1 \tag{26}
$$

For all $x\in X_2$,

$$
\mathbb{P}_{h\sim A(S,f)}\left[h(x)\neq f(x)\right] \le \max_{x\in X_2}\max_{y\in\mathcal{Y}}\mathbb{P}_{f\sim\mathcal{F}}[f(x)\neq y] = \tilde{c}_2 \tag{27}
$$

As $m_1 \ge \frac{pm}{2}$, $m_2 \ge \frac{(1-p)m}{2}$ occurs with probability at least 0.99 (Lemma 3), we can get the following upper bound on error with probability at least 0.99.

$$
\begin{aligned}
\text{err}\left(\mathcal{A},\Pi_{p,c},\mathcal{F}\right) =& \sum_{l=0}^{1}\mathbb{E}\left[\sum_{x\in S_{\#,l}}\Pi_{p,c}(x)\mathbb{P}_{h\sim\mathcal{A}(S,f)}\left[h(x)\neq f(x)\right]\right] \\
=& \sum_{l=0}^{1}\mathbb{E}\left[\sum_{x\in S_{1,\#\ell}}\Pi_{p,c}(x)\mathbb{P}_{h\sim\mathcal{A}(S,f)}[h(x)\neq f(x)]\right] \\
&+\sum_{l=0}^{1}\mathbb{E}\left[\sum_{x\in S_{2,\#\ell}}\Pi_{p,c}(x)\mathbb{P}_{h\sim\mathcal{A}(S,f)}[h(x)\neq f(x)]\right] \\
\le& \frac{\tilde{c}_1 p}{N}\mathbb{E}\left[|S_{1,\#0}|+|S_{1,\#1}|\right] + \frac{\tilde{c}_2}{k}\mathbb{E}[|S_{2,\#0}|+|S_{2,\#1}|] \qquad \text{By (26), (27)} \\
=& \frac{\tilde{c}_1 p}{N}\sum_{\ell=0}^{1}\sum_{x\in X_1}\mathbb{I}\{x \text{ occurs } \ell \text{ times in } S\}\mathbb{P}[x \text{ occurs } \ell \text{ times in } S]\} \qquad \because \mathbb{E}[|S_{2,\#0}|],\mathbb{E}[|S_{2,\#1}|]\to\infty \\
=& \frac{\tilde{c}_1 p}{N}\left(Nm_1\frac{1}{N}(1-\frac{1}{N})^{m_1-1} + N\left(1-\frac{1}{N}\right)^{m_1}\right) \\
\le& \tilde{c}_1\frac{p}{N}\left(m\frac{1+p}{2}e^{-\frac{p}{2c}} + Ne^{-\frac{p}{2c}}\right) \qquad \because \frac{pm}{2}\ge m_1 = 1-m_2 \le \frac{1+p}{2}m \\
=& \tilde{c}_1\left(\frac{p(1+p)}{2c}e^{-\frac{p}{2c}} + pe^{-\frac{p}{2c}}\right)
\end{aligned} \tag{28}
$$

Then, using the fact that almost all points occur more than once in the majority group (i.e. $\lim_{m\to\infty}\mathbb{E}\left[|S_{2,\#0}|\right],\mathbb{E}\left[|S_{2,\#1}|\right] = 0$ from (25)), we show that the fairness of such algorithm is worse or the disparate accuracy is large.

Let $\tilde{f}$ be sampled from a class of functions that only differ from f at x, then (S,f) and $(S,\tilde{f})$ are neighboring datasets as they only differ at x. As $\mathcal{A}$ is (ϵ,δ)-differentially private, we have the following inequality by the definition of differential privacy.

$$\mathbb{P}_{h\sim\mathcal{A}(S,f)}[h(x)=f(x)] \leq e^{\epsilon}\mathbb{P}_{h\sim\mathcal{A}(S,\tilde{f})}[h(x)=f(x)]+\delta$$

Thus, let $||\mathcal{F}(x)||_\infty = \max_{y\in\mathcal{Y}} \mathbb{P}_{f\sim\mathcal{F}}[f(x)=y]$ and $\tilde{c} := \min_{x\in X_1} 1-e^{\epsilon}||\mathcal{F}(x)||_\infty - \delta$, then for all $x \in X_1$,

$$\begin{aligned}
\mathbb{P}_{h\sim\mathcal{A}(S,f)}[h(x)\neq f(x)] &= 1-\mathbb{P}_{h\sim\mathcal{A}(S,f)}[h(x)=f(x)] \\
&\geq 1-e^{\epsilon}\mathbb{P}_{h\sim\mathcal{A}(S,\tilde{f})}[h(x)=f(x)]-\delta \\
&\geq \min_{x\in X_1} 1-e^{\epsilon}\max_{y\in\mathcal{Y}}\mathbb{P}_{f\sim\mathcal{F}}[f(x)=y]-\delta \\
&= \min_{x\in X} 1-e^{\epsilon}||\mathcal{F}(x)||_\infty-\delta=\tilde{c}
\end{aligned} \tag{29}$$

$$\begin{aligned}
\Gamma\left(\mathcal{A},\Pi_{p,c},\mathcal{F}\right) =& (1-p)\left[\mathrm{err}\left(\mathcal{A},\Pi^1_{p,c},\mathcal{F}\right)-\mathrm{err}\left(\mathcal{A},\Pi^2_{p,c},\mathcal{F}\right)\right] && \text{By (10)} \\
=& (1-p)\sum_{l=0}^{1}\mathbb{E}\left[\sum_{x\in S_{1,\#l}}\Pi^1_{p,c}(x)\mathbb{P}_{h\sim\mathcal{A}(S,f)}[h(x)\neq f(x)]\right] \\
&-\mathbb{E}\left[\sum_{x\in S_{2,\#l}}\Pi^2_{p,c}(x)\mathbb{P}_{h\sim\mathcal{A}(S,f)}[h(x)\neq f(x)]\right] \\
\geq& (1-p)\sum_{l=0}^{1}\mathbb{E}\left[\frac{1}{N}\sum_{x\in S_{1,\#l}}\mathbb{P}_{h\sim\mathcal{A}(S,f)}[h(x)\neq f(x)]\right] && \because \mathbb{E}\left[|S_{2,\#0}|\right],\mathbb{E}\left[|S_{2,\#1}|\right]\to\infty \\
\geq& \frac{(1-p)}{N}\mathbb{E}\left[\sum_{x\in S_{1,\#0}}\tilde{c}+\sum_{x\in S_{1,\#1}}\tilde{c}\right] && \text{By (29)} \\
\geq& \frac{\tilde{c}(1-p)}{N}\mathbb{E}_{S\sim\Pi^m_{p,c}}\left[|S_{1,\#1}|+|S_{1,\#0}|\right] \\
\geq& \tilde{c}(1-p)\left(\frac{p}{2c}e^{-\frac{1+p}{2c}}+e^{-\frac{1+p}{2c}}\right)
\end{aligned} \tag{30}$$

Then, we set $c=10$ and show that for any $\alpha\in[0,1]$, there is some p such that $\mathrm{err}\left(\mathcal{A},\Pi_{p,10},\mathcal{F}\right)\leq\alpha$ and $\Gamma\left(\mathcal{A},\Pi_{p,10},\mathcal{F}\right)\geq 0.5$.

First, we note that $\Gamma\left(\mathcal{A},\Pi_{p,10},\mathcal{F}\right)\geq 0.5$ is always satisfied for $p\leq 0.26$ because $\Gamma\left(\mathcal{A},\Pi_{0.26,10},\mathcal{F}\right)\geq 0.5$ and the lower bound of disparate minority accuracy increase when p is smaller. From (30),

$$\begin{aligned}
\Gamma\left(\mathcal{A},\Pi_{0.26,10},\mathcal{F}\right) &\geq \left[1-e^{\epsilon}\max_{x\in X_1}||\mathcal{F}||_\infty-\delta\right]\times 0.7039 \\
&\geq \left[1-e^{1}\times 0.1-0.01\right]\times 0.7039 \because \max_{x\in X_1}||\mathcal{F}||_\infty\leq 0.1, \epsilon\leq 1, \delta\geq 0.01 \\
&= 0.5055\geq 0.5
\end{aligned} \tag{31}$$

For $p\in(0,0.26)$, the lower bound of disparate minority accuracy decreases with p as the its derivative is negative.

$$\begin{aligned}
\frac{d}{dp}\tilde{c}\left(1-p\right)\left(\frac{p}{20}+1\right)e^{-\frac{1+p}{20}} &= \tilde{c}e^{-\frac{1+p}{20}}\left[\frac{1}{20}\left(\frac{p}{20}+1\right)(p-1)-\left(\frac{p}{20}+1\right)-\left(\frac{p-1}{20}\right)\right] \\
&= \tilde{c}e^{-\frac{1+p}{20}}\left[-\frac{p(1-p)}{20^2}\left(\frac{p}{20}+1\right)-\left(\frac{p}{20}+1\right)\right]\leq 0 \qquad \because \tilde{c}\geq 0
\end{aligned} \tag{32}$$

Then, we show that for $\alpha\in[0,1]$, there exists a distribution $\Pi_{p,c}$ with $p\leq 0.26, c=10$ such that $\mathrm{err}\left(\mathcal{A},\Pi_{p,10},\mathcal{F}\right)\leq\alpha$. Specifically, we first argue that the upper bound of error is monotonically increasing with p for $p\in[0,0.26]$, then the upper bound achieves its maximum and minimum at 0 and 0.26 respectively. From (28), we take the derivative of the upper bound of error, for $p\in[0,0.26]$,

$$\begin{aligned}\frac{d}{dp}\tilde{c}_1\left(\frac{p\,(1+p)}{20}+p\right)e^{-\frac{p}{20}} &= \tilde{c}_1 e^{-\frac{p}{20}}\left[\frac{p}{10}+\frac{21}{20}-\frac{1}{20}\left(p+\frac{p(1+p)}{20}\right)\right] \\ &\geq 0 \qquad\qquad \because \frac{1}{20}\left(p+\frac{p(1+p)}{20}\right)\leq\frac{21}{20}\end{aligned} \tag{33}$$

Denote the maximum and minimum of the upper bound of the error as $\alpha_{\max}$ and $\alpha_{\min}$ respectively, then,

$$\begin{aligned}\alpha_{\max} &= \tilde{c}_1\left(\frac{0.26\times(1+0.26)}{20}+0.26\right)e^{-\frac{-0.26}{20}} \\ &= \max_{x\in X_1}\max_{y\in\mathcal{Y}}\mathbb{P}_{f\sim\mathcal{F}}\left[f(x)\neq y\right]\times 0.2728 \\ &= \left(1-\min_{x\in X_1}\min_{y\in\mathcal{Y}}\mathbb{P}_{f\sim\mathcal{F}}\left[f(x)\neq y\right]\right)\times 0.2728 \\ &\leq 0.95\times 0.2728 \qquad\qquad \because \min_{x\in X_1}\min_{y\in Y}\mathbb{P}_{f\sim\mathcal{F}}\left[f(x)\neq y\right]\geq 0.05 \\ &= 0.2592\end{aligned} \tag{34}$$

$$\alpha_{\min} = \tilde{c}_1\left(\frac{0\times(1+0)}{20}+0\right)e^{-\frac{(1+0)}{20}} = 0 \tag{35}$$

Therefore, for $\alpha\in[0,\alpha_{\max}]$, we conclude that there exists some $p_0\in[0,0.26]$ such that the upper bound on error is exactly α by Intermediate Value Theorem, i.e. $\operatorname{err}(\mathcal{A},\Pi_{p_0,10},\mathcal{F})\leq\alpha$. If $\alpha\in(\alpha_{\max},1]$, we can always choose $p_0=0.26$ such that $\operatorname{err}(\mathcal{A},\Pi_{p_0,10},\mathcal{F})\leq\alpha_{\max}\leq\alpha$. □

A.3 PROOF FOR THEOREM 3

Theorem 6 (Detailed version of Theorem 3)**.** *Consider $X,\mathcal{F}$ as defined above. Given a distribution as defined in Definition 3. For all $p\in(0,1), c>0$ such that $p/c\leq 1$, for all $\epsilon,\delta>0$, for any $s_0=\frac{2-p}{2k(1-p)}m+\alpha\sqrt{m}$, $\alpha>0$, if an algorithm $\mathcal{A}$ satisfies the following conditions*

1. *$\mathcal{A}$ is (ϵ,δ)-differential private*
2. *$\forall l>s_0$, $\mathbb{P}_{h\sim\mathcal{A}(S,f),f\sim\mathcal{F}}[h(x)\neq f(x)]\leq p_1$*
3. *$\forall l\leq s_0$, $\mathbb{P}_{h\sim\mathcal{A}(S,f),f\sim\mathcal{F}}[h(x)\neq f(x)]>1-p_2$*

Then, with probability greater than 0.99,

$$\operatorname{err}(\mathcal{A},\Pi_{p,c},\mathcal{F})\geq\tilde{c}_1 p+(1-p_2)(1-p)\left(1-e^{-\frac{4(1-p)\alpha^2}{(2-p)^2}}\right)$$

$$\Gamma(\mathcal{A},\Pi_{p,c},\mathcal{F})\leq(1-p)\left[1-(1-e^{-\frac{4(1-p)\alpha^2}{(2-p)^2}})(1-p_2)\right]$$

where $\tilde{c}_1:=\min_{x\in X_1}1-e^{\epsilon}||\mathcal{F}(x)||_\infty-\delta$.

Proof. Assume $m_1\geq\frac{pm}{2}, m_2\geq\frac{(1-p)m}{2}$, which occurs with probability at least 0.99 by Lemma 3, we first show that all

points in the minority group occur less than s_0 times in the dataset, i.e. $\sum_{\ell=s_0+1}^{m_1} \mathbb{E}[|S_{1,\#l}|] \to 0$.

$$
\begin{aligned}
\sum_{l=s_0+1}^{m_1} \mathbb{E}[|S_{1,\#1}|] &= \sum_{\ell=s_0+1}^{m_1} \sum_{x\in X_1} \mathbb{I}\{x \text{ occurs } \ell \text{ times in } S\} \mathbb{P}_{S_1\sim\Pi_{p,c}^{1,m_1}}[x \text{ occurs } \ell \text{ times in } S] \\
&= \sum_{\ell=s_0+1}^{m_1} N \binom{m_1}{\ell} \left(\frac{1}{N}\right)^{\ell} \left(1-\frac{1}{N}\right)^{m_1-\ell} \\
&= N\mathbb{P}_{\ell\sim\text{binom}(m_1,\frac{1}{N})}[\ell > s_0] \\
&\le Ne^{\left(-2m_1\left(1-\frac{1}{N}-\frac{m_1-s_0}{m_1}\right)\right)} && \text{By Hoeffding's inequality} \\
&\le cme^{\frac{2m_1}{N}-\frac{(2-p)m}{k(1-p)}-\alpha\sqrt{m}} && \because s_0 = \frac{2-p}{2k(1-p)}m + \alpha\sqrt{m} \\
&\le cme^{\frac{1+p}{c}-\frac{pm}{k(1-p)}-\alpha\sqrt{m}} && \because \frac{p}{2}m \le m_1 \le \frac{1+p}{2}m \\
&\to 0
\end{aligned}
\tag{36}
$$

Then, we can calculate a lower bound for fairness. The expectations are over $S \sim \Pi_{p,c}^m$, $f \sim \mathcal{F}$ and the probability is also over $f \sim \mathcal{F}$, but we omit them for simplicity. As $\sum_{l=s_0+1}^{m_1} \mathbb{E}[|S_{1,\#l}|] \to 0$, we consider a simple upper bound for error in the minority group.

$$
\text{err}\left(\mathcal{A}, \Pi_{p,c}^1, \mathcal{F}\right) = \sum_{x\in X_1} \mathbb{E}\left[\Pi_{p,c}^1 \mathbb{P}_{h\sim\mathcal{A}(S,\mathcal{F})}[h(x) \ne f(x)]\right] \le 1 \tag{37}
$$

To calculate the error in the majority group, we apply Hoeffding's inequality on Binomial distribution to get the following results,

$$
\begin{aligned}
\mathbb{P}_{\ell\sim\text{binom}(m_2,\frac{1}{k(1-p)})}[\ell \le s_0] &= 1 - \mathbb{P}_{\ell\sim\text{binom}(m_2,\frac{1}{k(1-p)})}[\ell \ge s_0] \\
&\ge 1 - e^{-2m_2\left(1-\frac{1}{k(1-p)}-\frac{m_2-s_0}{m_2}\right)^2} && \text{By Hoeffding's inequality} \\
&\ge 1 - e^{-2m_2\left(1-\frac{1}{k(1-p)}-\frac{m_2-s_0}{\frac{2-p}{2}m}\right)^2} && \because m_2 = 1-m_1 \le \frac{2-p}{2}m \\
&= 1 - e^{-2m_2\left(\frac{2\alpha}{(2-p)\sqrt{m}}\right)^2} && \because s_0 = \frac{2-p}{2k(1-p)}m + \alpha\sqrt{m} \\
&= 1 - e^{-\frac{4(1-p)\alpha^2}{(2-p)^2}} && \because m_2 \ge \frac{1-p}{2}m
\end{aligned}
\tag{38}
$$

Then, we calculate a lower bound for the error in the majority group. If $c \ge 0$,

$$
\begin{aligned}
\text{err}\left(\mathcal{A}, \Pi_{p,c}^2, \mathcal{F}\right) &= \sum_{x\in X_2} \left[\Pi_{p,c}^2(x) \mathbb{P}_{h\sim\mathcal{A}(S,\mathcal{F})}[h(x) \ne f(x)]\right] \\
&\ge \mathbb{E}\left[\sum_{l=1}^{s_0} \sum_{x\in S_{2,\#l}} \frac{1}{k(1-p)}(1-p_2)\right] \\
&\ge \sum_{l=1}^{s_0} \frac{(1-p_2)}{k(1-p)} k(1-p) \binom{m_2}{l} \left(\frac{1}{k(1-p)}\right)^l \left(1-\frac{1}{k(1-p)}\right)^{m_2-l} \\
&= \mathbb{P}_{l\sim\text{binom}(m_2,\frac{1}{k(1-p)})}[l \le s_0](1-p_2) \\
&\ge \left(1 - e^{-\frac{4(1-p)\alpha^2}{(2-p)^2}}\right)(1-p_2) && \because \text{By (38)}
\end{aligned}
\tag{39}
$$

Thus, fairness is upper bounded

$$
\Gamma\left(\mathcal{A}, \Pi_{p,c}, \mathcal{F}\right) = (1-p)[\text{err}\left(\mathcal{A}, \Pi_{p,c}^1, \mathcal{F}\right) - \text{err}\left(\mathcal{A}, \Pi_{p,c}^2, \mathcal{F}\right)] \le (1-p)\left[1-(1-p_2)\left(1-e^{-\frac{4(1-p)\alpha^2}{(2-p)^2}}\right)\right] \tag{40}
$$

Then, we show that the algorithm must sacrifice its accuracy to reach a small fairness score.

$$\mathbb{P}_{h\sim\mathcal{A}(S,f)}[h(x)=f(x)] \leq e^{\epsilon}\mathbb{P}_{h\sim\mathcal{A}(S,\tilde{f})}[h(x)=f(x)]+\delta$$

Thus, let $||\mathcal{F}(x)||_\infty = \max_{y\in\mathcal{Y}} \mathbb{P}_{f\sim\mathcal{F}}[f(x)=y]$ and $\tilde{c}_1 := \min_{x\in X_1} 1-e^{\epsilon}||\mathcal{F}(x)||_\infty - \delta$, then for all $x \in X_1$,

$$\begin{aligned}
\mathbb{P}_{h\sim\mathcal{A}(S,f)}[h(x)\neq f(x)] &= 1-\mathbb{P}_{h\sim\mathcal{A}(S,f)}[h(x)=f(x)] \\
&\geq 1-e^{\epsilon}\mathbb{P}_{h\sim\mathcal{A}(S,\tilde{f})}[h(x)=f(x)]-\delta \\
&\geq \min_{x\in X} 1-e^{\epsilon}\max_{y\in\mathcal{Y}}\mathbb{P}_{f\sim\mathcal{F}}[f(x)=y]-\delta \\
&= \min_{x\in X} 1-e^{\epsilon}||\mathcal{F}(x)||_\infty-\delta=\tilde{c}_1
\end{aligned} \tag{41}$$

$$\begin{aligned}
\operatorname{err}(\mathcal{A},\Pi_{p,c},\mathcal{F}) &= \mathbb{E}\left[\sum_{x\in X}\Pi_{p,c}(x)\mathbb{P}_{h\sim\mathcal{A}(S,f)}[h(x)\neq f(x)]\right] \\
&\geq \mathbb{E}\left[\sum_{l=0}^{s_0}\sum_{x\in S_{\#,l}}\Pi_{p,c}(x)\mathbb{P}_{h\sim\mathcal{A}(S,f)}[h(x)\neq f(x)]\right] \\
&= \mathbb{E}\left[\sum_{l=0}^{s_0}\sum_{x\in S_{1,\#l}}\frac{p}{N}\mathbb{P}_{h\sim\mathcal{A}(S,f)}[h(x)\neq f(x)]\right] \\
&+ \mathbb{E}\left[\sum_{l=0}^{s_0}\sum_{x\in S_{2,\#l}}\frac{1}{k}\mathbb{P}_{h\sim\mathcal{A}(S,f)}[h(x)\neq f(x)]\right] \\
&\geq \tilde{c}_1 p + \mathbb{E}\left[\sum_{l=0}^{s_0}\sum_{x\in S_{2,\#l}}\frac{1}{k}\mathbb{P}_{h\sim\mathcal{A}(S,f)}[h(x)\neq f(x)]\right] \quad \because \sum_{l=s_0+1}^{m_1}\mathbb{E}[|S_{1,\#l}|]\to 0, \text{ and by (41)} \\
&= \tilde{c}_1 p + (1-p_2)(1-p)\,k\frac{1}{k}\mathbb{P}_{l\sim\text{binom}(m_2,\frac{1}{k})}[l\leq s_0] \\
&\geq \tilde{c}_1 p + (1-p_2)(1-p)\left(1-e^{-\frac{4(1-p)\alpha^2}{(2-p)^2}}\right) \qquad \text{By (38)}
\end{aligned} \tag{42}$$

□

B EXPERIMENTAL DETAILS

Constructing the synthetic data distribution Given $m \in \mathbb{N}, p \in (0, 0.5), C > 0$, and $k \in \mathbb{N}$, we first create two boolean hypercubes of dimension $d_{\min}$ and d_{maj} respectively, where $d_{\min} = O(\lceil \log Cm \rceil)$ and $d_{\text{maj}} = O(\lceil \log(k(1-p)) \rceil)$ respectively. For the minority subpopulation, the data is distributed uniformly across a mixture of N gaussian clusters positioned on a randomly chosen set of $d_{\min}$ vertices of this hyper-cube. Similarly, for the majority class the data is distributed across a uniform mixture of d_{maj} gaussian clusters situated on randomly chosen vertices of the second hypercube. Each of the clusters is randomly assigned a binary label. Finally, m samples are sampled from the minority distribution, along with their labels, and each data covariate is appended with a d_{maj}-dimensional vector of all 10^{-4} to create a $d_{\min} + d_{\text{maj}}$ dimensional dataset of size m. Similarly, m samples are sampled from the majority distribution and the covariates are appended with a $d_{\min}$ vector of all 10^{-4} to create a $d_{\min} + d_{\text{maj}}$ dimensional dataset of size m. Finally, m_1 points are chosen randomly from the minority dataset and $m - m_1$ points are chosen randomly from the majority dataset where m_1 is sampled from a binomial distribution with parameters m and p.

DP Algorithm for synthetic data We use PyTorch Opacus [38] for all our experiments. We use multiple values of ϵ in the range $(0.1, 10)$. We clip the maximum gradient norm to 1, and then train the algorithm for 10^3 epochs. Note that the clipping is necessary to control the lipschitzness of the gradient descent algorithm update step. The DP-SGD algorithm uses the value of δ, the maximum gradient norm, and the number of epochs mentioned above to compute the noise-multiplier of the algorithm that is necessary to attain the required privacy guarantee.

C ADDITIONAL EXPERIMENTS

C.1 ADDITIONAL WEIGHT OF MINORITY GROUP

In this section, we plot additional experiments with the same setup as Section 4.1. In particular, we show results from the same experiment as Figure 2 but with $p = 0.5$ instead of $p = 0.2$. Our experiments show that the same trend is held in particular for how the disparate accuracy changes with c and with ϵ.

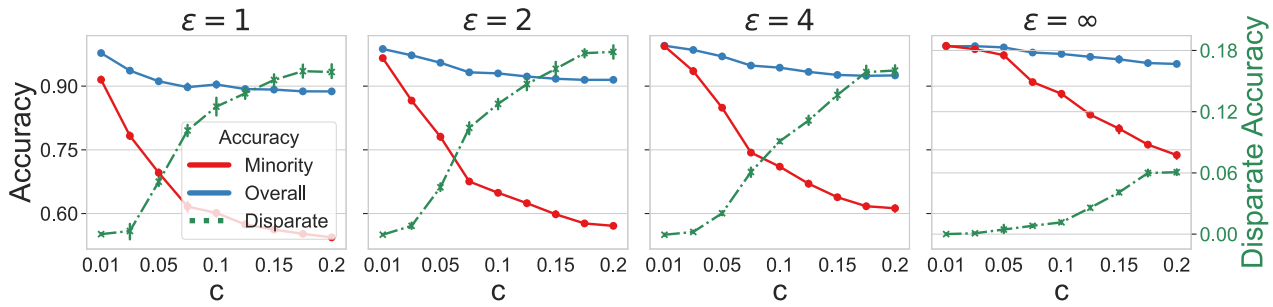

Figure 7: Here $p = 0.5$. Each figure plots the disparate accuracy (Definition 5; higher is less fair) in green dashed line, the accuracy of the minority group with red, and the overall accuracy with blue on the y-axis and the parameter c in the X-axis. The left most $(\epsilon = 1)$ achieves the strictest level of privacy and the right most $(\epsilon = \infty)$ is vanilla training without any privacy constraints. The two figures in between achieve intermediate levels of privacy.

C.2 ADDITIONAL SIZES OF SUBPOPULATIONS IN CELEBA

To construct the subpopulations in CelebA, we use the following attributes 1. Pointy Nose, 2. Wearing Earrings, 3. Wavy Hair, 4. Wearing Lipstick, 5. Heavy Makeup, 6. Attractive, 7. Receding Hairline, 8. Blurry, 9. Bangs, 10. Wearing Hat, 11. Eyeglasses.. This also corresponds to Set A used in Figure 5.

We use the following features for Set B in our experiments in Figure 5. 1. Arched Eyebrows 2. Bags Under Eyes 3. Blond Hair 4. Double Chin 5. High Cheekbones 6. Pale Skin 7. Rosy Cheeks 8. Straight Hair 9. Wearing Necklace 10. Wearing Necktie 11. Young

In Figure 4, we reported results with the subpopulation size set to 40.In this section, we report results where the size belongs to $\{5, 10, 20, 60, 80, 100\}$. In Figure 8, we show how disparate accuracy changes with ϵ for various subpopulation sizes. in Figure 9, we show how the minority group accuracy and the overall accuracy changes with ϵ for various subpopulation sizes. Overall the results, reflect the same trend as Figure 4 but the worst case disparate accuracy decreases for increasing subpopulation sizes.

C.3 ADDITIONAL DETAILS AND EXPERIMENTS ON CIFAR10

In Figure 10, we plot the same results as Figure 6 but with the threshold ρ set to 0.01.

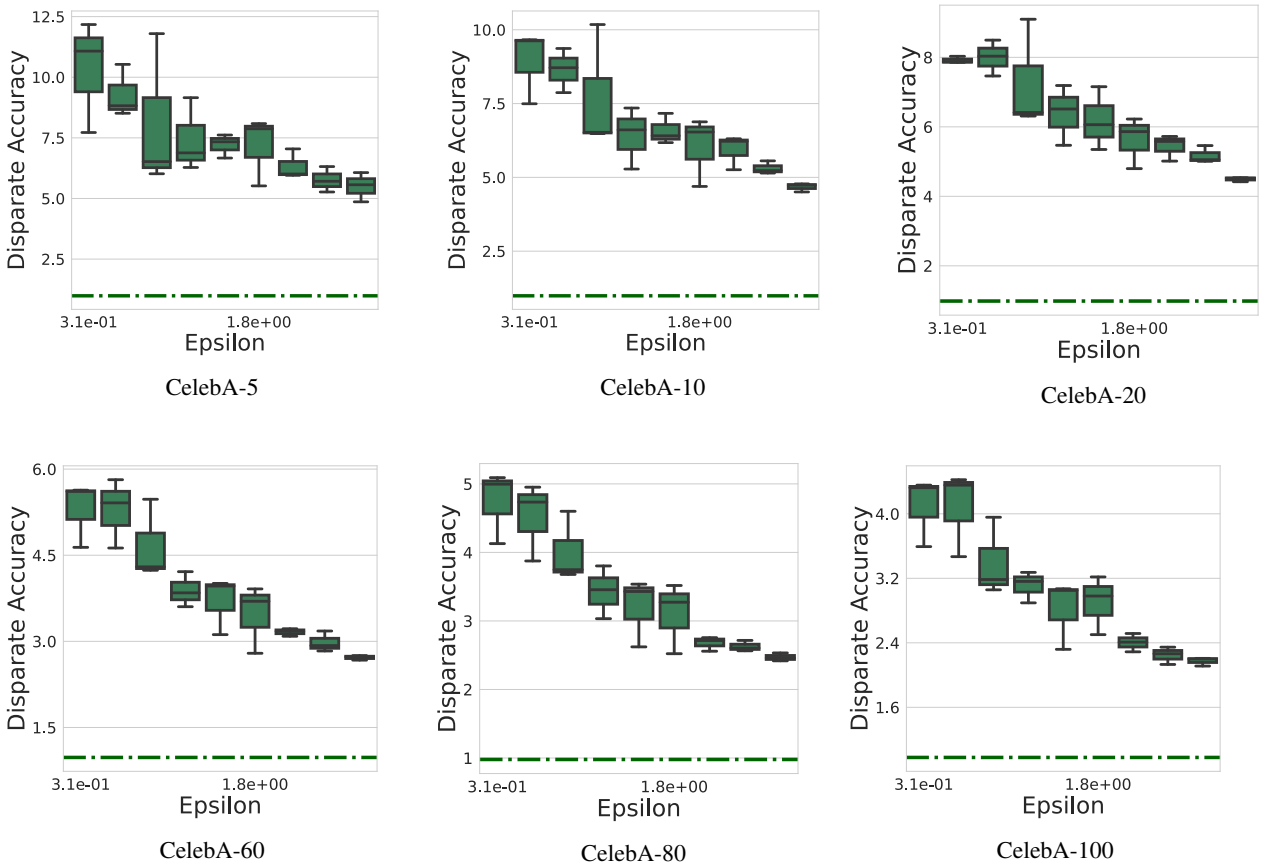

Figure 8: CelebA-x represents a minority majority division where the minority group is comprised of sub-populations whose size is atmost x. Disparate accuracy decreases (low is fairer) with increasing ϵ (high is less privacy) i.e. fairness is worse for stricter privacy for all sub-population sizes. Dashed lines show the corresponding disparate accuracy of a vanilla model (no privacy criterion).

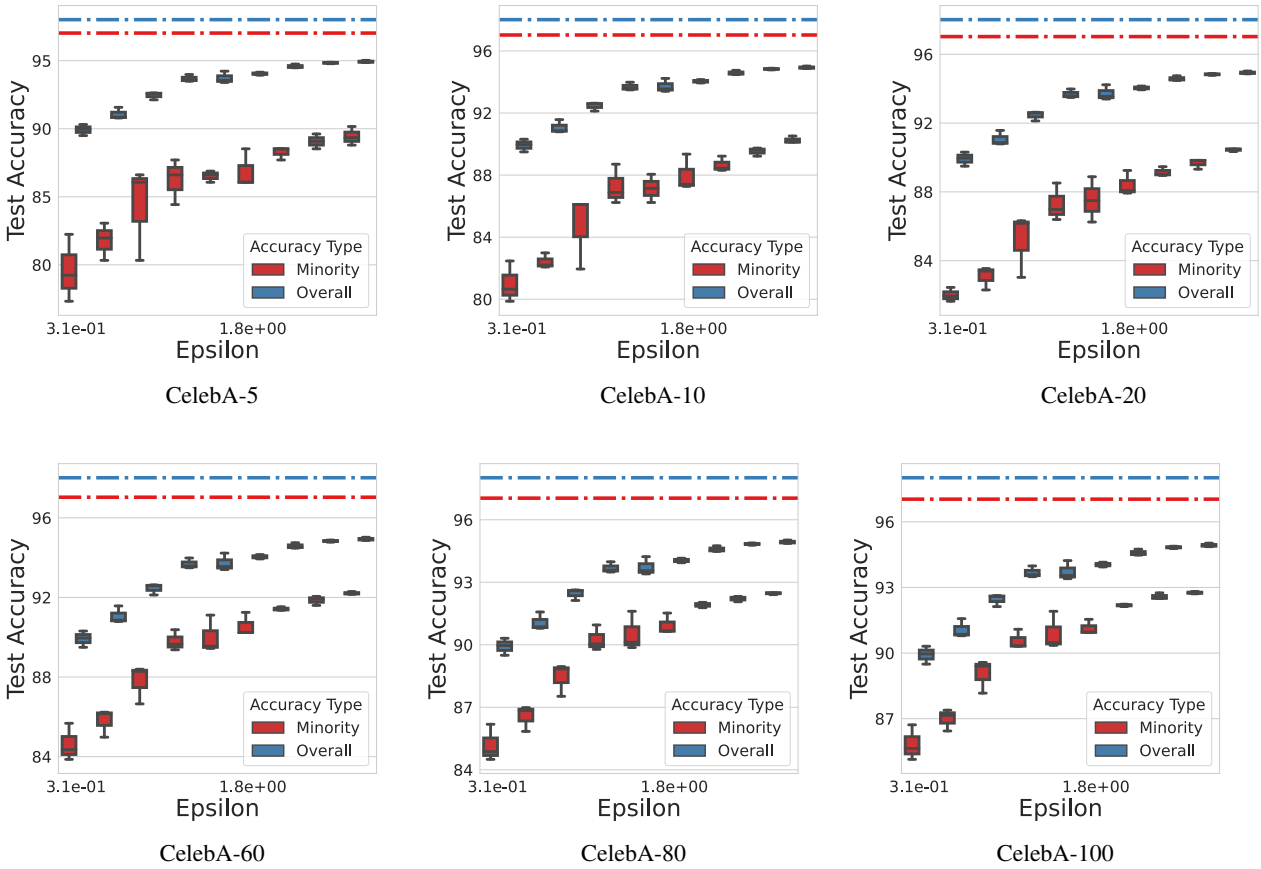

Figure 9: CelebA-x represents a minority majority division where the minority group is comprised of sub-populations whose size is at most x. Both minority and overall accuracy increases with increasing ϵ (high is less privacy) but overall accuracy increases faster. Dashed lines show the corresponding accuracy of a vanilla model (no privacy criterion).

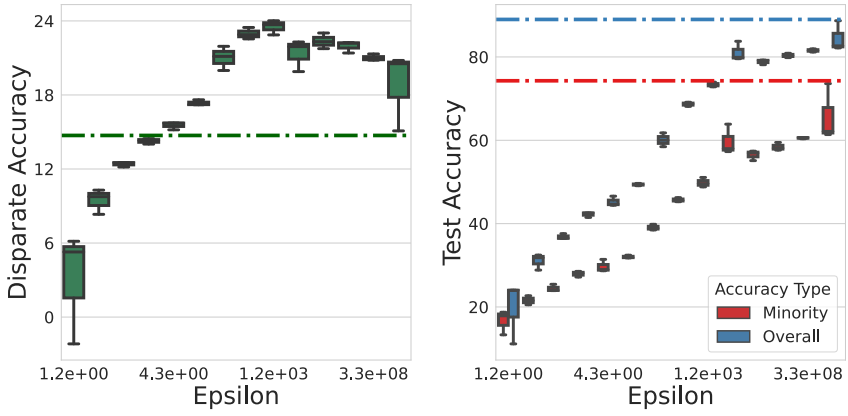

Figure 10: CIFAR10—**Left** Disparate accuracy decreases (low is fairer) with increasing ϵ (high is less privacy) i.e. fairness is worse for stricter privacy. **Right** figure shows that both minority and overall accuracy increases with ϵ. Dashed lines show the corresponding metric of a vanilla model (no privacy criterion).